INCREMENTAL CLUSTERING AND DYNAMIC INFORMATION RETRIEVAL *

M. CHARIKAR[†], C. CHEKURI[‡], T. FEDER[§], AND R. MOTWANI[¶]

Abstract. Motivated by applications such as document and image classification in information retrieval, we consider the problem of clustering *dynamic* point sets in a metric space. We propose a model called *incremental clustering* which is based on a careful analysis of the requirements of the information retrieval application, and which should also be useful in other applications. The goal is to efficiently maintain clusters of small diameter as new points are inserted. We analyze several natural greedy algorithms and demonstrate that they perform poorly. We propose new deterministic and randomized incremental clustering algorithms which have a provably good performance, and which we believe should also perform well in practice. We complement our positive results with lower bounds on the performance of incremental algorithms. Finally, we consider the dual clustering problem where the clusters are of fixed diameter, and the goal is to minimize the number of clusters.

Key words. incremental clustering, dynamic information retrieval, minimum diameter clustering, agglomerative clustering, k-center, performance guarantee.

AMS subject classifications. 68Q25, 68W40

1. Introduction.

We consider the following problem: as a sequence of points from a metric space is presented, efficiently maintain a clustering of the points so as to minimize the maximum cluster diameter. Such problems arise in a variety of applications, in particular in document and image classification for information retrieval. We propose a model called *incremental clustering* based primarily on the requirements for the information retrieval application, although our model should also be relevant to other applications. We begin by analyzing several natural greedy algorithms and discover that they perform rather poorly in this setting. We then identify some new deterministic and randomized algorithms with provably good performance. We complement our positive results with lower bounds on the performance of incremental algorithms. We also consider the dual clustering problem where the clusters are of fixed diameter, and the goal is to minimize the number of clusters. Before describing our results in any greater detail, we motivate and formalize our new model.

Clustering is used for data analysis and classification in a wide variety of applications [1, 20, 29, 36, 44]. It has proved to be a particularly important tool in *information retrieval* for constructing a taxonomy of a corpus of documents by forming groups of closely-related documents [21, 24, 44, 45, 47, 48]. For this purpose, a distance metric is imposed over documents, enabling us to view them as points in a

*A preliminary version of this paper appeared in the *Proc. of the 29th ACM STOC*, 1997.

[†]Department of Computer Science, Princeton University, Princeton, NJ 08544. E-mail: `moses@cs.princeton.edu`. This work was mostly done while the author was at Stanford University and was supported by Rajeev Motwani's NSF Award CCR-9357849. Currently supported by NSF ITR CCR-0205594, DOE Award DE-FG02-02ER25540, NSF CAREER Award CCR-0237113 and an Alfred P. Sloan Fellowship.

[‡]Bell Labs, 600 Mountain Avenue, Murray Hill, NJ 07974. E-mail: `chekuri@research.bell-labs.com`. This work was mostly done while the author was at Stanford University and was supported by Rajeev Motwani's NSF Award CCR-9357849.

[§]E-mail: `tomas@theory.stanford.edu`

[¶]Department of Computer Science, Stanford University, Stanford, CA 94305-9045. E-mail: `rajeev@cs.stanford.edu`. Supported by an Alfred P. Sloan Research Fellowship, an IBM Faculty Partnership Award, an ARO MURI Grant DAAH04-96-1-0007, and NSF Young Investigator Award CCR-9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

metric space. The central role of clustering in this application is captured by the so-called **cluster hypothesis**: *documents relevant to a query tend to be more similar to each other than to irrelevant documents and hence are likely to be clustered together.* Typically, clustering is used to accelerate query processing by considering only a small number of representatives of the clusters, rather than the entire corpus. In addition, it is used for classification [19] and has been suggested as a method for facilitating browsing [16, 17].

The current information explosion, fueled by the availability of hypermedia and the World-wide Web, has led to the generation of an ever-increasing volume of data, posing a growing challenge for information retrieval systems to efficiently store and retrieve this information [50]. A major issue that document databases are now facing is the extremely high rate of update. Several practitioners have complained that existing clustering algorithms are not suitable for maintaining clusters in such a dynamic environment, and they have been struggling with the problem of updating clusters without frequently performing complete re-clustering [7, 8, 9, 15, 45]. From a theoretical perspective, many different formulations are possible for this dynamic clustering problem, and it is not clear a priori which of these best addresses the concerns of the practitioners. After a careful study of the requirements, we propose the model described below.

Hierarchical Agglomerative Clustering. The clustering strategy employed almost universally in information retrieval is *Hierarchical Agglomerative Clustering (HAC)*, see [20, 44, 45, 47, 48, 49]. This is also popular in other applications such as biology, medicine, image processing, and geographical information systems. The basic idea is: initially assign the n input points to n distinct clusters; repeatedly merge pairs of clusters until their number is sufficiently small. Many instantiations have been proposed and implemented, differing mainly in the rule for deciding which clusters to merge at each step. Note that HAC computes hierarchy trees of clusters (also called *dendograms*) whose leaves are individual points and internal nodes correspond to clusters formed by merging clusters at their children. A key advantage of these trees is that they permit refinement of responses to queries by moving down the hierarchy. Typically, the internal nodes are labeled with indexing information (sometimes called *conceptual information*) used for processing queries and in associating semantics with clusters (e.g., for browsing). Experience shows that HAC performs extremely well both in terms of efficiency and cluster quality. In the dynamic setting, it is desirable to retain the hierarchical structure while ensuring efficient update and high-quality clustering. An important goal is to avoid any major modifications in the clustering while processing updates, since any extensive re-computation of the index information will swamp the cost of clustering itself. The input size in typical applications is such that super-quadratic time is impractical, and in fact it is desirable to obtain close to linear time.

A Model for Incremental Clustering. Various measures of distance between documents have been proposed in the literature, but we will not concern ourselves with the details thereof; for our purposes, it suffices to note that these distance measures induce a metric space. Since documents are usually represented as high-dimensional vectors, we cannot make any stronger assumption than that of an arbitrary metric space, although, as we will see, our results improve significantly in geometric spaces.

Formally, the *clustering* problem is: given n points in a metric space $\mathcal{M}$, partition the points into k clusters so as to minimize the maximum cluster diameter. The *diameter* of a cluster is defined to be the maximum inter-point distance in it.

Sometimes the objective function is chosen to be the maximum cluster radius. In Euclidean spaces, *radius* denotes the radius of the minimum ball enclosing all points in the cluster. To extend the notion of radius to arbitrary metric spaces, we first select a *center* point in each cluster, whereupon the radius is defined as the maximum distance from the center to any point in the cluster. We will assume the diameter measure as the default. We mention that there are several other measures of cluster quality that have ben considered in the literature (e.g. sum of squares of distances to cluster centers, etc). In this paper, we shall only consider the radius and diameter measures.

We define the *incremental clustering* problem as follows: for an update sequence of n points in $\mathcal{M}$, maintain a collection of k clusters such that as each input point is presented, either it is assigned to one of the current k clusters, or it starts off a new cluster while two existing clusters are merged into one. We define the *performance ratio* of an incremental clustering algorithm as the maximum over all update sequences of the ratio of its maximum cluster diameter (or, radius) to that of the optimal clustering for the input points.

Our model enforces the requirement that *at all times* an incremental algorithm should maintain a HAC for the points presented up to that time. As before, an algorithm is free to use any rule for choosing the two clusters to merge at each step. This model preserves all the desirable properties of HAC while providing a clean extension to the dynamic case. In addition, it has been observed that such incremental algorithms exhibit good paging performance when the clusters themselves are stored in secondary storage, while cluster representatives are preserved in main memory [42].

We have avoided labeling this model as *the online clustering problem* or referring to the performance ratio as a *competitive ratio* [34] for the following reasons. Recall that in an online setting, we would compare the performance of an algorithm to that of an adversary which knows the update sequence in advance but must process the points in the order of arrival. Our model has a stronger requirement for incremental algorithms, in that they are compared to adversaries which do not need to respect the input ordering, i.e., we compare our algorithms to optimal clusterings of the final point set, and no intermediate clusterings need be maintained. Also, online algorithms are permitted super-polynomial running times. In contrast, our model essentially requires polynomial-time approximation algorithms which are constrained to incrementally maintain a HAC. It may be interesting to explore the several different formulations of online clustering; for example, when the newly inserted point starts off a new cluster, we could allow the points of one old cluster to be redistributed among the remaining, rather than requiring that two clusters be merged together. The problem with such formulations is that they do not lead to HACs; moreover, they entail the re-computation of the index structures for *all clusters*, which renders the algorithms useless from the point of view of real applications. We note that since the incremental clustering model is more restrictive than the online model, our algorithms can also be viewed as online clustering algorithms; the competitive ratios are the same as the performance ratios that we prove here.

Previous Work in Static Clustering. The closely-related problems of clustering to minimize diameter and radius are also called *pairwise clustering* and the *k-center problem*, respectively [4, 30]. Both are NP-hard [25, 38], and in fact hard to approximate to within factor 2 for arbitrary metric spaces [4, 30]. For Euclidean spaces, clustering on the line is easy [6], but in higher dimensions it is NP-hard to approximate to within factors close to 2, regardless of the metric used [22, 23, 27, 39, 40]. The *furthest* *point heuristic* due to Gonzalez [27] (see also Hochbaum and Shmoys [32, 33]) gives a 2-approximation in all metric spaces. This algorithm requires $O(kn)$ distance computations, and when the metric space is induced by shortest-path distances in weighted graphs, the running time is $O(n^2)$. Feder and Greene [22] gave an implementation for *Euclidean* spaces with running time $O(n \log k)$.

Clustering problems have been extensively studied in different communities with a variety of different perspectives. The optimization viewpoint of clustering formulates the problem as that of finding a solution that optimizes a certain objective function that measures cluster quality. In addition to the minimum diameter and radius measures described above, several other objective functions have also been considered in the literature. A lot of recent work has focused on the k-median objective [11, 37, 2]. Here the goal is to assign points to k centers such that the sum of distances of points to their centers is minimized. Other objectives that have been studied include the objective of minimizing the sum of all distances within each cluster [3, 18] and that of minimizing the sum of cluster diameters [14]. In addition to this, outlier formulations of clustering problems have also been studied [12]. Here the algorithm is allowed to discard a fraction of the input as outliers and is required to obtain a clustering solution that minimizes a given objective function on the remaining input points.

Overview of Results. Our results for incremental clustering show that it is possible to obtain algorithms that are comparable to the best possible in the static setting, both in terms of efficiency and performance ratio. We begin in Section 2 by considering natural greedy algorithms that choose clusters to merge based on some measure of the resulting cluster. We establish that greedy algorithms behave poorly by proving that a Center-Greedy algorithm has a tight performance ratio of $2k-1$, and a Diameter-Greedy algorithm has a lower bound of $\Omega(\log k)$. It seems likely that greedy algorithms behave better in geometric spaces, and we discover some evidence in the case of the line. We show that Diameter-Greedy has performance ratio 2 for $k = 2$ on the line. This analysis suggests a variant of Diameter-Greedy, and this is shown to achieve ratio 3 for all k on the line. In Section 3 we present the Doubling Algorithm and show that its performance ratio is 8, and that a randomized version has ratio 5.43. While the obvious implementation of these algorithms is expensive, we show that they can be implemented so as to achieve amortized time $O(k \log k)$ per update. These results for the Doubling Algorithm carry over to the radius measure. We also give a variant of this algorithm that is oblivious to the number of clusters k. We maintain a hierarchy from which, for any given k, at most k clusters can be obtained such that the diameter of the clusters is at most a factor of 8 times the optimal diameters for that value of k. Then, in Section 4, we present the Clique Algorithm and show that it has performance ratio 6, and that a randomized version has ratio 4.33. While the Clique Algorithm may appear to dominate the Doubling Algorithm, this is not the case since the former requires computing clique partitions, an NP-hard problem, although it must be said in its defense that the clique partitions need only be computed in graphs with $k+1$ vertices. While the performance ratio of the Clique Algorithm is 8 for the radius measure, improved bounds are possible for d-dimensional Euclidean spaces; specifically, we show that the radius performance ratio of the Clique Algorithm in $\Re^d$ improves to $4(1 + \sqrt{d/(2d+2)})$, which is 6 for $d = 1$, and is asymptotic to 6.83 for large d. In Section 5, we provide lower bounds for incremental clustering algorithms. We show that even for $k = 2$ and on the line, no deterministic or randomized algorithm can achieve a ratio better than 2. We improve this lower bound to 2.414 for deterministic algorithms in general metric spaces. Finally, in Section 6 we consider the dual clustering problem of minimizing the number of clusters of a fixed radius. Since it is impossible to achieve bounded ratios for general metric spaces, we focus on d-dimensional Euclidean spaces. We present an incremental algorithm that has performance ratio $O(2^d d \log d)$, and also provide a lower bound of $\Omega(\log d/\log\log\log d)$.

Measure	Algorithm	Upper Bound	Lower Bound
Diameter			$1+\sqrt{2}$
	Center-Greedy	$2k-1$	
	Diameter-Greedy		$\Omega(\log k)$
	Doubling	8	
	Clique Partition	6	
			$(2-\epsilon)^*$
	Doubling	5.437^*	
	Clique Partition	4.33^*	
Radius			3
	Doubling	8	
	Clique Partition	8	
			$(3-\epsilon)^*$
	Doubling	5.437^*	
	Clique Partition	5.77^*	
Dual Clust			$\Omega(\frac{\log d}{\log\log\log d})$ in $\Re^d$
		$O(2^d d\log d)$ in $\Re^d$	$\Omega(2^d d\log d)$

FIG. 1.1. *Summary of results: a star indicates randomization. All algorithms have lower bounds on their performance ratios that match the upper bounds we establish.*

Many interesting directions for future research are suggested by our work. There are the obvious questions of improving our upper and lower bounds, particularly for the dual clustering problem. An important theoretical question is whether the geometric setting permits better ratios than do metric spaces. Our model can be generalized in many different ways. Depending on the exact application, we may wish to consider other measures of clustering quality, such as: minimum variance in cluster diameter, and the sum of squares of the inter-point distances within a cluster. Then, there is the issue of handling deletions which, though not important for our motivating application of information retrieval, may be relevant in other applications. Finally, there is the question of formulating a model for *adaptive clustering*, wherein the clustering may be modified as a result of queries and user feedback, even without any updates.

Recently, there has been considerable attention devoted to the streaming data model which was formalized and extensively studied after the appearance of the conference version of this paper. The streaming model is motivated by the goal of designing highly efficient algorithms for very large data sets. In this model, an algorithm is required to perform its computation in one pass or a few passes over the data while using very little memory. Our algorithms can be viewed as streaming algorithms for clustering problems where the goal is to minimize the maximum diameter or radius of the clusters produced. Subsequent to this work, other clustering objectives have also been studied in the streaming model, most notably the k-median objective [28, 13] and the sum of cluster diameters objective [14].

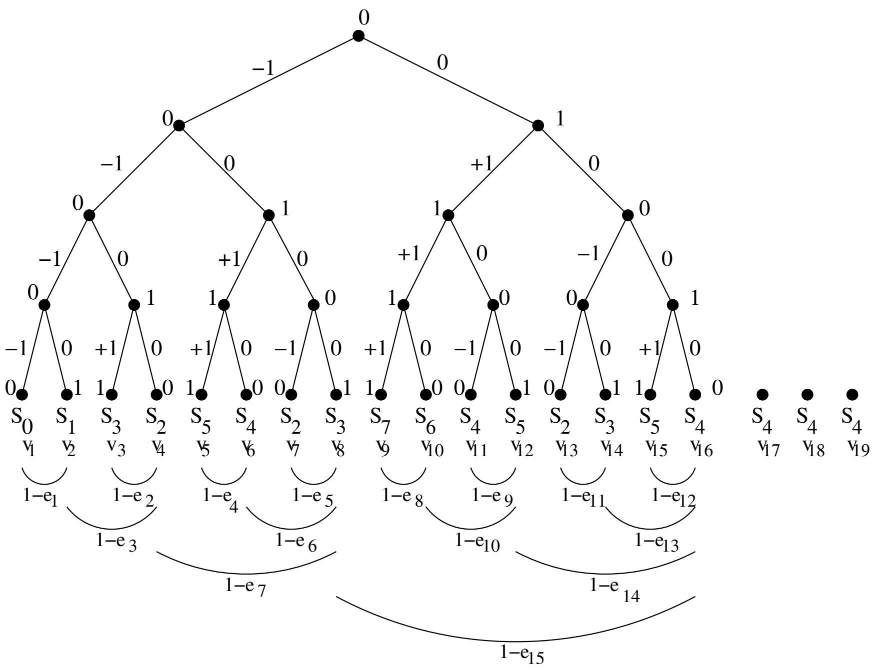


FIG. 2.1. *Illustration of the lower bound construction for Center-Greedy for $k = 4$. The instance consists of $2^k + k - 1 = 19$ vertices labeled $v_1, \ldots v_{19}$. Each vertex belongs to one of the sets $S_0, \ldots, S_7$ as indicated. Vertex and edge labels defined on the tree structure are illustrated. The curved edges at the bottom of the figure indicate the additional edges added. (Note that distances within each S_i are all 1 and are not shown in the figure). The curved edges also indicate the order in which clusters are merged by the Center-Greedy algorithm: cluster centers at the end points of the edge labeled $1 - \epsilon_1$ are first merged followed by centers at the end points of the edge labeled $1 - \epsilon_2$, then $1 - \epsilon_3$ and so on.*

2. Greedy Algorithms. We begin by examining some natural greedy algorithms. A *greedy incremental clustering algorithm* always merges clusters to minimize some fixed measure. Our results indicate that such algorithms perform poorly.

DEFINITION 2.1. *The Center-Greedy Algorithm associates a center for each cluster and merges the two clusters whose centers are closest. The center of the old cluster with the larger radius becomes the new center.*

It is possible to define variants of Center-Greedy based on how the centers of the clusters are picked but we restrict ourselves to the above definition for reasons of simplicity and intuitiveness.

DEFINITION 2.2. *The Diameter-Greedy Algorithm always merges those two clusters which minimize the diameter of the resulting merged cluster.*

We can establish the following lower bounds on the performance ratio of these two greedy algorithms.

THEOREM 2.3. *The Center-Greedy Algorithm's performance ratio has a lower bound of $2k - 1$.*

Proof. We first construct a graph $G = (V, E)$ which defines the metric space on which Center-Greedy has a performance ratio of $2k - 1$. The vertex set of G is the union of $2k$ disjoint sets $S_0, S_1, \ldots, S_{2k-1}$ with a total of $2^k + k - 1$ vertices. The request sequence consists of the entire set of vertices $v_1, v_2, \ldots, v_{2^k+k-1}$. We use a complete binary tree $T = (V_T, E_T)$ whose leaves in a left to right order are the vertices

$v_1, \ldots, v_{2^k}$ to describe our construction. Each internal node x of T has a binary *label* associated with it, denoted by $label(x)$. Let x_l and x_r be the left and right children of node x. We recursively define the labels of nodes of T as follows. The root of the tree is labeled with 0. If $label(x) = i$, we set $label(x_l) = i$ and $label(x_r) = 1-i$. With each edge (x, y) of the tree we associate an integer weight of value $label(x) + label(y) - 1$. Note that the weights belong to $\{-1, 0, 1\}$ by our labeling procedure. We now specify the sets $S_0, S_1, \ldots, S_{2k-1}$. A leaf v_i belongs to the set S_j if and only if the sum of the weights of the edges on the path from v_i to the root of T is equal to $j - k$. The vertices $v_{2^k+1}, \ldots, v_{2^k+k-1}$ belong to S_k. With each node x of T we also associate a vertex $v(x)$ of G as follows. If x is a leaf, we set $v(x) = x$; for an internal node x, we set $v(x) = v(x_r)$. For a node x of T let $po(x) = i$ if x is the ith vertex to be visited in a *post order* traversal of T, where we do not include the leaves in the traversal.

Now we specify the distances among the vertices of G. We first specify the edges present in G and then associate lengths with them. The other distances are induced by these edges. The edge set E is defined by the sets $S_0, \ldots, S_{2k-1}$ as follows:

$$\{(v_i, v_j) \mid v_i, v_j \in S_r, 0 \le r \le 2k-1\} \bigcup \{(v_i, v_j) \mid v_i \in S_r, v_j \in S_{r+1}, 0 \le r < 2k-1\}.$$

Edges between two vertices in the same set S_i are called *clique* edges. All the clique edges have length 1. The lengths of the other edges are defined as follows. We do a post order traversal of T and, as we process each internal node x of T, we set $d(v(x), v(x_l)) = 1 - \epsilon_{po(x)}$ such that $1 \gg \epsilon_1 > \epsilon_2 > \ldots > \epsilon_{2^k-1} > 0$. By our placement of vertices in the sets S_i, we are guaranteed that $(v(x), v(x_l)) \in E$ for all x in T. We set to 1 any edge-length of an edge in E which is not already determined by the above process. See Figure 2.1 for an illustration of the construction.

Let C_x denote the cluster with $v(x)$ as the center and the leaves in the subtree rooted at x as the elements. Let $po^{-1}(i)$ denote the vertex x where $po(x) = i$. For $1 \le i \le 2^k - 1$, let A_i be the set of clusters C_y such that y is the left child of some node on the path from $po^{-1}(i)$ to the root of T, but does not itself lie on the path. We also include C_x in A_i, where $x = po^{-1}(i)$.

The lower bound is based on the following two claims.

CLAIM 2.4. *For $1 \le i \le 2k-1$, A_i is the set of clusters of Center-Greedy which contain more than one vertex after the $k+i$ vertices $v_1, \ldots, v_{k+i}$ are given.*

Observe that $v_1 \in S_0$ and $v_{2^{k-1}+1} \in S_{2k-1}$, and the distance between them is at least $(2k-1)(1-\epsilon_1)$. From the above claim it follows that at the end of the request sequence, all the vertices $v_1, \ldots, v_{2^k}$ are in one cluster. Therefore, the diameter of this cluster is at least $(2k-1)(1-\epsilon_1)$.

CLAIM 2.5. *There is a k-clustering of G of diameter* 1.

The clustering which achieves the above diameter is $\{S_0 \cup S_1, \ldots, S_{2k-2} \cup S_{2k-1}\}$. This finishes the proof of the theorem. □

THEOREM 2.6. *The Diameter-Greedy Algorithm's performance ratio is $\Omega(\log k)$, even on the line.*

Proof. We will prove a lower bound of $\Omega(\log k)$ for Diameter-Greedy on the real line. Let F_j be the jth Fibonacci number. Let k and r be such that $k = 2\sum_{i=1}^{r} F_i$. We will prove a lower bound of $r+1$ on Diameter-Greedy. To this end, we define sets of points A_{ij}, for $1 \le i \le r$ and $1 \le j \le F_j$. Fix ϵ and δ such that $\epsilon < \delta \ll 1$. The set A_{ij} consists of the points p_{ij}, q_{ij}, r_{ij}, and s_{ij} sorted by their x-coordinates; the distances between them are as follows: $d(p_{ij}, q_{ij}) = 1 + \epsilon$, $d(q_{ij}, r_{ij}) = i$, and $d(r_{ij}, s_{ij}) = 1 + \delta$. Further, the distance between two distinct A_{ij}'s is set to ∞ (although any sufficiently large value will do). For $i = 1 \ldots r$, let U_i, V_i, W_i, X_i, Y_i, and Z_i denote the following clusters:

$$
\begin{aligned}
U_i &= \cup_{j=1}^{F_i}\{\{p_{ij}, q_{ij}\}, \{r_{ij}, s_{ij}\}\},\\
V_i &= \cup_{j=1}^{F_i}\{\{q_{ij}\}, \{r_{ij}\}\},\\
W_i &= \cup_{j=1}^{F_i}\{\{p_{ij}\}, \{q_{ij}, r_{ij}\}\},\\
X_i &= \cup_{j=1}^{F_i}\{\{p_{ij}\}, \{q_{ij}, r_{ij}\}, \{s_{ij}\}\},\\
Y_i &= \cup_{j=1}^{F_i}\{\{p_{ij}, q_{ij}, r_{ij}\}, \{s_{ij}\}\},\\
Z_i &= \cup_{j=1}^{F_i}\{\{p_{ij}, q_{ij}, r_{ij}, s_{ij}\}\}.
\end{aligned}
$$

Initially, the points given to Diameter-Greedy are the points q_{ij} and r_{ij} for each of the sets A_{ij}. Let P_i denote the sequence of requests $p_{i1}, \ldots, p_{iF_i}$ and S_i denote the sequence $s_{i1}, \ldots, s_{iF_i}$. Define $K_i = P_1S_1P_2S_2\ldots P_iS_i$. The request sequence is K_{r-1}. Note that $K_t = K_{t-1}P_tS_t$. We show inductively that the following invariant is maintained.

When the last element of K_t is received, Diameter-Greedy's $k+1$ clusters are

$$(\cup_{i=1}^{t-2}Z_i)\bigcup Y_{t-1}\bigcup X_t\bigcup(\cup_{i=t+1}^{r}V_i).$$

Since there are $k+1$ clusters, two of the clusters have to be merged and the algorithm merges two clusters in V_{t+1} to form a cluster of diameter $(t+1)$. Without loss of generality, we may assume that the clusters merged are $\{q_{(t+1)1}\}$ and $\{r_{(t+1)1}\}$.

Suppose that the situation is as described above at the end of the sequence K_t. We will show that the the invariant holds at the end of the sequence K_{t+1}. Note that $K_{t+1} = K_tP_{t+1}S_{t+1}$. Observe that merging any two clusters in Y_{t-1} will increase their diameter to $(t+1+\epsilon+\delta)$. Merging any two clusters in X_t will increase their diameter to $(t+1+\epsilon)$. Because we start with the situation where there is a cluster $\{q_{(t+1)1}, r_{(t+1)1}\}$, the request $p_{(t+1)1}$ forces the formation of the cluster $\{q_{(t+1)2}, r_{(t+1)2}\}$, since any other merging results in a diameter of more than $t+1$. Thus, at the end of the sequence of requests P_{t+1}, we have $k+1$ clusters $(\cup_{i=1}^{t-2}Z_i)\bigcup Y_{t-1}\bigcup X_t\bigcup X_{t+1}\bigcup(\cup_{i=t+2}^{r}V_i)$.

Since there are $k+1$ clusters, Diameter-Greedy merges two clusters in X_t to form a cluster of diameter $(t+1+\epsilon)$. Thus, as we give points in S_{t+1}, it will form the clusters Y_t. Since $F_{t+1} > F_t$, after F_t requests in the sequence S_{t+1} all the clusters in X_t would have merged and Diameter-Greedy starts merging clusters in Y_{t-1} to form clusters Z_{t-1} of diameter $(t+1+\epsilon+\delta)$. Since $F_{t+1} = F_t + F_{t-1}$, at the end of the request sequence S_{t+1} it will have the $k+1$ clusters $(\cup_{i=1}^{t-1}Z_i)\bigcup Y_t\bigcup X_{t+1}\bigcup(\cup_{i=t+2}^{r}V_i)$. This shows that the invariant is maintained at the end of the request sequence K_{t+1}.

It is easy to verify that the invariant is true for the base case of $t = 3$. Thus, at the end of the request sequence K_{r-1}, the maximum diameter of the clusters of Diameter-Greedy is r. Since $k = 2\sum_{i=1}^{r} F_i = 2F_{r+2} - 2$, it follows that $r = \Omega(\log k)$. To finish the proof we observe that the optimal clusters are $\cup_{i=1}^{r-1}U_i\bigcup V_r$, and the diameter of these clusters is at most $1+\delta$. □

We now give a tight upper bound for the Center-Greedy Algorithm. Note that for $k = 3$ it has ratio 5, but for larger k its performance is worse than that of the algorithms to be presented later.

THEOREM 2.7. *The Center-Greedy Algorithm has performance ratio of $2k-1$ in any metric space.*

Proof. Suppose that a set P of n points is inserted. Let their optimal clustering be the partition $S = \{C_1, \ldots, C_k\}$, with d as the optimal diameter. We will show that the diameter of any cluster produced by Center-Greedy is at most $(2k-1)d$.

We define a graph G on the set S of the optimal clusters, where two clusters are connected by an edge if the minimum distance between them is at most d, where the distance between two clusters is the minimum distances between points in them. Consider the connected components of G. Note that two clusters in different connected components have minimum distance strictly greater than d. We say that a cluster X intersects a connected component consisting of the optimal clusters $C_{i_1}, \ldots, C_{i_r}$ if X intersects $\cup_{j=1}^{r} C_{i_j}$.

We claim that at all times, any cluster produced by Center-Greedy intersects exactly one connected component of G. We prove this claim by induction over n. Suppose the claim is true before a new point p arrives. Initially, p is in a cluster of its own and Center-Greedy has $k+1$ clusters, each of which intersect exactly one connected component of G. Since there are $k+1$ cluster centers, two of them must be in the same optimal cluster. This implies that the distance between the two closest centers is at most d. If X_1 and X_2 are the clusters that Center-Greedy merges at this stage, the centers of X_1 and X_2 must be at most d apart. Hence, both clusters' centers must lie in the same connected component of G, say $\mathcal{C}$. By the inductive hypothesis, all points in X_1 and X_2 must be in $\mathcal{C}$. Hence, all points in the new cluster $X_1 \cup X_2$ must lie in $\mathcal{C}$, establishing the inductive hypothesis.

Since each cluster produced by Center-Greedy lies in exactly one connected component of G, the diameter is bounded by the maximum diameter of a connected component, which is at most $(2k-1)d$. □

For Diameter-Greedy in general metric spaces, we can only prove the following weak upper bound.

THEOREM 2.8. *For $k = 2$, the Diameter-Greedy Algorithm has a performance ratio* 3 *in any metric space.*

Proof. Let d be the diameter in the optimal clustering. If the minimum distance between the two clusters (in the optimal clustering) is greater than d, then Diameter-Greedy keeps the two clusters separate, and achieves the optimum. If the minimum distance between the two clusters is at most d, then all the points together form a cluster of diameter at most $3d$. □

In spite of the lower bounds for greedy algorithms, they may not be entirely useless since some variant may perform well in geometric spaces. We obtain some positive evidence in this regard via the following analysis for the line. The upper bounds given here should be contrasted with the lower bound of 2 for the line shown in Section 5. The following definitions underlie the analysis.

DEFINITION 2.9. *Given a set S of n points in the line, a* t-partition *subdivides the interval between the first and last points of S into t subintervals whose endpoints are in S. The* t-diameter *of S is the minimum over all t-partitions of the maximum interval length in a t-partition of S. The* 1*-diameter is the diameter, while the* 2*-diameter is the radius of S where the center is constrained to be a point of S.*

We define the following family of algorithms based on the notion of the t-diameter.

DEFINITION 2.10. *The t-Diameter Greedy Algorithm merges those two clusters which minimize the t-diameter of the merged cluster. Note that* 1*-Diameter Greedy is the same as Diameter-Greedy.*

A few preliminary results on the t-Diameter Greedy Algorithm can be found in the Appendix. We do not have good lower or upper bounds on the performance of this parameterized class of algorithms for general metric spaces.

3. The Doubling Algorithm. We now describe the Doubling Algorithm which has performance ratio 8 for incremental clustering in general metric spaces. The algorithm works in phases and uses two parameters α and β to be specified later, such that $\alpha/(\alpha-1) \leq \beta$. At the start of phase i, it has a collection of $k+1$ clusters $C_1, C_2, \ldots, C_{k+1}$ and a lower bound d_i on the optimal clustering's diameter (denoted by OPT). Each cluster C_i has a center c_i which is one of the points of the cluster. The following invariants are assumed at the start of phase i:

1. for each cluster C_j, the radius of C_j defined as $\max_{p \in C_j} d(c_j, p)$ is at most αd_i;
2. for each pair of clusters C_j and C_l, the inter-center distance $d(c_j, c_l) \geq d_i$;
3. and, $d_i \leq$ OPT.

Each phase consists of two stages: the first is a *merging stage* in which the algorithm reduces the number of clusters by merging certain pairs; the second is the *update stage* in which the algorithm accepts new updates and tries to maintain at most k clusters without increasing the radius of the clusters or violating the invariants (clearly, it can always do so by making the new points into separate clusters). A phase ends when the number of clusters again exceeds k.

DEFINITION 3.1. *The* t-threshold graph *on a set of points* $P = \{p_1, p_2, \ldots, p_n\}$ *is the graph* $G = (P, E)$ *such that* $(p_i, p_j) \in E$ *if and only if* $d(p_i, p_j) \leq t$.

The merging stage works as follows. Define $d_{i+1} = \beta d_i$, and let G be the d_{i+1}-threshold graph on the $k+1$ cluster centers $c_1, c_2, \ldots, c_{k+1}$. The graph G is used to merge clusters by repeatedly performing the following steps while the graph is non-empty: pick an arbitrary cluster C_i in G and merge all its neighbors into it; make c_i the new cluster's center; and, remove C_i and its neighbors from G. Let $C'_1, C'_2, \ldots, C'_m$ be the new clusters at the end of the merging stage. Note that it is possible that $m = k+1$ when the graph G has no edges, in which case the algorithm will be forced to declare the end of phase i without going through the update stage.

LEMMA 3.2. *The pairwise distance between cluster centers after the merging stage of phase* i *is at least* d_{i+1}.

LEMMA 3.3. *The radius of the clusters after the merging stage of phase* i *is at most* $d_{i+1} + \alpha d_i \leq \alpha d_{i+1}$.

Proof. Prior to the merging, the distance between two clusters which are adjacent in the threshold graph is at most d_{i+1}, and their radius is at most αd_i. Therefore, the radius of the merged clusters is at most

$$d_{i+1} + \alpha d_i \leq (1 + \alpha/\beta) d_{i+1} \leq \alpha d_{i+1},$$

where the last inequality follows from the choice that $\alpha/(\alpha-1) \leq \beta$. □

The update stage continues while the number of clusters is at most k. When a new point arrives, the algorithm attempts to place it in one of the current clusters without exceeding the radius bound αd_{i+1}: otherwise, a new cluster is formed with the update as the cluster center. When the number of clusters reaches $k+1$, phase i ends and the current set of $k+1$ clusters along with d_{i+1} are used as the input for the $(i+1)$st phase.

All that remains to be specified about the algorithm is the initialization. The algorithm waits until $k+1$ points have arrived and then enters phase 1 with each point as the center of a cluster containing just itself, and with d_1 set to the distance between the closest pair of points. It is easily verified that the invariants hold at the start of phase 1. The following lemma shows that the clusters at the end of the ith phase satisfy the invariants for the $(i+1)$st phase.

LEMMA 3.4. *The* $k+1$ *clusters at the end of the* i*th phase satisfy the following conditions:*

1. *The radius of the clusters is at most* αd_{i+1}.
2. *The pairwise distance between the cluster centers is at least* d_{i+1}.
3. $d_{i+1} \leq \text{OPT}$, *where* OPT *is the diameter of the optimal clustering for the current set of points.*

Proof. We have $k+1$ clusters at the end of the phase since that is the terminating condition for a phase. From Lemma 3.3, the radius of the clusters after the merging stage is at most αd_{i+1}. When adding new points, the algorithm ensures that the radius bound is respected. From Lemma 3.2, the distance between the clusters centers after the merging stage is d_{i+1}, and a new cluster is created only if a request point is at least d_{i+1} away from all current cluster centers. Therefore, the cluster centers have pairwise distance at least d_{i+1}. Since at the end of the phase we have $k+1$ cluster centers that are d_{i+1} apart, the optimal clustering if forced to put at least two of them in the same cluster. It follows that $\text{OPT} \geq d_{i+1}$. □

We make the following observations. The algorithm ensures the invariant that $d_i \leq \text{OPT}$ at the start of phase i. The radius of the clusters during phase i is at most αd_{i+1}. Therefore, the performance ratio at any time during the phase is at most $2\alpha d_{i+1}/\text{OPT} \leq 2\alpha\beta d_i/\text{OPT} \leq 2\alpha\beta$. We choose $\alpha = \beta = 2$; note that this choice satisfies the condition that $\alpha/(\alpha - 1) \leq \beta$. This leads to the following performance bound, which can be shown to be tight.

THEOREM 3.5. *The Doubling Algorithm has performance ratio* 8 *in any metric space.*

We give a simple example showing that our analysis of the Doubling Algorithm is tight. Here we assume that $k \geq 3$. The input sequence consists of $k+3$ points, $p_1, \ldots, p_{k+1}, p_{k+2}, p_{k+3}$ given in that order. The points $p_1, \ldots, p_{k+1}$ form a uniform metric space with distance 1. The points p_{k+2} and p_{k+3} are distance 4 from the points $p_1, \ldots, p_{k+1}$ and are at distance 8 from each other. It is easy to ensure that both p_{k+2} and p_{k+3} will be added to the same cluster during the update stage of phase 1, thus the diameter of the largest cluster formed by the Doubling Algorithm is 8. However an optimum clustering can achieve diameter 1 by having p_{k+2} and p_{k+3} in their own clusters and the rest of the points in a single cluster of diameter 1.

An examination of the proof of the preceding theorem shows that the radius of the clusters is within factor 4 of the diameter of the optimal clustering, leading to the following result.

COROLLARY 3.6. *The Doubling Algorithm has performance ratio* 8 *for the radius measure.*

A simple modification of the Doubling Algorithm, in which we pick the new cluster centers by a simple left-to-right scan, improves the ratio to 6 for the case of the line.

While the obvious implementation of this algorithm appears to be inefficient, we can establish the following time bound, which is close to the best possible.

THEOREM 3.7. *The Doubling Algorithm can be implemented to run in* $O(k \log k)$ *amortized time per update.*

Proof. First of all, we assume that there is a black-box for computing the distance between two points in the metric space in unit time. This is a reasonable assumption in most applications, and in any case even the static algorithms' analysis requires such an assumption. In the information retrieval application, the documents are represented as vectors and the black-box implementation will depend on the vector length as well as the exact definition of distance.

We now show how the Doubling Algorithm may be implemented so that the amortized time required for processing each new update is bounded by $O(k \log k)$.

We maintain the edge lengths of the complete graph induced by the current cluster centers in a heap. Since there at most k clusters the space requirement is $O(k^2)$. When a new point arrives, we compute the distance of this point to the each of the current cluster centers, which requires $O(k)$ time. If the point is added to one of the current clusters, we are done. If, on the other hand, the new point initiates a new cluster, we insert into the heap edges labeled with the distances between this new center and the existing cluster centers. For accounting purposes in the amortized analysis, we associate $\log k$ credits with each inserted edge. We will show that it is possible to charge the cost of implementing the merging stage of the algorithm to the credits associated with the edges. This implies the desired time bound.

We can assume, without loss of generality, that the merging stage merges at least two clusters. Let t be the threshold used during the phase. The algorithm extracts all the edges from the heap which have length less than t. Let m be the number of edges deleted from the heap. The deletion from the heap costs $O(m \log k)$ time. The t-threshold graph on the cluster centers is exactly the graph induced by these m edges. It is easy to see that the procedure described to find the new cluster centers using the threshold graph takes time linear in the number of edges of the graph, assuming that the edges are given in the form of an adjacency list. Forming the adjacency list from the edges takes linear time. Therefore, the total cost of the merging phase is bounded by $O(m \log k + m) = O(m \log k)$ time. The credit of $\log k$ placed with each edge when it is inserted into the heap accounts for this cost, completing the proof. ☐

Finally, we describe a Randomized Doubling Algorithm with significantly better performance ratio. The algorithm is essentially the same as before, the main change being in the value of d_1 which is the lower bound for phase 1. In the deterministic case we chose d_1 to be the minimum pairwise distance of the first $k+1$ points, say x. We now choose a random value r from $[1/e, 1]$ according to the probability density function $1/r$, set d_1 to rx, and redefine $\beta = e$ and $\alpha = e/(e-1)$. Similar randomization of doubling algorithms was used earlier in scheduling [41], and later in other applications [10, 26].

THEOREM 3.8. *The Randomized Doubling Algorithm has expected performance ratio* $2e \approx 5.437$ *in any metric space. The same bound is also achieved for the radius measure.*

Proof. Let σ be the sequence of updates and let the optimal cluster diameter for σ be γx for some $\gamma \geq 1$; by the definition of x, the optimal value is at least x. Suppose we choose $d_1 = rx$ for some $r \in (0, 1]$. Let ρ_r be the radius of the clusters created for σ with this value of r. Using arguments similar to those in the proof of Theorem 3.5, we can show that ρ_r is at most $d_{i+1} + \alpha d_i = e^{i+1} d_1/(e-1)$, where i is the largest integer such that $d_i = e^{i-1} d_1 = e^{i-1} rx \leq \text{OPT} = \gamma x$. Let i^* be the integer such that $e^{i^*-1} \leq \gamma < e^{i^*}$ and $\delta = \gamma / e^{i^*}$. Then, $\rho_r \leq \frac{rex\gamma}{(e-1)\delta}$ when $r > \delta$, and $\rho_r \leq \frac{re^2 x\gamma}{(e-1)\delta}$ when $r \leq \delta$. Let X_r^- and X_r^+ be indicator variables for the the events $[r \leq \delta]$ and $[r > \delta]$, respectively. We claim that the expected value of ρ_r is bounded by

$$
\begin{aligned}
E[\rho_r] &\leq \int_{1/e}^{1} \frac{re\gamma x(eX_r^- + X_r^+)}{\delta r(e-1)} \mathrm{d}r \\
&= \text{OPT} \frac{e}{\delta(e-1)} \int_{1/e}^{1} (eX_r^- + X_r^+) \mathrm{d}r \\
&= \text{OPT} \frac{e\delta(e-1)}{\delta(e-1)} \\
&= e\text{OPT}.
\end{aligned}
$$

Therefore, the expected diameter is at most $2e$OPT. □

REMARK 3.9. *The randomized version of the Doubling algorithm can be converted to a deterministic algorithm for the offline case which runs in* $O(\frac{1}{\epsilon}nk^2)$ *time and has a performance ratio of* $4(1+\epsilon)$.

3.1. An oblivious clustering algorithm. In this section we describe an incremental hierarchical clustering algorithm that does not need an a priori bound on the number of clusters. From the hierarchy, given an integer k, we can obtain a k-clustering of the points such that the diameter of the clustering is at most a factor of 8 times the optimum diameter for the given value of k.

For convenience, assume that we have an a priori upper bound on the maximum distance between points, which we take to be 1. We maintain the current points in a tree, where every point is a vertex of the tree. For convenience, we assume that if a point is a vertex of the tree, then one of its children corresponds to the same vertex. The root, at depth 0, is a single vertex. The vertices at depth $i \geq 1$ are points within distance $1/2^{i-1}$ of their parents, and all the vertices at depth i are at distance greater than $1/2^i$ from each other.

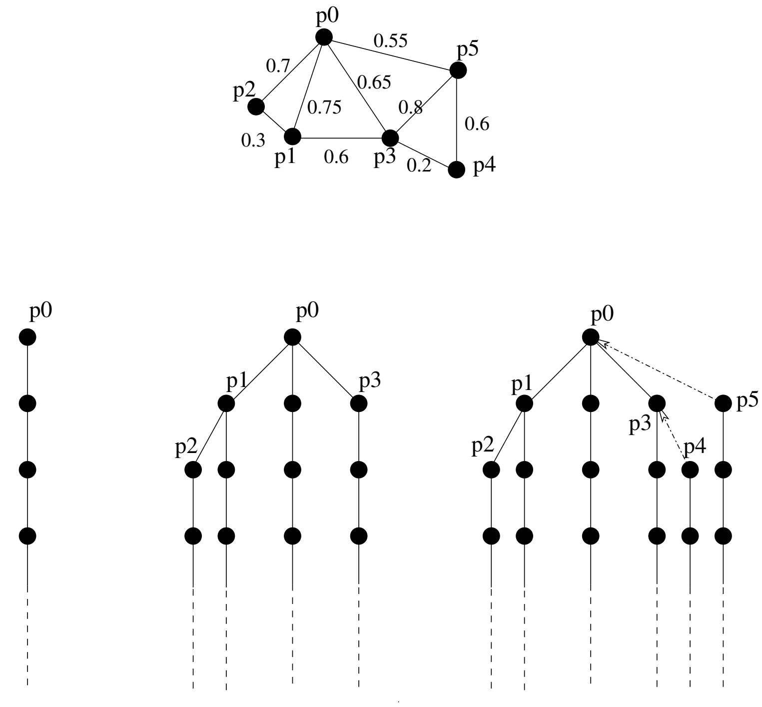
Initially, we have a single point at the root, which also occurs as the only child, the only child of this child, and so on; for conceptual clarity, we may assume that the depth of the tree is infinite. Suppose we have a tree representing the points at some stage, and a new point p comes in. Let i be the largest integer such that p is within $1/2^i$ of some point q at depth i. Then p is added at depth $i+1$, with q as its parent, with p as its only child, p as the only child of the child, and so on. So p is at depth $i+1$ with parent within distance $1/2^i$ as desired, and also the vertices representing p at depth $j \geq i+1$ are at distance greater than $1/2^j$ of other points at depth j. See Fig 3.1 for an example.

It remains to indicate how we obtain the k clusters from the tree when k is given. Let i be the greatest depth containing at most k points. These are the k cluster centers. The subtrees of the vertices at depth i are the clusters. As points are added, the number of vertices at depth i increases; if it goes beyond k, then we change i to $i-1$, collapsing certain clusters; otherwise, the new point is inserted in one of the existing clusters.

THEOREM 3.10. *The algorithm that outputs the k clusters obtained from the tree construction has performance ratio* 8 *for the diameter measure and the radius measure.*

Proof. Suppose the optimal diameter is $1/2^{i+1} < d \leq 1/2^i$. Then points at depth i are in different clusters, so there are at most k of them. Let $j \geq i$ be the greatest depth containing at most k points. Then these are the cluster centers, and the vertices in the corresponding subtrees are at distance of the root within $1/2^j + 1/2^{j+1} + 1/2^{j+2} + \cdots \leq 1/2^{i-1} < 4d$. Hence the radius of the clusters is at most $4d$, and thus the diameter is at most $8d$. This proof also implies that if the optimal radius is r, the radius of the clusters output is at most $8r$. □

The randomization technique used in the Randomized Doubling Algorithm can also be applied here to get a better performance ratio (in expectation). The construction of the tree described above used distance threshold $1/2^i$ for depth i. Instead, we use distance threshold r/e^i for depth i, where r is chosen at random from the interval $[1, e]$, according to the probability density function $1/r$. Note that the value r is chosen once at the beginning of the tree construction. By performing an analysis very similar to the that for the Randomized Doubling Algorithm, we can show that the expected diameter of the k clustering obtained from the tree is at most $2e \approx 5.437$



(a) (b) (c)

FIG. 3.1. *Illustration of the oblivious clustering algorithm. The points are numbered according to their arrival order and the distance between points is the minimum of* 1 *and the distance induced by the weighted graph shown. (a) Tree after first point* p0 *arrives. (b) Tree after four points. (c) Addition of* p4 *and* p5*. If* $k = 3$*, in (b), the clusters are centered around* p0, p1, *and* p3*; in (c) there is a single cluster around* p0.

times the optimal diameter. The same bound holds for the radius as well.

4. The Clique Partition Algorithm. We now describe the Clique Algorithm which has performance ratio 6. This does not totally improve upon the Doubling Algorithm since the new algorithm involves solving the NP-hard clique partition problem, even though it is only on a graph with $k+1$ vertices. Finding a minimum clique partition is NP-hard even for graphs induced by points in the Euclidean plane [25], although it is in polynomial time for points on the line. Since the algorithm needs to solve the clique partition problem on graphs with $k+1$ vertices, this may not be too inefficient for small k.

DEFINITION 4.1. *Given an undirected unweighted graph* $G = (V, E)$*, an* l*-clique partition is a partition of* V *into* $V_1, V_2, \ldots, V_l$ *such that for* $1 \leq i \leq l$*, the induced graph* $G[V_i]$ *is a clique. A* minimum clique partition *is an* l*-clique partition with the minimum possible value of* l*.*

The Clique Algorithm is similar to the Doubling algorithm in that it also operates in phases which have a merging stage followed by an update stage. The invariants maintained by the algorithm are different though. At the start of the ith phase we have $k+1$ clusters $C_1, C_2, \ldots, C_{k+1}$ and a value d_i such that:

1. the radius of each cluster C_j is at most $2d_i$;

2. the diameter of each cluster C_j is at most $3d_i$;
3. and, $d_i \leq$ OPT.

The merging works as follows. Let $d_{i+1} = 2d_i$. We form the minimum clique partition of the d_{i+1}-threshold graph G of the cluster centers. The new clusters are then formed by merging the clusters in each clique of the clique partition. We arbitrarily choose one cluster from each clique and make its center the cluster center of the new merged cluster. Let $C'_1, C'_2, \ldots, C'_{l_i}$ be the resulting clusters. In the rest of the phase we also need to know which old clusters merged to form each of the new clusters.

LEMMA 4.2. *After the merging stage, the radius of the new clusters is at most $2d_{i+1}$ and the diameter is at most $3d_{i+1}$.*

Proof. Let $C_{j_1}, C_{j_2}, \ldots, C_{j_{n_j}}$ be the clusters whose union is the new cluster C'_j and without loss of generality assume that the center of C_{j_1} was chosen to be the center of C'_j. Since the centers of $C_{j_1}, C_{j_2}, \ldots, C_{j_{n_j}}$ induce a clique in the d_{i+1}-threshold graph, the distance from the new center to each of the old cluster centers is at most d_{i+1}. The radius of each of $C_{j_1}, C_{j_2}, \ldots, C_{j_{n_j}}$ is at most $2d_i$. Therefore it follows that the new radius is at most $d_{i+1} + 2d_i \leq 2d_{i+1}$ and the diameter is at most $2d_i + d_{i+1} + 2d_i \leq 3d_{i+1}$. □

During the update phase, a new point p is handled as follows. Let the current number of clusters be m, where $l_i \leq m \leq k$. Recall that $C'_1, C'_2, \ldots, C'_{l_i}$ are the clusters formed during the merging stage. If there exists j such that $j > l_i$ and $d(p, c'_j) \leq d_{i+1}$, or if $j \leq l_i$ and $d(p, c_{j_s}) \leq d_{i+1}$ where C_{j_s} is a cluster which merged to form C'_j, add p to the cluster C'_j. If no such j exists, make a new cluster with p as the center. The phase ends when the number of clusters exceeds $k+1$, or if there are $k+1$ clusters at the end of the merging phase.

The following lemmas show that the clusters at the end of phase i satisfy the invariants for phase $i+1$.

LEMMA 4.3. *The radius of the clusters at the end of the phase i is at most $2d_{i+1}$ and the diameter of the clusters is at most $3d_{i+1}$.*

Proof. From Lemma 4.2, the radius of the clusters at the end of the merging stage of phase i is at most $2d_{i+1}$ and the diameter is at most $3d_{i+1}$. We now argue that these bounds are maintained during the update stage. If a point p is added to a cluster C_j, $j > l_i$ it satisfies the condition $d(p, c'_j) \leq d_{i+1}$, hence the radius of such a cluster is at most d_{i+1}, and hence the diameter is at most $2d_{i+1}$. If p is added to a cluster C'_j, $j \leq l_i$, then p satisfies the condition $d(p, c_{j_s}) \leq d_{i+1}$ where c_{j_s} is the center of a cluster that was merged to form the cluster C'_j. The center c'_j of C'_j is at distance at most d_{i+1} from the centers of all the clusters that were merged to form C'_j. Therefore p is at distance $2d_{i+1}$ from c'_j. This shows the radius bound. For the diameter bound, consider any two points p and q in the cluster C'_j. There must be cluster centers c_{j_s} and c_{j_t} of clusters that merged to form C'_j such that $d(p, c_{j_s}) \leq d_{i+1}$ and $d(q, c_{j_t}) \leq d_{i+1}$. The diameter bound follows since $d(c_{j_s}, c_{j_t}) \leq d_{i+1}$. □

LEMMA 4.4. *At the end of phase i, $d_{i+1} \leq$ OPT.*

Proof. Suppose $d_{i+1} >$ OPT. Let $S = \{c_1, c_2, \ldots, c_{k+1}\}$ be the cluster centers at the beginning of the phase i. Note that the centers $c'_1, \ldots, c'_{l_i}$ belong to S. Let $S' = \{c'_j \mid j > l_i\}$ be the set of cluster centers of the clusters which are formed in phase i after the merging stage. Since each of the centers c'_j in S' started a new cluster $d(c'_j, p) > d_{i+1}$ for all $p \in S \cup S' - \{c'_j\}$. Therefore in the optimal solution each center in S' is in a cluster which contains no center in S. This implies that the centers in S are contained in at most $l_i - 1$ clusters of diameter d_{i+1}. This is a contradiction since

l_i was the size of the minimum clique partition of the d_{i+1}-threshold graph on S. □

THEOREM 4.5. *The Clique Algorithm has performance ratio* 6 *in any metric space.*

Proof. The diameter of the clusters during phase i is at most $3d_{i+1}$ and we maintain the invariant that $d_i \leq \text{OPT}$ at the start of the phase. Therefore the performance ratio of the algorithm is at most $3d_{i+1}/d_i \leq 6$. □

The analysis of the Clique algorithm is tight as shown by the following example. We assume that $k \geq 4$. The sequence consists of $k+3$ points $p_1, \ldots, p_{k+1}, p_{k+2}, p_{k+3}$. The first k points form a uniform metric space with distance 1. The point p_{k+1} is at distance 2 from p_2 and distance 1 from p_i for $2 \leq i \leq k$. The point p_{k+2} has the following properties: $d(p_{k+2}, p_1) = 2$, $d(p_{k+2}, p_i) = 3$ for $2 \leq i \leq k$, $d(p_{k+2}, p_{k+1}) = 4$, and $d(p_{k+2}, p_{k+3}) = 6$. The point p_{k+3} has the following properties: $d(p_{k+3}, p_1) = 4$, $d(p_{k+3}, p_i) = 3$ for $2 \leq i \leq k$, $d(p_{k+3}, p_{k+1}) = 2$, and $d(p_{k+3}, p_{k+2}) = 6$. After the first $k+1$ points are given the algorithm enters phase 1 with $d_1 = 1$ and $d_2 = 2$. In the merging stage the first $k+1$ points are merged into one single cluster since they form a clique in the d_2-threshold graph. It is easy to see that points p_{k+2} and p_{k+3} will be added to this cluster in the update stage and the diameter of this cluster is seen to be 6. There is however an optimal clustering that achieves diameter 1 for all clusters: p_{k+2} and p_{k+3} are in their own clusters, p_1 and p_{k+1} are in different clusters and the rest of the points can go either into the cluster containing p_1 or the one containing p_{k+1}.

Since the radius of the clusters is within 4 of the optimal diameter, we obtain the following corollary.

COROLLARY 4.6. *The Clique Algorithm has performance ratio* 8 *in any metric space for the radius measure.*

As in the case of the Doubling Algorithm, we can use randomization to improve the bound. Let x be the minimum distance among the first $k+1$ points. The randomized algorithm sets $d_1 = rx$ in phase 1 of the deterministic algorithm, where r is chosen from $[1/2, 1]$ according to the probability density function $\frac{1}{r \ln 2}$. The analysis is similar to that of Theorem 3.8 and we omit the details.

THEOREM 4.7. *The Randomized Clique Algorithm has performance ratio* $\frac{3}{\ln 2} \approx$ 4.33 *in any metric space.*

COROLLARY 4.8. *The Randomized Clique Algorithm has performance ratio* $\frac{4}{\ln 2} \approx$ 5.77 *for the radius measure in any metric space.*

The special structure of the clusters in the Clique Algorithm can be used to show that the performance ratio for the radius measure is better than 8 for the geometric case. This is based on the following result in geometry, known as Jung's theorem ([5],p.84-85).

PROPOSITION 4.9. *Any convex set in* R^d *of diameter at most* 1 *can be circumscribed by a sphere of radius* $r_d = \sqrt{d/(2d+2)}$,

THEOREM 4.10. *The Clique Algorithm has performance ratio* $4(1 + r_d)$ *for the* radius *measure in* R^d. *This implies performance ratio* 6 *for* $d = 1$, 6.3 *for* $d = 2$, *and* 6.83 *asymptotically for large* d.

Proof. Let a_i be the maximum radius of the clusters of the algorithm in phase i and let a_i^* be the radius achievable by an optimal clustering. We claim that $a_i \leq d_{i+1}(1 + r_d) = 2d_i(1 + r_d)$. Trivially, we have that $a_i^* \geq \text{OPT}_i/2$ where OPT_i the optimal diameter achievable for points at start of phase i. From the invariants of the algorithm we have that $d_i \leq \text{OPT}_i$, hence we get that $a_i \leq 4a_i^*(1 + r_d)$.

Now we prove the claim by induction. Assuming that the claim is true in phase

$i-1$ we prove that it is true in phase i. In phase i let $C'_1, C'_2, \ldots, C'_{l_i}$ be clusters formed in the merging stage. It is easy to see that any new cluster formed during the update stage has radius at most d_{i+1} hence we can focus on the clusters $C'_1, C'_2, \ldots, C'_{l_i}$. Without loss of generality consider C'_1 which is formed by the merging of clusters $C_1, C_2, \ldots, C_j$ from the start of the phase. Let c_i be the center of C_i, $1 \leq i \leq j$. Since the clusters were merged, it follows that the diameter of the point set $\{c_1, c_2, \ldots, c_j\}$ is at most d_{i+1}. Hence their radius, by Proposition 4.9, is at most $d_{i+1} r_d$. From the invariants at the start of the phase, $d(p, c_i) \leq 2d_i = d_{i+1}$ for $p \in C_i$. Further any new point added to the cluster C'_1 is at most a distance of d_{i+1} from a point in $\{c_1, c_2, \ldots, c_j\}$. It follows that the radius of the cluster C'_1 through out the phase is at most $d_{i+1} r_d + d_{i+1}$ which proves the claim. ∎

5. Lower Bounds. We present some lower bounds on the performance of incremental clustering. The lower bounds apply to both diameter and radius measures but our proofs are given for the diameter case. The following theorem shows that even for the simplest geometric space, we cannot expect a ratio better than 2.

THEOREM 5.1. *For $k = 2$, there is a lower bound of 2 and $2 - 1/2^{k/2}$ on the performance ratio of deterministic and randomized algorithms, respectively, for incremental clustering on the line.*

Proof. Start with the three points $1, 2, 3$. Two consecutive points are necessarily merged, say 2 and 3. Add a new point at 4. Then 1 or 4 is merged with $[2, 3]$. This gives diameter 2, while the optimum is 1. In fact, this construction can be repeated to obtain n points $1, 2, \ldots, n$ clustered into 1 and $[2, n]$.

For the randomized lower bound, consider the instance in the preceding paragraph. To convert this to a randomized lower bound, the adversary places the fourth point at 0 or 4 with equal probability, and now the algorithm has probability $1/2$ of creating the wrong cluster (that is with diameter 2). This gives a lower bound of 1.5. For general k, the algorithm makes $k/2$ copies of the above scenario far enough from each other that the above analysis applies locally. The probability that the algorithm succeeds in creating clusters of diameter 1 on all $k/2$ copies is $2^{-k/2}$. This implies a lower bound of $2 - 2^{-k/2}$. ∎

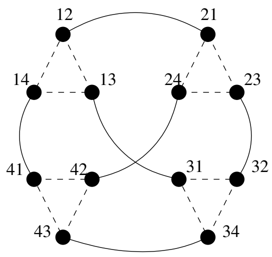
In the case of general metric spaces, we can establish a stronger lower bound.

THEOREM 5.2. *There is a lower bound of $1+\sqrt{2} \approx 2.414$ on the performance ratio of any deterministic incremental clustering algorithm for arbitrary metric spaces.*

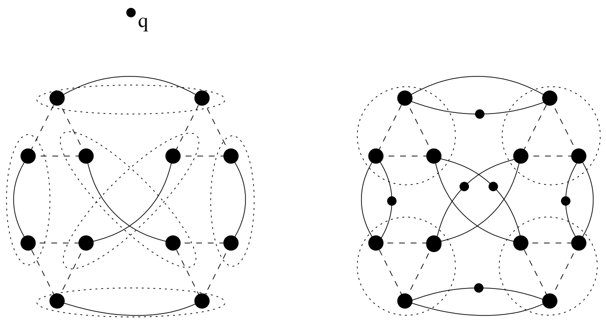
Proof. Consider a metric space consisting of the points p_{ij}, $1 \leq i, j \leq 4$, $i \neq j$. The distances between the points are the shortest path distances in the graph with the following distances specified: $d(p_{ij}, p_{ji}) = 1$, and $d(p_{ij_1}, p_{ij_2}) = \sqrt{2}$ for $j_1 \neq j_2$. Let $B_i = \{p_{ij} \mid 1 \leq j \leq 4, i \neq j\}$. Note that the sets B_i, $1 \leq i \leq 4$, partition the metric space into 4 clusters of diameter $\sqrt{2}$. See Fig 5.1. Let A be any deterministic algorithm for the incremental clustering problem. Let $k = 6$. Consider the clusters produced by A after it is given the 12 points p_{ij} described above.

Case 1: Suppose the maximum diameter of A's clusters is 1. Then A's clusters must be the 6 sets $\{p_{ij}, p_{ji}\}$. Now the adversary gives a point q such that $d(q, p_{ij}) = 10$ for $1 \leq i, j \leq 4$. The optimal clustering is $\{q\}$ and the sets B_1, B_2, B_3, B_4. The optimal diameter is $\sqrt{2}$. We claim that the maximum diameter of A is at least $2 + \sqrt{2}$. If the cluster that contains q contains any other point then our claim is clearly true. If on the other hand, the cluster that contains q does not contain any other point, A must have merged two of its existing clusters. Then the maximum diameter of A's resulting clusters is at least $2 + \sqrt{2}$. Thus the performance ratio of A is at least $1 + \sqrt{2}$.

Case 2: Suppose the maximum diameter of A's clusters is greater than 1. Then some



The first 12 points



Case 1 Case 2

FIG. 5.1. *Lowerbound instance. Solid and dashed lines show edges of length* 1 *and* $\sqrt{2}$ *respectively. Dotted ellipses show potential clusters. New points inserted after the first batch of* 12 *points are shown as smaller circles.*

cluster of A contains 2 points which are at least distance $\sqrt{2}$ apart. Let these points be p_{wx} and p_{yz}, $(w,x) \neq (z,y)$. Now the adversary gives 6 points r_{ij}, $1 \leq i < j \leq 4$ such that $d(r_{ij}, p_{ij}) = d(r_{ij}, p_{ji}) = 1$. The optimal clustering consists of the 6 sets $\{r_{ij}, p_{ij}, p_{ji}\}$. The optimal diameter is 1. We claim that the maximum diameter of A's clusters must be at least $1+\sqrt{2}$. Note that $d(r_{i_1 j_1}, p_{i_2 j_2}) \geq 1+\sqrt{2}$ for $(i_2, j_2) \neq (i_1, j_1)$, $(i_2, j_2) \neq (j_1, i_1)$. Also $d(r_{i_1 j_1}, r_{i_2 j_2}) \geq 2+\sqrt{2}$ for $(i_1, j_1) \neq (i_2, j_2)$. If A puts any two r_{ij} in the same cluster, our claim is clearly true. If all the r_{ij} are in separate clusters, each of the 6 clusters must contain one of them. Also one of the 6 clusters, say C must contain both the points p_{wx} and p_{yz}. Then C must have diameter at least $1+\sqrt{2}$, since the r_{ij} in C must be at distance at least $1+\sqrt{2}$ from one of p_{wx} and p_{yz}. Hence the performance ratio of A is at least $1+\sqrt{2}$.

This proves a lower bound of $1+\sqrt{2}$ on the performance ratio of any deterministic incremental algorithm. ▯

Finally, we can improve the randomized lower bound slightly for the case of arbitrary metric spaces.

THEOREM 5.3. *For any* $\epsilon > 0$ *and* $k \geq 2$, *there is a lower bound of* $2-\epsilon$ *on the performance ratio of any randomized incremental algorithm.*

Proof. We use Yao's technique and show a lower bound on deterministic algorithms on an appropriately chosen distribution. Let A be a deterministic algorithm for incremental clustering. The distribution on inputs is as follows. Initially, the adversary provides n points $P_1, P_2 \ldots P_n$ such that the distance between any two of them is 1. Then the adversary partitions the n points into k disjoint sets $S_1, S_2 \ldots S_k$ at random, such that all partitions are equally likely. Finally the adversary provides k points $Q_1, Q_2, \ldots Q_k$, such that $d(Q_i, P_j) = 1$ if $P_j \in S_i$, $d(Q_i, P_j) = 2$ if $P_j \notin S_i$, $d(Q_i, Q_j) = 3$. Now, the diameter of the optimal solution for any input in the distribution is 1, obtained by constructing the k clusters $S_i \cup \{Q_i\}$. However, the incremental algorithm can produce a clustering with diameter 1 only if the clusters it produces after it sees points $P_1, P_2 \ldots P_n$ are precisely the sets S_i (selected at random by the adversary). Let $X_k(n)$ be the number of ways to partition the n points into k sets. Then the probability that the incremental algorithm produces a clustering of diameter 1 is at most $p = 1/X_k(n)$. With probability at least $1 - p$, the incremental algorithm produces a clustering of diameter at least 2. Thus the expected value of the diameter of the clustering produced is at least $2 - p$. Hence the expected value of the performance ratio is at least $2 - p$. By choosing n suitably large, $X_k(n)$ can be made arbitrarily large, and hence p can be made arbitrarily small, in particular smaller than ϵ for any fixed $\epsilon > 0$. □

For the radius measure we have the following theorem.

THEOREM 5.4. *For the radius measure, no deterministic incremental clustering algorithm has a performance ratio better than* 3 *and no randomized algorithm has a ratio better than* $3 - \epsilon$ *for any fixed* $\epsilon > 0$.

Proof. We first consider the randomized case. The instances are very similar to the ones used in the proof of Theorem 5.3. The only difference is that $d(Q_i, P_j) = 0.5$ if $P_j \in S_i$ and $d(Q_i, P_j) = 1.5$ if $P_j \notin S_i$. The optimal clusters are $S_i \cup \{Q_i\}$ for $1 \leq i \leq k$ and each of them has a radius of 0.5. Any other clustering has radius at least 1.5 and the algorithm has a probability of $(1 - 1/X_k(n))$ of having such a cluster.

The instance used above can be adapted easily to show a bound of 3 for the deterministic case, we leave the details to the reader. □

6. Dual Clustering. We now consider the dual clustering problem: for a sequence of points $p_1, p_2, \ldots, p_n \in \Re^d$, cover each point with a unit ball in $\Re^d$ as it arrives, so as to minimize the total number of balls used. In the static case, this problem is NP-complete and a PTAS is achievable in any fixed dimension [31]. We note that in general metric spaces, it is impossible to achieve any bounded ratio (for example, consider the uniform metric space).

Our algorithm's analysis is based on a theorem from combinatorial geometry called Roger's theorem [46] (see also Theorem 7.17 [43]), which says that R^d can be covered by any convex shape with covering density $O(d \log d)$. Since the volume of a radius 2 ball is 2^d times the volume of a unit-radius ball, the number of balls needed to cover a ball of radius 2 using balls of unit radius is $f(d) = O(2^d d \log d)$. We first describe an incremental algorithm which has performance ratio $f(d)$. We also establish an asymptotic lower bound of $\Omega(\frac{\log d}{\log \log \log d})$; for $d = 1$ and 2, our proof yields lower bounds of 2 and 4, respectively.

THEOREM 6.1. *For the dual clustering problem in* $\Re^d$*, there is an incremental algorithm with performance ratio* $O(2^d d \log d)$.

Proof. Our incremental algorithm maintains a set $\mathcal{C}$ of *centers* which is a subset of the points that have arrived so far; initially, $\mathcal{C} = \emptyset$. Define the *range* $R(p)$ of a center p to be the sphere of radius 2 about p. For any two centers p_1 and p_2, we ensure that $d(p_1, p_2) > 2$. Associated with each center p is a set of points $\Gamma(p)$ called the *neighbors* of p. For convenience, we assume that $p \in \Gamma(p)$. We ensure that all neighbors of p lie in $R(p)$. When a new point x is received, if $x \in R(p)$ for some center

p, we add x to $\Gamma(p)$, breaking ties arbitrarily. If no such center exists, x must be at a distance greater than 2 from all the existing centers. In this case, we make x a new center and set $\Gamma(x) = \{x\}$.

From Roger's theorem on packing density, a sphere of radius 2 in $\Re^d$ can be covered by $f(d) = O(2^d d \log d)$ spheres of radius 1. When a new center p is created, we fix a set of spheres $S_1(p), S_2(p), \ldots, S_{f(d)}(p)$ which cover $R(p)$. Whenever a point x is added to $\Gamma(p)$, if it is not already covered by some previously placed sphere, we add the sphere $S_r(p)$ where r is any value such that $x \in S_r(p)$. Note that such a sphere must exist as $x \in R(p)$ and the spheres $S_1(p), S_2(p), \ldots, S_{f(d)}(p)$ cover $R(p)$ completely.

Since no two centers can be covered by a sphere of unit radius, any solution must use a separate sphere to cover each center. Hence, the number of centers is a lower bound for the number of spheres used by the optimal offline algorithm. For each center p, the incremental algorithm uses at most $f(d)$ spheres to cover the points in $\Gamma(p)$. Hence, the performance ratio of the incremental algorithm is bounded by $f(d) = O(2^d d \log d)$. □

THEOREM 6.2. *For the dual clustering problem in $\Re^d$, any incremental algorithm must have performance ratio $\Omega(\frac{\log d}{\log \log \log d})$.*

Proof. The idea is as follows. At time t, when t points have been given by the adversary, it will be the case that the points $p_1, \ldots, p_t$ can be covered by a ball of radius $R_t < 1$. Then, the adversary will find a point p_{t+1} lying outside the t unit balls laid down by the algorithm so as to minimize the radius R_{t+1} of the ball required to cover all $t+1$ points and present that as a request. The game terminates when at some time $k+1$, we have for the first time that $R_{k+1} > 1$. Clearly, k is a lower bound on the performance ratio since the points $p_1, \ldots, p_k$ can be covered by a ball of radius $R_k \leq 1$, and the algorithm has used k balls up to that point. It remains to analyze the worst-case growth rate of R_t as a function of t. Note that $R_1 = 0$ and $R_2 = 1/2$.

Let α_d denote the volume of a unit ball in $\Re^d$. At time t, let D_t be any ball of radius (at most) R_t that covers the points $p_1, \ldots, p_t$. For some δ_t to be specified later, define the ball D_t^* as a ball with the same center as D_t and with radius $R_t + \delta_t$. We will choose δ_t such that the volume of D_t^* is at least $t\alpha_d$, implying that the current t unit balls placed by the algorithm cannot cover the entire volume of D_t^*. This would imply that there is a choice of a point p_{t+1} inside D_t^* which is not covered by the current t balls. It is also clear that the new set of $t+1$ points now can be covered by a ball of radius at most $R_t + \delta_t/2$. implying that

$$R_{t+1} \;=\; R_t + \frac{\delta_t}{2}.$$

To determine the value of δ_t is easy, since we have the inequality that

$$\alpha_d(R_t + \delta_t)^d \;>\; t\alpha_d$$

from the requirement that the ball D_t have volume equal to that of t unit balls. Now let $R_t = 1 - \epsilon_t$. Substituting in the above equations we obtain that $\delta_t = 2(\epsilon_t - \epsilon_{t+1})$ and hence $R_t + \delta_t = 1 + \epsilon_t - 2\epsilon_{t+1}$. Therefore

$$\alpha_d(1 + \epsilon_t - 2\epsilon_{t+1})^d \;>\; t\alpha_d$$

which implies that

$$\ln(1 + \epsilon_t - 2\epsilon_{t+1}) \;>\; \frac{\ln t}{d} \;.$$

Note that $\epsilon_t - 2\epsilon_{t+1} < 1$. Using the fact that $\ln(1+x) > x/2$ for $x < 1$, we see that choosing ϵ_i such that

$$\frac{\epsilon_t - 2\epsilon_{t+1}}{2} = \frac{\ln t}{d}$$

will satisfy our requirements. Unfolding the recurrence,

$$\frac{\epsilon_1}{2^t} - \epsilon_{t+1} = \sum_{i=1}^{t} \frac{\ln i}{d2^{t-i}} = \sum_{i=1}^{t} \frac{\ln i}{d2^{t-i}} \leq \sum_{i=1}^{t} \frac{\ln t}{d2^{t-i}} \leq \frac{2\ln t}{d} .$$

Noting that $\epsilon_1 = 1$, we obtain that $\epsilon_{t+1} \geq 2^{-t} - 2d^{-1}\ln t$. The lower bound is the smallest value of t for which ϵ_{t+1} is negative. Let t_{max} be the largest value of t for which

$$\frac{1}{2^t} - \frac{2\ln t}{d} \geq 0 .$$

This implies that $2^{t+1}\ln t \leq d$ and hence

$$t_{max} = \Omega\left(\frac{\log d}{\log\log\log d}\right).$$

This gives the desired lower bound. $\Box$

Acknowledgments. We thank Pankaj Agarwal and Leonidas Guibas for helpful discussions, and for suggesting that we consider the dual clustering problem. We thank Bernard Chazelle for pointing out reference [5].

REFERENCES

[1] M.S. Aldenderfer and R.K. Blashfield. *Cluster Analysis.* Sage, Beverly Hills, 1984.

[2] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristic for k-median and facility location problems. *Proc. 33rd STOC*, pp. 21-29, 2001.

[3] Y. Bartal, M. Charikar, and D. Raz, Approximating min-sum k-clustering in metric spaces. *Proc. 33rd STOC*, pp. 11-20, 2001.

[4] M. Bern and D. Eppstein. Approximation Algorithms for Geometric Problems. In: D.S. Hochbaum, editor, *Approximation Algorithms for NP-Hard Problems.* PWS Publishing Company, 1996.

[5] T. Bonnesen and W. Fenchel. *Theorie der konvexen Körper.* Springer, Berlin 1934; English translation: BCS Associates, Moscow, Idaho, 1987.

[6] P. Brucker. On the complexity of clustering problems. In: R. Henn, B. Korte, and W. Oletti, editors, *Optimization and Operations Research,* Heidelberg, New York, NY, 1977, pp. 45–54.

[7] F. Can. Incremental Clustering for Dynamic Information Processing. *ACM Transactions on Information Processing Systems*, 11 (1993), pp. 143–164.

[8] F. Can and E.A. Ozkarahan. A Dynamic Cluster Maintenance System for Information Retrieval. In *Proceedings of the Tenth Annual International ACM SIGIR Conference,* 1987, pp. 123-131.

[9] F. Can and N.D. Drochak II. Incremental Clustering for Dynamic Document Databases. In *Proceedings of the 1990 Symposium on Applied Computing,* 1990, pp. 61–67.

[10] S. Chakrabarti, C. Phillips, A. Schulz, D.B. Shmoys, C. Stein, and J. Wein. Improved Scheduling Algorithms for Minsum Criteria. In *Proceedings of the 23rd International Colloquium on Automata, Languages and Programming*, Springer, 1996.

[11] M. Charikar, S. Guha, E. Tardos, and D. Shmoys. A Constant-Factor Approximation Algorithm for the k-Median Pro blem. *Journal of Computer and System Sciences*, 65(1) (2002), pp. 129-149.

[12] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. *Proc. 12th SODA* pp. 642-651, 2001.

[13] M. Charikar, L. O´Callaghan and R. Panigrahy, *Better Streaming Algorithms for Clustering Problems*, in *Proc. 35th Annual ACM Symposium on Theory of Computing (STOC)*, (2003), pp. 30-39.

[14] M. Charikar and R. Panigrahy. Clustering to minimize the sum of cluster diameters. *Proc. 33rd STOC*, pp. 1-10, 2001.

[15] B.B. Chaudhri. Dynamic clustering for time incremental data. *Pattern Recognition Letters*, 13 (1994), pp. 27-34.

[16] D.R. Cutting, D.R. Karger, J.O. Pederson, and J.W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference*, 1992, pp. 318–329.

[17] D.R. Cutting, D.R. Karger, and J.O. Pederson. Constant interaction-time Scatter/Gather Browsing of Very Large Document Collections. In *Proceedings of the 16th Annual International ACM SIGIR Conference*, 1993, pp. 126–135.

[18] W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for clustering problems. In *Proceedings of the 35th Annual ACM Sympoium on Theory of Computing*, pp. 50-58 (2003).

[19] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, NY, 1973.

[20] B. Everitt. *Cluster Analysis.* Heinemann Educational, London, 1974.

[21] C. Faloutsos and D.W. Oard. A Survey of Information Retrieval and Filtering Methods. Technical Report CS-TR-3514, Department of Computer Science, University of Maryland, College Park, 1995.

[22] T. Feder and D.H. Greene. Optimal Algorithms for Approximate Clustering. *In Proceedings of the Twentieth Annual Symposium on Theory of Computing*, 1988, pp. 434–444.

[23] R.J. Fowler, M.S. Paterson, and S.L. Tanimoto. Optimal packing and covering in the plane are NP-complete. *Information Processing Letters,* 12 (1981), pp. 133–137.

[24] W. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, 1992.

[25] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, 1979.

[26] M. Goemans and J. Kleinberg. An improved approximation ratio for the minimum latency problem. *In Proceedings of the Seventh ACM-SIAM Symposium on Discrete Algorithms*, 1996, pp. 152–157.

[27] T.E. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science,* 38 (1985), pp. 293–306.

[28] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. *Proc. 41st FOCS*, pp. 359–366, 2000.

[29] J.A. Hartigan. *Clustering Algorithms.* Wiley, New York, NY, 1975.

[30] D. Hochbaum. Various Notions of Approximations: Good, Better, Best, and More. In: D.S. Hochbaum, editor, *Approximation Algorithms for NP-Hard Problems.* PWS Publishing Company, 1996.

[31] D.S. Hochbaum and W. Maas. Approximation Schemes for Covering and Packing Problems in Image Processing and VLSI. *Journal of the ACM*, 32 (1985), pp. 130–135.

[32] D.S. Hochbaum and D.B. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research,* 10 (1985), pp. 180–184.

[33] D.S. Hochbaum and D.B. Shmoys. A unified approach to approximation algorithms for bottleneck problems. *Journal of the ACM,* 33 (1986), pp. 533–550.

[34] S. Irani and A. Karlin. Online Computation. In: D.S. Hochbaum, editor, *Approximation Algorithms for NP-Hard Problems.* PWS Publishing Company, 1996.

[35] N. Jardine and C.J. van Rijsbergen. The Use of Hierarchical Clustering in Information Retrieval. *Information Storage and Retrieval*, 7 (1971), pp. 217–240.

[36] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice-Hall, NJ, 1988.

[37] K. Jain and V. Vazirani. Approximation algorithms for metric facility location and k-Median pro blems using the primal-dual schema and Lagrangian relaxation. JACM 48(2): 274-296, (2001).

[38] O. Kariv and S.L. Hakimi. An algorithmic approach to network location problems, part I: the p-centers problem. *SIAM Journal of Applied Mathematics,* 37 (1979), pp. 513–538.

[39] N. Megiddo and K.J. Supowit. On the complexity of some common geometric problems. *SIAM Journal on Computing,* 13 (1984), pp. 182–196.

[40] S.G. Mentzer. Lower bounds on metric k-center problems. Manuscript, 1988.

[41] R. Motwani, S. Phillips and E. Torng. Non-Clairvoyant Scheduling, In *Proceedings of the Fourth ACM-SIAM Symposium on Discrete Algorithms*, 1993, pp. 422–431. See also: *Theoretical Computer Science,* 130 (1994), pp. 17–47.
[42] E. Omiecinski and P. Scheuermann. A Global Approach to Record Clustering and File Organization. In *Proceedings of the Third BCS-ACM Symposium on Research and Development in Information Retrieval,* 1984, pp. 201–219.
[43] J. Pach and P.K. Agarwal. *Combinatorial Geometry.* John Wiley & Sons, New York, NY, 1995.
[44] E. Rasmussen. Clustering Algorithms. Chapter 16 in: W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms,* Prentice-Hall, Englewood Cliffs, NJ, 1992.
[45] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.
[46] C. Rogers. A note on coverings. *Mathematika,* 4 (1957), pp. 11–6.
[47] G. Salton. *Automatic Text Processing.* Addison-Wesley, Reading, MA, 1989.
[48] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, New York, NY, 1983.
[49] P. Willett. Recent Trends in Hierarchical Document Clustering: A Critical Review. *Information Processing & Management,* 24 (1988), pp. 577–597.
[50] I.H. Witten, A. Moffat, and T.C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, New York, NY, 1994.

Appendix A. t-Diameter-Greedy Algorithm.

We give a few preliminary results on the t-Diameter-Greedy Algorithm defined in Section 2.

THEOREM A.1. *For $k = 3$, there is a lower bound of 3 on the performance ratio of the Diameter-Greedy Algorithm on the line.*

Proof. We first show that Diameter-Greedy achieves ratio 3 for $k = 3$. Suppose that the optimal clustering is $[r, s], [t, u], [v, w]$ with $\max(s - r, u - t, w - v) = d$. It is sufficient to show that a merging of two out of four clusters does not create a cluster of diameter greater than $3d$. There are two cases: if $t - s, v - u > d$ then this algorithm actually produces the optimal solution; conversely, if $t - s \leq d$, then either the first two out of four clusters are contained in $[r, u]$ with $u - r \leq 3d$, or the last two out of four clusters are contained in $[v, w]$ with $w - v \leq d$.

For the case $k = 2$, we will in fact show that Diameter-Greedy creates two intervals whose radius (the 2-diameter) is at most the diameter of the optimal solution. Suppose the two intervals obtained by the algorithm are $[0, a]$ and $[b, 1]$, with $a < b$. The optimum diameter is achieved when $a \leq 1/2 \leq b$, in which case it is $\max(a, 1 - b)$. The other case has a and b on the same side of $1/2$, say $b < 1/2$. We claim that in this case, there are no gaps greater than b between consecutive points. Consequently, the two consecutive points $x \leq 1/2 \leq y$ satisfy $y - x \leq b$, and the optimum diameter is $d = \max(x, 1 - y)$. The interval $[0, a]$ has radius at most $a < b \leq x \leq d$ as needed. The interval $[b, 1]$ has radius at most $\max(y - b, 1 - y) \leq \max(x, 1 - y) = d$ as needed.

It remains to verify the claim that there are no gaps greater than b. If there is such a gap, it must be inside $[b, 1]$. At some point, a merge crossing this gap was performed. That is, we had three intervals $[r, s], [t, u], [v, w]$ with $t - s > b$, and a merge producing $[r, u]$ was carried out. This can only happen if $w - u > b$. Thus, we obtain two intervals $[r, u]$ and $[v, w]$ with $b < (w - r)/2$, $v - r > b$, and $w - u > b$. We shall show that these three inequalities are preserved for the two current intervals until the end of the algorithm. However, they are false at the end for the two intervals $[0, a]$ and $[b, 1]$, a contradiction.

We show that the three inequalities are preserved. A new point z is added either between the two intervals or outside the two intervals, say after w. If z is added after w, then the resulting intervals are either $[r, u]$ and $[v, z]$, in which case the inequalities still hold, or $[r, w]$ and $[z, z]$, which can only happen if $z - w > b$ and so the inequalities still hold. If z is added between the two intervals, say $z \leq (w+r)/2$ and the two resulting intervals are $[r, z]$ and $[v, w]$, then $w - z \geq (w-r)/2 > b$, completing the proof. □

For $k = 3$, we give an example where greedy based on diameter cannot do strictly better than 3, and show that greedy based on diameter achieves a factor of 3.

THEOREM A.2. *For $k = 3$, there is a lower bound of 3 on the performance ratio of the Diameter-Greedy Algorithm on the line.*

Proof. The adversary gives following sequence of points: $1+\epsilon, 2+\epsilon, 3, 4, 6-2\epsilon, 4-\epsilon, 7-2\epsilon$. The optimal intervals are $[1+\epsilon, 2+\epsilon], [3,4], [6-2\epsilon, 7-2\epsilon]$ giving diameter 1. However, when the first four points are introduced, the interval $[2+\epsilon, 3]$ is created by greedy, when $6-2\epsilon$ is added the interval $[4, 6-2\epsilon]$ is created, when $4-\epsilon$ is added then the enlarged interval $[2+\epsilon, 4-\epsilon]$ is created, and finally when $7-2\epsilon$ is added either $[1+\epsilon, 4-\epsilon]$ or $[4, 7-2\epsilon]$ is created, for a factor of $3-2\epsilon$]. □

The proof only gives a lower bound of 2 for t-Diameter Greedy Algorithm when $t > 1$, leaving open the possibility that these algorithms may perform better than Diameter Greedy. Indeed, we have the following result.

THEOREM A.3. *The 3-Diameter Greedy Algorithm has performance ratio 3 on the line.*

Proof. In fact, we show that it produces a clustering with 3-diameter at most the optimal diameter, and the factor of 3 follows. Assume this holds before the last two clusters are merged. Let $I_1, I_2, \ldots, I_k$ be the intervals in the optimal clustering, with maximum diameter d. Let $C_1, C_2, \ldots, C_{k+1}$ be the current clusters, each with 3-diameter at most d, of which two must be merged. If C_i starts in I_a and ends in I_b, let $x_i = b - a$; notice that $x_1 + \cdots + x_{k+1} \leq k - 1$. We assume that if C_i ends in I_b then C_{i+1} starts in I_b; otherwise, we could replace the argument in the k intervals I_j by an argument either in the first b intervals $I_1, \ldots, I_b$, if there are at least $b+1$ clusters C_i in this region, or in the last $k-b$ intervals $I_{b+1}, \ldots, I_k$, if there are at least $k-b+1$ current clusters C_i in this region. Now, the bounds imply that for some i, we have $x_i + x_{i+1} < 2$. If $x_i = x_{i+1} = 0$, then the merging of C_i and C_{i+1} is contained in a single interval I_j and has diameter at most d. If say $x_i = 0$ and $x_{i+1} = 1$, then the gap G between the two consecutive intervals I_j and I_{j+1} involved is at most d, since C_{i+1} has 3-diameter at most d, so the merger of C_i and C_{i+1} has 3-diameter at most d given by the 3-partition I_j, G, I_{j+1}. This completes the proof.

We comment briefly on the running time of this algorithm. In the above proof, the 3-diameter of an interval may be replaced by an easily-computed upper bound: at the time of creation of interval $[a, b]$, let $[x, y]$ be the gap containing $(a+b)/2$, and let the upper bound be $\max(x-a, y-x, b-y)$. Maintaining the n points sorted in a balanced tree, the running time is $O(\log n)$ for each of the n points inserted. □