2017

Streaming Algorithms for k-Means Clustering with Fast Queries

Yu Zhang
Iowa State University, yuz1988@iastate.edu

Kanat Tangwongsan
Mahidol University International College

Srikanta Tirthapura
Iowa State University, snt@iastate.edu

Streaming Algorithms for k-Means Clustering with Fast Queries

Abstract
We present methods for k-means clustering on a stream with a focus on providing fast responses to clustering queries. When compared with the current state-of-the-art, our methods provide a substantial improvement in the time to answer a query for cluster centers, while retaining the desirable properties of provably small approximation error, and low space usage. Our algorithms are based on a novel idea of "coreset caching" that reuses coresets (summaries of data) computed for recent queries in answering the current clustering query. We present both provable theoretical results and detailed experiments demonstrating their correctness and efficiency.

Streaming Algorithms for k-Means Clustering with Fast Queries

Yu Zhang* Kanat Tangwongsan† Srikanta Tirthapura‡

January 17, 2017

Abstract

We present methods for k-means clustering on a stream with a focus on providing fast responses to clustering queries. When compared with the current state-of-the-art, our methods provide a substantial improvement in the time to answer a query for cluster centers, while retaining the desirable properties of provably small approximation error, and low space usage. Our algorithms are based on a novel idea of "coreset caching" that reuses coresets (summaries of data) computed for recent queries in answering the current clustering query. We present both provable theoretical results and detailed experiments demonstrating their correctness and efficiency.

1 Introduction

Clustering is a fundamental method for understanding and interpreting data. The goal of clustering is to partition input objects into groups or "clusters" such that objects within a cluster are similar to each other, and objects in different clusters are not. A popular formulation of clustering is k-means clustering. Given a set of points S in an Euclidean space and a parameter k, the goal is to partition S into k "clusters" in a way that minimizes a cost metric based on the ℓ_2 distance between points. The k-means formulation is widely used in practice.

We consider *streaming k-means clustering*, where the inputs to clustering are not all available at once, but arrive as a continuous, possibly unending sequence. The algorithm needs to maintain enough state to be able to incrementally update the clusters as more tuples arrive. When a query is posed, the algorithm is required to return k cluster centers, one for each cluster within the data observed so far.

While there has been substantial prior work on streaming k-means clustering (e.g. [1, 2, 3, 4]), the major focus of prior work has been on optimizing the memory used by the streaming algorithm. In this respect, these works have been successful, and achieve a provable guarantee on the approximation quality of clustering, while using space polylogarithmic in the size of the input stream [1, 2]. However, for all these algorithms, when a query for cluster centers is posed, an expensive computation is needed at time of query. This can be a serious problem for applications that need answers in (near) real-time, such as in network monitoring and sensor data analysis. Our work aims at designing a *streaming clustering algorithm that significantly improves the clustering query runtime compared to the current state-of-the-art, while maintaining other desirable properties enjoyed by current algorithms, such as provable accuracy and limited memory.*

To understand why current solutions have a high query runtime, let us review the framework used in current solutions for streaming k-means clustering. At a high level, incoming data stream $\mathcal{S}$ is divided into

*Department of Electrical and Computer Engineering, Iowa State University, yuz1988@iastate.edu
†Computer Science Program, Mahidol University International College, kanat.tan@mahidol.edu
‡Department of Electrical and Computer Engineering, Iowa State University, snt@iastate.edu

smaller "chunks" $\mathcal{S}_1, \mathcal{S}_2, \ldots,$. Each chunk is summarized using a "coreset" (for example, see [5]). The resulting coresets may still not all fit into the memory of the processor, so multiple coresets are further merged recursively into higher level coresets forming a hierarchy of coresets, or a "coreset tree". When a query arrives, all active coresets in the coreset tree are merged together, and a clustering algorithm such as k-means++ [6] is applied on the result, outputting k cluster centers. The query runtime is proportional to the number of coresets that need to be merged together. In prior algorithms, the total size of all these coresets could be as large as the memory of the processor itself, and hence the query runtime can be very high.

1.1 Our Contributions

We present algorithms for streaming k-means whose query runtime is substantially smaller than the current state-of-the-art, while maintaining the desirable properties of a low memory footprint and provable approximation guarantees on the result. Our algorithms are based on the idea of "coreset caching" that to our knowledge, has not been used before in streaming clustering. The idea in coreset caching is to *reuse coresets that have been computed during previous (recent) queries to speedup the computation of a coreset for the current query*. This way, when a query arrives, it is not needed to combine all coresets currently in memory; it is sufficient to only merge a coreset from a recent query (stored in the coreset cache) along with coresets of points that arrived after this query. We show that this leads to substantial savings in query runtime.

Name	Query Cost (per point)	Update Cost (per point)	Memory Used	Coreset level returned at query after N batches
Coreset Tree (CT)	$O\left(\frac{kdm}{q} \cdot \frac{r \log N}{\log r}\right)$	$O(kd)$	$O\left(md\frac{r \log N}{\log r}\right)$	$\log_r N$
Cached Coreset Tree (CC)	$O\left(\frac{kdm}{q} \cdot r\right)$	$O(kd)$	$O\left(md\frac{r \log N}{\log r}\right)$	$2 \log_r N$
Recursive Cached Coreset Tree (RCC)	$O\left(\frac{kdm}{q} \log \log N\right)$	$O(kd \log \log N)$	$O(mdN^{1/8})$	$O(1)$
Online Coreset Cache (OnlineCC)	usually $O(1)$ worst case $O\left(\frac{kdm}{q} \cdot r\right)$	$O(kd)$	$O\left(md\frac{r \log N}{\log r}\right)$	$2 \log_r N$

Table 1: The accuracy and query cost of different clustering methods. k is the number of centers desired, d is the dimension of a data point, m is the size of a coreset (in practice, this is a constant factor times k), r is a parameter used for CC and CT, showing the "merge degree" of the coreset tree, q is the number of points between two queries. The "level" of a coreset is indicative of the number of recursive merges of prior coresets to arrive at this coreset. Smaller the level, more accurate is the coreset. For example, a batch algorithm that sees the entire input can return a coreset at level 0.

Let n denote the number of points observed in the stream so far. Let $N = \frac{n}{m}$ where m is the size of a coreset, a parameter that is independent of n. Our main contributions are as follows:

- We present an algorithm "Cached Coreset Tree" (CC) whose query runtime is a factor of $O(\log N)$ smaller than the query runtime of a state-of-the-art current method, "Coreset Tree" (CT), [1] while using similar memory and maintaining the same quality of approximation.

[1] CT is essentially the streamkm++ algorithm [1] and [2] except it has a more flexible rule for merging coresets.

- We present a recursive version of CC, "Recursive Cached Coreset Tree" (RCC), which provides more flexible tradeoffs for the memory used, quality of approximation, and query cost. For instance, it is possible to improve the query runtime by a factor of $O(\log N / \log \log N)$, and improve the quality of approximation, at the cost of greater memory.
- We present an algorithm OnlineCC, a combination of CC and a simple sequential streaming clustering algorithm (due to [7]), which provides further improvements in clustering query runtime while still maintaining the provable clustering quality, as in RCC and CC.
- For all algorithms, we present proofs showing that the k centers returned in response to a query form an $O(\log k)$ approximation to the optimal k-means clustering cost. In other words, the quality is comparable to what we will obtain if we simply stored all the points so far in memory, and ran an (expensive) batch k-means++ algorithm at time of query.
- We present a detailed experimental evaluation. These show that when compared with streamkm++ [1], a state-of-the-art method for streaming k-means clustering, our algorithms yield substantial speedups (5x-100x) in query runtime and in total time, and match the accuracy, for a broad range of query arrival frequencies.

Our theoretical results are summarized in Table 1.

1.2 Related Work

In the batch setting, when all input is available at the start of computation, Lloyd's algorithm [8], also known as the k-means algorithm, is a simple iterative algorithm for k-means clustering that has been widely used for decades. However, it does not have a provable approximation guarantee on the quality of the clusters. k-means++ [6] presents a method to determine the starting configuration for Lloyd's algorithm that yields a provable guarantee on the clustering cost. [9] proposes a parallelization of k-means++ called k-meansII.

The earliest streaming clustering method, Sequential k-means (due to [7]), maintains the current cluster centers and applies one iteration of Lloyd's algorithm for every new point received. Because it is fast and easy to implement, Sequential k-means is commonly used in practice (e.g., Apache Spark mllib [10]). However, it cannot provide any approximation guarantees [11] on the cost of clustering. BIRCH [12] is a streaming clustering method based on a data structure called the "CF Tree", and returns cluster centers through agglomerative hierarchical clustering on the leaf nodes of the tree. CluStream[13] constructs "microclusters" that summarize subsets of the stream, and further applies a weighted k-means algorithm on the microclusters. STREAMLS [3] is a divide-and-conquer method based on repeated application of a bicriteria approximation algorithm for clustering. A similar divide-and-conquer algorithm based on k-means++ is presented in [2]. However, these methods have a high cost of query processing, and are not suitable for continuous maintenance of clusters, or for frequent queries. In particular, at the time of query, these require merging of multiple data structures, followed by an extraction of cluster centers, which is expensive.

[5] presents coresets of size $O\left(\frac{k \log n}{\epsilon^d}\right)$ for summarizing n points k-means, and also show how to use the merge-and-reduce technique based on the Bentley-Saxe decomposition [14] to derive a small-space streaming algorithm using coresets. Further work [15, 16] has reduced the size of a k-means coreset to $O\left(\frac{kd}{\epsilon^6}\right)$. streamkm++ [1] is a streaming k-means clustering algorithm that uses the merge-and-reduce technique along with k-means++ to generate a coreset. Our work improves on streamkm++ w.r.t. query runtime.

Roadmap. We present preliminaries in Section 2, background for streaming clustering in Section 3 and then the algorithms CC, RCC, and OnlineCC in Section 4, along with their proofs of correctness and quality guarantees. We then present experimental results in Section 5.

2 Preliminaries

2.1 Model and Problem Description

We work with points from the d-dimensional Euclidean space $\mathbb{R}^d$ for integer $d > 0$. A point can have a positive integral weight associated with it. If unspecified, the weight of a point is assumed to be 1. For points $x, y \in \mathbb{R}^d$, let $D(x, y) = \|x - y\|$ denote the Euclidean distance between x and y. For point x and a point set $\Psi \subseteq \mathbb{R}^d$, the distance of x to Ψ is defined to be $D(x, \Psi) = \min_{\psi \in \Psi} \|x - \psi\|$.

Definition 1 (k-means clustering problem) *Given a set $P \subseteq \mathbb{R}^d$ with n points, an associated weight function $w\colon P \to Z^+$, find a point set $\Psi \subseteq \mathbb{R}^d$, $|\Psi| = k$, that minimizes the objective function*

$$\phi_\Psi(P) = \sum_{x \in P} w(x) \cdot D^2(x, \Psi) = \sum_{x \in P} \min_{\psi \in \Psi} \left(w(x) \cdot \|x - \psi\|^2 \right).$$

Streams: A stream $\mathcal{S} = e_1, e_2, \ldots$ is an ordered sequence of points, where e_i is the i-th point observed by the algorithm. For $t > 0$, let $\mathcal{S}(t)$ denote the prefix $e_1, e_2, \ldots, e_t$. For $0 < i \leq j$, let $\mathcal{S}(i, j)$ denote the substream $e_i, e_{i+1}, \ldots, e_j$. Let n denote the total number of points observed so far. Define $S = \mathcal{S}(1, n)$ be the whole stream observed until e_n.

We have written our analysis as if a query for cluster centers arrives every q points. This parameter (q) captures the query rate in the most basic terms, we note that our results on the amortized query processing time still hold as long as the average number of points between two queries is q. The reason is that the cost of answering each query does not relate to when the query arrives, and the total query cost is simply the number of queries times the cost of each query. Suppose that the queries arrived according to a different probability distribution, such that the expected interval between two queries is q points. Then, the same results will hold in expectation.

In the theoretical analysis of our algorithms, we measure the performance in both terms of computational runtime and memory consumed. The computational cost can be divided into two parts, the query runtime, and the update runtime, which is the time to update internal data structures upon receiving new points. We typically consider *amortized* processing cost, which is the average per-point processing cost, taken over the entire stream. We express the memory cost in terms of *words*, while assuming that each point in $\mathbb{R}^d$ can be stored in $O(d)$ words.

2.2 The k-means++ Algorithm

Our algorithm uses as a subroutine the k-means++ algorithm [6], a batch algorithm for k-means clustering with provable guarantees on the quality of the objective function. The properties of the algorithm are summarized below.

Theorem 2 (Theorem 3.1 in [6]) *On an input set of n points $P \subseteq \mathbb{R}^d$, the k-means++ algorithm returns a set Ψ of k centers such that $\mathbf{E}[\phi_\Psi(P)] \leq 8(\ln k + 2) \cdot \phi_{OPT}(P)$ where $\phi_{OPT}(P)$ is the optimal k-means clustering cost for P. The time complexity of the algorithm is $O(nkd)$.*

2.3 Coresets

Our clustering method builds on the concept of a *coreset*, a small-space representation of a weighted set of points that (approximately) preserves certain properties of the original set of points.

Algorithm 1: Stream Clustering Driver

```
def StreamCluster-Update(H, p)
    ▷ Insert points into D in batches of size m
    H.n ← H.n + 1
    Add p to H.C
    if (|H.C| = m) then
        H.D.Update(H.C, H.n/m)
        H.C ← ∅
def StreamCluster-Query()
    C_1 ← H.D.Coreset()
    return k-means++(k, C_1 ∪ H.C)
```

Definition 3 (k-means Coreset) *For a weighted point set $P \subseteq \mathbb{R}^d$, integer $k > 0$, and parameter $0 < \epsilon < 1$, a weighted set $C \subseteq \mathbb{R}^d$ is said to be a (k, ϵ)-coreset of P for the k-means metric, if for any set Ψ of k points in $\mathbb{R}^d$, we have*

$$(1 - \epsilon) \cdot \phi_\Psi(P) \leq \phi_\Psi(C) \leq (1 + \epsilon) \cdot \phi_\Psi(P)$$

When k is clear from the context, we simply say an ϵ-coreset. In this paper we use term "coreset" to mean a k-means coreset. For integer $k > 0$, parameter $0 < \epsilon < 1$, and weighted point set $P \subseteq \mathbb{R}^d$, we use the notation $\texttt{coreset}(k, \epsilon, P)$ to mean a (k, ϵ)-coreset of P. We use the following observations from [5].

Observation 4 ([5]) *If C_1 and C_2 are each (k, ϵ)-coresets for disjoint multi-sets P_1 and P_2 respectively, then $C_1 \cup C_2$ is a (k, ϵ)-coreset for $P_1 \cup P_2$.*

Observation 5 ([5]) *If C_1 is (k, ϵ)-coreset for C_2, and C_2 is a (k, δ)-coreset for P, then C_1 is a $(k, (1 + \epsilon)(1 + \delta) - 1)$-coreset for P.*

While our algorithms can work with any method for constructing coresets, one concrete construction due to [16] provides the following guarantees.

Theorem 6 (Theorem 15.7 in [16]) *Let $0 < \delta < \frac{1}{2}$ and let n denote the size of point set P. There exists an algorithm to compute $\texttt{coreset}(k, \epsilon, P)$ with probability at least $1 - \delta$. The size of the coreset is $O\left(\frac{kd + \log(\frac{1}{\delta})}{\epsilon^6}\right)$, and the construction time is $O\left(ndk + \log^2\left(\frac{1}{\delta}\right)\log^2 n + |\texttt{coreset}(k, \epsilon, P)|\right)$.*

3 Streaming Clustering and Coreset Trees

To provide context for how algorithms in this paper will be used, we describe a generic "driver" algorithm for streaming clustering. We also discuss the coreset tree (CT) algorithm. This is both an example of how the driver works with a specific implementation and a quick review of an algorithm from prior work that our algorithms build upon.

Algorithm 2: Coreset Tree Update

```
▷ Input: bucket b
def CT-Update(b)
    Append b to Q_0
    j ← 0
    while |Q_j| ≥ r do
        U ← coreset(k, ε, ∪_{B∈Q_j} B)
        Append U to Q_{j+1}
        Q_j ← ∅
        j ← j + 1
def CT-Coreset()
    return ⋃_j {⋃_{B∈Q_j} B}
```

3.1 Driver Algorithm

The "driver" algorithm (presented in Algorithm 1) is initialized with a specific implementation of a clustering data structure $\mathcal{D}$ and a batch size m. It internally keeps state inside an object $\mathcal{H}$. It groups arriving points into batches of size m and inserts into the clustering data structure at the granularity of a batch. $\mathcal{H}$ stores additional state, the number of points received so far in the stream $\mathcal{H}.n$, and the current batch of points $\mathcal{H}.C$. Subsequent algorithms in this paper, including CT, are implementations for the clustering data structure $\mathcal{D}$.

3.2 CT: r-way Merging Coreset Tree

The *r-way coreset tree* (CT) turns a traditional batch algorithm for coreset construction into a streaming algorithm that works in limited space. Although the basic ideas are the same, our description of CT generalizes the coreset tree of Ackermann et al. [1], which is the special case when $r = 2$.

The Coreset Tree: A coreset tree Q maintains *buckets* at multiple levels. The buckets at level 0 are called *base buckets*, which contain the original input points. The size of each base bucket is specified by a parameter m. Each bucket above that is a coreset summarizing a segment of the stream observed so far. In an r-way CT, level ℓ has between 0 and $r - 1$ (inclusive) buckets, each a summary of r^ℓ base buckets.

Initially, the coreset tree is empty. After observing n points in the stream, there will be $N = \lfloor n/m \rfloor$ base buckets (level 0). Some of these base buckets may have been merged into higher-level buckets. The distribution of buckets across levels obeys the following invariant:

> If N is written in base r as $N = (s_q, s_{q-1}, \ldots, s_1, s_0)_r$, with s_q being the most significant digit (i.e., $N = \sum_{i=0}^{q} s_i r^i$), then *there are exactly s_i buckets in level i.*

How is a base bucket added? The process to add a base bucket is reminiscent of incrementing a base-r counter by one, where merging is the equivalent of transferring the carry from one column to the next. More specifically, CT maintains a sequence of sequences $\{Q_j\}$, where Q_j is the buckets at level j. To incorporate a new bucket into the coreset tree, CT-Update, presented in Algorithm 2, first adds it at level 0. When the number of buckets at any level i of the tree reaches r, these buckets are merged, using the coreset algorithm, to form a single bucket at level $(i + 1)$, and the process is repeated until there are fewer than r buckets at all levels of the tree. An example of how the coreset tree evolves after the addition of base buckets is shown in Figure 1.

How to answer a query? The algorithm simply unions all the (active) buckets together, specifically $\bigcup_j \{\bigcup_{B \in Q_j} B\}$. Notice that the driver will combine this with a partial base bucket before deriving the k-means centers.

We present lemmas stating the properties of the CT algorithm and we use the following definition in proving clustering guarantees.

Definition 7 (Level-ℓ Coreset) *For $\ell \in \mathbb{Z}_{\geq 0}$,* a (k, ϵ, ℓ)-coreset *of a point set $P \subseteq \mathbb{R}^d$, denoted by* $\texttt{coreset}(k, \epsilon, \ell, P)$*, is as follows:*

- *The level-0 coreset of P is P.*
- *For $\ell > 0$, a level-ℓ coreset of P is a coreset of the C_i's (i.e., $\texttt{coreset}(k, \epsilon, \cup_{i=1}^{t} C_i)$), where each C_i is a level-ℓ_i coreset, $\ell_i < \ell$, of P_i such that $\{P_j\}_{j=1}^{t}$ forms a partition of P.*

We first determine the number of levels in the coreset tree after observing N base buckets. Let the maximum level of the tree be denoted by $\ell(Q) = \max\{j \mid Q_j \neq \emptyset\}$.

Lemma 8 *After observing N base buckets, $\ell(Q) \leq \frac{\log N}{\log r}$.*

Proof: As was pointed out earlier, for each level $j \geq 0$, a bucket in Q_j is a summary of r^j base buckets. Let $\ell^* = \ell(Q)$. After observing N base buckets, the coverage of a bucket at level ℓ^* cannot exceed N, so $r^{\ell^*} \leq N$, which means $\ell(Q) = \ell^* = \frac{\log N}{\log r}$. ■

Lemma 9 *For a point set P, parameter $\epsilon > 0$, and integer $\ell \geq 0$, if $C = \texttt{coreset}(k, \epsilon, \ell, P)$ is a level ℓ-coreset of P, then $C = \texttt{coreset}(k, \epsilon', P)$ where $\epsilon' = (1+\epsilon)^{\ell} - 1$.*

Proof: We prove this by induction using the proposition $\mathcal{P}$: *For a point set P, if $C = \texttt{coreset}(k, \epsilon, \ell, P)$, then $C = \texttt{coreset}(k, \epsilon', P)$ where $\epsilon' = (1+\epsilon)^{\ell} - 1$.*

To prove the base case of $\ell = 0$, consider that, by definition, $\texttt{coreset}(k, \epsilon, 0, P) = P$, and $\texttt{coreset}(k, 0, P) = P$.

Now consider integer $L > 0$. Suppose that for each positive integer $\ell < L$, $\mathcal{P}(\ell)$ was true. The task is to prove $\mathcal{P}(L)$. Suppose $C = \texttt{coreset}(k, \epsilon, L, P)$. Then there must be an arbitrary partition of P into sets $P_1, P_2, \ldots, P_q$ such that $\cup_{i=1}^{q} P_i = P$. For $i = 1 \ldots q$, let $C_i = \texttt{coreset}(k, \epsilon, \ell_i, P_i)$ for $\ell_i < L$. Then C must be of the form $\texttt{coreset}(k, \epsilon, \cup_{i=1}^{q} C_i)$.

By the inductive hypothesis, we know that $C_i = \texttt{coreset}(k, \epsilon_i, P_i)$ where $\epsilon_i = (1+\epsilon)^{\ell_i} - 1$. By the definition of a coreset and using $\ell_i \leq (L-1)$, it is also true that $C_i = \texttt{coreset}(k, \epsilon'', P_i)$ where $\epsilon'' = (1+\epsilon)^{(L-1)} - 1$. Let $C' = \cup_{i=1}^{q} C_i$. From Observation 4 and using $P = \cup_{i=1}^{q} P_i$, it must be true that $C' = \texttt{coreset}(k, \epsilon'', P)$. Since $C = \texttt{coreset}(k, \epsilon, C')$ and using Observation 5, we get $C = \texttt{coreset}(k, \gamma, P)$ where $\gamma = (1+\epsilon)(1+\epsilon'') - 1$. Simplifying, we get $\gamma = (1+\epsilon)(1+(1+\epsilon)^{(L-1)} - 1) - 1 = (1+\epsilon)^{L} - 1$. This proves the inductive case for $\mathcal{P}(L)$, which completes the proof. ■

The accuracy of a coreset is given by the following lemma, since it is clear that a level-ℓ bucket is a level-ℓ coreset of its responsible range of base buckets.

Lemma 10 *Let $\epsilon = (c \log r)/\log N$ where c is a small enough constant. After observing stream $S = \mathcal{S}(1, n)$, a clustering query* `StreamCluster-Query` *returns a set of k centers Ψ of S whose clustering cost is a $O(\log k)$-approximation to the optimal clustering for S.*

Proof: After observing N base buckets, Lemma 8 indicates that all coresets in Q are at level no greater than $(\log N/\log r)$. Using Lemma 8, the maximum level coreset in Q is an ϵ'-coreset where

$$\epsilon' = \left[\left(1 + \frac{c\log r}{\log N}\right)^{\frac{\log N}{\log r}} - 1\right] \leq \left[e^{\left(\frac{c\log r}{\log N}\right)\cdot\frac{\log N}{\log r}} - 1\right] < 0.1.$$

Consider that `StreamCluster-Query` computes k-`means++` on the union of two sets, one of the result is `CT-Coreset`and the other the partially-filled base bucket $\mathcal{H}.C$. Hence, $\Theta = (\cup_j \cup_{B\in Q_j} B) \cup \mathcal{H}.C$ is the coreset union that is given to k-`means++`. Using Observation 4, Θ is a ϵ'-coreset of S. Let Ψ be the final k centers generated by running k-`means++` on Θ, and let Ψ_1 be the set of k centers which achieves optimal k-`means` clustering cost for S. From the definition of coreset, when $\epsilon' < 0.1$, we have

$$0.9\phi_\Psi(S) \leq \phi_\Psi(\Theta) \leq 1.1\phi_\Psi(S) \tag{1}$$

$$0.9\phi_{\Psi_1}(S) \leq \phi_{\Psi_1}(\Theta) \leq 1.1\phi_{\Psi_1}(S) \tag{2}$$

Let Ψ_2 denote the set of k centers which achieves optimal k-`means` clustering cost for Θ. Using Theorem 2, we have

$$\mathbf{E}[\phi_\Psi(\Theta)] \leq 8(\ln k + 2)\cdot\phi_{\Psi_2}(\Theta) \tag{3}$$

Since Ψ_2 is the optimal k centers for coreset Θ, we have

$$\phi_{\Psi_2}(\Theta) \leq \phi_{\Psi_1}(\Theta) \tag{4}$$

Using Equations 2, 3 and 4 we get

$$\mathbf{E}[\phi_\Psi(\Theta)] \leq 9(\ln k + 2)\cdot\phi_{\Psi_1}(S) \tag{5}$$

Using Equations 1 and 5,

$$\mathbf{E}[\phi_\Psi(S)] \leq 10(\ln k + 2)\cdot\phi_{\Psi_1}(S) \tag{6}$$

We conclude that Ψ is a factor-$O(\log k)$ clustering centers of S compared to the optimal. ■

The following lemma quantifies the memory and time cost of `CT`.

Lemma 11 *Let N be the number of buckets observed so far. Algorithm `CT`, including the driver, takes amortized $O(kd)$ time per point, using $O\left(\frac{mdr\log N}{\log r}\right)$ memory. The amortized cost of answering a query is $O\left(\frac{kdm}{q}\cdot\frac{r\log N}{\log r}\right)$ per point.*

Proof: First, the cost of arranging n points into level-0 buckets is trivially $O(n)$, resulting in $N = n/m$ buckets. For $j \geq 1$, a level-j bucket is created for every mr^j points, so the number of level-j buckets ever created is N/r^j. Hence, across all levels, the total number of buckets created is $\sum_{j=1}^{\ell}\frac{N}{r^j} = O(N/r)$. Furthermore, when a bucket is created, `CT` merges rm points into m points. By Theorem 6, the total cost of creating these buckets is $O(\frac{N}{r}\cdot(kdrm + \log^2(rm) + dk)) = O(nkd)$, hence $O(kd)$ amortized time per point. In terms of space, each level must have fewer than r buckets, each with m points. Therefore, across $\ell \leq \frac{\log N}{\log r}$ levels, the space required is $O(\frac{\log N}{\log r}\cdot mdr)$. Finally, when answering a query, the union of all the buckets has at most $O(mr\cdot\frac{\log N}{\log r})$ points, computable in the same time as the size. Therefore, k-`means++`, run on these points plus at most one base bucket, takes $O(kdrm\cdot\frac{\log N}{\log r})$. The amortized bound immediately follows. This proves

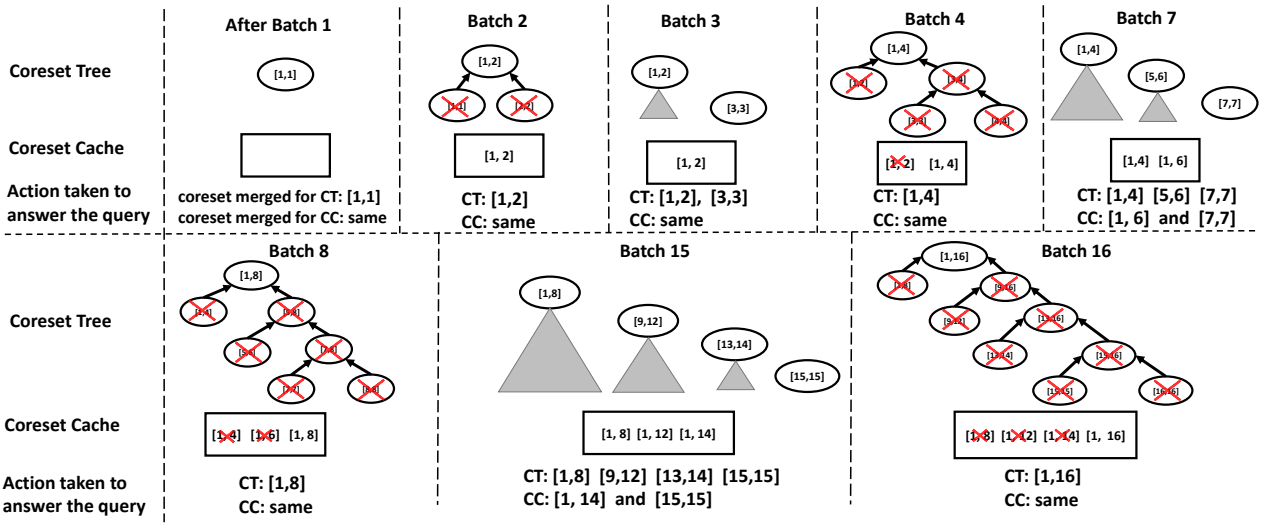


Figure 1: Illustration of Algorithm CC, showing the states of coreset tree and cache after batch 1, 2, 3, 4, 7, 8, 15 and 16. The notation $[l, r]$ denotes a coreset of all points in buckets l to r, both endpoints inclusive. The coreset tree consists of a set of coresets, each of which is a base bucket or has been formed by merging multiple coresets. Whenever a coreset is merged into a another coreset (in the tree) or discarded (in the cache), the coreset is marked with an "X". We suppose that a clustering query arrives after seeing each batch, and describe the actions taken to answer this query (1) if only CT was used, or (2) if CC was used along with CT.

the theorem. ■

As evident from the above lemma, answering a query using CT is expensive compared to the cost of adding a point. More precisely, when queries are made rather frequently—every q points, $q < O(rm \cdot \log_r N) = \widetilde{O}(rkd \cdot \log_r N)$—the cost of query processing is asymptotically greater than the cost of handling point arrivals. We address this issue in the next section.

4 Clustering Algorithms with Fast Queries

This section describes algorithms for streaming clustering with an emphasis on query time.

4.1 Algorithm CC: Coreset Tree with Caching

The CC algorithm uses the idea of "coreset caching" to speed up query processing by reusing coresets that were constructed during prior queries. In this way, it can avoid merging a large number of coresets at query time. When compared with CT, the CC algorithm can answer queries faster, while maintaining nearly the same processing time per element.

In addition to the coreset tree CT, the CC algorithm also has an additional *coreset cache*, `cache`, that stores a subset of coresets that were previously computed. When a new query has to be answered, CC avoids the cost of merging coresets from multiple levels in the coreset tree. Instead, it reuses previously cached coresets and retrieves a small number of additional coresets from the coreset tree, thus leading to

less computation at query time.

However, the level of the resulting coreset increases linearly with the number of merges a coreset is involved in. For instance, suppose we recursively merged the current coreset with the next arriving batch to get a new coreset, and so on, for N batches. The resulting coreset will have a level of $\Theta(N)$, which can lead to a very poor clustering accuracy. Additional care is needed to ensure that the level of a coreset is controlled while caching is used.

Details: Each cached coreset is a summary of base buckets 1 through some number u. We call this number u as the *right endpoint* of the coreset and use it as the key/index into the cache. We call the interval $[1, u]$ as the "span" of the bucket. To explain which coresets are cached by the algorithm, we introduce the following definitions.

For integers $n > 0$ and $r > 0$, consider the unique decomposition of n according to powers of r as $n = \sum_{i=0}^{j} \beta_i r^{\alpha_i}$, where $0 \leq \alpha_0 < \alpha_1 \ldots < \alpha_j$ and $0 < \beta_i < r$ for each i. The β_is can be viewed as the non-zero digits in the representation of n as a number in base r. Let $\texttt{minor}(n, r) = \beta_0 r^{\alpha_0}$, the smallest term in the decomposition, and $\texttt{major}(n, r) = n - \texttt{minor}(n, r)$. Note that when n is of the form βr^{α} where $0 < \beta < r$ and $\alpha \geq 0$, $\texttt{major}(n) = 0$.

For $\kappa = 1 \ldots j$, let $n_\kappa = \sum_{i=\kappa}^{j} \beta_i r^{\alpha_i}$. n_κ can be viewed as the number obtained by dropping the κ smallest non-zero digits in the representation of n as a number in base r. The set $\texttt{prefixsum}(n, r)$ is defined as $\{n_\kappa | \kappa = 1 \ldots j\}$. When n is of the form βr^{α} where $0 < \beta < r$, $\texttt{prefixsum}(n, r) = \emptyset$.

For instance, suppose $n = 47$ and $r = 3$. Since $47 = 1 \cdot 3^3 + 2 \cdot 3^2 + 2 \cdot 3^0$, we have $\texttt{minor}(47, 3) = 2, \texttt{major}(47, 3) = 45$, and $\texttt{prefixsum}(47, 3) = \{27, 45\}$.

We have the following fact on the `prefixsum`.

Fact 12 *Let $r \geq 2$. For each $N \in \mathbb{Z}_+$, $\texttt{prefixsum}(N + 1, r) \subseteq \texttt{prefixsum}(N, r) \cup \{N\}$.*

Proof: There are three cases:

Case I: $N \not\equiv (r - 1) \pmod{r}$. Consider N in r-ary representation, and let δ denote the least significant digit. Since $\delta < (r-1)$, in going from N to $(N+1)$, the only digit changed is the least significant digit, which changes from δ to $\delta + 1$ and no carry propagation takes place. For all elements $y \in \texttt{prefixsum}(N + 1, r)$, y is also in $\texttt{prefixsum}(N)$. The only exception is when $N \equiv 0 \pmod{r}$, when one element of $\texttt{prefixsum}(N + 1, r)$ is N itself. In this case, it is still true that $\texttt{prefixsum}(N + 1, r) \subseteq \texttt{prefixsum}(N, r) \cup \{N\}$.

Case II: $N \equiv (r - 1) \pmod{r}$ and for r-ary representation of N, all digits are $(r - 1)$. In this case, $(N + 1)$ should be power of r and can be represented as term r^{α} where $\alpha \geq 0$, then $\texttt{prefixsum}(N + 1, r)$ is empty set so our claim holds.

Case III: $N \equiv (r - 1) \pmod{r}$ but Case II does not hold. Consider the r-ary representation of N. There must exist at least one digit less than $(r - 1)$. $N + 1$ will change a streak of $(r - 1)$ digits to 0 starting from the least significant digit, until it hits the first digit that is not $(r - 1)$ which should be less than $(r - 1)$. We refer such digit as β_k. N can be expressed in r-ary form as $N = (\beta_j \beta_{j-1} \cdots \beta_{k+1} \beta_k \beta_{k-1} \cdots \beta_1 \beta_0)_r$. Correspondingly, $N + 1 = (\beta_j \beta_{j-1} \cdots \beta_{k+1} (1 + \beta_k) 00 \cdots 00)_r$. Comparing the `prefixsum` of $(N + 1)$ with N, the part of digits $\beta_j \beta_{j-1} \cdots \beta_{k+1}$ remains unchanged, thus $\texttt{prefixsum}(N + 1, r) \subset \texttt{prefixsum}(N, r)$. ■

`CC` caches every coreset whose right endpoint is in $\texttt{prefixsum}(N, r)$. When a query arrives after N batches, the task is to compute a coreset whose span is $[1, N]$. `CC` partitions $[1, N]$ as $[1, N_1] \cup [N_1 + 1, N]$ where $N_1 = \texttt{major}(N, r)$. Out of these two intervals, $[1, N_1]$ is already available in the cache, and $[N_1 + 1, N]$ is retrieved from the coreset tree, through the union of no more than $(r - 1)$ coresets. This needs a merge of no more than r coresets. This is in contrast with `CT`, which may need to merge as many as $(r - 1)$ coresets at each level of the tree, resulting in a merge of up to $(r - 1) \log N$ coresets at query time. The algorithm

for maintaining the cache and answering clustering queries is shown in Algorithm 3. See Figure 1 for an example of how the CC algorithms updates the cache and answers queries using cached coresets.

Note that to keep the size of the cache small, as new base buckets arrive, `CC-Update` will ensure that "stale" or unnecessary coresets are removed.

Algorithm 3: Coreset Tree with Caching: Algorithm Description

```
1  def CC-Init(r, k, ε)
2      Remember the parameters r, k, and ε.
       ▷ The coreset tree
3      Q ← CT-Init(r, k, ε)
4      cache ← ∅
5  def CC-Update(b, N)
       ▷ b is a batch and N is the number of batches so far.
6      Remember N
7      Q.CT-Update(b, N)
       ▷ May need to insert a coreset into cache
8      if r divides N then
9          c ← CC-Coreset()
10         Add coreset c to cache using key N
11         Remove from cache each bucket whose key does not appear in prefixsum(N + 1)
12 def CC-Coreset()
       ▷ Return a coreset of points in buckets 1 till N
13     N_1 ← major(N, r) and N_2 ← minor(N, r)
14     Let N_2 = βr^α where α and β < r are positive integers
       ▷ a is the coreset for buckets N_1 + 1, N_1 + 2, ..., (N_1 + N_2) = N and is retrieved from the
         coreset tree
15     a ← ∪_{B∈Q_α} B
       ▷ b is the coreset spanning [1, N_1], retrieved from the cache
16     if N_1 is 0 then
17         b ← ∅
18     else
19         b ← cache.lookup(N_1)
20     C ← coreset(k, ε, a ∪ b)
21     return C
```

The following lemma relates what the cache stores with the number of base buckets observed so far, guaranteeing that Algorithm 3 can find the required coreset.

Lemma 13 *Immediately before base bucket N arrives, each $y \in$* `prefixsum`(N, r) *appears in the key set of* `cache`.

Proof: Proof is by induction on N. The base case $N = 1$ is trivially true, since `prefixsum`$(1, r)$ is empty. For the inductive step, assume that at the beginning of batch N, each $y \in$ `prefixsum`(N, r) appears in `cache`. By Fact 12, we know that `prefixsum`$(N + 1, r) \subseteq$ `prefixsum`$(N, r) \cup \{N\}$. Using this, every bucket with a

right endpoint in $\texttt{prefixsum}(N+1, r)$ is present in `cache` at the beginning of batch $(N+1)$, except for the coreset with right endpoint N. But the algorithm adds the coreset for this bucket to the cache, if r divides N. Hence, the inductive step is proved. ■

Since $\texttt{major}(N, r) \in \texttt{prefixsum}(N, r)$ for each N, we can always retrieve the bucket with span $[1, \texttt{major}(N, r)]$ from `cache`.

Lemma 14 *When queried after inserting batch N, Algorithm* `CC-Coreset` *returns a coreset whose level in no more than* $\left\lceil \frac{2 \log N}{\log r} \right\rceil - 1$.

Proof: Let $\chi(N)$ denote the number of non-zero digits in the representation of N as a number in base r. We show that the level of the coreset returned by Algorithm `CC-Coreset` is no more than $\left\lceil \frac{\log N}{\log r} \right\rceil + \chi(N) - 1$. Since $\chi(N) \leq \left\lceil \frac{\log N}{\log r} \right\rceil$, the lemma follows.

The proof is by induction on $\chi(N)$. If $\chi(N) = 1$, then $\texttt{major}(N, r) = 0$, and the coreset is retrieved directly from the coreset tree Q. By Lemma 8, each coreset in Q is at a level no more than $\lceil \frac{\log N}{\log r} \rceil$, and the base case follows. Suppose the claim was true for all N such that $\chi(N) = t$. Consider N such that $\chi(N) = (t+1)$. The algorithm computes $N_1 = \texttt{major}(N, r)$, and retrieves the coreset with span $[1, N_1]$ from the cache. Note that $\chi(N_1) = t$. By the inductive hypothesis, b, the coreset for span $[1, N_1]$ is at a level $\left\lceil \frac{\log N}{\log r} \right\rceil + t - 1$. The coresets for span $[N_1 + 1, N]$ are retrieved from the coreset tree; note there are multiple such coresets, but each of them is at a level no more than $\left\lceil \frac{\log N}{\log r} \right\rceil$, using Lemma 8. Their union is denoted by a. The level of the final coreset for span $[1, N]$ is no more than $\left\lceil \frac{\log N}{\log r} \right\rceil + t$, proving the inductive case. ■

Let the accuracy parameter $\epsilon = \frac{c \log r}{2 \log N}$, where $c < \ln 1.1$. We have the following lemma on the accuracy of clustering centers returned by `CC`.

Lemma 15 *After observing N batches, Algorithm* `StreamCluster-Query` *when using clustering data structure* `CC`*, returns a set of k points whose clustering cost is within a factor of* $O(\log k)$ *of the optimal k-means clustering cost.*

Proof: From Lemma 14, we know that the level of a coreset returned is no more than $\left\lceil \frac{2 \log N}{\log r} \right\rceil - 1$. Following an argument similar to Lemma 10, we arrive at the result. ■

Lemma 16 *Algorithm 3 processes a stream of points using amortized time* $O(kd)$ *per point, using memory of* $O\left(\frac{mdr \log N}{\log r}\right)$*. The amortized cost of answering a query is* $O\left(\frac{kdm}{q} \cdot r\right)$.

Proof: The runtime for Algorithm `CC-Update` is the sum of the times to update the coreset tree Q and to update `cache`. We know from Lemma 11 that the time to update the coreset is $O(kd)$ per point. To update the cache, note that `CC-Update` inserts a new coreset into the cache every r batches. The cost of computing this coreset is $O(kmdr)$. Averaged over the mr points in r batches, the cost of maintaining `cache` is $O(kd)$ per point. The overall update time for Algorithm `CC-Update` is $O(kd)$ per point.

The coreset tree Q uses space $O\left(\frac{mdr \log N}{\log r}\right)$. After processing batch N, `cache` only stores those buckets b corresponding to $\texttt{prefixsum}(N+1, r)$. The number of such buckets possible is $O(\log N / \log r)$, so that the space cost of `cache` is $O(md \log N / \log r)$. The space complexity follows.

At query time, Algorithm `CC-Coreset` combines no more than r buckets, out of which there is no more than one bucket from the cache, and no more than $(r-1)$ from the coreset tree. It is necessary to run k-`means++` on $O(mr)$ points using time $O(kdmr)$. Since there is a query every q points, the amortized query time per point is $O\left(\frac{kdmr}{q}\right)$. ■

4.2 Algorithm RCC: Recursive Coreset Cache

There are a few issues with the CC data structure. One is that the level of the coreset finally generated is $O(\log_r N)$. Since theoretical guarantees on the approximation quality of clustering worsen with an increase in the level of the coreset, it is natural to ask if the level can be reduced further to $O(1)$. Further, the time taken to process a query is linearly proportional to r; it would be interesting to reduce the query time further. While it is desirable to simultaneously reduce the level of the coreset as well as the query time, at first glance, these two seem to be inversely related. It seems that if we decreased the level of a coreset, leading to better accuracy, then we will have to increase the merge degree, which would in turn increase the query time. For example, if we set $r = \sqrt{N}$, then the level of the resulting coreset is $O(1)$, but the query time will increase to $O\left(\sqrt{N}\right)$.

In the following, we present a solution RCC that uses the idea of coreset caching in a recursive manner to achieve both a low level of the coreset, as well as a small query time. In our approach, we keep the merge degree of nodes relatively high, thus keeping the levels of coresets low. At the same time, we use coreset caching even within a single level of a coreset tree, so that it is not necessary to merge r coresets at query time. The coreset caching has to be done carefully, so that the level of the coreset does not increase significantly.

For instance, suppose we built another coreset tree with merge degree 2 for the $O(r)$ coresets within a single level of the current coreset tree, this would lead to a level of $\log r$. At query time, we will need to aggregate $O(\log r)$ coresets from this tree, in addition to a coreset from the coreset cache. So, this will lead to a level of $O\left(\max\left\{\frac{\log N}{\log r}, \log r\right\}\right)$, and a query time proportional to $O(\log r)$. This is an improvement from the coreset cache, which has a query time proportional to r and a level of $O\left(\frac{\log N}{\log r}\right)$.

We can take this idea further by recursively applying the same idea to the $O(r)$ buckets within a single level of the coreset tree. Instead of having a coreset tree with merge degree 2, we use a tree with a higher merge degree, and then have a coreset cache for this tree to reduce the query time, and apply this recursively within each tree. This way we can approach the ideal of a small level and a small query time. We are able to achieve interesting tradeoffs, as shown in Table 2. In order to keep the level of the resulting coreset low, along with the coreset cache for each level, we also maintain a list of coresets at each level, like in the CT algorithm. In merging coresets to a higher level, the list is used, rather than the recursive coreset cache.

The RCC data structure is defined inductively as follows. For integer $i \geq 0$, the RCC data structure of order i is denoted by RCC(i). RCC(0) is a CC data structure with a merge degree of $r_0 = 2$. For $i > 0$, RCC(i) consists of:

- cache(i), a coreset cache that stores previously computed coresets.
- For each level $\ell = 0, 1, 2, \ldots$, there are two structures. One is a list of buckets L_ℓ, similar to the structure Q_ℓ in a coreset tree. The maximum length of a list is $r_i = 2^{2^i}$. Another is an RCC_ℓ structure which is a RCC structure of a lower order $(i-1)$, which stores the same information as L_ℓ, except, in a way that can be quickly retrieved during a query.

The main data structure $\mathcal{R}$ is initialized as $\mathcal{R}$ = RCC-Init(ι), for a parameter ι, to be chosen. Note that ι is the highest order of the recursive structure. This is also called the "nesting depth" of the structure.

Lemma 17 *When queried after inserting N batches, Algorithm 6 using RCC(ι) returns a coreset whose level is $O\left(\frac{\log N}{2^\iota}\right)$. The amortized time cost of answering a clustering query is $O\left(\frac{kdm}{q} \cdot \iota\right)$ per point.*

Proof: Algorithm 6 retrieves a few coresets from RCC of different orders. From the outermost structure R_ι = RCC(ι), it retrieves one coreset c from cache(ι). Using an analysis similar to Lemma 14, the level of b_ι is no more than $2\frac{\log N}{\log r_\iota}$.

Algorithm 4: RCC-Init(i)

$\mathcal{R}$.order $\leftarrow i$, $\mathcal{R}$.cache $\leftarrow \emptyset$, $\mathcal{R}.r \leftarrow 2^{2^i}$
/* N is the number of batches so far */
$\mathcal{R}.N \leftarrow 0$
foreach $\ell = 0, 1, 2, \ldots$ **do**
    $\mathcal{R}.L_\ell \leftarrow \emptyset$
    if $\mathcal{R}.order > 0$ **then**
        $\mathcal{R}.R_\ell \leftarrow$ RCC-Init(order $-$ 1)
return R

Algorithm 5: $\mathcal{R}$.RCC-Update(b)

/* b is a batch of points */
$\mathcal{R}.N \leftarrow \mathcal{R}.N + 1$
/* Insert b into $\mathcal{R}.L_0$ and merge if needed */
Append b to $\mathcal{R}.L_0$.
if $\mathcal{R}.order > 0$ **then**
    recursively update $\mathcal{R}.R_0$ by $\mathcal{R}.R_0$.RCC-Update(b)

$\ell \leftarrow 0$
while ($|\mathcal{R}.L_\ell| = \mathcal{R}.r$) **do**
    $b' \leftarrow$ BucketMerge($\mathcal{R}.L_\ell$)
    Append b' to $\mathcal{R}.L_{\ell+1}$
    if $\mathcal{R}.order > 0$ **then**
        recursively update $\mathcal{R}.R_{\ell+1}$ by $\mathcal{R}.R_{\ell+1}$.RCC-Update(b)
    ▷ *Empty the list of coresets*
    $\mathcal{R}.L_\ell \leftarrow \emptyset$
    ▷ *Empty the cache*
    if $\mathcal{R}.order > 0$ **then**
        $\mathcal{R}.R_\ell \leftarrow$ RCC-Init($\mathcal{R}$.order $-$ 1)
if $\mathcal{R}.r$ *divides* $\mathcal{R}.N$ **then**
    Bucket $b' \leftarrow \mathcal{R}$.RCC-Coreset()
    Add b' to $\mathcal{R}$.cache with right endpoint $\mathcal{R}.N$
    From $\mathcal{R}$.cache, remove all buckets b'' such that right(b'') $\notin$ prefixsum($\mathcal{R}.N + 1$)

Note that for $i < \iota$, the maximum number of coresets that will be inserted into RCC(i) is $r_{i+1} = r_i^2$. The reason is that inserting r_{i+1} buckets into RCC(i) will lead to the corresponding list structure for RCC(i) to become full. At this point, the list and the RCC(i) structure will be emptied out in Algorithm 5. From each recursive call to RCC(i), it can be similarly seen that the level of a coreset retrieved from the cache is at level

Algorithm 6: $\mathcal{R}$.RCC-Coreset()

$B \leftarrow \mathcal{R}$.RCC-Getbuckets
$C \leftarrow \texttt{coreset}(k, \epsilon, B)$
return bucket $(C, 1, \mathcal{R}.N, 1 + \max_{b \in B} \texttt{level}(b))$

Algorithm 7: $\mathcal{R}$.RCC-Getbuckets()

$N_1 \leftarrow \texttt{major}(\mathcal{R}.N, \mathcal{R}.r)$
$b_1 \leftarrow$ retrieve bucket with right endpoint N_1 from $\mathcal{R}$.cache
Let ℓ^* be the lowest numbered non-empty level among $\mathcal{R}.L_i, i \geq 0$.
if $\mathcal{R}.order > 0$ **then**
    $B_2 \leftarrow \mathcal{R}.R_{\ell^*}$.RCC-Getbuckets()
else
    $B_2 \leftarrow \mathcal{R}.L_{\ell^*}$
return $\{b_1\} \cup B_2$

$2\frac{\log r_i}{\log r_{i-1}}$, which is $O(1)$. The algorithm returns a coreset formed by the union of all the coresets, followed by a further merge step. Overall, the level of the coreset is one more than the maximum of the levels of all the coresets returned. This is $O(\frac{\log N}{\log r_\iota})$.

For the query cost, note that the number of coresets merged at query time is equal to the nesting depth of the structure ι. The query time equals the cost of running k-means++ on the union of all these coresets, for a total time of $O(kdm\iota)$. The amortized per-point cost of a query follows. ■

Lemma 18 *The memory consumed by* RCC(ι) *is* $O(mdr_\iota)$. *The amortized processing time is* $O(kd\iota)$ *per point.*

Proof: First, we note in RCC(i) for $i < \iota$, there are $O(1)$ lists L_ℓ. The reason is as follows. It can be seen that in order to get a single bucket in list L_2 within RCC(i), it is necessary to insert $r_i^2 = r_{i+1}$ buckets into RCC(i). Since this is the maximum number of buckets that will be inserted into RCC(i), there are no more than three levels of lists within each RCC(i) for $i < \iota$.

We prove by induction on i that RCC(i) has no more than $6r_i$ buckets. For the base case, $i = 0$, and we have $r_0 = 2$. In this case, RCC(0) has three levels, each with no more than 2 buckets. The number of buckets in the cache is also a constant for r_0, so that the total memory is no more than 6 buckets, due to the lists in different levels, and no more than 2 buckets in the cache, for a total of $8 = 4r_0$ buckets. For the inductive case, consider that RCC(i) has no more than three levels. The list at each level has no more than r_i buckets. The recursive structures R_ℓ within RCC(i) themselves have no more than $6r_{i-1}$ buckets. Adding the constant number of buckets within the cache, we get the total number of buckets within RCC(i) to be $3r_i + 4r_{i-1} + 2 = 3r_i + 6\sqrt{r_i} + 2 \leq 6r_i$, for $r_i \geq 16$, i.e. $i \geq 2$. Thus if ι is the nesting depth of the structure, the total memory consumed is $O(mdr_\iota)$, since each bucket requires $O(md)$ space.

For the processing cost, when a bucket is inserted into $\mathcal{R} =$ RCC(ι), it is added to list L_0 within $\mathcal{R}$. The cost of maintaining these lists in $\mathcal{R}$ and $\mathcal{R}$.cache, including merging into higher level lists, is amortized $O(kd)$ per point, similar to the analysis in Lemma 16. The bucket is also recursively inserted into a RCC($\iota - 1$) structure, and a further structure within, and the amortized time for each such structure is $O(kd)$ per point. The total time cost is $O(kd\iota)$ per point. ■

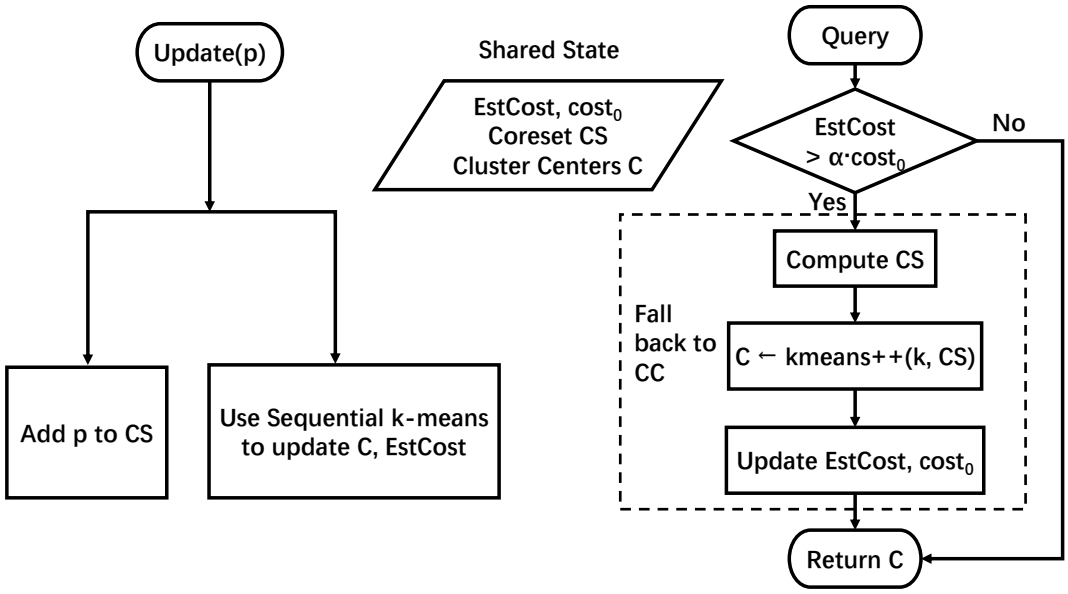


Figure 2: Illustration of Algorithm `OnlineCC`.

Different tradeoffs are possible by setting ι to specific values. Some examples are shown in the Table 2.

ι	coreset level at query	Query cost (per point)	update cost per point	Memory
$\log\log N - 3$	$O(1)$	$O(\frac{kdm}{q}\log\log N)$	$O(kd\log\log N)$	$O(mdN^{1/8})$
$\log\log N/2$	$O(\sqrt{\log N})$	$O(\frac{kdm}{q}\log\log N)$	$O(kd\log\log N)$	$O(md2^{\sqrt{\log N}})$

Table 2: Possible tradeoffs for the RCC(ι) algorithm, based on the parameter ι, the nesting depth of the structure.

4.3 Online Coreset Cache: a Hybrid approach of CC and Sequential k-means

If we breakdown the query runtime of the algorithms considered so far, we observe two major components. The first component is the construction of the coreset of all points seen so far, through merging stored coresets. The second component is the k-means++ algorithm applied on the resulting coreset. The algorithms discussed so far, CC and RCC, are focused on decreasing the runtime of the first component, coreset construction, by reducing the number of coresets to be merged at query time. But they still have to pay the cost of the second component k-means++, which is substantial in itself, since the runtime of k-means++ is $O(kdm)$ where m is the size of the coreset. To make further progress, we have to reduce this component. However, the difficulty in eliminating k-means++ at query time is that without an approximation algorithm such as k-means++ we do not have a way to guarantee that the returned clustering is an approximation to the optimal.

We present an algorithm, `OnlineCC`, which only occasionally uses k-means++ at query time, and most of the time, uses a much cheaper method of cost $O(1)$ to compute the clustering centers. `OnlineCC` uses a

combination of CC and the Sequential k-means algorithm [7] (a.k.a Online Lloyd's algorithm) to maintain the cluster centers quickly while also providing a guarantee on the quality of clustering. OnlineCC continuously maintains cluster centers in a manner similar to [7], where each arriving point incrementally updates the current set of cluster centers. While Sequential k-means can process incoming points (and answer queries) extremely quickly, it cannot provide any guarantees on the quality of answers, and in some cases, the clustering quality can be very poor when compared with say, k-means++. To guard against such deterioration in clustering quality, our algorithm (1) falls back to a provably accurate clustering algorithm CC occasionally, and (2) runs Sequential k-means only so long as the clustering cost does not get much larger than the previous time CC was used. This ensures that our clusters always have a provable quality with respect to the optimal.

In order to achieve the above, OnlineCC also processes incoming points using CC, thus maintaining coresets of substreams of data seen so far. When a query arrives, it typically answers them in $O(1)$ time using the centers maintained using Sequential k-means. If however the clustering cost is significantly higher (by more than a factor of α for a parameter $\alpha > 1$) than the previous time the algorithm fell back to CC, then the query processing again returns to regenerate a coreset using the CC algorithm. One difficulty in implementing this idea is that (efficiently) maintaining an estimate of the current clustering cost is not easy, since each change in cluster centers can affect the contribution of a number of points to the clustering cost. To reduce the cost of maintenance, our algorithm maintains an upper bound on the clustering cost; as we show further, this is sufficient to give a provable guarantee on the quality of clustering. Further details on how the upper bound on the clustering cost is maintained, and how Sequential k-means and CC interact are shown in Algorithm 8, with a schematic in Figure 2.

We state the properties of Algorithm OnlineCC.

Lemma 19 *In Algorithm 8, after observing point set P, if C is the current set of cluster centers, then* EstCost *is an upper bound on* $\phi_C(P)$.

Proof: Consider the value of EstCost between every two consecutive switches to CC. Without loss of generality, suppose there is one switch happens at time 0, let P_0 denote the points observed until time 0 (including the points received at 0). We will do induction on the number of points received after time 0, we denote this number as i. Then P_i is P_0 union the i points received after time 0.

When i is 0, we compute C from the coreset CS, from the coreset definition

$$cost_0 = \phi_C(CS) \geq (1 - \epsilon) \cdot \phi_C(P_0)$$

where ϵ is the approximation factor of coreset CS. So for dataset P_0, the estimation cost EstCost is greater than the k-means cost $\phi_C(P_0)$.

At time i, denote C_i as the cluster centers maintained and $\texttt{EstCost}_i$ as the estimation of k-means cost. Assume the statement is true such that $\texttt{EstCost}_i > \phi_{C_i}(P_i)$.

Consider when a new point p comes, c_p is the nearest center in C_i to p. We compute c'_p which is the new position of the center c_p, let C_{i+1} denote the new center set where $C_{i+1} = C_i \setminus \{c_p\} \cup \{c'_p\}$.

We know that

$$\phi_{C_i}(S(i)) + \|p - c'_p\|^2 \geq \phi_{C_{i+1}}(S(i))$$

As c'_p is the centroid of c_p and p, we have

$$\|p - c'_p\| < \|p - c_p\| = min_{c \in C_i} \|p - c\|$$

So c'_p is the nearest center in C_{i+1} to p. Adding up together, we get:

$$\phi_{C_i}(S(i)) + \|p - c_p\|^2 \geq \phi_{C_{i+1}}(S(i+1))$$

Algorithm 8: The Online Coreset Cache: A hybrid of CC and Sequential k-means algorithms

def *OnlineCC-Init*(k, ϵ, α)
  Remember coreset approximation factor ϵ, merge-degree r, and parameter $\alpha > 1$ the threshold to switch the query processing to CC
  ▷ *C is the current set of cluster centers*
  Initialize C by running k-means++ on set $\mathcal{S}_0$ consisting of the first $O(k)$ points of the stream
  ▷ *$cost_0$ is the clustering cost during the previous "fallback" to CC; EstCost is an estimate of the clustering cost of C on the stream so far*
  $cost_0 \leftarrow$ clustering cost of C on $\mathcal{S}_0$
  EstCost $\leftarrow cost_0$
  $Q \leftarrow$ CC-Init(r, k, ϵ)
▷ *On receiving a new point p from the stream*
def *OnlineCC-Update*(p)
  Assign p to the nearest center c_p in C
  EstCost $\leftarrow$ EstCost $+ \|p - c_p\|^2$
  ▷ *c'_p is the centroid of c_p and p where w is the weight of c_p*
  $c'_p \leftarrow (w \cdot c_p + p)/(w + 1)$
  Update center c_p in C to c'_p
  Add p to the current batch b. If $|b| = m$, then execute Q.CC-Update(b)
def *OnlineCC-Query*()
  if *EstCost* $> \alpha \cdot cost_0$ **then**
    $CS \leftarrow Q$.CC-Coreset() $\cup\, b$, where b is the current batch that has not been inserted into Q
    $C \leftarrow k$-means++(k, CS)
    $cost_0 \leftarrow \phi_C(CS)$, the k-means cost of C on CS
    EstCost $\leftarrow cost_0/(1 - \epsilon)$
  return C

From the assumption $\texttt{EstCost}_i \geq \phi_{C_i}(S(i))$, we prove that $\texttt{EstCost}_{i+1} \geq \phi_{C_{i+1}}(S(i+1))$. ■

Lemma 20 *When queried after observing point set P, the OnlineCC algorithm (Algorithm 8) returns a set of k points C whose clustering cost is within $O(\log k)$ of the optimal k-means clustering cost of P, in expectation.*

Proof: Let $\phi^*(P)$ denote the optimal k-means cost for P. Our goal is to show that $\phi_C(P) = O(\log k)\phi^*(P)$. There are two cases for handling the query.

Case I: When C is directly retrieved from CC, using Lemma 15 we have $\mathbf{E}[\phi_C(P)] \leq O(\log k) \cdot \phi^*(P)$. This case is handled through the correctness of CC.

Case II: The query algorithm does not fall back to CC. We first note from Lemma 19 that $\phi_C(P) \leq$ EstCost. Since the algorithm did not fall back to CC, we have EstCost $\leq \alpha \cdot cost_0$. Since $cost_0$ was the result of applying CC to P_0, we have from Lemma 15 that $cost_0 \leq O(\log k) \cdot \phi^*(P_0)$. Since $P_0 \subseteq P$, we know that $\phi^*(P_0) \leq \phi^*(P)$. Putting together the above four inequalities, we have $\phi_C(P) = O(\log k) \cdot \phi^*(P)$.

■

5 Experimental Evaluation

In this section, we present results from an empirical evaluation of the performance of algorithms proposed in this paper. Our goals are twofold: to understand the relative clustering accuracy and running time of different algorithms in the context of continuous queries, and to investigate how they behave under different settings of parameters.

5.1 Datasets

We work with the following real-world or semi-synthetic datasets, based on data from the UCI Machine Learning Repositories [17], all of which have been used in the past for benchmarking clustering algorithms. A summary of the datasets used appears in Table 3.

The *Covtype* dataset models the forest cover type prediction problem from cartographic variables. The dataset contains 581,012 instances and 54 integer attributes. The *Power* dataset measures electric power consumption in one household with a one-minute sampling rate over a period of almost four years. We remove the instances with missing values, resulting in a dataset with 2,049,280 instances and 7 real attributes. The *Intrusion* dataset is the 10% subset of the *KDD Cup 1999* data. The competition task was to build a predictive model capable of distinguishing between normal network connections and intrusions. We ignore symbolic attributes, resulting in a dataset with 494,021 instances and 34 real attributes. For the above datasets, to erase any potential special ordering within data, we randomly shuffle each dataset before consuming it as a data stream.

The above datasets, as well as most datasets used in previous works on streaming clustering have been static datasets that have been converted into streams by reading them in some sequence. To better model the evolving nature of data streams and drift in location of centers, we generate a semi-synthetic dataset that we call `Drift` based on the `USCensus1990` dataset from [17]. The method of data generation is as follows, and is inspired by [18]. We first cluster the `USCensus1990` dataset to compute 20 cluster centers and for each cluster, the standard deviation of the distances to the cluster center. Then we generate the synthetic dataset using the Radial Basis Function (RBF) data generator from the MOA stream mining framework [19]. The RBF generator moves the drifting centers with a user-given direction and speed. For each time step, the RBF generator creates 100 random points around each center using a Gaussian distribution with the cluster standard deviation. In total, the synthetic dataset contains 200,000 and 68 real attributes.

Dataset	Number of Points	Dimension
Covtype	581,012	54
Power	2,049,280	7
Intrusion	494,021	34
Drift	200,000	68

Table 3: Summary of Datasets.

5.2 Experimental Setup and Implementation Details

We implemented all the clustering algorithms using Java, and ran experiments on a desktop with Intel Core i5-4460 3.2GHz processor and 8GB main memory.

Algorithms: For comparison, we used two prominent streaming clustering algorithm. (1) One is the `Sequential k-means` algorithm due to [7], which is also implemented in clustering packages today. For `Sequential k-means` clustering, we used the implementation in Apache Spark MLLib [10], except that our implementation is sequential. The initial centers are set by the first k points in the stream instead of setting by random Gaussians, because that may cause some clusters be empty. (2) We also implemented `streamkm++` [1], a current state-of-the-art algorithm with good practical performance. `streamkm++` can be viewed as a special case of `CT` where the merge degree r is 2. The bucket size is set to be $10k$ where k is the number of centers. [2]

For `CC`, we set the merge degree to 2, in line with `streamkm++`. For `RCC`, the maximum nesting depth is 3, so the merge degrees for different structures are $N^{\frac{1}{2}}$, $N^{\frac{1}{4}}$ and $N^{\frac{1}{8}}$ respectively, where N is the total number of buckets. For `OnlineCC`, the threshold α is set to 1.2. To compute the `EstCost` after each fall back to `CC`, we need to know the value of cluster standard deviation D. This value is estimated using the coreset which stands for the whole points received.

Finally, as the baseline on the accuracy of stream clustering algorithms, we use the batch k-`means++` algorithm, which is expected to outperform every streaming algorithm. We report the median error due to five independent runs of each algorithm for each setting. The same applies to runtime as well.

To compute a coreset, we use the k-`means++` algorithm (similar to [1, 2]). Note since the size of the coreset is greater than k, k-`means++` is used with multiple centers chosen in each iteration, to control the number of iterations. We also use k-`means++` as the final step to construct k centers of from the coreset, and take the best clustering out of five independent runs of k-`means++`; each instance of k-`means++` is followed by up to 20 iterations of Lloyd's algorithm to further improve clustering quality. The number of clusters k is chosen from the set $\{10, 15, 20, 25, 30\}$.

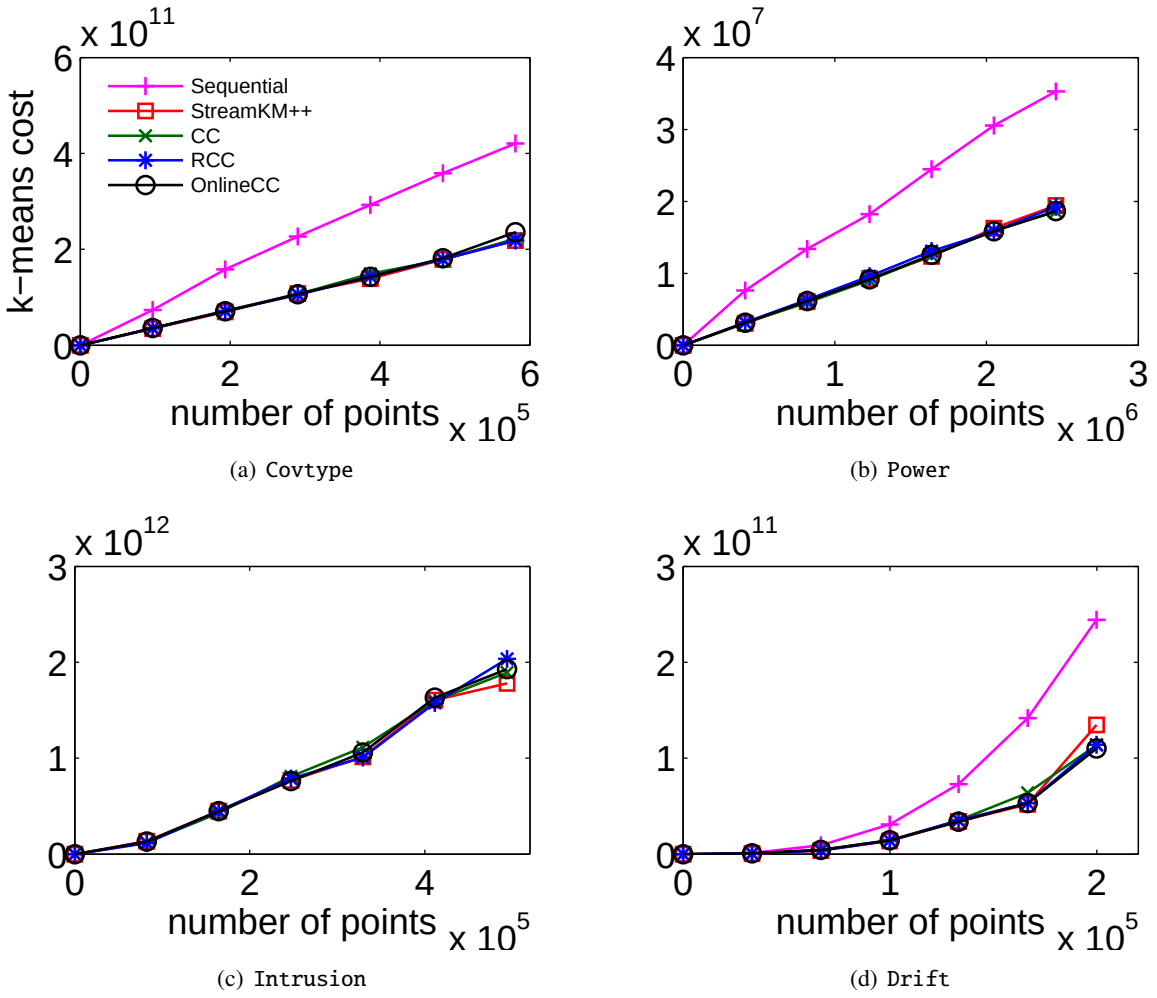
Metrics: We evaluate the clustering accuracy through the standard within cluster sum of squares (SSQ) metric, which is also the k-`means` objective function. We also measure the runtime of each algorithm and the runtime is measured for the entire datasets. Further, runtime is split into two parts, (1) update time, the time required to update internal data structures upon receiving new data (2) query time, the time required to answer the clustering queries. There is a query posed for cluster centers for every q points observed.

5.3 Discussion of Experimental Results

Accuracy (k-means cost): Consider Figures 3 and 4. Figure 4 shows the k-`means` cost versus k when the query interval is 100 points. For the `Intrusion` data, the result of `Sequential k-means` is not shown since its cost much larger (by a factor of about 10^5) than the other methods. Not surprisingly, for all algorithms studied, the clustering cost decreases with k. For all the datasets, `Sequential k-means` always achieves the highest k-`means` cost, in some cases (such as `Intrusion`), much higher than other methods. This shows that `Sequential k-means` is consistently worse than the other methods, when it comes to clustering accuracy – this is as expected, since unlike the other methods, `Sequential k-means` does not have a theoretical guarantee on clustering quality. A similar trend is also observed on the plot with the k-means cost versus the number of points received, Figure 3.

The other algorithms, `streamkm++`, `CC`, `RCC`, and `OnlineCC` all achieve very similar clustering cost, on all data sets. In Figure 4, we also show the cost of running a batch algorithm k-`means++` (followed by iterations of Lloyd's algorithm). We found that the clustering costs of the streaming algorithms are nearly the same as that of running the batch algorithm, which can see the input all at once! Indeed, we cannot

[2] A larger bucket size such as $200k$ can yield slightly better clustering quality. But this led to a high runtime for `streamkm++`, especially when queries are frequent, hence we stay with a smaller bucket size.



(a) Covtype (b) Power

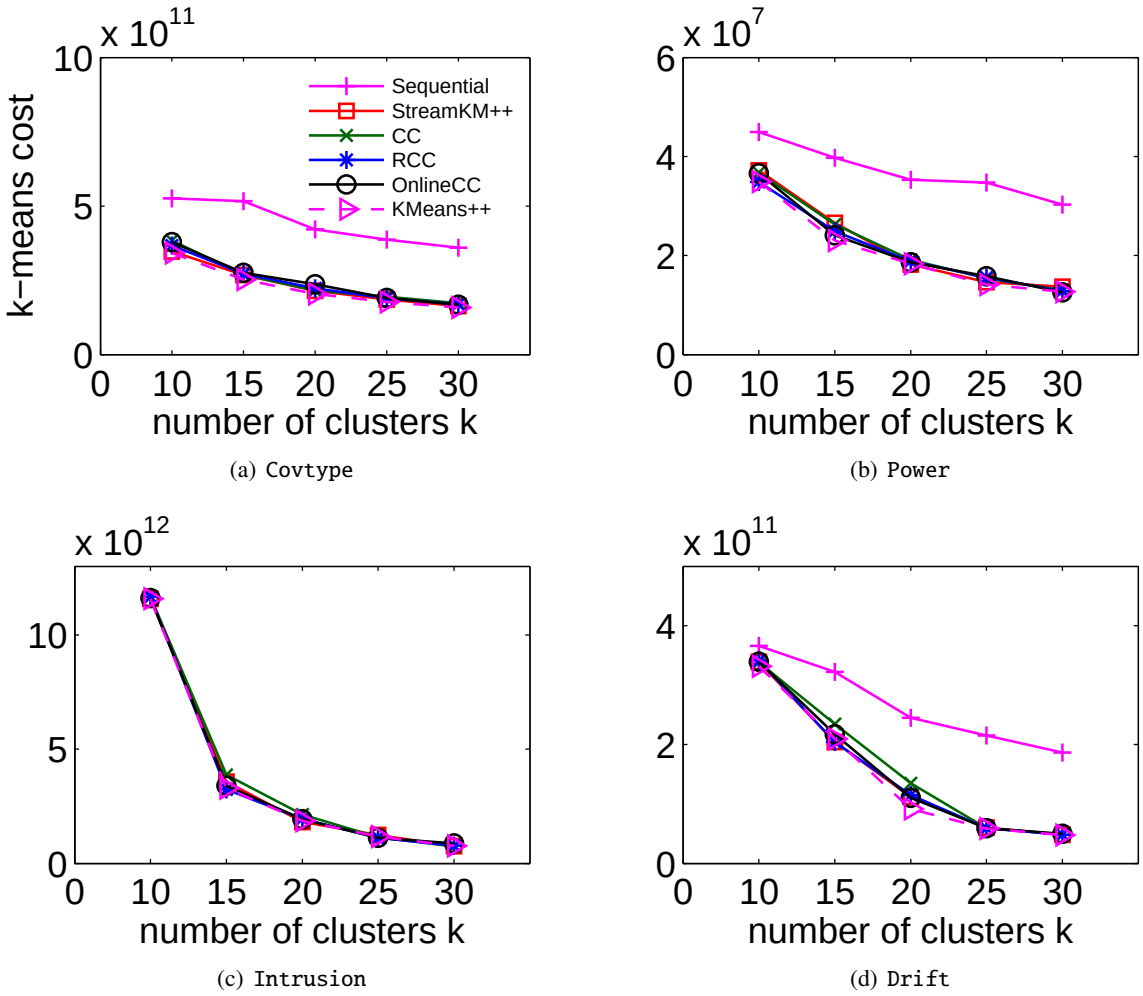
(c) Intrusion (d) Drift

Figure 3: k-means cost vs. number of points. The number of centers $k = 20$. The k-means cost of Sequential k-means on Intrusion dataset is not shown in Figure (c), since it was orders of magnitude larger than the other algorithms.

expect the streaming clustering algorithms to perform any better than this.

According to theory, as the merge degree r increases, the clustering accuracy increases. With this reasoning, we should see RCC achieve the highest clustering accuracy (lowest clustering cost), better than that of CC and streamkm++; e.g. for Covtype when k is 20, the merge degree of RCC is 53, compared with 2 for streamkm++. But our experimental results do not show such behavior, and RCC and streamkm++ show similar accuracy. Further, their accuracy matches that of batch k-means++. A possible reason for this may be that our theoretical analyses of streaming clustering methods is too conservative, and/or there is structure within real data that we can better exploit to predict clustering accuracy.

Update Time: Figure 5–7 show the results of run time versus the number of clusters, when the query interval is 100 points. Due to the inferior accuracy of Sequential k-means, the run time result is not shown in these figures. The update runtime of algorithm streamkm++, CC and OnlineCC all increase linearly with the number of centers, as the amortized update time is proportional to k. The update time of



(a) Covtype (b) Power

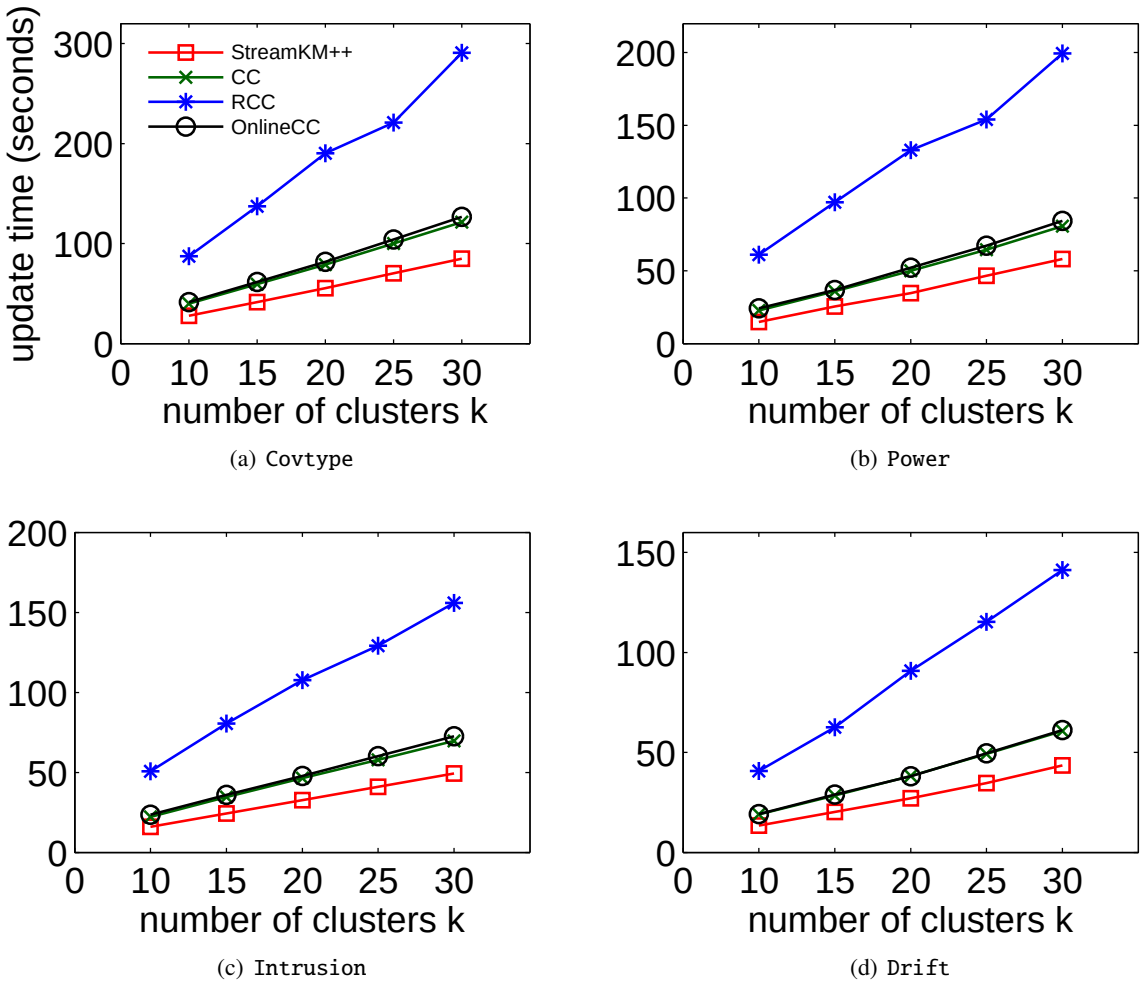
(c) Intrusion (d) Drift

Figure 4: k-means cost vs. number of centers k for different algorithms. The cost is computed at the end of observing all the points. The k-means cost of Sequential k-means on Intrusion dataset is not shown in Figure (c), since it was orders of magnitude larger than the other algorithms.

OnlineCC is nearly the same as CC, since OnlineCC-Update calls CC-Update, with a small amount of additional computation. Among the four algorithms that we compare, RCC has the largest update time, since it needs to update multiple levels of the cache as well as the coreset tree.

Query Time: From Figure 6, we see that OnlineCC has the fastest query time, followed by RCC, and CC, and finally by streamkm++. Note that the y-axis in Figure 6 is in log scale. We note that OnlineCC is significantly faster than the rest of the algorithms. For instance, it is about two orders of magnitude faster than streamkm++ for $q = 100$. This shows that the algorithm succeeds in achieving significantly faster queries than streamkm++, while maintaining the same clustering accuracy.

Total Time: Figure 7 shows the total runtime, the sum of the update time and the query time, as a function of k, for $q = 100$. For streamkm++, update time dominates the query time, hence the total time is close to its query time. For OnlineCC, however, the update time is greater than the query time, hence the total time is substantially larger than its query time. Overall, the total time of OnlineCC is still nearly 5-10



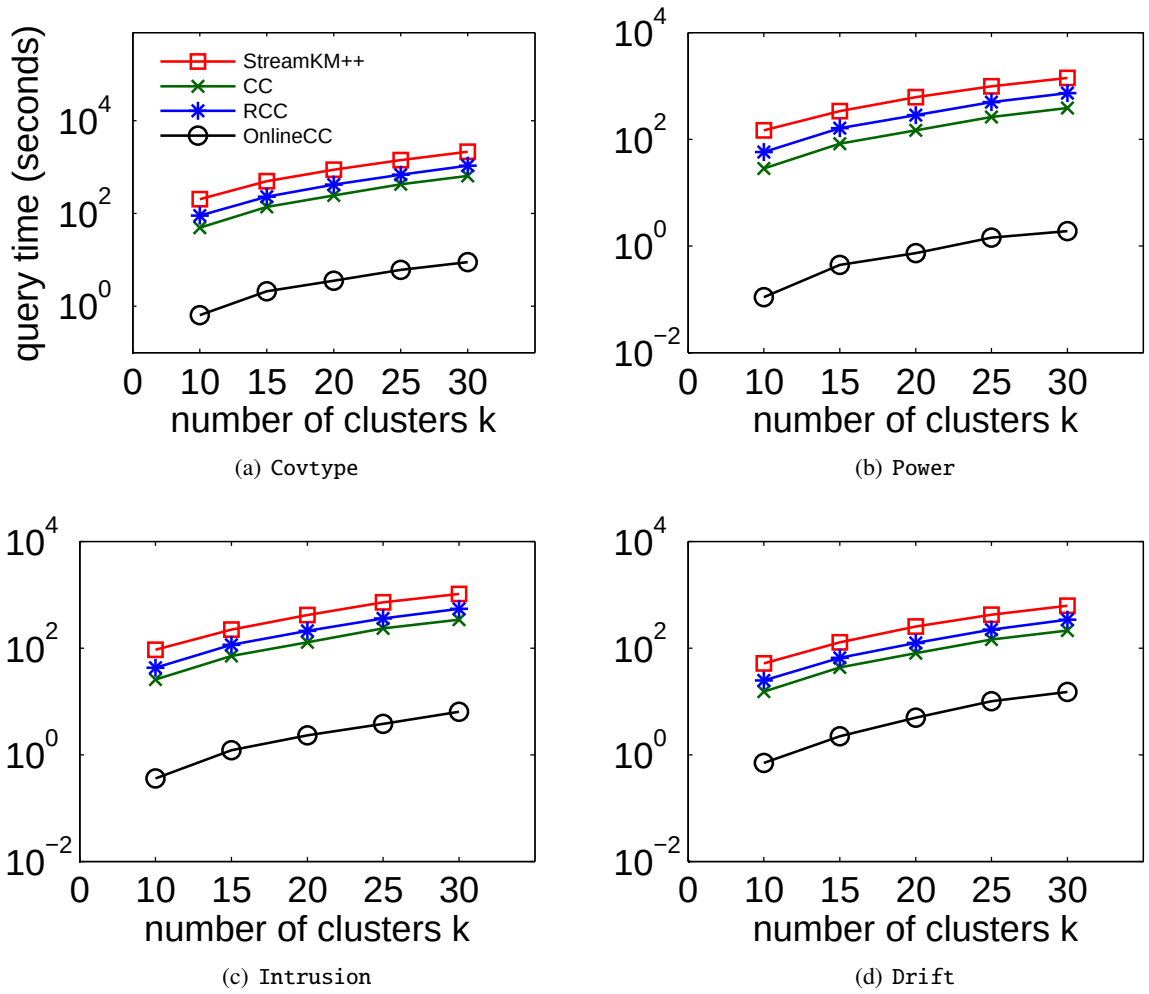
(a) `Covtype` (b) `Power`

(c) `Intrusion` (d) `Drift`

Figure 5: Update time (seconds) vs. number of centers k for different algorithms. The query interval q is 100 points.

times faster than `streamkm++`.

We next consider how the runtime varies with q, the query interval. Figure 8 shows the algorithm total run time as a function of the query interval q. Note that the update time does not change with q, and is not shown here. The trend for query time is similar to that shown for total time, except that the differences are more pronounced. We note that the total time for `OnlineCC` is consistently the smallest, and does not change with an increase in q. This is because `OnlineCC` essentially maintains the cluster centers on a continuous basis, while occasionally falling back to `CC` to recompute coresets, to improve its accuracy. For the other algorithms including `CC`, `RCC`, and `streamkm++`, the query time and the total time decrease as q increases (and queries become less frequent). As q approaches 5000, the total time stabilizes, since at this point update time dominates the query time for all algorithms.



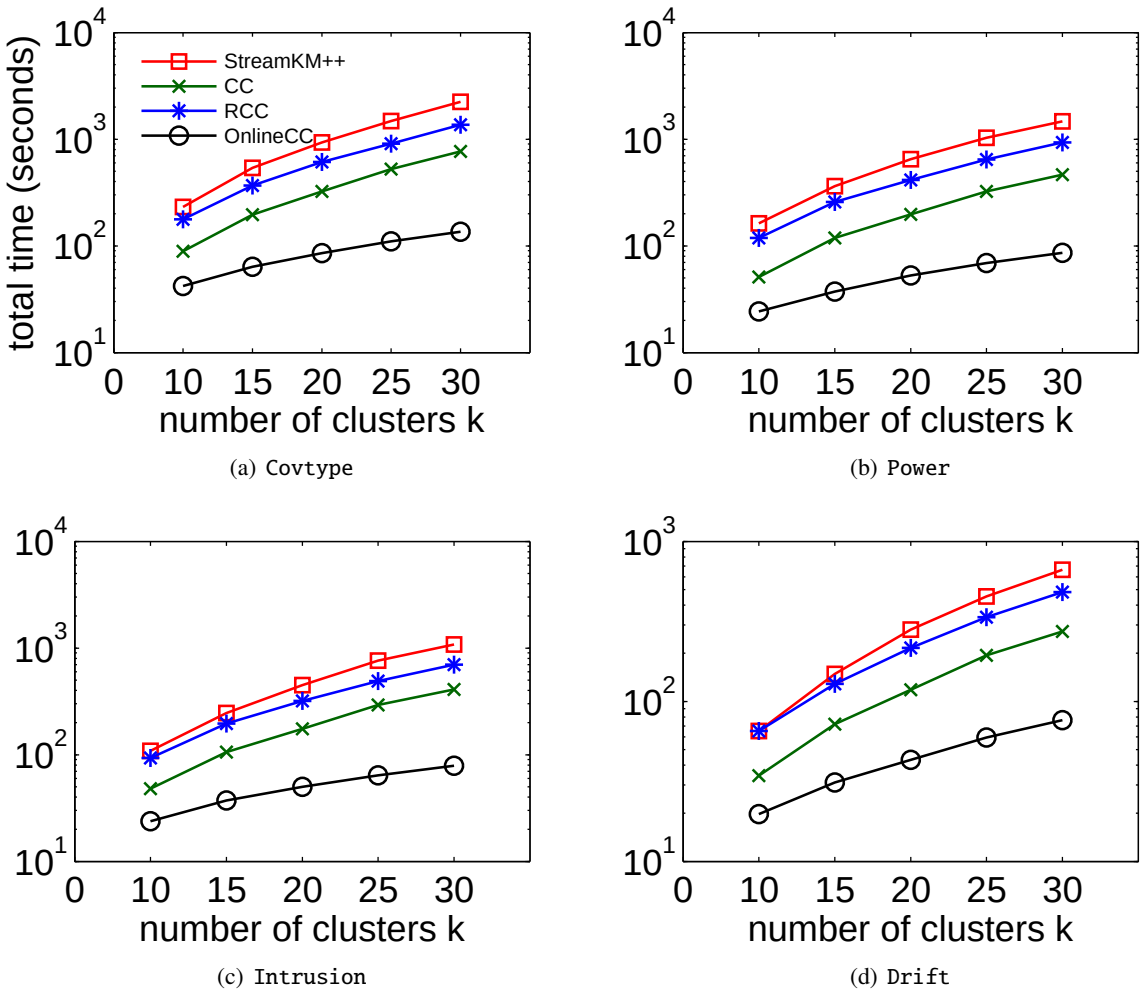
(a) Covtype (b) Power

(c) Intrusion (d) Drift

Figure 6: Query time (seconds) vs. number of centers k for different algorithms. The query interval q is 100 points.

6 Conclusion

We have presented fast algorithms for streaming k-means clustering that are capable of answering queries quickly. When compared with prior methods, our method provides a significant speedup—both in theory and practice—in query processing while offering provable guarantees on accuracy and memory cost. The general framework that we present for "coreset caching" maybe applicable to other streaming algorithms that are built around the Bentley-Saxe decomposition. Many open questions remain, including (1) improved handling of concept drift, through the use of time-decaying weights, and (2) clustering on distributed and parallel data streams.



(a) Covtype (b) Power

(c) Intrusion (d) Drift

Figure 7: Total time (seconds) vs. number of centers k. Total time is the sum of update time and the query time. The query interval q is 100 points.

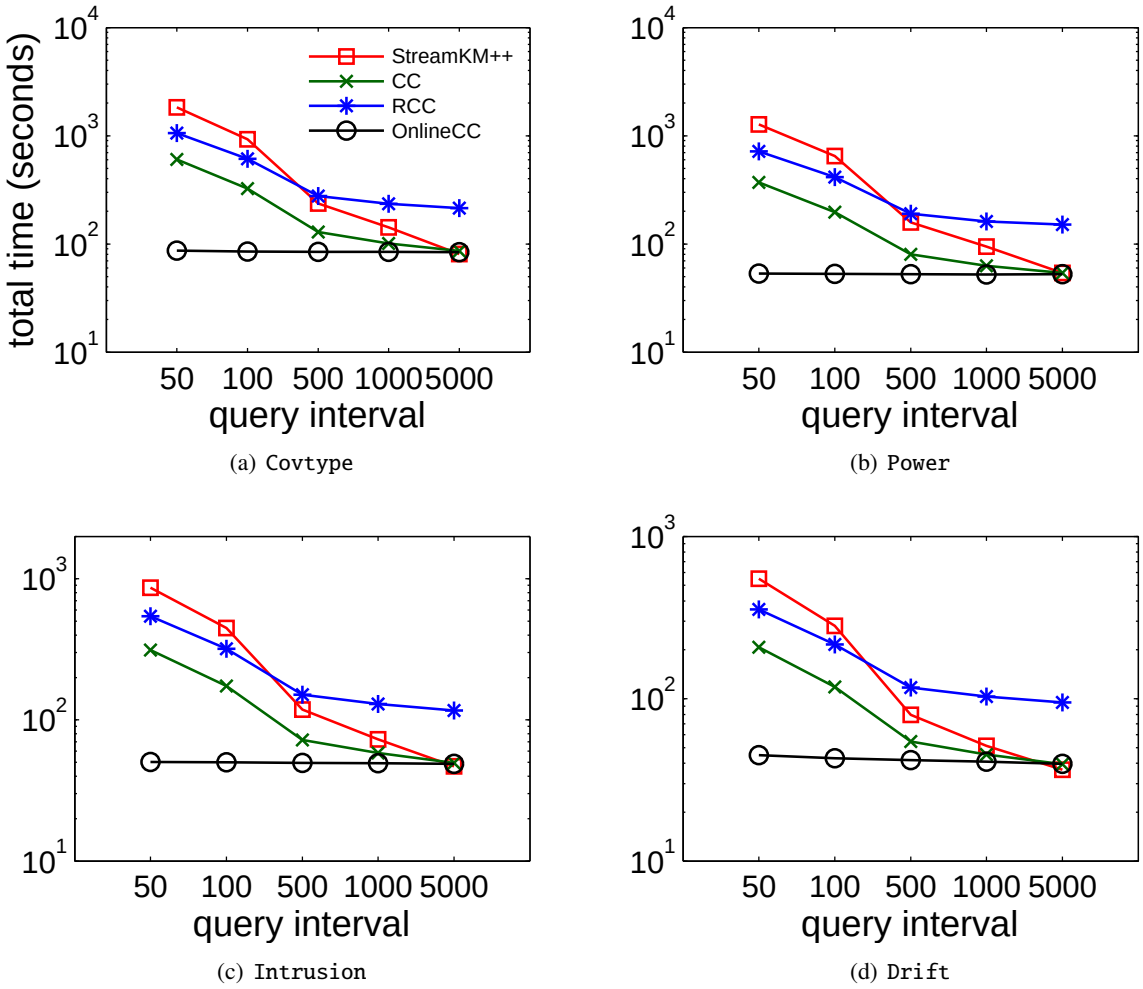
References

[1] M. R. Ackermann, M. Märtens, and C. R. et al., "Streamkm++: A clustering algorithm for data streams," *J. Exp. Algorithmics*, vol. 17, no. 1, pp. 2.4:2.1–2.4:2.30, 2012.

[2] N. Ailon, R. Jaiswal, and C. Monteleoni, "Streaming k-means approximation," in *NIPS*, 2009, pp. 10–18.

[3] S. Guha, A. Meyerson, and N. M. et al., "Clustering data streams: Theory and practice," *IEEE TKDE*, vol. 15, no. 3, pp. 515–528, 2003.

[4] M. Shindler, A. Wong, and A. Meyerson, "Fast and accurate k-means for large datasets," in *NIPS*, 2011, pp. 2375–2383.



(a) Covtype (b) Power

(c) Intrusion (d) Drift

Figure 8: Total time as a function of the query interval q. For every q points, there is a query for the cluster centers. The number of centers k is set to 20.

[5] S. Har-Peled and S. Mazumdar, "On coresets for k-means and k-median clustering," in *STOC*, 2004, pp. 291–300.

[6] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *SODA*, 2007, pp. 1027–1035.

[7] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[8] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[9] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *PVLDB*, vol. 5, pp. 622–633, 2012.

[10] X. Meng, J. K. Bradley, and B. Y. et al., "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, pp. 1235–1241, 2016.

[11] T. Kanungo, D. M. Mount, and N. S. N. et al., "A local search approximation algorithm for k-means clustering," *Computational Geometry*, vol. 28, no. 23, pp. 89 – 112, 2004.

[12] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An efficient data clustering method for very large databases," in *SIGMOD*, 1996, pp. 103–114.

[13] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *PVLDB*, 2003, pp. 81–92.

[14] J. L. Bentley and J. B. Saxe, "Decomposable searching problems i. static-to-dynamic transformation," *Journal of Algorithms*, vol. 1, pp. 301 – 358, 1980.

[15] S. Har-Peled and A. Kushal, "Smaller coresets for k-median and k-means clustering," *Discrete Comput. Geom.*, vol. 37, no. 1, pp. 3–19, 2007.

[16] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," in *STOC*, 2011, pp. 569–578.

[17] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[18] J. P. Barddal, H. M. Gomes, F. Enembreck, and J.-P. Barths, "Sncstream+: Extending a high quality true anytime data stream clustering algorithm," *Information Systems*, vol. 62, pp. 60 – 73, 2016.

[19] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 1601–1604, 2010.