

Fast k -clustering Queries on Road Networks

James W. McClain*, Piyush Kumar†
Department of Computer Science
Florida State University
Tallahassee, Florida
mcclain@cs.fsu.edu*, piyush@cs.fsu.edu†

Abstract—In this article, we study the k -clustering query problem on road networks, an important problem in Geographic Information Systems. Using Euclidean embeddings and reduction to fast nearest neighbor search, we devise approximation algorithms for these problems. Since these problems are difficult to solve exactly – and even hard to approximate for most variants – we compare our constant factor approximation algorithms to exact answers on small synthetic datasets and on a dataset representing Tallahassee, Florida, a small city. We have implemented a web application that demonstrates our method for road networks in the same small city.

Keywords— k -clustering, k -means, k -medians, k -centers, embeddings, Computational Geometry, GIS.

I. Introduction

Clustering entails partitioning a set of points into smaller subsets by optimizing a given objective function. In this paper we discuss our study of the k -clustering query problem on road networks. Let the sets Q and P both be finite collections of locations (say, longitude and latitude) on a given road network. The goal is to find, given a set of query points Q , a set $C \subset P$ of size k such that $g_Q(C)$ is minimized.

Here $g_Q(C)$ is an objective function that assigns a real value to a candidate solution C . In our formulation, centers can only be located on the nodes of the network. To concretize and better-understand the general problem that we wish to discuss, we will start by giving some realistic examples of specific 1-clustering problems.

The 1-center problem on road networks is the problem of computing a ball of minimum radius enclosing a given set of locations on the network. The measurements on such networks are done using the shortest path metric. For this problem, the objective function is $g_Q(x) = \max_{q \in Q} d(x, q)$ where x is used instead of C to emphasize the fact that in this instance $|C| = 1$. $d(x, q)$ is used to denote the shortest path from x to q . The ability to solve this problem allows one to determine on which intersection to build a fire station so that its distance from the furthest building in its area of responsibility

is minimized.

A simple variant of this problem emerges when the “max” in the objective function is replaced by a “ \sum ” to yield $g_Q(x) = \sum_{q \in Q} d(x, q)$. That corresponds to the 1-median problem. If a set of people on the road network want to meet at a restaurant, being able to solve this problem will allow them to choose a meeting place such that the total distance traveled by the group is minimized.

Another problem, the 1-mean problem, is represented by the objective function $g_Q(x) = \sum_{q \in Q} d(x, q)^2$. The 1-mean is the “average” location of the query points Q .

The examples given above are instances of k -means, k -medians, and k -centers where the answer is constrained ($C \subset P$) and $k = 1$. How to solve similarly constrained k -means, k -medians, and k -centers problems with $k > 1$ is the question that we examine in this article. Here, we are particularly focused on k -means but will give some brief discussion of the other two, as well. Neither k -means, nor k -medians, nor k -centers is known to be solvable in polynomial time and most variants of these problems are NP-hard.

An example of the constrained k -means problem is depicted in Figure 1. The query set Q contains ten points and the set P contains all of the street intersections in the road network. The problem is to minimize $g_Q(C) = \sum_{q \in Q} d(C, q)^2$, where $d(C, q) = \min_{c \in C} d(c, q)$.

Chatterjee et.al. [1] have recently shown that the shortest path metric on road networks, $d(\cdot)$, can be embedded in Euclidean spaces with low distortions. In this paper, we follow the same approach and embed the road network in Euclidean space using multi dimensional scaling with the Sammon criterion [2]. Hence, in the rest of the paper, we use the Euclidean norm, $\|\cdot\|$, for measuring distances instead of the shortest path metric $d(\cdot)$. This paper is a generalization of the results in [1] which concentrates on only the 1-center problem.

The rest of the paper is organized as follows: We prove in Section II that our scheme offers a constant-factor approximation for k -means and we continue in

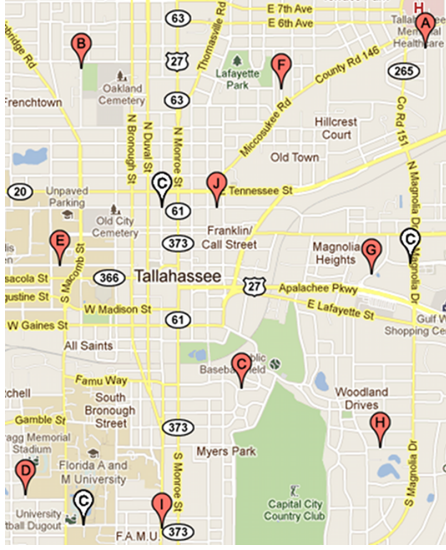


Figure 1: An example of k -means on a road network. In this example, $|Q| = 10$ and $k = 3$.

§III with a discussion of how to prove that our strategy provides constant-factor approximations for different g_Q . In particular we focus on functions g_Q which are positively related with the distance function and for which the gradient $\nabla g_Q(C)$ exists piecewise in a particular way. Both of those nebulous criteria are made specific in Section III. We provide experimental results in Section IV. The rest of this section is spent describing the algorithm and introducing notation (Section I-A) and briefly summarizing previous and related work (§I-B).

§I-A. ALGORITHM AND NOTATION

Algorithm 1 Aggregate Nearest Neighbor Algorithm

- Require:** k and $Q, P \subset \mathbb{R}^n$, $|Q|, |P| < \infty$
- 1: $C' \leftarrow S$ where $g_Q(S)$ is minimal with respect to sets $S \subset \mathbb{R}^n$, $|S| = k$
 - 2: $C \leftarrow \cup_i a_i$ where $a_i \in P$, $a_i = \text{nearest_neighbor}(b_i, P)$, and $C' = \cup_i b_i$
 - 3: **return** C
-

The very simple algorithm (or recipe for algorithms) can be found in Algorithm 1. First, compute the unconstrained set $C' \subset \mathbb{R}^n$ for which $g_Q(C')$ is minimal. We will say that $C' = \cup_i b_i$. The output $C \subset P$ is computed by unioning the set of points $a_i \in P$ such that a_i is the (not necessarily unique, but arbitrarily chosen if it is not) point closest to $b_i \in C'$. Thus $C = \cup_i a_i$.

Note that it is possible that $|C|$ may be less than k . We will show in Section II-D that if one adds the constraint $|C| = k$ to this type of algorithm, then the algorithm

is not guaranteed to return a constant-factor approximation. This is not of much consequence, however, because if a set of size k is an absolute requirement, one can compute C then add random elements of P to it as needed without damaging the approximation factor – at least if it can be said of g_Q that $g_Q(C \cup x) \leq g_Q(C)$ (this is expected if g_Q is defined in terms of distances from the input to the set Q .)

In addition to the symbols defined above, we will use \hat{C} to refer to the set which minimizes $g_Q(\hat{C})$ subject to the constraints $\hat{C} \subset P$ and $|\hat{C}| = k$ with $\hat{C} = \cup_i c_i$. The output C from the algorithm is an approximation of the true solution, \hat{C} . Finally, we will use $Q_{b_i} := \mathcal{V}(b_i) \cap Q$ and $Q_{c_i} := \mathcal{V}(c_i) \cap Q$ to refer to Voronoi cells with members in Q relative to the collections of centers C' and \hat{C} , respectively. Said differently, Q_{b_i} is those members of Q which are closer to b_i than they are to any other member of C' , and similarly for Q_{c_i} .

Initially, we will be assuming that the clustering subroutine which produces C' gives exact results and that the nearest-neighbor subroutine which produces C from C' does the same. Those assumptions will be removed in Section II-E.

§I-B. PREVIOUS AND RELATED WORK

This algorithm is not new here it was described in [3; 1] in the context of the 1-center problem and in context of the 1-median and 1-center problem in [4]. The primary contribution of this article is to demonstrate the algorithm gives a constant-factor approximation for k -clustering with $k > 1$, and hence, show that this simple algorithm is practical for the $k > 1$ case. In [4], the authors give a proof that the approximation factor for the algorithm in the 1-centers case is $\sqrt{2}$ and that this bound is tight. An approximation bound of 3 for the 1-median problem is reported in [3]. These are the only two published results for approximation bounds on this family of algorithms of which we are aware.

The study of aggregate nearest neighbor queries was initiated in [5] and the k -median waterfront was more-or-less covered in that work. This continued in a more general context in [6] and [7]. The question was also considered in [8].

The problem of clustering of data sets relative to various g is well-studied, and our algorithm requires such a clustering algorithm as a subroutine. Notable algorithms for the k -medians problem are given in [9] and [10]. Many of the best of these are based on the idea of core-sets which was introduced in [11]. More general thoughts on the subject of clustering were presented in [12]. We also relied heavily on chapter four of [13] for the constant-factor approximation algorithms which were used as subroutines of Algorithm 1 in the experiments described in Section IV.

II. k -means

In this section, we show that the algorithm provides a constant-factor approximation for the k -means problem. In the first subsection we show that the algorithm actually provides an exact answer when $k = 1$. We then reuse part of that computation to prove our claim for $k > 1$.

§ II-A. RATIO FOR $k = 1$

To show that the algorithm is exact for the 1-mean problem, we first compute the gradient of the function $g_Q(x) = \sum_{q \in Q} \|x - q\|^2$, the objective function for the problem. We find that isoshapes of g_Q are spheres centered at \bar{q} , and conclude from this that the algorithm is exact.

Lemma 1. *If $f(x) := \sum_{q \in Q} \|x - q\|^2$ then $\nabla f(x) = 2|Q|(x - \bar{q})$.*

Proof: If $f(x) := \sum_{q \in Q} \|x - q\|^2$ then:

$$\begin{aligned} \frac{\partial}{\partial x_i} f(x) &= \frac{\partial}{\partial x_i} \sum_{q \in Q} \left(\sum_j (x_j - q_j)^2 \right) \\ &= \frac{\partial}{\partial x_i} \sum_{q \in Q} (x_i - q_i)^2 \\ &= \frac{\partial}{\partial x_i} \sum_{q \in Q} (x_i^2 - 2x_i q_i + q_i^2) \\ &= 2 \left(\sum_{q \in Q} x_i - \sum_{q \in Q} q_i \right) \end{aligned}$$

So that

$$\nabla f(x) = 2|Q|(x - \bar{q})$$

Corollary 1. *The algorithm is exact for $k = 1$.*

Proof: For this problem, $g_Q(x) = f(x)$. Lemma 1 shows that the sets of x values satisfying $f(x) = c$ and $\|x - \bar{q}\| = c'$ are spherical, centered at \bar{q} , and identical for some constants c and c' . That means that $p \in P$ has minimal $f(p)$ if and only if $\|p - \bar{q}\|$ is also minimal. ■

§ II-B. RATIO FOR $k > 1$

Say that $f_i(x) := \sum_{q \in Q_{b_i}} \|x - q\|^2$ is the sum of squares value for some center x chosen for the subset Q_{b_i} . That would mean that $g_Q(C) = \left(\sum_{c \in C} f_{i(c)}(c) \right)$ where $i(c)$ is an index such that $c \in \mathcal{V}(b_{i(c)})$. Let $ALG^* := \sum_i f_i(b_i)$ be the sum-of-squares value for the unconstrained choice of means made by the clustering subroutine. Let $ALG := \sum_i f_i(a_i)$ be the sum-of-squares value for the choice of means made by the algorithm (the set $C \subset P$). Finally, let $OPT := \sum_i \sum_{q \in Q_{b_i}} \|x - q\|^2$ be the sum of squares value

for the best possible choice of means (the set $\hat{C} \subset P$). We need to understand the ratio $\frac{ALG}{OPT}$.

Lemma 2. $\frac{ALG}{OPT} \leq 1 + \frac{\sum_i |Q_{b_i}| \|a_i - b_i\|^2}{|Q|}$

Proof:

$$\begin{aligned} ALG &:= \sum_i f_i(a_i) \\ &= \sum_i \left(f_i(b_i) + \int_{b_i}^{a_i} \|\nabla f_i(x)\| dx \right) \end{aligned} \quad (II.1)$$

$$= \sum_i f_i(b_i) + \sum_i \int_0^{a_i - b_i} 2|Q_{b_i}| \|x\| dx \quad (II.2)$$

$$\begin{aligned} &= \sum_i f_i(b_i) + \\ &\quad \frac{1}{2} \sum_i 2|Q_{b_i}| \|a_i - b_i\| \|a_i - b_i\| \end{aligned} \quad (II.3)$$

$$= ALG^* + \sum_i |Q_{b_i}| \|a_i - b_i\|^2$$

where (II.1) comes from writing the function as the integral of its derivative, (II.2) comes from rewriting the integral with the help of Lemma 1, and (II.3) comes from thinking of the integral as the area under a curve. We also have the inequality

$$OPT \geq ALG^*$$

which is true because the unconstrained k -means chosen by the subroutine have an objective value no more than the optimal answer which is constrained. Hence:

$$ALG \leq OPT + \sum_i |Q_{b_i}| \|a_i - b_i\|^2$$

Scaling the problem so that $OPT = |Q|$ and dividing through gives the desired ratio. ■

§ II-C. UNIQUE a_i

If the a_i are required to be unique, then there is no bound on the approximation provided by the algorithm. One example is the following: $k = 3$, a Q composed of two clusters of two points each, and a P composed of three points, two in the vicinity of one cluster and one near the other. Imagine that the cluster in Q with only one member of P near it is more dispersed than the other cluster. In such a circumstance, the more dispersed cluster of Q would be split into two by the clustering subroutine (because $k = 3$) and a half of it matched with a $p \in P$ close to the opposite cluster. This happens because the subroutine which clusters Q does not consider P so it has no way of knowing how harmful this choice of clusters is.

Because the clusters can be arbitrarily far apart, there is no upper bound on $\|a_i - b_i\|^2$ and therefore the

approximation factor provided by the algorithm has no upper bound. Please see Figure 2 for a visual rendering. In that figure, the circular items are members of Q and the squares are members of P . The colored patterns indicate which cluster each item is associated with.



Figure 2: Splitting of a cluster.

Similar arguments can be made for any function g_Q defined in terms of distance from Q . For that reason, in all discussion of the algorithm from this point on, we will assume that the elements a_i need not be unique, that $|C| < k$ is allowed.

§ II-D. NON-UNIQUE a_i

If the a_i are not required to be unique (that is, a $p \in P$ can be matched to more than one b_i), then the quality of the approximation can be described by a constant factor. The Q_{b_i} partition Q , but in order to go forward it is helpful to partition each Q_{b_i} . We will use the definition $Q_{ij} := Q_{b_i} \cap Q_{c_j}$. If that is the case, then $Q_{b_i} = \cup_{j=1 \dots k} Q_{ij}$ and $|Q_{b_i}| = \sum_{j=1 \dots k} |Q_{ij}|$. The mean of each Q_{ij} is called \bar{q}_{ij} .

We will also need to define α_{ij} , β_{ij} , and γ_{ij} such that $\sum_{ij} \alpha_{ij} \leq 1$, $\sum_{ij} \beta_{ij} := 1$, and $\sum_{ij} \gamma_{ij} := 1$. Q_{ij} contributes $\alpha_{ij}|Q|$ to ALG^* , Q_{ij} contributes $\beta_{ij}|Q|$ to OPT , and $|Q_{ij}| = \gamma_{ij}|Q|$.

Recall that previously we scaled the problem so that $\text{OPT} = |Q|$. We are using that in the definitions in the previous paragraph and will continue to do so for the balance of the article.

Lemma 3. *Given the definitions above, $\|\bar{q}_{ij} - b_i\| \leq \sqrt{\frac{\alpha_{ij}}{\gamma_{ij}}}$ and similarly, $\|\bar{q}_{ij} - c_j\| \leq \sqrt{\frac{\beta_{ij}}{\gamma_{ij}}}$.*

Proof: If $\|\bar{q}_{ij} - b_i\| > \sqrt{\frac{\alpha_{ij}}{\gamma_{ij}}}$, then

$$\begin{aligned} \|\bar{q}_{ij} - b_i\|^2 &> \frac{\alpha_{ij}}{\gamma_{ij}} \\ \Rightarrow \|\bar{q}_{ij} - b_i\|^2 &> \alpha_{ij} \frac{|Q|}{|Q_{ij}|} \\ \Rightarrow |Q_{ij}| \|\bar{q}_{ij} - b_i\|^2 &> \alpha_{ij} |Q| \\ \Rightarrow \sum_{x \in Q_{ij}} \|x - b_i\|^2 &> \alpha_{ij} |Q| \end{aligned}$$

which is a contradiction. ■

Lemma 4. *Given the definitions and lemma above, $\sum_i |Q_{b_i}| \|a_i - b_i\|^2 \leq 4|Q|$.*

Proof: We know that $\|a_i - b_i\| \leq \min_j \|b_i - c_j\|$ because b_i is a nearest-neighbor of a_i in P and $c_j \in P$. That allows us to say

$$\begin{aligned} \sum_i |Q_{b_i}| \|a_i - b_i\|^2 &\leq \sum_i |Q_{b_i}| \min_j (\|b_i - \bar{q}_{ij}\| + \|\bar{q}_{ij} - c_j\|)^2 \quad (\text{II.4}) \end{aligned}$$

$$\leq \max_{\alpha, \beta, \gamma} \sum_i |Q_{b_i}| \min_j \left(\sqrt{\frac{\alpha_{ij}}{\gamma_{ij}}} + \sqrt{\frac{\beta_{ij}}{\gamma_{ij}}} \right)^2 \quad (\text{II.5})$$

$$= \max_{\alpha, \beta, \gamma} \sum_i |Q_{b_i}| \min_j \left(\left(\alpha_{ij} + \beta_{ij} + 2\sqrt{\alpha_{ij}\beta_{ij}} \right) \frac{1}{\gamma_{ij}} \right)$$

$$= |Q| \max_{\alpha, \beta, \gamma} \sum_i \left(\sum_j \gamma_{ij} \right)$$

$$\min_j \left(\left(\alpha_{ij} + \beta_{ij} + 2\sqrt{\alpha_{ij}\beta_{ij}} \right) \frac{1}{\gamma_{ij}} \right)$$

where (II.4) comes from the triangle inequality and (II.5) comes from Lemma 3.

Each term of $\sum_j \gamma_{ij}$ is multiplied with a minimum over j , so we can get rid of the minimum and say

$$\begin{aligned} \sum_i |Q_{b_i}| \|a_i - b_i\|^2 &\leq |Q| \max_{\alpha, \beta, \gamma} \sum_{ij} \left(\gamma_{ij} \left(\alpha_{ij} + \beta_{ij} + 2\sqrt{\alpha_{ij}\beta_{ij}} \right) \frac{1}{\gamma_{ij}} \right) \\ &= |Q| \max_{\alpha, \beta} \sum_{ij} \left(\alpha_{ij} + \beta_{ij} + 2\sqrt{\alpha_{ij}\beta_{ij}} \right) \\ &= 4|Q| \end{aligned}$$

which is the desired bound. ■

Notice that throughout the analysis, we have used a definition of ALG which is based on the clusters of Q_{b_i} rather than the more sensible Q_{a_i} . This choice was made for convenience, but does not invalidate the bound. In fact, if one were to compute the sum-of-squares where the clustering was done with respect to a_i rather than b_i , ALG would only get smaller.

Given Lemma 2 which establishes the ratio in symbolic form and lemma 4 which bounds the numerator of the second term of the ratio, it is possible to state the following theorem:

Theorem 1. *The approximation factor provided by the algorithm is $\frac{\text{ALG}}{\text{OPT}} \leq 5$.*

§ II-E. APPROXIMATE VERSION

Until this point, we have been assuming that the subroutine that computes C' and the subroutine which

finds the nearest neighbors in P to each b_i in Algorithm 1 are both exact. The second assumption might be possible, but in general the first assumption is not realistic. In this section, we will remove both of those assumptions and see what the effect is on the approximation factor provided by the algorithm. We will need to modify Lemma 2, Lemma 3, and Lemma 4.

If the clustering subroutine provides a C' which is a $(1 + \epsilon)$ approximation instead of an exact answer, we can no longer say $\text{OPT} \geq \text{ALG}^*$ in Lemma 2 as we did before, but instead must say that $(1 + \epsilon)\text{OPT} \geq \text{ALG}^*$. We also have to change $\sum_{ij} \alpha_{ij} \leq 1$ to $\sum_{ij} \alpha_{ij} \leq (1 + \epsilon)$ in Lemma 3. Also, if the available nearest neighbor subroutine returns a $(1 + \delta)$ approximation instead of an exact answer, we will need to multiply our bound on $\|a_i - b_i\|$ by that factor in Lemma 4. All of that implies the following:

Theorem 2. *If the clustering subroutine provides a $(1 + \epsilon)$ approximate answer and the nearest neighbor subroutine provides a $(1 + \delta)$ approximate answer, then*

$$\frac{\text{ALG}}{\text{OPT}} \leq (1 + \epsilon) + (1 + \delta)^2 (\epsilon + 2 + 2\sqrt{1 + \epsilon})$$

III. Generalizing to other functions g_Q

It is possible to use the same type of analysis used above for similar problems. We will be using the k -medians problem and k -centers problems throughout this section as positive and negative examples of how to extend the ideas presented in Section II. Because the proof for k -medians problem almost exactly follows a pattern suggested by that of the k -means problem, we will not present complete proofs for everything said here.

§ III-A. PIECEWISE CONTINUITY OF THE GRADIENT

The gradient ∇g_Q of the objective function should be continuous on the sets Q_{b_i} . It is this property that allows the statement $\text{ALG} \leq \text{ALG}^* + \sum_i \int_{b_i}^{d_i} \|\nabla f_i(x)\| dx$ to be made which is important for being able to establish a bound on the ratio as is done in Lemma 2. That is convenient to be able to do, because $(1 + \epsilon)\text{OPT} \geq \text{ALG}^*$ immediately removes the first term as a concern and allows one to relate $\|a_i - b_i\|$ to ALG (this is because the integral can be approximated by an area with one of its sides having length $\|a_i - b_i\|$.) That, in turn, is potentially helpful for relating ALG to OPT.

It can be seen that the g_Q corresponding to the k -centers problem does not enjoy this property, so k -centers is not well suited for this pattern that we are discussing. The k -median problem, by contrast, works well within this scheme. For that problem, we can define $f_i(x) := \sum_{y \in Q_{b_i}} \|x - y\|$, $\text{ALG} := \sum_i f_i(a_i)$, and

$g_Q = \left(\sum_{c \in C} f_i(c) \right)$ as in Section II-B. It can be seen that $\nabla f_i = \sum_{q \in Q_{b_i}} \frac{x - q}{\|x - q\|}$ which is the equivalent in this context of Lemma 1.

All of that eventually gives

$$\text{ALG}_{\text{Medians}} \leq (1 + \epsilon) + \sum_i |Q_{b_i}| (1 + \delta) \|a_i - b_i\|$$

for the k -medians problem.

§ III-B. POSITIVE RELATIONSHIP WITH THE DISTANCE FUNCTION

We also wrote earlier about the need for g_Q to have a positive relationship with distance. What we meant by this is that we need an analogue of Lemma 3. Even though ∇g_Q for the k -centers problem is not acceptable as we saw in the last subsection, the criterion of this subsection still applies to it. We can say that

$$\|b_i - q\| \leq (1 + \epsilon)\text{OPT}_{\text{Centers}}$$

and

$$\|q - a_j\| \leq (1 + \delta)\text{OPT}_{\text{Centers}}$$

for $q \in Q_{ij}$. This serves the same purpose for k -centers as Lemma 3 does for k -means.

Lemma 3 goes through almost unchanged for k -medians. The result is that we have $\|\bar{q}_{ij} - b_i\| \leq \frac{\alpha_{ij}}{\gamma_{ij}}$ and $\|\bar{q}_{ij} - c_j\| \leq \frac{\beta_{ij}}{\gamma_{ij}}$. Those can be used in an equivalent of Lemma 4.

§ III-C. PUTTING IT ALL TOGETHER

For the k -centers problem, the bounds on $\|b_i - q\|$ and $\|q - a_j\|$ can be combined together to say

$$\frac{\text{ALG}_{\text{Centers}}}{\text{OPT}_{\text{Centers}}} \leq (2 + \epsilon + \delta)$$

in a way similar to the process used in Lemma 4. For the k -medians problem, the bounds $\|\bar{q}_{ij} - b_i\| \leq \frac{\alpha_{ij}}{\gamma_{ij}}$ and $\|\bar{q}_{ij} - c_j\| \leq \frac{\beta_{ij}}{\gamma_{ij}}$ can be plugged into the estimate of $\text{ALG}_{\text{Medians}}$ given above to get

$$\frac{\text{ALG}_{\text{Medians}}}{\text{OPT}_{\text{Medians}}} \leq (1 + \epsilon) + (1 + \delta)(2 + \epsilon)$$

with the help of reasoning analogous to Lemma 4.

IV. Empirical Findings

For our experiments, we used the following clustering subroutines: constant factor approximation algorithms for the k -centers problem, the k -medians problem, and the k -means problem (all three from chapter four of [13]), the “kmltest” program for k -means [14; 15], and a subroutine which randomly chooses a k -means solution from the possible universe of them. The last one is for the purpose of comparison.

	Average	Median
k -centers	1.0830	1.0623
k -medians	1.0396	1.0346
constant-factor k -means	1.0804	1.0691
“kmltest” k -means	1.0665	1.0571
random k -means	1.4599	1.4146

(a) Uniformly random data.

	Average	Median
constant-factor k -means	1.0128	1.0015
“kmltest” k -means	1.0068	1.0000
random k -means	1.3444	1.3292

(b) The Tallahassee dataset.

Table I: Approximation ratios for the two experiments.

The constant-factor k -centers algorithm provides a 2-factor approximation. It builds a collection of centers by greedily selecting as an additional center the member of Q furthest from the current collection of centers. The k -medians algorithm provides a $(5 + \epsilon)$ -factor approximation. It starts with an initial set of centers (those centers are the output from the k -centers algorithm), then it treats those centers as medians and swaps current medians with presently unused members of Q to improve the approximation. It continues doing this until it fails to find a swap which has an approximation which is a factor of $(1 - \tau)$ of the present one. The k -means algorithm works in exactly the same way as the k -medians algorithm, and provides a $(25 + \epsilon)$ -factor approximation. For all of our experiments, we used the parameter value $\tau = 0.10$. We should note that the quoted approximation factors apply to the problem of finding centers $C \subset Q$, so the approximation factor provided by these algorithms when used as subroutines in this particular context could be somewhat higher.

We did not want our subroutines to return answers constrained to the set Q , but instead we wanted unconstrained answers. Unconstrained solutions were generated from the constrained ones by computing the centroids of the clusters returned by the algorithms. For k -means, this meant finding the mean of each cluster, for k -medians we used Weiszfeld’s algorithm [16] on each cluster, and for k -centers we used the algorithm from [17].

The “kmltest” program also requires parameters. For this, we used the “ez-hybrid” modality with 777 maximum stages and a minimum total RDL of 0.10. We have found empirically that these values give a good mixture of speed and accuracy for our purpose.

We did two different experiments. The first experiment was on uniformly random data generated in a six-dimensional hypercube. The size of $|P|$ was fixed at 20 while k was allowed to range from 1 to 9 and $n = |Q|$ went from 10 to 30. We generated 1000 random instances for each combination of k and n . The approximation ratio achieved by the algorithm as a function of k can be seen in Figure 3. The same as a function of $|Q|$ can be seen in Figure 4. The results are

summarized in tabular form in Table *Ia*.

We also did an experiment on a dataset representing the road network in Tallahassee, Florida. In that dataset, $|P| = 1580$, and for the experiment we let $n = |Q|$ range from 3 to 103 with $k = 2$ fixed. For each value of n , we generated 1000 random query sets uniformly within the bounding box of P and calculated approximation ratios and execution times. We did not test k -centers or k -medians on the Tallahassee data set. We found that computing approximation ratios for larger or more values of k was not practical because $\binom{|P|}{k}$ possible solutions must be exhaustively checked to find the optimal one. The approximation ratio results are summarized in Table *Ib*. The results for the constant-factor algorithm are graphed in Figure 5, the behavior of the “kmltest” algorithm can be seen in Figure 6, and the statistics for a randomly chosen solution are in Figure 7.

As the data show, the approximation scheme that we have presented works well. In the k -means tests that we performed, the objective value of the solution found was on average no more than 8 percent larger than that of the optimal solution for data and queries randomly generated in a hypercube. By contrast, a random solution on average gives an objective value that is more than 45 percent larger. On a dataset representing a city, the spread is 1 percent versus more than 33 percent.

V. Conclusions and Future Work

In this article we also proved that the algorithm provides a constant factor approximation in the context of k -means. In addition, we have provided an informal discussion of how to extend the ideas used to prove Theorem 1 to many other situations, including k -medians. Our experiments clearly show that in most cases, the approximation ratio provided by the algorithm can be expected to be far below the theoretical bound that we derive.

We do not believe that the bounds calculated in this article are tight. The empirical evidence and the use of an ALG based on Q_{b_i} rather than Q_{a_i} in the analysis both point in that direction. The tightening of these

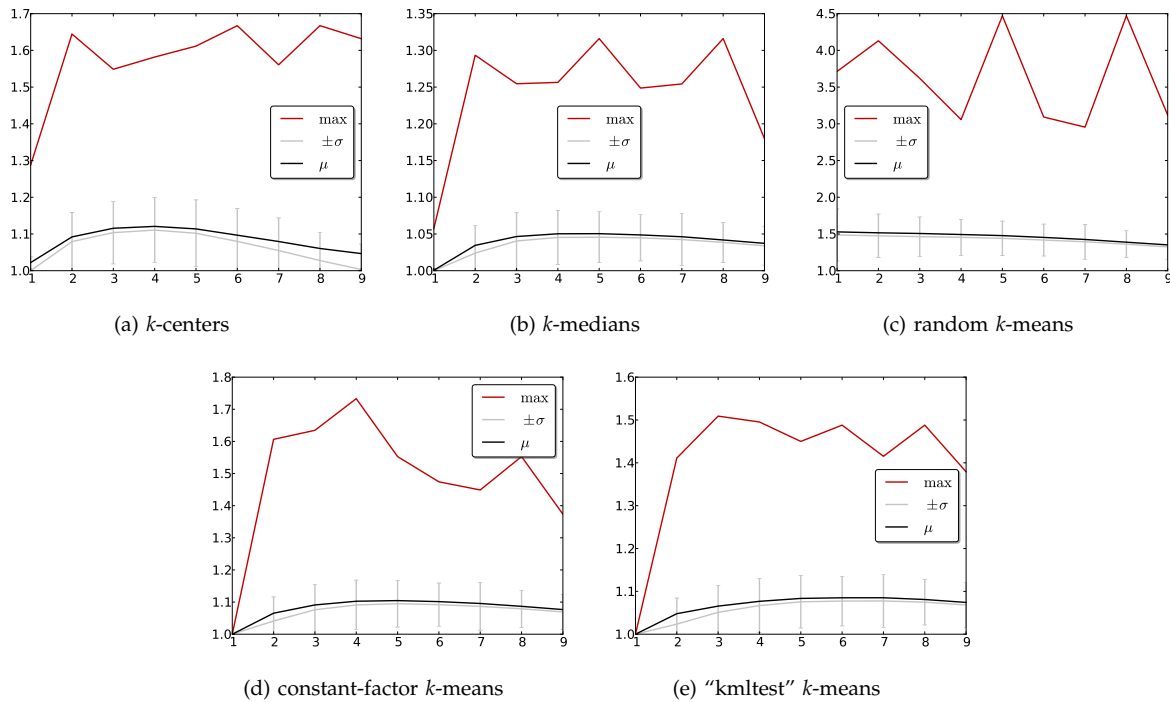


Figure 3: Approximation ratios as a function of k .

bounds is an avenue for future work. We also believe that the extension of this scheme to handle data sets with outliers might also be useful.

Acknowledgements: The authors were partially supported by NSF through CAREER Grant CCF-0643593 and the AFOSR Young Investigator Research Program. The authors would like to thank Dr. David Mount for initial discussions on this problem and for pointing out that in case of 1-means, our technique could actually give the exact answer instead of an approximation. We would like to thank Bradley Neff and Samidh Chatterjee for discussions, help with the implementation and code that they made available to us [1].

References

- [1] S. Chatterjee, B. Neff, and P. Kumar, "Instant approximate 1-center on road networks via embeddings," ser. SIGSPATIAL GIS '11. Chicago, IL, USA: ACM, 2011, to appear.
- [2] J. Venna and S. Kaski, "Local multidimensional scaling," *Neural Networks*, vol. 19, pp. 889–899, July 2006. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1167870.1167889>
- [3] F. Li, B. Yao, and P. Kumar, "Group enclosing queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1526–1540, 2011.
- [4] Y. Li, F. Li, K. Yi, B. Yao, and M. Wang, "Flexible aggregate similarity search," in *Proceedings of the 2011 international conference on Management of data*, ser. SIGMOD '11. New York, NY, USA: ACM, 2011, pp. 1009–1020. [Online]. Available: <http://doi.acm.org/10.1145/1989323.1989429>
- [5] D. Papadias, Y. Tao, K. Mouratidis, and C. K. Hui, "Aggregate nearest neighbor queries in spatial databases," *ACM Transactions on Database Systems*, vol. 30, pp. 529–576, June 2005. [Online]. Available: <http://doi.acm.org/10.1145/1071610.1071616>
- [6] K. Mouratidis, D. Papadias, and S. Papadimitriou, "Tree-based partition querying: a methodology for computing medoids in large spatial datasets," *The VLDB Journal*, vol. 17, pp. 923–945, 2008, 10.1007/s00778-007-0045-2. [Online]. Available: <http://dx.doi.org/10.1007/s00778-007-0045-2>
- [7] D. Papadias, Q. Shen, Y. Tao, and K. Mouratidis, "Group nearest neighbor queries," *Data Engineering, International Conference on*, vol. 0, p. 301, 2004.
- [8] H. Li, H. Lu, B. Huang, and Z. Huang, "Two ellipse-based pruning methods for group nearest neighbor queries," in *Proceedings of the 13th annual ACM international workshop on Geographic*

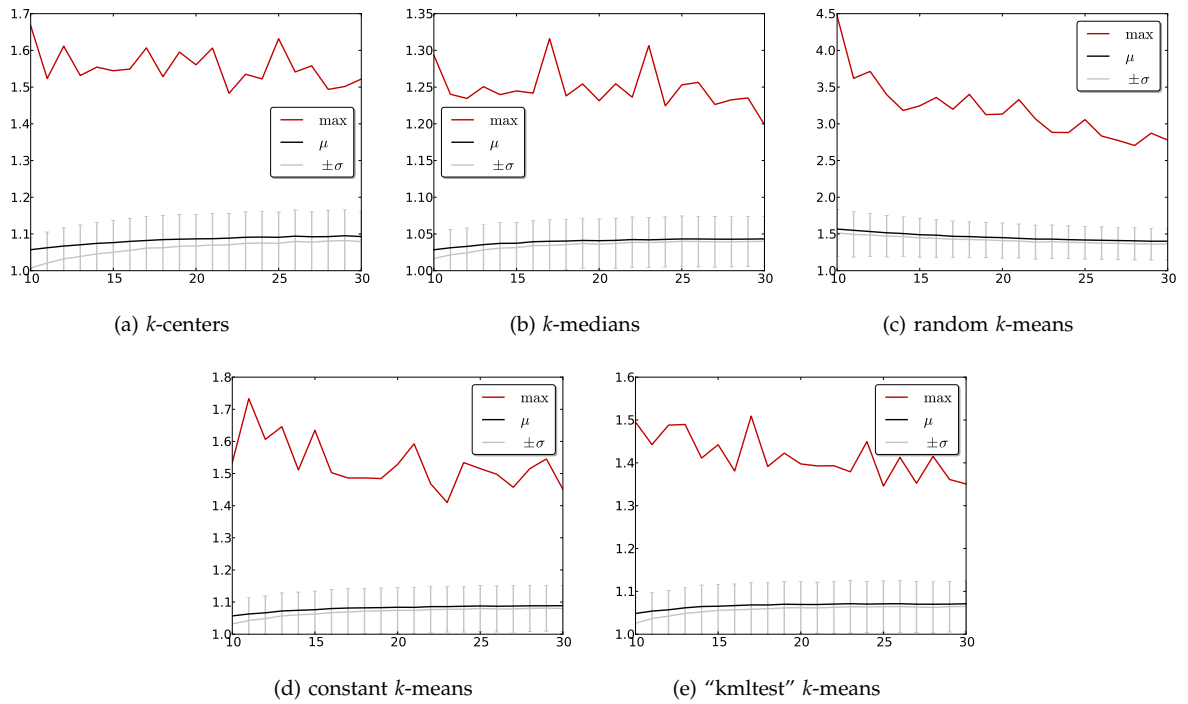


Figure 4: Approximation ratios as a function of $|Q|$.

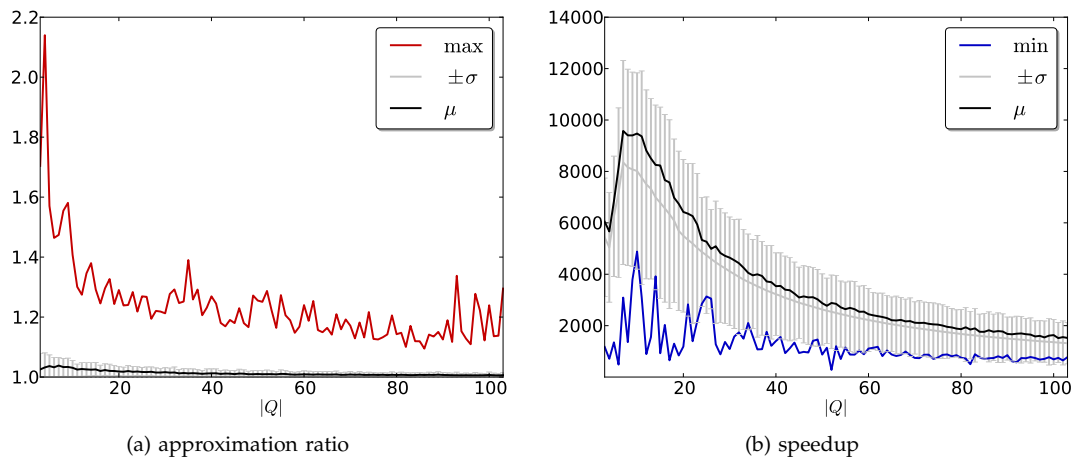


Figure 5: The constant-factor algorithm on the Tallahassee data set.

- information systems*, ser. GIS '05. New York, NY, USA: ACM, 2005, pp. 192–199. [Online]. Available: <http://doi.acm.org/10.1145/1097064.1097092>
- [9] P. Bose, A. Maheshwari, and P. Morin, “Fast approximations for sums of distances, clustering and the fermat–weber problem,” *Computational Geometry Theory and Applications*, vol. 24, pp. 135–146, April 2003. [Online]. Available: <http://dl.acm.org/citation.cfm?id=639211.639213>
- [10] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys, “A constant-factor approximation algorithm for the k-median problem (extended abstract),” in *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, ser. STOC '99. New York, NY, USA: ACM, 1999, pp. 1–10. [Online]. Available: <http://doi.acm.org/10.1145/301250.301257>
- [11] M. Bădoiu, S. Har-Peled, and P. Indyk, “Approx-

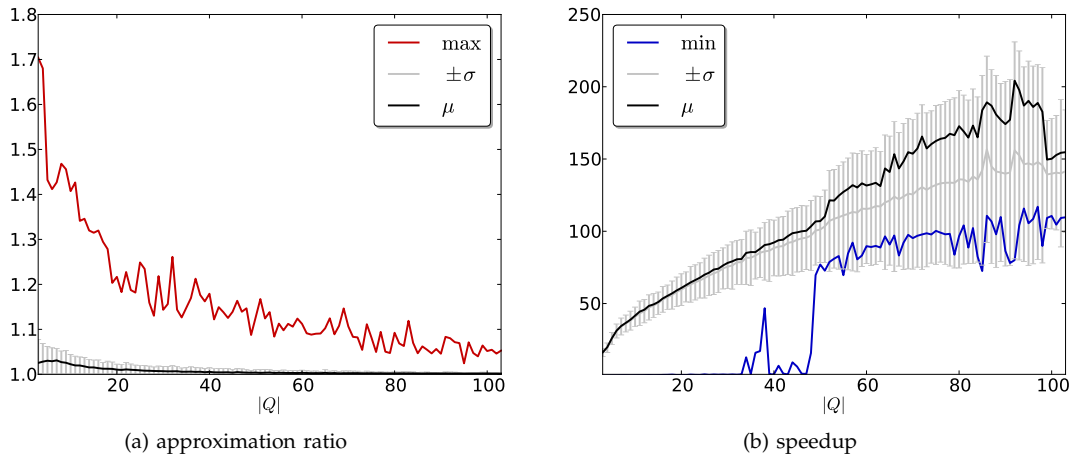


Figure 6: The “kmltest” algorithm on the Tallahassee data set.

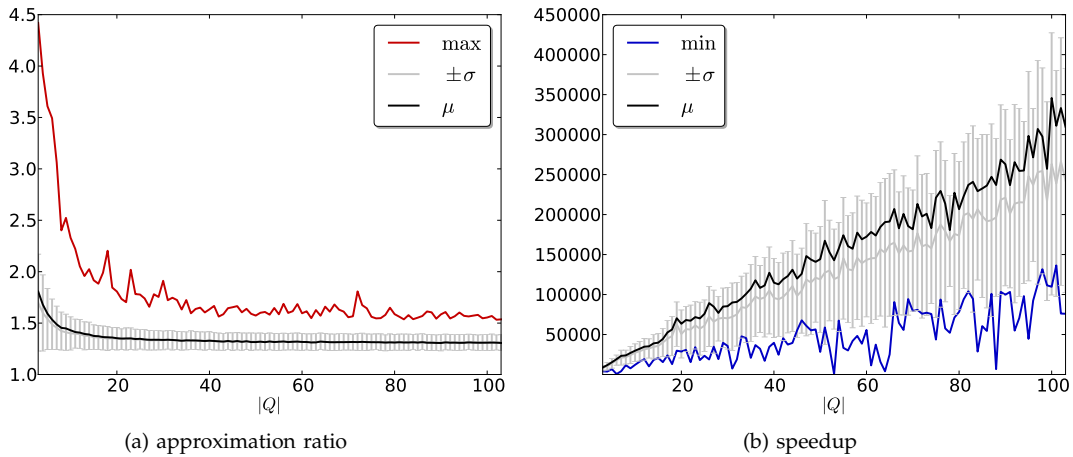


Figure 7: Random solutions on the Tallahassee data set.

imate clustering via core-sets,” in *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, ser. STOC ’02. New York, NY, USA: ACM, 2002, pp. 250–257. [Online]. Available: <http://doi.acm.org/10.1145/509907.509947>

- [12] D. Feldman and M. Langberg, “A unified framework for approximating and clustering data,” in *Proceedings of the 43rd annual ACM symposium on Theory of computing*, ser. STOC ’11. New York, NY, USA: ACM, 2011, pp. 569–578.
- [13] S. Har-Peled, *Geometric Approximation Algorithms*. American Mathematical Society, 2011.
- [14] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, vol. 24, pp. 881–892, July 2002.

- [15] —, “A local search approximation algorithm for k-means clustering,” in *Proceedings of the eighteenth annual symposium on Computational geometry*, ser. SCG ’02. New York, NY, USA: ACM, 2002, pp. 10–18. [Online]. Available: <http://doi.acm.org/10.1145/513400.513402>
- [16] E. Weiszfeld, “Sur le point pour lequel la somme des distances de n points donnees est minimum,” *Tohoku Mathematical Journal*, vol. 43, pp. 355–386, 1937.
- [17] M. Bădoiu and K. L. Clarkson, “Optimal core-sets for balls,” *Computational Geometry Theory and Applications*, vol. 40, pp. 14–22, May 2008.