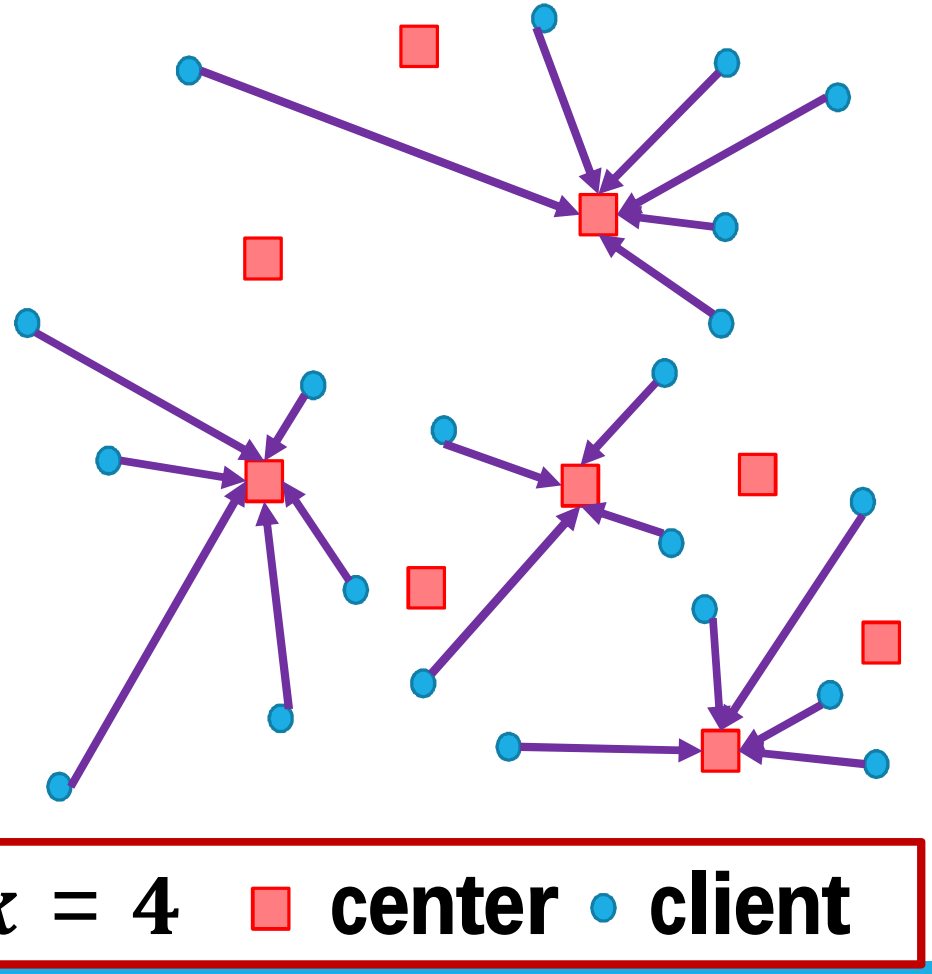
Approximation Algorithms for Clustering Problems with Lower Bounds and Outliers

SARA AHMADIAN, UNIVERSITY OF WATERLOO

JOINT WORK WITH CHAITANYA SWAMY

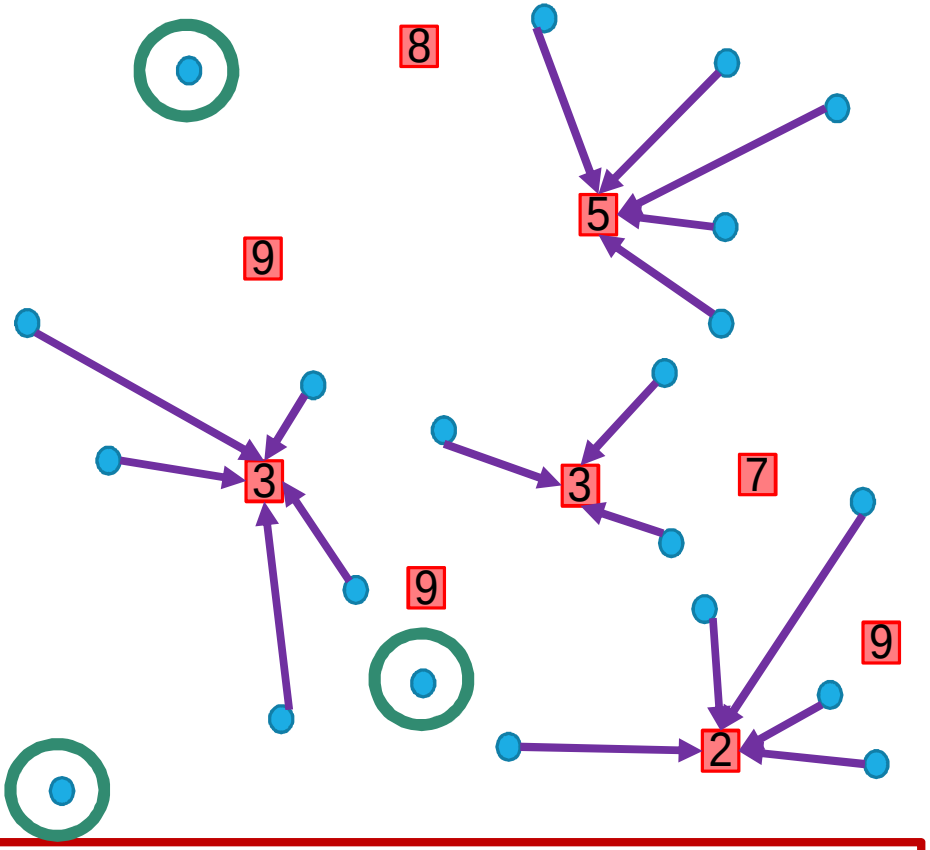
Clustering points into k groups

- Input:
  - Set $\mathcal{D}$ of clients
  - Set $\mathcal{F}$ of centers
  - Metric c on $\mathcal{D} \cup \mathcal{F}$
  - Bound k



Clustering points into k groups with lower-bounds and outliers

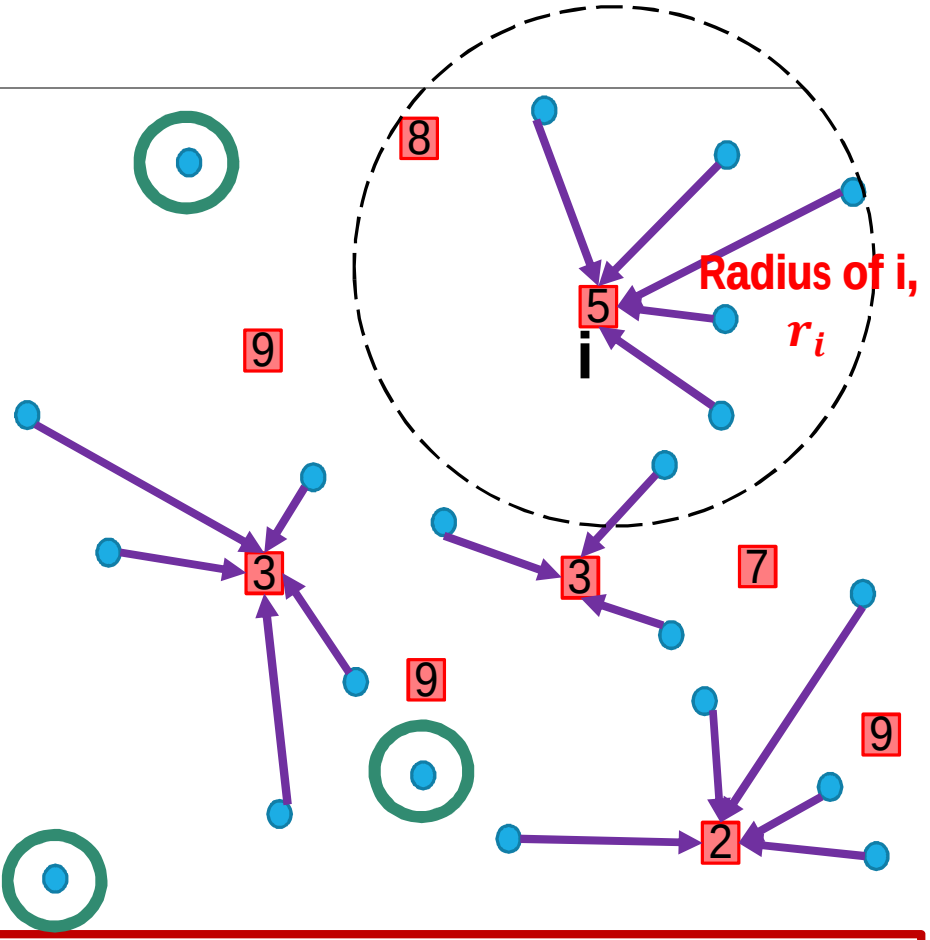
- Input:
  - Set $\mathcal{D}$ of clients
  - Set $\mathcal{F}$ of centers with lower-bounds L_i
  - Metric c on $\mathcal{D} \cup \mathcal{F}$
  - Bound k and m
- Feasible solution $S \subseteq \mathcal{F}, |S| \leq k$ and $\sigma: \mathcal{D} \to S \cup \{out\}, |\{j: \sigma(j) = out| \leq m, |\sigma^{-1}(i)| \geq L_i$ for each $i \in S$.



$k = 4, m = 3$ center client outlier

Clustering points into k groups with lower-bounds and outliers

- Input:
  - Set $\mathcal{D}$ of clients
  - Set $\mathcal{F}$ of centers with lower-bounds
  - Metric c on $\mathcal{D} \cup \mathcal{F}$
  - Bound k and m
- Feasible solution $S \subseteq \mathcal{F}, |S| \leq k$ and $\sigma: \mathcal{D} \to S \cup \{out\}, |\{\sigma^{-1}(out)| \leq m,$ $|\sigma^{-1}(i)| \geq L_i$ for each $i \in S$.
- Objective function:
  - **Lower-bounded k-supplier:** $min\ max_{i \in S}\ r_i$
  - **Lower-bounded min-sum-of-radii:** $min \sum_{i \in S} r_i$



Motivation

❖ Why Lower-Bounds (LBs)?

❑ Anonymity:

- ❑ Samarati[2001]: perturb attributes of data points and cluster, each cluster has $\geq L$ identical perturbed data points ⇒ **difficult to identify original data!**
- ❑ Aggrawal et al.[2005, 2010]: this can be abstracted as LB clustering where clustering objective captures the cost of perturbing data.

❑ Facility Location:

- ❑ LB models infeasible or unprofitable to use services unless they satisfy certain min demand.

❖ Why Outliers?

- ❑ Ignore data points that are quite dissimilar and degrade the quality of any clustering of entire set. Focus of data points of interest.

Our results:

- Lower-bounded k-supplier (LBkSup): $\boldsymbol{min\ max_{i \in S}\ r_i}$
  - First result for non-uniform LBs, 3 and 5-app for non-outlier and outlier
  - Previously: uniform LBs 2 and 4-app when $\mathcal{F} = \mathcal{D}$ by Aggrawal et al[2010], factor-3 hardness for non-outlier ($\mathcal{F} \neq \mathcal{D}$) by Charkiar et al[2001]
  - Main technique: use notion of skeleton introduced by Cygan et al[2014] for cap k-center with outliers.
- Lower-bounded min-sum-of-radii (LBkSR): $\boldsymbol{min \sum_{i \in S}\ r_i}$
  - First result for either LBs or outlier, 3.8 and 12.3-app for non-outlier and outlier
  - Previously, 3.53-app for non-outlier (no LBs) by Charikar et al[2010]
  - Main technical contribution: For outlier version, consider Lagrangian relaxation, even though not an LMP algorithm, constant factor approximation!

Related work

- Incorporate LB and outliers:
  - Aggrawal et al[2010]: 4 and 2-app when $\mathcal{F} = \mathcal{D}$ and uniform LBs.
- Sum of radii:
  - Doddi et al[2000] introduced the problem, Charikar and Panigrahi[2001]: 3.53-app
  - $\mathcal{F} = \mathcal{D}$, QPTAS by Gibson et al[2010].
- Facility location with Outlier or LBs:
  - Charikar et al[2010]: outlier version of uncap FL, k-supplier, k-median.
  - Chen[2008]: k-median with outliers
  - Cygan and Kociumaka[2014]: Cap k-center with outliers
  - Ene et al[2013]: LBkSup in Euclidean space

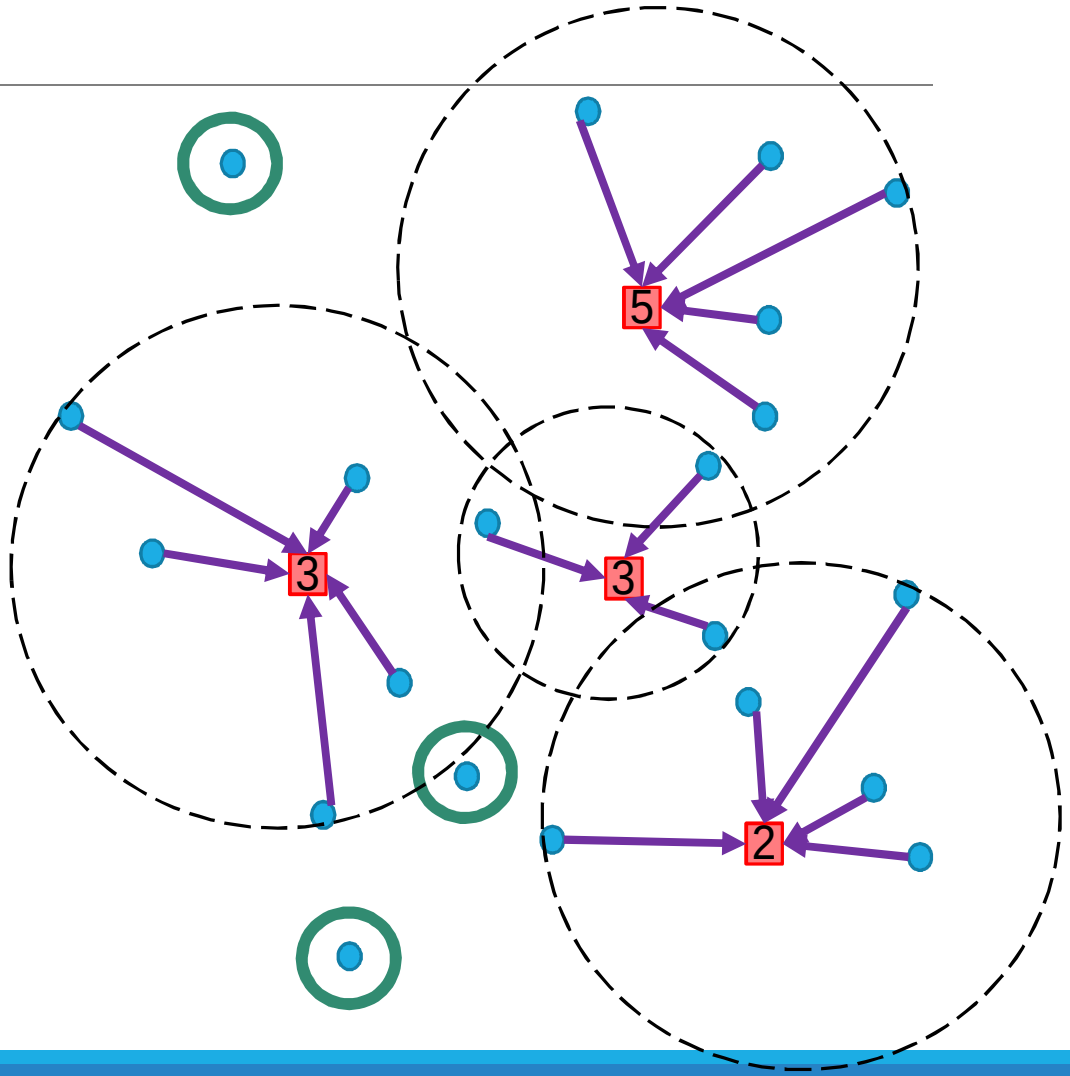
Lower-bounded k-sum-of-radii with outliers (LBkSRO)

Main ideas for LBkSRO

- Reduction to Ball-selection problem
- Lagrangian relaxation of ball-selection problem, and Design a Primal-Dual Algorithm
- Binary search for opening cost and adequate combination routines

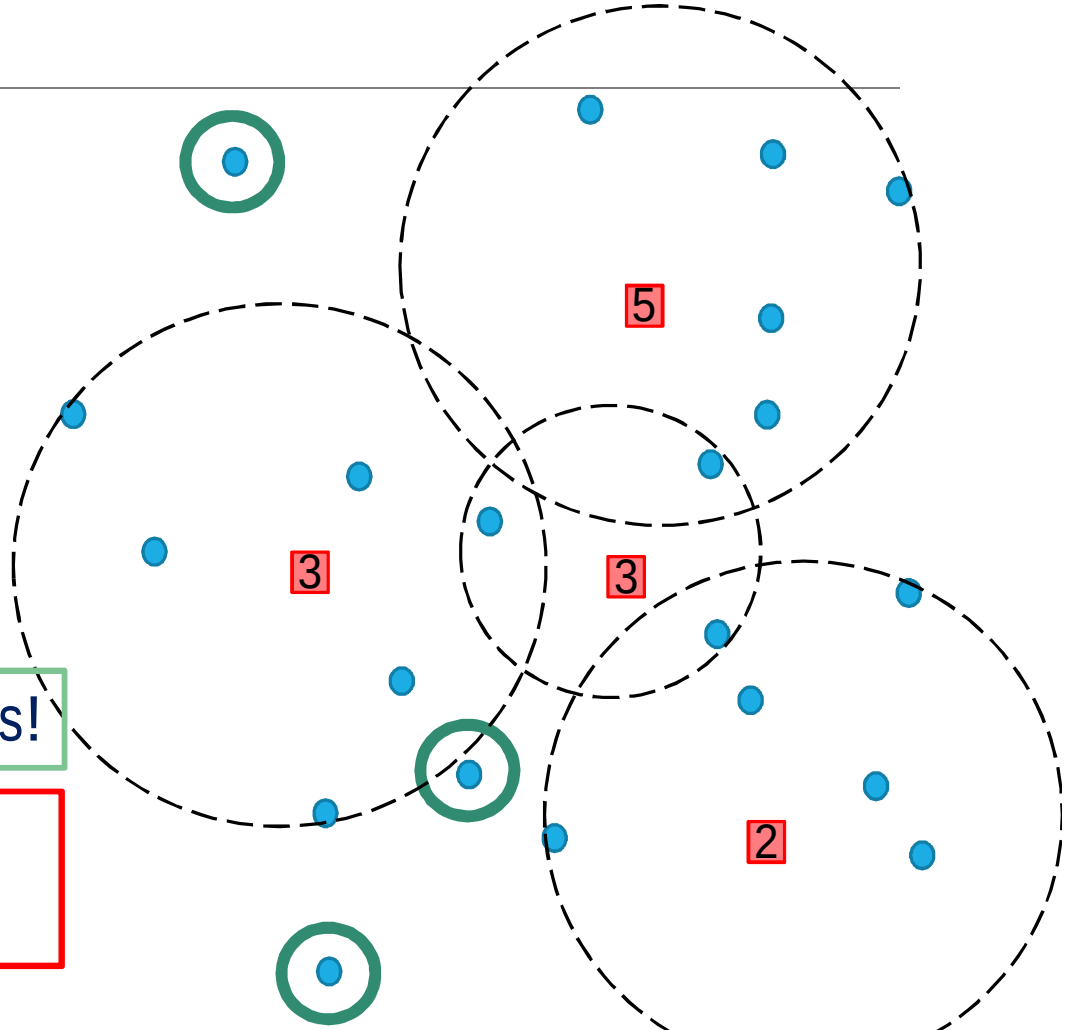
Reduction to k-ball selection problem

- Create list $\mathcal{L}$ of balls:
  - $B(i, r)$: ball with center i, radius r.
  - Consider B(i, r) in $\mathcal{L}$ if $|B(i, r)| \geq L_i$
  - Goal: pick k balls such that there are $\leq m$ clients not covered while minimizing total radii.



Reduction to k-ball selection problem

- Create list $\mathcal{L}$ of balls:
  - Consider B(i, r) in $\mathcal{L}$ if $|B(i, r)| \geq L_i$
  - Goal: pick k balls such that there are $\leq m$ clients not covered while minimizing total radii.



Advantage: No worries about LBs!

Issue: Balls may overlap, how to translate to LBkSRO solution?

LP formulation for k-BS

$y_{i,r}$: $B(i,r)$ is picked, w_j: Client j is outlier. Only consider balls with $r \leq R^*$ our guess of max radius in optimal solution.

$$\min \sum_{(i,r)\in\mathcal{L}} r \cdot y_{i,r} + \sum_{(i,r)\in\mathcal{L}} z \cdot y_{i,r}$$

$$\text{s.t.} \sum_{(i,r):j\in B(i,r)} y_{i,r} + w_j \geq 1 \quad \forall j$$

$$\sum_j w_j \leq m$$

~~$$\sum_{(i,r)} y_{i,r} \leq k$$~~

$$y, w \geq 0$$

LP formulation for k-BS

$y_{i,r}$: $B(i,r)$ is picked, w_j: Client j is outlier.

Only consider balls with $r \leq R^*$.

$$\min \sum_{(i,r)\in\mathcal{L}} (r + z) \cdot y_{i,r}$$

$$\text{s.t.} \sum_{(i,r):j\in B(i,r)} y_{i,r} + w_j \geq 1 \quad \forall j$$

$$\sum_j w_j \leq m$$

$$y, w \geq 0$$

.

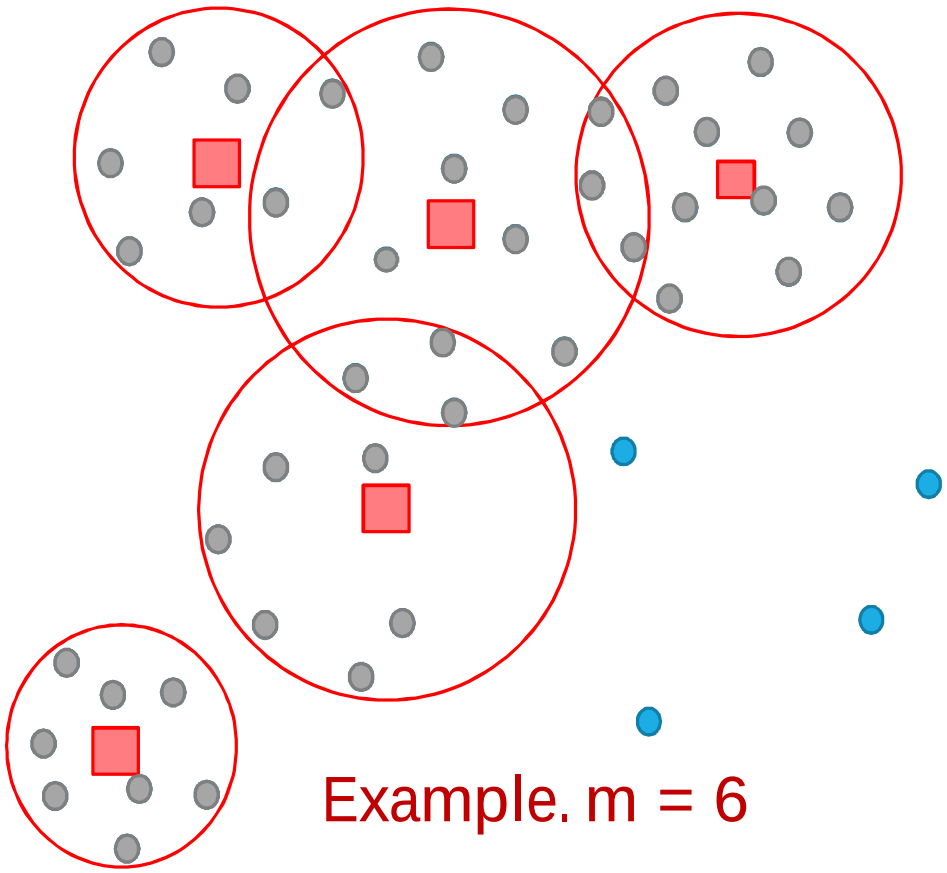
$$\max \sum_{j\in\mathcal{D}} \alpha_j - m\gamma$$

$$\text{s.t.} \sum_{j\in B(i,r)} \alpha_j \leq r + z \quad \forall (i,r)$$

$$\alpha_j \leq \gamma$$

$$\alpha, \gamma, z \geq 0$$

Primal Dual Algorithm



Example. m = 6

$$\max \sum_{j \in \mathcal{D}} \alpha_j - m\gamma$$

$$\text{s.t.} \sum_{j \in B(i,r)} \alpha_j \leq r + z \quad \forall (i,r)$$

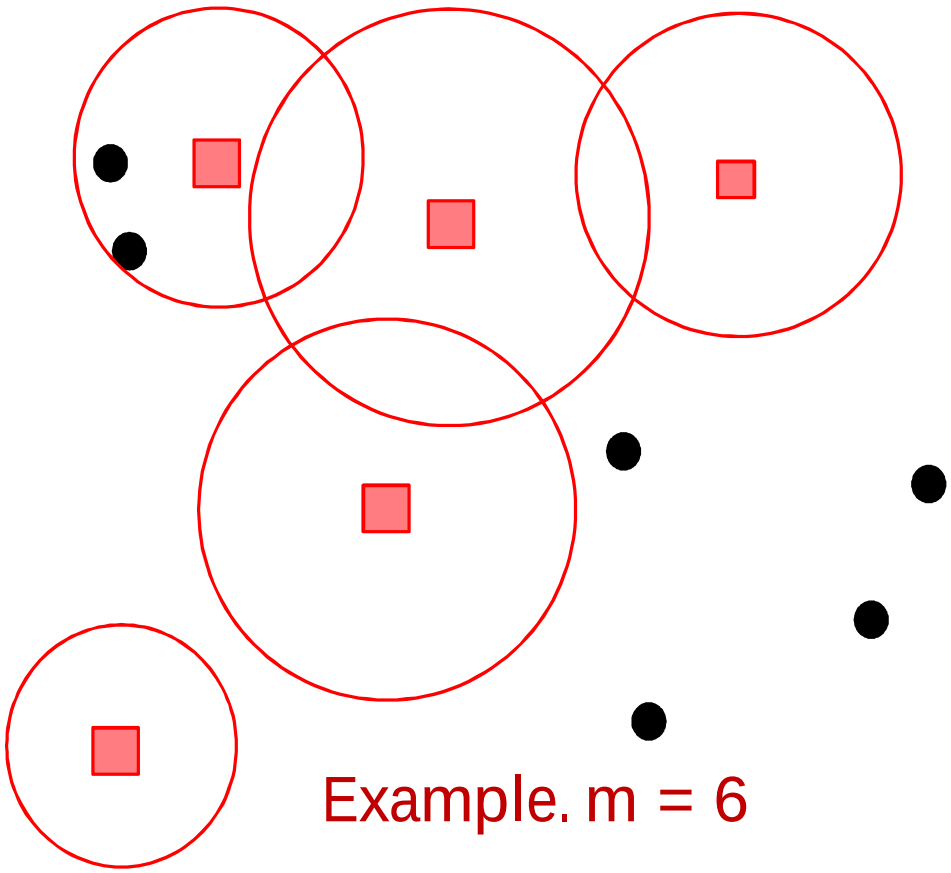
$$\alpha_j \leq \gamma$$

$$\alpha, \gamma, z \geq 0$$

Dual Ascent Phase

1. $\alpha = 0$, all clients are active.
2. Inc α_j active clients until ★ becomes tight for some (i,r)
3. Freeze clients in tight ball
4. Repeat until there are at most m active clients.

Primal Dual Algorithm



$$\max \sum_{j \in \mathcal{D}} \alpha_j - m\gamma$$

$$\text{s.t.} \sum_{j \in B(i,r)} \alpha_j \leq r + z \quad \forall (i,r)$$

$$\alpha_j \leq \gamma$$

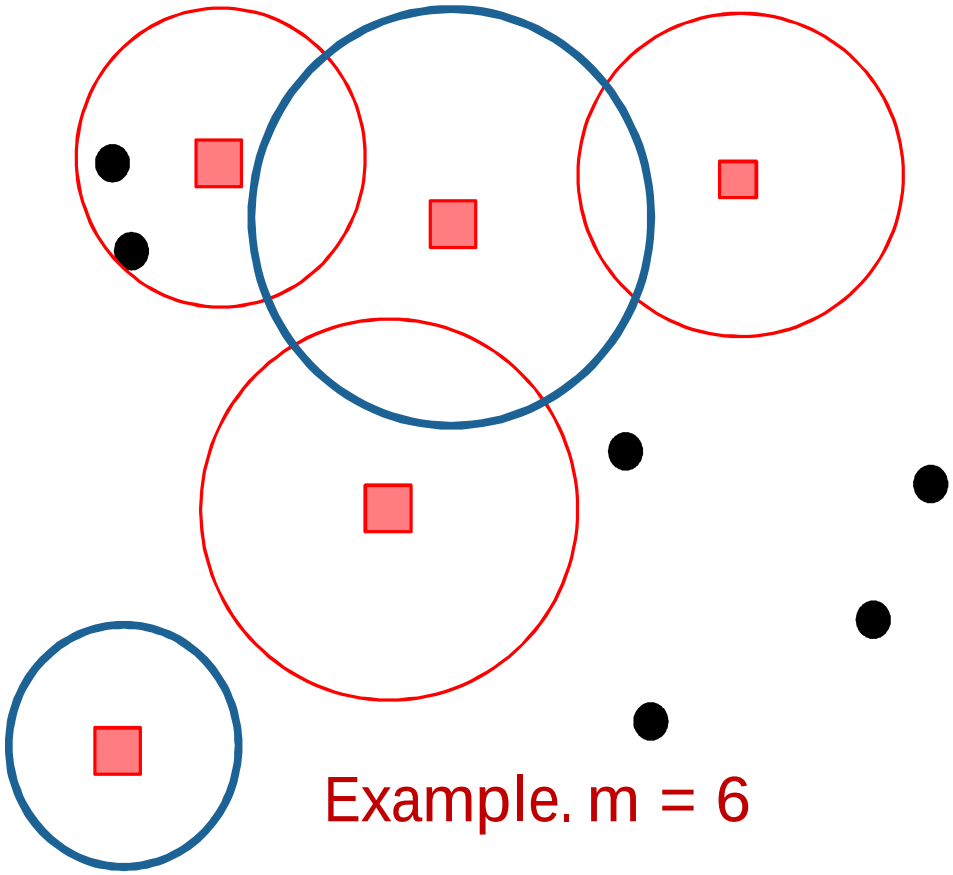
$$\alpha, \gamma, z \geq 0$$

Dual Ascent Phase

1. $\alpha = 0$, all clients are active.
2. Inc α_j active clients until ★ becomes tight for some (i,r)
3. Freeze clients in tight ball
4. Repeat until there are at most m active clients.
5. Declare m of lastly active clients outlier!
6. Set $\gamma = \max \alpha_j$

Primal Dual Algorithm

Last ball f



Example. m = 6

$$\max \sum_{j \in \mathcal{D}} \alpha_j - m\gamma$$

$$\text{s.t.} \sum_{j \in B(i,r)} \alpha_j \leq r + z \quad \forall (i,r)$$

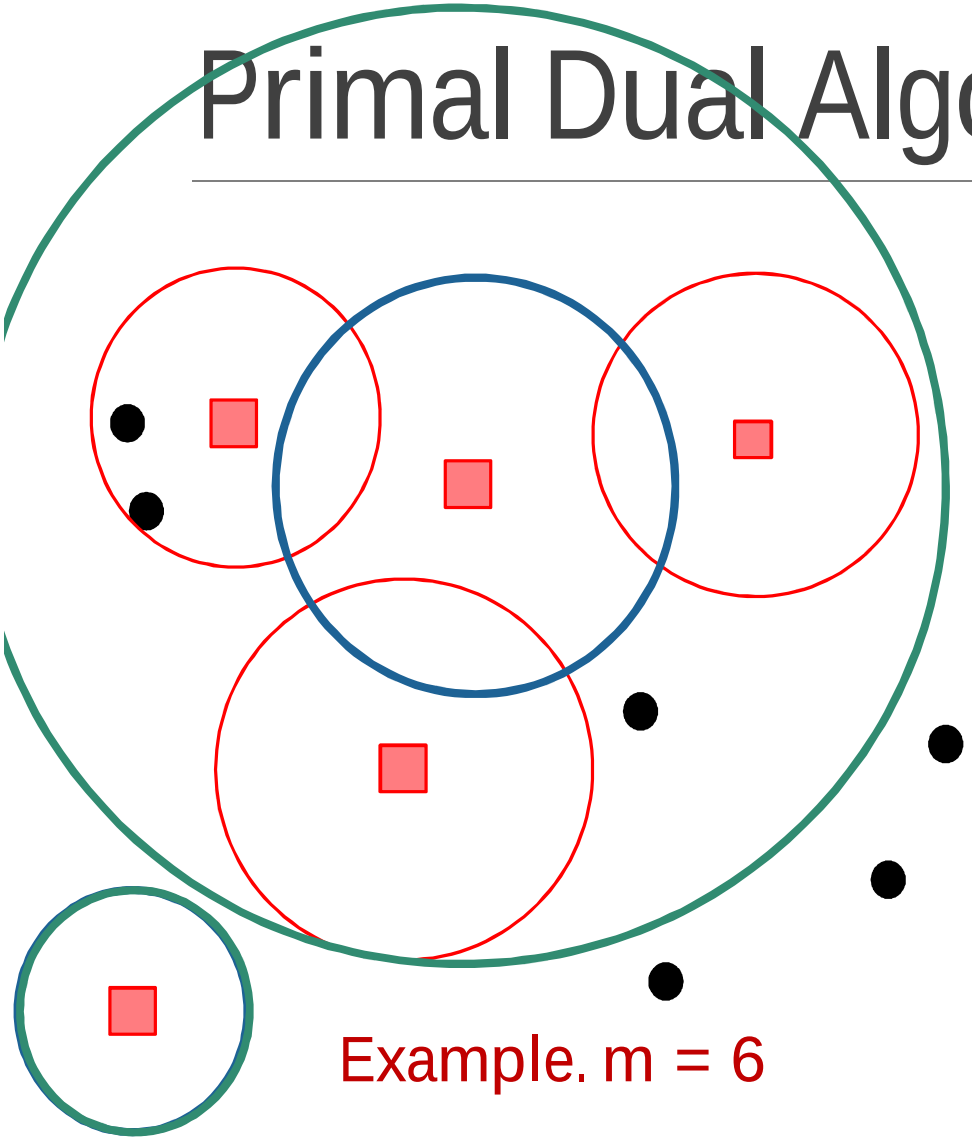
$$\alpha_j \leq \gamma$$

$$\alpha, \gamma, z \geq 0$$

Pruning Phase

1. Pick a maximal non-intersecting subset T_I of tight sets (largest radius first)preferably not containing f.

Primal Dual Algorithm



Example. m = 6

$$\max \sum_{j \in \mathcal{D}} \alpha_j - m\gamma$$

$$\text{s.t.} \sum_{j \in B(i,r)} \alpha_j \leq r + z \quad \forall (i,r)$$

$$\alpha_j \leq \gamma$$

$$\alpha, \gamma, z \geq 0$$

Pruning Phase

1. Pick a maximal non-intersecting subset T_I of tight balls (largest radius first).
2. Expand radius of each $(i,r) \in T_I$ (by at most $2r$) to cover clients of intersecting smaller balls.
3. Drop f if cost increase is $O(R^*)$.

Our Primal-Dual is not LMP!

Although we try to avoid picking the last ball f, in some circumstances, last ball is far from other tight balls, we have to pick it!

Theorem. Solution F of our PD algorithm, f is the last tight ball:

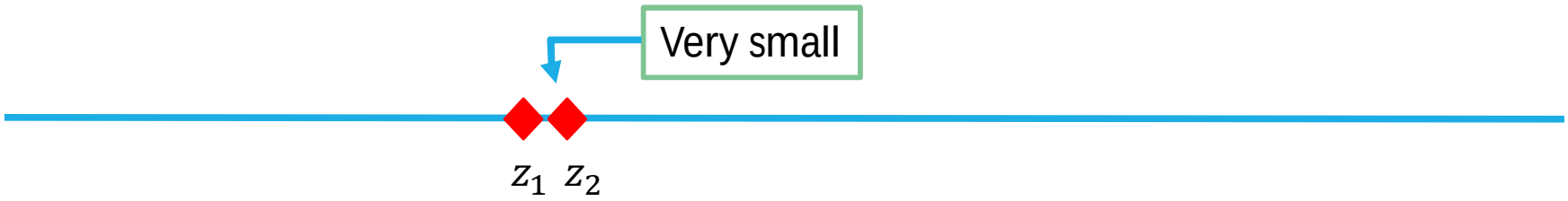
$$\sum_{(i,r)\in F\setminus \mathrm{f}} (r + 3z) \leq 3(OPT + kz) + O(R^*)$$

Not LMP as we can't pay for 3z cost of last facility f! Radius of f is at most R^* by filtering in LP formulation. --> Even if a solution with k facilities, if it includes f, cost(F) $\leq 3OPT + O(R^*) + 3z$

Can only say :

1. If $|F\setminus \mathrm{f}| \geq k$ then, $cost(F) \leq O(OPT)$
2. Either have to bound z or found some other way!

Binary search



Solution F_1 of size $\geq k + 1$ Solution F_2 of size $\leq k$

Combination subroutine

Theorem. Solution F of our PD algorithm, f is the last tight ball:

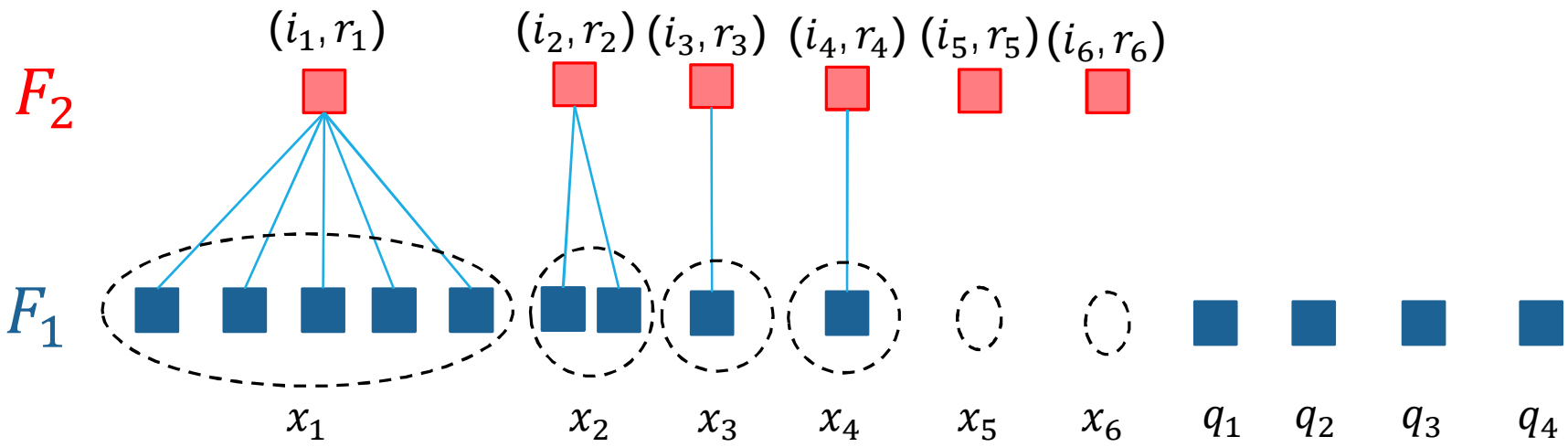
$$\sum_{(i,r)\in F} r \leq 3OPT + 3(k - |F \backslash \mathrm{f}|)z + O(R^*)$$

Two possible combination subroutines depending on

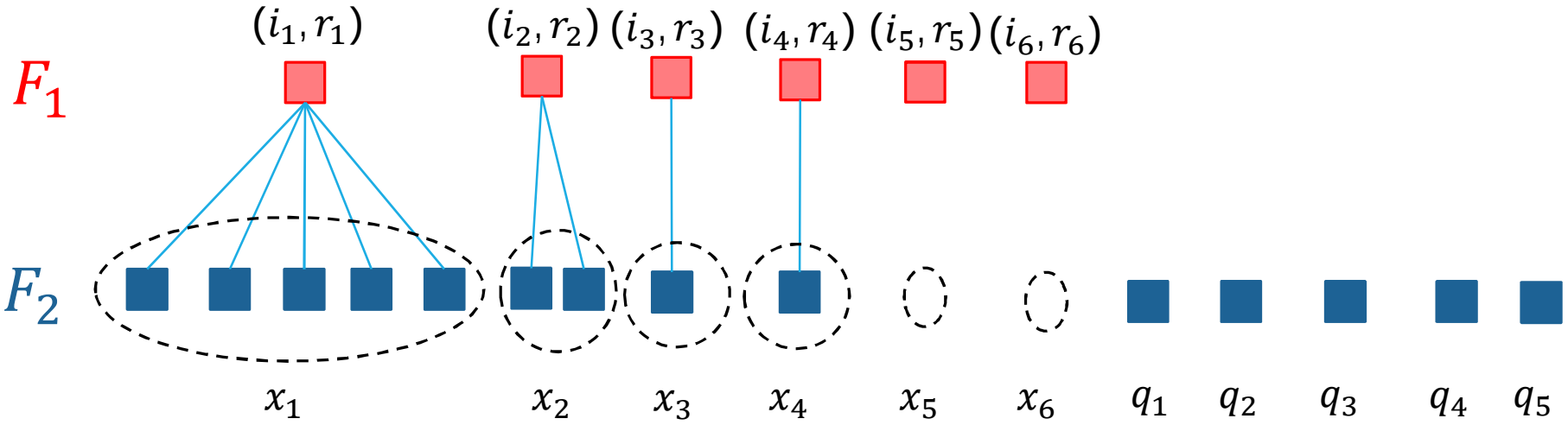
- $z \leq OPT$: Combination subroutine A
- $z > OPT$: Combination subroutine B, Bigger solution F_1 has size $k + 1$ and includes last tight facility.

Combination subroutine A: $z \leq OPT$

Construct graph, put an edge from $(i, r) \in F_1$ to some intersecting $(i', r') \in F_2$.



Combination subroutine A: $z \leq OPT$



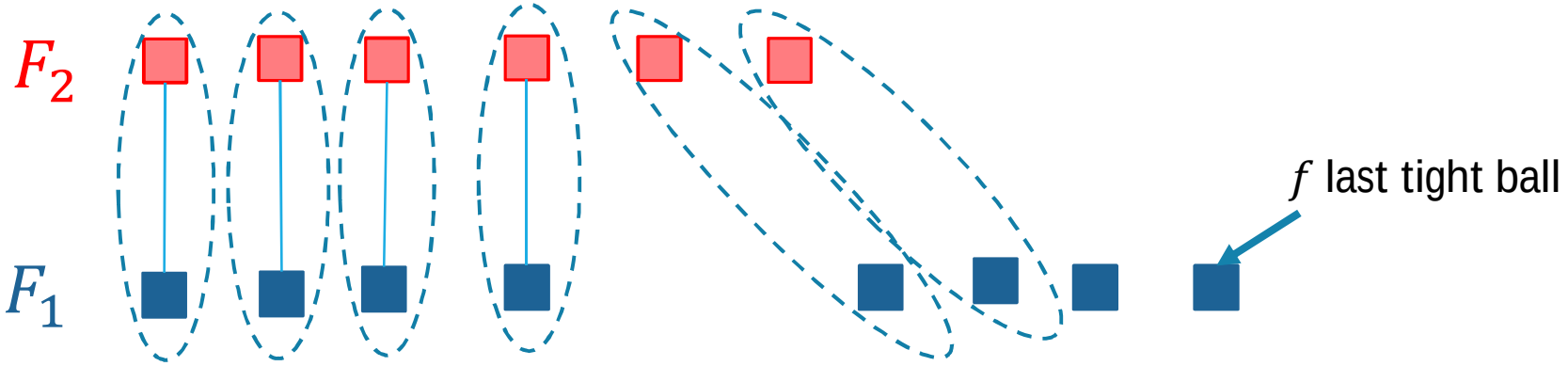
Covering Knapsack LP,

1. Only two constraints and some bound constraints -> Optimal solution hast $\leq$ 2 fractional components
2. Optimal value of LP is $O(OPT)$ -> fractional combination of F_1 and F_2.

Combination subroutine B: $z > OPT$

Solution F_1 has size at least k + 1 and includes last tight facility f, $cost(F_1) \leq 3OPT + O(R^*)$.

1. Can assume F_1 balls are far otherwise can merge two balls.
2. Can assume each F_2-ball intersects with at most one ball in F_1 -> can define a mapping ψ from F_2 to $F_1 \setminus$f

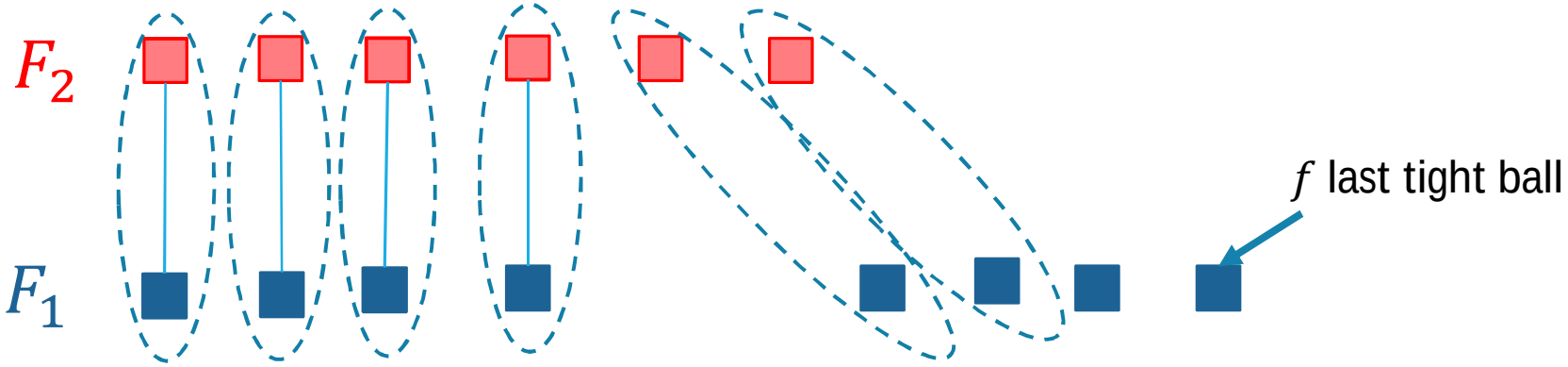


Idea: (a) Balls in F_1 and F_2 are almost tight by Charikar et al
(b) Solution $F = (F_1 \cup A) \setminus \psi(A)$ and j $\in uncov(F)$: $\alpha_j \approx \gamma$, m clients with $\alpha_j \approx \gamma$

$$\boldsymbol{cost(F) = O(OPT)}$$

Combination subroutine B: $z > OPT$

Solution F_1 has size at least k + 1 and includes last tight facility f, $cost(F_1) \leq O(OPT)$.



Idea: Solution $F = (F_1 \cup A) \setminus \psi(A)$ and j ∈ $uncov(F)$: $\alpha_j \approx \gamma$, m clients with $\alpha_j \approx \gamma$

$$\boldsymbol{cost(F) = O(OPT)}$$

Start with $F_1 \setminus$ f:

- Swap in F_2-ball and swap out its mapped F_1-ball.
- Either at some point, realize f is not needed -> Solution of size k with good cost!
- Or at the end of process can show F_2 has a good cost!

Summary and Open Questions

- Chen's algorithm for k-median with outliers encounter similar difficulties in spirit
  - Can we use our idea for that problem?
- Better approximation ratio for LBkSR, LBkSRO, LBkSupO?
- Better hardness results for LBkSR/LBkSRO?

Thank You

QUESTIONS?