Summarizing Spatial Data Streams Using ClusterHulls

JOHN HERSHBERGER
Mentor Graphics Corp.
NISHEETH SHRIVASTAVA
Bell-Labs Research
and
SUBHASH SURI
University of California, Santa Barbara

2.4

We consider the following problem: given an on-line, possibly unbounded stream of two-dimensional (2D) points, how can we summarize its spatial distribution or *shape* using a small, bounded amount of memory? We propose a novel scheme, called *ClusterHull*, which represents the shape of the stream as a dynamic collection of convex hulls, with a total of at most m vertices, where m is the size of the memory. The algorithm dynamically adjusts both the number of hulls and the number of vertices in each hull to best represent the stream using its fixed-memory budget. This algorithm addresses a problem whose importance is increasingly recognized, namely, the problem of summarizing real-time data streams to enable on-line analytical processing. As a motivating example, consider habitat monitoring using wireless sensor networks. The sensors produce a steady stream of geographic data, namely, the locations of objects being tracked. In order to conserve their limited resources (power, bandwidth, and storage), the sensors can compute, store, and exchange ClusterHull summaries of their data, without losing important geometric information. We are not aware of other schemes specifically designed for capturing shape information in geometric data streams and so we compare ClusterHull with some of the best general-purpose clustering schemes, such as CURE, k-medians, and LSEARCH. We show through experiments that ClusterHull is able to represent the shape of two-dimensional data streams more faithfully and flexibly than the stream versions of these clustering algorithms.

A preliminary version of this paper appeared in the proceedings of ALENEX '06; a partial summary of the work was presented as a poster at ICDE '06, and was represented in the proceedings by a three-page abstract. The research of Nisheeth Shrivastava and Subhash Suri was supported in part by National Science Foundation grants IIS-0121562 and CCF-0514738.
Authors' addresses: John Hershberger, Mentor Graphics Corp., 8005 SW Boeckman Road, Wilsonville, Ohio 97070, and by courtesy of Computer Science Department, University of California at Santa Barbara, Santa Barbara, California 93106; email: john.hershberger@mentor.com; Nisheeth Shrivastava, Bell-Labs Research, Bangalore, India 560095; email: nisheeth@alcatel-lucent.com; Subhash Suri, Computer Science Department, University of California at Santa Barbara, Santa Barbara, California 93106; email: suri@cs.ucsb.edu.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Convex hull, data streams, geometric data

ACM Reference Format:

Hershberger, J., Shrivastava, N., and Suri, S. 2008. Summarizing spatial data streams using ClusterHulls. ACM J. Exp. Algor. 13, Article 2.4 (September 2008), 28 pages DOI 10.1145/1412228.1412238 http://doi.acm.org 10.1145/1412228.1412238

1. INTRODUCTION

The extraction of meaning from data is perhaps the most important problem in all of science. Algorithms that can aid in this process by identifying useful structure are valuable in many areas of science, engineering, and information management. The problem takes many forms in different disciplines, but in many settings a *geometric abstraction* can be convenient: for instance, it helps formalize many informal, but visually meaningful concepts, such as similarity, groups, shape, etc. In many applications, geometric coordinates are a natural and integral part of data: e.g., locations of sensors in environmental monitoring, objects in location-aware computing, digital battlefield simulation, or meteorological data. Even when data have no intrinsic geometric association, many natural data-analysis tasks, such as clustering, are best performed in an appropriate artificial coordinate space: e.g., data objects are mapped to points in some Euclidean space using certain attribute values, where similar objects (points) are grouped into spatial clusters for efficient indexing and retrieval. Thus, we see that the problem of finding a simple characterization of a distribution known only through a collection of sample points is a fundamental one in many settings.

Recently there has been a growing interest in detecting patterns and analyzing trends in data that are generated continuously, often delivered in some fixed order and at a rapid rate. Some notable applications of such data processing include monitoring and surveillance using sensor networks, transactions in financial markets and stock exchanges, web logs and click streams, monitoring and traffic engineering of IP networks, telecommunication call records, retail and credit card transactions, and so on. Imagine, for instance, a surveillance application, where a remote environment instrumented by a wireless sensor network is being monitored through sensors that record the movement of objects (e.g., animals). The data gathered by each sensor can be thought of as a stream of 2D points (geographic locations). Given the severe resource constraints of a wireless sensor network, it would be rather inefficient for each sensor to send its entire stream of raw data to a remote base station. Indeed, it would be far more efficient to compute and send a compact geometric summary of the trajectory. One can imagine many other remote monitoring applications like forest fire hazards, marine life, etc., where the shape of the observation point cloud is a natural and useful data summary. Thus, there are many sources of "transient"

geometric data, where the key goal is to spot important trends and patterns, where only a small summary of the data can be stored, and where a "visual" summary such as shape or distribution of the data points is quite valuable to an analyst.

A common theme underlying these data-processing applications is the continuous, real-time, large-volume, transient, single-pass nature of data. As a result, *data streams* have emerged as an important paradigm for designing algorithms and answering database queries for these applications. In the data-stream model, one assumes that data arrive as a continuous stream, in some arbitrary order possibly determined by an adversary; the total size of the data stream is quite large; the algorithm may have memory to store only a tiny fraction of the stream; and any data not explicitly stored are essentially lost. Thus, data-stream processing necessarily entails *data reduction*, where most of the data elements are discarded and only a small representative sample is kept. At the same time, the patterns or queries that the applications seek may require knowledge of the entire history of the stream, or a large portion of it, not just the most recent fraction of the data. The lack of access to full data significantly complicates the task of data analysis, because patterns are often hidden and easily lost unless care is taken during the data-reduction process. For simple database aggregates, subsampling can be appropriate, but for many advanced queries or patterns, sophisticated synopses or summaries must be constructed. Many such schemes have recently been developed for computing quantile summaries [Greenwald and Khanna 2001], most frequent or top-k items [Manku and Motwani 2002], distinct item counts [Alon et al. 1996; Muthukrishnan 2005], etc.

When dealing with geometric data, an analyst's goal is often not as precisely stated as many of these *numerically oriented* database queries. The analyst may wish to understand the general structure of the data stream, look for unusual patterns, or search for certain "qualitative" anomalies before diving into a more precisely focused and quantitative analysis. The "shape" of a point cloud, for instance, can convey important qualitative aspects of a data set more effectively than many numerical statistics. In a stream setting, where the data must be constantly discarded and compressed, special care must be taken to ensure that the sampling faithfully captures the overall shape of the point distribution.

Shape is an elusive concept, which is quite challenging even to define precisely. Many areas of computer science, including computer vision, computer graphics, and computational geometry deal with representation, matching and extraction of shape. However, techniques in those areas tend to be computationally expensive and unsuited for data streams. One of the more successful techniques in processing of data streams is clustering. Clustering algorithms are mainly concerned with identifying dense groups of points, and are not specifically designed to extract the boundary features of the cluster groups. Nevertheless, by maintaining some sample points in each cluster, one can extract some information about the geometric shape of the clusters.

Of the clustering algorithms used for this type of analysis, k-medians clustering is one of the most popular. Motivated by the need for stream mining,

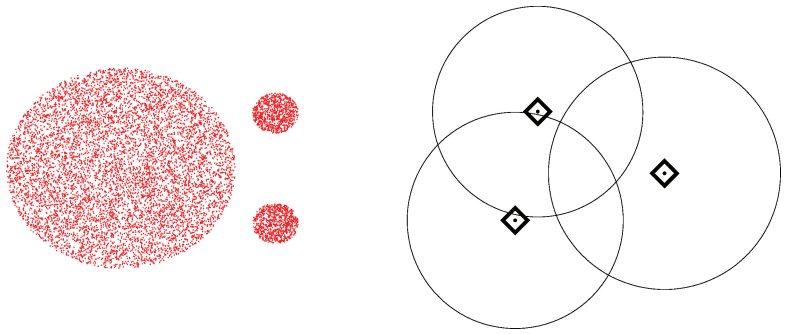


Fig. 1. The right figure shows the result of applying k-means clustering, for $k = 3$, to the input points on the left.

k-medians clustering has recently been extended to the data-stream setting [Guha et al. 2003]. In this case k-medians clustering computes k approximate cluster centers for the portion of the stream seen so far. When dealing with large data sets, cluster centers (or even cluster balls) can be a rather crude representation of the clusters. In Figure 1, we show a simple example of three natural point clusters and its three cluster centers as computed by the stream k-medians algorithm (for $k = 3$). While the stream k-medians algorithm computes only cluster centers, we also computed the radii of the three cluster balls using a second pass over the data, as shown in the figure. (Recall that the centers selected by k-medians are not necessarily points from the input set, so the rightmost center is not an input point.) We observe that even the three-ball representation is rather crude and essentially fails to capture the tight *spatial structure* of the input points.

While admittedly contrived, this example helps motivate our research goal. Traditional clustering methods, such as k-medians clustering, are optimized for determining the *center locations* of clusters and they are not concerned with the spatial shape or the structure of clusters themselves. Yet not all clusters are created equal: clusters vary in size—large versus small; they vary in *shape*—round and fat versus long and skinny; they vary in spatial organization—closely packed versus widely separated, and so on. Therefore, stream-mining applications in which such fine cluster details are important clues in data analysis can benefit from a richer form of spatial summary. We will show that the ClusterHull scheme sketched in the next section, which explicitly aims to summarize the geometric shape of the input point stream using a limited memory budget, is more effective than general-purpose stream-clustering schemes, such as CURE, k-medians, and LSEARCH.

1.1 ClusterHull

Given an on-line, possibly unbounded stream of 2D points, we propose a scheme for summarizing its spatial distribution or *shape* using a small, bounded amount of memory m. Our scheme, called *ClusterHull*, represents the shape of the stream as a dynamic collection of convex hulls, with a total of, at most, m

vertices. The algorithm dynamically adjusts both the number of hulls and the number of vertices in each hull to represent the stream using its fixed-memory budget. Thus, the algorithm attempts to capture the shape by decomposing the stream of points into groups or clusters and maintaining an approximate convex hull of each group. Depending on the input, the algorithm adaptively spends more points on clusters with complex (potentially more interesting) boundaries and fewer on simple clusters. Because each cluster is represented by its convex hull, the ClusterHull summary is particularly useful for preserving such geometric characteristics of each cluster as its boundary shape, orientation, and volume. Because hulls are objects with spatial extent, we can also maintain additional information such as the number of input points contained within each hull or their approximate *data density* (e.g., population divided by the hull volume). By shading the hulls in proportion to their density, we can then compactly convey a simple visual representation of the data distribution. By contrast, such information seems difficult to maintain in stream-clustering schemes, because the cluster centers in those schemes constantly move during the algorithm.

For illustration, in Figure 2 we compare the output of our ClusterHull algorithm with those produced by two popular stream-clustering schemes, k-medians [Guha et al. 2003] and CURE [Guha et al. 1998]. The top row shows the input data (left), and output of ClusterHull (right) with memory budget set to $m = 45$ points. The middle row shows outputs of k-medians, while the bottom row shows the outputs of CURE. One can see that both the boundary shapes and the densities of the point clusters are quite accurately summarized by the cluster hulls.

We implemented ClusterHull and experimented with both synthetic and real data to evaluate its performance. In all cases, the representation by ClusterHull appears to be more information-rich than those by clustering schemes, such as CURE, k-medians, or LSEARCH, even when the latter are enhanced with some simple mechanisms to capture cluster shape. Thus, our general conclusion is that ClusterHull can be a useful tool for summarizing geometric data streams.

ClusterHull is computationally efficient and thus well-suited for streaming data. At the arrival of each new point, the algorithm must decide whether the point lies in one of the existing hulls (actually, within a certain ring around each hull), and possibly merge two existing hulls. With appropriate data structures, this processing can be done in amortized time $O(\log m)$ per point.

ClusterHull is a general paradigm that can be extended in several orthogonal directions and adapted to different applications. For instance, if the input data are *noisy*, then covering all points by cluster hulls can lead to poor shape results. We propose an *incremental cleanup* mechanism, in which we periodically discard low-density hulls, that deals with noise in the data very effectively. Similarly, the performance of a shape summary scheme can depend on the order in which input is presented. If points are presented in a bad order, the ClusterHull algorithm may create long, skinny, interpenetrating hulls early in the stream processing. We show that a *period-doubling cleanup* is effective in correcting the effects of these early mistakes. When there is spatial coherence within the data stream, our scheme is able to exploit that coherence. For instance,

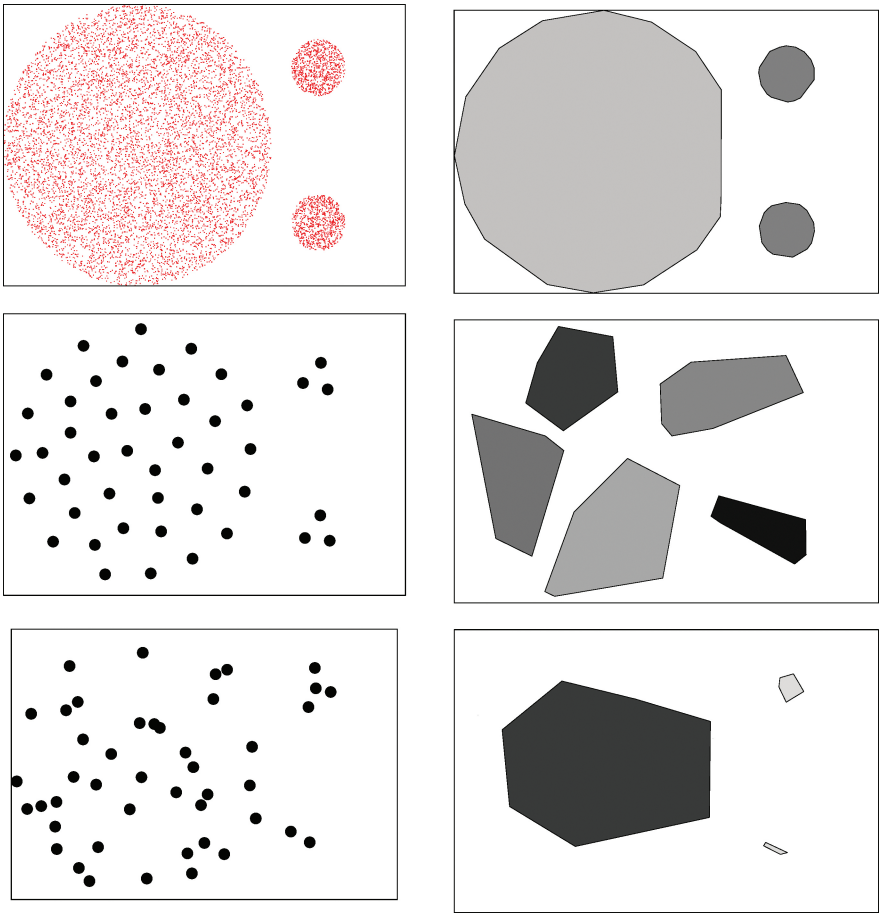


Fig. 2. The top row shows the input data (left) and the output of ClusterHull (right) with memory budget of $m = 45$. The hulls are shaded in proportion to their estimated point density. The middle row shows two different outputs of the stream k-medians algorithm, with $m = 45$: in one case (left), the algorithm simply computes $k = 45$ cluster centers; in the other (right), the algorithm computes $k = 5$ centers, but maintains nine (random) sample points from the cluster to get a rough approximation of the cluster geometry. (This is a simple enhancement we implemented to give more expressive power to the k-medians algorithm.) To maintain the uniformity of display, we show the convex hull of the sample points and shade the hulls with the corresponding density estimate (the ratio of the number of points per cluster given by the algorithms and the area of the hull). The bottom row shows the outputs of **CURE**: in the left figure, the algorithm computes $k = 45$ cluster centers; in the right figure, the algorithm computes $k = 5$ clusters, with $c = 9$ samples per cluster. CURE has a tunable shrinkage parameter, α, which we set to 0.4, in the middle of the range suggested by its authors [Guha et al. 1998].

imagine a point stream generated by a sensor field monitoring the movement of an *unknown* number of vehicles in a 2D plane. The data naturally cluster into a set of spatially coherent trajectories, which our algorithm is able to isolate and represent more effectively than general-purpose clustering algorithms.

1.2 Related Work

Inferring shape from an unordered point cloud is a well-studied problem that has been considered in many fields, including computer vision, machine learning, pattern analysis, and computational geometry [Amenta et al. 1998; Curless and Levoy 1996; Dey et al. 2000; O'Rourke and Toussaint 1997]. However, the classical algorithms from these areas tend to be computationally expensive and require full access to data, making them unsuited for use in a data-stream setting.

An area where significant progress has occurred on stream algorithms is *clustering*. Our focus is somewhat different from classical clustering—we are mainly interested in low-dimensional data and capturing the "surface" or boundary of the point cloud, while clustering tends to focus on the "volume" or density and moderate and large dimensions. While classical clustering schemes of the past have focused on cluster centers, which work well for spherical clusters, some recent work has addressed the problem of nonspherical clusters and tried to pay more attention to the geometry of the clusters. Still this attention to geometry does not extend to the *shape* of the boundary.

Our aim is not to exhaustively survey the clustering literature, which is immense and growing, but only to comment briefly on those clustering schemes that could potentially be relevant to the problem of summarizing shape of 2- or 3D point streams. Many well-known clustering schemes (e.g., [Bradley et al. 1998; Charikar et al. 1997; Ester et al. 1995; Ng and Han 1994]) require excessive computation and require multiple passes over the data, making them unsuited for our problem setting. There are machine learning-based clustering schemes [Domingos and Hulten 2001a, 2001b; Pelleg and Moore 2000] that use classification to group items into clusters. These methods are based on statistical functions and not geared toward shape representation. Clustering algorithms based on spectral methods [Cheng et al. 2005; Drineas et al. 1999; Frieze et al. 1998; Kannan et al. 2000] use the singular value decomposition on the similarity graph of the data and are good at clustering statistical data, especially in high dimensions. Work on projective clustering finds lower-dimensional clusters in higher dimensional spaces [Agarwal and Procopiuc 2000; Har-Peled and Varadarajan 2002]. "Information-theoretic clustering" can improve the representation quality of an existing clustering [Böhm et al. 2006]. We are unaware of any results showing that these methods are particularly effective at capturing boundary shapes and, more importantly, streaming versions of these algorithms are not available. Thus, we now focus on clustering schemes that work on streams and are designed to capture some of the geometric information about clusters.

One of the popular clustering schemes for large data sets is BIRCH [Zhang et al. 1996], which also works on data streams. An extension of BIRCH by Aggarwal et al. [2003] also computes multiresolution clusters in evolving streams. While BIRCH appears to work well for spherical-shaped clusters of uniform size, Guha et al. [1998] experimentally show that it performs poorly when the data are clustered into groups of unequal sizes and different shapes. The CURE-clustering scheme proposed by Guha et al. [1998] addresses this

problem and is better at identifying nonspherical clusters. CURE also maintains a number of sample points for each cluster, which can be used to deduce the geometry of the cluster. It can also be extended easily for streaming data (as noted by Guha et al. [2003]). Thus, CURE is one of the clustering schemes we compare against ClusterHull.

Guha et al. [2003] propose two stream variants of k-center clustering, with provable theoretical guarantees as well as experimental support for their performance. The stream k-medians algorithm attempts to minimize the sum of the distances between the input points and their cluster centers. Guha et al. [2003] also propose a variant where the number of clusters k can be relaxed during the intermediate steps of the algorithm. They call this algorithm LSEARCH (local search). Through experimentation, they argue that the stream versions of their k-medians and LSEARCH algorithms produce better quality clusters than BIRCH, although the latter is computationally more efficient. Since we are chiefly concerned with the quality of the shape, we compare the output of ClusterHull against the results of k-medians and LSEARCH (but not BIRCH). Recent work [Charikar et al. 2003; Chen 2006; Frahling and Sohler 2005; Har-Peled and Kushal 2005] on streaming algorithms for k-medians clustering has focused on improving the cluster quality by giving PTASs for the problem. While these algorithms improve the quantitative performance guarantees of k-medians, they do not differ qualitatively from the algorithm by Guha et al. [2003]. Our experiments compare ClusterHull against the algorithm of Guha et al. [2003], but we believe comparing against a k-medians PTAS would give qualitatively similar results.

1.3 Organization

The paper is organized in seven sections. Section 2 describes the basic algorithm for computing cluster hulls. In Section 3 we discuss the cost function used in refining and unrefining our cluster hulls. Section 4 provides extensions to the basic ClusterHull algorithm. In Sections 5 and 6, we present some experimental results. We conclude in Section 7.

2. REPRESENTING SHAPE AS A CLUSTER OF HULLS

We are interested in simple, highly efficient algorithms that can identify and maintain *bounded-memory approximations* of a stream of points. Some techniques from computational geometry appear especially well-suited for this. For instance, the *convex hull* is a useful shape representation of the *outer boundary* of the whole data stream. Although the convex hull accurately represents a convex shape with an arbitrary aspect ratio and orientation, it loses all the internal details. Therefore, when the points are distributed nonuniformly within the convex hull, the outer hull is a poor representation of the data.

Clustering schemes, such as k-medians, partition the points into groups that may represent the distribution better. However, because the goal of many clustering schemes is typically to minimize the maximum or the sum of distance functions, there is no explicit attention given to the shape of clusters—each

cluster is conceptually treated as a ball, centered at the cluster center. Our goal is to mediate between the two extremes offered by the convex hull and k-medians. We would like to combine the best features of the convex hull—its ability to represent convex shapes with any aspect ratio accurately—with those of ball-covering approximations such as k-medians—their ability to represent nonconvex and disconnected point sets. With this motivation, we propose the following measure for representing the shape of a point set under the bounded memory constraint.

> *Given a two-dimensional set of N points and a memory budget of m, where $m \ll N$, compute a set of convex hulls such that (1) the collection of hulls uses, at most, m vertices, (2) the hulls together cover all the points of S, and (3) the total area covered by the hulls is minimized.*

Intuitively, this definition interpolates between a single convex hull, which potentially covers a large area, and k-medians clustering, which fails to represent the shape of individual clusters accurately. Later, we will relax the condition of "covering all the points" to deal with *noisy data*—in the relaxed problem, *a constant fraction* of the points may be dropped from consideration. However, the general goal will remain the same: to compute a set of convex hulls that attempts to cover the important geometric features of the data stream using least possible area, under the constraint that the algorithm is allowed to use, at most, m vertices.

2.1 Geometric Approximation in Data Streams

Even the classical convex hull (outer boundary) computation involves some subtle and nontrivial issues in the data-stream setting. What should one do when the number of extreme vertices in the convex hull exceeds the memory available? Clearly, some of the extreme vertices must be dropped. But which ones and how shall we measure the *error* introduced in this approximation? This problem of summarizing the convex hull of a point stream using a fixed memory m has been studied recently in computational geometry and data streams [Agarwal et al. 2004; Chan 2004; Cormode and Muthukrishnan 2003; Feigenbaum et al. 2002; Hershberger and Suri 2004]. An adaptive sampling scheme proposed by Hershberger and Suri [2004] achieves an optimal memory-error trade-off in the following sense: given memory m, the algorithm maintains a hull that (1) lies within the true convex hull, (2) uses, at most, m vertices, and (3) approximates the true hull well—any input point not in the computed hull lies within distance $O(D/m^2)$ of the hull, where D is the diameter of the point stream. Moreover, the error bound of $O(D/m^2)$ is the best possible in the worst case.

In our problem setting, we will maintain not one but many convex hulls, depending on the geometry of the stream, with each hull roughly corresponding to a cluster. Moreover, the locations of these hulls are not determined *a priori*—rather, as in k-medians, they are dynamically determined by the algorithm. Unlike k-medians clusters, however, each hull can use a different fraction of the available memory to represent its cluster boundary. One of the key challenges in designing the ClusterHull algorithm is to formulate a good policy for this memory allocation. For this we will introduce a cost function that the various

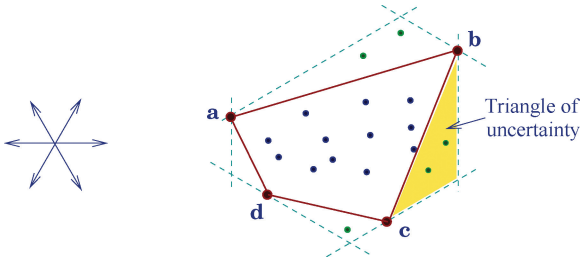


Fig. 3. An approximate hull, with 6 sampling directions. The sample hull's vertices are a, b, c, d.

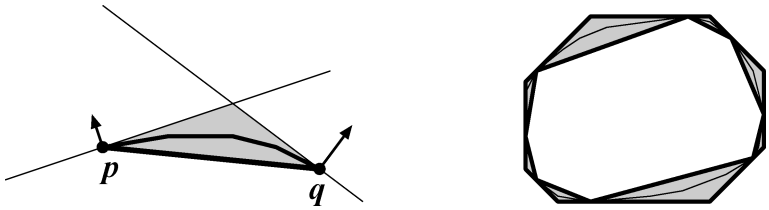


Fig. 4. The true hull is sandwiched in a ring of uncertainty triangles.

hulls use to decide how many hull vertices each gets. Let us first begin with an outline of our scheme.

2.2 The Basic Algorithm

The available memory m is divided into two pools: a fixed pool of k groups, each with a constant number of vertices, and a shared pool of $O(k)$ points, from which different cluster hulls draw additional vertices. The number k has the same rôle as the parameter fed to k-medians clustering—it is set to some number at least as large as the number of native clusters expected in the input. (Thus our representation will maintain a more refined view of the cluster structure than necessary, but simple postprocessing can clean up the unnecessary subclustering.) The exact constants in this division are tunable and we show their effect on the performance of the algorithm through experimentation. For the sake of concreteness, we can assume that each of the k groups is initially allocated eight vertices, and the common pool has a total of $8k$ vertices. Thus, if the available memory is m, then we must have $m \geq 16k$.

Our algorithm approximates the convex hull of each group by its extreme vertices in selected (sample) directions: among all the points assigned to this cluster group, for each sample direction, the algorithm retains the extreme vertex in that direction. (See Figure 3 for an example.) Each edge of this sampled hull supports what we call an *uncertainty triangle*—the triangle formed by the edge and the tangents at the two endpoints of the edge in the sample directions for which those endpoints are extreme. A simple, but important, property of the construction is that the boundary of the true convex hull is sandwiched in the ring of uncertainty triangles defined by the edges of the computed hull. (See Figure 4 for an illustration.) The extremal directions are divided into two sets, one containing uniformly spaced fixed directions, corresponding to the initial

endowment of memory, and another containing adaptively chosen directions, corresponding to additional memory drawn from the common pool. The adaptive directions are added incrementally, bisecting previously chosen directional intervals, to minimize the error of the approximation.

Each hull has an individual cost associated with it and the whole collection of k hulls has a total cost that is the sum of the individual costs. Our goal is to choose the cost function such that minimizing the total cost leads to a set of approximate convex hulls that represent the shape of the point set well. Furthermore, because our minimization is performed online, assigning each new point in the stream to a convex hull when the point arrives, we want our cost function to be *robust*: as much as possible, we want it to reduce the chance of assigning early-arriving points to hulls in a way that forces late-arriving points to incur high cost. We leave the technical details of our choice of the cost function to the following section.

Let us now describe the high-level organization of our algorithm. Suppose that the current point set S is partitioned among k convex hulls $H_1, \ldots, H_k$. The cost of hull H_i is $w(H_i)$, and the total cost of the partition $\mathcal{H} = \{H_1, \ldots, H_k\}$ is $w(\mathcal{H}) = \sum_{H \in \mathcal{H}} w(H)$. We process each incoming point p with the following algorithm:

Algorithm ClusterHull

if p is contained in any $H \in \mathcal{H}$, or in the ring of uncertainty triangles
  for any such H, **then**
  Assign p to H without modifying H.
else
  Create a new hull containing only p and add it to $\mathcal{H}$.
  if $|\mathcal{H}| > k$ **then**
    Choose two hulls $H, H' \in \mathcal{H}$ such that merging H and H' into a single
      convex hull will result in the minimum increase to $w(\mathcal{H})$.
    Remove H and H' from $\mathcal{H}$, merge them to form a new hull H^*,
      and put that into $\mathcal{H}$.
    If H^* has an uncertainty triangle over any edge joining points of the
      former H and H' whose height exceeds the previous maximum
      uncertainty triangle height, refine (repeatedly bisect) the angular interval
      associated with that uncertainty triangle by choosing new adaptive
      directions until the triangle height is less than the previous maximum.
    while the total number of adaptive directions in use exceeds ck
      Unrefine (discard one of the adaptive directions for some $H \in \mathcal{H}$) so that
        the uncertainty triangle created by unrefinement has minimum height.
  endif
endif

The last two steps (refinement and unrefinement) are technical steps for preserving the approximation quality of the convex hulls that were introduced by Hershberger and Suri [2004]. The key observation is that an uncertainty triangle with "large height" leads to a poor approximation of a convex hull.

Ideally, we would like uncertainty triangles to be flat. The height of an uncertainty triangle is determined by two key variables: the length of the convex hull edge and the angle difference between the two sampling directions that form that triangle. More precisely, consider an edge $\overline{pq}$. We can assume that the extreme directions for p and q, namely, θ_p and θ_q, point toward the same side of $\overline{pq}$, and, hence, the intersection of the supporting lines projects perpendicularly onto $\overline{pq}$. Therefore the height of the uncertainty triangle is at most the edge length $\ell(\overline{pq})$ times the tangent of the smaller of the angles between $\overline{pq}$ and the supporting lines. Observe that the sum of these two angles equals the angle between the directions θ_p and θ_q. If we define $\theta(\overline{pq})$ to be $|\theta_p - \theta_q|$, then the height of the uncertainty triangle at $\overline{pq}$ is at most $\ell(\overline{pq}) \cdot \tan(\theta(\overline{pq})/2)$, which is closely approximated by

$$\frac{\ell(\overline{pq}) \cdot \theta(\overline{pq})}{2}. \tag{1}$$

This formula forms the basis for adaptively choosing new sampling directions: we devote more sampling directions to cluster hull edges whose uncertainty triangles have large height. Refinement is the process of introducing a new sampling direction that bisects two consecutive sampling directions; unrefinement is the converse of this process. The analysis by Hershberger and Suri [2004] showed that if a single convex hull is maintained using $m/2$ uniformly spaced sampling directions and $m/2$ adaptively chosen directions (using the policy of minimizing the maximum height of an uncertainty triangle), then the maximum distance error between true and approximate hulls is $O(D/m^2)$. Because in ClusterHull we share the refinement directions among k different hulls, we choose them to explicitly minimize the global maximum uncertainty triangle height. We point out that the allocation of adaptive directions is independent of the cost function $w(\mathcal{H})$. The cost function guides the partition into convex hulls; once that choice is made, we allocate adaptive directions to minimize the error for that partition. One could imagine making the assignment of adaptive directions dependent on the cost function but, for simplicity, we have chosen not to do so.

3. CHOOSING A COST FUNCTION

In this section we describe the cost function we apply to the convex hulls that ClusterHull maintains. We discuss the intuition behind the cost function, experimental support for that intuition, and variants on the cost function that we considered.

The *α hull* is a well-known structure for representing the shape of a set of points [Edelsbrunner 1987]. It can be viewed as an extension of the convex hull in which half planes are replaced by the complements of fixed-radius disks (i.e., the regions outside the disks). In particular, the convex hull is the intersection of all half planes containing the point set and the α hull is the intersection of all disk complements with radius ρ that contain the point set.[1] See Figure 5

[1]In the definition of α hulls, the disk radius $\rho = 1/|\alpha|$, and $\alpha \leq 0$, but we are not concerned with these technical details.

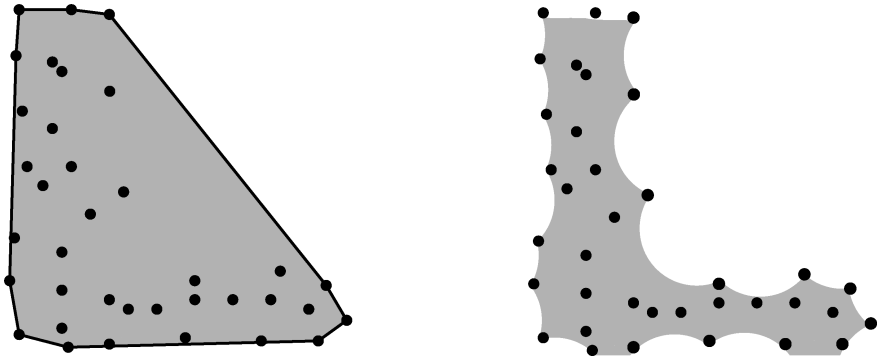


Fig. 5. Shape representations for a set of points: (left) convex hull, (right) α hull.

for examples of the convex hull and α hull on an L-shaped point set. The α hull minimizes the area of the shape that covers the points, subject to the radius constraint on the disks.

The α hull is not well suited to represent the shape of a stream of points, because an unbounded number of input points may appear on the boundary of the shape. Our goal of covering the input points with bounded-complexity convex hulls of minimum total area is an attempt to mimic the modeling power of the α hull in a data-stream setting.

Although our goal is to minimize the total area of our convex hull representation, we use a slightly more complex function as the cost of a convex hull H:

$$w(H) = \texttt{area}(H) + \mu \cdot (\texttt{perimeter}(H))^2. \tag{2}$$

Here μ is a constant, chosen empirically as described below. Note that the perimeter is squared in this expression to match units: if the perimeter term entered linearly, then simply changing the units of measurement would change the relative importance of the area and perimeter terms, which would be undesirable.

We want to minimize total area and so defining $w(H) = \texttt{area}(H)$ seems natural; however, this proves to be infeasible in a stream setting. If a point set has only two points, the area of its convex hull is zero; thus all such hulls have the same cost. The first $2k$ points that arrive in a data stream are paired up into k two-point convex hulls, each with cost zero, and the pairing will be arbitrary. In particular, some convex hulls are likely to cross natural cluster boundaries. When these clusters grow as more points arrive, they will have higher cost than the optimal hulls that would have been chosen by an off-line algorithm. This effect is clearly visible in the clusters produced by our algorithm in Figure 7 (right) for the *ellipses* data set of Figure 6 (right). By contrast, the L-shaped distribution of Figure 6 (left) is recovered well using the area cost function, as shown in Figure 7 (left).

We can avoid the tendency of the area cost to create long thin needles in the early stages of the stream by minimizing the perimeter. If we choose

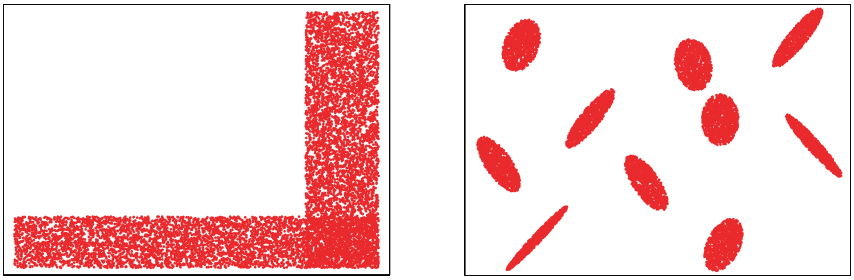


Fig. 6. Input distributions: L-shaped and *ellipses*.

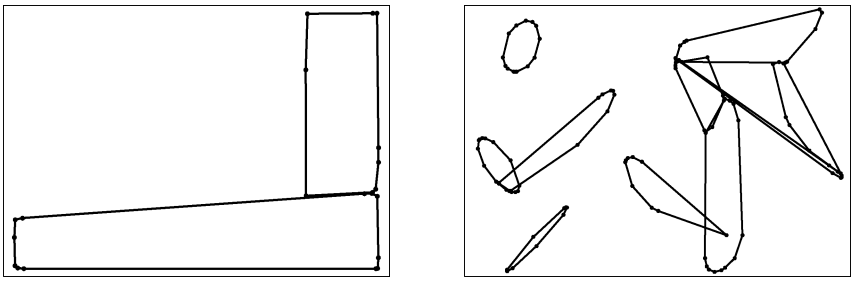


Fig. 7. With the area cost function, ClusterHull faithfully recovers the L-shaped distribution of points. However, it performs poorly on a set of $n = 10{,}000$ points distributed among ten elliptical clusters; it merges pairs of points from different groups and creates intersecting hulls.

$w(H) = \texttt{perimeter}(H)$, then the well-separated clusters of the *ellipses* data set are recovered perfectly, even when the points arrive on-line—see Figure 8 (right).

However, as the poor recovery of the L distribution shows (Figure 8 (left)), the perimeter cost has its own liabilities. The total perimeter of two hulls that are relatively near each other can often be reduced by merging the two into one. Furthermore, merging two large hulls reduces the perimeter more than merging two similar small ones, and so the perimeter cost applied to a stream often results in many small hulls and a few large ones that contain multiple "natural" clusters.

We need to incorporate both area and perimeter into our cost function to avoid the problems shown in Figures 7 and 8. Because our overall goal is to minimize area, we choose to keep the area term primary in our cost function (Equation 2). In that function, $(\texttt{perimeter}(H))^2$ is multiplied by a constant μ, which is chosen to adjust the relative importance of area and perimeter in the cost. Experimentation shows that choosing $\mu = 0.05$ gives good shape reconstruction on a variety of inputs. With μ substantially smaller than 0.05, the perimeter effect is not strong enough and, with μ greater than 0.1, it is too strong. (Intuitively, we want to add just enough perimeter dependence to avoid creating needle convex hulls in the early stages of the stream.)

We can understand the combined area–perimeter cost by modeling it as the area of a fattened convex hull. If we let $\rho = \mu \cdot \texttt{perimeter}(H)$, we see that the area–perimeter cost (2) is very close to the area obtained by fattening H by ρ.

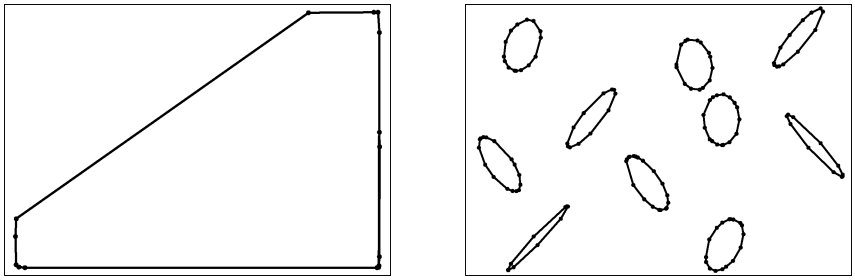


Fig. 8. With the perimeter cost function, ClusterHull faithfully recovers the disjoint elliptical clusters, but performs poorly on the L-shaped distribution.

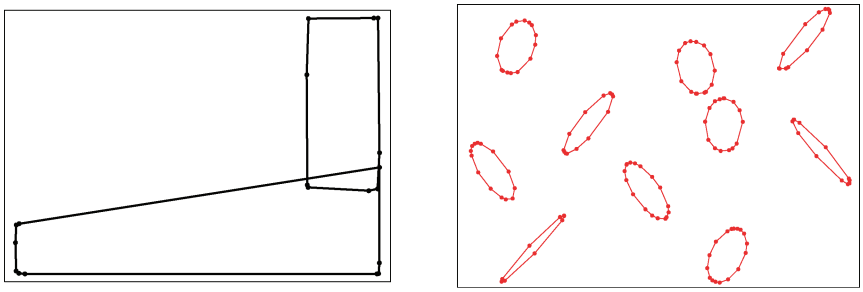


Fig. 9. With the combined area and perimeter cost function, the algorithm ClusterHull recovers both the ellipses and L distributions. The choice of $\mu = 0.05$ gives good shape reconstruction.

The true area is $\texttt{area}(H) + \rho \cdot \texttt{perimeter}(H) + \pi\rho^2 = \texttt{area}(H) + \rho^2(\frac{1}{\mu} + \pi)$; if μ is small, then $1/\mu$ is relatively large compared to π and the extra $\pi\rho^2$ term is not very significant.

Because the cost (2) may fatten long thin clusters more than is desirable, we also experimented with replacing the constant μ in (2) by a value inversely related to the aspect ratio of H. The aspect ratio of H is $\texttt{diam}(H)/\texttt{width}(H) = \Theta((\texttt{perimeter}(H))^2/\texttt{area}(H))$. Thus, if we simply replaced μ by $1/\texttt{aspectRatio}(H)$ in (2), we would essentially obtain the area cost. We compromised by using the cost

$$w(H) = \texttt{area}(H) + \mu \cdot (\texttt{perimeter}(H))^2/(\texttt{aspectRatio}(H))^x$$

for various values of x ($x = 0.5$, $x = 0.1$). The aspect ratio is conveniently approximated as $(\texttt{perimeter}(H))^2/\texttt{area}(H)$, since the quantities in that expression are already maintained by our convex-hull approximation. Except in extreme cases, the results with this cost function were not enough different from the basic area–perimeter cost to merit a separate figure.

The cost (2) fattens each hull by a radius proportional to its own perimeter. This is appropriate if the clusters have different natural scales and we want to fatten each according to its own dimensions. However, in our motivating structure the α hull, a uniform radius is used to define all the clusters. To fatten hulls uniformly, we could use the cost function

$$w(H) = \texttt{area}(H) + \rho \cdot \texttt{perimeter}(H) + \pi\rho^2.$$

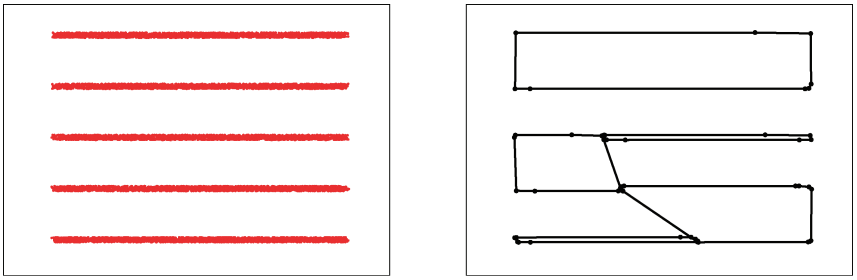


Fig. 10. Processing the *stripes* input (left) in random order leads to errors for our algorithm (right).

However, the choice of the fattening radius ρ is problematic. We might like to choose ρ such that α hulls defined using radius ρ disks form exactly k clusters, but then the optimum value of ρ would decrease and increase as the stream points arrived. We can avoid these difficulties by sticking to the simpler cost of definition (2).

4. EXTENSIONS AND ENHANCEMENTS

In this section, we discuss how to enhance the basic ClusterHull algorithm to improve the quality of shape representation.

4.1 Spatial Incoherence and Period-Doubling Cleanup

In many data streams the arriving points are ordered arbitrarily, possibly even adversarily. The ClusterHull scheme (and indeed any on-line clustering algorithm) is vulnerable to early errors,[2] in which an early-arriving point is assigned to a hull that later proves to be the wrong one.

Figure 10 (left) shows a particularly bad input consisting of points located in five thin parallel stripes. We used ClusterHull with $\mu = 0.05$ to maintain five hulls, with the input points ordered randomly. A low-density sample from the stripe distribution (such as a prefix of the stream) looks to the algorithm very much like uniformly distributed points. Early hull merges combine hulls from different stripes and the ClusterHull algorithm cannot recover from this mistake. (See Figure 10 (right).)

If the input data arrive in random order, the idea of *period-doubling cleanup* may help identify and amplify the true clusters. The idea is to process the input stream in *rounds* in which the number of points processed doubles in each round. At the end of each round, we identify low-density hulls and discard them—these likely group points from several true clusters. The dense hulls are retained from round to round and are allowed to grow.

Formally, the period-doubling cleanup operates as follows: For each $H \in \mathcal{H}$ we maintain the number of points it represents, denoted by $\texttt{count}(H)$. The

[2]For an evolving stream, the optimal clusters for the first one-half of the stream may be very different from those for the entire stream. Decisions made while processing the first one-half (e.g., cluster locations, point assignments) may lead to suboptimal clustering for the whole stream. Even for clustering methods with performance guarantees, clustering performance will be closer to the worst-case bound if the stream is adversarily ordered.

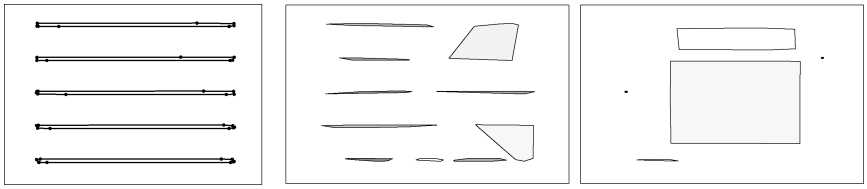


Fig. 11. Period-doubling cleanup (left) on ClusterHull corrects the errors in the *stripes* distribution; the middle figure shows the output of k-medians and the right figure shows the output of CURE.

density of any hull H is $\texttt{density}(H) = \texttt{count}(H)/\texttt{area}(H)$. The algorithm also maintains an approximate convex hull G of all the input points. After each round, it discards from $\mathcal{H}$ every hull H for which any of the following holds:

—$\texttt{count}(H) < \delta \cdot N/k$
—$\texttt{density}(H) < \texttt{density}(G)$
—$\texttt{density}(H) < \frac{1}{2} \cdot \texttt{median}\{\texttt{density}(\texttt{A}) : \texttt{A} \in \mathcal{H}\}$

Here N is the number of points seen so far. In our experiments, we set the tunable parameter δ to 0.1.

The first test takes care of hulls with a very small number of tightly clustered points (these may have high densities because of their smaller area and will not be caught by density pruning). The second test discards hulls that have less than average density. The intuition is that each cluster should be at least as dense as the entire input space (otherwise it is not an interesting cluster). In case the points are distributed over a very large area, but the individual clusters are very compact, the average density may not be very helpful for discarding hulls. Instead, we should discard hulls that have low densities relative to other hulls in the data structure; the third test takes care of this case—it discards any hull with density significantly less than the median density. As a caveat, we note that the period-doubling cleanup is counterproductive in cases when the natural clusters in the input have very different cardinalities or densities and the user wishes to preserve the low cardinality/density clusters in the output.

Figure 11 (left) shows the result of period-doubling cleanup on the *stripes* distribution; the sparse hulls that were initially found have been discarded and five dense hulls have been correctly computed. We note that, with the same amount of memory, neither CURE nor the k-medians clustering is able to represent the *stripes* distribution well (cf. Figure 11). Our experiments show that applying period-doubling cleanup helps improve clustering on almost all data sets.

4.2 Noisy Data and Incremental cleanup

Sampling error and outliers cause difficulty for nearly all clustering algorithms. Likewise, a few outliers can adversely affect the ClusterHull shape summary. An algorithm needs some way to distinguish between dense regions of the input distribution (the true clusters) and sparse ones (noise). In this section,

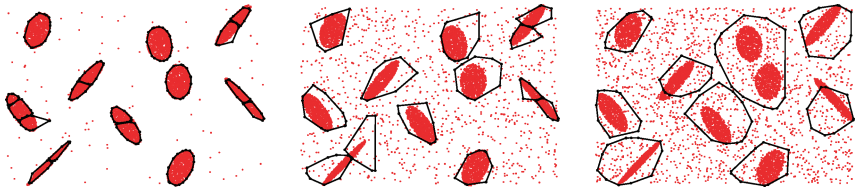


Fig. 12. Incremental cleanup, with estimated noise frequency $\epsilon = 0.1$, applied to distributions with 1, 10, and 20% actual background noise.

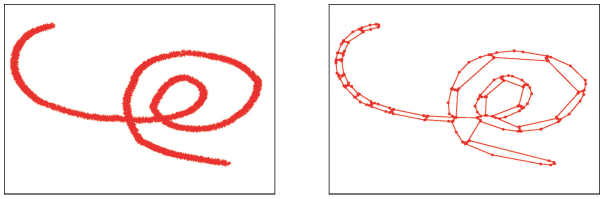
we propose an *incremental cleanup* mechanism that can improve the performance of our algorithm in the presence of noise. Both the period-doubling and the incremental cleanup are inspired by sampling techniques used in frequency estimation in streams. In particular, period-doubling is inspired by sticky sampling and incremental cleanup is inspired by lossy counting [Manku and Motwani 2002]. The incremental cleanup also processes the input in rounds, but the rounds do not increase in length. This is because noise is not limited to the beginning of the input; if we increased the round length, all the hulls would be corrupted by noisy points. Instead, we fix the size of each round depending on the (estimated) noise in the input.

Specifically, the incremental cleanup assumes that the input stream consists of points drawn randomly from a fixed distribution, with roughly $(1 - \epsilon)N$ of them belonging to high-density clusters and ϵN of them being low-density noise. We do not assume that the noise is uniformly distributed, but we do assume that it is significantly less dense than the true clusters. The expected noise frequency ϵ affects the quality of the output. (It should be estimated conservatively if it is unknown.) The idea is to set the value of δ to be roughly equal to ϵ and process the input in rounds of $k/(2\epsilon)$ points. The logic is that in every round, only about $k/2$ hulls will be corrupted by noisy points, still leaving one-half of the hulls untouched and free to track the true distribution of the input. If we set k to be more than twice the expected number of natural clusters in the input, we obtain a good representation of the clusters.

This scheme propagates the good hulls (those with high density and high cardinality) from one round of the algorithm to the next, while discarding hulls that are sparse or belong to outliers. (See Figure 12 for an example of how this scheme identifies true clusters and discards noisy regions.) Of course, if noise is underestimated significantly (Figure 12 (right)), the quality of the cluster hulls suffers.

4.3 Spatial Coherence and Trajectory tracking

Section 4.1 considered spatially incoherent input streams. If the input *is* spatially coherent, as occurs in some applications, ClusterHull performs particularly well. If the input stream consists of locations reported by sensors detecting some moving entity (a light pen on a tablet, a tank in a battlefield, and animals in a remote habitat), our algorithm effectively finds a covering of the trajectory by convex "lozenges." The algorithm also works well when there are multiple simultaneous trajectories to represent, as might occur when sensors track



(a) (b)

Fig. 13. Input along a trajectory in (a); the shape is recovered well using $m = 150$ in (b).

multiple independent entities. If the stripes of Figure 10 are fed to the algorithm in left-to-right order, they are recovered perfectly; likewise, in Figure 13, a synthetic trajectory is represented accurately.

4.4 Density Estimation and Display

Stream versions of traditional clustering schemes (including k-medians and CURE) do not include an estimate of the density of points associated with each cluster center, whereas each cluster hull H can easily maintain $\texttt{count}(H)$.[3] As in Section 4.1, this gives an estimate of the density of the points in each hull. If this information is displayed graphically (cf. Figure 14) it conveys more insight about the distribution of the input data than does a simple cluster-center output or even cluster-sample output. Note that in Figures 2, 14, and 15 we have enhanced the output of k-medians and CURE to display shaded hulls, thereby making a fairer comparison to ClusterHull.

5. IMPLEMENTATION AND EXPERIMENTS

We implemented the convex-hull algorithm of Hershberger and Suri [2004] and the ClusterHull algorithm on top of it. The convex-hull algorithm takes logarithmic time per inserted point, on average, but our ClusterHull implementation is more simple-minded, optimized for ease of implementation instead of runtime performance. Because our implementation is known to be suboptimal in runtime, we did not consider performance statistics to be meaningful and did not collect them.

The bottleneck in our implementation is neighborhood queries/point location, taking time proportional to the number of hulls. By storing the hull edges in a quad tree, we could speed up these operations to $O(\log m)$ time. When a new point arrives, we must check which hull it belongs to, if any. Using a quad tree, this reduces to a logarithmic-time search,[4] followed by logarithmic

[3]Most k medians implementations *do* maintain the total distance of points to their centers, which gives another kind of "density measure."

[4]A naïve application of quad trees may take more than $O(\log m)$ time for searching, because a pathological point distribution may force the quad tree to have many levels. To cope with this potential problem a sophisticated implementation would need to use level compression, in which quad tree nodes with only one populated child are eliminated and centroid balancing, in which the quad tree is decomposed hierarchically by cutting a centroid edge and point-location searches follow the logarithmic-height centroid tree. Such refinements are beyond the scope of this paper. For more details see Aluru [2005] and Arya et al. [1998].

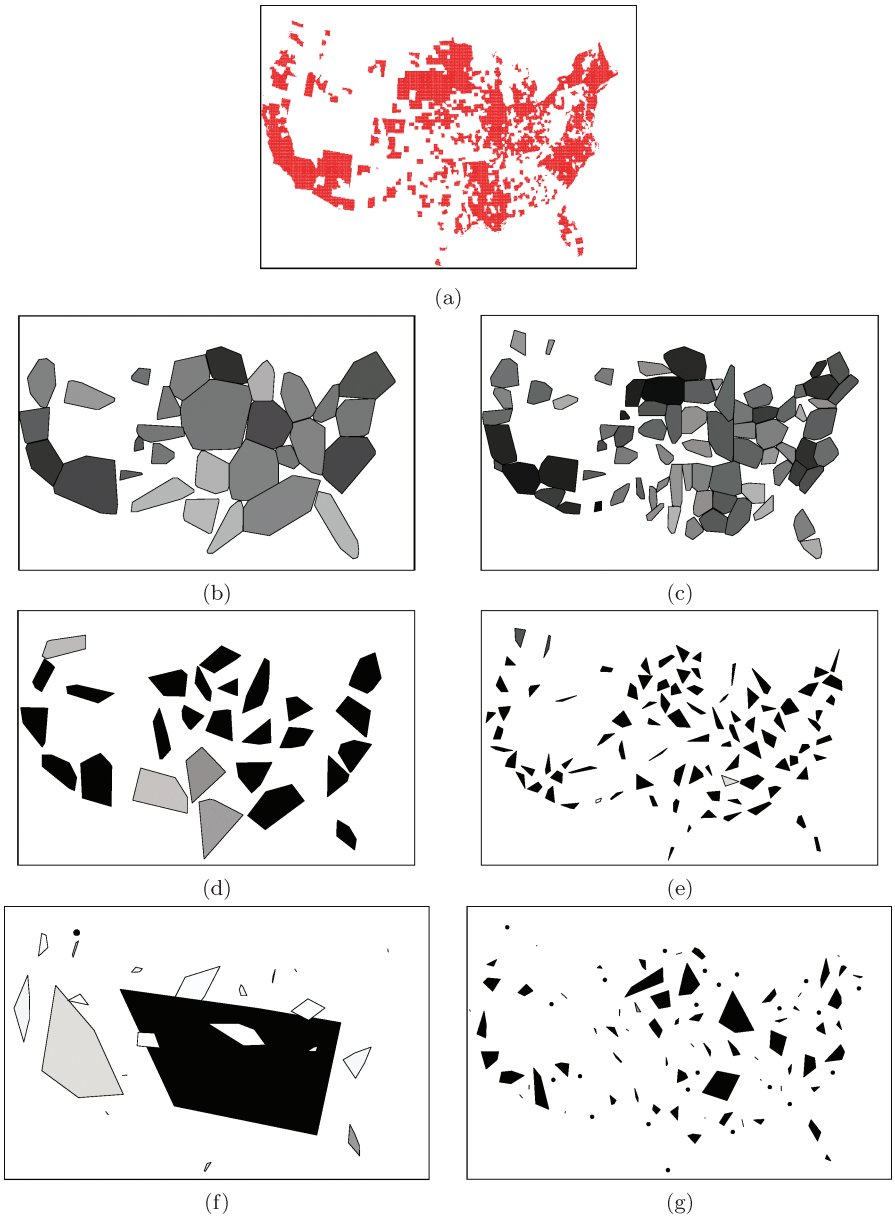


Fig. 14. The *westnile* data set is shown in the top figure (a). Figures (b) and (c) show the outputs of ClusterHull for $m = 256$ and $m = 512$. Figures (d) and (e) show the corresponding outputs for k-medians. Figures (f) and (g) show the corresponding outputs for CURE.

time point-in-polygon tests with an expected constant number of hulls. Each new hull H must compute its optimum merge cost—the minimum increment to $w(\mathcal{H})$ caused by merging H with another hull. On average, this increment is greater for more distant hulls. Using the quad tree, we can compute the increment for $O(1)$ nearby hulls first. Simple lower bounds on the incremental

cost of distant merges then let us avoid computing the costs for distant hulls. Computing the incremental cost for a single pair of hulls reduces to computing tangents between the hulls, which takes logarithmic time [Hershberger and Suri 2004].

Merging hulls eliminate some vertices forever and so we can charge the time spent performing the merge to the deleted vertices. Thus, a careful implementation of the ClusterHull algorithm would process stream points in $O(\log m)$ amortized time per point, where m is the total number of hull vertices.

In the remainder of this section we evaluate the performance of our algorithm on different data sets. When comparing our scheme with k-medians clustering [Guha et al. 2003], we used an enhanced version of the latter. The algorithm is allowed to keep a constant number of sample points per cluster, which can be used to deduce the approximate shape of that cluster. We ran the k-medians clustering using k clusters and total memory (number of samples) equal to m. CURE already has a parameter for maintaining samples in each cluster, so we used that feature. In this section, we analyze the output of these three algorithms (ClusterHull, k medians, and CURE) on a variety of different data sets, and as a function of m, the memory.

Throughout this section, we use the *period-doubling cleanup* along with the area-perimeter cost (Equation 2) to compute the hulls. We use $\mu = 0.05$ and r, the number of initial sample directions per hull, equal to 8. The values of these parameters are critical for our algorithm; however, in this section *we use the same set of parameters for all data sets.* This shows that when tuned properly, our algorithm can generate good-quality clusters for a variety of input distributions using a single set of parameters. In the next section, we will analyze in detail the effects of these parameters on the results of our scheme. To visualize the output, we also shade the hulls generated by our algorithm according to their densities (darker regions are more dense). For a fair visual comparison, we also display the convex hulls of random samples maintained by the k-medians and CURE-clustering algorithms. Since those algorithms also keep track of the number of points in each cluster, we calculated the density of each cluster and shaded the hulls accordingly. This density is an overestimate, since the sample convex hull may lie far inside the true cluster boundary, but the visual result is nonetheless better than simply displaying the cluster center and sample points.

5.1 West Nile Virus Spread

Our first data set, *westnile* (Figure 14(a)), contains about 68,000 points corresponding to the locations of the *West Nile* virus cases reported in the U.S. as collected by the CDC and the USGS [USGS 2003]. We randomized the input order to eliminate any spatial coherence that might give an advantage to our algorithm (cf. Section 4.3); however, we should note that the same randomization probably improved the effectiveness of the period-doubling cleanup. We ran ClusterHull to generate output of total size $m = 256$ and 512; k was 27 and 90 in the two cases (Figures (b) and (c)). The clustering algorithms k-medians and CURE were used to generate clusters with the same amount of memory. The results are shown in Figure 14.

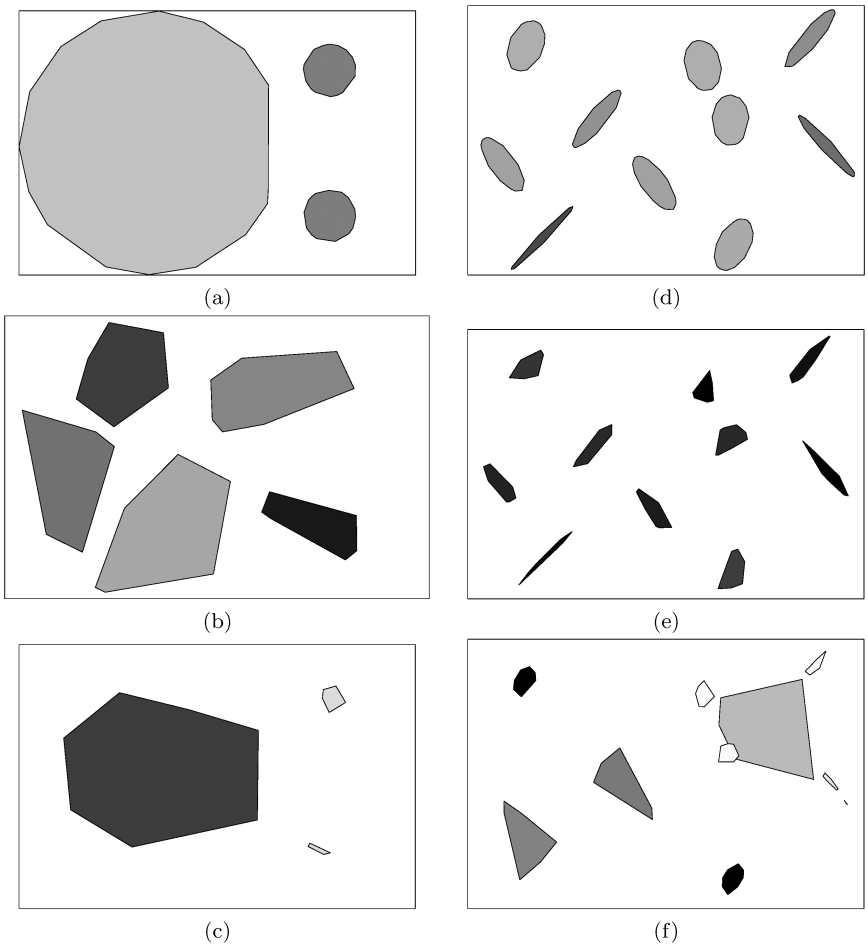


Fig. 15. The left column shows output of ClusterHull (top), k-medians (middle), and CURE (bottom) for the *circles* dataset with $m = 64$. The right column shows corresponding outputs for the *ellipses* dataset with $m = 128$.

All three algorithms are able to track high-density regions in coherent clusters, but there was little information about the shapes of the clusters in the output of k-medians or CURE. Visually the output of ClusterHull looks strikingly similar to the input set, offering the analyst a faithful, yet compact, representation of the geometric shapes of important regions.

5.2 The *Circles* and the *Ellipses* Datasets

In this experiment, we compared ClusterHull with k-medians and CURE on the *circles* and the *ellipses* datasets described earlier. The circles set contains $n = 10{,}000$ points generated inside three circles of different sizes. We ran the three algorithms with a total memory $m = 64$ and $k = 5$. The output of ClusterHull is shown in Figure 15(a); the output of k-medians is shown in (b); and the output of CURE is shown in (c).

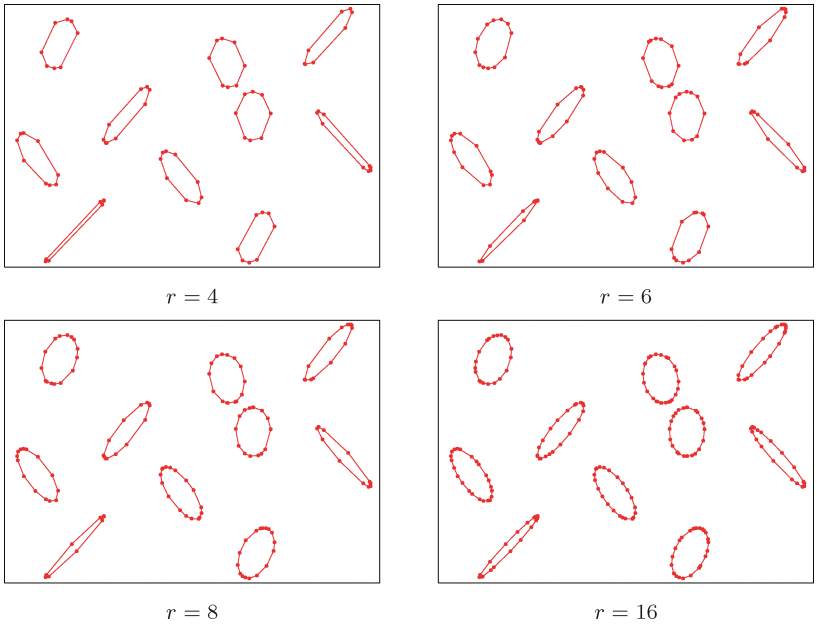


Fig. 16. The result of varying r on the *ellipses* dataset.

Similarly, Figures 15(d), (e), and (f), respectively, show the outputs of ClusterHull, k-medians, and CURE on the ellipses dataset with memory $m = 128$ and $k = 10$. The ellipses dataset contains $n = 10{,}000$ points distributed among ten ellipse-shaped clusters.

In all cases, ClusterHull output is more accurate, visually informative, and able to compute the boundaries of clusters with remarkable precision. The outputs of other schemes are ambiguous, inaccurate, and lacking in detail of the cluster shape boundary. For the circles data, the k-medians algorithm does a poor job of determining the true cluster structure. For the ellipses data, CURE does a poor job of separating the clusters. (CURE needed a much larger memory—a "window size" of at least 500 to separate the clusters correctly.)

6. TUNING CLUSTERHULL PARAMETERS

In this section, we study the effects of various parameters on the quality of clusters.

6.1 Variation with r

We first consider the effect of changing r, the number of initial directions assigned to each hull. To isolate the effects of r, we fixed the values of $\mu = 0.05$ and $k = 10$. We ran the experiments on two datasets, *ellipses* and *circles*. The results are shown in Figures 16 and 17, respectively.

The results show that the shape representation with four initial directions is very crude: ellipses are turned into pointy polygons. As we increase r, the representation of clusters becomes more refined. This contrast can be seen if we

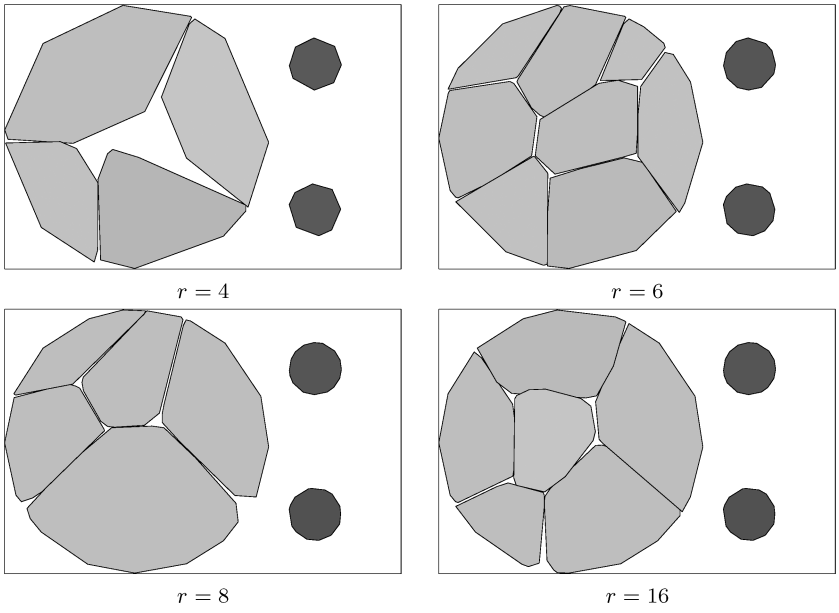


Fig. 17. The result of varying r on the *ellipses* dataset.

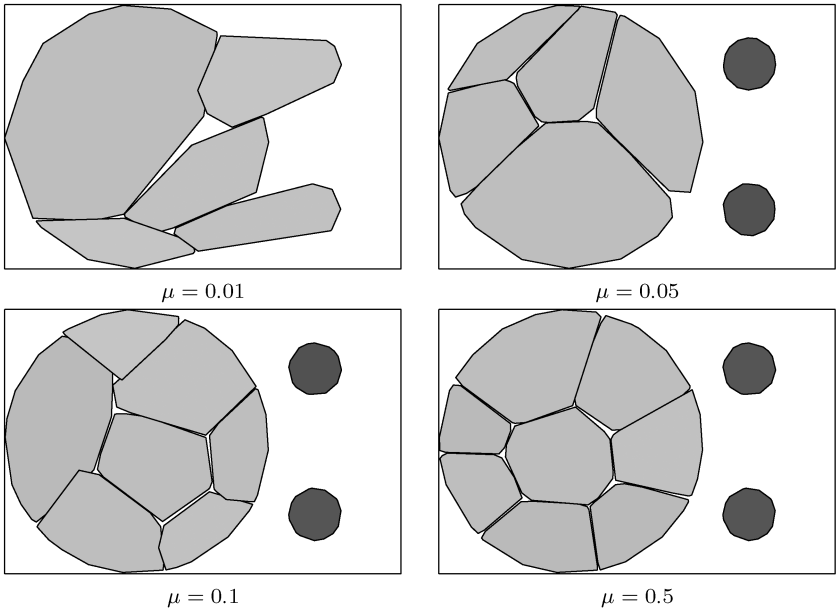


Fig. 18. For the *circles* data set, ClusterHull recovers clusters correctly for $\mu \in [0.05, 0.5]$, but fails for $\mu \leq 0.01$.

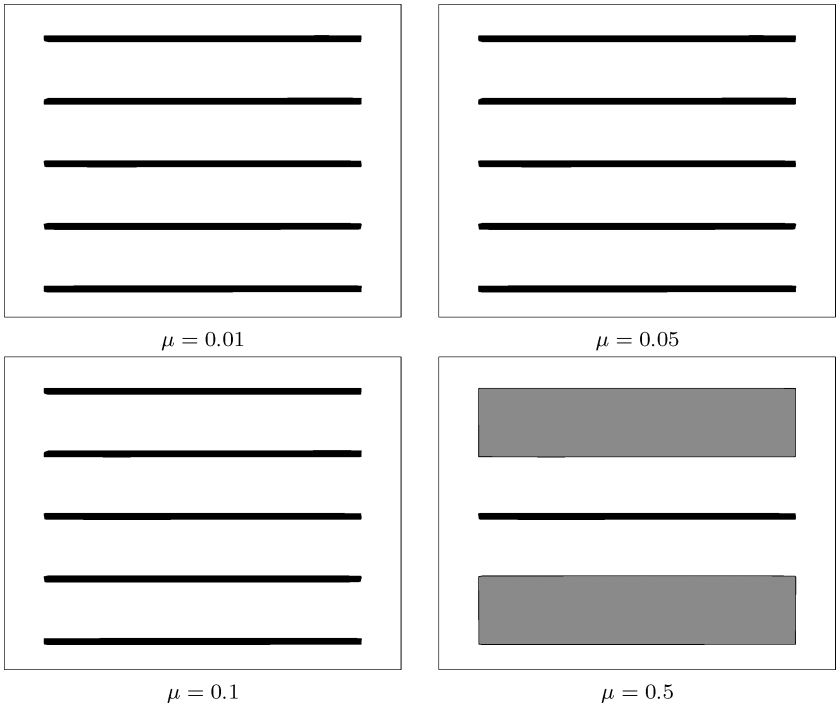


Fig. 19. For the *stripes* data set, ClusterHull recovers clusters correctly for $\mu \in [0.01, 0.1]$, but fails for $\mu \geq 0.5$.

compare the boundary of the big circle in Figure 17 for $r = 4$ and 8. However, increasing the number of directions means that we need more memory for the hulls (memory grows linearly with r). For $r = 8$, we get a good balance between memory usage and the quality of the shapes.

6.2 Variation with μ

We considered values of μ in the range $[0.01, 0.5]$. We fixed $r = 8$, and ran our algorithm for three datasets, *circles*, *ellipses*, and *stripes*. We fixed the number of hulls at $k = 10$ ($m = 128$) for *circles* and *ellipses*, and $k = 5$ ($m = 64$) for *stripes*.

If the value of μ is too small, the area dominates the cost function. This causes distant hulls to merge into long skinny hulls spanning multiple clusters. Although the period-doubling cleanup gets rid of most of them by discarding hulls with small densities, the output still contains some hulls spanning multiple natural clusters. Figure 18 shows this effect when $\mu = 0.01$.

On the other hand, if μ is increased too much, the cost function prefers decreasing the total perimeter and it is hard to prevent large neighboring clusters from merging together. In Figure 19, neighboring stripes are merged into a single hull for $\mu = 0.5$. The results show that choosing μ in the range $[0.05, 0.1]$ gives good clusters for most input sets.

7. CONCLUSION

We developed a novel framework for summarizing the geometric shape and distribution of a two-dimensional point stream. We also proposed an area-based quality measure for such a summary. Unlike existing stream-clustering methods, our scheme adaptively allocates its fixed-memory budget to represent different clusters with different degrees of detail. Such an adaptive scheme can be particularly useful when the input has widely varying cluster structures and the boundary shape, orientation, or volume of those clusters can be important clues in the analysis.

Our scheme uses a simple and natural cost function to control the cluster structure. Experiments show that this cost function performs well across widely different input distributions. The overall framework of ClusterHull is flexible and easily adapted to different applications. For instance, we show that the scheme can be enhanced with period-doubling and incremental cleanup to deal effectively with noise and extreme data distributions. In those settings, especially when the input has spatial coherence, our scheme performs noticeably better than general-purpose clustering methods like CURE and k-medians.

Because our hulls tend to be more stable than, for instance, the centroids of k-medians, we can maintain other useful data statistics, such as *population count* or *density* of individual hulls. (Our hulls grow by merging with other hulls, whereas the centroids in k-medians potentially shift after each new point arrival. The use of incremental cleanup may cause some of our hulls to be discarded, but that happens only for very low-density and, hence, less-interesting hulls.) Thus the cluster hulls can capture some important frequency statistics, such as which five hulls have the most points, or which hulls have the highest densities, etc.

Although ClusterHull is inspired by the α hull and built on top of an optimal convex hull structure, the theoretical guarantees of those structures do not extend to give approximation bounds for ClusterHull. Providing a theoretical justification for ClusterHull's practical performance is a challenge for future work.

This paper describes ClusterHull in a two-dimensional setting. The basic ideas extend to dimensions greater than two, but the data structure is limited in practice for use in low dimensions. As noted in Hershberger and Suri [2004], extending the adaptive convex hull representation to higher dimensions would require a number of samples that grows exponentially with the dimension to achieve a given accuracy. This size blowup would make ClusterHull unattractive in high-dimensional spaces.

REFERENCES

AGARWAL, P. AND PROCOPIUC, C. 2000. Approximation algorithms for projective clustering. In *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA. 538–547.

AGARWAL, P. K., HAR-PELED, S., AND VARADARAJAN, K. R. 2004. Approximating extent measures of points. *J. ACM 51,* 4, 606–635.

AGGARWAL, C. C., HAN, J., WANG, J., AND YU., P. S. 2003. A framework for clustering evolving data streams. In *Proceedings of the 29th International Conference on Very Large Databases*. Morgan Kaufmann, San Francisco, CA.

ALON, N., MATIAS, Y., AND SZEGEDY, M. 1996. The space complexity of approximating the frequency moments. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*. ACM Press, New York. 20–29.

ALURU, S. 2005. Quadtrees and octrees. In *Handbook of Data Structures and Applications*, D. P. Mehta and S. Sahni, Eds. CRC Press LLC, Boca Raton, FL. Chapter 19, 1–26.

AMENTA, N., BERN, M., AND EPPSTEIN, D. 1998. The crust and the β-skeleton: Combinatorial curve reconstruction. *Graphical Models and Image Processing 60*, 125–135.

ARYA, S., MOUNT, D. M., NETANYAHU, N. S., SILVERMAN, R., AND WU, A. 1998. An optimal algorithm for approximate nearest neighbor searching. *J. ACM 45*, 891–923.

BÖHM, C., FALOUTSOS, C., PAN, J.-Y., AND PLANT, C. 2006. Robust information-theoretic clustering. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York. 65–75.

BRADLEY, P. S., FAYYAD, U. M., AND REINA, C. 1998. Scaling clustering algorithms to large databases. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. AAAI, New York. 9–15.

CHAN, T. M. 2004. Faster core-set constructions and data stream algorithms in fixed dimensions. In *Proceedings of the 20th Annual Symposium on Computational Geometry*. ACM Press, New York. 152–159.

CHARIKAR, M., CHEKURI, C., FEDER, T., AND MOTWANI, R. 1997. Incremental clustering and dynamic information retrieval. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*. ACM Press, New York. 626–635.

CHARIKAR, M., O'CALLAGHAN, L., AND PANIGRAHY, R. 2003. Better streaming algorithms for clustering problems. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*. ACM Press, New York. 30–39.

CHEN, K. 2006. On k-median clustering in high dimensions. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA. 1177–1185.

CHENG, D., VEMPALA, S., KANNAN, R., AND WANG, G. 2005. A divide-and-merge methodology for clustering. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM Press, New York. 196–205.

CORMODE, G. AND MUTHUKRISHNAN, S. 2003. Radial histogram for spatial streams. Technical Report 2003-11, DIMACS.

CURLESS, B. AND LEVOY, M. 1996. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. ACM Press, New York. 303–312.

DEY, T. K., MEHLHORN, K., AND RAMOS, E. 2000. Curve reconstruction: Connecting dots with good reason. *Comput Geom. Theory Appl. 15*, 229–244.

DOMINGOS, P. AND HULTEN, G. 2001a. A general method for scaling up machine learning algorithms and its application to clustering. In *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA. 106–113.

DOMINGOS, P. AND HULTEN, G. 2001b. Learning from infinite data in finite time. In *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA. 673–680.

DRINEAS, P., FRIEZE, A., KANNAN, R., VEMPALA, S., AND VINAY, V. 1999. Clustering in large graphs and matrices. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA. 291–299.

EDELSBRUNNER, H. 1987. *Algorithms in Combinatorial Geometry*. EATCS Monographs on Theoretical Computer Science, vol. 10. Springer-Verlag, New York.

ESTER, M., KRIEGEL, H.-P., AND XU, X. 1995. A database interface for clustering in large spatial databases. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. AAAI, New York. 94–99.

FEIGENBAUM, J., KANNAN, S., AND ZHANG, J. 2002. Computing diameter in the streaming and sliding-window models. Manuscript.

FRAHLING, G. AND SOHLER, C. 2005. Coresets in dynamic geometric data streams. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*. ACM Press, New York. 209–217.

FRIEZE, A., KANNAN, R., AND VEMPALA, S. 1998. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Proceedings of the 39th Symposium on Foundations of Computer Science*. IEEE Computer Society, Los Alamitos, CA.

GREENWALD, M. AND KHANNA, S. 2001. Space-efficient online computation of quantile summaries. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, New York. 58–66.

GUHA, S., RASTOGI, R., AND SHIM, K. 1998. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, New York. 73–84.

GUHA, S., MEYERSON, A., MISHRA, N., MOTWANI, R., AND O'CALLAGHAN, L. 2003. Clustering data streams: Theory and practice. *IEEE Trans. Knowledge Data Engi. 15,* 3, 515–528.

HAR-PELED, S. AND VARADARAJAN, K. 2002. Projective clustering in high dimensions using core-sets. In *Proceedings of the 18th Annual Symposium on Computational Geometry*. ACM Press, New York. 312–318.

HAR-PELED, S. AND KUSHAL, A. 2005. Smaller coresets for k-median and k-means clustering. In *Proceedings of the 21st Annual Symposium on Computational Geometry*. ACM Press, New York. 126–134.

HERSHBERGER, J. AND SURI, S. 2004. Adaptive sampling for geometric problems over data streams. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM Press, New York. 252–262.

KANNAN, R., VEMPALA, S., AND VETA, A. 2000. On clusterings—good, bad and spectral. In *Proceedings of the 41st Symposium on Foundations of Computer Science*. IEEE Computer Society, Los Alamitos, CA. 367.

MANKU, G. S. AND MOTWANI, R. 2002. Approximate frequency counts over data streams. In *Proceedings of the 28th International Conference on Very Large Databases*. Morgan Kaufmann, San Francisco, CA. 346–357.

MUTHUKRISHNAN, S. 2005. *Data Streams: Algorithms and Applications*. Foundations and Trends in Theoretical Computer Science, vol. 1, issue 2. Now Publishers, Delft, Netherlands.

NG, R. T. AND HAN, J. 1994. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Databases*. Morgan Kaufmann, San Francisco, CA. 144–155.

O'ROURKE, J. AND TOUSSAINT, G. T. 1997. Pattern recognition. In *Handbook of Discrete and Computational Geometry*, J. E. Goodman and J. O'Rourke, Eds. CRC Press LLC, Boca Raton, FL. Chapter 43, 797–814.

PELLEG, D. AND MOORE, A. W. 2000. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA. 727–734.

USGS. 2003. West Nile virus maps - 2002. `http://cindi.usgs.gov/hazard/event/west_nile/`.

ZHANG, T., RAMAKRISHNAN, R., AND LIVNY, M. 1996. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, New York. 103–114.

Received July 2006; revised April 2007; accepted September 2007