Clustering with Outliers and Generalizations

Samir Khuller

University of Maryland

College Park, Maryland

UNIVERSITY OF MARYLAND

Clustering Problems

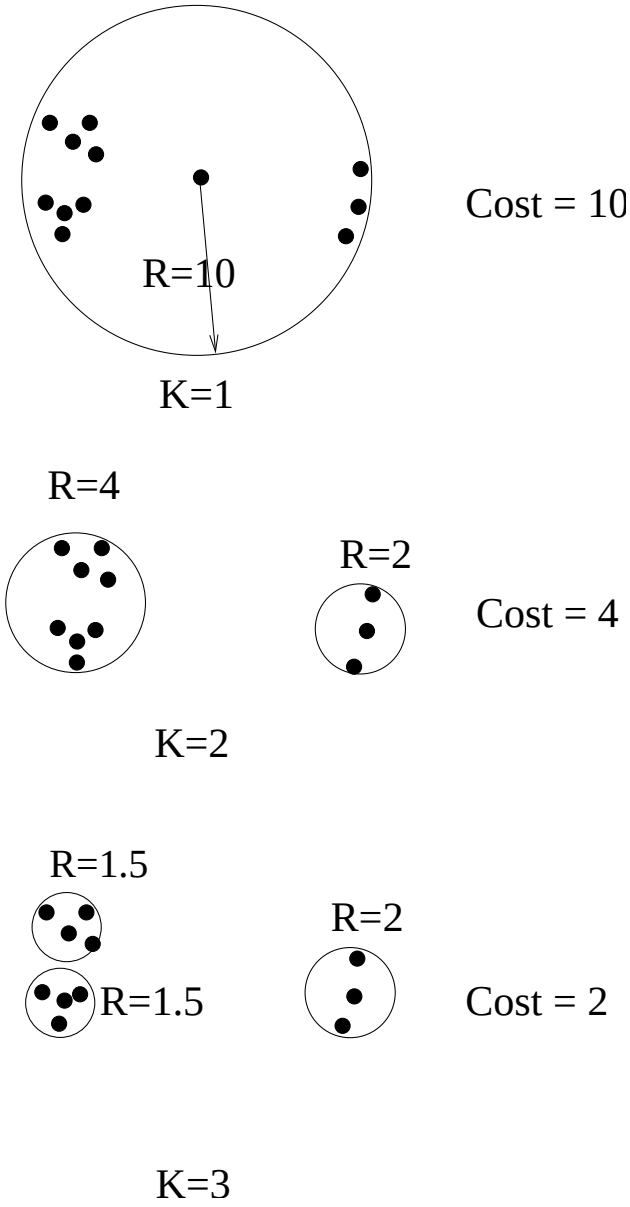
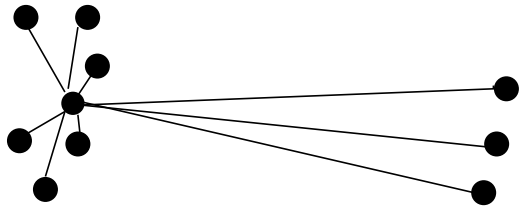


Figure 1: K-Center Clustering

Clustering Problems



K=1

K=2

Figure 2: K-Median Clustering

The K-Center problem

Select locations for K fire stations so that no house is too far from its nearest fire station.

Formally: Given a graph $G = (V, E)$ and integer K, find a subset S ($|S| \leq K$) of centers that minimizes the following:

$$\text{Radius } R = \max_{u \in V} \min_{v \in S} d(u, v).$$

- NP-Hard — $(2 - \epsilon)$-approximation also NP-Hard (reduction from Dominating Set).
- 2-approximable (Gonzalez (85), Hochbaum-Shmoys (85)).
- Can also be extended to weighted K-centers.

$$\text{Radius } R = \max_{u \in V} \min_{v \in S} w(u) \cdot d(u, v).$$

Observations

Radius R^* of OPT must be the distance between a pair of nodes in the graph (when $S \subset V$).

$\Longrightarrow$ "Guess" each possible value for R^*.

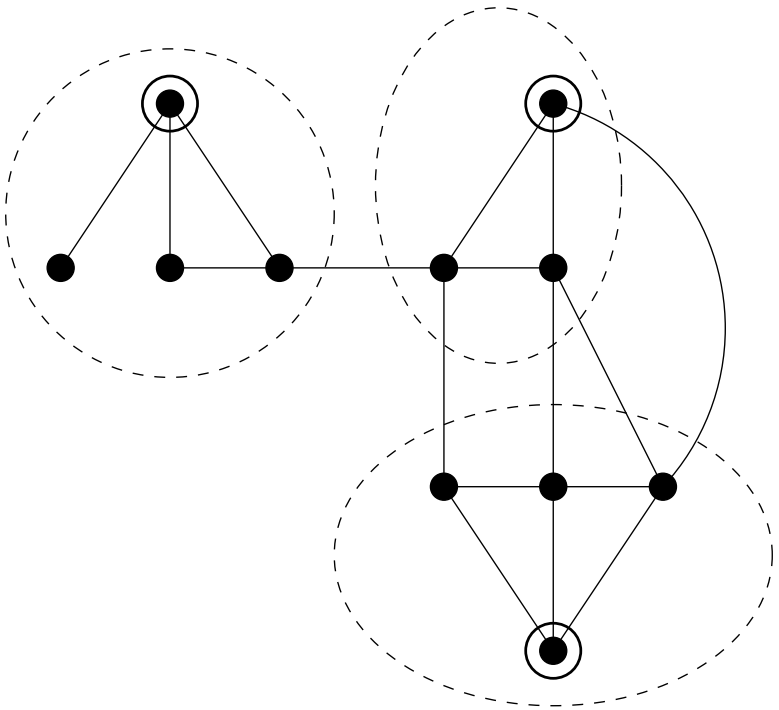
(At most ($O(n^2)$.)

Definition 1 *G_δ is the unweighted graph with all the nodes of G and edges (x, y) such that $d(x, y) \leq \delta$.*

Goal

$\delta = 5$

G_δ:



K=3

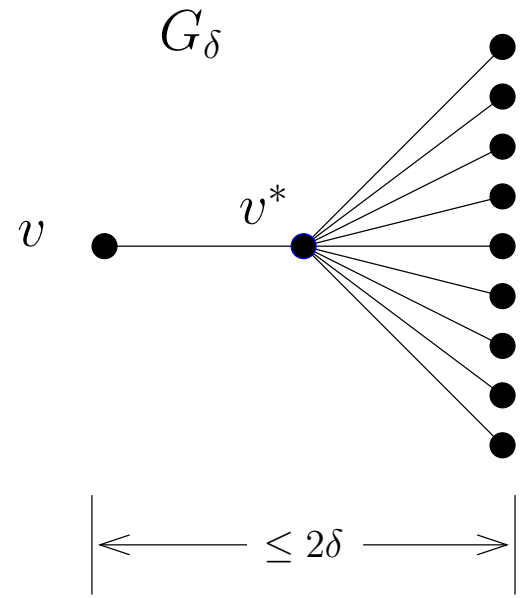
Assume solution of radius δ exists.

Goal: find a solution with radius at most $c \cdot \delta$ *using at most* K *centers*.

Intuition

If we select v as a center, and v is covered in OPT by node v^* within radius δ, then v covers all nodes covered by v^* within distance 2δ.

Pick an uncovered node v as a center. Mark all nodes within 2 hops in G_δ of v as covered. Repeat.

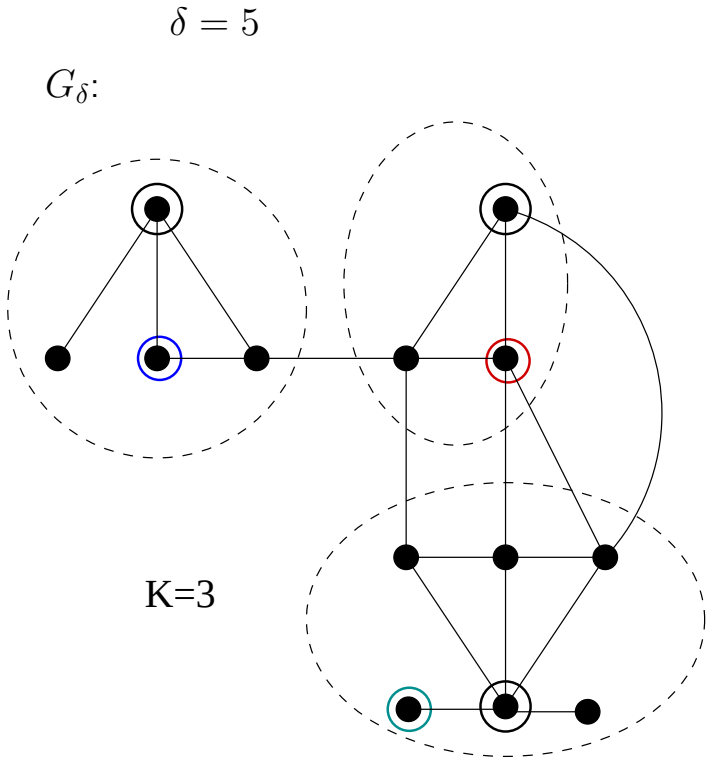


Algorithm

Try increasing values of δ.

Find a MIS S in G_δ^2.

If $|S| \leq K$ then S is the solution.



Proof

Distance of each node from a node in S is at most 2δ.

At the correct radius, the algorithm must succeed, since G_δ^2 cannot have any MIS $> |S|$.

If R_i is the smallest radius for which the algorithm succeeds, then $R_i \leq \delta^*$. Our cost is at most $2R_i$.

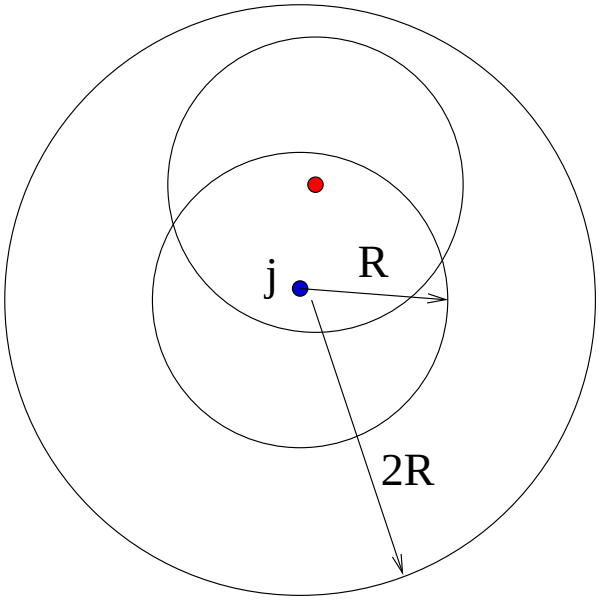


Figure 3: Hochbaum-Shmoys Method

Generalizations

1. (Capacities) Each center has an upper bound of L points that can be assigned to it. Parameters: K, L.

2. (Outliers) Cluster at least p points ($\leq n$) into one of K clusters. Parameters: K, p. Or we can assume we are allowed z points to be dropped as outliers.

3. (Anonymity) Each cluster should have at least r points in it. Parameters: K, r. Problem is hard even if K is unrestricted!
   r-Gather problem: Unbounded K.

We can also study these problems in the streaming model in which the main restriction is that the volume of data is extremely large and we can only use very limited amount of memory and are permitted a single scan on the data.

Capacties on Cluster Sizes

(Bar-Ilan, Kortsarz, Peleg (93)) Develop a factor 10 approximation for the capacitated K-center problem.

(Khuller, Sussmann (96)) Improve to factor 5 approximation.

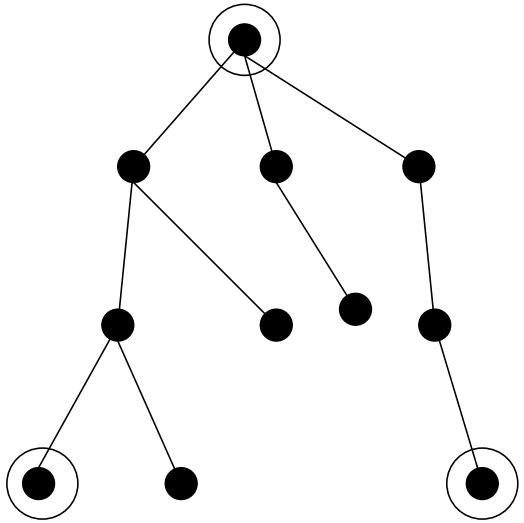
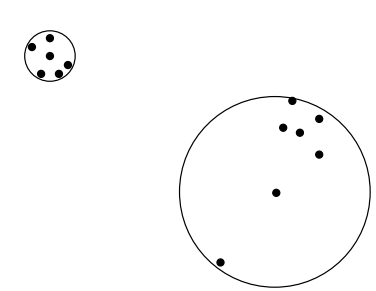


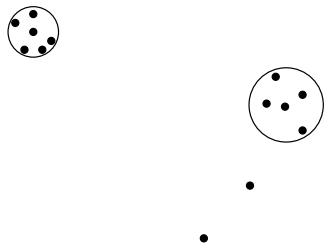
Figure 4: Tree of Centers

Uses BFS to build a “tree” of centers, and then uses network flow for coming up with a good lower bound on the optimal solution. Easy to get a bound of 7. More work to improve that.

Outliers



2-center solution (K=2)



2-center robust solution (K=2, p=11)

Figure 5: We are only required to cluster p points.

Why is the outlier version tricky?

- Hochbaum-Shmoys method inherently assumes that *each* point is covered. This is not true anymore.
- Main difficulty comes from “fragmentation” of an OPT clustering.
- In the asymmetric distance function case, NO approximation is possible.

Outliers

(Charikar, Khuller, Mount, Narasimhan (01)) There is a factor 3 approximation for the K-center problem with outliers.

We also prove a $3 - \epsilon$ hardness for any $\epsilon > 0$ for the problem when some locations are forbidden.

The upper bound works even with forbidden locations!

Observations

Suppose we know the optimal solution radius (R) (try them all!).

For each point $v_i \in V$, let D_i (E_i, resp.) denote the set of points that are within distance R ($3R$, resp.) from v_i. D_i are *disks* of radius R and the sets E_i are the corresponding *expanded disks* of radius $3R$. *Size* of a disk (or expanded disk) is its cardinality.

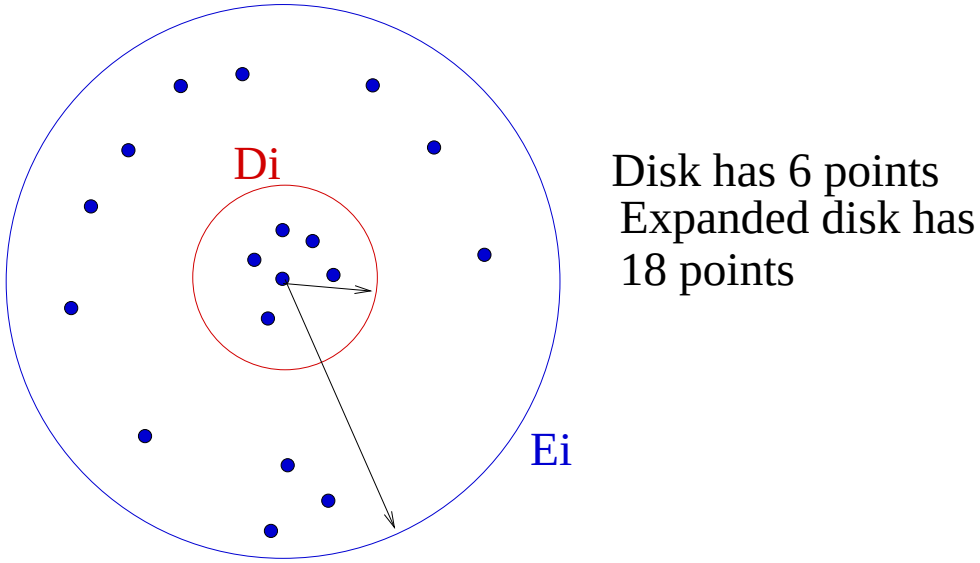


Figure 6: Disks and Expanded Disks.

New Algorithm (Outlier K-centers)

1. Initially all points are uncovered.
2. Construct all disks and corresponding expanded disks.
3. Repeat the following K times:
   - Let D_j be the disk containing the most uncovered points.
   - Mark as covered all points in the corresponding expanded disk E_j after placing facility at j.
   - Update all the disks and expanded disks (i.e., remove covered points).
4. If at least p points of V are marked as covered, then answer YES, else answer NO.

Bad Example

The algorithm fails if we greedily pick the heaviest expanded disk instead!

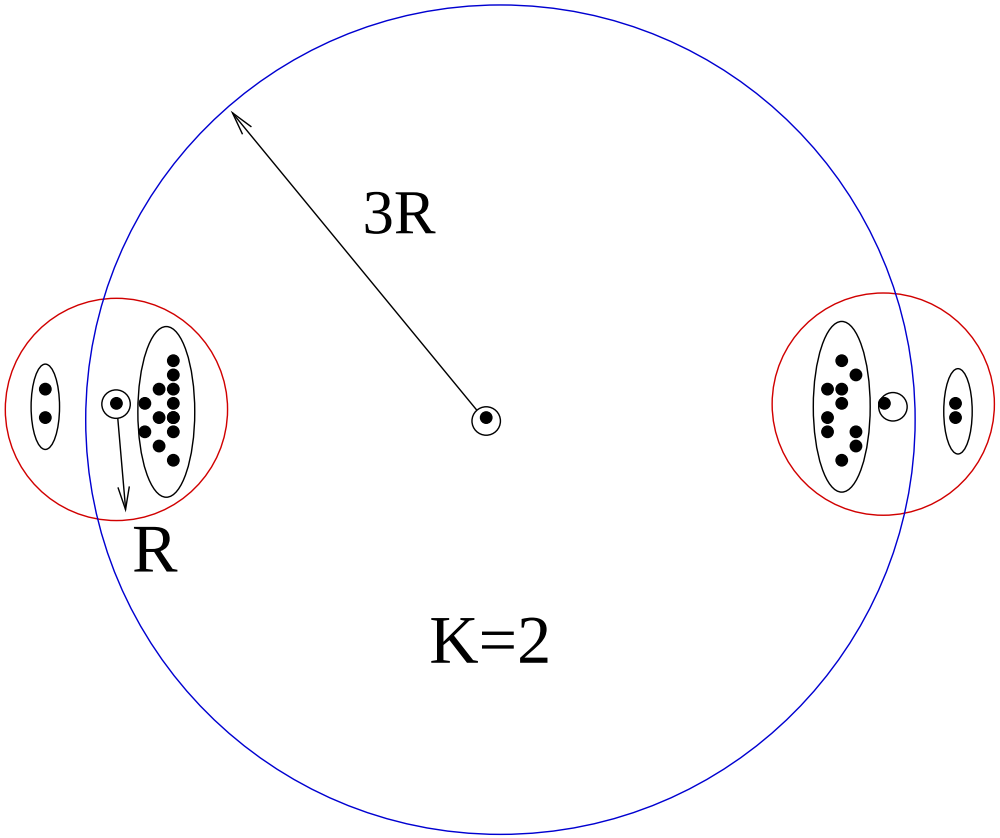


Figure 7: Bad example for choosing based on E_j.

Proof Idea

Let the sets of points covered by the OPTIMAL solution be $O_1, \ldots, O_K$.

The key observation is that if we ever pick a set D_j that covers a point in some O_i, then E_j covers all points in O_i.

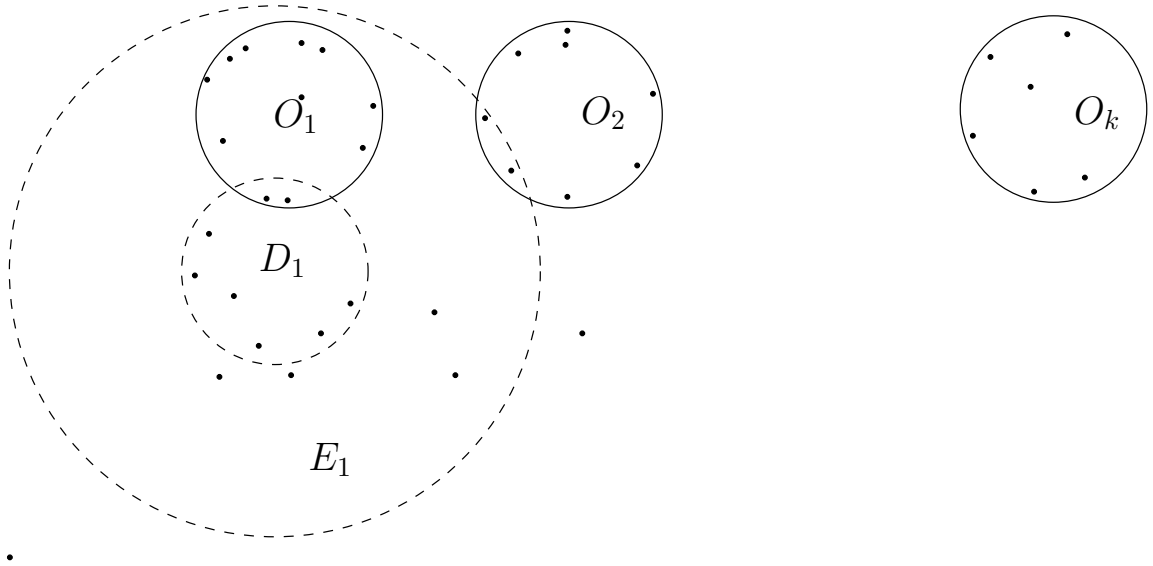


Figure 8: Optimal Clusters and the Greedy Step

Proof Idea

Theorem 1 *With radius R if there exists a placement of K centers that covers p customers, then the algorithm finds a placement of K centers that with a radius of $3R$ cover at least p customers.*

$$|E_1| \geq |O_1| + \sum_{i=2}^{k} |E_1 \cap O_i|. \tag{1}$$

Consider the $(k-1)$-center problem on the set $S - E_1$. We choose $E_2, E_3, \ldots, E_k$. For $S - E_1$, it is clear that $O_2 - E_1, O_3 - E_1, \ldots, O_k - E_1$ is a solution, although not an optimal one. By induction, we know that

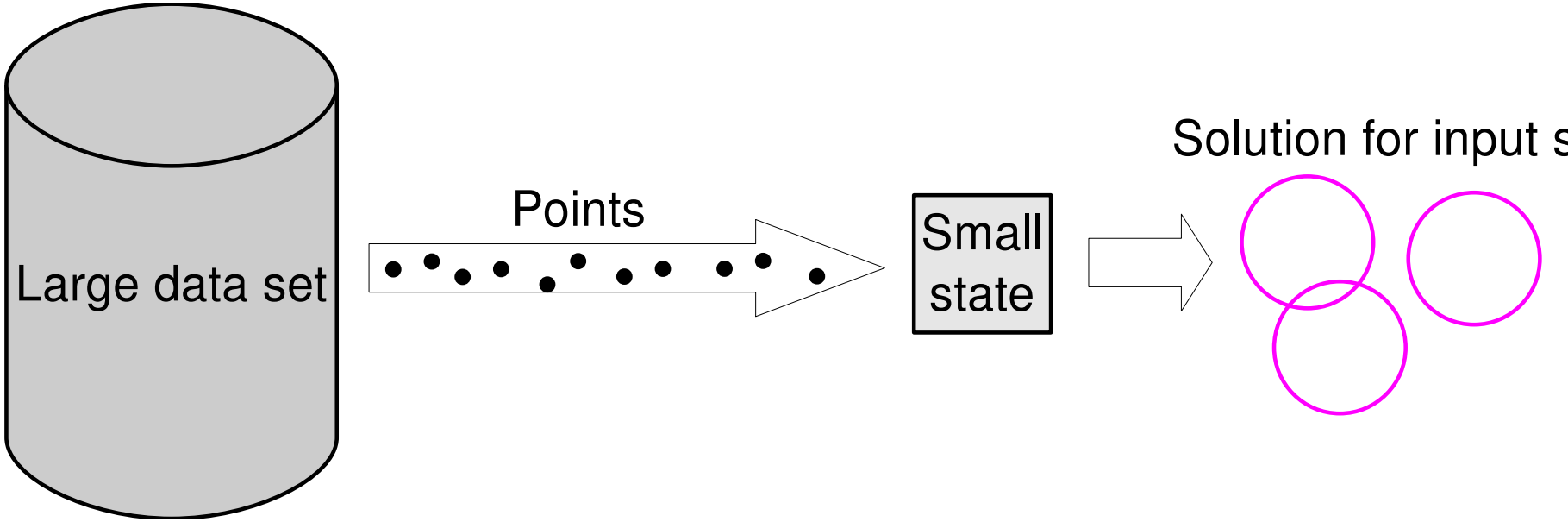
$$|E_2 \cup \ldots \cup E_k| \geq |\bigcup_{i=2}^{k} (O_i - E_1)| \tag{2}$$

Adding gives the result.

Streaming Model of Computation

Points presented one at a time. We have only $O(K)$ memory to store a subset of points.

Goal: Retain a subset S of points, so that every input point is within cR of a point in S, where R is the radius of the optimal K-center clustering.



Seems a bit hard at first glance!

K-Center Streaming

Factor 8 approximation (Charikar, Chekuri, Feder, Motwani (97))

Algorithm maintains a lower bound on the radius r (initially 0), and selects the first K points as S.

Let $S = \{c_1, c_2, \ldots c_\ell\}$. When a new point i arrives.

- If $d(i, c_j) \leq 8r$, assign i to c_j. Exit.
- Let $c_{\ell+1} = i$ add it to S.
- If $\ell = K$ then
  1. Let $r \leftarrow \frac{t}{2}$, where t = smallest distance between a pair of points in S.
  2. Pick a point from S and let it be the new center c'_1. Remove all points from S that are within $4r$ from c'_1. All removed clusters are merged into c'_1.
  3. Repeat step (2) until S is empty. This is the new S.

Streaming

Let $r_1, r_2, \ldots$ be the sequence of values of r.

Observation: The minimum pairwise distance between points in S is $\geq 4r_i$. Thus $t \geq 4r_i$ and thus $r_{i+1} \geq 2r_i$.

This also implies that we do reduce the size of S, certainly the closest pair are at most $t = 2r_{i+1}$ apart.

Proof

The distance of a point from its center was at most $8r_i$. This center may be merged with another cluster center at distance at most $4r_{i+1}$. The radius may go up to $8r_i + 4r_{i+1} \leq 8r_{i+1}$!

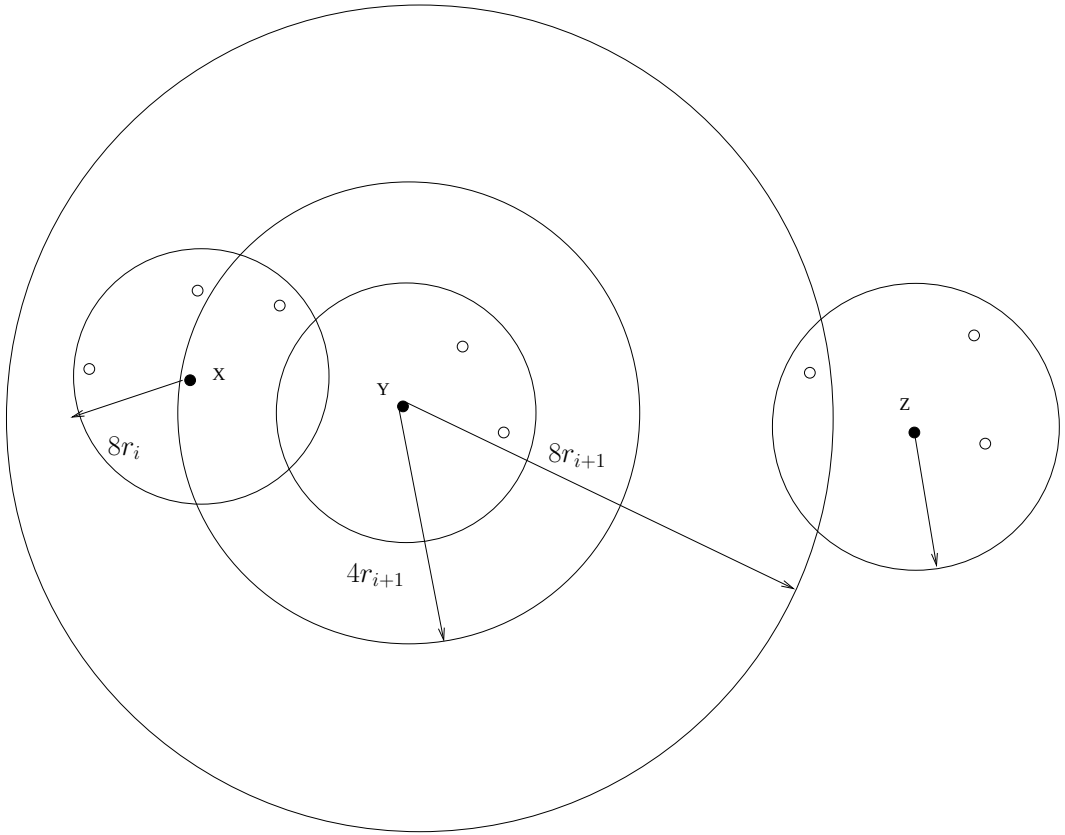


Figure 9: Streaming Model Clustering Merging

Streaming

Using “parallel” copies of the algorithm we can improve the bound to $(2+\epsilon)$ (Guha (07), McCutchen & Khuller (08)). This still uses $O(K)$ memory but with a dependence on ϵ.

New: Guha (09) shows that we cannot get rid of the $+\epsilon$ factor.

Streaming with Outliers (McCutchen Khuller (08))

Recall that we assume that there are z outliers. Using $O(Kz)$ memory we can get a $4+\epsilon$ approximation via complex generalization of the doubling algorithm.

Previous random sampling method due to Charikar, O'Callaghan, Panigrahy leaves out $O(z)$ outliers. Their memory usage is very high when z is small.

Streaming with Outliers

New result: Using $O(K+z)$ memory we can get a $14+\epsilon$ approximation.

Uses a simple shifting idea, that can get constant approximation even for the anonymity application.

Streaming with Outliers

Using $O(K + z)$ memory we can get a $14 + \epsilon$ approximation.

CENTRAL IDEA: As each point is assigned to a cluster center, we keep track of a count of the number of points that were assigned to the cluster center. These counts are maintained when the clusters are merged. We use $K + z$ clusters.

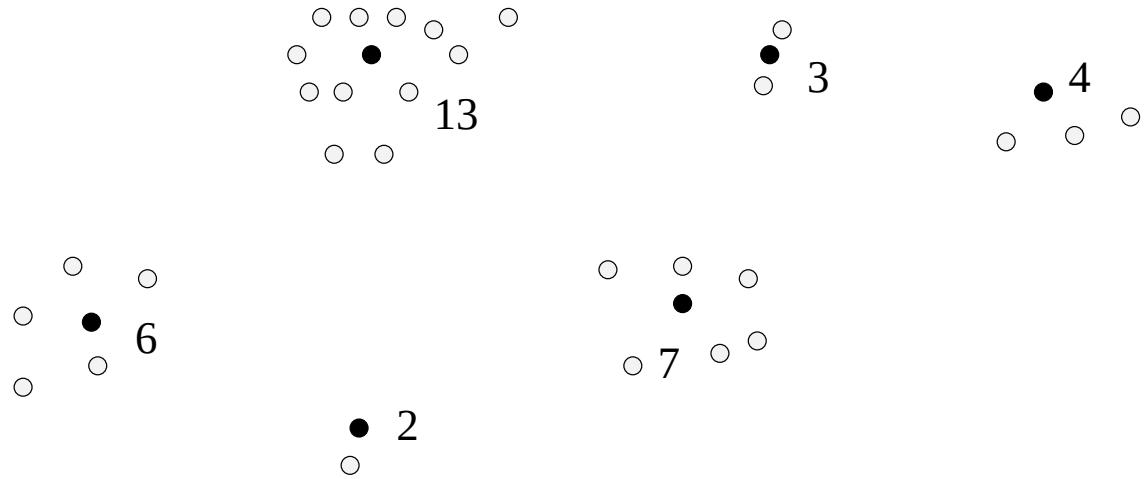
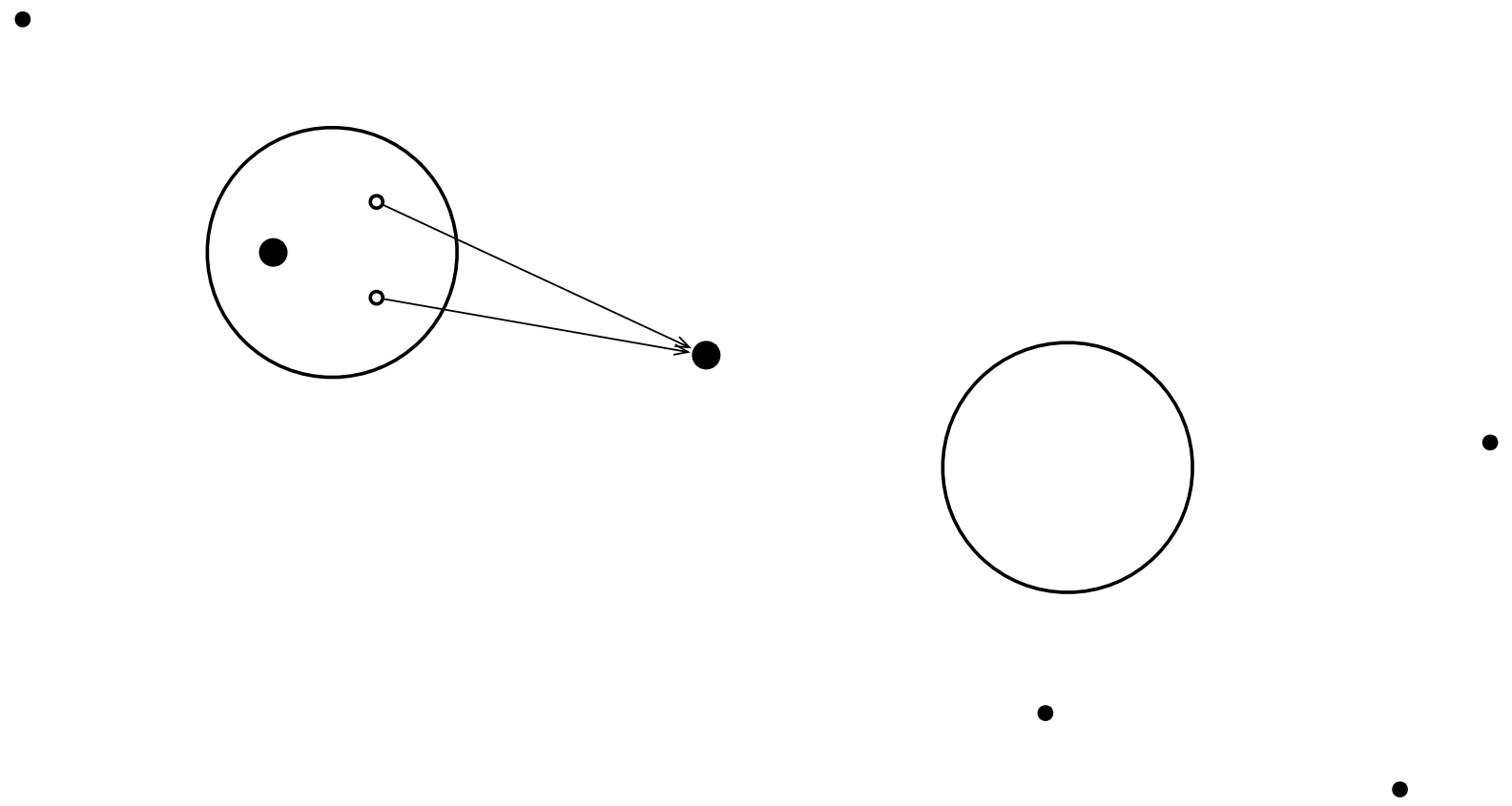


Figure 10: Streaming Model Clustering

Streaming with Outliers

We run an offline K center algorithm with outliers on the compressed representation, to select K centers that cover all but z points. This is our final output.



Conclusions

1. Concept of outliers can also be used for standard facility location (Charikar, Khuller, Mount, Narasimhan).

2. Extensions for the two metric case (Bhatia, Guha, Khuller, Sussmann). Fix K centers so that everyone is close to a center in each of two metrics.

   Approximation factor: 3. Uses matchings.

Lower bound on Cluster Size (Anonymity)

How do we publish data about individuals?

One solution: Remove identifying information (names) and then publish the information.

Problem: using public databases (voter records) people are able to infer information about individuals (or narrow the options down to a very small number).

Another approach (Agarwal, Feder, Kentapadhi, Khuller, Panigrahy, Thomas, Zhu) is to fudge the data slightly to provide anonymity.

Lower bound on Cluster Size (Anonymity)

Another approach: cluster data into dense clusters of small radius. Publish information about the cluster centers.

Problem is NP-complete even when the number of clusters is not specified!

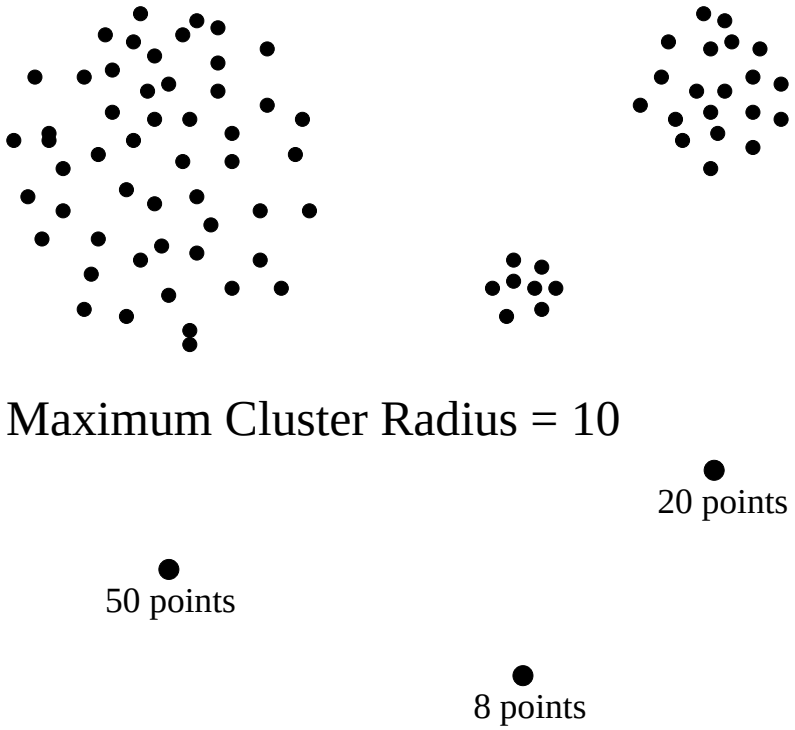


Figure 11: Publishing anonymized data

(K, r)-Center Problem

Cluster data into K clusters and minimize the largest radius.

Moreover, each cluster should have size at least r.

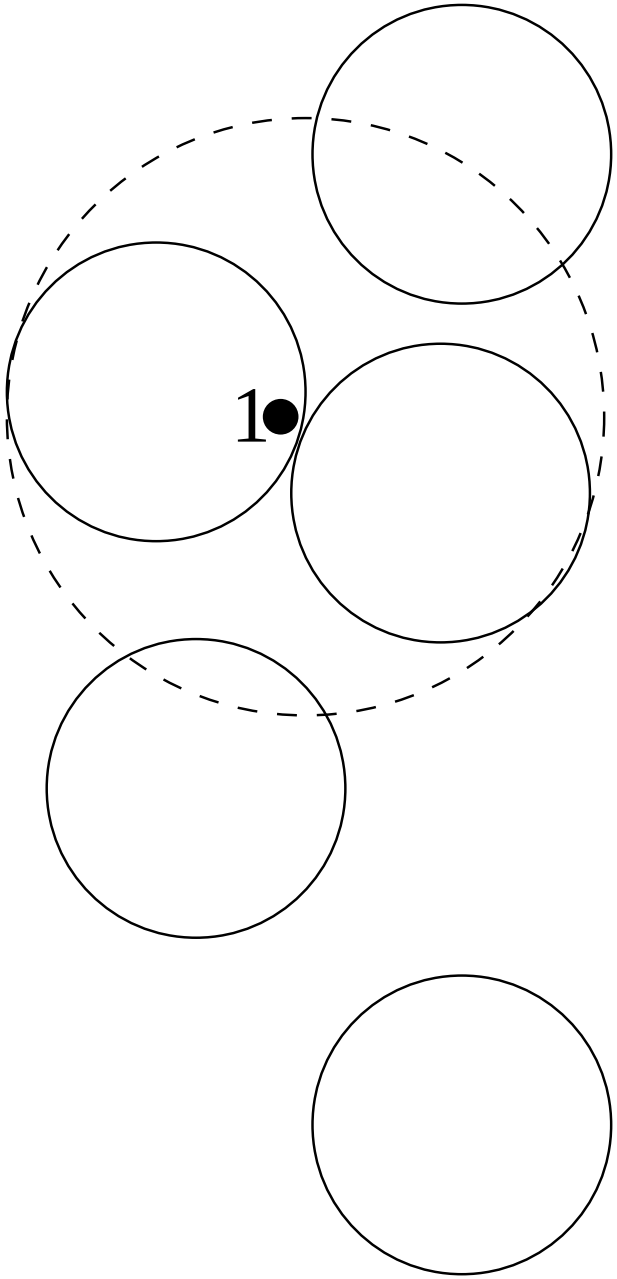
Condition (1) Each point in the database should have at least $r - 1$ other points within distance $2R$.

Condition (2) Let all nodes be unmarked initially.

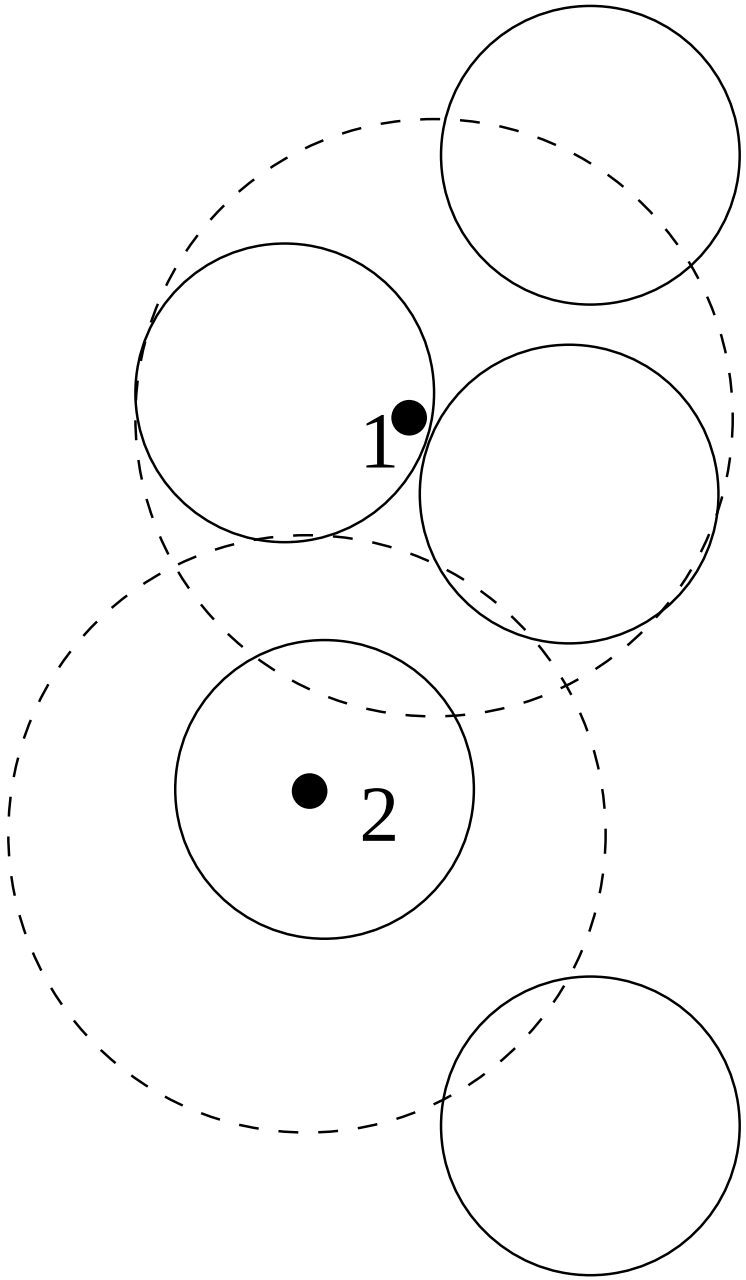
Select an arbitrary unmarked point as a center. Select all unmarked points within distance $2R$ to form a cluster and mark these points.

Repeat this as long as possible, until all points are marked.

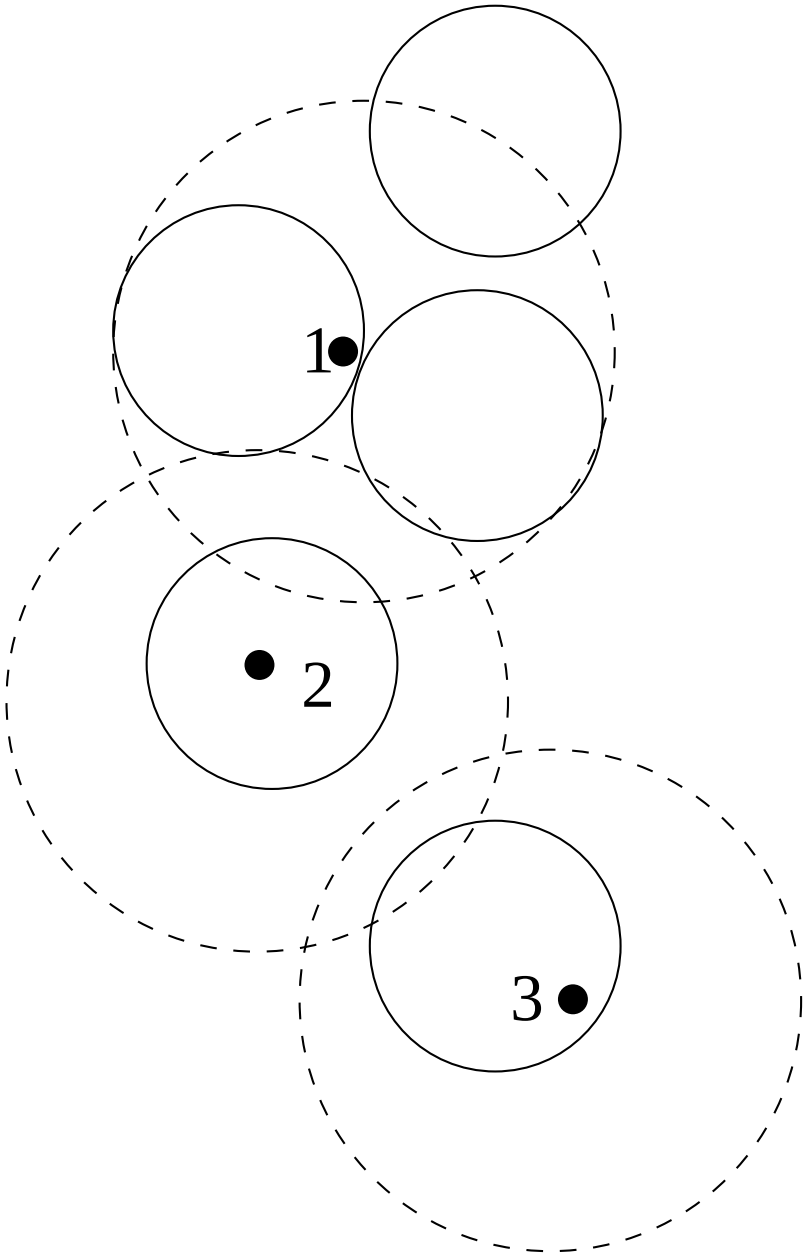
Example



Example

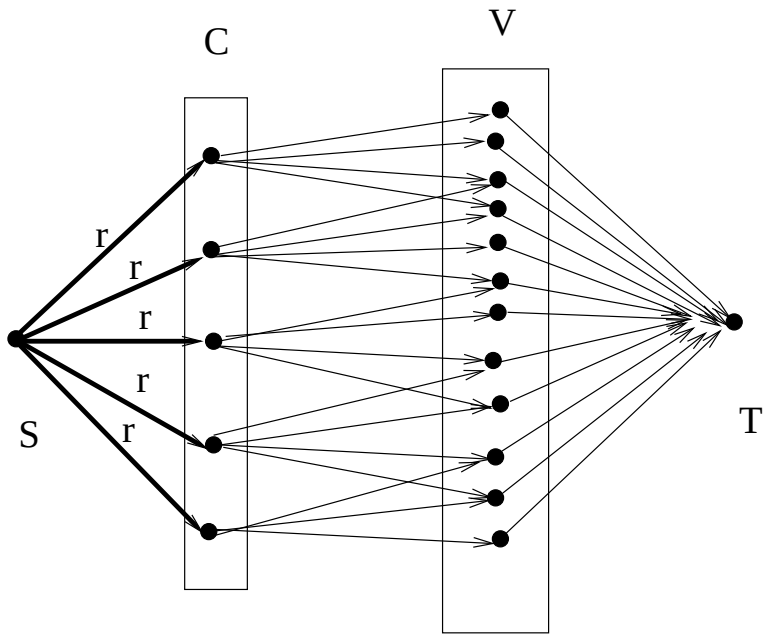


Example



Re-assignment Step

Reassign points to clusters to get at least r in each cluster.



Let C be the set of centers that were chosen. Add edges (capacity r) from s to each node in C. Add an edge of unit capacity from a node $c \in C$ to a node $v \in V$ $d(v, c) \leq 2R$. Check to see if a flow of value $r|C|$ can be found.

Re-assignment

Suppose r units of flow enter a node $v \in C$. The nodes of V through which the flow goes to the sink are assigned to v. Nodes of V through which no flow goes to the sink can be assigned anywhere.

(K, r, p)-Centers

Find K small clusters of size at least r so that at least p points are clustered.

Algorithm:

(Filtering Step) Let S be points v such that $|N(v, 2R)| \geq r$. Check if $|S| \geq p$, otherwise exit. We only consider points in S.

(Greedy Step) Choose up to K centers. Initially Q is empty. All points are uncovered initially. Let $N(v, \delta)$ be the set of *uncovered points* within distance δ of v. Once a point is covered it is removed.

Algorithm

At each step i, pick a center c_i that satisfies the following criteria:

(a) c_i is uncovered.

(b) $|N(c_i, 2R)|$ is maximum.

All uncovered points in $N(c_i, 4R)$ are then marked as covered.

After Q is chosen, check to see if at least p points are covered, otherwise exit with failure.

(Assignment step): Form clusters as follows. For each $c_i \in Q$, form a cluster C_i centered at c_i. Each covered point is assigned to its closest cluster center.

Denote $G_i = N(c_i, 2R)$ and $E_i = N(c_i, 4R)$, which are uncovered points within distance $2R$ and $4R$ of c_i, when c_i is chosen.

(K, r, p)-Centers

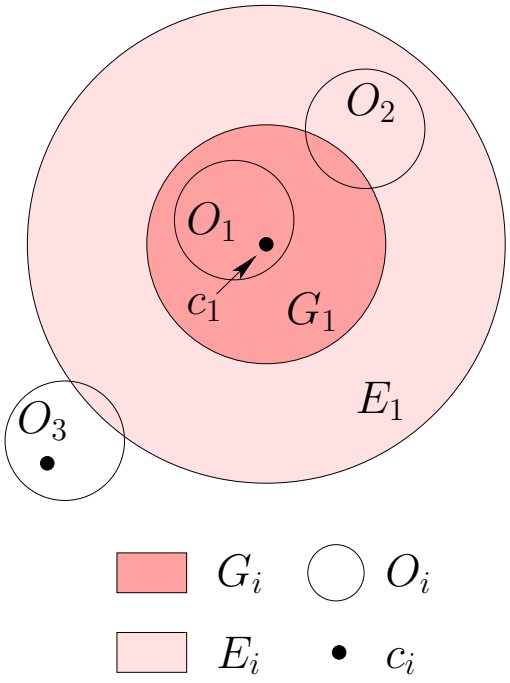


Figure 12: Optimal Clusters and the Greedy Step

Observations

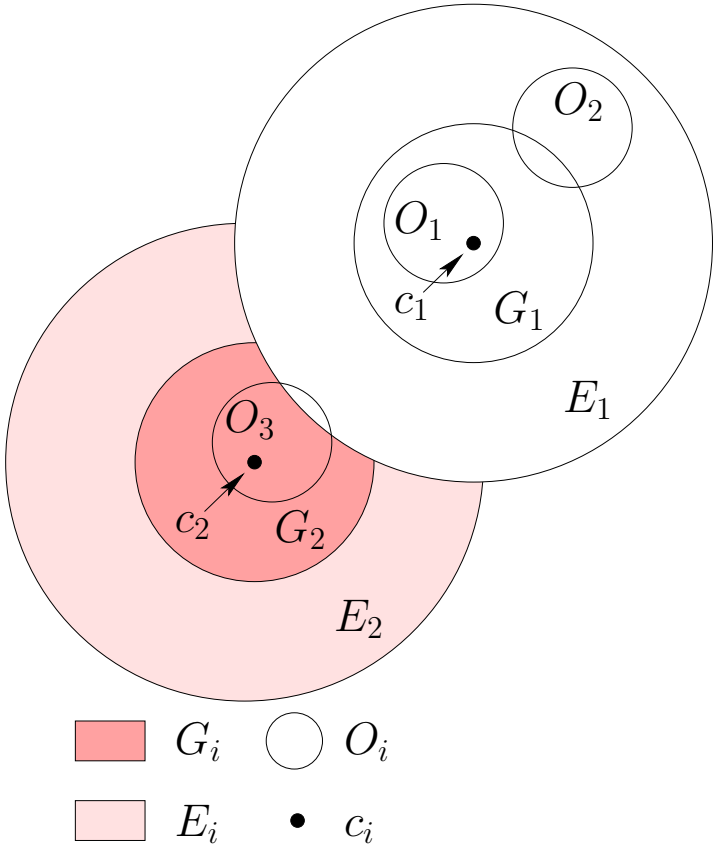


Figure 13: Optimal Clusters and the Greedy Step

Observations

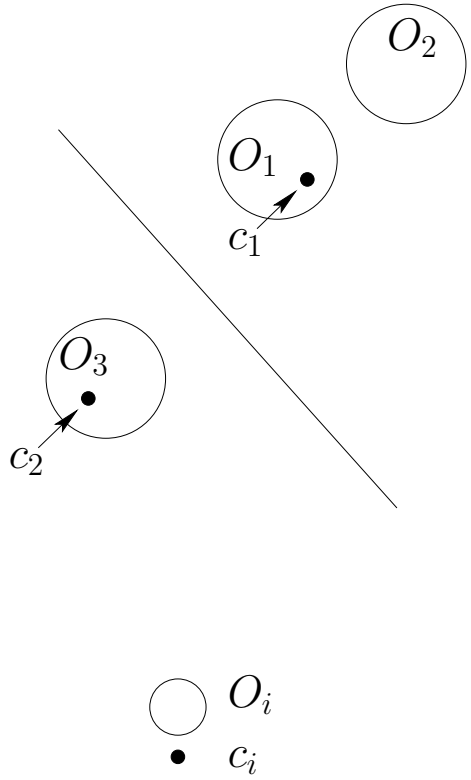


Figure 14: Optimal Clusters and the Greedy Step

Proof

Key Points:

- Cluster centers are far apart ($> 4R$), so we get all the points within radius $2R$ (at least r).
- Once a cluster is covered by G_i, it is completely covered by E_i (get all the points).
- E_i may grab a few points from any cluster making it sparse. However, these points will eventually be re-assigned to the center in this cluster if all the points are not covered by $E_j, j \geq i$.
- Proof that we get at least p points is similar to the proof done earlier.

Lower Bound on Cluster Sizes

For facility location Karger, Minkoff and Guha, Meyerson, Munagala give a $(\frac{r}{2}, 3\rho OPT_r)$ bound.

ρ is the approximation guarantee for facility location.

Currently $\rho \approx 1.5$.

r-Cellular Clustering

Find clusters such that each cluster has at least r points. The cost for cluster C_i is $R_i \cdot n_i$ (upper bound on distortion of data) and a facility cost of f_i.

$$Min \sum_i cost(C_i) + f_i$$

Use primal-dual methods to get a $O(1)$ approximation for this problem.

Conclusions

1. Concept of outliers can also be used for standard facility location (Charikar, Khuller, Mount, Narasimhan).

2. Extensions for the two metric case (Bhatia, Guha, Khuller, Sussmann). Fix K centers so that everyone is close to a center in each of two metrics.

   Approximation factor: 3. Uses matchings.