Facility Location in Evolving Metrics

David Eisenstat, Claire Mathieu, Nicolas Schabanel

Facility Location in Evolving Metrics[†]

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

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

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.

The full version of this article with complete proofs may be found in [2].

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 [6, 8]. These networks evolve over time, and their dynamics have a considerable impact on their structure and effectiveness [7, 4]. 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 [9].

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

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

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).

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 [8], 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 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 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:* $cost \leqslant 8\log(2nT)\cdot \mathrm{LP} \leqslant 8\log(2nT)\cdot \mathrm{OPT}$,
where OPT *is the cost of an optimal solution and* LP *is the value of LP* (1)*, defined at the end of Section 2.*

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 then show (omitted, see [2]) that this approximation ratio is asymptotically optimal, even for a very special case.

Theorem 2 (Hardness for fixed opening cost)**.** *Unless P = NP, there is no* $o(\log T)$*-approximation even for the metric case with only one client and only two possible positions.*

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 [5]. We finally show (omitted) 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.

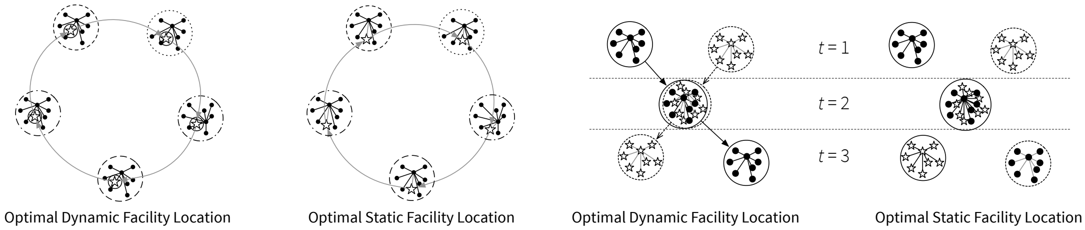
Theorem 3 (Hourly opening cost)**.** *For the dynamic facility location problem with hourly opening cost, there exists a polynomial-time randomized algorithm that, on all inputs, with probability at least* $1/4$*, outputs a solution satisfying:* $cost \leqslant 8\log(2nT)\cdot \mathrm{LP} \leqslant 8\log(2nT)\cdot \mathrm{OPT}$,
where OPT *is the cost of an optimal solution and* LP *is the value of LP (2) in [2].*

2 Facility Location in Evolving Metrics

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 $(d_t)_{1\leqslant t\leqslant T}$ over $F\times C$, a non-negative *facility opening cost* f and a non-negative *client switching cost* g. The goal is to output a subset $A\subseteq F$ of facilities and, for each time step $t\in\{1,\dots,T\}$, an assignment $\phi_t : C\to A$ of facilities to clients, so as to minimize:

$$f\cdot \#A + \sum_{1\leqslant t\leqslant T,\, j\in C} d_t(\phi_t(j),j) + g\cdot \sum_{1\leqslant t<T}\sum_{j\in C} \mathbb{1}\{\phi_t(j)\neq\phi_{t+1}(j)\},$$

that is to say the sum of the opening cost (f for each open facility), of the total *distance cost* to connect each client to its assigned facility at every time step, and of the *switching cost* for each client (g per change of facility per client).


t = 1
t = 2
t = 3
Optimal Dynamic Facility Location
Optimal Static Facility Location
Optimal Dynamic Facility Location
Optimal Static Facility Location

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

(b) Two groups crossing.

Fig. 1: Dynamic versus static Facility Location.

Examples. 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 of each snapshot isolates the five groups and moves the teacher from one group to the next between snapshots; whereas 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 crossing each other (on a street for instance): a static facility location would first output the two groups, then merge them into a single group, then split it into two groups again; whereas a dynamic facility location would keep the same groups for the whole time period, with the same representatives.

Fact 4. *The ratio between the cost of an optimal dynamic facility location solution and the (dynamic) cost of a sequence of optimal static facility location solutions for each snapshot can be as large as* $\Omega(T)$ *and* $\Omega(n)$.

A linear relaxation. For an integer programming formulation, we define indicator 0-1 variables y_i, x_{ij}^t, and z_{ij}^t for $i \in F$, $j \in C$, and $t \in \{1,\ldots,T\}$: $y_i = 1$ iff facility i is open; $x_{ij}^t = 1$ iff client j is connected to facility i at time t; and $z_{ij}^t = 1$ iff client j is connected to facility i at time t but no more at time $t+1$. The dynamic facility location problem is then equivalent to finding an integer solution to the following linear programming relaxation.

$$\left\{\begin{array}{lll} \text{Minimize} & \multicolumn{2}{l}{f \cdot \sum_{i \in F} y_i + \sum_{1 \leqslant t \leqslant T, i \in F, j \in C} x_{ij}^t \cdot d_t(i,j) + g \cdot \sum_{1 \leqslant t < T, i \in F, j \in C} z_{ij}^t} \\ \text{subject to} & (\forall 1 \leqslant t \leqslant T,\ i \in F,\ j \in C) & x_{ij}^t \leqslant y_i \\ & (\forall 1 \leqslant t \leqslant T,\ j \in C) & \sum_{i \in F} x_{ij}^t = 1 \\ & (\forall 1 \leqslant t < T,\ i \in F,\ j \in C) & z_{ij}^t \geqslant x_{ij}^t - x_{ij}^{t+1} \\ & (\forall 1 \leqslant t \leqslant T,\ i \in F,\ j \in C) & y_i, x_{ij}^t, z_{ij}^t \geqslant 0 \end{array}\right. \tag{1}$$

The 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 $\sum_i x_{ij}^t = 1$, one can view $(x_{ij}^t)_i$ as the desired distribution for the facility assigned to client j at timestep t. The **for** loop partitions time, independently for each client j, into intervals during which the distribution $(x_{ij}^t)_i$ remains stable enough, i.e., the distributions $(x_{ij}^t)_i$ *share a large enough common probability mass* during each time interval of the partition. The common probability mass of the distributions $(x_{ij}^t)_i$ during a time interval U is defined as the sum over all facilities i of the minimum probability $\hat{x}_{ij}^U = \min_{t \in U} x_{ij}^t$ 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.

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 $A = \{i \in F : Y_i \leqslant y_i\}$.
for each client $j$ do
  • Partition time greedily into $\ell_j$ intervals $[t^j_k, t^j_{k+1})$ where $\ell_j$ and $(t^j_k)_{k\in[\ell_j+1]}$ are defined as follows: $t^j_1 = 1$, and $t^j_{k+1}$ is defined inductively as the greatest $t \in (t^j_k, T+1]$ such that $\sum_{i\in F}\left(\min_{t^j_k \leqslant u < t} x^u_{ij}\right) \geqslant 1/2$. Let $t^j_{\ell_j+1} = T+1$.
  • For each time interval $U = [t^j_k, t^j_{k+1})$, assign client $j$ to argument of $\min_{i\in F}(Y_i/\hat{x}^U_{ij})$, where $\hat{x}^U_{ij} = \min_{u\in U} x^u_{ij}$.
end for
```

Conclusion. A 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., [1] 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 [3].

References

[1] M. Charikar and R. Panigrahy. Clustering to minimize the sum of cluster diameters. In *STOC*, pages 1–10, 2001.

[2] David Eisenstat, Claire Mathieu, and Nicolas Schabanel. Facility location in evolving metrics. In *ICALP*, 2014. To appear.

[3] 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).

[4] J. M. Kleinberg. The small-world phenomenon and decentralized search. *SIAM News*, 37(3), 2004.

[5] Shi Li. A 1.488-approximation algorithm for the uncapacitated facility location problem. In *ICALP*, pages 77–88, 2011.

[6] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[7] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, 2001.

[8] 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.

[9] C. Tantipathananandh, T. Y. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD*, pages 717–726, 2007.