Local k-median and k-means clustering

Aversion k-clustering

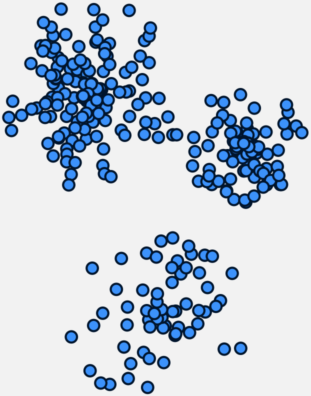
Melanie Schmidt

Anupam Gupta, Guru Guruganesh

6th Colloquium of the Research Area KL

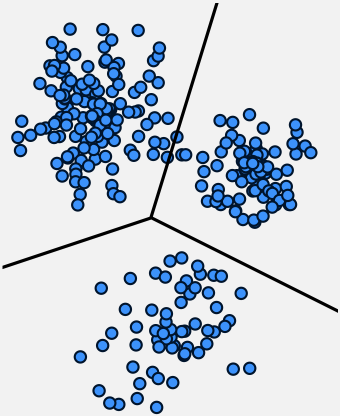
29.04.2016

Partitional Clustering



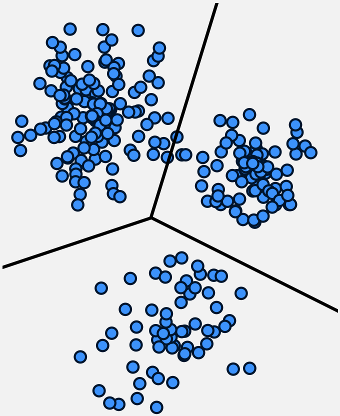
Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,



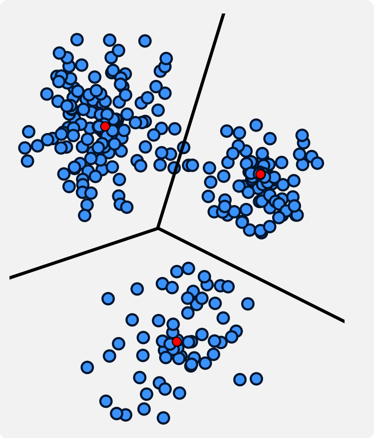
Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$



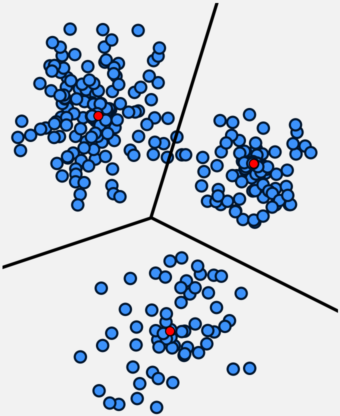
Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$
- to optimize some objective



Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$
- to optimize some objective



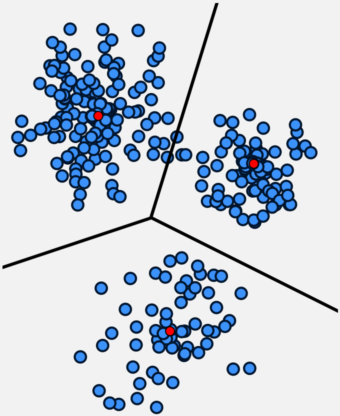
Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$
- to optimize some objective

Center based objectives

Pick k centers, assign to closest

- k-center: maximum distance, metric
- k-median: sum of distances, metric
- k-means: sum of squared distances, Euclidean space



Partitional Clustering

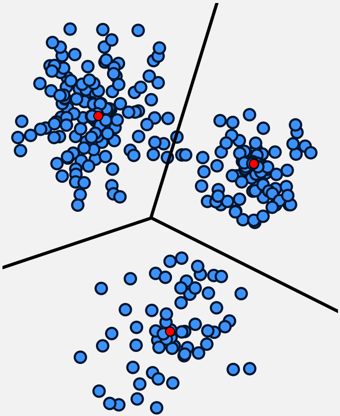
- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$
- to optimize some objective

Center based objectives

Pick k centers, assign to closest

- k-center: maximum distance, metric
- k-median: sum of distances, metric
- k-means: sum of squared distances, Euclidean space

Facility location flavor: k unbounded, but centers have opening cost



Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$
- to optimize some objective

Center based objectives

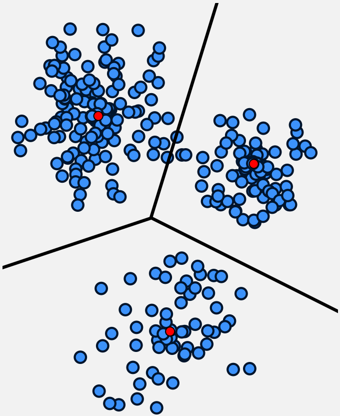
Pick k centers, assign to closest

- k-center: maximum distance, metric
- k-median: sum of distances, metric
- k-means: sum of squared distances, Euclidean space

Facility location flavor: k unbounded, but centers have opening cost

Not center based (metric)

- min sum k-clustering: sum of all pairwise distances
- aversion k-clustering: sum of distances to furthest away point



Partitional Clustering

- Given a point set P and $k \in \mathbb{N}$,
- partition P into $C_1, \ldots, C_k$
- to optimize some objective

Center based objectives

Pick k centers, assign to closest

- k-center: maximum distance, metric
- k-median: sum of distances, metric
- k-means: sum of squared distances, Euclidean space

Facility location flavor: k unbounded, but centers have opening cost

Not center based (metric)

- min sum k-clustering: sum of all pairwise distances
- aversion k-clustering: sum of distances to furthest away point

Approximation: State of the art

	Approximation	Hardness
k-center		
facility location		
k-median		
k-means		
min sum *k*-clustering		
aversion *k*-clustering		

Approximation: State of the art

	Approximation	Hardness
k-center	2 [G85,HS85]	2 [HN79]
facility location	1.488 [L11]	1.463 [GK99]
k-median	$2.675 + \varepsilon$ [BPRST15]	$1 + 2/e \approx 1.74$ [HN79]
k-means	$9 + \varepsilon$ [KMN+02]	1.0013 [ACKS15], [LSW15]
min sum k-clustering	$\mathcal{O}(\log n)$ [BFSS15]	–
aversion k-clustering		–

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[BFSS15] Behsaz, Friggstad, Salavatipour, Sivakumar. Appr. Alg. for Min-Sum k-Clustering and Balanced k-Median, ICALP 2015.
[BPRST15] Byrka, Pensyl, Rybicki, Srinivasan, Trinh. An Improved Approximation for k-median [. . .], SODA 2015.
[G85] Gonzalez. Clustering to minimize the maximum intercluster distance, Theoretical Computer Science 1985.
[GK99] Guha, Khuller. Greedy strikes back: Improved facility location algorithms, J. Algorithms 1999.
[HN79] Hsu, Nemhauser, Easy and hard bottleneck location problems. Discrete Applied Mathematics 1979.
[HS85] Hochbaum, Shmoys, A best possible heuristic for the k-center problem, Mathematics of Operations Research 1985.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Wu. A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright. Improved and Simplified Inapproximability for k-means, CORR 2015.
[L11] Li. A 1.488-approximation algorithm for the uncapacitated facility location problem. ICALP 2011.

Approximation: State of the art

	Approximation	Hardness
k-center	2 [G85,HS85]	2 [HN79]
facility location	1.488 [L11]	1.463 [GK99]
k-median	$2.675 + \varepsilon$ [BPRST15]	$1 + 2/e \approx 1.74$ [HN79]
k-means	$9 + \varepsilon$ [KMN+02]	1.0013 [ACKS15], [LSW15]
min sum k-clustering	$\mathcal{O}(\log n)$ [BFSS15]	–
aversion k-clustering	$\mathcal{O}(1)$	–

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[BFSS15] Behsaz, Friggstad, Salavatipour, Sivakumar. Appr. Alg. for Min-Sum k-Clustering and Balanced k-Median, ICALP 2015.
[BPRST15] Byrka, Pensyl, Rybicki, Srinivasan, Trinh. An Improved Approximation for k-median [...], SODA 2015.
[G85] Gonzalez. Clustering to minimize the maximum intercluster distance, Theoretical Computer Science 1985.
[GK99] Guha, Khuller. Greedy strikes back: Improved facility location algorithms, J. Algorithms 1999.
[HN79] Hsu, Nemhauser, Easy and hard bottleneck location problems. Discrete Applied Mathematics 1979.
[HS85] Hochbaum, Shmoys, A best possible heuristic for the *k*-center problem, Mathematics of Operations Research 1985.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Wu. A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright. Improved and Simplified Inapproximability for k-means, CORR 2015.
[L11] Li. A 1.488-approximation algorithm for the uncapacitated facility location problem. ICALP 2011.

Many techniques

Center based

- k-center somewhat easier
- Often facility location is first, k-means last

Many techniques

Center based

- *k*-center somewhat easier
- Often facility location is first, *k*-means last

- Embedding into tree metrics
- Local Search
- LP-based rounding
  - Filtering
  - Dual fitting
  - Randomized rounding
- Primal-dual framework

Many techniques

Center based

- k-center somewhat easier
- Often facility location is first, k-means last

- Embedding into tree metrics
- Local Search
- LP-based rounding
  - Filtering
  - Dual fitting
  - Randomized rounding
- Primal-dual framework

Not center based

- Reductions to center-based objectives
- min sum k-clustering: balanced k-median (cluster cost times $|C_i|$)
- aversion k-clustering: local k-median

Basic clustering objectives are well studied,
but need solutions with constraints

Basic clustering objectives are well studied,
but need solutions with constraints

Capacitated clustering

Every center has a maximal capacity for assigned points

Basic clustering objectives are well studied,
but need solutions with constraints

Capacitated clustering

Every center has a maximal capacity for assigned points

Lower bounded clustering

It is not worthwhile to open a center unless a lower bound is met

Basic clustering objectives are well studied,
but need solutions with constraints

Capacitated clustering

Every center has a maximal capacity for assigned points

Lower bounded clustering

It is not worthwhile to open a center unless a lower bound is met

Clustering with outliers

The data is noisy and p outliers should be found and ignored

Basic clustering objectives are well studied,
but need solutions with constraints

Capacitated clustering

Every center has a maximal capacity for assigned points

Lower bounded clustering

It is not worthwhile to open a center unless a lower bound is met

Clustering with outliers

The data is noisy and p outliers should be found and ignored

Clustering with locality constraints

A point can only be assigned to a center if it is within its radius

Constraints make problems difficult. . .

Capacitated facility location

Constraints make problems difficult. . .

Capacitated facility location, all distances zero

Constraints make problems difficult. . .

Capacitated facility location, all distances zero

Should be simple?

Constraints make problems difficult...

Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity

Constraints make problems difficult. . .

Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity
- minimize total opening cost

Constraints make problems difficult. . .

Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity
- minimize total opening cost

$\rightsquigarrow$ Minimum Knapsack Problem

Constraints make problems difficult. . .

Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity
- minimize total opening cost

$\rightsquigarrow$ Minimum Knapsack Problem

Standard LP: integrality gap

Constraints make problems difficult. . .

Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity
- minimize total opening cost

$\rightsquigarrow$ Minimum Knapsack Problem

Standard LP: integrality gap

i_1 zero distance i_2

n clients, cap. $n-1$, opening cost 0

1 client, cap. $n-1$, opening cost 1

Constraints make problems difficult. . .

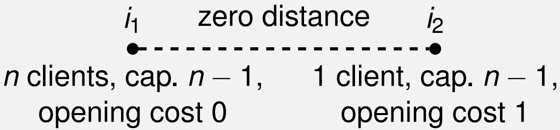
Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity
- minimize total opening cost

$\rightsquigarrow$ Minimum Knapsack Problem

Standard LP: integrality gap



- integral solution: Assign one client to $i_2 \rightsquigarrow$ cost 1
- fractional solution: Assign all clients to i_2 by $\frac{1}{n} \rightsquigarrow$ cost $\frac{1}{n}$

Constraints make problems difficult. . .

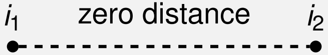
Capacitated facility location, all distances zero

Should be simple?

- find facilities with $\geq n$ total capacity
- minimize total opening cost

$\rightsquigarrow$ Minimum Knapsack Problem

Standard LP: integrality gap unbounded



n clients, cap. $n-1$, opening cost 0 | 1 client, cap. $n-1$, opening cost 1

- integral solution: Assign one client to $i_2 \rightsquigarrow$ cost 1
- fractional solution: Assign all clients to i_2 by $\frac{1}{n} \rightsquigarrow$ cost $\frac{1}{n}$

Capacitated clustering

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]
- $\mathcal{O}(1)$-approximation with knapsack based LP [ASS14]

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]
- $\mathcal{O}(1)$-approximation with knapsack based LP [ASS14]
- capacitated k-center: ≥ 3 [CHK12] and ≤ 9 [ABC+]

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]
- $\mathcal{O}(1)$-approximation with knapsack based LP [ASS14]
- capacitated k-center: ≥ 3 [CHK12] and ≤ 9 [ABC+]
- No constant factor known for capacitated k-median or k-means

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]
- $\mathcal{O}(1)$-approximation with knapsack based LP [ASS14]
- capacitated k-center: ≥ 3 [CHK12] and ≤ 9 [ABC+]
- No constant factor known for capacitated k-median or k-means
- Various bicriteria approximations

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]
- $\mathcal{O}(1)$-approximation with knapsack based LP [ASS14]
- capacitated k-center: ≥ 3 [CHK12] and ≤ 9 [ABC+]
- No constant factor known for capacitated k-median or k-means
- Various bicriteria approximations
- Research on easier variants: uniform or soft capacities

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Capacitated clustering

- Local search based 5-approximation for cap. fac. location [BGG12]
- $\mathcal{O}(1)$-approximation with knapsack based LP [ASS14]
- capacitated k-center: ≥ 3 [CHK12] and ≤ 9 [ABC+]
- No constant factor known for capacitated k-median or k-means
- Various bicriteria approximations
- Research on easier variants: uniform or soft capacities

[ABC+] An, Bhaskara, Chekuri, Gupta, Madan, Svensson. Centrality of trees for capacitated k-center, Math. Progr. 2015.
[ASS14] An, Singh, Svensson. LP-Based Algorithms for Capacitated Facility Location, FOCS 2014.
[BGG12] Bansal, Garg, Gupta. A 5-Approximation for Capacitated Facility Location, ESA 2012.
[CHK12] Cygan, Hajiaghayi, Khuller. LP rounding for k-centers with non-uniform hard capacities, FOCS 2012.

Lower bounds

- No constant factor approximations for nonuniform bounds at all

Local k-median

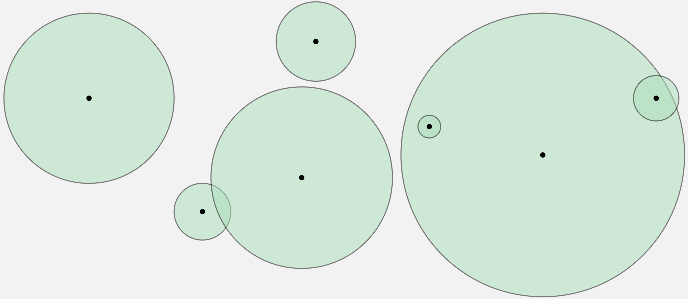
- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

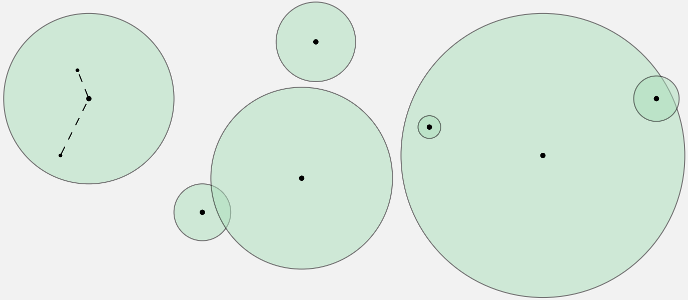
$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$



Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

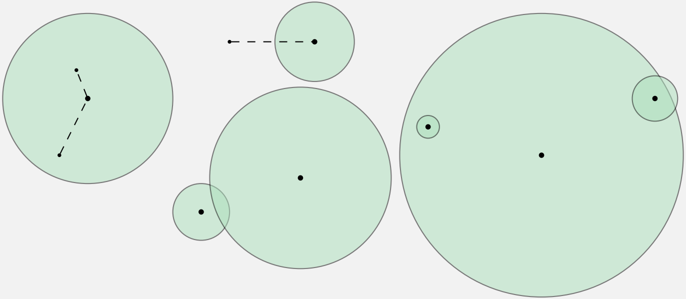
$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$



Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

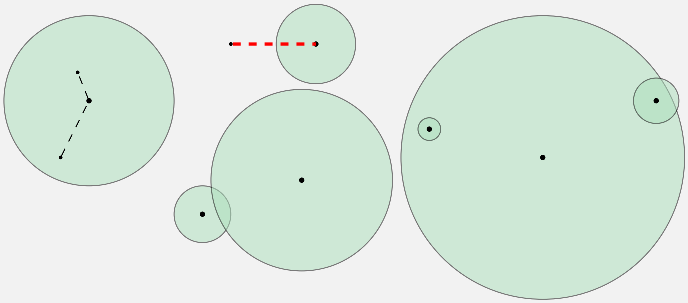
$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$



Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$



Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i, j)$$

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i, j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Reduction from set cover

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Reduction from set cover

- Can k sets from $\mathcal{S}$ cover universum U?

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Reduction from set cover

- Can k sets from $\mathcal{S}$ cover universum U?
- clients=U, facilities=$\mathcal{S}$, radii are all one

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Reduction from set cover

- Can k sets from $\mathcal{S}$ cover universum U?
- clients=U, facilities=$\mathcal{S}$, radii are all one
- distance between $e \in U$ and $S \in \mathcal{S}$: 1 iff $e \in S$, 2 else

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Reduction from set cover

- Can k sets from $\mathcal{S}$ cover universum U?
- clients=U, facilities=$\mathcal{S}$, radii are all one
- distance between $e \in U$ and $S \in \mathcal{S}$: 1 iff $e \in S$, 2 else

$\rightsquigarrow$ feasible solution for local k-median $\Leftrightarrow$ set cover answer is yes

Local k-median

- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \leq \mathbb{R}_i}} D(i,j)$$

- Some points cannot be assigned $\rightsquigarrow$ preprocessing $\rightsquigarrow$ easy
- But: Are there k centers such that all points can be assigned?

Reduction from set cover $\rightsquigarrow$ Approximation is NP-hard

- Can k sets from $\mathcal{S}$ cover universum U?
- clients=U, facilities=$\mathcal{S}$, radii are all one
- distance between $e \in U$ and $S \in \mathcal{S}$: 1 iff $e \in S$, 2 else

$\rightsquigarrow$ feasible solution for local k-median $\Leftrightarrow$ set cover answer is yes

Conclusion (Part I)

Conclusion (Part I)

- Constraints make problems significantly harder

Conclusion (Part I)

- Constraints make problems significantly harder
- Active research topic, lots of new results

Conclusion (Part I)

- Constraints make problems significantly harder
- Active research topic, lots of new results
- Many open questions, in particular about upper/lower bounds

Conclusion (Part I)

- Constraints make problems significantly harder
- Active research topic, lots of new results
- Many open questions, in particular about upper/lower bounds
- Local k-median cannot be approximated

Conclusion (Part I)

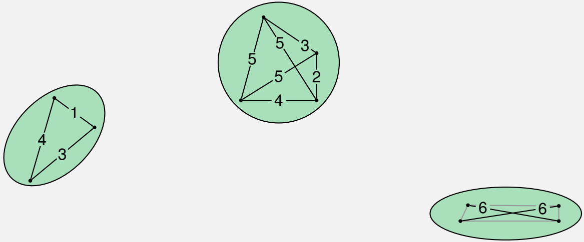
- Constraints make problems significantly harder
- Active research topic, lots of new results
- Many open questions, in particular about upper/lower bounds
- Local k-median cannot be approximated
- (Open: Local facility location?)

Aversion k-clustering

Aversion k-clustering

- Metric space (X, d), points=clients C, no facilities
- Find partitioning $C_1, \ldots, C_k$ of C
- Minimize

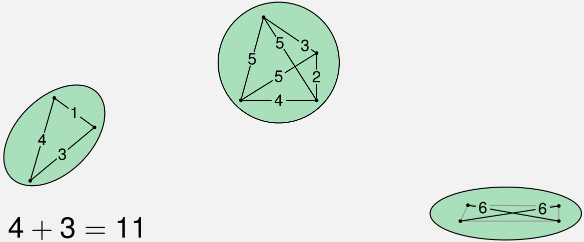
$$\sum_{\ell=1}^{k} \sum_{j_1 \in C_\ell} \max_{j_2 \in C_\ell} D(j_1, j_2)$$



Aversion k-clustering

- Metric space (X, d), points=clients C, no facilities
- Find partitioning $C_1, \ldots, C_k$ of C
- Minimize

$$\sum_{\ell=1}^{k} \sum_{j_1 \in C_\ell} \max_{j_2 \in C_\ell} D(j_1, j_2)$$

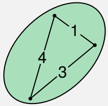


$4 + 4 + 3 = 11$

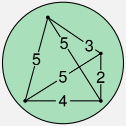
Aversion k-clustering

- Metric space (X, d), points=clients C, no facilities
- Find partitioning $C_1, \ldots, C_k$ of C
- Minimize

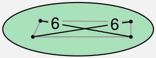
$$\sum_{\ell=1}^{k} \sum_{j_1 \in C_\ell} \max_{j_2 \in C_\ell} D(j_1, j_2)$$



$4 + 4 + 3 = 11$



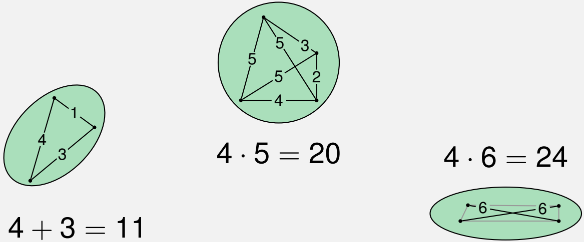
$4 \cdot 5 = 20$



Aversion k-clustering

- Metric space (X, d), points=clients C, no facilities
- Find partitioning $C_1, \ldots, C_k$ of C
- Minimize

$$\sum_{\ell=1}^{k} \sum_{j_1 \in C_\ell} \max_{j_2 \in C_\ell} D(j_1, j_2)$$



$4 + 4 + 3 = 11$ $\quad$ $4 \cdot 5 = 20$ $\quad$ $4 \cdot 6 = 24$

Where does this problem come from?

Aversion *k*-clustering

Where does this problem come from?

Aversion *k*-clustering

- Name: Cost of a point is its maximal dislike to any other member

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions
- More precisely: Paper by Kroer, Sandholm at EC 2014

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions
- More precisely: Paper by Kroer, Sandholm at EC 2014

Extensive-Form Game Imperfect-Recall Abstractions

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions
- More precisely: Paper by Kroer, Sandholm at EC 2014

Extensive-Form Game Imperfect-Recall Abstractions

- Large Extensive-Form Games: Computing Equilibria too slow

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions
- More precisely: Paper by Kroer, Sandholm at EC 2014

Extensive-Form Game Imperfect-Recall Abstractions

- Large Extensive-Form Games: Computing Equilibria too slow
- Practical solution: Imperfect-Recall Abstractions

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions
- More precisely: Paper by Kroer, Sandholm at EC 2014

Extensive-Form Game Imperfect-Recall Abstractions

- Large Extensive-Form Games: Computing Equilibria too slow
- Practical solution: Imperfect-Recall Abstractions
- [KS14] provide first approximation bounds

Where does this problem come from?

Aversion k-clustering

- Name: Cost of a point is its maximal dislike to any other member
- Origin: Extensive-Form Game Imperfect-Recall Abstractions
- More precisely: Paper by Kroer, Sandholm at EC 2014

Extensive-Form Game Imperfect-Recall Abstractions

- Large Extensive-Form Games: Computing Equilibria too slow
- Practical solution: Imperfect-Recall Abstractions
- [KS14] provide first approximation bounds
- Computing a bound-minimizing abstraction for a single level $=$ clustering problem in a metric space

Imperfect-Recall Abstractions by Kroer/Sandholm

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point
- Cost function depends on game

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point
- Cost function depends on game
- Two extremes: k-center and aversion k-clustering

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point
- Cost function depends on game
- Two extremes: k-center and aversion k-clustering

- k-center

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point
- Cost function depends on game
- Two extremes: *k*-center and aversion *k*-clustering

- *k*-center ☺

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point
- Cost function depends on game
- Two extremes: k-center and aversion k-clustering

- k-center ☺
- aversion k-center

Imperfect-Recall Abstractions by Kroer/Sandholm

- Different game states are collapsed into one state
- These form a cluster, accuracy loss for each point
- Cost function depends on game
- Two extremes: k-center and aversion k-clustering

- k-center ☺
- aversion k-center ⇝ not yet studied, and non-linear ☹

Constant approximation via local k-median

Constant approximation via local k-median

Aversion k-clustering

- Metric space (X, d), points=clients C, no facilities
- Find partitioning $C_1, \ldots, C_k$ of C
- Minimize

$$\sum_{\ell=1}^{k} \sum_{j_1 \in C_\ell} \max_{j_2 \in C_\ell} D(j_1, j_2)$$

Constant approximation via local k-median

Aversion k-clustering

- Metric space (X, d), points=clients C, no facilities
- Find partitioning $C_1, \ldots, C_k$ of C
- Minimize

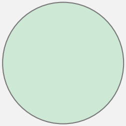
$$\sum_{\ell=1}^{k} \sum_{j_1 \in C_\ell} \max_{j_2 \in C_\ell} D(j_1, j_2)$$

Local k-median

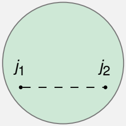
- Metric space (X, d), points=clients C, possible centers=facilities $\mathcal{F}$
- each facility $i \in F$ has a radius R_i
- find k facilities, $F \subset \mathcal{F}$, k is fixed
- Minimize

$$\sum_{j \in C} \min_{\substack{i \in F, \\ D(i,j) \le \mathbb{R}_i}} D(i, j)$$

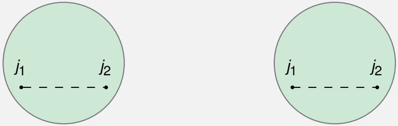
Converting aversion k-clustering instances for local k-median



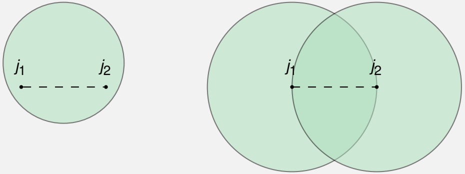
Converting aversion k-clustering instances for local k-median



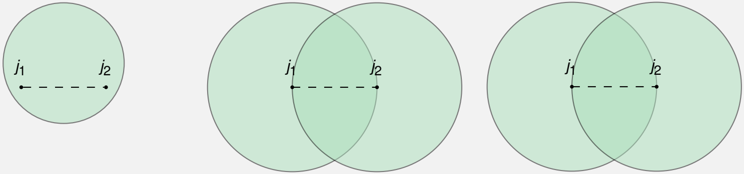
Converting aversion k-clustering instances for local k-median



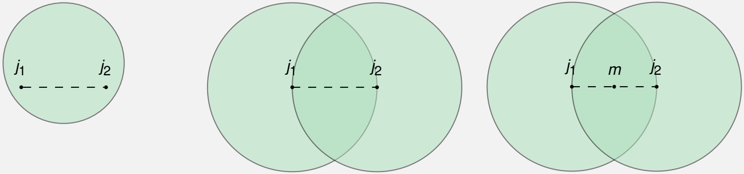
Converting aversion k-clustering instances for local k-median



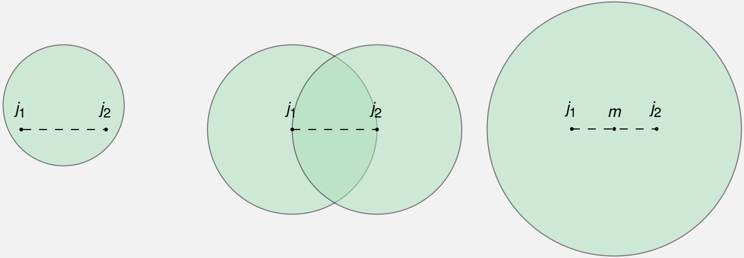
Converting aversion k-clustering instances for local k-median



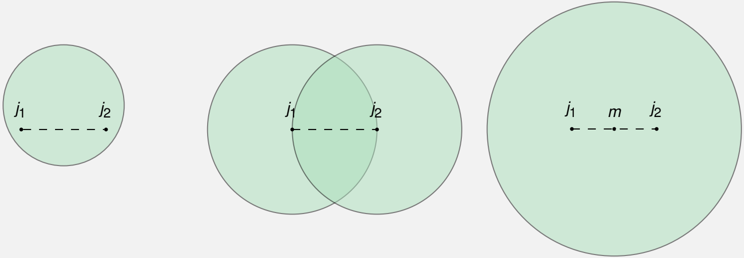
Converting aversion k-clustering instances for local k-median



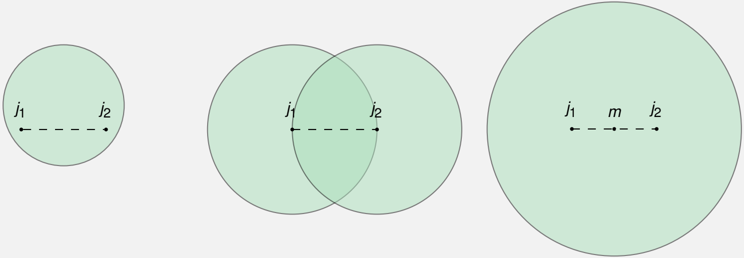
Converting aversion *k*-clustering instances for local *k*-median



Converting aversion k-clustering instances for local k-median

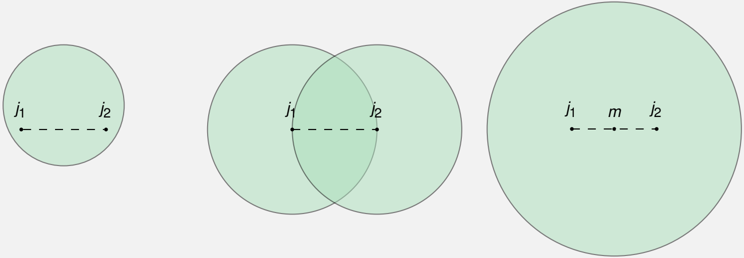


Converting aversion k-clustering instances for local k-median



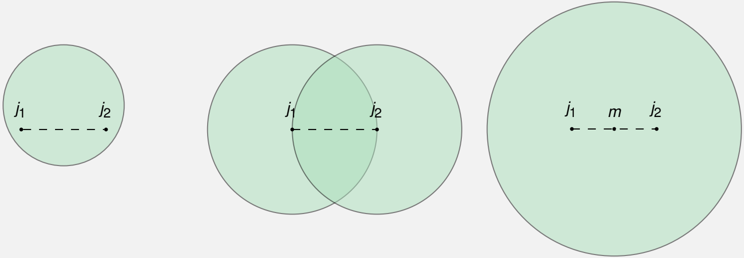
- Add midpoint for all $j_1, j_2 \in C$

Converting aversion k-clustering instances for local k-median



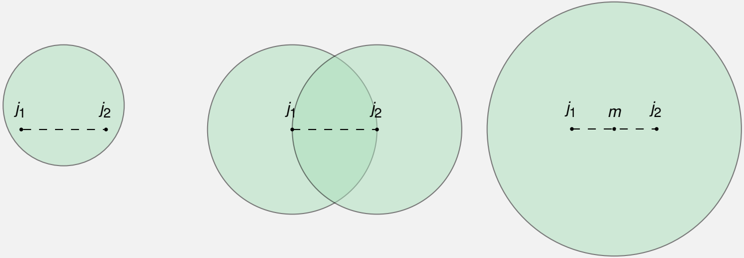
- Add midpoint for all $j_1, j_2 \in C$ $\rightsquigarrow$ facility, radius $\frac{3}{2}d(j_1, j_2)$

Converting aversion k-clustering instances for local k-median



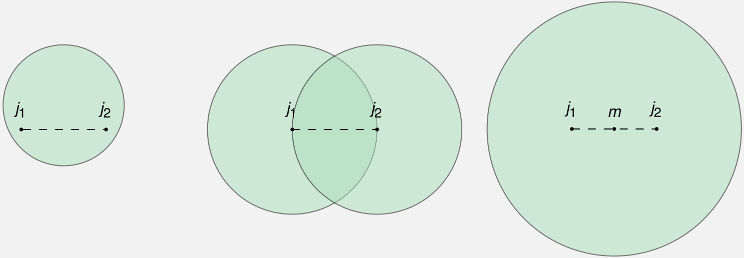
- Add midpoint for all $j_1, j_2 \in C$ $\rightsquigarrow$ facility, radius $\frac{3}{2}d(j_1, j_2)$
- Distance of j to new facility is $\min\{d(j, j_1), d(j, j_2)\} + \frac{1}{2}d(j_1, j_2)$

Converting aversion k-clustering instances for local k-median



- Add midpoint for all $j_1, j_2 \in C$ $\rightsquigarrow$ facility, radius $\frac{3}{2}d(j_1, j_2)$
- Distance of j to new facility is $\min\{d(j, j_1), d(j, j_2)\} + \frac{1}{2}d(j_1, j_2)$
- crucial property: distance always $\geq \frac{1}{2}d(j_1, j_2)$ and $\leq \frac{3}{2}d(j_1, j_2)$ by def.

Converting aversion k-clustering instances for local k-median



- Add midpoint for all $j_1, j_2 \in C \rightsquigarrow$ facility, radius $\frac{3}{2}d(j_1, j_2)$
- Distance of j to new facility is $\min\{d(j, j_1), d(j, j_2)\} + \frac{1}{2}d(j_1, j_2)$
- crucial property: distance always $\geq \frac{1}{2}d(j_1, j_2)$ and $\leq \frac{3}{2}d(j_1, j_2)$ by def.

Lemma

c-approximate solution to local k-median with radii violation c'

$$\Downarrow$$

$(3c' + 1) \cdot 2c \cdot OPT$ solution for aversion k-clustering

Bicriteria approximation for induced instances

Bicriteria approximation for induced instances

Aversion k-clustering instance

Bicriteria approximation for induced instances

Aversion k-clustering instance

$$\downarrow$$

Local k-median instance

Bicriteria approximation for induced instances

Aversion k-clustering instance

↓

Local k-median instance

↓

Primal-dual LMP framework by Jain/Vazirani [JV01]

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Bicriteria approximation for induced instances

Aversion k-clustering instance

↓

Local k-median instance

↓

Primal-dual LMP framework by Jain/Vazirani [JV01]

↓

Two solutions with $k_1 < k < k_2$ centers

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Bicriteria approximation for induced instances

Aversion k-clustering instance

↓

Local k-median instance

↓

Primal-dual LMP framework by Jain/Vazirani [JV01]

↓

Two solutions with $k_1 < k < k_2$ centers

Output

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Bicriteria approximation for induced instances

Aversion k-clustering instance

↓

Local k-median instance

↓

Primal-dual LMP framework by Jain/Vazirani [JV01]

↓

Two solutions with $k_1 < k < k_2$ centers

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Bicriteria approximation for induced instances

Aversion k-clustering instance

$\downarrow$

Local k-median instance

$\downarrow$

Primal-dual LMP framework by Jain/Vazirani [JV01]

$\downarrow$

Two solutions with $k_1 < k < k_2$ centers

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Bicriteria approximation for induced instances

Aversion k-clustering instance

$\downarrow$

Local k-median instance

$\downarrow$

Primal-dual LMP framework by Jain/Vazirani [JV01]

$\downarrow$

Two solutions with $k_1 < k < k_2$ centers

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$
- Then $\rho k_1 + (1 - \rho)k_2 = k$, and

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Bicriteria approximation for induced instances

Aversion k-clustering instance

$\downarrow$

Local k-median instance

$\downarrow$

Primal-dual LMP framework by Jain/Vazirani [JV01]

$\downarrow$

Two solutions with $k_1 < k < k_2$ centers

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$
- Then $\rho k_1 + (1 - \rho) k_2 = k$, and
- also $\rho \cdot cost(A) + (1 - \rho) \cdot cost(B) \leq 3 \cdot OPT$

[JV01] Jain, Vazirani. Approx. alg. for metric fac. loc. & k-Median pr. using the primal-dual schema and Lagr. relax., J. ACM 2001.

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$
- Then $\rho k_1 + (1 - \rho)k_2 = k$, and
- also $\rho \cdot cost(A) + (1 - \rho) \cdot cost(B) \leq 3 \cdot OPT$

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$
- Then $\rho k_1 + (1 - \rho)k_2 = k$, and
- also $\rho \cdot cost(A) + (1 - \rho) \cdot cost(B) \leq 3 \cdot OPT$

Fractional solution with good cost

Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$
- Then $\rho k_1 + (1 - \rho)k_2 = k$, and
- also $\rho \cdot cost(A) + (1 - \rho) \cdot cost(B) \leq 3 \cdot OPT$

Fractional solution with good cost $\rightsquigarrow$ Round it? Mix A and B?

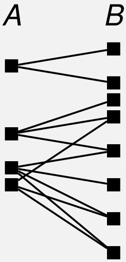
Output

- solution A, B, $k_1 < k < k_2$ centers, radii violation by factor 3
- Set $\rho = (k_2 - k)/(k_2 - k_1)$, it will be $\rho \ll 1/2$
- Then $\rho k_1 + (1 - \rho)k_2 = k$, and
- also $\rho \cdot cost(A) + (1 - \rho) \cdot cost(B) \leq 3 \cdot OPT$

Fractional solution with good cost $\rightsquigarrow$ Round it? Mix A and B?

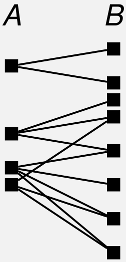
- Want k centers
- Centers from B are cheap, but B has $k_2 > k$ centers
- Can take $\approx \rho k_1$ centers from A
- Must assign all clients, respect radii!

Think of this as a graph



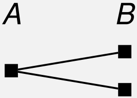
- means a client is connected to i_1 in A and i_2 in B
- edge can represent multiple clients ($\rightsquigarrow$ weight)

Think of this as a graph



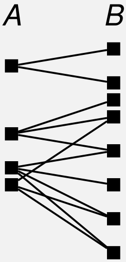
- means a client is connected to i_1 in A and i_2 in B
- edge can represent multiple clients ($\rightsquigarrow$ weight)

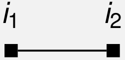
This part looks nice!



- open left center
- two centers from B can go!

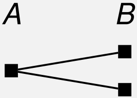
Think of this as a graph





- means a client is connected to i_1 in A and i_2 in B
- edge can represent multiple clients ($\rightsquigarrow$ weight)

This part looks nice!



- open left center
- two centers from B can go!

- B has k_2 centers $\rightsquigarrow$ Need to shutdown $k_2 - k$ centers
- If one center from A closes two from B, we need $k_2 - k$ A-centers
- cheapest $k_2 - k$ centers from A cost less than $\rho \cdot \text{cost}(A)$

Step 1: Turn graph into a tree

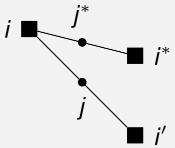
Step 1: Turn graph into a tree

- (i_1, i_2) is 'down' edge for i_1 if radius of i_2 is larger

Step 1: Turn graph into a tree

- (i_1, i_2) is 'down' edge for i_1 if radius of i_2 is larger

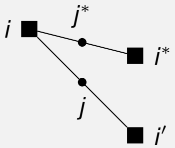
R increases ↓



Step 1: Turn graph into a tree

- (i_1, i_2) is 'down' edge for i_1 if radius of i_2 is larger
- all down edges become a single edge

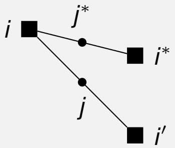
R increases $\downarrow$



Step 1: Turn graph into a tree

- (i_1, i_2) is 'down' edge for i_1 if radius of i_2 is larger
- all down edges become a single edge

R increases $\downarrow$



$d(j, i^*) \leq (R_i + R_i + R_{i^*}) \leq 3 \cdot R_{i^*}$

(violation of R_{i^*} by factor 3)

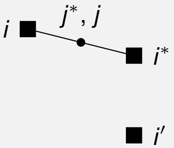
also: $3 \cdot R_{i^*} \leq 3 \cdot R_{i'}$

thus cost goes up by a factor of ≤ 3

Step 1: Turn graph into a tree

- (i_1, i_2) is 'down' edge for i_1 if radius of i_2 is larger
- all down edges become a single edge

R increases $\downarrow$



$d(j, i^*) \leq (R_i + R_i + R_{i^*}) \leq 3 \cdot R_{i^*}$

(violation of R_{i^*} by factor 3)

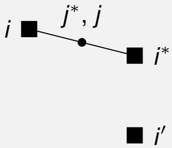
also: $3 \cdot R_{i^*} \leq 3 \cdot R_{i'}$

thus cost goes up by a factor of ≤ 3

Step 1: Turn graph into a tree

- (i_1, i_2) is 'down' edge for i_1 if radius of i_2 is larger
- all down edges become a single edge

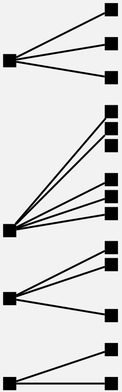
R increases $\downarrow$



$d(j, i^*) \leq (R_i + R_i + R_{i^*}) \leq 3 \cdot R_{i^*}$
(violation of R_{i^*} by factor 3)
also: $3 \cdot R_{i^*} \leq 3 \cdot R_{i'}$
thus cost goes up by a factor of ≤ 3

- In new graph, no center has two down edges
- $\rightsquigarrow$ graph is a forest
  - Assume there is a cycle. Start at some client i.
  - Follow up edge. Is down edge for the other endpoint.
  - Can only go up and never get back to i.

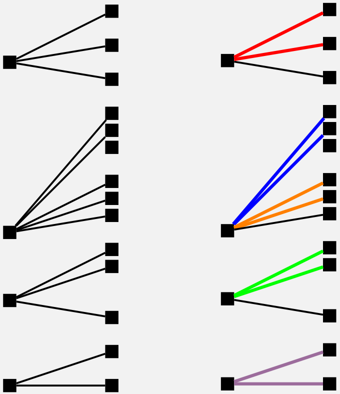
Step 2: Find enough suitable subtrees



$k_2 - k_1 = 10$

Suitable: Trees with exactly one more facility from B than from A

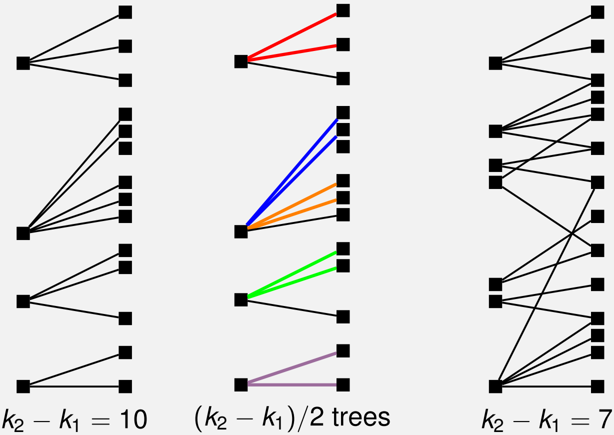
Step 2: Find enough suitable subtrees



$k_2 - k_1 = 10$ $(k_2 - k_1)/2$ trees

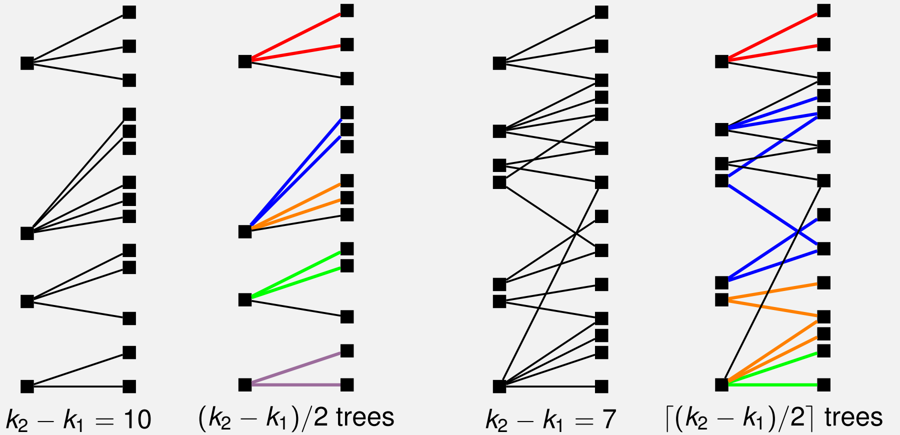
Suitable: Trees with exactly one more facility from B than from A

Step 2: Find enough suitable subtrees



Suitable: Trees with exactly one more facility from B than from A

Step 2: Find enough suitable subtrees



Suitable: Trees with exactly one more facility from B than from A

Main Technical Lemma: Can find $\frac{1}{2}(k_2 - k_1)$ trees T_i with ...

- with $|A(T_i)| = |B(T_i)| - 1$
- $B(T_i) \cap B(T_j) = \emptyset$
- $A(T_i)$ covers all edge in $\delta(B(T_i))$,
  i.e. opening $A(T_i)$ means we do not need $B(T_i)$

Main Technical Lemma: Can find $\frac{1}{2}(k_2 - k_1)$ trees T_i with ...

- with $|A(T_i)| = |B(T_i)| - 1$
- $B(T_i) \cap B(T_j) = \emptyset$
- $A(T_i)$ covers all edge in $\delta(B(T_i))$,
  i.e. opening $A(T_i)$ means we do not need $B(T_i)$

- need to close $k_2 - k$ centers from B
- choose $k_2 - k$ cheapest trees with respect to cost in A
- costs $(k_2 - k)/(\frac{1}{2}(k_2 - k_1)) \cdot cost(A) = 2\rho \cdot cost(A)$

Main Technical Lemma: Can find $\frac{1}{2}(k_2 - k_1)$ trees T_i with ...

- with $|A(T_i)| = |B(T_i)| - 1$
- $B(T_i) \cap B(T_j) = \emptyset$
- $A(T_i)$ covers all edge in $\delta(B(T_i))$,
  i.e. opening $A(T_i)$ means we do not need $B(T_i)$

- need to close $k_2 - k$ centers from B
- choose $k_2 - k$ cheapest trees with respect to cost in A
- costs $(k_2 - k)/(\frac{1}{2}(k_2 - k_1)) \cdot cost(A) = 2\rho \cdot cost(A)$

$\rightsquigarrow$ integral solution with k centers and good cost

Theorem

Theorem

- For any local k-median instance

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

$\Rightarrow$ $\mathcal{O}(1)$-approximation for the aversion k-clustering problem.

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

$\Rightarrow$ $\mathcal{O}(1)$-approximation for the aversion k-clustering problem.

Open questions

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

$\Rightarrow \mathcal{O}(1)$-approximation for the aversion k-clustering problem.

Open questions

- True bicriteria approximation for local k-median?

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

$\Rightarrow$ $\mathcal{O}(1)$-approximation for the aversion k-clustering problem.

Open questions

- True bicriteria approximation for local k-median?
- Locality constraints for other clustering problems

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

$\Rightarrow$ $\mathcal{O}(1)$-approximation for the aversion k-clustering problem.

Open questions

- True bicriteria approximation for local k-median?
- Locality constraints for other clustering problems
- Applying non-uniform rounding technique to other objectives

Theorem

- For any local k-median instance
- arising from the aversion k-clustering problem
- we can find an integral solution
- that is a constant bicriteria approximation.

$\Rightarrow$ $\mathcal{O}(1)$-approximation for the aversion k-clustering problem.

Open questions

- True bicriteria approximation for local k-median?
- Locality constraints for other clustering problems
- Applying non-uniform rounding technique to other objectives

Thank you for your attention!