Clustering for Edge-Cost Minimization

[Extended Abstract]

Leonard J. Schulman[*]
College of Computing
Georgia Institute of Technology
Atlanta GA 30332-0280

ABSTRACT

We address the problem of partitioning a set of n points into clusters, so as to minimize the sum, over all intracluster pairs of points, of the cost associated with each pair. We obtain a randomized approximation algorithm for this problem, for the cost functions ℓ_2^2, ℓ_1 and ℓ_2, as well as any cost function isometrically embeddable in ℓ_2^2.

In the 2-cluster case the algorithm computes, with high probability, a solution which differs in its labelling of no more than an ε fraction of the points, from a clustering whose cost is within $(1+\varepsilon)$ times optimal. Given a fixed approximation parameter ε, the runtime is linear in n for ℓ_2^2 problems of dimension $o(\log n/\log\log n)$; and $n^{O(\log\log n)}$ in the general case.

The case ℓ_2^2 is addressed by combining three elements: (a) Variable-probability sampling of the given points, to reduce the size of the data set. (b) Near-isometric dimension reduction. (c) A deterministic exact algorithm which runs in time exponential in the dimension (rather than the number of points). The remaining cases are addressed by reduction to ℓ_2^2.

1. INTRODUCTION

We consider the problem of clustering, or classifying, a set of data points T. Clustering is a ubiquitous problem in the analysis of large data sets. It arises whenever there is a need to organise data points by similarity, and identify patterns. The wide variety of applications precludes any single mathematical formulation of the problem. Many straightforward formulations are, as optimization problems, NP-complete. It is therefore critical to identify a clustering criterion which on the one hand is *rich* enough to be useful in applications, and on the other hand is *tractable* enough that optimal or near-optimal clusterings can be efficiently computed. In this paper we propose such a criterion. We demonstrate its versatility by showing how various other clustering criteria reduce to it; and its amenability to computation by providing a fairly efficient algorithm identifying near-optimal clusterings. Crucially, the runtime of the algorithm grows only moderately even in unbounded dimension (rather than suffering the usual exponential "combinatorial explosion"); and in moderate dimension (up to $\log|T|/\log\log|T|$), it runs in linear time.

We adopt an edge-cost minimization approach to the clustering problem. A cost $\phi_{u,v}$ is charged for every pair of points $u, v \in T$ assigned to the same cluster. (Other terms for such a cost are "weight", "penalty", "energy", "dissimilarity" and "distance".) ϕ increases with dissimilarity of points, although it is not necessarily a metric. The clustering task is to partition T into clusters S and $\bar{S} = T - S$ so that the total cost is minimized. In other words, if we define $\phi(S) = \sum_{u,v\in S} \phi_{u,v}$, the task is to find a partition $(S, \bar{S})$ minimizing $\phi(S) + \phi(\bar{S})$ (which we will also write $\phi(S, \bar{S})$). More generally the task is to find a "k-partition" into clusters $\{S_1, ..., S_k\}$ whose sum of costs is minimal among k-partitions.

The edge-cost approach offers two benefits. First, like any approach in which an objective function is to be optimized, it allows clustering algorithms to be compared both on the basis of their runtimes and on the basis of the quality of their output (namely whether they find an optimum or only approximately optimum partition; and whether this is accomplished with certainty or only with high probability). Second, we find out more than just what is the best partition with respect to the stated criterion. The ratio $\phi(S, \bar{S})/\phi(T)$, or $\phi(S_1, ..., S_k)/\phi(T)$, gives an indication of the quality of the partition: the achievable ratio may guide the choice of the most appropriate k, including the possibility that no partitioning hypothesis is supported by the data ($k = 1$). At this level of generality, however, there is no way to find an optimum partition without an exhaustive examination of all $2^{|T|-1}$ (or generally Stirling number, second kind, $S_{|T|,k} \approx k^{|T|}$) partitions of the input set.

In this paper we provide an efficient randomized approximation algorithm for clustering with respect to a range of cost functions ϕ. This is achieved by a combination of sampling; near-isometric dimension reduction; and reduction to the interesting special case in which ϕ is the square of a Euclidean metric (written ℓ_2^2). This case occupies the central place in our approach. Besides ℓ_2^2 we address also any cost function isometrically embeddable in ℓ_2^2, including ℓ_1 and ℓ_2. (Near

[*]Supported in part by an NSF CAREER grant.
schulman@cc.gatech.edu

isometric embedding would also be sufficient.) Much of the paper focusses upon an algorithm which, for any such cost function and any fixed ε, δ, given a clustering problem on n points in any dimension, computes in time $n^{O(\log\log n)}$, with probability at least $1-\delta$, a 2-clustering that is "ε-close" to optimum. In the special case that $\phi = \ell_2^2$ and that the dimension is $o(\log n/\log\log n)$, the runtime is linear in n. That our output is "ε-close" to optimal means that either its cost is within $(1+\varepsilon)$ times optimal; or its cost is less than an ε fraction of the cost of the original data and it differs from an optimum clustering in its labelling of at most an ε fraction of the points. More precisely, if ϕ_{opt} is the cost of an optimal clustering, and S is the clustering that is output, then: (a) $|\phi(S,\bar{S}) - \phi_{\text{opt}}| \leq \varepsilon\phi(T)$ (which implies that if $\phi_{\text{opt}} \geq \varepsilon\phi(T)$ then ϕ_{opt} is multiplicatively approximated to within factor $(1+\varepsilon)$); and (b) if $\phi_{\text{opt}} \leq \varepsilon\phi(T)$ then the fraction of points whose membership must be switched between S and $\bar{S}$ in order to convert S into an optimal clustering, is less than ε. (For $k > 2$ the first of these guarantees is shown.) The second guarantee is especially important, since we may be most interested in the results precisely when a good clustering of the data exists (as signified by $\phi_{\text{opt}} < \varepsilon\phi(T)$). In this case the guarantee provides that, though the estimate of ϕ_{opt} may be inaccurate, the actual clustering output by the algorithm will differ from the optimal clustering only in the mislabelling of a very small fraction of outliers.

As a by-product, the second guarantee also provides validation of our clustering criterion. That is because it establishes that, if the data can be clustered very well according to our criterion, then any good clustering that we output makes a meaningful statement about the data, because any good clustering partitions the points in almost the same way as the optimal one.

It is worth emphasizing that the clustering criterion ϕ of the algorithm is, in an application, likely to be only a crude approximation of an ideal but intractable application-specific criterion; and that the usefulness of the algorithm will rest upon the insensitivity of the clustering task to this modification in the criterion. However, it cannot be hoped that the optimal clusterings according to the ideal criterion and to ϕ, will agree on outliers; therefore, mislabelling of a small fraction of outliers while still obtaining a low cost clustering is an acceptable sacrifice for the sake of an efficient algorithm.

For the case of metric spaces, an approximation algorithm for max cut (i.e. $k = 2$) has been provided in recent independent work by Fernandez de la Vega and Kenyon [14] (building upon [2; 12; 13]). Max cut (maximization of $\phi(T) - \phi(S,\bar{S})$) is NP-complete [31; 42; 18] (even for metric spaces); min cluster (minimization of $\phi(S,\bar{S})$) is equivalent, but multiplicative approximation of these quantities is not equivalent, with min cluster being harder since there is always a clustering for which $\phi(S,\bar{S}) \leq \phi(T)/2$.

The clustering problem is fully interesting already for the case in which the points of T are of equal "significance" or "weight". However, all our results go through, and are in fact more naturally stated, for the case in which the points are weighted by a nonnegative real valued function w; this possibility will also be very useful in the algorithm. So, the more general formulation of the cost function is

$$\phi(S) = \sum_{\{u,v\}\subseteq S} w_u w_v \phi_{u,v}. \tag{1}$$

Observe that without loss of generality the points may be assumed distinct. We will assume throughout that ϕ is symmetric and that $\phi_{u,u} = 0\ \forall u \in T$; hence equivalently $\phi(S) = \frac{1}{2}\sum_{u,v\in S} w_u w_v \phi_{u,v}$. We will also assume throughout that ϕ is nonnegative.

For general references in the field of clustering see [10; 28; 23; 47; 19; 25; 36; 37; 1; 7]; for discussions of a variety of interesting methods and application areas see [44; 40; 41; 46; 34; 33].

Other common approaches to clustering include k-means, expectation-maximization (EM), and agglomerative methods. These methods do not provide guarantees on the quality of their output. Another approach based upon graphs looks for min cuts (multiway cuts for $k > 2$) rather than max cuts; in this case edges of the graph indicate similarity rather than dissimilarity. Min cut is a computationally easier formulation, with polynomial time exact algorithms known for $k = 2, 3$ and $\log k$ approximation for larger k. In other formulations such as k-medians, substantial work has been required even in order to obtain approximations in the plane [3]. (For approaches to this formulation in more general spaces, achieving logarithmic and constant factor approximations, see see [4; 9; 26].)

A key role in our method is played by a random sampling process which, given T, picks a very small weighted collection of points. We show that for a range of cost functions, the cost of this collection is with high probability close to that of the original collection T. Moreover in the case $\phi = \ell_2^2$, a clustering computation for such small samples can also be used to induce a good clustering of T. We therefore begin by describing the sampling process.

2. SAMPLING PROCESS

An optimistic idea for the clustering problem is simply to select each point with some small probability p; hopefully then good partitions of the sample will "lift" to good partitions of the original data. Indeed there is some promise in this approach, for if we let T' denote the selected set, then $E(\phi(T')) = p^2 E(\phi(T))$. However, this is futile as a method of identifying good clusterings of T. For, there are sets T with the following property: for almost all small subsets S of T, the partition of S inherited from the optimum partition of T, is much more expensive than the optimal partition of S. So examining the partitions of random subsets of T does not contribute substantially toward partitioning T.

We propose instead a more interesting sampling method. In this method the selection probabilities are determined by the original weights and by the location of points within T; a variable-weighting method is used to balance the effects of the uneven sampling probabilities.

We analyze the sampling process for any nonnegative cost function ϕ on the edges (pairs of points) satisfying the following "c-metric" condition: there is a positive constant c such that $\phi^{1/c}$ is a metric. Thus $\phi_{x,y}^{1/c} \leq \phi_{x,z}^{1/c} + \phi_{z,y}^{1/c}$ and consequently also $\phi_{x,y} \leq 2^c \max\{\phi_{x,z}, \phi_{z,y}\}$. With a little more care note that

$$\phi_{x,y} \leq 2^{c-1}(\phi_{x,z} + \phi_{z,y}). \tag{2}$$

Given a collection of points T with weight function w, form a new variable-weights collection T' in the following random process. For each point $u \in T$ let

$$\alpha_u = \frac{s w_u \sum_{v\in T} w_v \phi_{u,v}}{2\phi(T)}. \tag{3}$$

(A satisfactory choice for s is, for example, 4. In practice it may be desirable to adjust this value to optimize the performance of the algorithm.) Observe that $s = \sum_u \alpha_u$. Let $\beta_u = \frac{\alpha_u}{1+\alpha_u}$. To each point u assign an independently chosen random variable K_u with the integral exponential distribution with expectation α_u; namely, $P(K_u = i) = (1-\beta_u)\beta_u^i$. We will also denote this quantity $p_{u,i}$.
Observe for future reference that $\alpha_u = \frac{\beta_u}{1-\beta_u} = \sum_0^\infty i(1-\beta_u)\beta_u^i = \sum_0^\infty i p_{u,i} = \sum_1^\infty \beta^i$. Moreover, the variance of an integer exponential r.v. with expectation α is $\alpha(1+\alpha)$.
Now form the collection T' by assigning weight $w'_u = w_u K_u/\alpha_u$ to each point u of T. We examine the r.v. $\phi(T')$. We begin with its first moment, establishing that it is an estimator of the desired quantity. This relies only on the independence of the random variables $\{K_u\}$ and the weight selection $w'_u = w_u K_u / E(K_u)$.

$$E(\phi(T')) = \frac{1}{2}\sum_{x,y\in T} \phi_{x,y} E(w'_x w'_y) \tag{4}$$

$$= \frac{1}{2}\sum_{x,y\in T} \phi_{x,y} E(w'_x) E(w'_y)$$

$$= \frac{1}{2}\sum_{x,y\in T} \phi_{x,y} w_x w_y = \phi(T)$$

Next we turn to the second moment.

$$E(\phi(T')^2) = \frac{1}{4}\sum_{x,y,z,t\in T} \phi_{x,y}\phi_{z,t}\frac{w_x w_y w_z w_t}{\alpha_x\alpha_y\alpha_z\alpha_t}\times$$

$$\sum_{i,j,k,\ell\geq 0} ijk\ell P((K_x = i)\wedge(K_y = j)\wedge(K_z = k)\wedge(K_t = \ell)).$$

We calculate this by beginning as if the variables K_x, K_y, K_z, K_t are independent even when some of x, y, z, t collide; and then correcting for the effects of collisions. The calculation is involved, and is omitted entirely from this extended abstract. The result is

$$E(\phi(T')^2) = \phi(T)^2 + [\frac{2\phi(T)}{s}]^2 \sum_{y\in T}\alpha_y(1+\alpha_y) + \sum_{x,y\in T}\phi^2_{x,y}\frac{w_x^2 w_y^2}{\beta_x\beta_y} \tag{5}$$

Now we analyze this expression. Beginning with the second term, recall that $\sum \alpha_u = s$, therefore $\sum \alpha_u^2 \leq s^2$, and so

$$[\frac{2\phi(T)}{s}]^2 \sum_{y\in T}\alpha_y(1+\alpha_y) \leq 4\phi(T)^2(1+1/s). \tag{6}$$

Next we examine the third term. Recalling equation (3) and that $\beta_u = \alpha_u/(1+\alpha_u)$, we write

$$\sum_{x,y\in T}\phi^2_{x,y}\frac{w_x^2 w_y^2}{\beta_x\beta_y} = \sum_{x,y\in T}\phi^2_{x,y}\frac{w_x^2 w_y^2(1+\alpha_x)(1+\alpha_y)}{\alpha_x\alpha_y}$$

$$= \frac{4\phi(T)^2}{s^2}\sum_{x,y\in T}\phi^2_{x,y}\frac{w_x w_y(1+\alpha_x)(1+\alpha_y)}{(\sum_z w_z\phi_{x,z})(\sum_z w_z\phi_{y,z})}$$

and recalling $\sum \alpha_u = s$ we have

$$\sum_{x,y\in T}\phi^2_{x,y}\frac{w_x^2 w_y^2}{\beta_x\beta_y} \leq 4\frac{(1+s)^2}{s^2}\phi(T)^2\times$$

$$\sum_{x,y\in T}\phi^2_{x,y}\frac{w_x w_y}{\min\{\sum_z w_z\phi_{x,z}, \sum_z w_z\phi_{y,z}\}\max\{\sum_z w_z\phi_{x,z}, \sum_z w_z\phi_{y,z}\}}$$

Let us lower bound each of the terms in the last denominator. For the min term, recalling inequality 2, consider that for any $x \in T$

$$\phi(T) = \frac{1}{2}\sum_{u,v\in T} w_u w_v \phi_{u,v} \leq \frac{1}{2}\sum_{u,v\in T} w_u w_v 2^{c-1}(\phi_{u,x} + \phi_{x,v})$$

$$= \frac{1}{2}\sum_{u,v\in T} w_u w_v 2^c \phi_{u,x} = 2^{c-1} w_T \sum_{u\in T} w_u\phi_{u,x}$$

Hence

$$\min\{\sum_z w_z\phi_{x,z}, \sum_z w_z\phi_{y,z}\} \geq 2^{1-c}\phi(T)/w_T.$$

For the max term, write (again using inequality 2)

$$\max\{\sum_z w_z\phi_{x,z}, \sum_z w_z\phi_{y,z}\} \geq \frac{1}{2}\sum_z w_z(\phi_{x,z} + \phi_{y,z})$$

$$\geq \frac{1}{2}\sum_z w_z 2^{1-c}\phi_{x,y} = 2^{-c} w_T \phi_{x,y}$$

Now combine the min and max analyses to continue from equation 2:

$$\sum_{x,y\in T}\phi^2_{x,y}\frac{w_x^2 w_y^2}{\beta_x\beta_y} \leq 2^{2+2c}(1+s)^2\phi(T)^2/s^2 \tag{7}$$

Unifying equations 5, 6 and 7 we find that

Theorem 1. *If ϕ is a c-metric then for the random collection T' selected as described above,*

$$E(\phi(T')^2) \leq \phi(T)^2(5 + 4/s + 2^{2+2c}(1+s)^2/s^2)$$

□

For the clustering procedure we will need a slight extension of this statement. Let G be an undirected graph whose vertices are the points of a collection S, and let

$$\phi_G(S) = \sum_{\{u,v\}\in G} w_u w_v \phi_{u,v}. \tag{8}$$

If $S' \subseteq S$ then we let $\phi_G(S')$ equal $\phi_{G'}(S')$ where G' is the induced graph on vertex set S'. By an analysis identical to equation 4, the prescribed sampling procedure on a set T yields

$$E(\phi_G(T')) = \phi_G(T). \tag{9}$$

Moreover, since $0 \leq \phi_G(T') \leq \phi(T')$,

$$E(\phi_G(T')^2) \leq E(\phi(T')^2). \tag{10}$$

We therefore have:

Theorem 2. *If ϕ is a c-metric then for any graph G and for the random collection T' selected as described above,*

$$E(\phi_G(T')^2) \leq \phi(T)^2(5 + 4/s + 2^{2+2c}(1+s)^2/s^2)$$

□

3. THE CASE $\phi = \ell_2^2$

3.1 Preliminaries and related literature

We focus now on the central special case in which the cost function is the square of Euclidean distance $\rho_{u,v}$ between points u, v thus in this section $\phi_{u,v} = \rho^2_{u,v}$.

We say that two sets $S_1, S_2 \subset \mathbb{R}^d$ are *separated* (or *strictly separated*) by sphere C if one of these sets is contained in the interior and the other in the exterior of C (the components of $\mathbb{R}^d - C$, halfspaces labelled arbitrarily in case C is a hyperplane). We say that S_1, S_2 are *weakly separated* by sphere C if one of these sets is disjoint from the interior and the other from the exterior of C. The key to the deterministic algorithm computing the optimal 2-partition or k-partition of a point set, is the following corollary of proposition 7:

Corollary 3. *If $(S, T - S)$ is an optimal partition with respect to $\phi = \ell^2_2$ of a (possibly variable-weight) point set T, then there exists a sphere separating S and $T - S$.*

Let $c(S)$ be the center of gravity of a set of points S, $c(S) = w_S^{-1} \sum_{u \in S} w_u u$. (Here $w_S = \sum_{u \in S} w_u$, or simply $|S|$ in the uniform-weights case.) Let $\mathrm{Var}(S) = w_S^{-1} \sum_{v \in S} w_v \rho^2_{v,c(S)}$. (For brevity we will also, given sets S_1 and S_2, let ρ_{S_1,S_2} denote $\rho_{c(S_1),c(S_2)}$. Thus $\mathrm{Var}(S) = w_S^{-1} \sum_{v \in S} w_v \rho^2_{v,S}$.) Calculation shows that an equivalent formulation of the ℓ^2_2 cost function is $\phi(S) = w^2_S \mathrm{Var}(S)$.

Weak separation was shown previously by Boros and Hammer [8]. The distinction between the kinds of separation was overlooked. Later Inaba, Katoh and Imai [24] proposed examining all sphere partitions to find an optimal partition. That proposal is justified only on the basis of the present work, because weak separation does not imply a sub-exponential time algorithm.

The proper handling of data that is "singular" in the sense that it contains more than $d + 1$ co-spherical points (not necessarily a rarity in integer-coordinate data) has turned out to be an aspect requiring substantial care both in the description of the deterministic exact algorithm, and in its implementation (by the author and by students).

On the other hand, since perturbations of the point locations affect ϕ continuously, this issue can, in the case of approximation algorithms, be circumvented (though there is no need to) by first perturbing the points into general position with respect to spheres.

Based upon corollary 3 we provide a deterministic algorithm computing an optimal 2-partition in section 3.3.

The first mention of ℓ^2_2 as a clustering criterion may be by Kiseleva, Muchnik and Novikov [32], for point sets in one dimension.

3.2 Necessary condition for local optimality

Since points are allowed varying weights, it is natural to allow clusterings in which the weight of a point is allocated among several clusters. However as can easily be verified, to any such clustering there corresponds another of lesser or equal cost which splits no points. (This is true of any cost function, requiring only $\phi_{u,u} = 0$ $\forall u$; and also for criteria such as ψ discussed in section 5.) Hence in the sequel only partitions which assign points to unique clusters will be considered.

Definition 4. *The distance between two k-partitions $\mathcal{S} = \{S_1, ..., S_k\}$ and $\mathcal{R} = \{R_1, ..., R_k\}$ of a set T, is the least number of elements whose memberships must be changed so that $\forall i\ \exists j\ S_i = R_j$. (Equivalently, the least Hamming distance between the vectors in $\{1, ..., k\}^n$ specifying the partitions $\mathcal{S}$ and $\mathcal{R}$, that can be obtained by permutation of the alphabet $\{1, ..., k\}$.)*

Definition 5. *A k-partition $\mathcal{S} = \{S_1, ..., S_k\}$ is j-stable if its cost $\phi(\mathcal{S}) = \sum \phi(S_i)$ is minimal among all k-partitions within distance j.*

If $\mathcal{S}$ is optimal then it is j-stable for any j. If $\mathcal{S}$ is $(n-1)$-stable then it is optimal.

Consider a set R and a point v. Then $\phi(R \cup \{v\}) - \phi(R) = w_v \sum_{u \in R} w_u \rho^2_{u,v}$. This can be rewritten

$$\phi(R \cup \{v\}) - \phi(R) = w_v[w_R^{-1}\phi(R) + w_R \rho^2_{v,R}]. \qquad (11)$$

For, place $c(R)$ at the origin of the coordinate system, and v at the position $\rho_{v,R}$ on the first axis. Then $w_v \sum_{u \in R} w_u \rho^2_{u,v} = w_v \sum_{u \in R} w_u[\sum_2^d u_i^2 + (u_1 - \rho_{v,R})^2] = w_v[\sum_{u \in R} w_u \sum_1^d u_i^2 + w_R \rho^2_{v,R} - 2\rho_{v,R} \sum_u w_u u_1] = w_v[\sum_{u \in R} w_u \rho^2_{u,R} + w_R \rho^2_{v,R}] = w_v[w_R^{-1}\phi(R) + w_R \rho^2_{v,R}]$.

Note that equation 11 is a special case of the more general

$$\phi(R) = \sum_{i,j}[\frac{w_{R_i}\phi(R_j)}{w_{R_j}} + \frac{1}{2} w_{R_i} w_{R_j} \rho^2_{R_i,R_j}]. \qquad (12)$$

expressing the cost of R in terms of constituents $\{R_i\}$ with weights $\{w_{R_i}\}$ and costs $\{\phi(R_i)\}$. (Weights in R are summed in case of repeated points.) This expression in turn can be written in the following way, which will be useful in the sequel:

$$\phi(R) = \sum \phi(R_i) + \sum_{i<j} (\phi(R_i \cup R_j) - \phi(R_i) - \phi(R_j)) \qquad (13)$$

Now consider an existing partition $\{S, \bar{S}\}$ and a new point v. To which cluster is it preferable to adjoin v? Define three regions partitioning space as follows: the region in which it is preferable to adjoin the new point to S, $\eta(S) = \{v \in \mathbb{R}^d : \phi(S \cup \{v\}) - \phi(S) < \phi(\bar{S} \cup \{v\}) - \phi(\bar{S})\}$; the region in which it is preferable to adjoin to $\bar{S}$, $\eta(\bar{S}) = \{v \in \mathbb{R}^d : \phi(S \cup \{v\}) - \phi(S) > \phi(\bar{S} \cup \{v\}) - \phi(\bar{S})\}$; and the boundary between these two regions, where there is a tie, $\nu = \eta(S)^c \cap \eta(\bar{S})^c$. The surface ν is defined by the equation $\phi(S \cup \{v\}) - \phi(S) = \phi(\bar{S} \cup \{v\}) - \phi(\bar{S})$, equivalently $\sum_{u \in S} w_u \rho^2_{u,v} = \sum_{u \in \bar{S}} w_u \rho^2_{u,v}$, equivalently

$$w_S \rho^2_{v,S} + w_S^{-1}\phi(S) = w_{\bar{S}} \rho^2_{v,\bar{S}} + w_{\bar{S}}^{-1}\phi(\bar{S}). \qquad (14)$$

Examination of the last condition shows that ν, if not empty, is a sphere (a hyperplane if $w_S = w_{\bar{S}}$).

Proposition 6. *S is 1-stable if and only if $S \subseteq \eta(S)^c$ and $\bar{S} \subseteq \eta(\bar{S})^c$.*

Proof: If S is 1-stable then, for any point $v \in S$, $\sum_{u \in S - \{v\}} w_u \rho^2_{u,v} \leq \sum_{u \in \bar{S}} w_u \rho^2_{u,v}$. Equivalently, $\sum_{u \in S} w_u \rho^2_{u,v} \leq \sum_{u \in \bar{S}} w_u \rho^2_{u,v}$, implying that $v \in \eta(S)^c$. The argument for $\bar{S}$ is identical. The converse is immediate. □

Proposition 7. *If S is 2-stable then either $S \cap \nu = \emptyset$ or $\bar{S} \cap \nu = \emptyset$.*

Proof: Suppose that $u \in S \cap \nu$ and $v \in \bar{S} \cap \nu$. Now exchange the memberships of these points. The change in ϕ is $\Delta\phi = w_u(\sum_{r \in \bar{S}-\{v\}} w_r \rho^2_{u,r} - \sum_{s \in S} w_s \rho^2_{u,s}) + w_v(\sum_{s \in S-\{u\}} w_s \rho^2_{v,s} - \sum_{r \in \bar{S}} w_r \rho^2_{v,r})$. Since u and v are on ν, each of these terms equals $-w_u w_v \rho^2_{u,v}$. Hence $\Delta\phi = -2 w_u w_v \rho^2_{u,v} < 0$, contradicting 2-stability. □

Let $\Phi_d(n) = \sum_0^d \left(\begin{array}{c} n \\ i \end{array} \right)$.

Proposition 8. *A set of n points in $\mathbb{R}^d$ has at most $\Phi_{d+1}(n-1)$ 2-stable 2-partitions. The clusters of every such partition are separated by a sphere.* Proof: An $O(n^{d+1})$ bound is immediate; further details omitted from this extended abstract. □

The 2-stable partitions are even further restricted. A subset of a poset is termed a j-family if it contains no chains of length $j+1$ [21]. (A 1-family is an antichain.)

Proposition 9. *If $S \subset R$ then $\eta(R) \subset \eta(S)$. Furthermore $\nu(R) \cap \nu(S)$ can contain at most one point.*

Proof: A consequence of the equation $\phi(S \cup \{v\}) - \phi(S) = w_v \sum_{u \in S} w_u \rho^2_{u,v}$. □

Corollary 10. *(a) The collection of sets which occur as clusters in 1-stable partitions of a set T are a 2-family in the poset of subsets of T. (b) The collection of sets $S \cap \eta(S)$ for 1-stable partitions $(S, \bar{S})$ of a set T are an antichain in the poset of subsets of T.* **Proof:** Omitted from this extended abstract. □

3.3 Exact deterministic algorithm for 2-partitions

Theorem 11. *For fixed d an optimal 2-partition in $\mathbb{R}^d$ may be found in $O(n^{d+1})$ time.*

Proof: A time bound of $O(n^{d+2})$ is easily obtained by expending $O(n)$ time computing ϕ for each sphere partition. The description of the method achieving time $O(n^{d+1})$ is omitted from this extended abstract.

3.4 Exact deterministic algorithm for k-partitions

Proposition 12. *To a 1-stable k-partition there corresponds a set of $\left(\begin{array}{c} k \\ 2 \end{array} \right)$ spheres, such that each cluster region is the union of regions defined by intersections of interiors or exteriors of the spheres. If the k-partition is 2-stable then (just as for the 2-partition case) the boundary region between two of these clusters can only contain points belonging to one of them; the same number of spheres therefore also suffice in order to separate the clusters.* □

Let $F(n,d,k) = n^{(d+2)k+1} e^{O((d+2)k)}$.

Proposition 13. *There is an algorithm finding the optimal k-partition of a set of n points which runs in time $F(n,d,k)$.*

Essential use is made here of the "sample points" algorithm of Basu, Pollack and Roy, which produces representative points in the cells defined by a set of polynomials ([5] §3.1.3 p. 1028). The reduction is omitted from this extended abstract.

3.5 Simplification of ℓ_2^2 clustering by dimension reduction

The set T of n points to be clustered may lie in a high dimensional Euclidean space. We need never consider a space of dimension greater than $n-1$: input given in a higher dimensional space should be reduced to this case by projection onto the affine subspace containing T.

Proposition 14. *Fix any $\varepsilon > 0$. Given a set T of n points in $\mathbb{R}^{n-1}$, a k-partition of T can be computed whose cost ϕ is within a factor of $1+\epsilon$ of optimal, in time $n^{O(\varepsilon^{-2}\log n)}$ (for $k=2$) and $F(n, \varepsilon^{-2}\log n, k)$ (for general k).*

Proof: Johnson and Lindenstrauss showed that if a set T of n points in Euclidean space is mapped under a random orthogonal projection M to an $O(\frac{\log n}{\epsilon^2})$-dimensional subspace, then with high probability the distortion of the metric on these points is no more than $1+\epsilon$ [29] (a constant of 8 is achievable in this theorem, with log being the natural logarithm). The distortion is $\max_{a,b,c,d \in T}(\frac{\rho_{Ma,Mb}\rho_{c,d}}{\rho_{a,b}\rho_{Mc,Md}})$. Such a mapping may be found efficiently (in time $\tilde{O}(n^2)$) by trial and error. Once a suitable mapping has been found, 2-partition or k-partition algorithms (deterministic exact, for a guaranteed approximation, or randomized approximate, for a high probability result) can be applied. □

For the objective function ψ to be defined in section (5) one need range in the deterministic algorithm over hyperplane rather than sphere partitions. As shown in [22], improving on the immediate $O(n^{d+1})$, this can be done in time $O(n^d \log n)$ in dimension d. This can be improved:

Note 15. The optimal clustering for ψ can be found in time $O(n^d)$ for $d \geq 2$, by using geometric duality and computing the incidence graph of the arrangement of hyperplanes, in the same manner as in section 3.3.

3.6 Simplification of ℓ_2^2 clustering by sampling

We now show what is the consequence for $\phi = \ell_2^2$ of using the sampling process described in section 2. The cost function ℓ_2^2 is of course a 2-metric in the sense discussed in section 2. Hence theorem 1, with $s \geq 4$ and $c = 2$, implies that

$$E(\phi(T')^2) \leq 106\phi(T)^2 = 106 E(\phi(T'))^2.$$

Fix $s = 4$. Let $D(p;q)$ denote the information divergence or Kullback-Liebler divergence, $D(p;q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$.

Given ε, δ, set $a = 636\varepsilon^{-2}$, $b = (\log(n^{d+1}\delta^{-1}))/D(1/2;2/3)$, and repeat the sampling process described in section (2) $t = ab$ times. Let T_i (for $1 \leq i \leq t$) be the collection of points (with appropriate weights) obtained in trial i.

For a collection of points S and a sphere γ containing no point of S (with γ_1 and γ_2 denoting the two closed regions of space bounded by γ), let $\phi_\gamma(S)$ be the cost of a partition of S by γ, namely $\phi_\gamma(S) = \phi(S \cap \gamma_1) + \phi(S \cap \gamma_2)$; this corresponds to the notation of equation 8 with the understanding that the graph consists of all pairs of points not separated by γ. Let U be the set of points sampled with nonzero weight in any of the sampling processes. For every spherical partition of the set U (represented by a sphere γ passing through no point of U), consider the following quantity:

$$h(\gamma) = \text{median}_{j=1}^{b}\{\frac{1}{a}\sum_{i=1}^{a}\phi_\gamma(T_{a(j-1)+i})\}. \qquad (15)$$

LEMMA 16. *For any given sphere γ, the inequality $|\frac{h(\gamma)-\phi_\gamma(T)}{\phi(T)}| < \varepsilon/2$ holds with probability at least $1-\delta n^{-d-1}$. For the optimal sphere partition γ of U, the inequality $|\frac{h(\gamma)-\phi_\gamma(T)}{\phi_\gamma(T)}| < \varepsilon$ holds with probability at least $1-\delta n^{-d-1}$.*

Proof: For a fixed sphere γ, and any i, the random variable $\phi_\gamma(T_i)$ is an unbiased estimator of $\phi_\gamma(T)$; and, using theorem 2, its variance is at most $106\phi(T)^2$.
Correspondingly, for a fixed sphere γ, and any j, the random variable $M = \frac{1}{a}\sum_{i=1}^{a}\phi_\gamma(T_{a(j-1)+i})$ is an unbiased estimator of $\phi_\gamma(T)$, with variance at most $106\phi(T)^2/a$. Hence using the Chebychev inequality, $P(|\frac{M-\phi_\gamma(T)}{\phi(T)}| > \varepsilon/2) < \frac{4\mathrm{Var}(M)}{\varepsilon^2\phi(T)^2} = \frac{424}{a\varepsilon^2} = 2/3$. By an application of the Chernoff bound this implies the first statement of the lemma. The best sphere partition γ satisfies $\phi_\gamma(T) \le \phi(T)/2$, which implies the second statement of the lemma. □
As noted earlier, the number of distinct sphere partitions of T is bounded by $\Phi_{d+1}(n-1) < n^{d+1}$. We can now conclude that with high probability, the sphere partitions of U are an "ε-approximate" set of representatives for the sphere partitions of T, in the following sense:

THEOREM 17. *With probability at least $1-\delta$: $|\frac{h(\gamma)-\phi_\gamma(T)}{\phi(T)}| < \varepsilon/2$ for all spheres γ, and, for the optimal sphere cut, $|\frac{h(\gamma)-\phi_\gamma(T)}{\phi_\gamma(T)}| < \varepsilon$.* □

3.7 Randomized approximation algorithm for 2-partitions

1. Depending on the dimension d execute either 1A or 1B:

1A. If the dimension is low, $d \in o(\varepsilon^{-2}\log n)$:
Carry out the above sampling procedure $t = 636\varepsilon^{-2}\log(n^{d+1}\delta^{-1})/D(1/2;2/3)$ times, and proceed either to option 2A or 2B below.

1B. If the dimension is high, $d \in \Omega(\varepsilon^{-2}\log n)$:
First carry out the dimension-reduction procedure described in section (3.5), reducing the dimension to $d' = O((\varepsilon/3)^{-2}\log n)$ while distorting all distances by at most $1+\varepsilon/3$. Then we carry out the sampling procedure, again using the parameter $\varepsilon/3$, i.e. setting $t = 636(\varepsilon/3)^{-2}\log(n^{d'+1}\delta^{-1})/D(1/2;2/3)$. (These choices guarantee that the combined allowed error $(1+\varepsilon/3)^2$ is less than $1+\varepsilon$, provided $\varepsilon \le 1$; the case $\varepsilon > 1$ is less interesting.) Then we proceed to either option 2A or 2B below.

2. In the second step carry out either of the following options:

2A. For each sphere partition γ of U, calculate $h(\gamma)$. Select a sphere γ minimizing h and use it to partition T. This partition (which will generally include some arbitrary choices for points near γ) is the output of the algorithm.

2B. For each sphere partition γ of U, evaluate $\phi_\gamma(T)$ (again there will generally be some arbitrary choices for points near γ), and output the partition minimizing this quantity.

COROLLARY 18. *With probability at least $1-\delta$ the value of the cut output by the algorithm (using either option 2A or 2B) is within a multiplicative factor of $1-\varepsilon$ of the optimal value.* □

The output of option 2B is always of course at least as good as that of option 2A, but it necessitates a slightly higher runtime, about $n|U|^{d+1}$; both the improvement in output quality, and the increase in runtime, are fairly slight, so either option seems reasonable. Since option 2B comes with no better guarantees than option 2A, we evaluate the runtime in terms of option 2A.
Run time of the algorithm: linear time suffices to compute the quantities $\{\alpha_u\}_{u\in T}$ required for the sampling procedures. Generation in the simplest way of the r.v.s $\{K_u\}$ used for each of the trials, requires time $O(n\log n)$. However since almost all of these coefficients are likely to equal 0, they can be generated in sublinear expected time (without explicitly listing the zero-valued r.v.s).
Finally, time $|T_i||U|^{d+1}$ suffices to evaluate all spherical partitions γ of U with respect to each of the samples T_i $(1 \le i \le t)$, and so time $|U|^{d+2}$ suffices to compute $h(\gamma)$. $|U|$ is bounded by the sum of the variables K_i, and the expectation of this sum is st. Since the K_i are exponentially distributed and independent, the distribution of their sum has exponential tails, hence the expected runtime of the computation is $O((\varepsilon^{-2}\log(n^{d+1}\delta^{-1}))^{d+2})$. (Alternatively we can allow another probability δ of the algorithm failing, and simply restart it whenever the $|U|$ is too large.) Recall that due to section 3.5, d may be assumed here to be the minimum of $8\varepsilon^{-2}\log n$ and the original dimension.
In conclusion, we have shown (neglecting the cost of the linear algebra that may be required at initialization to isometrically reduce the dimension of the problem to $n-1$):

THEOREM 19. *Given a clustering problem for cost function $\phi = \ell_2^2$ on n points in dimension d, the above algorithm runs in time $O((\varepsilon^{-2}\log(n^{d'+1}\delta^{-1}))^{d'+2})$ (where $d' = \min\{d, 8\varepsilon^{-2}\log n\}$), and with probability at least $1-\delta$ outputs a clustering with a cut cost within a factor of $1-\varepsilon$ of optimum.* □

If we simplify somewhat by assuming ε and δ constant, this gives a runtime of $O(((d+1)\log n)^{d+2})$. In the worst case this is $n^{O(\log\log n)}$. If $d \in o(\frac{\log n}{\log\log n})$ then the runtime of the algorithm is linear, and is dominated by the time to compute the sampling probabilities α_u.

3.8 Few points of a good clustering are mislabelled

Multiplicative approximation of the maximum cut value, obtained above, does not imply multiplicative approximation of the min cluster value; however, it does imply, as we now show, that the minimum cluster has been determined correctly except for a small fraction of misidentified points. This is the main goal of an automated clustering method, since a realistic classification problem will generally be only roughly modeled by a simple criterion such as ϕ, so that there is little reason to think that the decree of the optimum ϕ-clustering, concerning points at the fringes of the clusters, carries much meaning. Nonetheless, multiplicative approximation of the min cluster would be a stronger result and remains, at the least, an outstanding theoretical problem.

Consider a cut $(S, \bar{S})$ of T (we will have in mind that this is the optimal clustering, although this plays no role in the following arguments, only in their application in note 23). Let G denote the graph containing all edges that do not cross this cut. Consider any other clustering $(S', \bar{S}')$ of T; and let G' be the graph containing all edges that do not cross the second cut. Hence $\phi_G(T)$ and $\phi_{G'}(T)$ are the costs of the two clusterings. Define the distance $\Delta(S, S')$ between the two cuts to be $\frac{1}{w_T} \min\{w_{S \cap S'} + w_{\bar{S} \cap \bar{S}'}, w_{S \cap \bar{S}'} + w_{\bar{S} \cap S'}\}$. In case the points have unit weights this is the same (after scaling) as the Hamming-type distance of definition 4.

Lemma 20. $\phi(T) \leq (9 + \frac{4}{\Delta(S,S')})(\phi_G(T) + \phi_{G'}(T))$.

This bound is not far from the truth; it is easy to construct an example with $\phi_G(T) = 0$ and $\phi(T) = \frac{1}{\Delta(S,S')}\phi_{G'}(T)$.
The lemma immediately implies:

Theorem 21. *Let a clustering $(S, \bar{S})$ be given (with corresponding graph G). Let c be any positive number. Let $0 < \varepsilon \leq \frac{1}{17(1+c)}$ and suppose that $\phi_G(T) \leq \varepsilon\phi(T)$. Let $(S', \bar{S}')$ be another clustering (with corresponding graph G'), such that $\Delta(S, S') \geq \frac{4(1+c)\varepsilon}{1-9(1+c)\varepsilon}$, i.e. the two clusterings differ in the labelling of a substantial part of the data. Then $\phi_{G'}(T) \geq c\varepsilon\phi(T)$.*

So $(S', \bar{S}')$ is a worse clustering than $(S, \bar{S})$, by a factor of at least c.
To simplify the above theorem, in the particular case $c = 1$ we can for example say:

Corollary 22. *If $\phi_G(T) \leq \varepsilon\phi(T) \leq \frac{1}{34}\phi(T)$ and $\Delta(S, S') \geq 17\varepsilon$, then $\phi_{G'}(T) \geq \varepsilon\phi(T)$.*

Theorem 21 and this corollary are a precise formulation, for ϕ, of the intuitive statement that should hold for any useful clustering criterion: that if the data set can be clustered very well, then that clustering must be "meaningful" or "nearly unique" — there cannot be an entirely different way of achieving a clustering of similar quality.

Note 23. The algorithmic implication is as follows: the algorithm of the preceding section can be run to identify a $(1-\varepsilon)$ multiplicative approximation of max cut. If the value of that cut is less than $(1-\varepsilon)\phi(T)$, then we can also obtain a multiplicative approximation of the min cluster by using ε^2 in place of ε in the algorithm. On the other hand if it is found that the max cut value is at least $(1-\varepsilon)\phi(T)$, then we can conclude that the clustering S' so identified is very close to the optimal clustering S, specifically $\Delta(S, S') < 17\varepsilon$.

Proof of lemma 20: We begin with a triangle inequality concerning ϕ. Given two weighted collections of points A and B, let $r_{AB} = [\frac{\phi(A \cup B) - \phi(A) - \phi(B)}{w_A w_B}]^{1/2}$. From equation 12 we read: $r^2_{AB} = \rho^2_{A,B} + \frac{\phi(A)}{w^2_A} + \frac{\phi(B)}{w^2_B}$. Now, we show the triangle inequality: $r_{AB} + r_{BC} \geq r_{AC}$. (Argument omitted from this extended abstract.) This does not make r a metric on clusters, because $r_{AA} > 0$ (unless A is a point).
We conclude also that

$$r^2_{AB} + r^2_{BC} \geq r^2_{AC}/2. \tag{16}$$

Now, consider the four subcollections defined by the cuts S and S': $A = S \cap S'$, $B = S \cap \bar{S}'$, $C = \bar{S} \cap S'$, and $D = \bar{S} \cap \bar{S}'$. In these terms, $\phi_G(T) = \phi(A \cup B) + \phi(C \cup D)$, $\phi_{G'}(T) = \phi(A \cup C) + \phi(B \cup D)$, and $\Delta(S, S') = \frac{1}{w_T} \min\{w_A + w_D, w_B + w_C\}$.
We assume without loss of generality that $\Delta(S, S') = \frac{w_B + w_C}{w_T}$.
Let $E \in \{B, C\}$ be the heavier of the two, so that $w_E = \max\{w_B, w_C\} \geq \Delta(S, S')w_T/2$. Similarly let $F \in \{A, D\}$ be the heavier of the two, so that $w_F = \max\{w_A, w_D\} \geq w_T/4$.
Employ equation 13 to write: $\phi(T) = \phi(A) + \phi(B) + \phi(C) + \phi(D) + (\phi(A \cup B) - \phi(A) - \phi(B)) + (\phi(C \cup D) - \phi(C) - \phi(D)) + (\phi(A \cup C) - \phi(A) - \phi(C)) + (\phi(B \cup D) - \phi(B) - \phi(D)) + (\phi(A \cup D) - \phi(A) - \phi(D)) + (\phi(B \cup C) - \phi(B) - \phi(C))$
$\leq \phi(A \cup B) + \phi(C \cup D) + \phi(A \cup C) + \phi(B \cup D) + (\phi(A \cup D) - \phi(A) - \phi(D)) + (\phi(B \cup C) - \phi(B) - \phi(C))$
$= \phi(A \cup B) + \phi(C \cup D) + \phi(A \cup C) + \phi(B \cup D) + w_A w_D r^2_{AD} + w_B w_C r^2_{BC}$.
This, by inequality 16, is
$\leq \phi(A \cup B) + \phi(C \cup D) + \phi(A \cup C) + \phi(B \cup D) + 2w_A w_D(r^2_{AE} + r^2_{ED}) + 2w_B w_C(r^2_{BF} + r^2_{FC})$
$= \phi(A \cup B) + \phi(C \cup D) + \phi(A \cup C) + \phi(B \cup D)$
$+ 2w_A w_D(\frac{\phi(A \cup E) - \phi(A) - \phi(E)}{w_A w_E} + \frac{\phi(E \cup D) - \phi(E) - \phi(D)}{w_E w_D})$
$+ 2w_B w_C(\frac{\phi(B \cup F) - \phi(B) - \phi(F)}{w_B w_F} + \frac{\phi(F \cup C) - \phi(F) - \phi(C)}{w_F w_C}))$
$\leq \phi(A \cup B) + \phi(C \cup D) + \phi(A \cup C) + \phi(B \cup D)$
$+ \frac{2w_D}{w_E}\phi(A \cup E) + \frac{2w_A}{w_E}\phi(E \cup D) + \frac{2w_C}{w_F}\phi(B \cup F) + \frac{2w_B}{w_F}\phi(F \cup C)$.
Every argument of ϕ in the last line is the same as one of those in the preceding line. Both $\frac{2w_D}{w_E}$ and $\frac{2w_A}{w_E}$ are bounded above by $\frac{2w_T}{w_E} \leq \frac{4}{\Delta(S,S')}$, while both $\frac{2w_C}{w_F}$ and $\frac{2w_B}{w_F}$ are bounded above by $\frac{2w_T}{w_F} \leq 8$. Hence

$$\begin{aligned}\phi(T) &\leq (9 + \frac{4}{\Delta(S, S')}) \times \\ &\quad (\phi(A \cup B) + \phi(C \cup D) + \phi(A \cup C) + \phi(B \cup D)) \\ &\leq (9 + \frac{4}{\Delta(S, S')})(\phi_G(T) + \phi_{G'}(T)).\end{aligned}$$

□

3.9 Randomized approximation algorithm for k-partitions

Similar to the case $k = 2$; omitted from the extended abstract.

4. THE CASES $\phi = \ell_1, \ell_2$, AND OTHER COST FUNCTIONS

Having obtained a clustering algorithm for ℓ^2_2, we are positioned to take advantage of the generosity of ℓ^2_2 as a host space.
By a cost function on a set of points T we mean a function $\lambda : T^2 \to \mathbb{R}$ which is symmetric, nonnegative, and 0 on the diagonal. An embedding of (T, λ) in (T', λ') with distortion C is a map $\iota : T \to T'$ such that $\sup_{a,b,c,d \in T}(\frac{\lambda'(\iota(a),\iota(b))\lambda(c,d)}{\lambda(a,b)\lambda'(\iota(c),\iota(d))}) = C$. If $C = 1$ we say ι is isometric (regardless of whether the domain and range are metric spaces). We will abbreviate by writing simply ℓ_1, ℓ_2 or ℓ^2_2 when all that matters is that the dimension of the space is finite. Note that for ℓ_2 and ℓ^2_2 no dimension beyond $n - 1$ need be considered. A space is finite if T is finite.

Theorem 24. **[Linial, London and Rabinovich] (a)** *There is an algorithm which, given a cost function λ on a set of n points T, identifies a minimum-distortion embedding of (T, λ) in ℓ^2_2, in time polyomial in n.* **(b)** *Every finite ℓ_1 or ℓ_2 space is isometrically embeddable in ℓ^2_2.*

Hence our approximation scheme solves also the cases $\phi = \ell_1$ and $\phi = \ell_2$ in the same asymptotic runtime (i.e. $n^{O(\log \log n)}$ for fixed ε, δ and k) guaranteed for the case ℓ_2^2. Note that this does not supply a way of taking advantage of an initially low dimension to obtain an improved runtime. Finally note that whether or not a given cost function is known to be ℓ_2^2-embeddable, one may solve the PSD program, and provide a good clustering opportunistically if a low distortion embedding exists.

5. DISCUSSION AND OTHER OBJECTIVE FUNCTIONS FOR CLUSTERING

Some interesting objective functions for clustering do not fall within the framework discussed in this paper, of the sum of a cost function over all intra-cluster pairs of points.

One such criterion which has attracted considerable attention is $\psi(S) = |S|\mathrm{Var}(S) = \sum_{v \in S} \rho_{v,S}^2 = \phi(S)/|S|$ (in this section we let $\phi = \ell_2^2$). Clustering so as to minimize $\sum \psi(S_i)$ (sometimes known as "sum of squares minimization") appears to have been discussed first for one dimension by Fisher in 1958 [17] and for higher dimensions by Ward in 1963 [48] and Shlezinger in 1965 [45]; a partial list of subsequent literature is [16; 27; 20; 35; 43; 6; 30; 11; 24; 22; 15], and some surveys touching on the subject are [19; 38]. (Point weights appear not to have been discussed in this literature, but can be accomodated without harm to any existing result.) The regions containing optimal clusters relative to the ψ criterion are Voronoi cells centered on the centers of gravity of the clusters; the ϕ and ψ criteria generally lead to quite different kinds of optimum partitions. Which criterion, if either, is preferable will depend on the application domain. The restriction that optimal regions for the ψ criterion must be convex can be regarded as either an advantage or a limitation of the criterion. The full version of this paper contains some examples which may help clarify the relative advantages of the two criteria.

The earliest related discussion we are aware of is by Neyman; it appears that his proposed clustering criterion corresponds to the function $\sum_{v \in S} \rho_{v,S}$ [39]. We are not aware of any existing algorithmic work specifically concerning this criterion. However, just as for any criterion of the form $\sum_{v \in S} g(\rho_{v,S})$ (g monotone increasing), optimal cluster regions must be Voronoi cells, so an exhaustive examination of such partitions will find an optimal partition in time $O(n^{d+1})$.

The Johnson-Lindenstrauss dimension-reduction step is useful for both of the above objective functions, since, for the same reasons described earlier in the paper, it reduces the effective dimension of a $1 \pm \varepsilon$ approximation problem to $O(\varepsilon^{-2} \log n)$.

Acknowledgments

Thanks to V. Vazirani, N. Linial, A. Wigderson, A. Magen, I. Muchnik, C. Kenyon, P. Indyk, U. Vazirani, D. Klain and S. Basu for very helpful discussions.

6. REFERENCES

[1] P. Arabie, L. J. Hubert, and G. De Soete, editors. *Clustering and Classification.* World Scientific, 1996.

[2] S. Arora, D. Karger, and M. Karpinski. Polynomial time approximation schemes for dense instances of NP-hard problems. In *27th Annual ACM Symposium on the Theory of Computing*, pages 284–293, Las Vegas, 1995.

[3] S. Arora, P. Raghavan, and S. Rao. Polynomial time approximation schemes for euclidean k-medians and related problems. In *Proc. 30'th Annual ACM Symposium on Theory of Computing*, 1998.

[4] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *Proc. 37'th Ann. Symp. on Foundations of Computer Science*, pages 184–193, 1996.

[5] S. Basu, R. Pollack, and M.-F. Roy. On the combinatorial and algebraic complexity of quantifier elimination. *J. ACM*, 43(6):1002–1045, 1996.

[6] J. P. Benzécri. Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Dannées*, VII(2):209–218, 1982.

[7] M. Bern and D. Eppstein. Approximation algorithms for geometric problems. In D. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, pages 296–345. PWS Publishing, 1996.

[8] E. Boros and P. L. Hammer. On clustering problems with connected optima in Euclidean spaces. *Discrete Mathematics*, 75:81–88, 1989.

[9] M. Charikar, S. Guha, E. Tardos, and D.B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the 31'st Annual ACM Symposium on Theory of Computing*, 1999.

[10] R. M. Cormack. A review of classification. *J. Roy. Stat. Soc. A*, 134:321–367, 1971.

[11] H. E. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1:7–24, 1984.

[12] W. Fernandez de la Vega. Max-cut has a randomized approximation scheme in dense graphs. *Random Structures and Algorithms*, 8(3):187–198, 1996.

[13] W. Fernandez de la Vega and M. Karpinski. Polynomial time approximation of dense weighted instances of max-cut. manuscript.

[14] W. Fernandez de la Vega and C. Kenyon. A randomized approximation scheme for metric max-cut. In *39'th Ann. Symp. on Foundations of Computer Science*, pages 468–471. IEEE, 1998.

[15] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[16] A. W. F. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.

[17] W. D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53:789–798, 1958.

[18] R. M. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theor. Comput. Sci.*, 1:237–267, 1976.

[19] A. D. Gordon. *Classification.* Chapman and Hall, 1981.

[20] J. C. Gower. A comparison of some methods of cluster analysis. *Biometrics*, 23:623–637, 1967.

[21] C. Greene and D. J. Kleitman. Proof techniques in the theory of finite sets. In G.-C. Rota, editor, *Studies in Combinatorics.* The Mathematical Association of America, 1978.

[22] P. Hansen, B. Jaumard, and N. Mladenović. Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, 15:37–55, 1998.

[23] J. A. Hartigan. *Clustering Algorithms.* Wiley, 1975.

[24] M. Inaba, N. Katoh, and H. Imai. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In *Proc. 10'th ACM Symp. Comp. Geom.*, pages 332–339, 1994.

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

[26] K. Jain and V. V. Vazirani. Primal-dual approximation algorithms for metric facility location and k-median problems. In *Proc. 40'th Ann. Symp. on Foundations of Computer Science*, 1999.

[27] R. C. Jancey. Multidimensional group analysis. *Australian Journal of Botany*, 14:127–130, 1966.

[28] N. Jardine and R. Sibson. *Mathematical Taxonomy.* Wiley, 1971.

[29] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984.

[30] J. Juan. Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Dannées*, VII(2):229–225, 1982.

[31] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[32] N. E. Kiseleva, I. B. Muchnik, and S. G. Novikov. Stratified samples in the problem of representative sampling. *Automation and Remote Control*, 47(5):684–693, 1986.

[33] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. 9'th ACM-SIAM Symp. on Discr. Alg.*, 1998.

[34] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. In *Proc. 35'th Annual Symposium on Foundations of Computer Science*, pages 577–591. IEEE Press, 1994.

[35] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the 5'th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. U. California Press, 1967.

[36] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, 1988.

[37] B. Mirkin. *Mathematical Classification and Clustering.* Kluwer, 1996.

[38] B. G. Mirkin and I. Muchnik. Clustering and multidimensional scaling in Russia (1960-1990): a review. In P. Arabie, L. J. Hubert, and G. De Soete, editors, *Clustering and Classification*, pages 295–339. World Scientific, 1996.

[39] J. Neyman. On the two different aspects of the representative model: the method of stratified sampling and the method of purposive selection. *J. R. Statis. Soc.*, 97:558–606, 1934.

[40] D. Pollard. Quantization and the method of k-means. *IEEE Trans. Inform. Theory*, IT-28:199–205, March 1982.

[41] M. J. Sabin and R.M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Trans. Inform. Theory*, IT-32(2):148–155, 1986.

[42] S. Sahni and T. Gonzales. P-complete problems and approximate solutions. In *15th Annual Symposium on Switching and Automata Theory*, pages 28–32. IEEE, 1974.

[43] A. J. Scott and M. J. Symons. On the Edwards and Cavalli-Sforza method of cluster analysis. *Biometrics*, 27:217–219, 1971.

[44] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[45] M. I. Shlezinger. On unsupervised pattern recognition. In V. M. Glushkov, editor, *Reading Automata*, pages 62–70. Naukova Dumka, 1965.

[46] R. R. Sokal and P. H. A. Sneath. *Principles of Numerical Taxonomy.* Freeman, 1963.

[47] J. van Ryzin, editor. *Classification and Clustering.* Academic Press, 1977.

[48] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.