K-Center Clusterings and Generalizations

Samir Khuller

University of Maryland

College Park, Maryland

UNIVERSITY OF MARYLAND

Clustering Problems

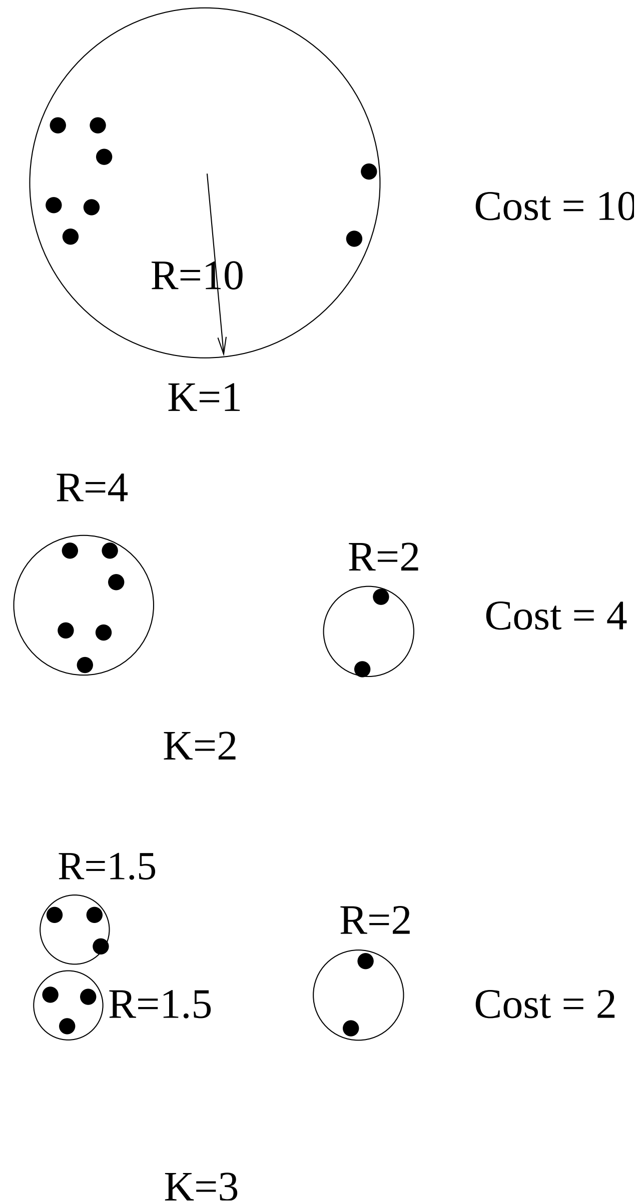
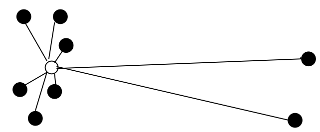


Figure 1: K-Center Clustering

Clustering Problems



K=1

K=2

K=3

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, Hochbaum-Shmoys).
- 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 $S \subset V$).

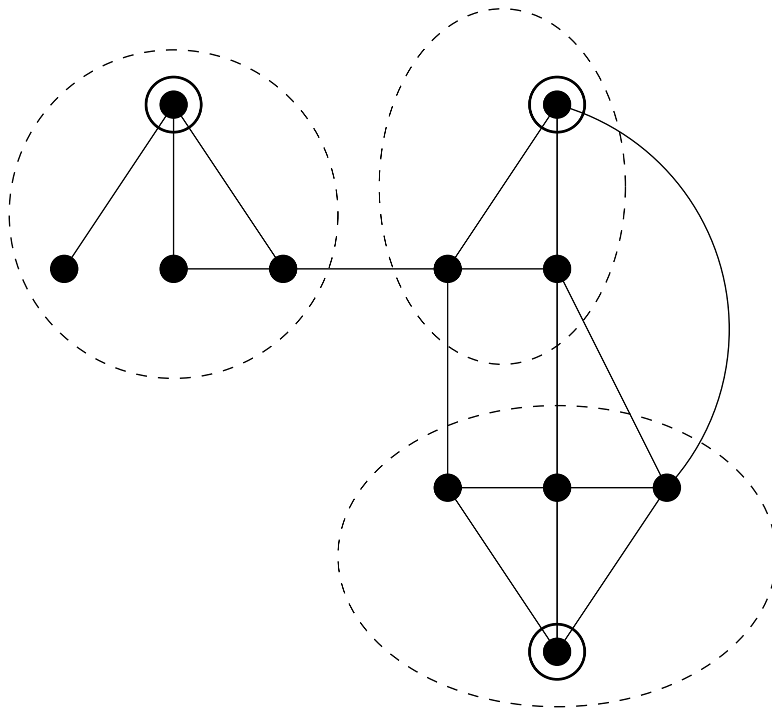
$\Longrightarrow$ “Guess” each possible value for R^*.
(At most ($O(n^2)$.)

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*.

Algorithm

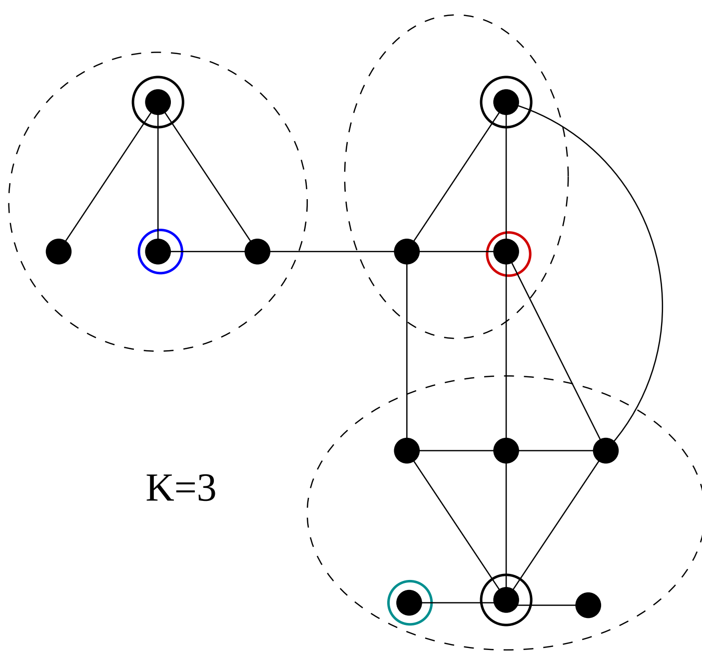
Try increasing values of δ.

Find a MIS S in G_δ^2.

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

$\delta = 5$

G_δ:

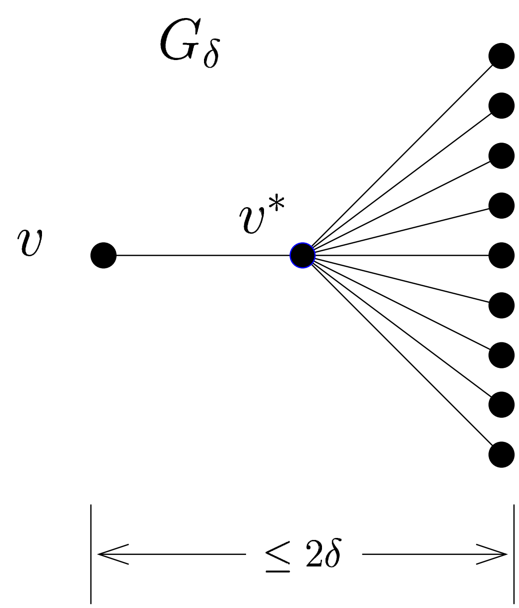


K=3

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.



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 $2R_i$.

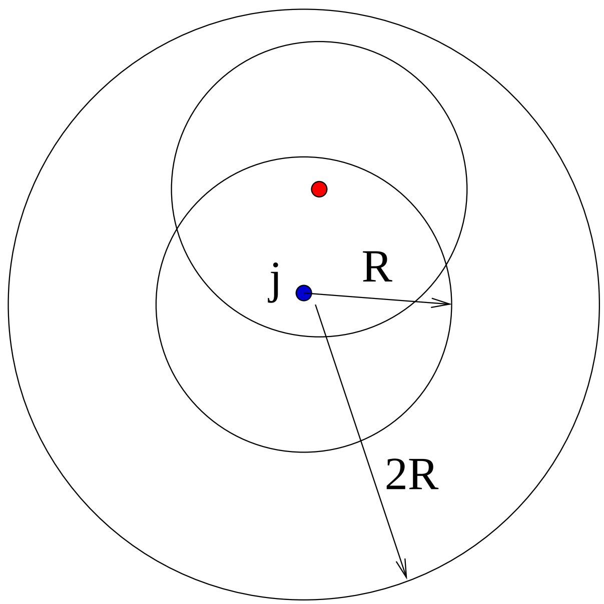


Figure 3: Hochbaum-Shmoys Method

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.

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.

Capacties on Cluster Sizes

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

(Khuller, Sussmann) Improve to factor 5 approximation.

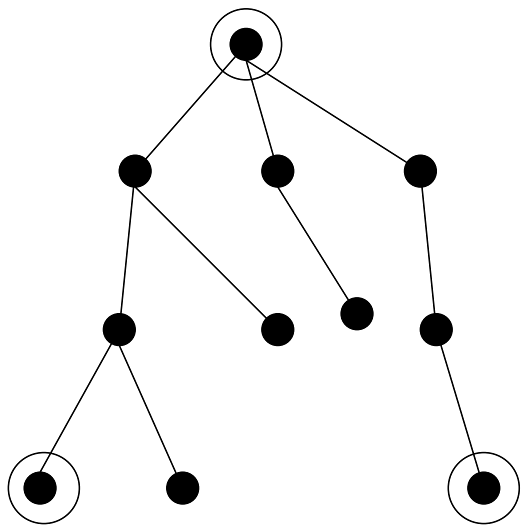


Figure 4: Tree of Centers

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.

Outliers

(Charikar, Khuller, Mount, Narasimhan) 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.

Extended to case $p = n$ (Cost K-Centers) recently (Chuzhoy, Halperin, Khanna, Kortsarz, Krauhtgamer, Naor).

OPEN: Can we get a $3 - \epsilon$ hardness for any $\epsilon > 0$ for the K-center problem with outliers?

Observations

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

For each point $v_i \in V$, let G_i (E_i, resp.) denote the set of points that are within distance R ($3R$, resp.) from v_i. G_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.

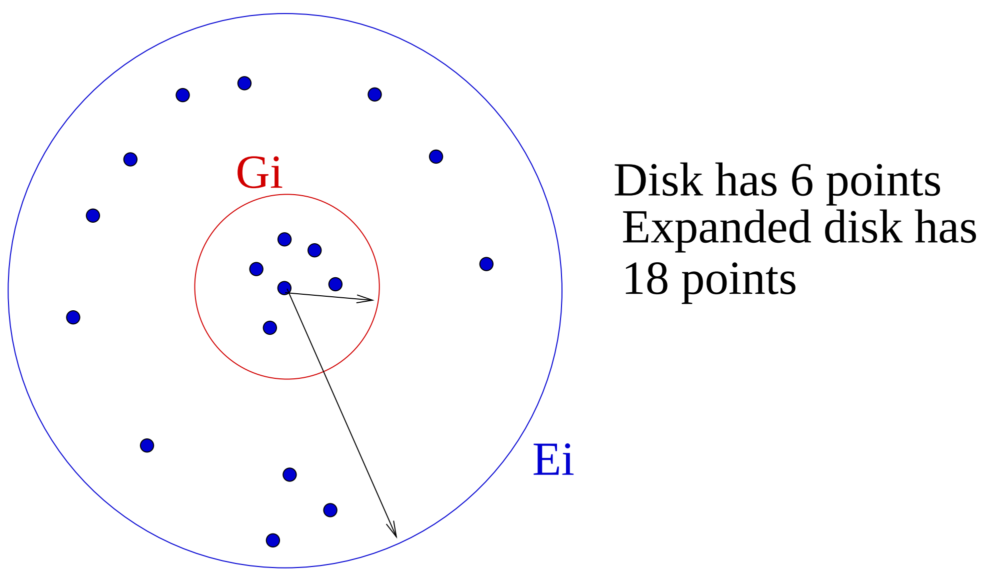


Figure 6: Disks and Expanded Disks.

New Algorithm (Robust K-centers/K-suppliers)

1. Initially all points are uncovered.
2. Construct all disks and corresponding expanded disks.
3. Repeat the following K times:
   - Let G_j be the heaviest disk, i.e. contains 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!

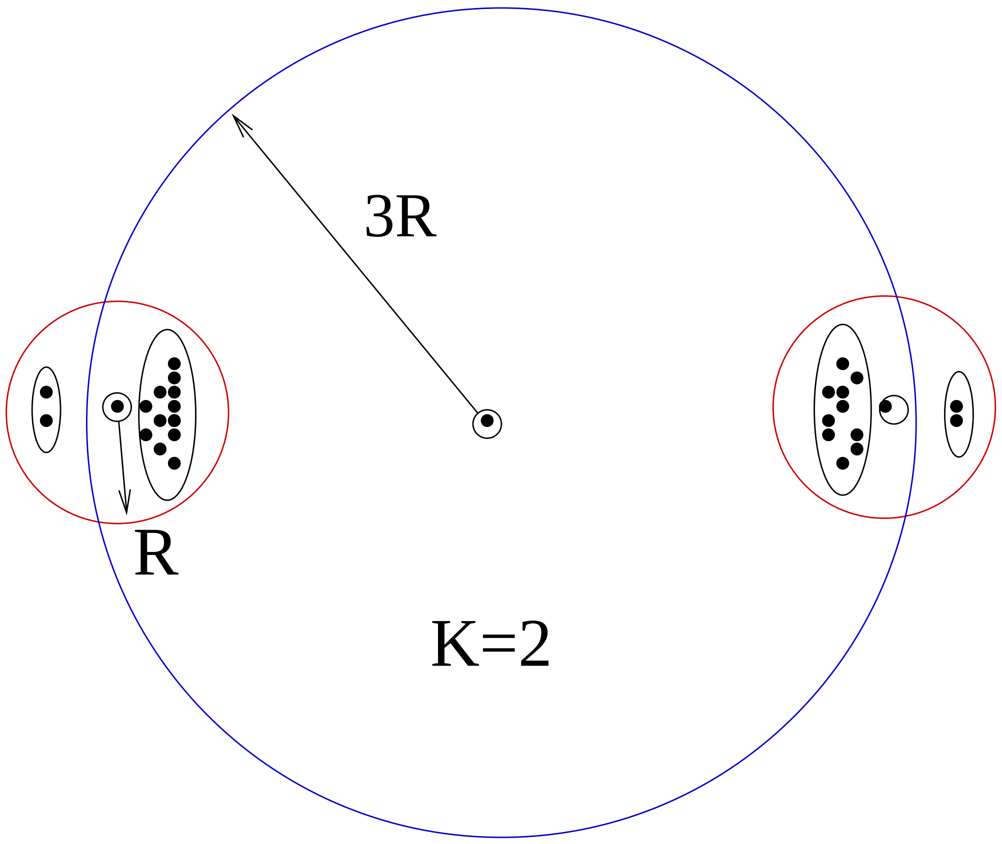


Figure 7: Bad example for choosing based on E_j.

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 G_j that covers a point in some O_i, then E_j covers all points in O_i.

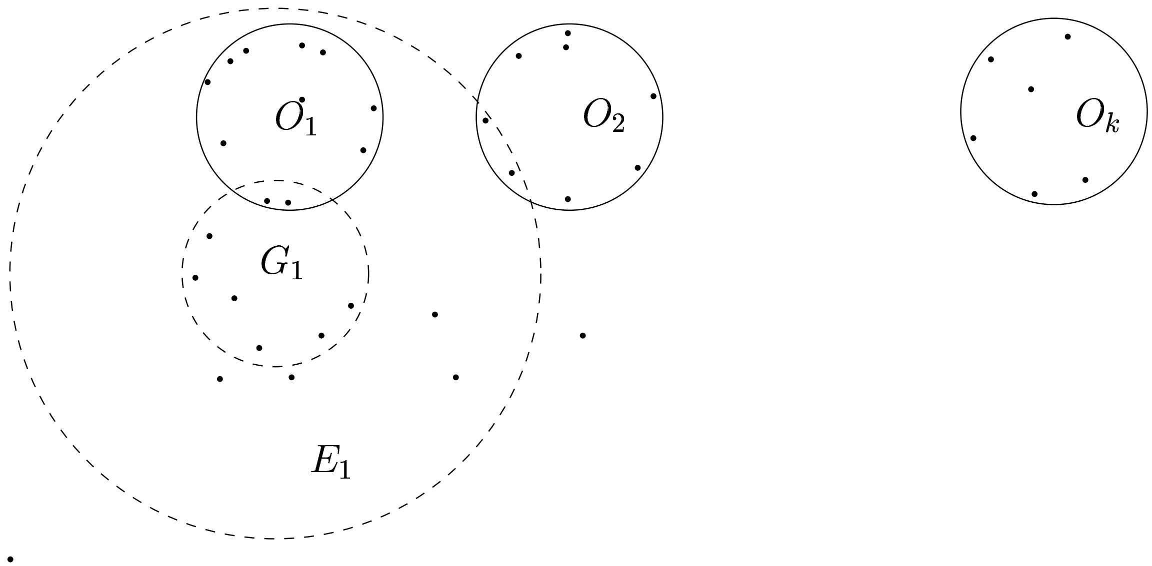


Figure 8: Optimal Clusters and the Greedy Step

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.

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!

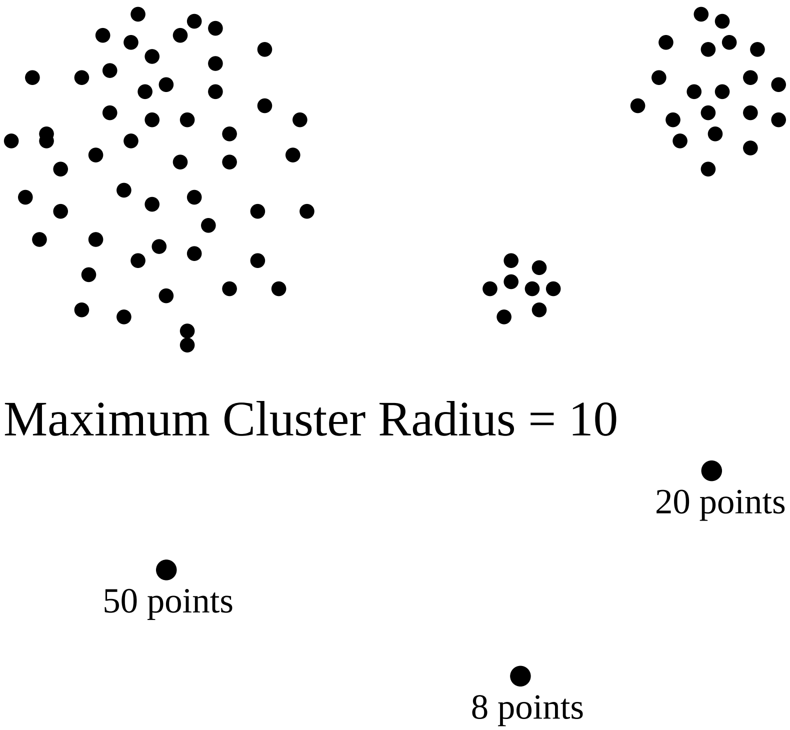


Figure 9: Publishing anonymized data

(K, r)-Center Problem

Cluster data into K clusters and minimize the largest radius.

Moreover, each cluster should have size at least r.

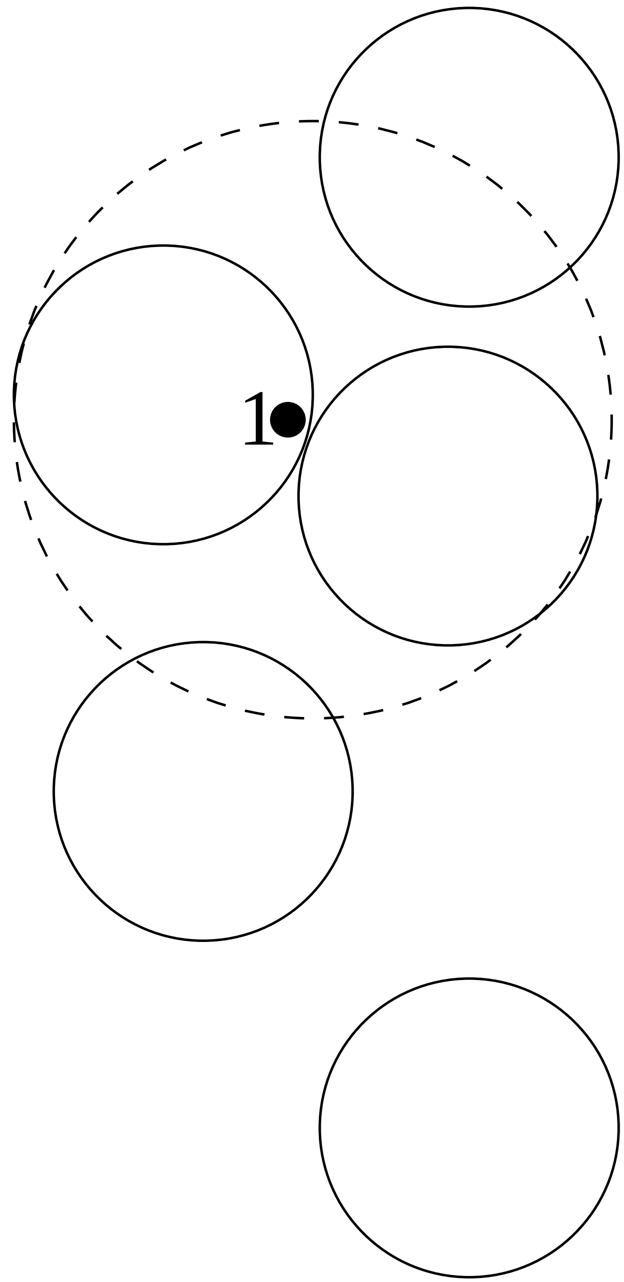
Condition (1) Each point in the database should have at least $r - 1$ other points within distance $2R$.

Condition (2) Let all nodes be unmarked initially.

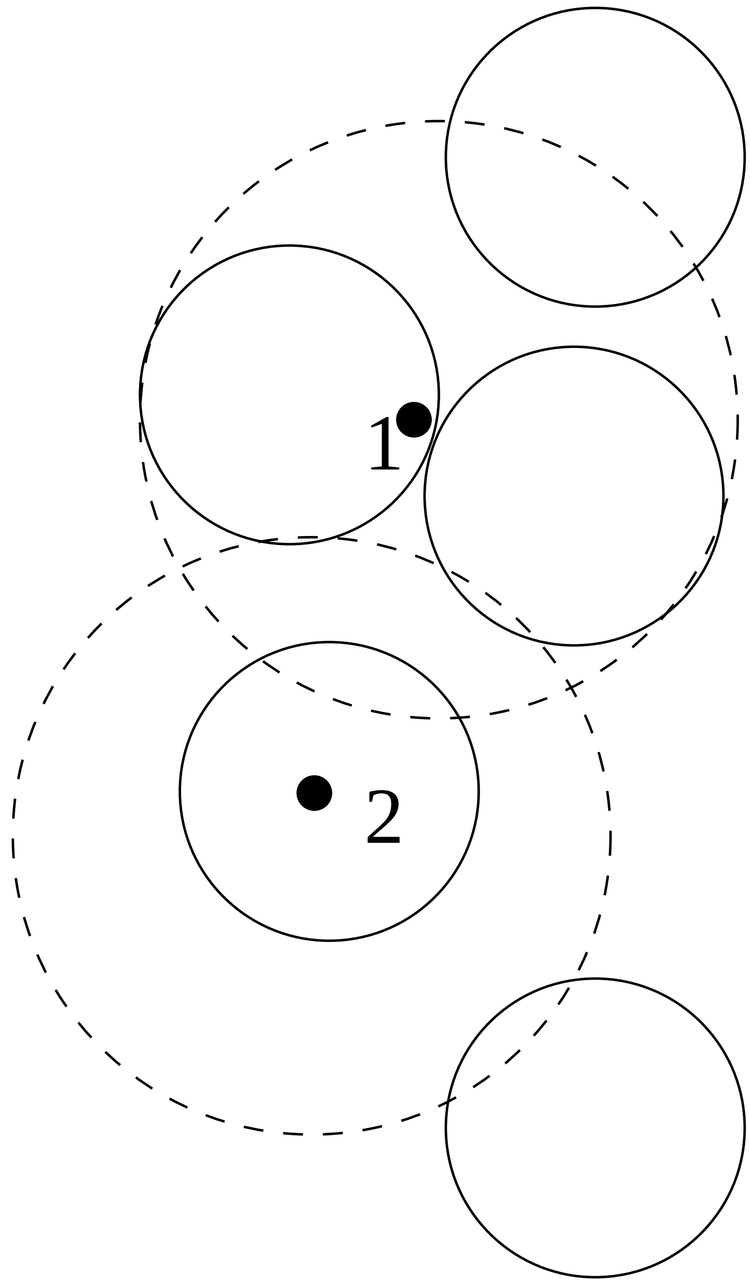
Select an arbitrary unmarked point as a center. Select all unmarked points within distance $2R$ to form a cluster and mark these points.

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

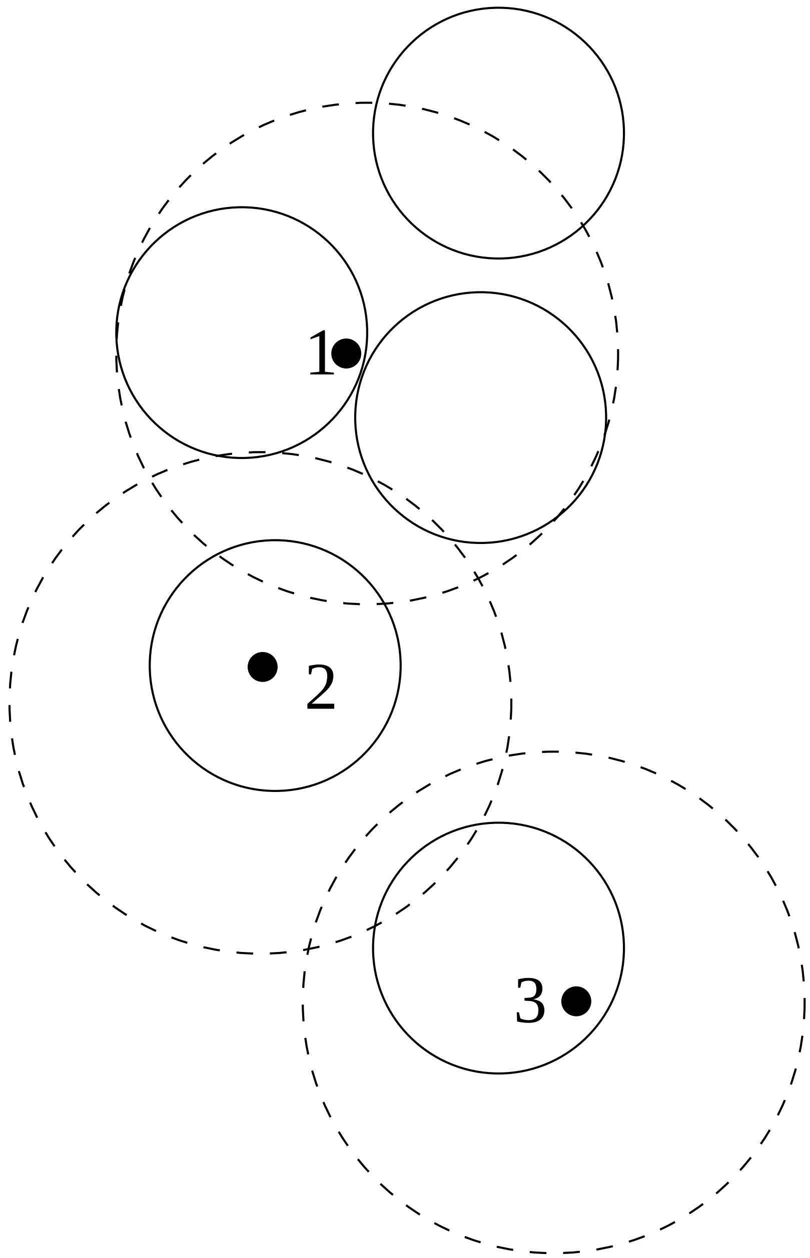
Example



Example

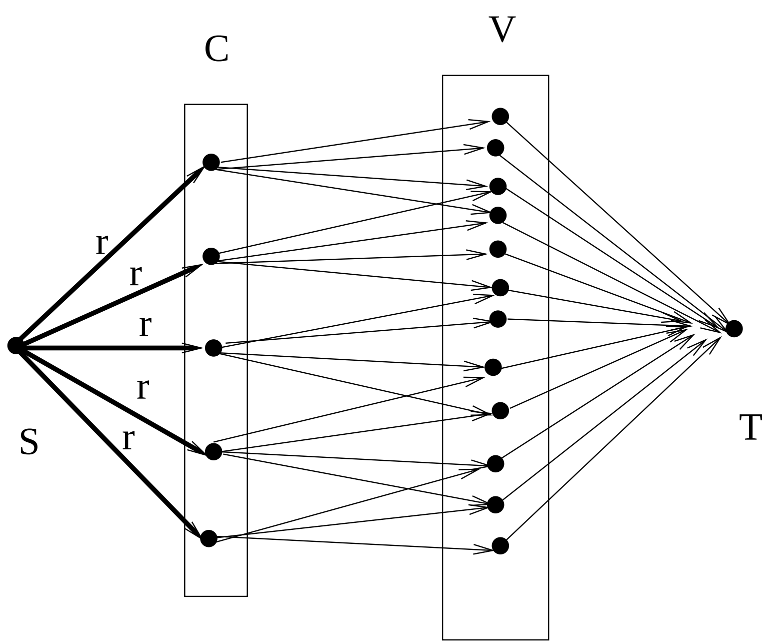


Example



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) $\left|N(c_i, 2R)\right|$ 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

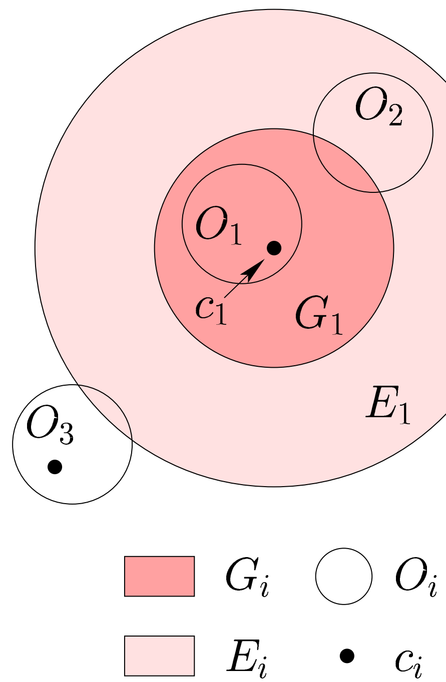


Figure 10: Optimal Clusters and the Greedy Step

Observations

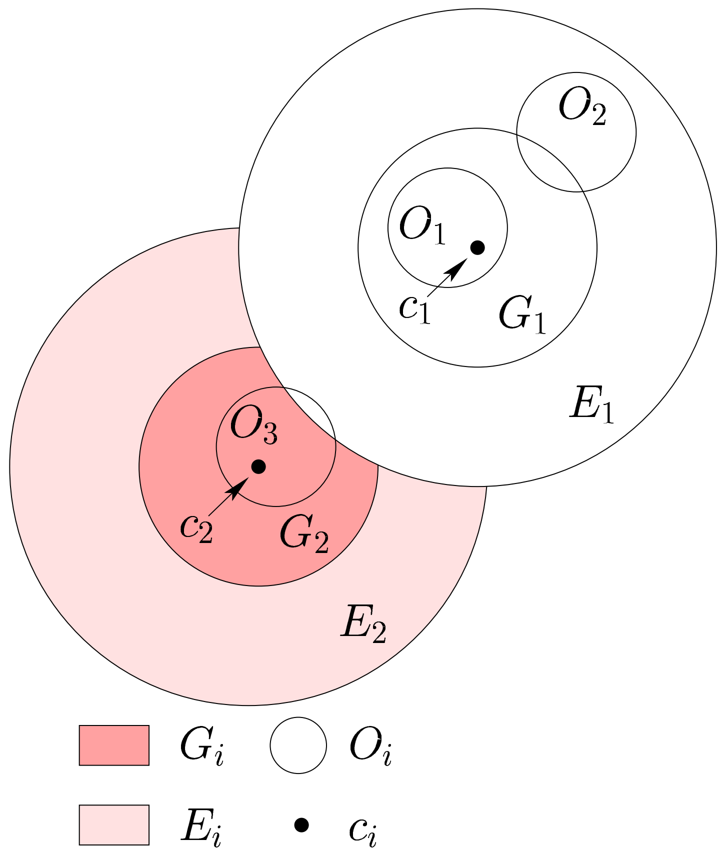


Figure 11: Optimal Clusters and the Greedy Step

Observations

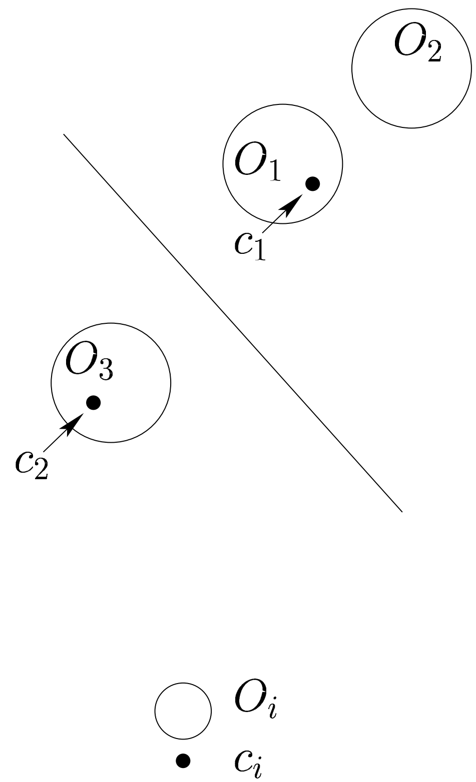


Figure 12: Optimal Clusters and the Greedy Step

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. K-centers can be solved with a single pass over the data. (Data stream clustering (Charikar,Chekuri,Feder,Motwani)). Approximation factor (randomized): $8(2e)$.

3. 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.

Thats all folks!