Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Clustering through Continuous Facility Location Problems[☆]

Luis A.A. Meira[a,*], Flávio K. Miyazawa[b], Lehilton L.C. Pedrosa[b,*]

[a] *School of Technology, University of Campinas, Brazil*
[b] *Institute of Computing, University of Campinas, Brazil*

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form 21 May 2016
Accepted 3 October 2016
Available online xxxx
Communicated by T. Erlebach

Keywords:
Continuous Facility Location Problem
k-Clustering
Approximate center sets
Random sampling procedure
Approximation algorithms

ABSTRACT

We consider the Continuous Facility Location Problem (ConFLP). Given a finite set of clients $C \subset \mathbb{R}^d$ and a number $f \in \mathbb{R}_+$, ConFLP consists in opening a set $F' \subset \mathbb{R}^d$ of facilities, each at cost f, and connecting each client to an open facility. The objective is to minimize the costs of opening facilities and connecting clients. We reduce ConFLP to the standard Facility Location Problem (FLP), by using the so-called approximate center sets. This reduction preserves the approximation, except for an error ε, and implies $1.488 + \varepsilon$ and $2.04 + \varepsilon$-approximations when the connection cost is given by the Euclidean distance and the squared Euclidean distance, respectively. Moreover, we obtain approximate center sets for the case that the connection cost is the αth power of the Euclidean distance, achieving approximations for the corresponding problems, for any $\alpha \geq 1$. As a byproduct, we also obtain a polynomial-time approximation scheme for the k-median problem with this cost function, for any fixed k.

1. Introduction

In the Facility Location Problem (FLP), given a set of clients and a set of locations, one must select a subset of locations, where to install facilities, and connect each client to an open facility. The objective is to find the solution that minimizes the costs of opening facilities, and serving clients. The Continuous FLP (ConFLP) is the variant in which a facility may be opened at any point of the Euclidean space, whilst in FLP, the set of possible locations is a given finite set. ConFLP has applications in clustering, and is closely related to k-means and k-median problems. Contrary to k-clustering, however, it does not impose a number k of clusters; instead, it fixes a penalty f for each created cluster.

The classical FLP is defined as follows. Consider finite sets C and F, that represent *clients* and *facilities*, respectively. The cost to connect each facility i to each client j is c_{ij}, and the cost to open a facility i is f_i. The objective is to find a subset F' of F such that $\sum_{i \in F'} f_i + \sum_{j \in C} \min_{i \in F'} c_{ij}$ is minimum. Hochbaum [17] presented a $O(\log n)$-approximation for FLP, that is known to be asymptotically tight unless $\mathrm{P} = \mathrm{NP}$, by using a reduction from Set Cover. The so-called metric FLP is the particular case in which $C \cup F$ is a metric space with distance c, i.e., c is symmetric and satisfies the triangle inequality. For this case, the best known approximation factor is 1.488 [26], and there is no factor better than 1.463, unless $\mathrm{NP} \subseteq \mathrm{DTIME}[n^{O(\log \log n)}]$ [13]. For the special case that clients and facilities are in the plane, and c is the Euclidean distance, Arora et al. [2] obtained a polynomial-time approximation scheme (PTAS). Kolliopoulos and Rao [23] obtained PTASs when points are in the d-dimensional space, for any fixed d.

[☆] This work was partially supported by FAPESP (grant 2010/20710-4) and CNPq (grants 477692/2012-5, and 311499/2014-7).

[*] Corresponding authors.
E-mail addresses: meira@ft.unicamp.br (L.A.A. Meira), fkm@ic.unicamp.br (F.K. Miyazawa), lehilton@ic.unicamp.br (L.L.C. Pedrosa).

http://dx.doi.org/10.1016/j.tcs.2016.10.001
0304-3975/© 2016 Published by Elsevier B.V.

When the connection cost function is not metric, there are few works in the literature of FLP. For the variant whose connection cost is the square of the Euclidean distance, Jain and Vazirani [21] presented a 9-approximation. This factor holds for the more general case in which clients and facilities are in a metric space, and the connection cost is the square of the metric function. This variant was studied by Fernandes et al. [11], who gave a 2.04-approximation, and showed that this factor is tight unless $\mathrm{P} = \mathrm{NP}$ [11]. Fernandes et al. also obtained approximations for the case that c is the power of a metric function [12].

The continuous FLP is defined as follows. Consider a finite set of points $C \subset \mathbb{R}^d$, and a number $f \in \mathbb{R}_+$. A facility may be opened at any point of $\mathbb{R}^d$ at cost f. The cost to connect a facility $i \in \mathbb{R}^d$ to a client $j \in C$ is given by the Euclidean distance between i and j, and is denoted by $\ell_2(i, j)$. The objective is to find a finite set $F' \subset \mathbb{R}^d$, such that $|F'|\, f + \sum_{j \in C} \min_{i \in F'} \ell_2(i, j)$ is minimum. This problem is denoted by ℓ_2-ConFLP. More generally, for $\alpha \geq 1$, we denote by ℓ_2^α-ConFLP the variant whose connection cost is given by the αth power of the Euclidean distance, ℓ_2^α. Work on ConFLP has been done by Meira and Miyazawa [30], who obtained approximations with factors $1.861 + \varepsilon$ and $9 + \varepsilon$ for the ℓ_2 and ℓ_2^2 cost functions, respectively. Later, Czumaj et al. [7] obtained a PTAS for the special case with ℓ_2 cost function and points in the plane.

1.1. Other related works

In the *k-means* problem, given a set P of n points in $\mathbb{R}^d$, the objective is to choose a set $K \subset \mathbb{R}^d$ of k cluster centers, such that the sum of the squared Euclidean distance from every point to its nearest center is minimized. The clustering obtained from k-means is the Voronoi partition of P with respect to K. Lloyd's algorithm [27] is a largely used heuristic to solve k-means. Although it has no guarantee of either execution time nor solution quality, it has been proved useful in practice. The k-means problem is NP-hard, even if k, or d is a fixed constant [8,28]. Inaba et al. [18] showed that the number of Voronoi partitions induced by k points in $\mathbb{R}^d$ is n^{dk}, and therefore k-means can be solved exactly in $O(n^{dk+1})$. Faster algorithms for k-means are given as polynomial-time approximation schemes when both k and d are fixed [29,16], and when k is fixed but d is part of the instance [9,25,6,10]. When both d and k are part of the instance, Kanungo et al. [22] used a local search to give a $9 + \varepsilon$-approximation.

The *k-median* problem is similar to k-means, but the cost to connect two points is the standard Euclidean distance. For this problem there exist polynomial-time approximation schemes for the case that k and d are fixed [2,23,16], and for the case that d is part of the input [25,6]. When k is part of the input and $d = \Omega(\log n)$, obtaining a PTAS to k-median is NP-hard [14]. The *metric k-median* problem is a widely studied variant in which the sets of clients and candidate centers are elements of a metric space. This problem is known to be 1.736-hard to approximate, unless $\mathrm{NP} \subseteq \mathrm{DTIME}[n^{O(\log \log n)}]$ [19], and currently the best approximation is due to Byrka et al. [5] and has factor 2.611.

In the approximation schemes for k-means, Matousek [29] used the concept of approximate *centroid sets*. The term centroid comes from the fact that the optimal center of a given cluster is given by the centroid of its points. The idea is to obtain a set of candidate centers, an *approximate center set*, that contains a subset of k centers that is an approximate solution for the optimal k-means clustering. Matousek gave ε-approximate center sets of size $n\varepsilon^{-1} \log(1/\varepsilon)$. Other constructions of center sets have sizes $k\varepsilon^{-2d} \log n \log 1/\varepsilon$ [16] and, independent of n, $k^3 \varepsilon^{-(2d+2)} \log 1/\varepsilon$ [15]. Also, center sets independent of d with size $n^{1/\varepsilon}$ are implicitly given by Inaba et al. [18]. For k-median, there are approximate center sets with size $k^2 \varepsilon^{-2d} \log^2 n$ [16].

Kumar et al. [25] formalized the existence of a *random sampling procedure* for certain k-clustering problems, and used a superset sampling algorithm to obtain linear-time approximations algorithms for k-means and k-median problems. The idea is that, for certain clustering problems, one may obtain, with constant probability, an approximate center set from a given set of points using only a constant-size subset. Ackermann et al. [1] extended the algorithm of Kumar et al., using a combinatorial analysis, and avoiding the need of a triangle inequality, therefore obtaining linear-time approximation schemes for k-clustering problems with different cost functions, such as Kullback–Leibler divergence and Mahalanobis distances.

1.2. Our contribution

In this work, we solve ConFLP problems using approximate center sets for its set of clients, and solving the discrete version of the problem. This reduction leads to approximations for ConFLP problems, provided that there exists a procedure to create ε-approximate center sets for the considered connection cost function, and that there exists an approximation algorithm for the corresponding discrete facility location problem. Using this reduction, we obtain $1.488 + \varepsilon$ and $2.04 + \varepsilon$-approximations for the ℓ_2 and ℓ_2^2 cost functions when d is part of the input, respectively, and a PTAS for the ℓ_2 cost function for fixed d.

A natural way to solve ℓ_2-ConFLP and ℓ_2^2-ConFLP is to execute the corresponding discrete k-clustering problems on the metric for all values of k. Such approach cannot improve the approximation factors, as there is a lower bound on the approximation factor of 1.735 for metric k-median [19], and a lower bound of 3.943 for squared metric k-median (Subsection 3.3).

We also obtain ε-approximate center sets of polynomial size for the more general ℓ_2^α cost function, thus obtaining approximation algorithms for the corresponding ConFLP problems, for any $\alpha \geq 1$. Using such center sets, we also obtain a PTAS for the k-clustering problems with ℓ_2^α cost function, for fixed k.