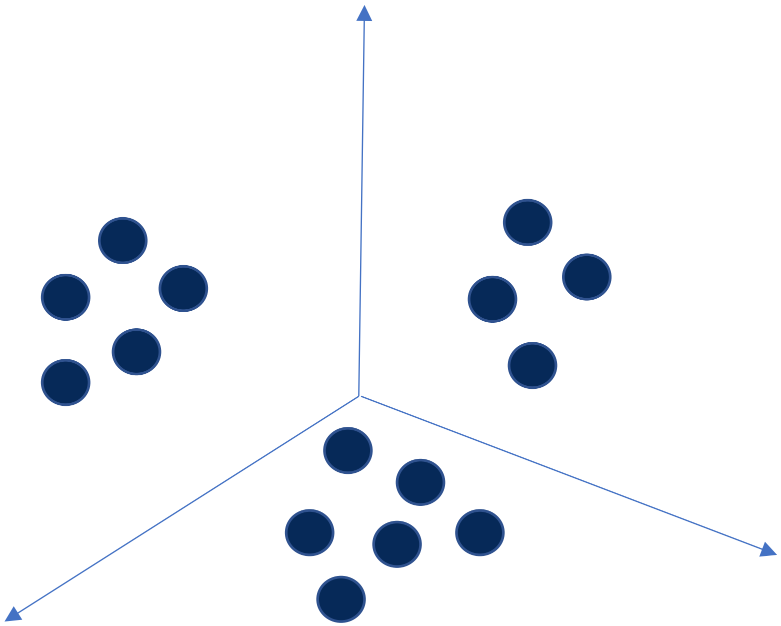
Fair and Efficient Social Decision-Making

CSCI 699

Fairness in Clustering

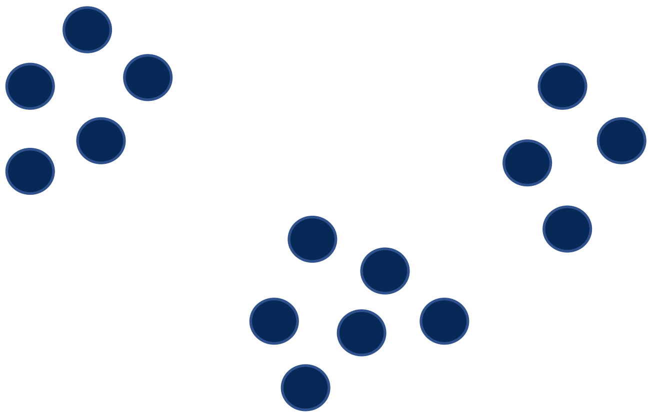
Evi Micha

Clustering



Clustering in ML/Data Analysis

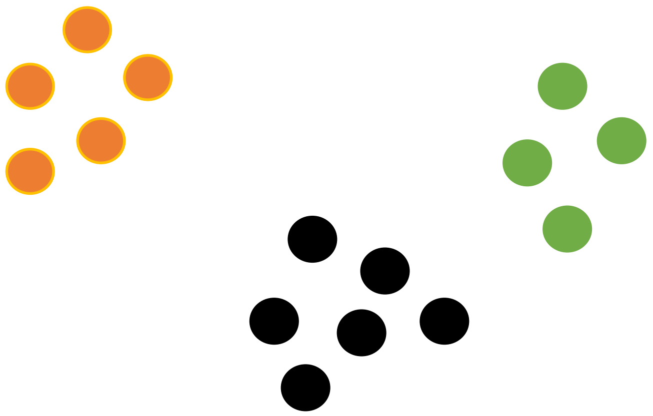
- Goal:
  - Analyze data sets to summarize their characteristics
  - Objects in the same group are similar



Clustering in ML/Data Analysis

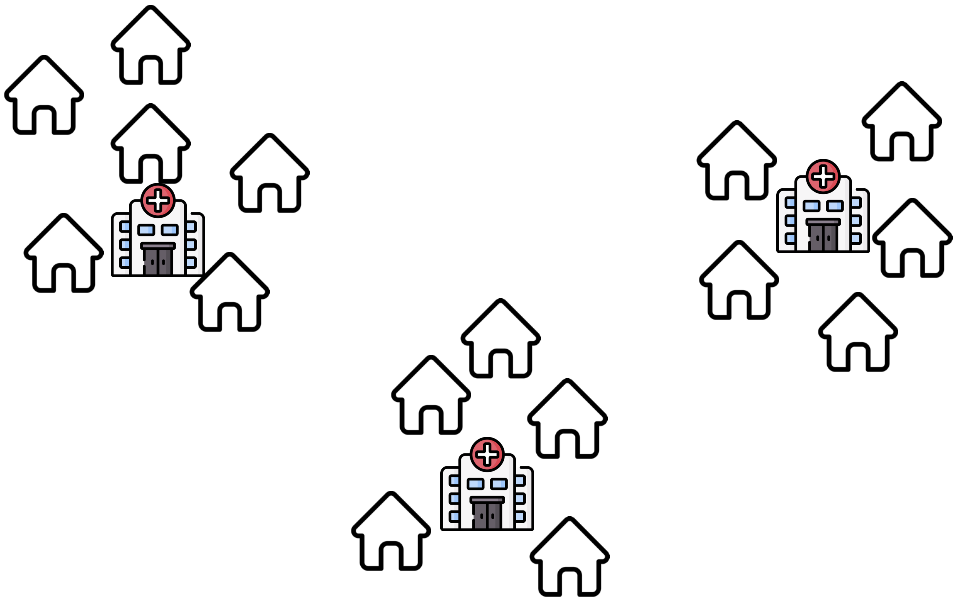
- Goal:
  - Analyze data sets to summarize their characteristics
  - Objects in the same group are similar

k=3



Clustering in Economics/OR

- Goal:
  - Allocate a set of facilities that serve a set of agents (e.g. hospitals)



Centroid Clustering

Center-Based Clustering

- Input:
  - Set N of n data points
  - Set M of m feasible cluster centers
  - $\forall i, j \in N \cup M$: we have $d(i,j)$ (which forms a ***Metric Space***)
    - $\boldsymbol{d(i,i) = 0}$, $\forall\, i \in N \cup M$
    - $\boldsymbol{d(i,j) = d(j,i)}$, $\forall\, i,j \in N \cup M$
    - $\boldsymbol{d(i,j) \leq d(i,\ell) + d(\ell,j)}$, $\forall\, i,j,\ell \in N \cup M$, (**Triangle Inequality**)
- Output:
  - A set $C \subseteq M$ of k centers, i.e. $C = \{c_1, \dots, c_k\}$
  - Each data point is assigned to its closest cluster center
    - $C(i) = argmin_{c \in C}\, d(i,c)$

Famous Objective Functions

- k-median: Minimizes the sum of the distances
  - $\min_{\substack{C \subseteq M: \\ |C| \leq k}} \sum_{i \in N} d(i, C(i))$
- k-means: Minimizes the sum of the square of the distances
  - $\min_{\substack{C \subseteq M: \\ |C| \leq k}} \sum_{i \in N} d^2(i, C(i))$
- k-center: Minimizes the maximum distance
  - $\min_{\substack{C \subseteq M: \\ |C| \leq k}} \max_{i \in N} d(i, C(i))$

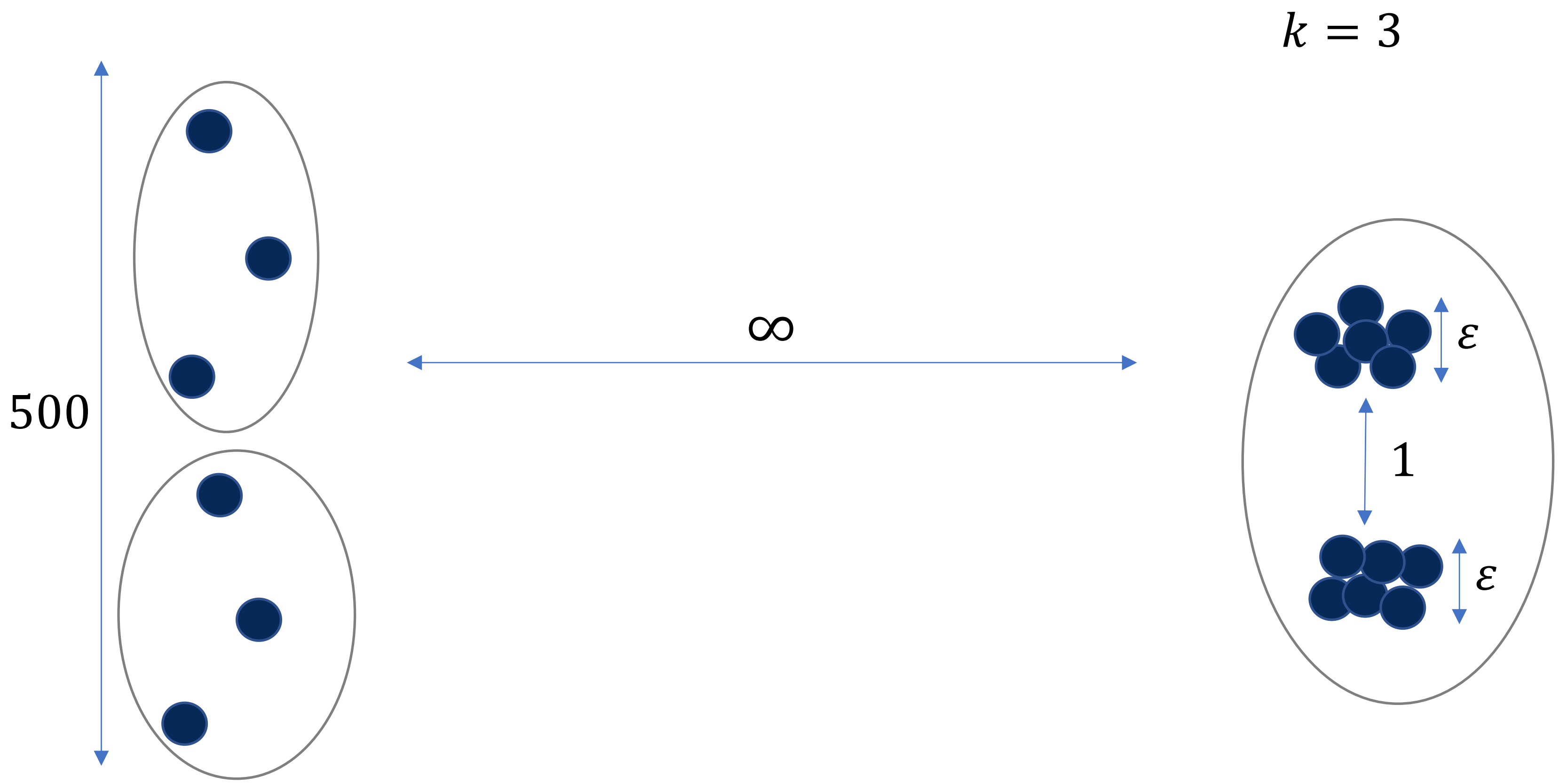
Fairness in Clustering

- **Why do we need fairness:**
  - Many decisions are made at least (partly) using algorithms

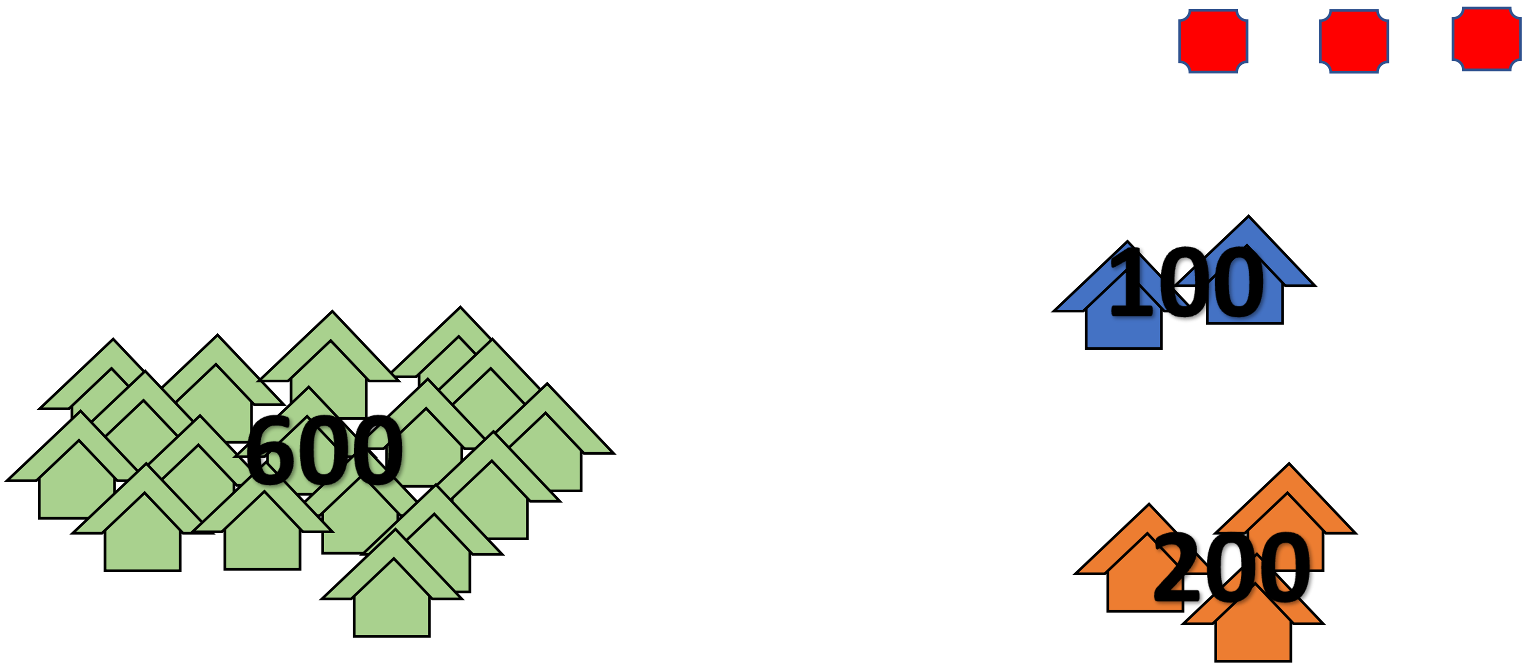
- Each point wishes to be as close as possible to some center

- **ML applications:** Closer to center $\Rightarrow$ better represented by the center
- **FL applications:** Closer to the center $\Rightarrow$ less travel distance to the facility

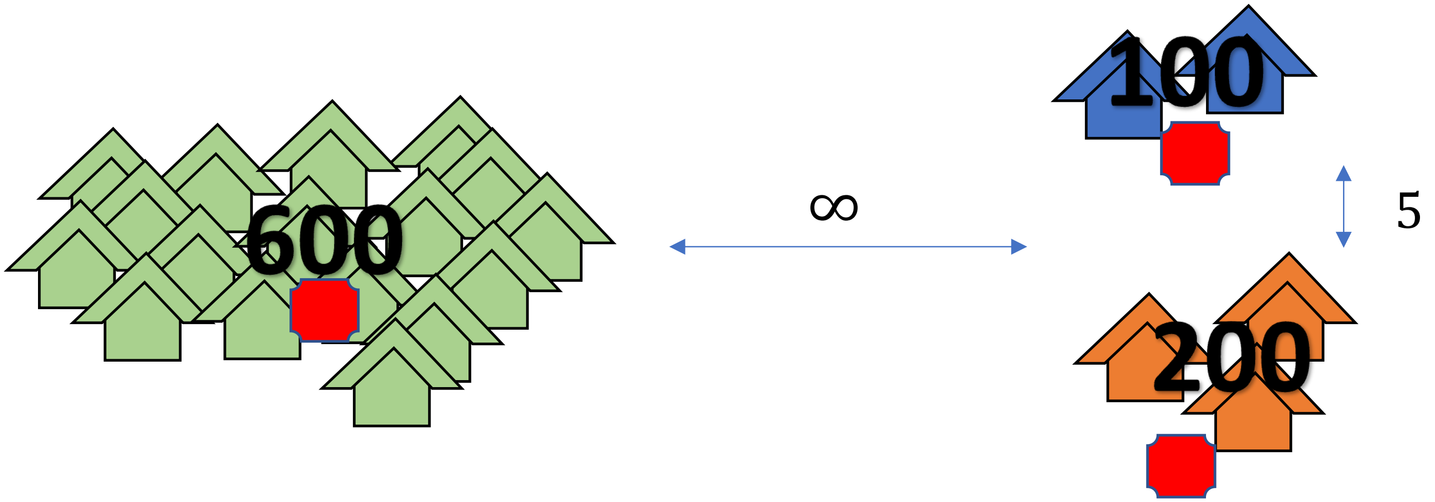
Fairness in Clustering



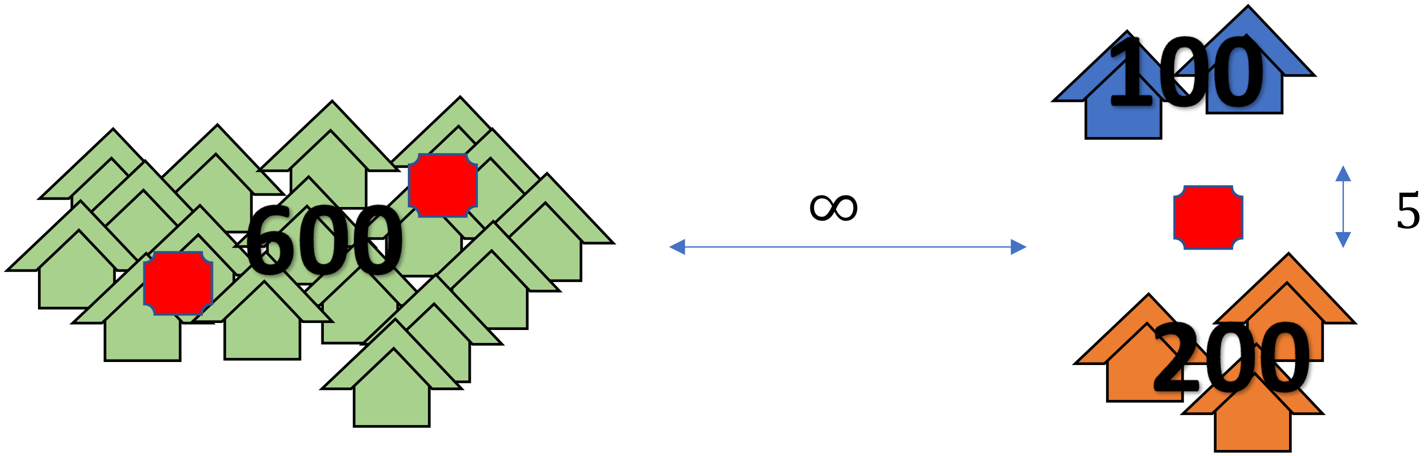
Fairness in Clustering



Fairness in Clustering



Fairness in Clustering



Fairness Through Proportionality

- *Proportionally Fair Clustering:*
  - *Every x% of the data points can select x% of the cluster centers*
  - *Every group of n/k agents "deserves" its own cluster center*

Core

- Definition in Committee Selection: W is in the core if
  - For all $S \subseteq N$ and $T \subseteq M$
  - If $|S| \geq |T| \cdot n/k$ (**large**)
  - Then, $u_i(W) \geq u_i(S)$ for some $i \in S$
  - "If a group can afford T, then T should not be a (strict) Pareto improvement for the group"

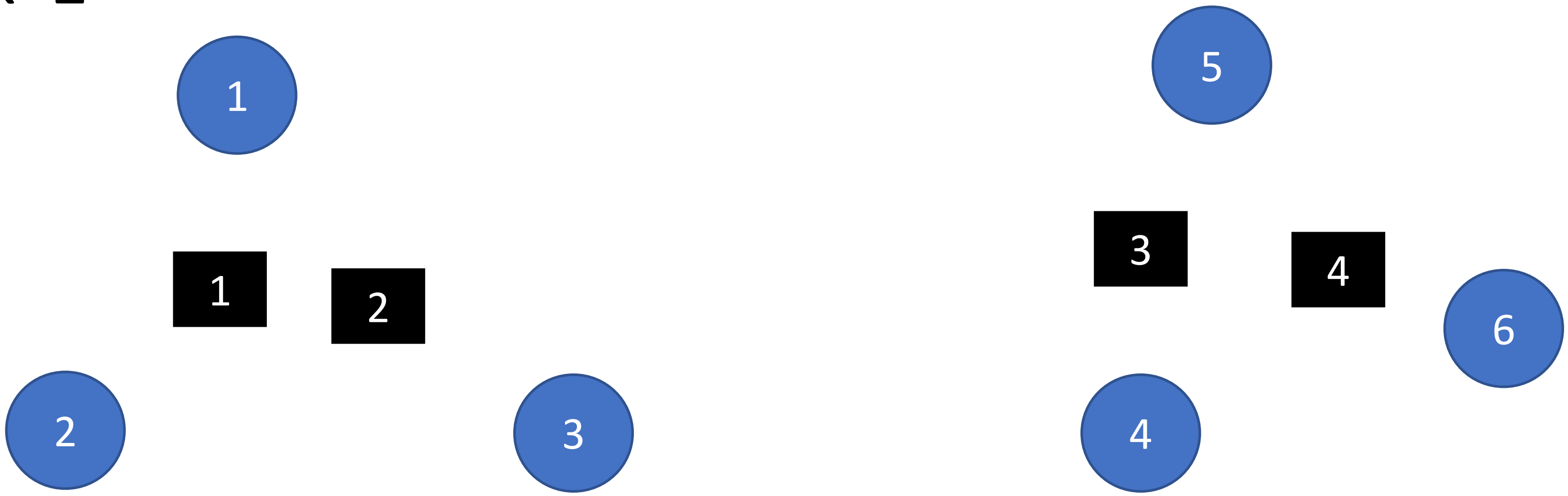
- Given clustering solution C, $C(i)$ denotes the closest center to $i \in N$

- Definition in Clustering: C is in the core if
  - For all $S \subseteq N$ and $y \subseteq M$
  - If $|S| \geq n/k$ (**large**)
  - Then, $d(i, C(i)) \leq d(i, y)$ for some $i \in S$
  - "If a group can afford a center y, then y should not be a (strict) Pareto improvement for the group"

Core

Example

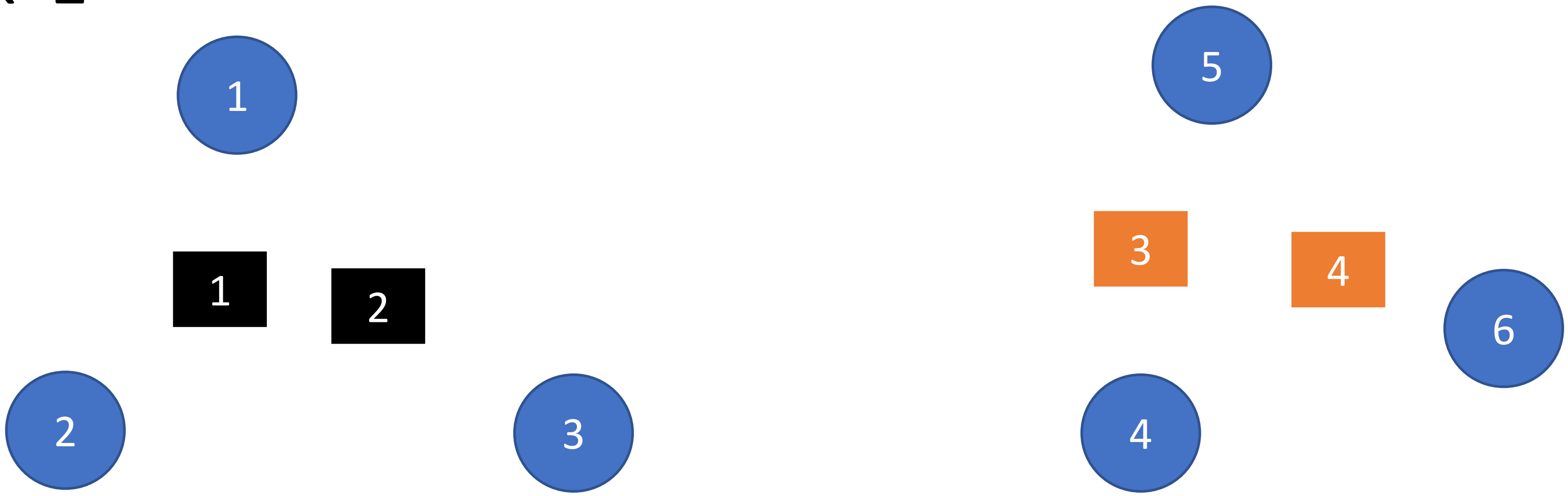
k=2



Core

Example

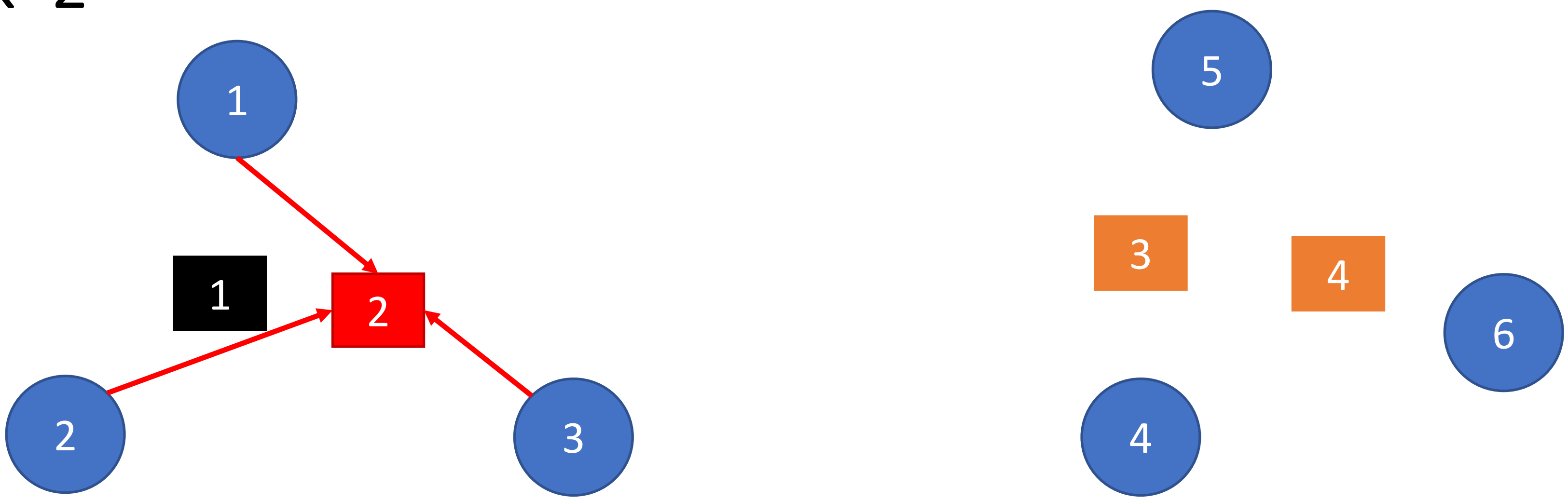
k=2



Core

Example

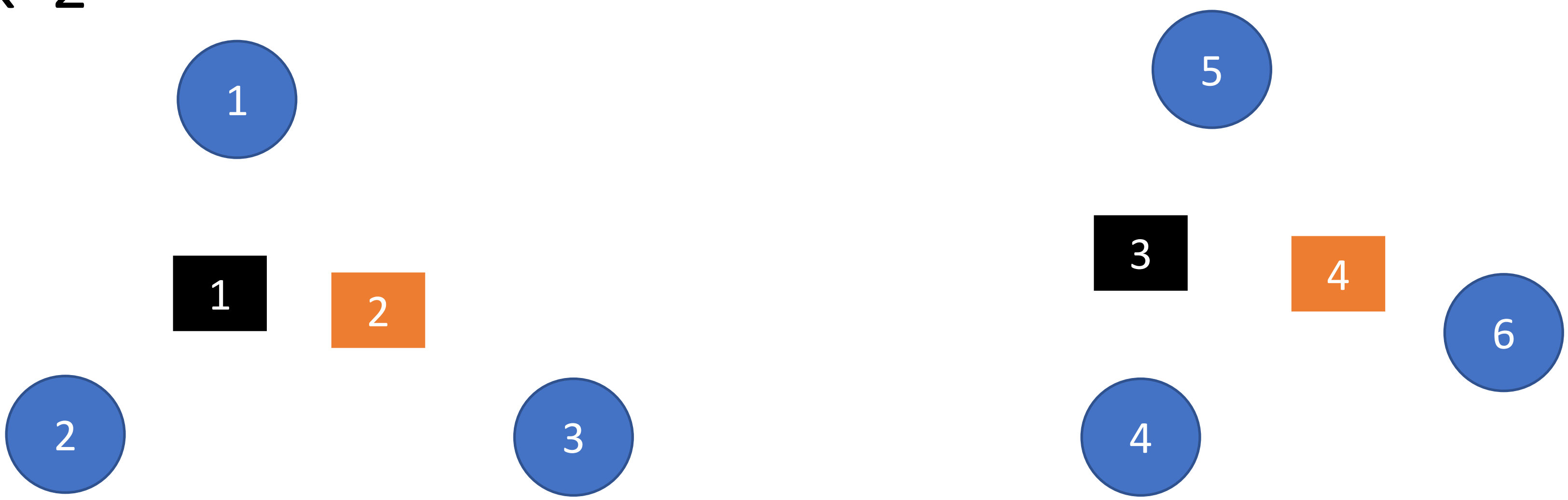
k=2



Core

Example

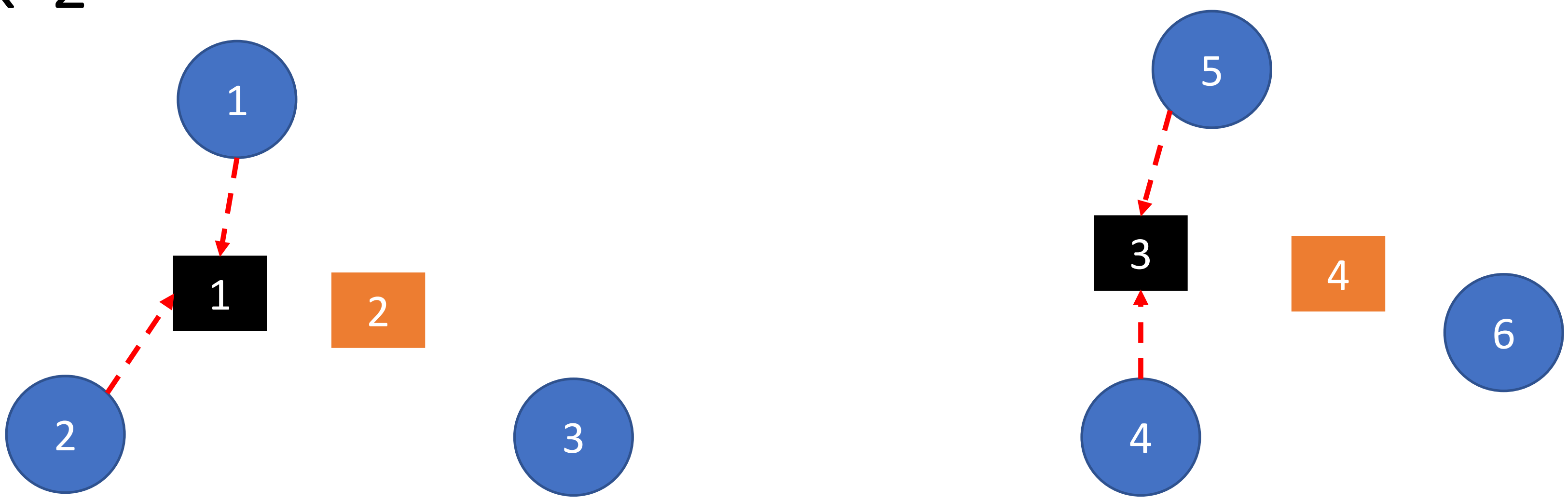
k=2



Core

Example

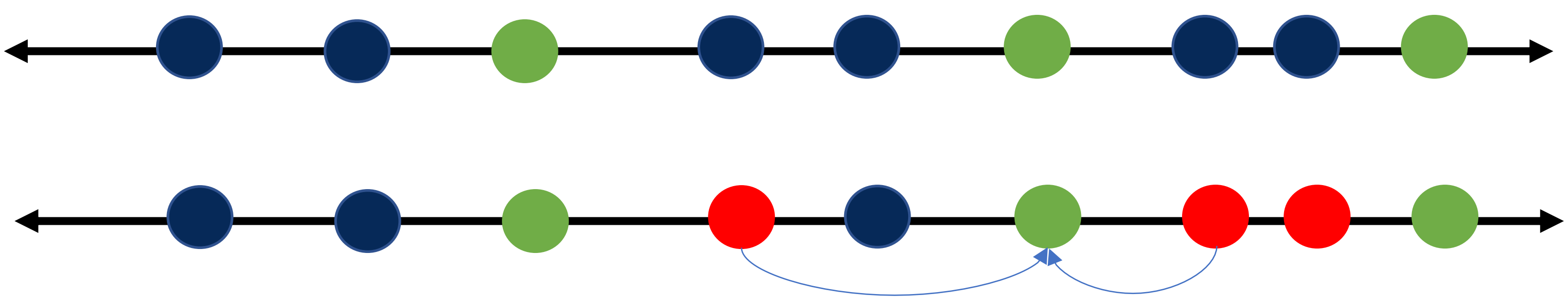
k=2



Core in the Line

- Theorem: In 1-D, a clustering solution in the core always exists
- Informal Proof:

- Move from left to right, making every $\lceil n/k \rceil$-th point a cluster center
- k=3



Core in Trees

- Tree G=(V, E)
- Every vertex is a ***data point*** and a ***feasible cluster center***
- Every edge has weight equal to 1
- ST(x) denotes the subtree rooted at node x

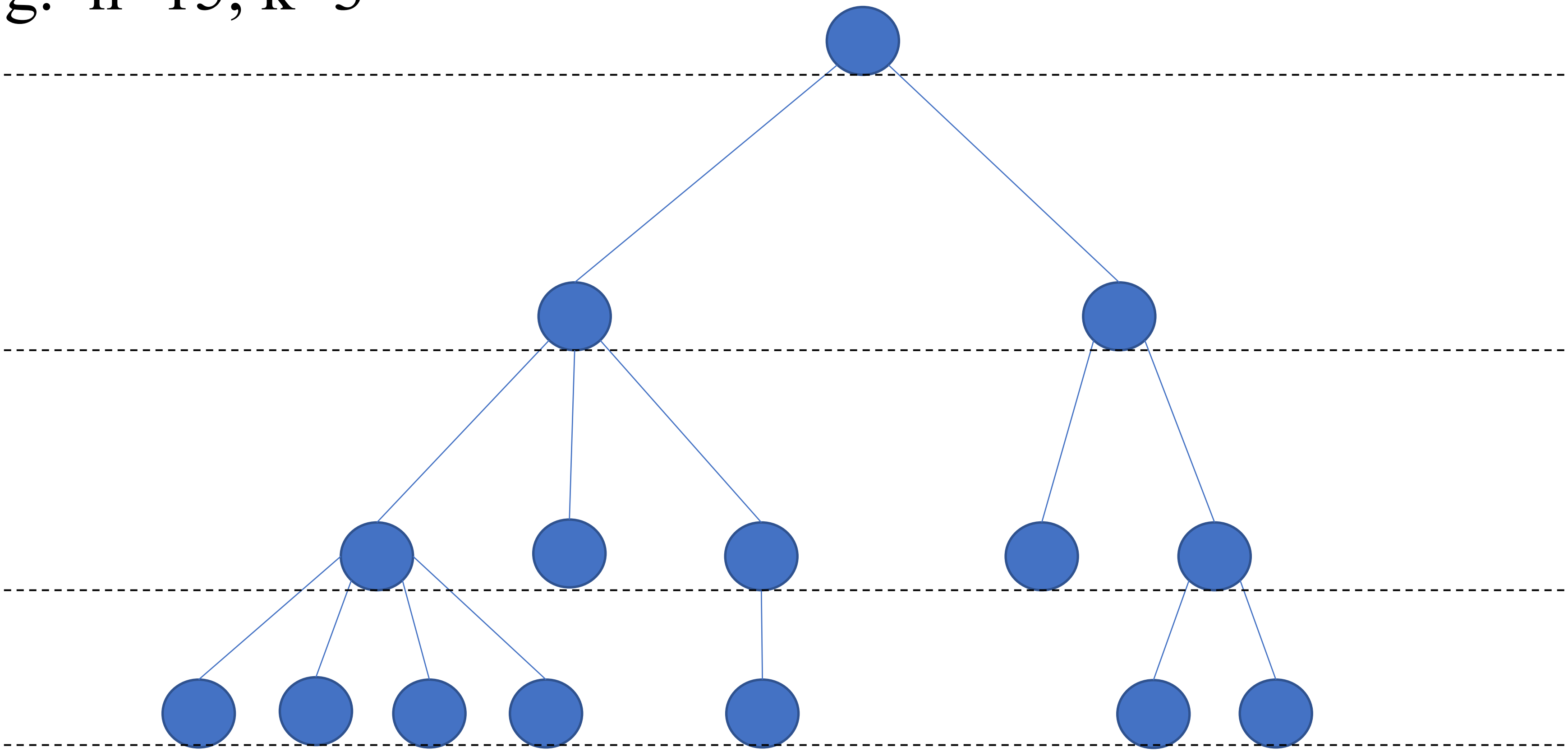
Core in Trees

Tree-Core Algorithm

1. $C = \emptyset$; *Root G at an arbitrary node r*
2. *For level equal to the height of the tree to 1 do*
3.   *For every node x in the current level*
4.   *If* $ST(x) \geq \frac{n}{k}$ *do*
5. $C = C \cup x$
6. $G = G \setminus ST(x)$
7. *If* $|G| > 0$
8. $C = C \cup r$
9. *Return C*

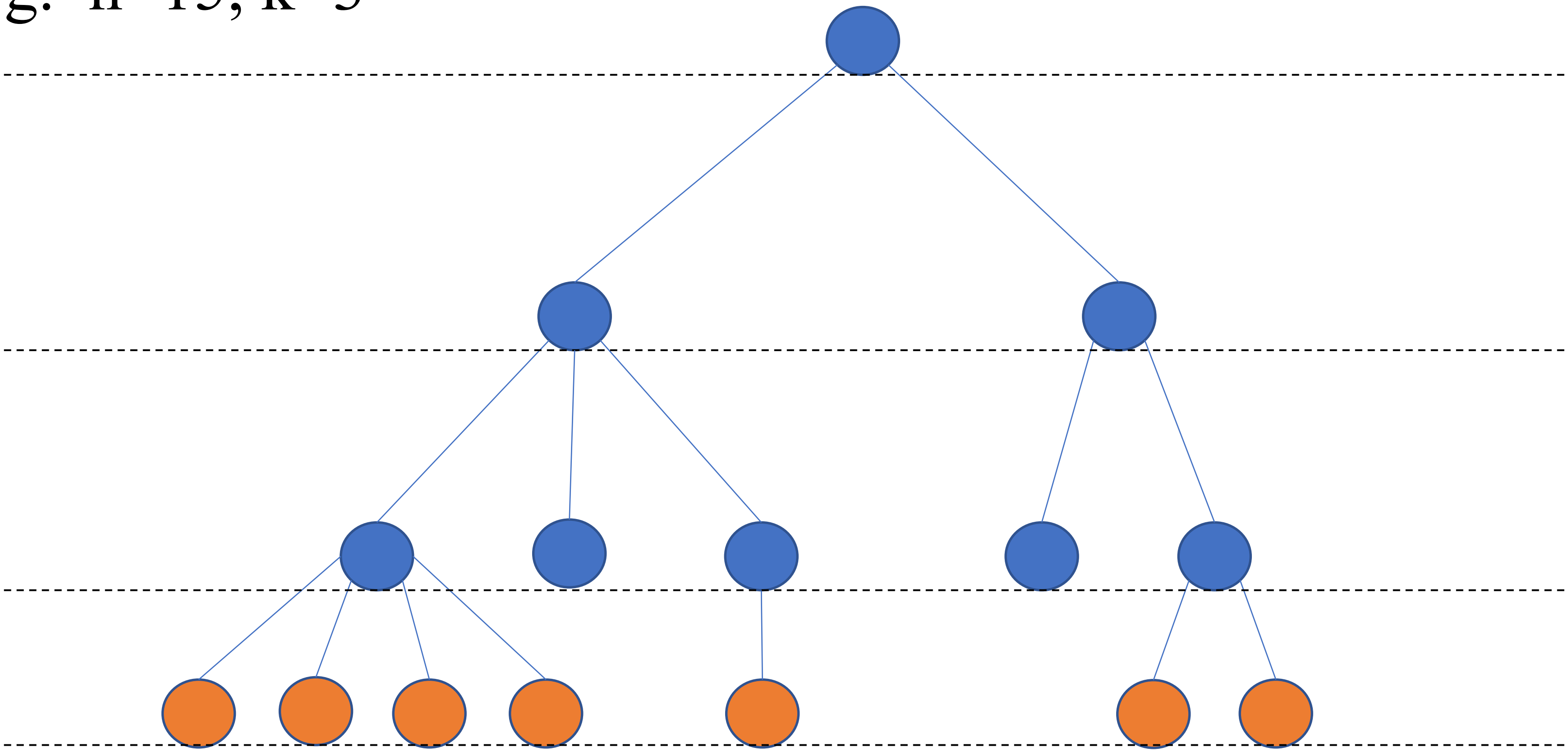
Core in Trees

e.g. $n=15$, $k=3$



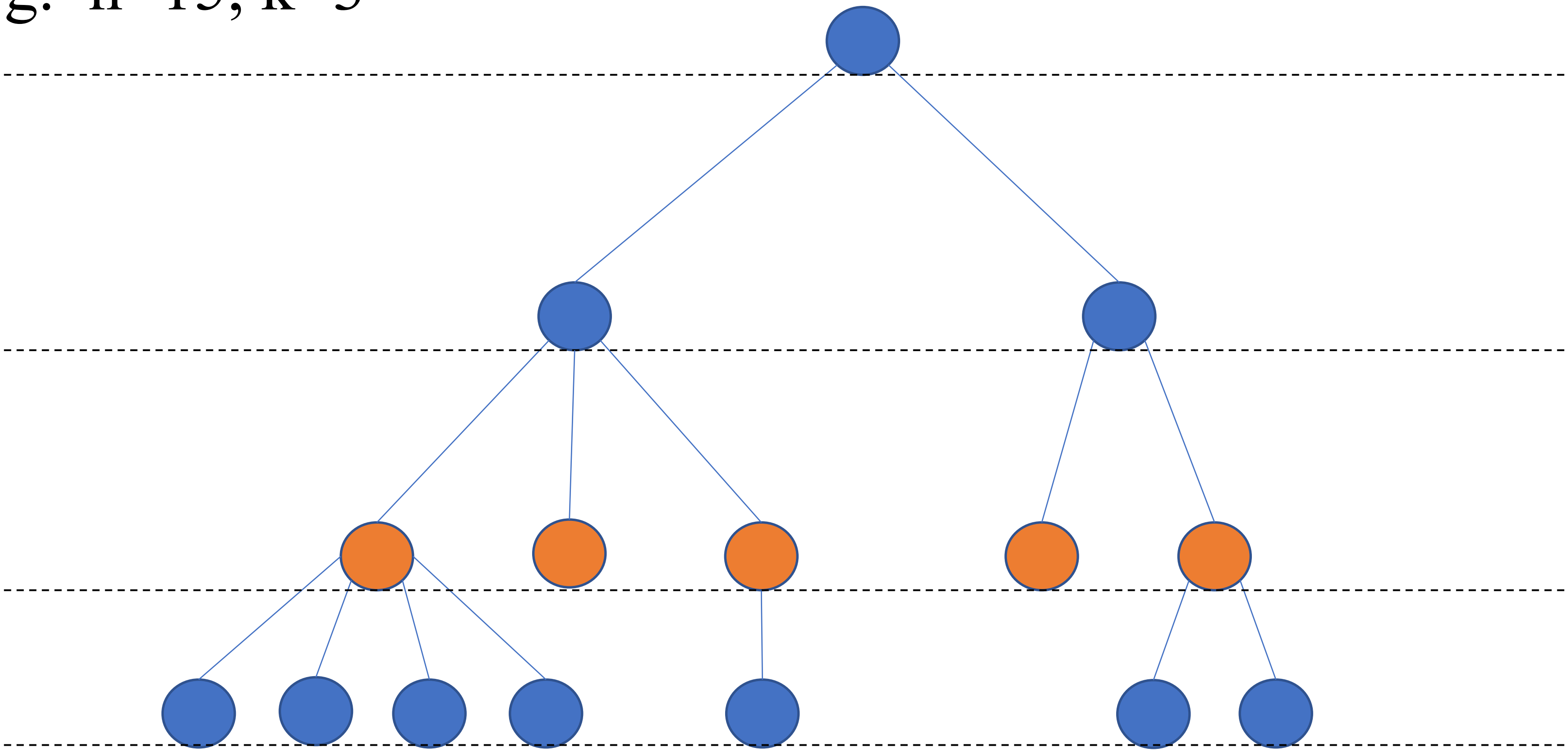
Core in Trees

e.g. $n=15$, $k=3$



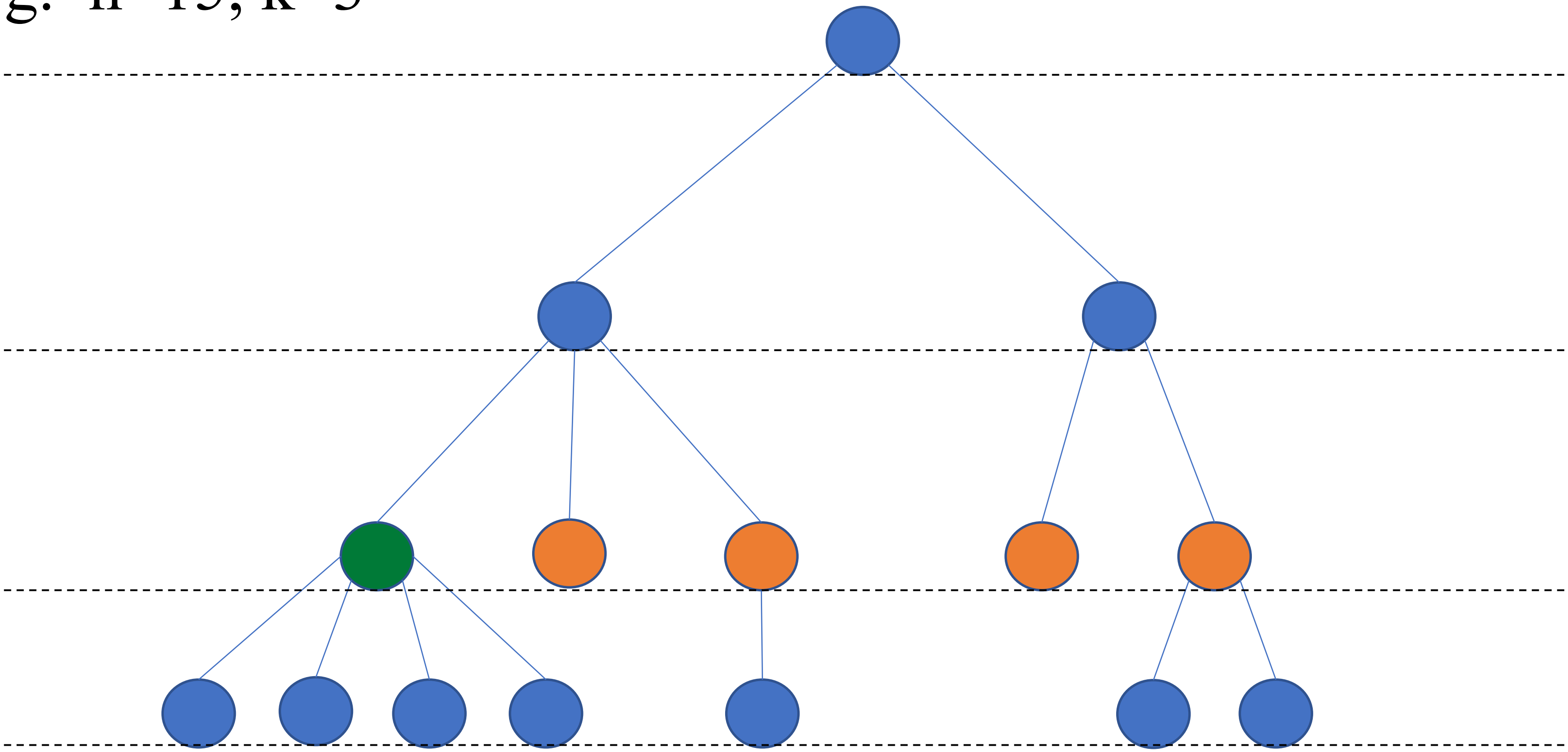
Core in Trees

e.g. $n=15$, $k=3$



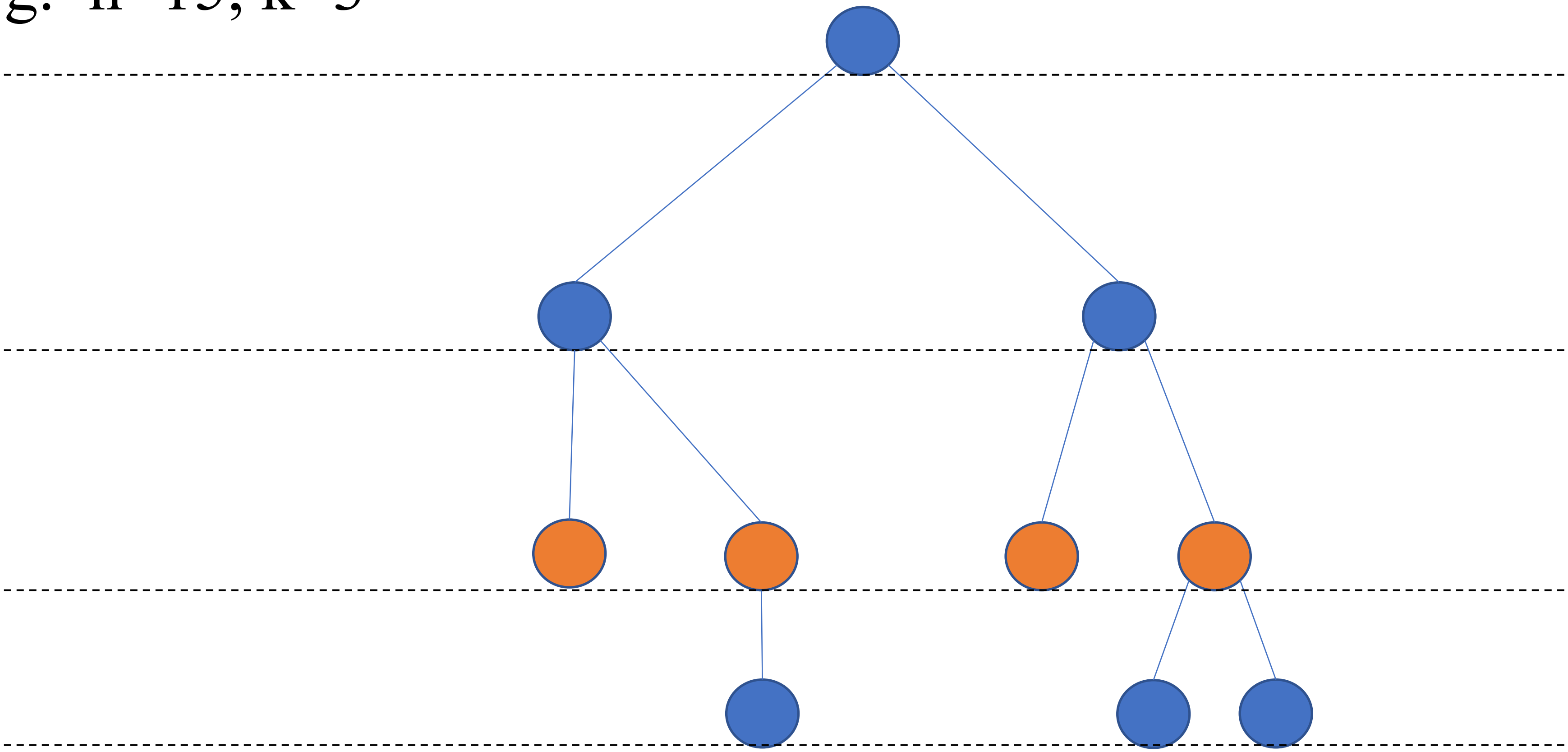
Core in Trees

e.g. $n=15$, $k=3$



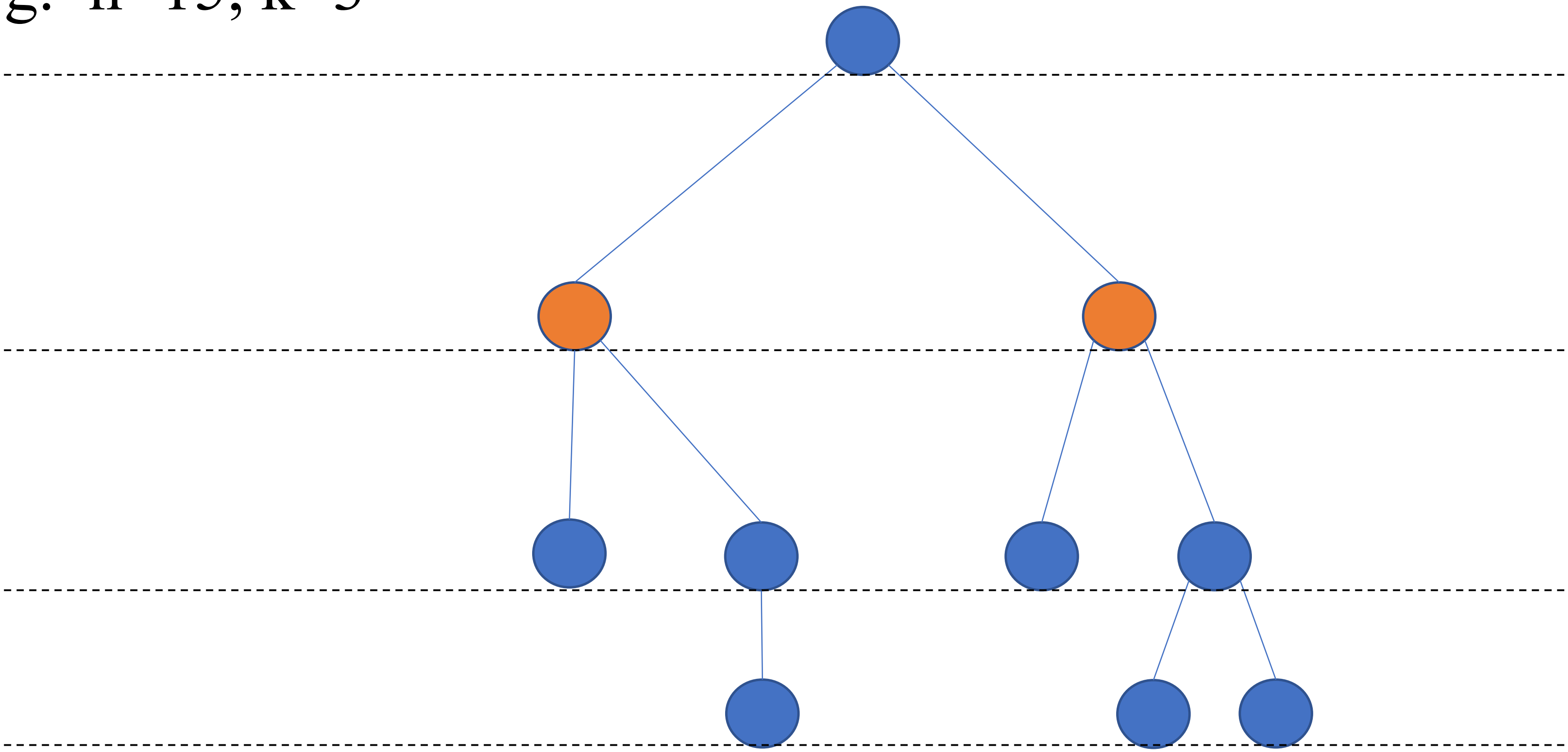
Core in Trees

e.g. n=15, k=3



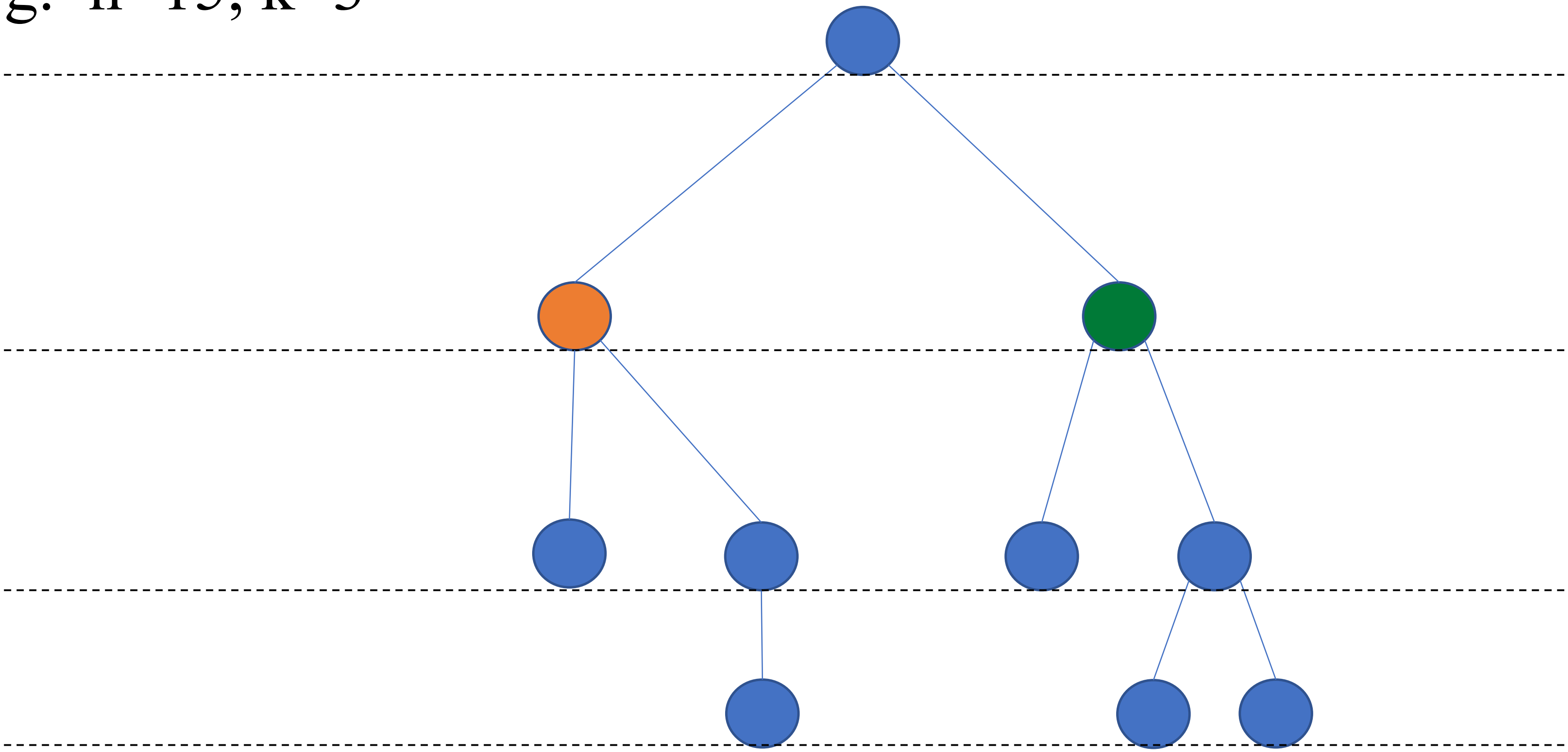
Core in Trees

e.g. $n=15, k=3$



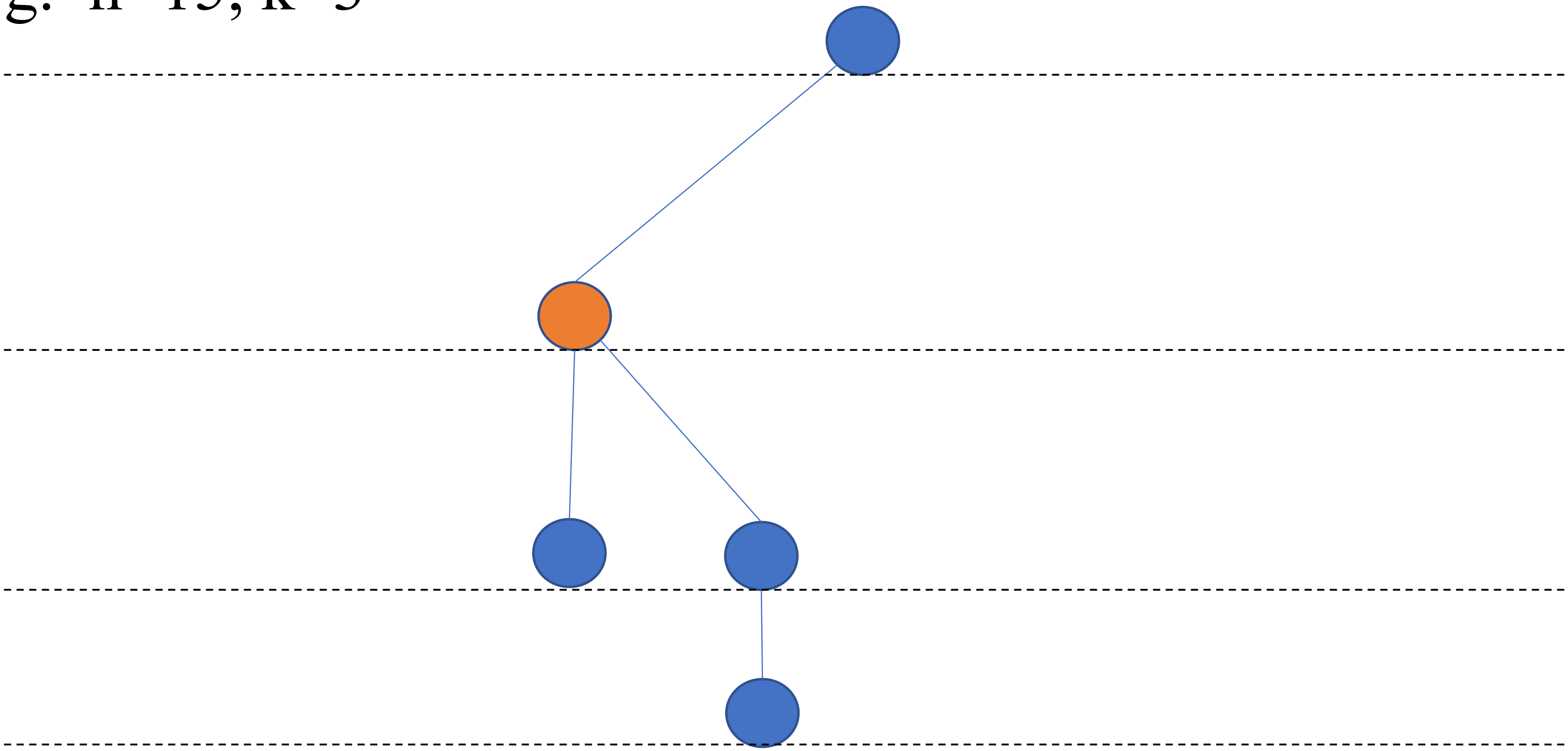
Core in Trees

e.g. $n=15$, $k=3$



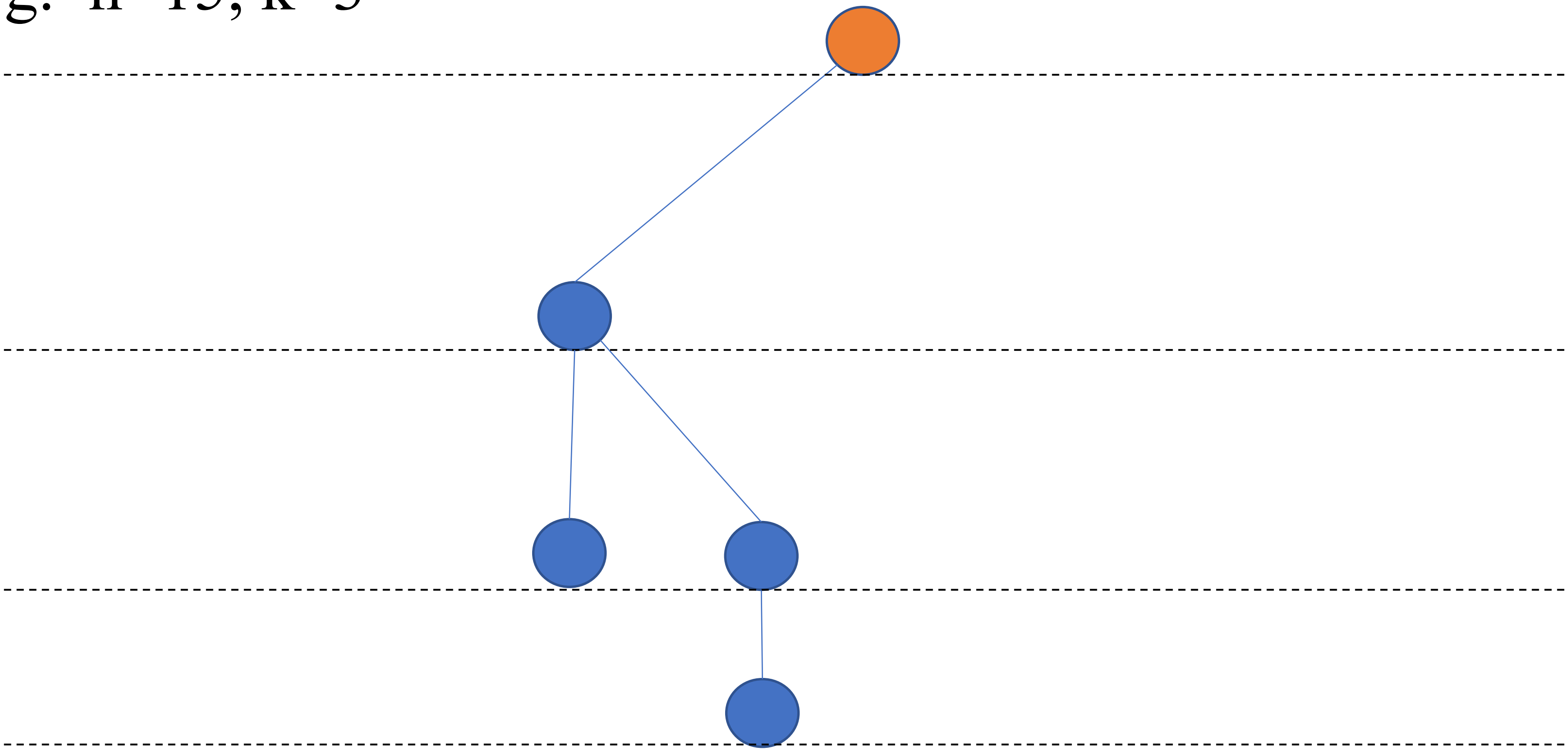
Core in Trees

e.g. $n=15$, $k=3$



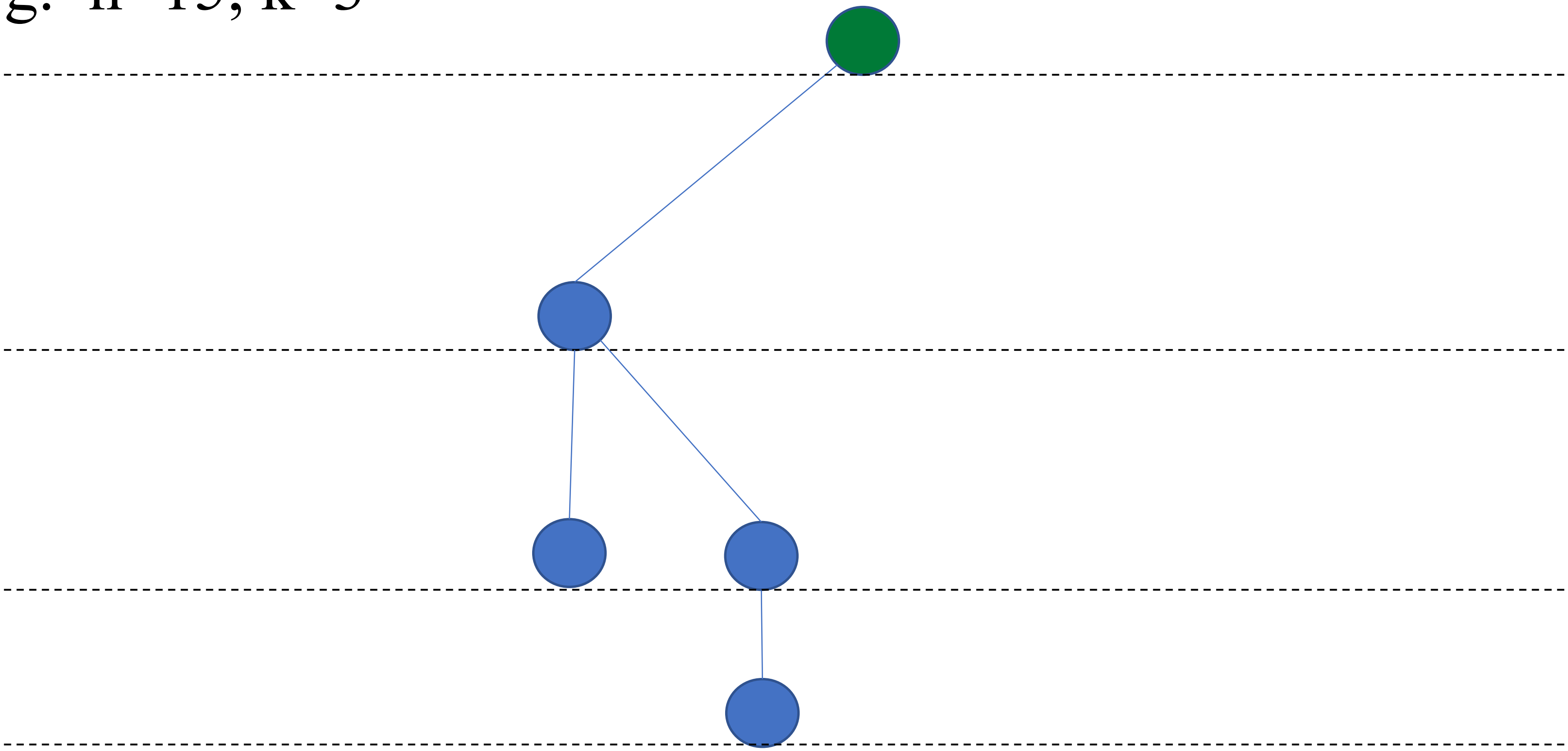
Core in Trees

e.g. n=15, k=3



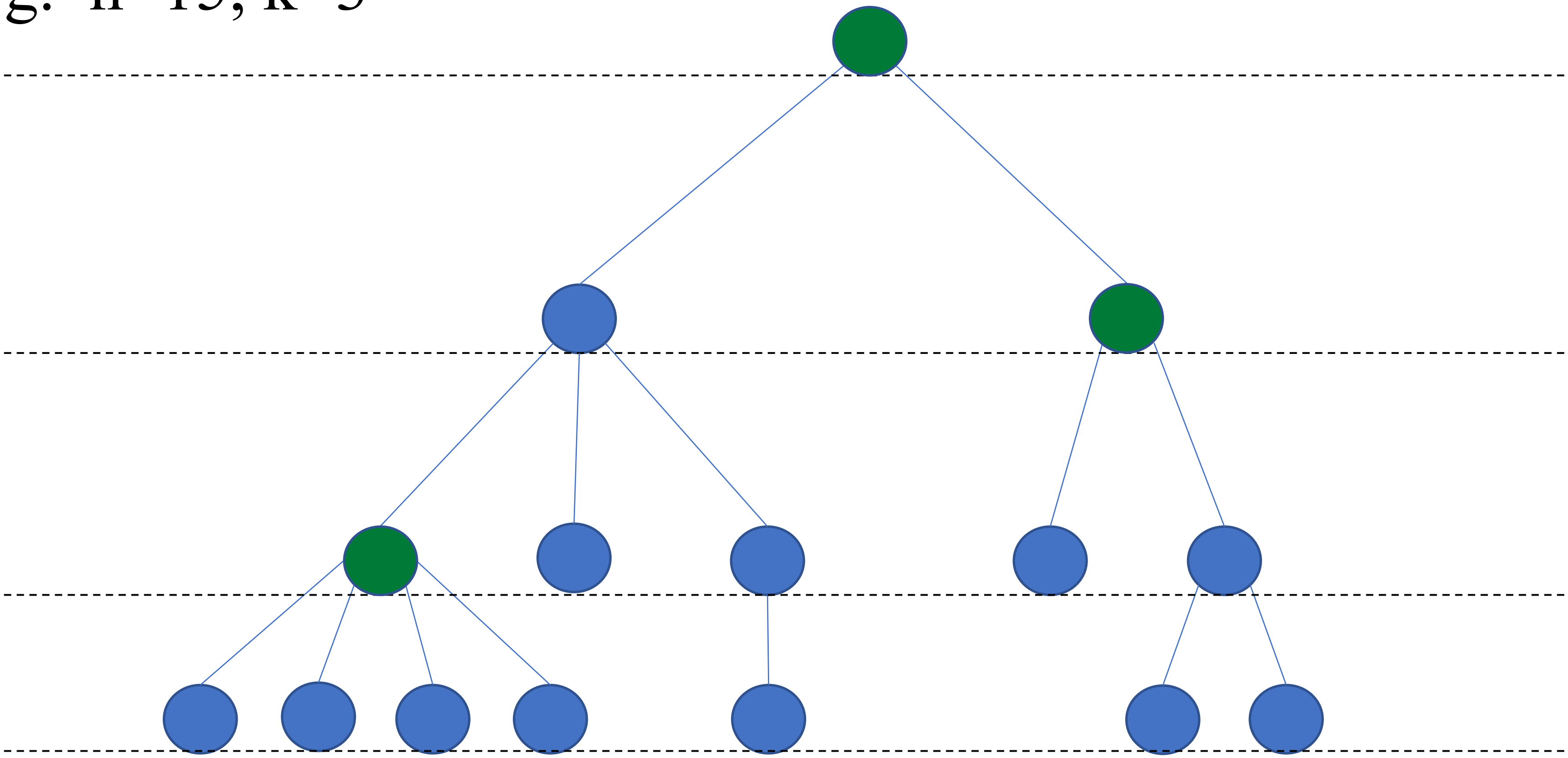
Core in Trees

e.g. $n=15$, $k=3$



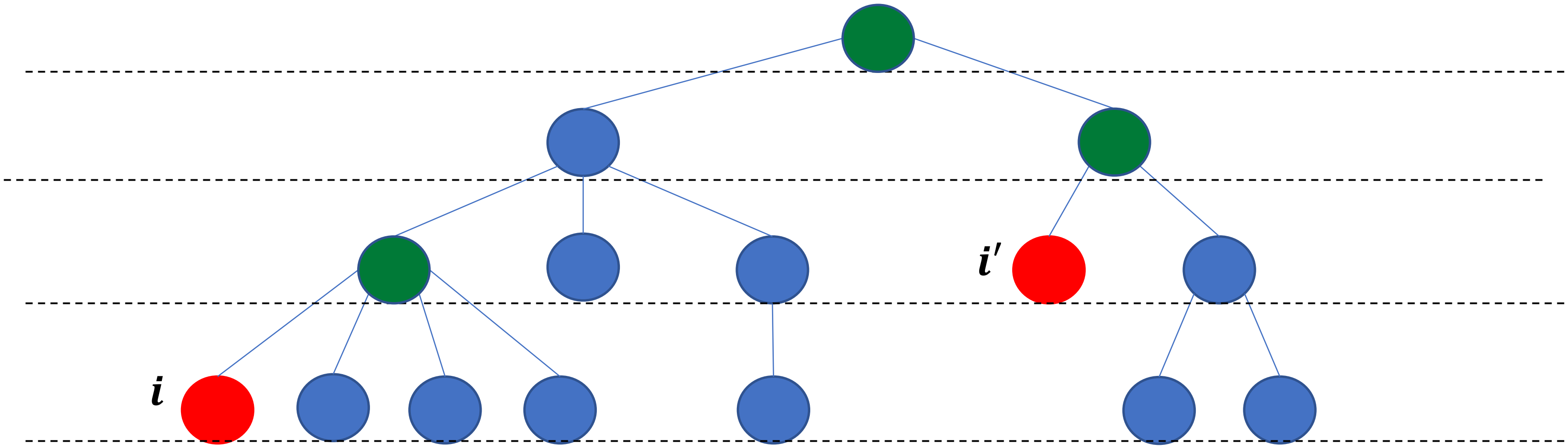
Core in Trees

e.g. $n=15$, $k=3$



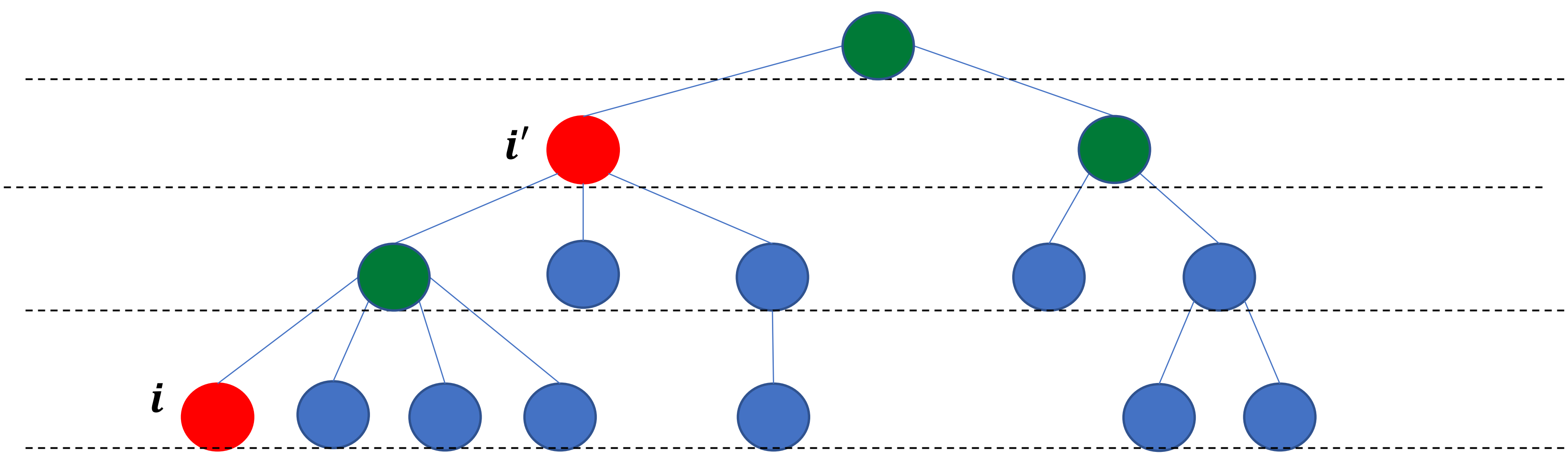
Core in Trees

- Theorem: Tree-Core Algorithm returns a clustering solution C in the core
- Informal Proof:
- Let $p(i)$ be the closest ancestor of i in C
- Observation: $d\big(i, p(i)\big) < d(i,j), \forall\, j \notin ST(p(i))$
- Case I: There are $i, i' \in S$, $ST\big(p(i)\big) \cap ST\big(p(i')\big) = \emptyset$



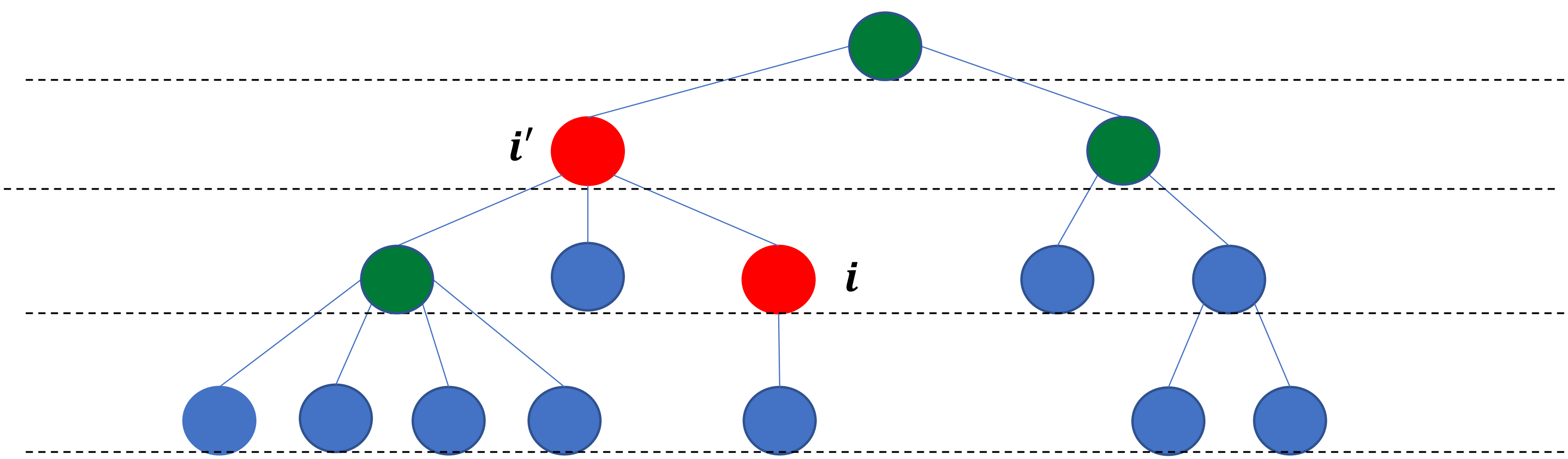
Core in Trees

- Theorem: Tree-Core Algorithm returns a clustering solution C in the core
- Informal Proof:
- Let $p(i)$ be the closest ancestor of i in C
- Observation: $d\big(i, p(i)\big) < d(i, j), \forall\, j \notin ST(p(i))$
- Case II: There are $i, i' \in S$, $p(i) \in ST\big(p(i')\big)$



Core in Trees

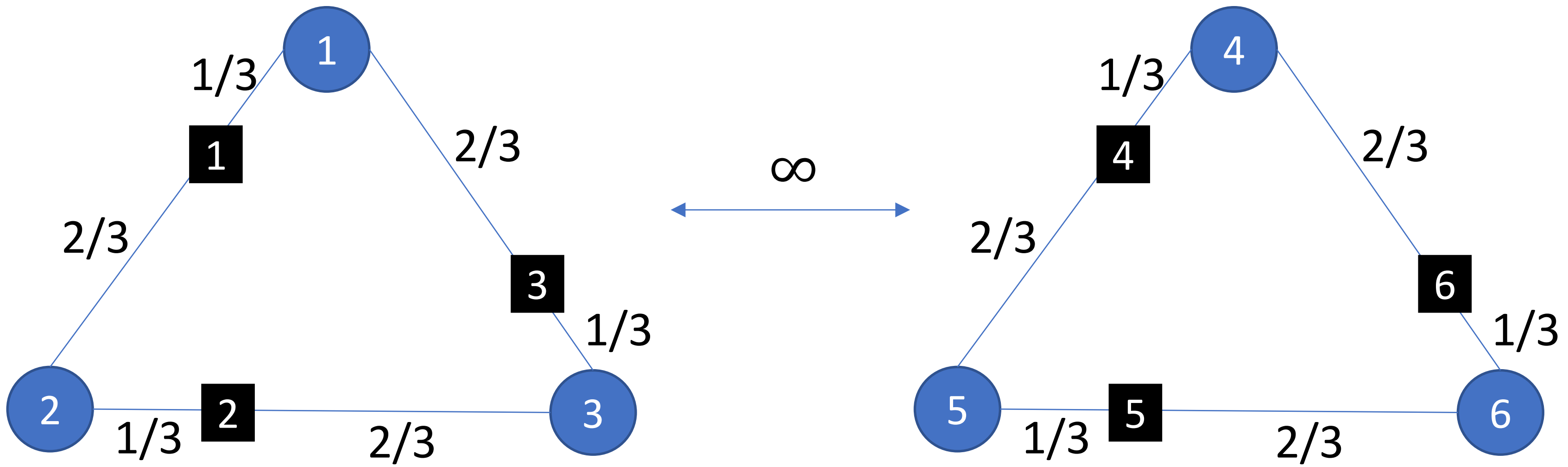
- Theorem: Tree-Core Algorithm returns a clustering solution C in the core
- Informal Proof:
- Let $p(i)$ be the closest ancestor of i in C
- Observation: $d\big(i, p(i)\big) < d(i, j), \forall\, j \notin ST(p(i))$
- Case III: There are $i, i' \in S$, $p(i)$= $p(i')$



Core in General Metric Spaces

- Theorem: A clustering solution in the core does not always exist
- Proof:

k=3

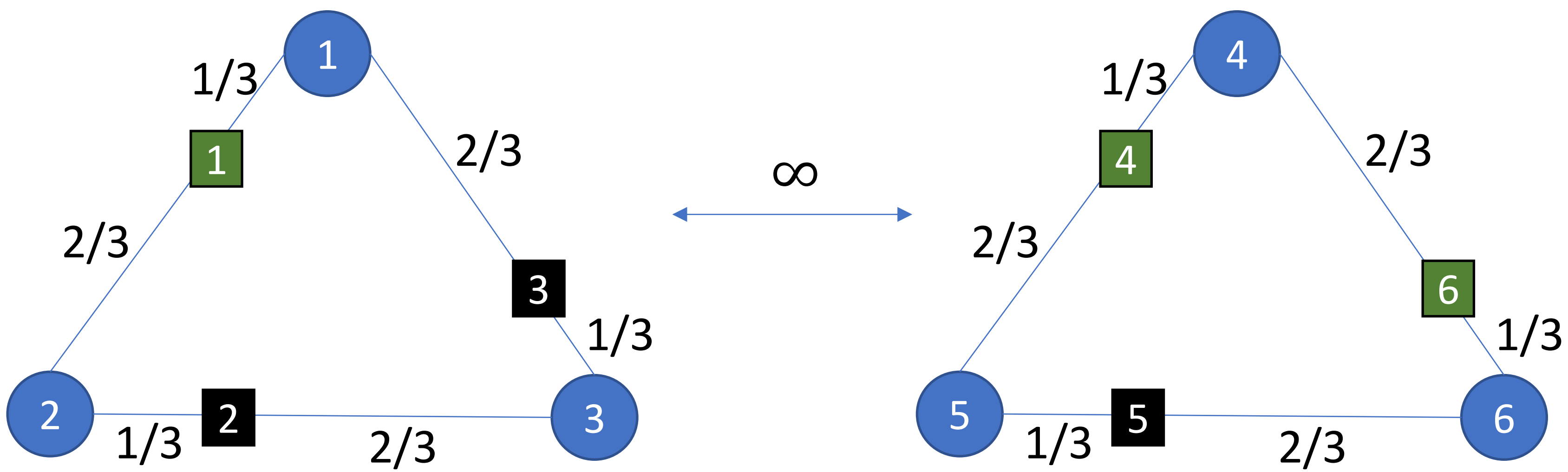


Core in General Metric Spaces

- Theorem: A clustering solution in the core does not always exist

- Proof:

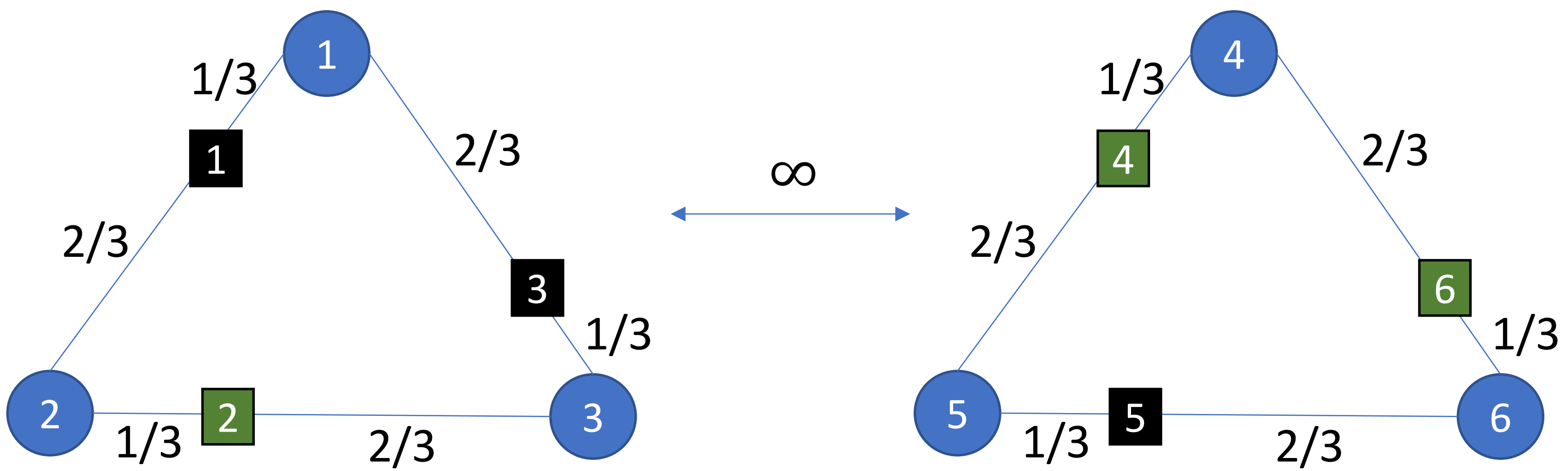
k=3



Core in General Metric Spaces

- **Theorem:** A clustering solution in the core does not always exist
- **Proof:**

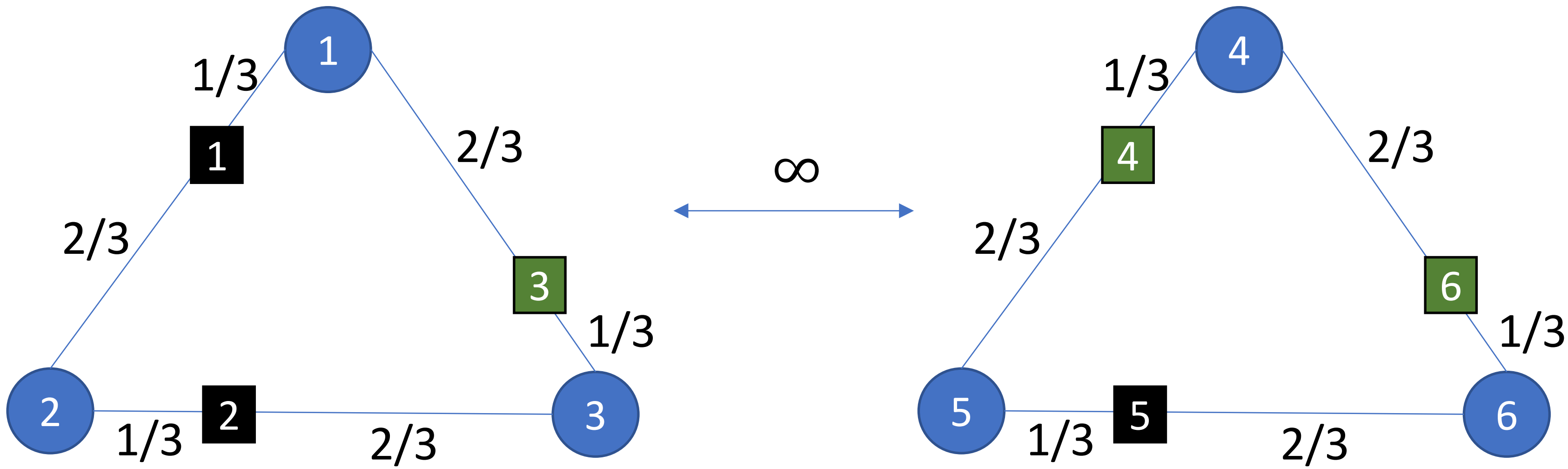
k=3



Core in General Metric Spaces

- Theorem: A clustering solution in the core does not always exist
- Proof:

k=3



α-Core

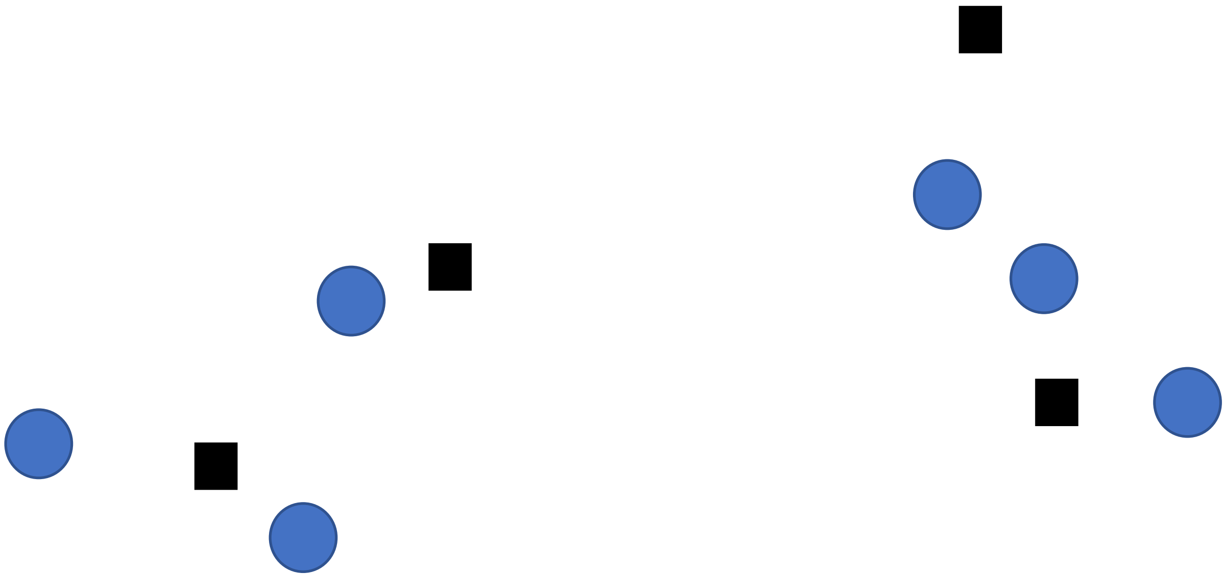
- Definition in Clustering: C is in the core if
  - For all $S \subseteq N$ and $y \subseteq M$
  - If $|S| \geq n/k$ (**large**)
  - Then, $d(i, C(i)) \leq \alpha \cdot d(i, y)$ for some $i \in S$
  - "If a group can afford a center y, then y should not be a (strict) Pareto improvement for the group"

α-Core:
A solution C is in the α-core, with α ≥ 1 if there is ***no*** *group of points* $S \subseteq N$ *with* $|S| \geq n/k$ *and* $y \in M$ *such that:*

$$\forall i \in S, \alpha \cdot d(i, y) < d(i, C(i))$$

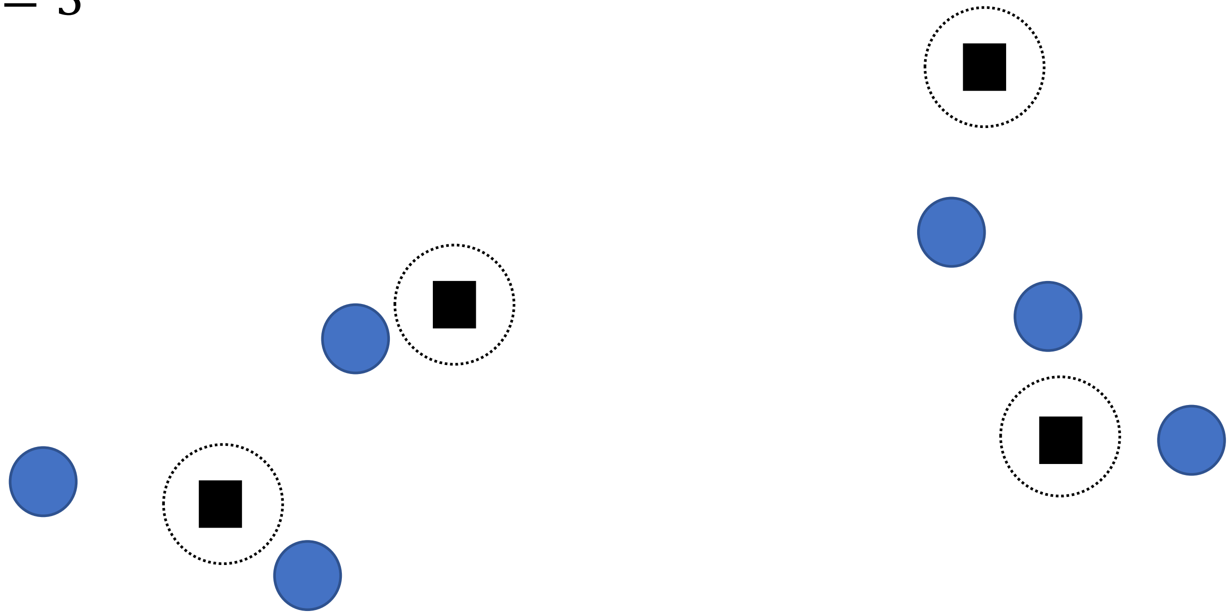
Greedy Capture

$k = 3$



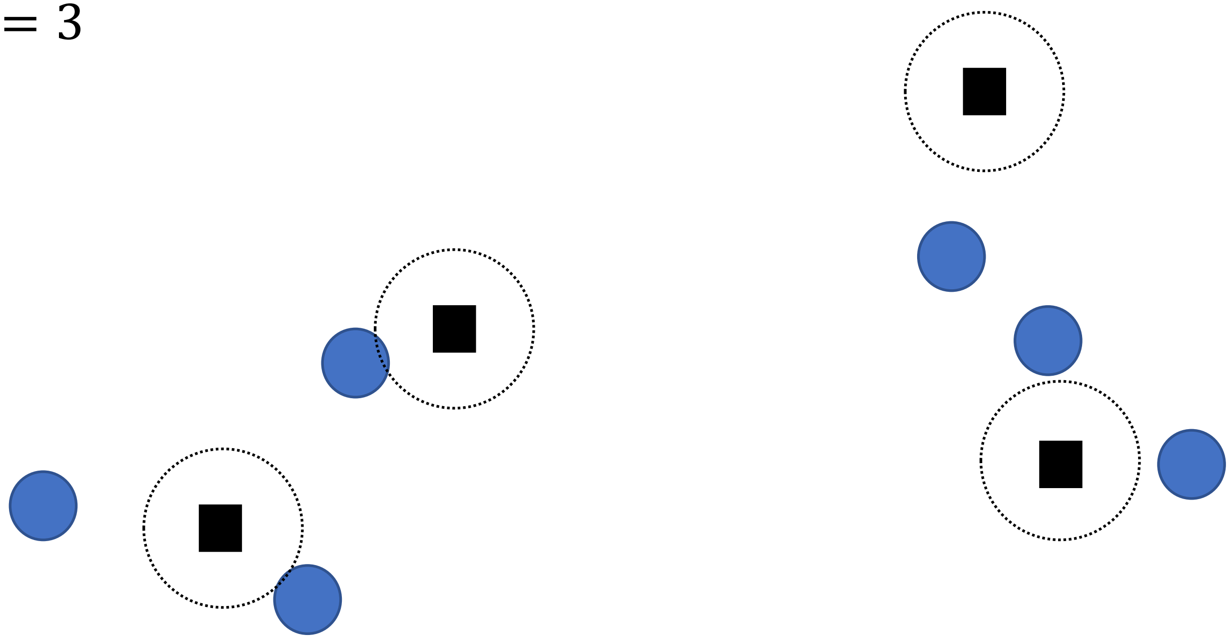
Greedy Capture

$k = 3$



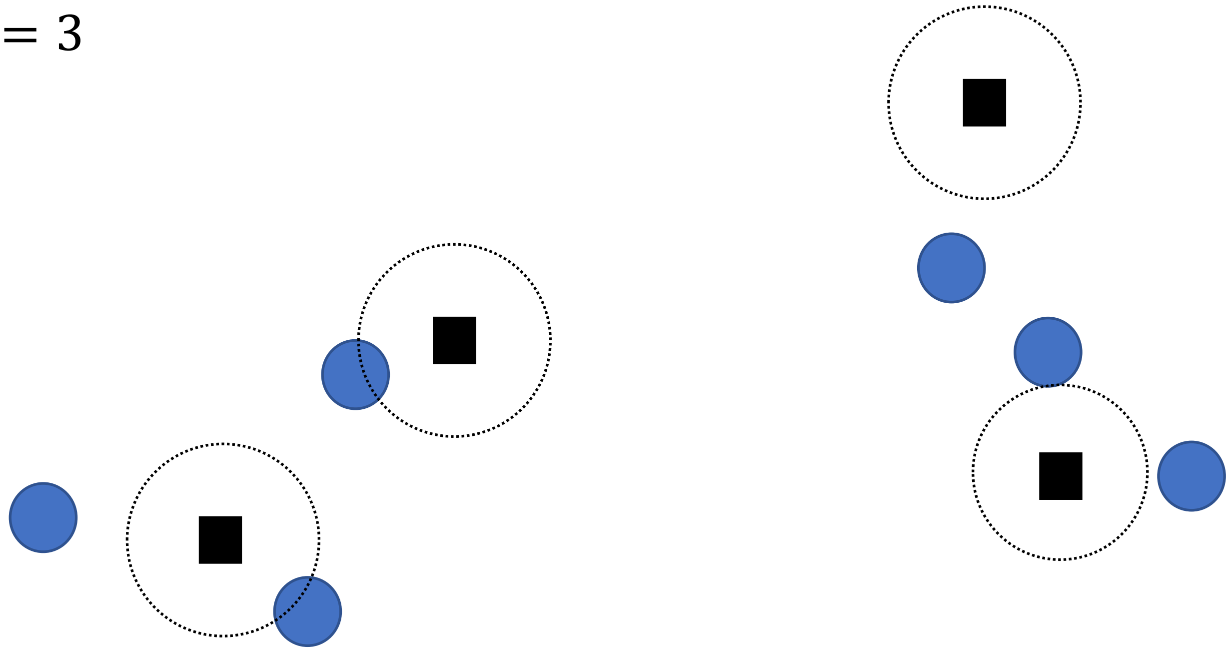
Greedy Capture

$k = 3$



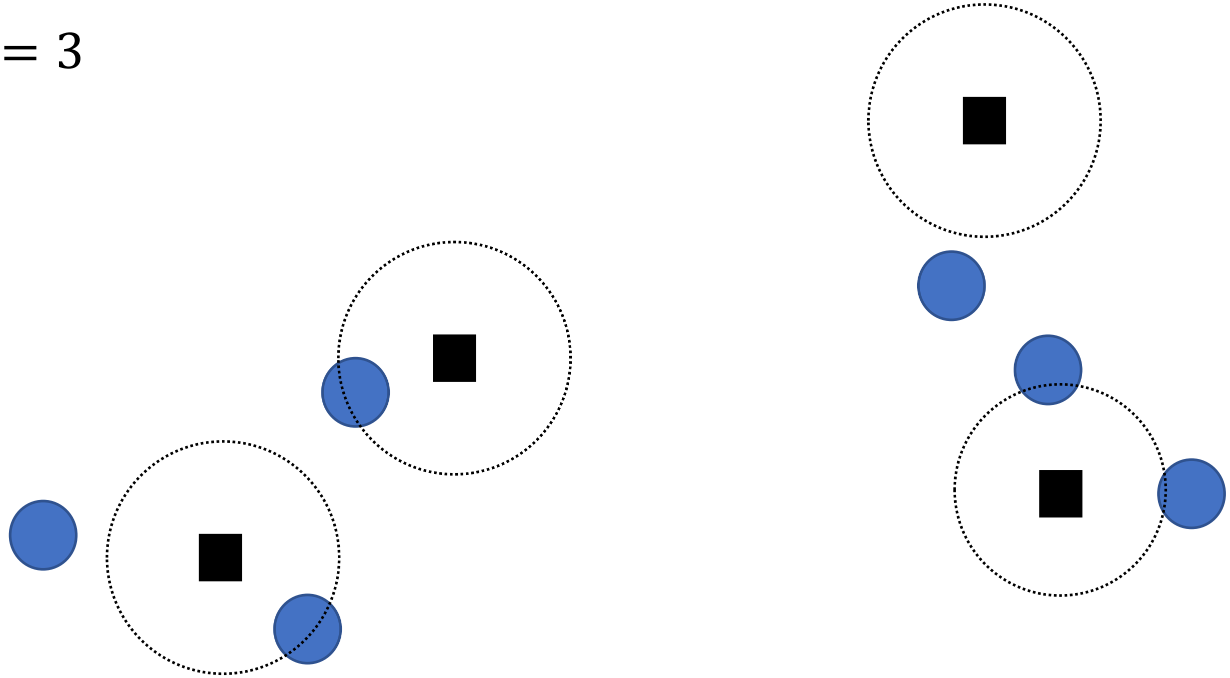
Greedy Capture

$k = 3$



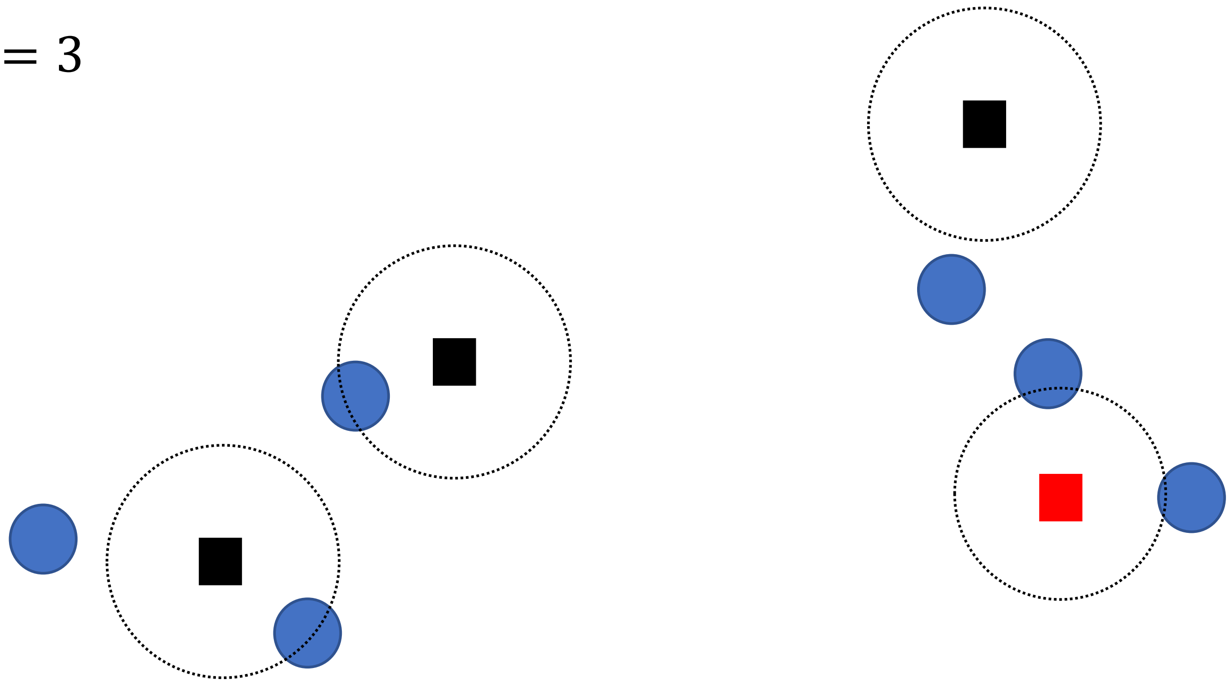
Greedy Capture

$k = 3$



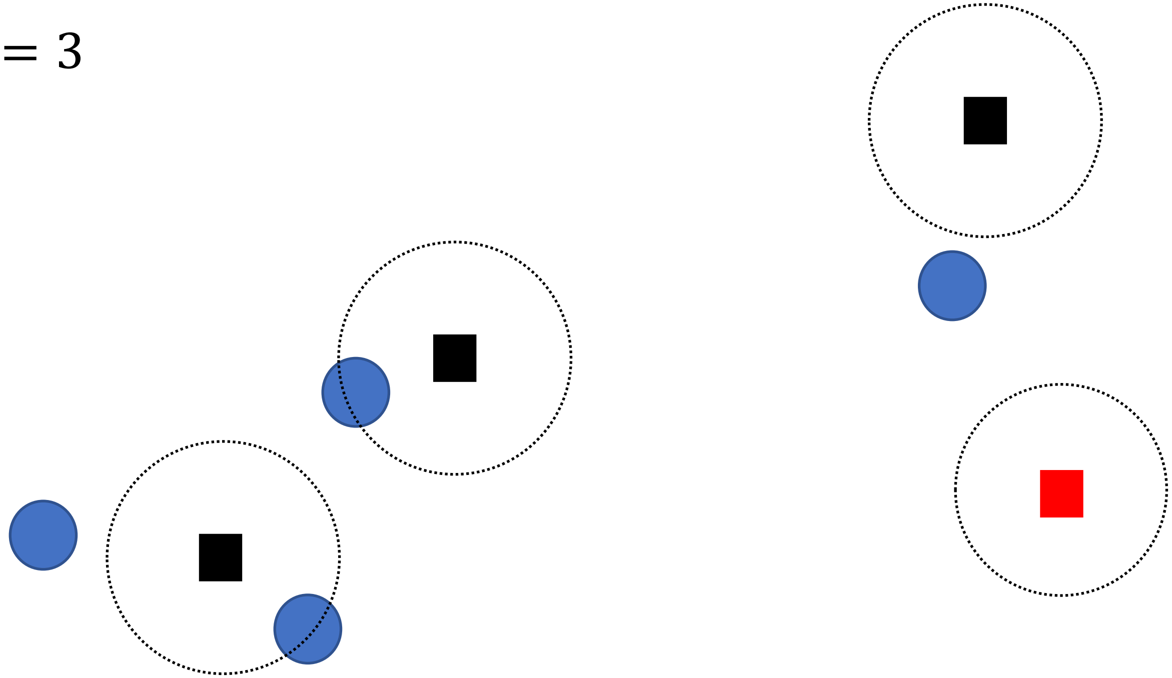
Greedy Capture

$k = 3$



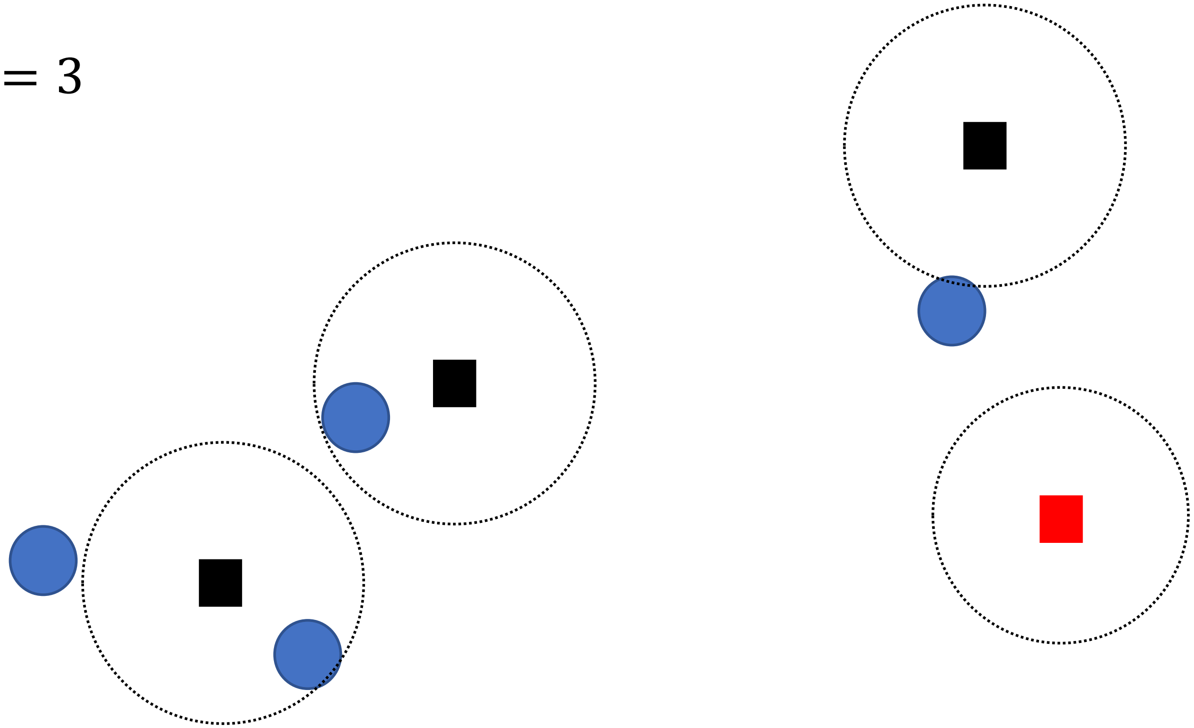
Greedy Capture

$k = 3$



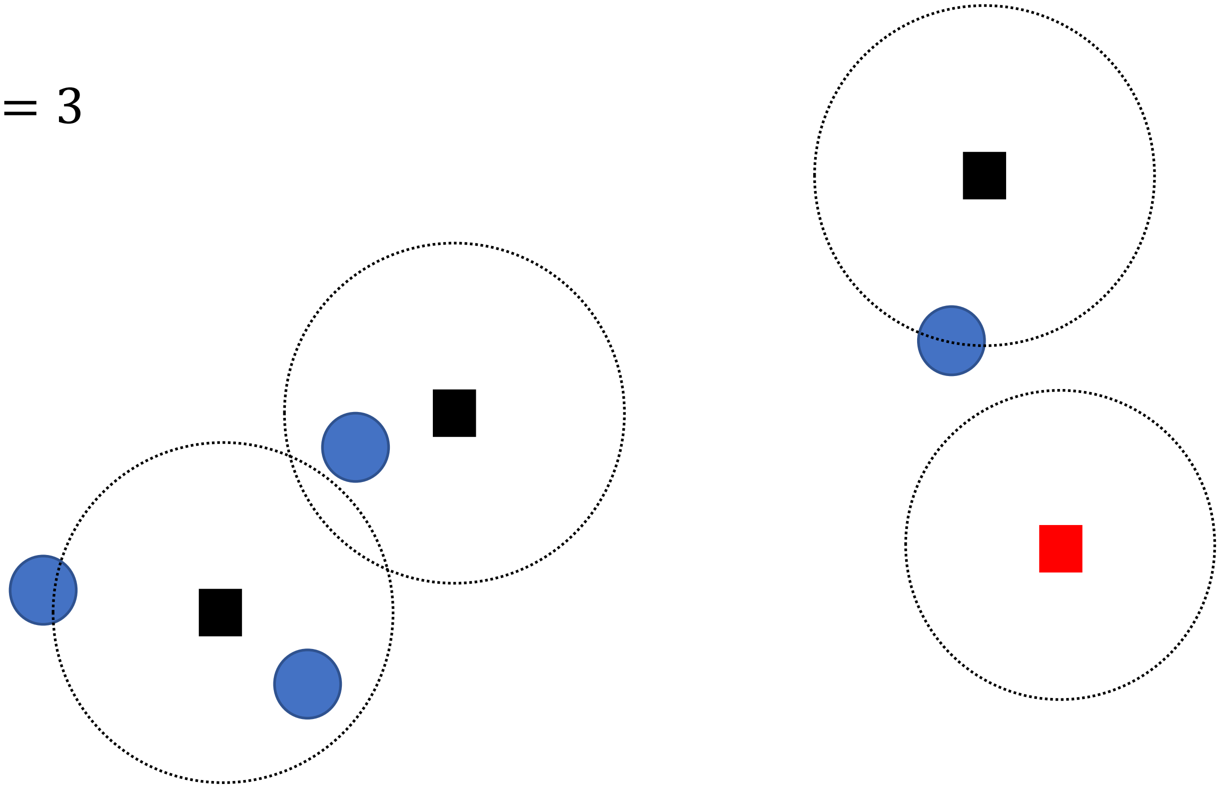
Greedy Capture

$k = 3$



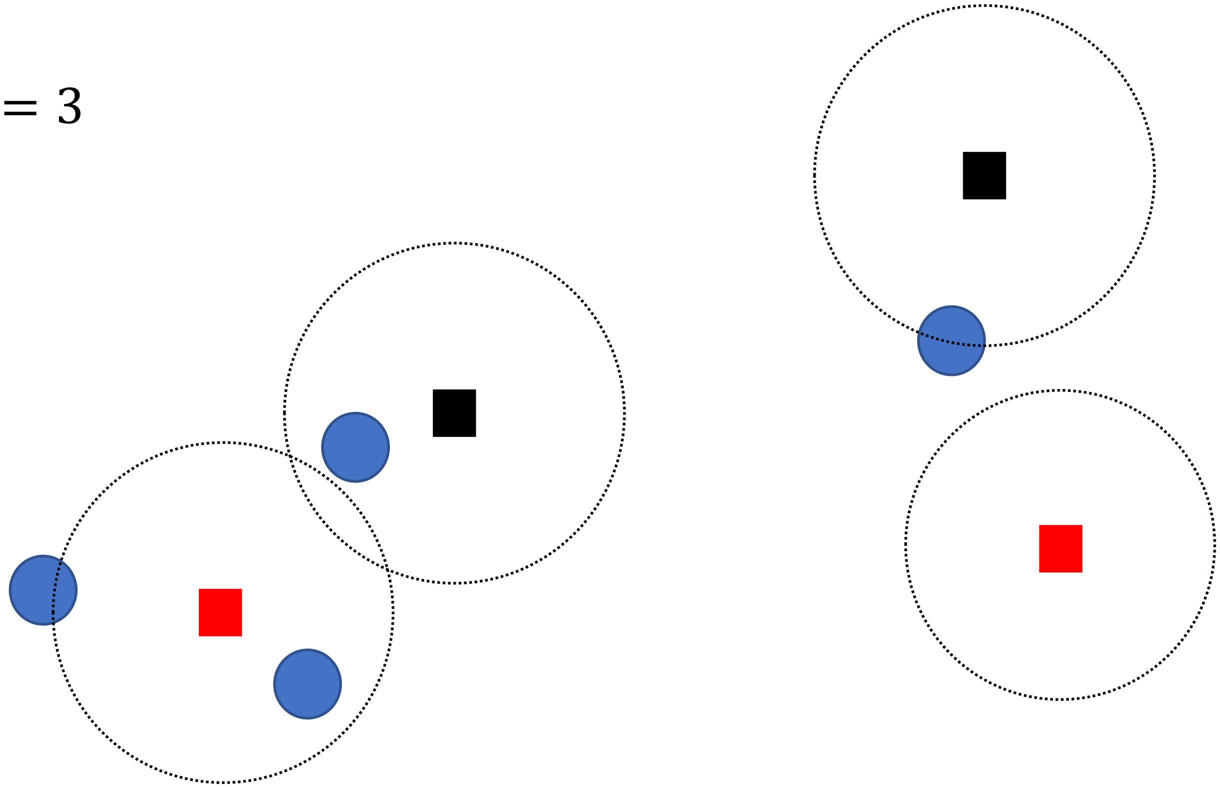
Greedy Capture

$k = 3$



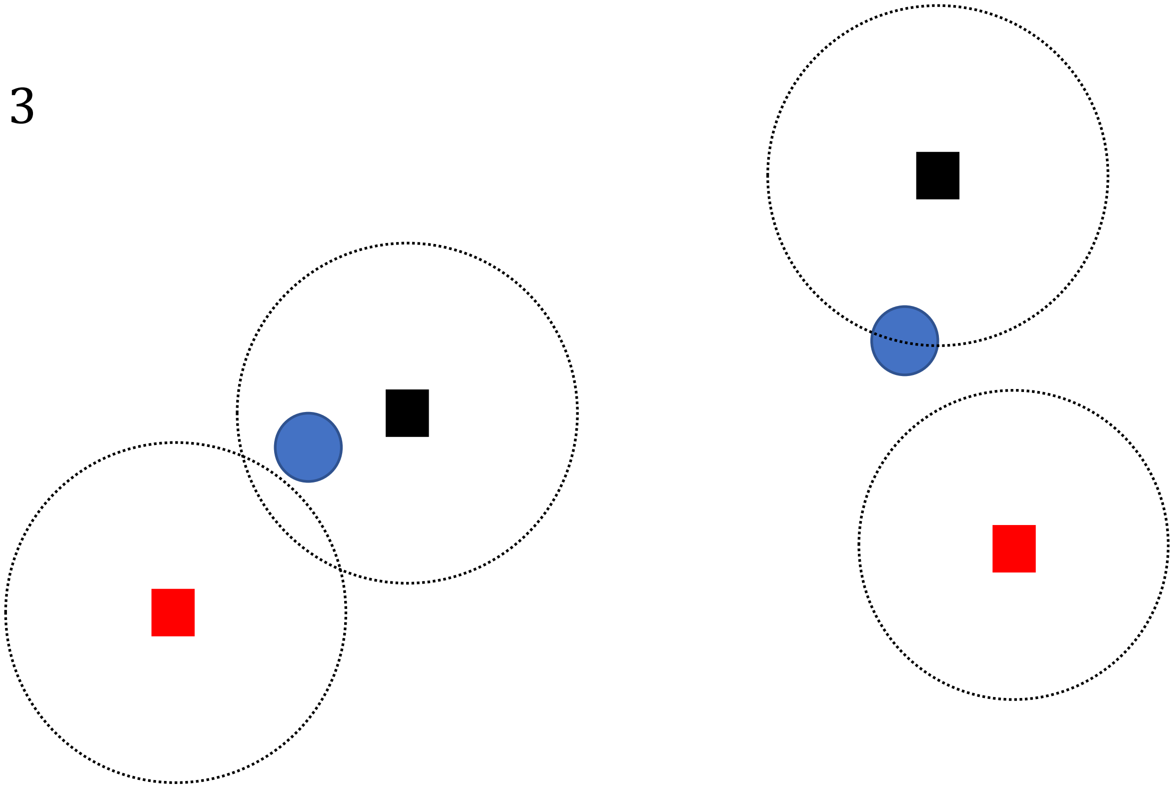
Greedy Capture

$k = 3$



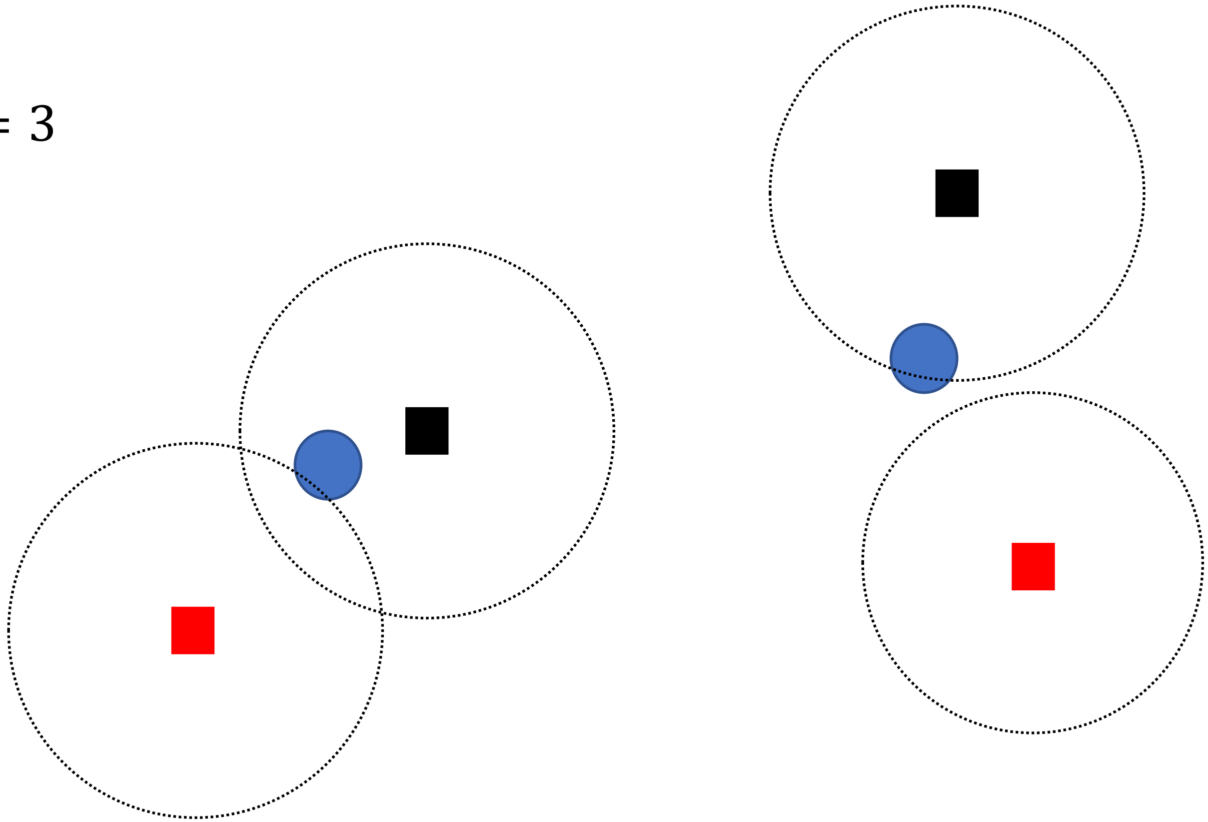
Greedy Capture

$k = 3$



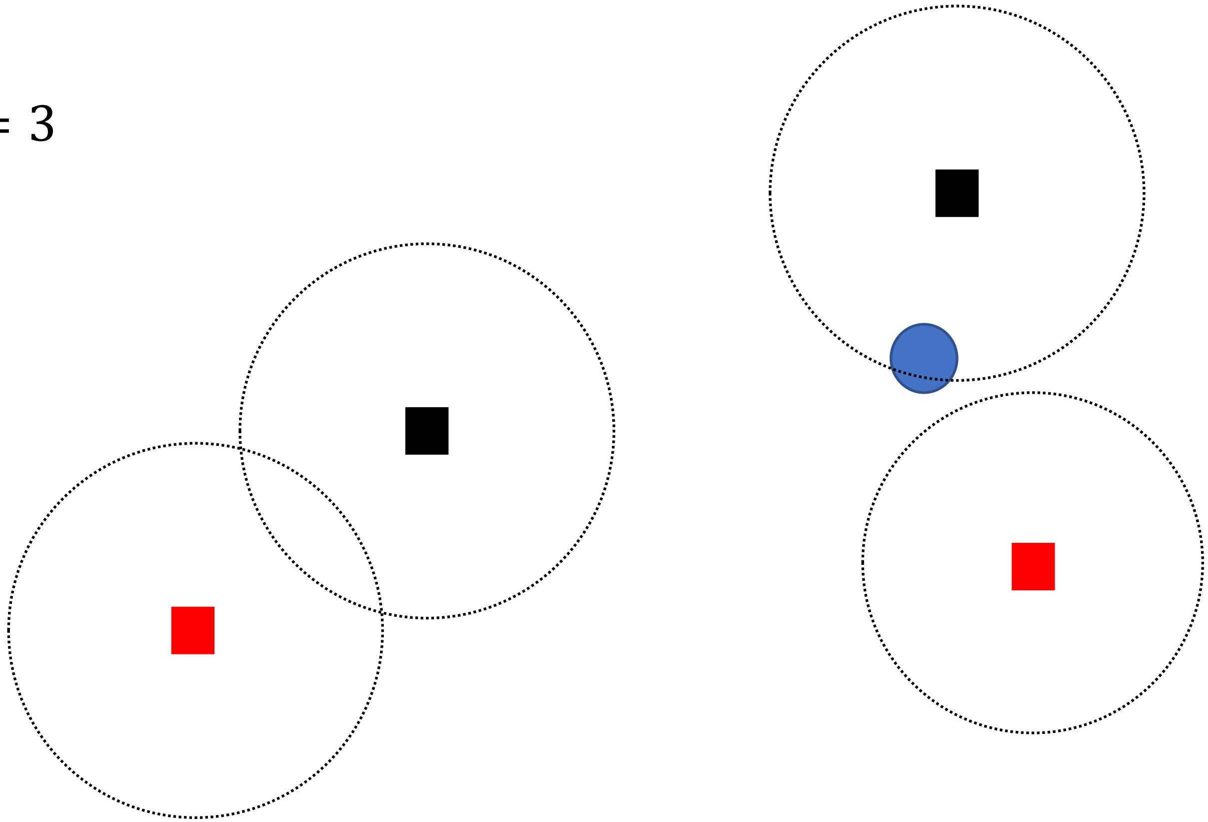
Greedy Capture

$k = 3$



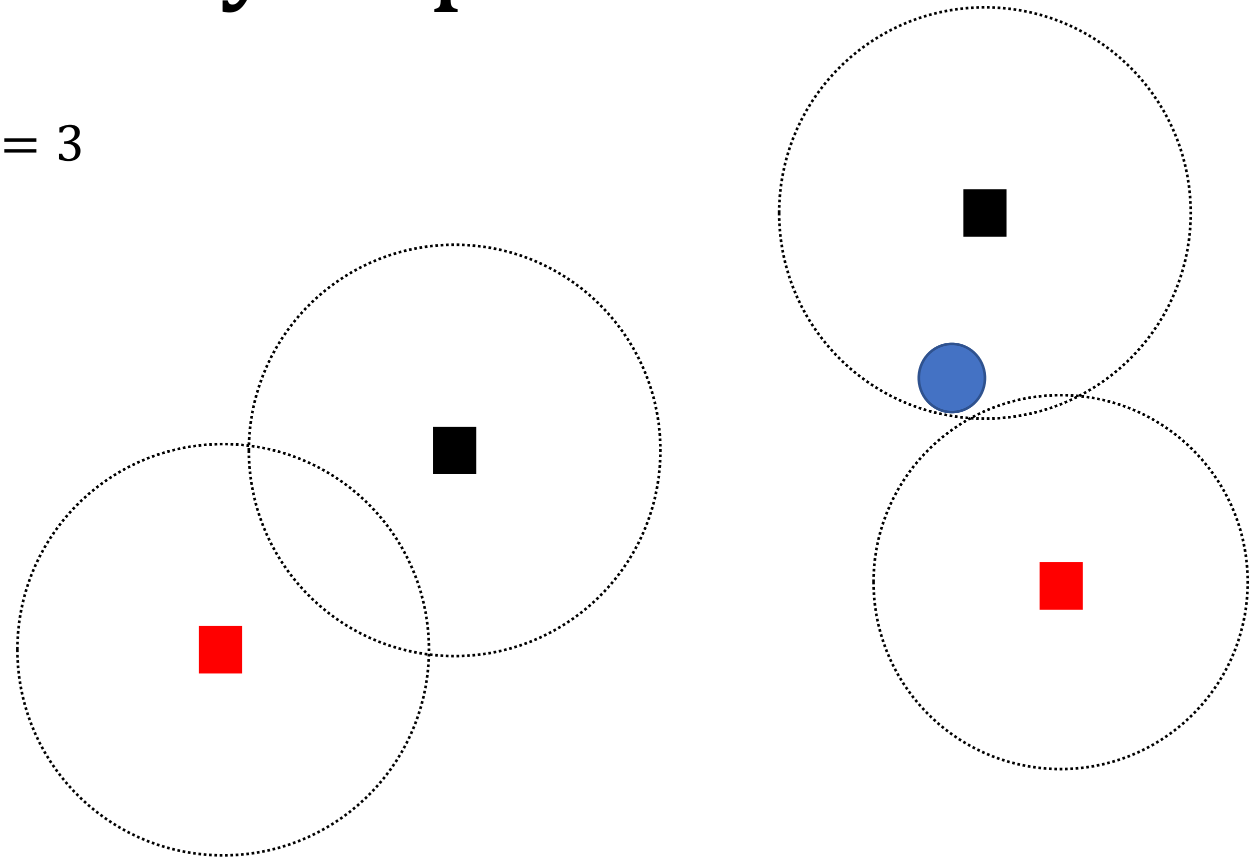
Greedy Capture

$k = 3$



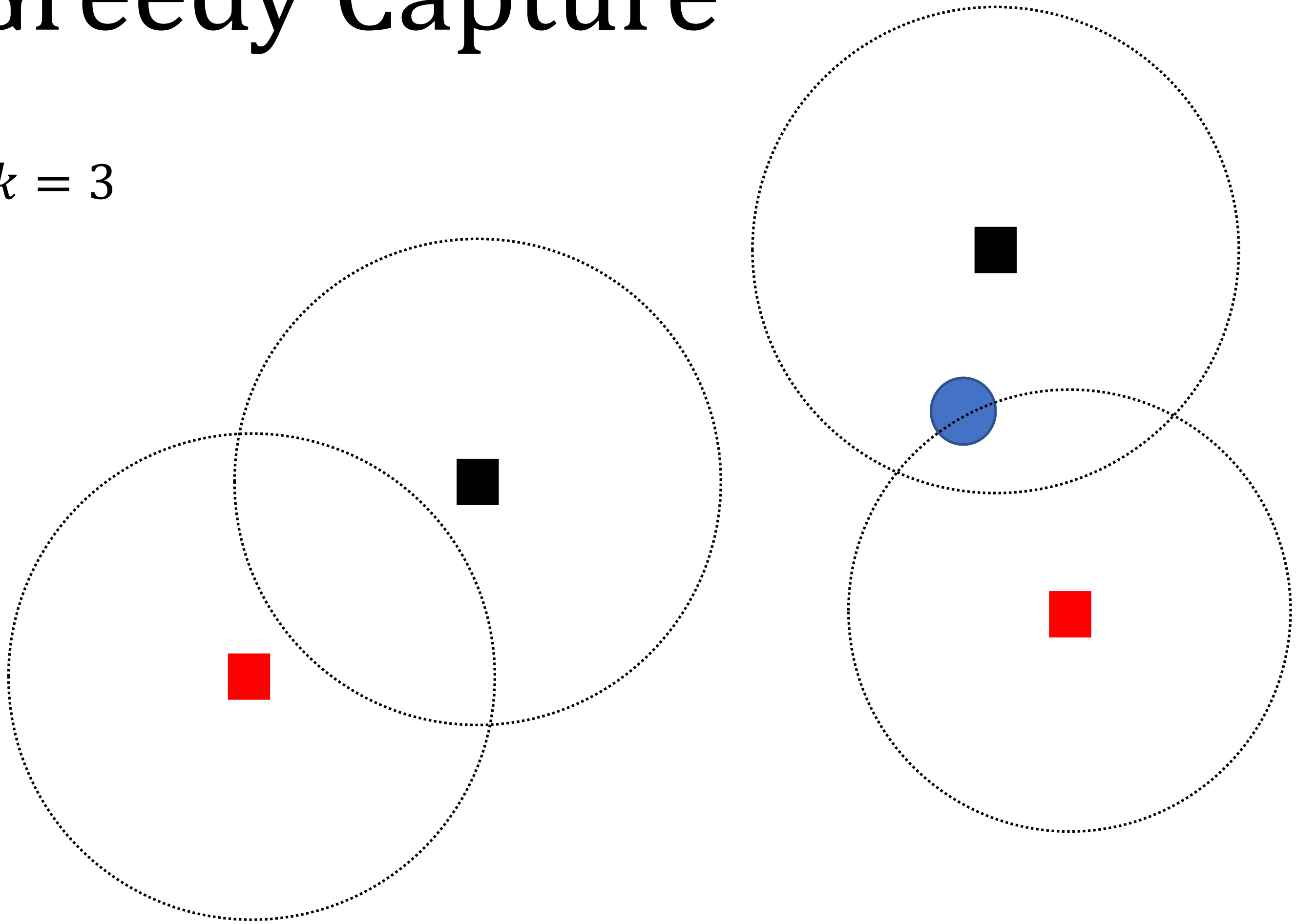
Greedy Capture

$k = 3$



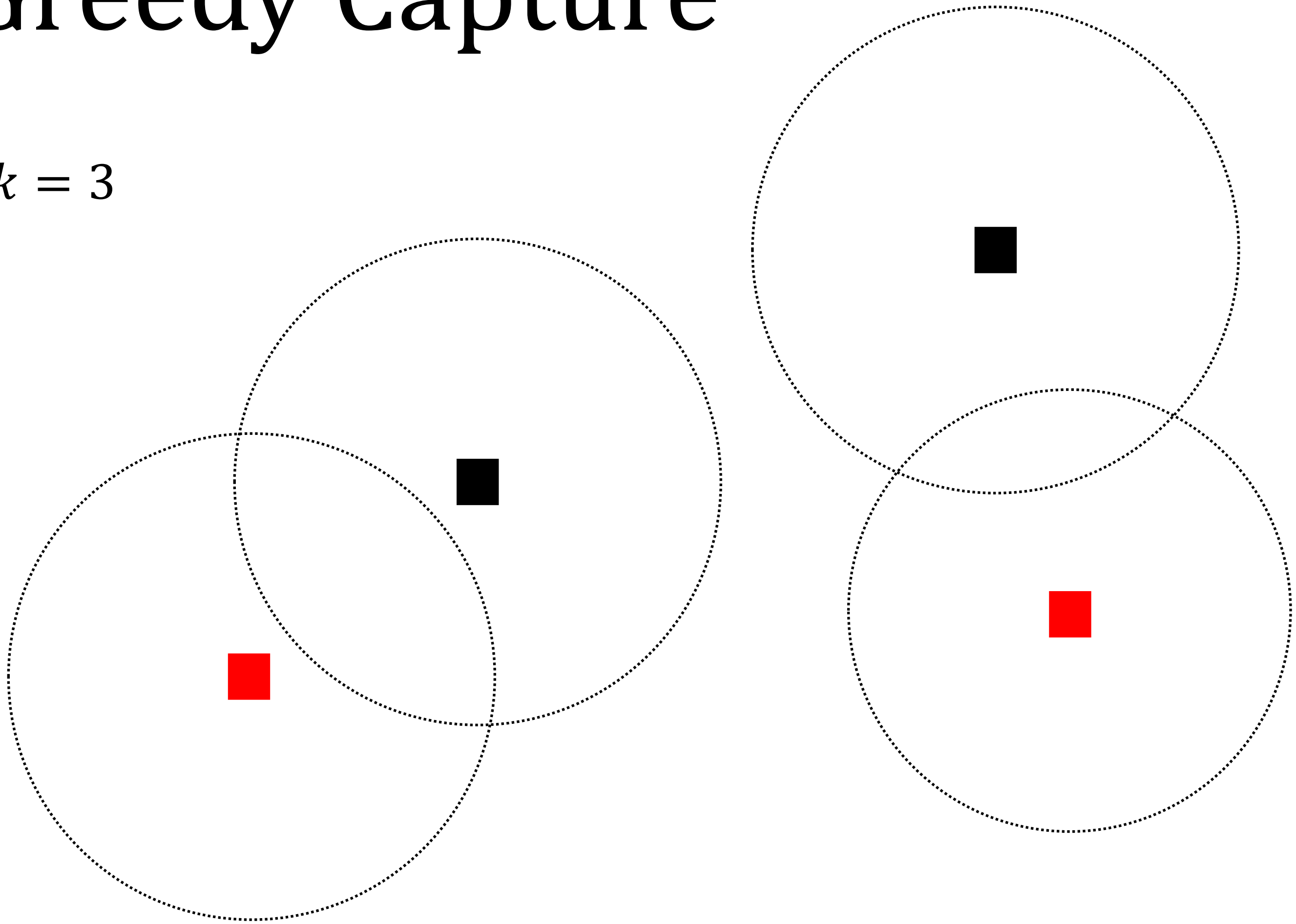
Greedy Capture

$k = 3$



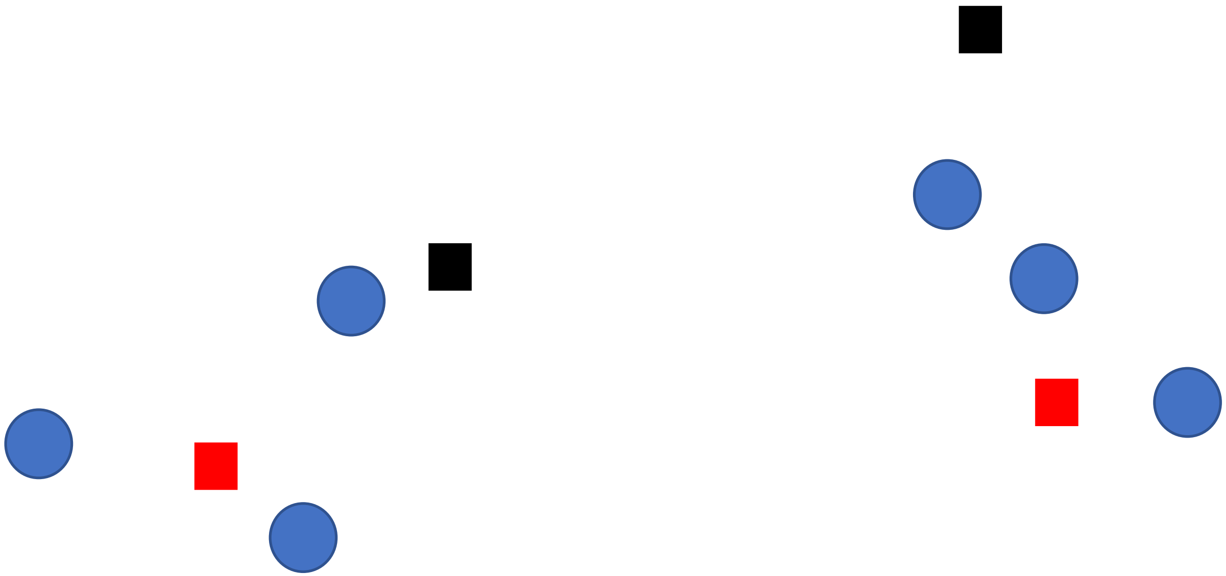
Greedy Capture

$k = 3$



Greedy Capture

$k = 3$



Greedy Capture

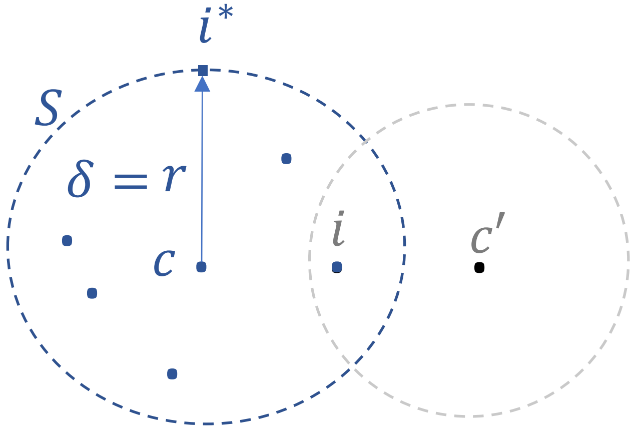
- $B(c, \delta)$ denotes the ball centered at c with radius δ

Greedy Capture

1. $\delta \leftarrow 0$; $C \leftarrow 0$
2. *While* $N \neq 0$ *do*
3. *Smoothly increase* δ
4. *While* $\exists\ c \in C$ *such that* $|B(c, \delta) \cap N| \geq 1$ *do*
5. $C: N \leftarrow N \setminus (B(c, \delta) \cap N)$
6. *While* $\exists\ c \in M \setminus C$ *such that* $|B(c, \delta) \cap N| \geq n/k$ *do*
7. $C \leftarrow C \cup c$
8. $N \leftarrow N \setminus (B(c, \delta) \cap N)$
9. *Return* C

Greedy Capture

- Theorem [Chen et al. '19]: Greedy Capture returns a clustering solution in the $(1+\sqrt{2})$-core.
- Proof:
- Let C be the solution that Greedy Capture returns
- Suppose for contradiction that there exists $S \subseteq N$, with $|S| \geq \frac{n}{k}$ and $c \in M \setminus C$, such that $\forall i \in S, \left(1+\sqrt{2}\right) \cdot d(i,c) < d(i, C(i))$



$$\min\left(\frac{d(i,c')}{d(i,c)}, \frac{d(i^*,c')}{d(i^*,c)}\right)$$

$$\leq \min\left(\frac{d(i,c')}{d(i,c)}, \frac{d(i^*,c)+d(c,c')}{d(i^*,c)}\right) \text{ (triangle inequality)}$$

$$\leq \min\left(\frac{d(i,c')}{d(i,c)}, \frac{d(i^*,c)+d(c,i)+d(i,c')}{d(i^*,c)}\right) \text{ (triangle inequality)}$$

$$\leq \min\left(\frac{d(i^*,c)}{d(i,c)}, 2+\frac{d(i,c)}{d(i^*,c)}\right) \left(d(i,c') \leq d(i^*,c)\right)$$

$$\leq \max_{z \geq 0}(\min(z, 2+1/z)) \leq 1+\sqrt{2}$$

Core

- Theorem [Chen et al. '19]: Greedy Capture returns a clustering solution in the $(1+\sqrt{2})$-core for any metric space
- Theorem [Chen et al. '19]: For all $\alpha < 2$ and all metric spaces, a clustering solution in the α-core is not guaranteed to exist
- Theorem [Chen et al. '19]: When $N = M$, for all $\alpha < 1.5$ and all metric spaces, a clustering solution in the α-core is not guaranteed to exist
- Theorem [M and Shah '20]: Greedy Capture returns a clustering solution in the 2-core for Euclidean metric space
- Theorem [M and Shah '20]: For Euclidean metric space, for all $\alpha < 1.155$, a clustering solution in the α-core is not guaranteed to exist
- Theorem [M and Shah '20]: For L_1 and L_∞ , for all $\alpha < 1.4$, a clustering solution in the α-core is not guaranteed to exist
- Theorem [M and Shah '20]: For Euclidean metric space, checking whether a clustering solution in the core exists is an NP-hard problem

Justified Representation

- Definition in Committee Selection: W satisfies JR if
  - For all $S \subseteq N$
  - If $|S| \geq n/k$ (large) and $|\cap_{i \in S} A_i| \geq 1$ (cohesive)
  - Then, $|A_i \cap W| \geq 1$ for some $i \in S$
  - "If a group deserves one candidate and has a commonly approved candidate, then not every member should get 0 utility"

- Definition in Clustering: C satisfies JR if
  - For all $S \subseteq N$
  - If $|S| \geq n/k$ (large) and $|\cap_{i \in S} B(i,r) \cap M| \geq 1$ (cohesive)
    - i.e. $\forall i \in S, d(i,c) \leq r$ for some $c \in M$
  - Then, $|B(i,r) \cap C| \geq 1$ for some $i \in S$
    - i.e. $d(i, C(i)) \leq r$ for some $i \in S$
  - "If a group deserves one cluster center and has a center that has distance at most r from each of them, then not every member should have distance larger than r from all the centers in the clustering "

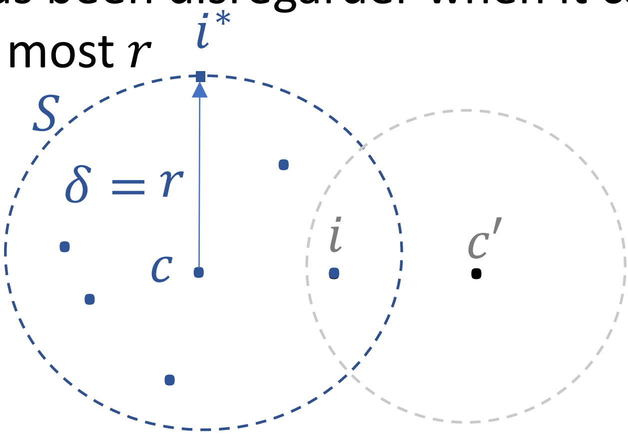
Justified Representation

- Question: What is the relationship between JR and core in clustering?

1. core $\Rightarrow$ JR
2. JR $\Rightarrow$ core
3. JR$=$core
4. JR $\neq$ core

Justified Representation

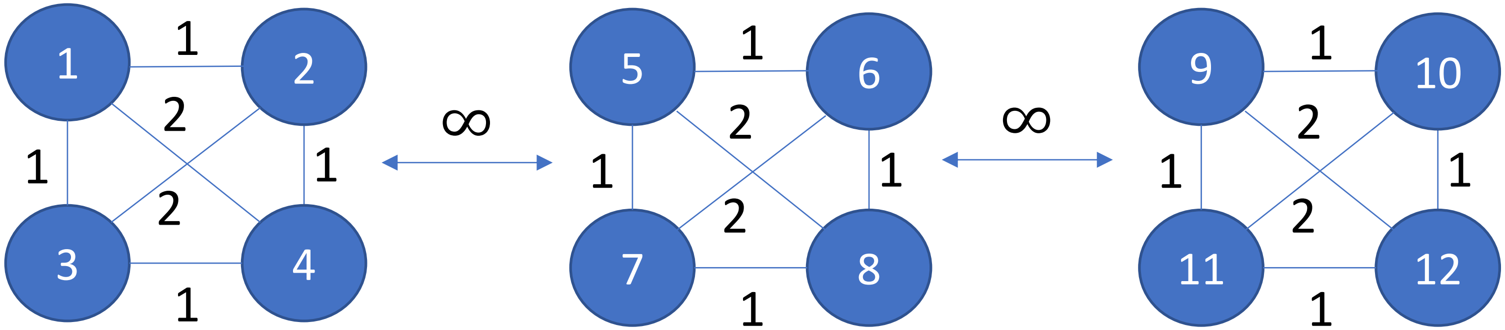
- Theorem [Kellerhals and Peters ‘24]: Greedy Capture returns a clustering solution that is JR
- Proof:
- Let C be the solution that Greedy Capture returns
- Suppose for contradiction that there exists $S \subseteq N$, with $|S| \geq \frac{n}{k}$ and $c \in M \setminus C$, such that $\forall i \in S,\ d(i,c) \leq r$ and $d\big(i, C(i)\big) > r$
- If none of $i \in S$ has been disregarded, then $|B(c,\delta)| \geq n/k$ and then c is included in the committee
- Otherwise, some of $i \in S$ has been disregarder when it captured from a ball centered at c with radius at most r



Individual Fairness

- Definition: C satisfies Individual Fairness (IF) if
  - $N = M$
  - Let $r_i = \min_{r \in \mathbb{R}} \{ |B(i,r) \cap N| \geq n/k \}$
  - For all $i \in N$, $|B(i, r_i) \cap C| \geq 1$
  - "Each individual expects a center within their proportional neighborhood"

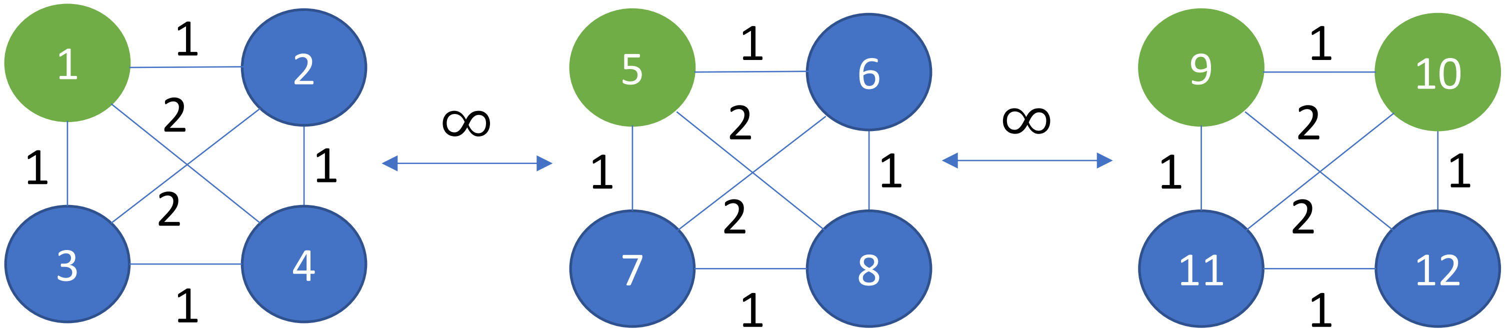
- Theorem [Jung et al. '19]: An individually fair clustering solution does not always exist
- Proof: $k = 4$



Individual Fairness

- Definition: C satisfies Individual Fairness (IF) if
  - $N = M$
  - Let $r_i = \min_{r \in \mathbb{R}} \{ |B(i, r) \cap N| \geq n/k \}$
  - For all $i \in N$, $|B(i, r_i) \cap C| \geq 1$
  - "Each individual expects a center within their proportional neighborhood"

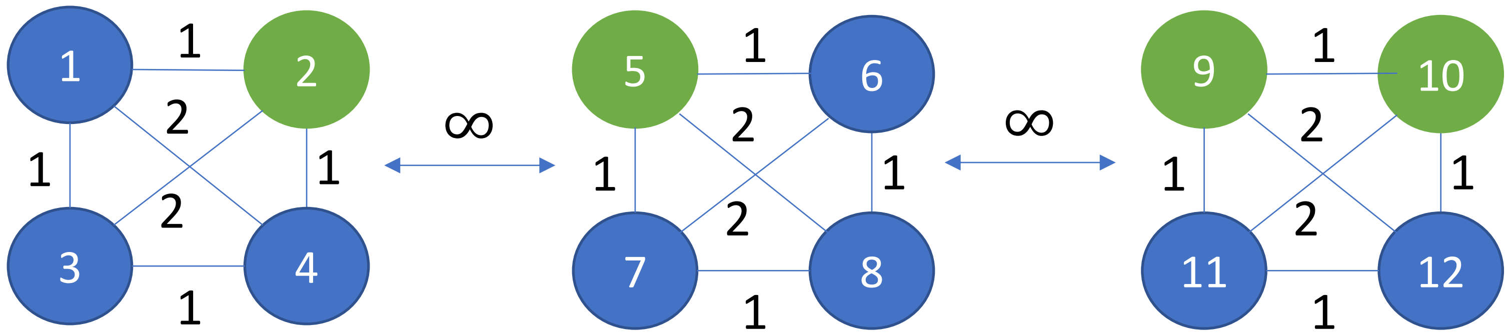
- Theorem [Jung et al. '19]: An individually fair clustering solution does not always exist
- Proof: $k = 4$



Individual Fairness

- Definition: C satisfies Individual Fairness (IF) if
  - $N = M$
  - Let $r_i = \min_{r \in \mathbb{R}} \{ |B(i, r) \cap N| \geq n/k \}$
  - For all $i \in N$, $|B(i, r_i) \cap C| \geq 1$
  - "Each individual expects a center within their proportional neighborhood"

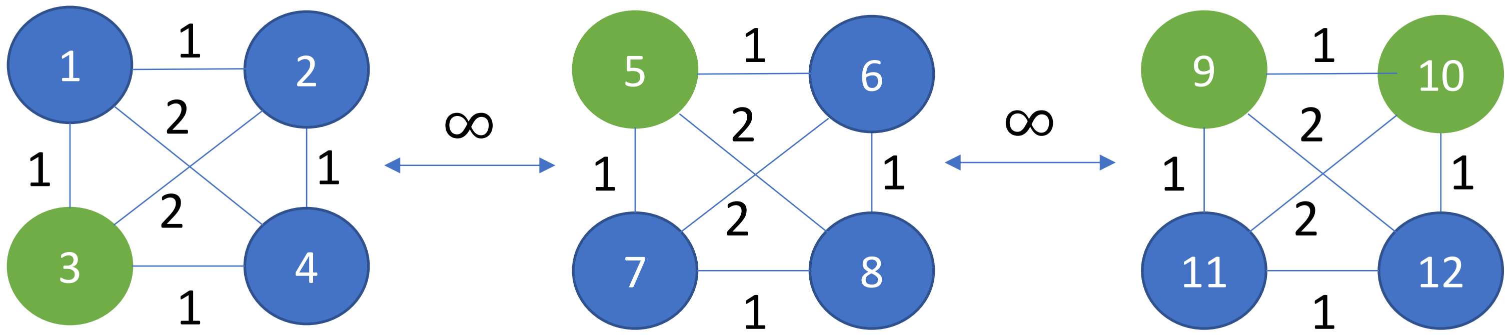
- Theorem [Jung et al. '19]: An individually fair clustering solution does not always exist
- Proof: $k = 4$



Individual Fairness

- Definition: C satisfies Individual Fairness (IF) if
  - $N = M$
  - Let $r_i = \min_{r \in \mathbb{R}} \{ |B(i,r) \cap N| \geq n/k \}$
  - For all $i \in N$, $|B(i,r_i) \cap C| \geq 1$
  - "Each individual expects a center within their proportional neighborhood"

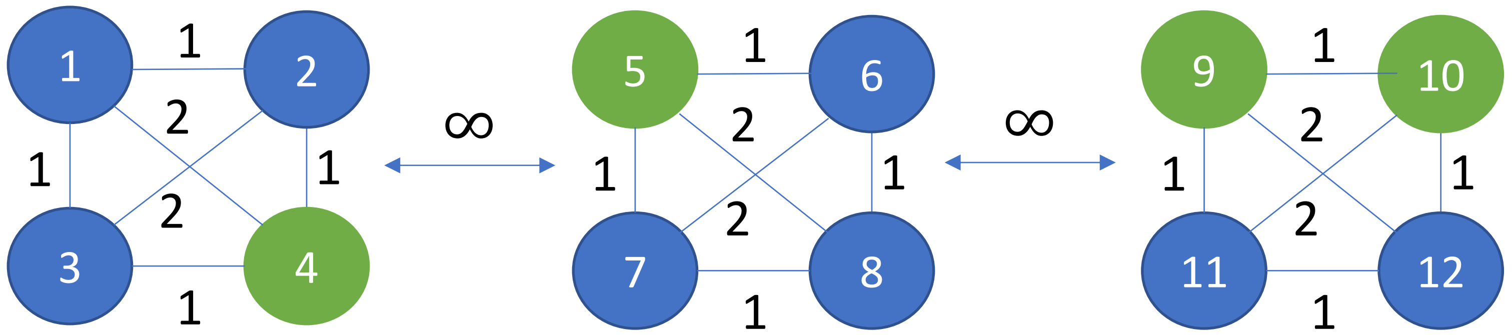
- Theorem [Jung et al. '19]: An individually fair clustering solution does not always exist
- Proof: $k = 4$



Individual Fairness

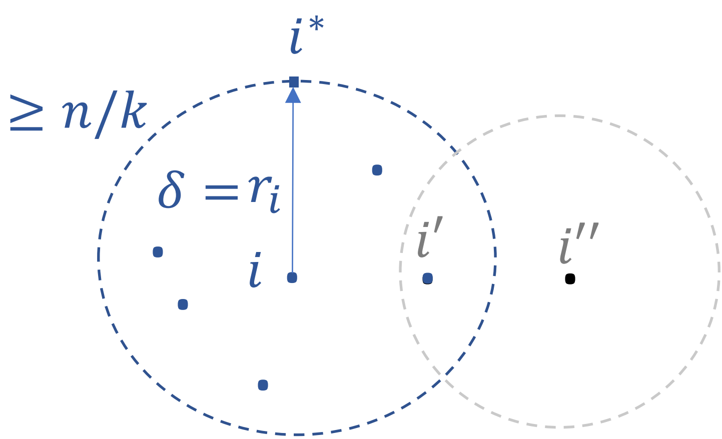
- Definition: C satisfies Individual Fairness (IF) if
  - $N = M$
  - Let $r_i = \min_{r \in \mathbb{R}} \{ |B(i,r) \cap N| \geq n/k \}$
  - For all $i \in N$, $|B(i, r_i) \cap C| \geq 1$
  - "Each individual expects a center within their proportional neighborhood"

- Theorem [Jung et al. '19]: An individually fair clustering solution does not always exist
- Proof: $k = 4$



Individual Fairness

- Theorem [Jung et al. '19]: Greedy Capture returns a clustering solution that is 2-IF
- Proof:
- Let C be the solution that Greedy Capture returns
- Suppose for contradiction that some $i \in N$, $|B(i, r_i) \cap C| = 0$
- If $|B(i, r_i)| \geq n/k$, then i is included in the solution
- Otherwise, some of $i' \in B(i, r_i)$has been disregarded when it captured from a ball centered at i'' with radius at most r_i
- From triangle inequality, $d(i, i'') \leq d(i, i') + d(i', i'') \leq 2 \cdot r_i$



Core, JR and IF

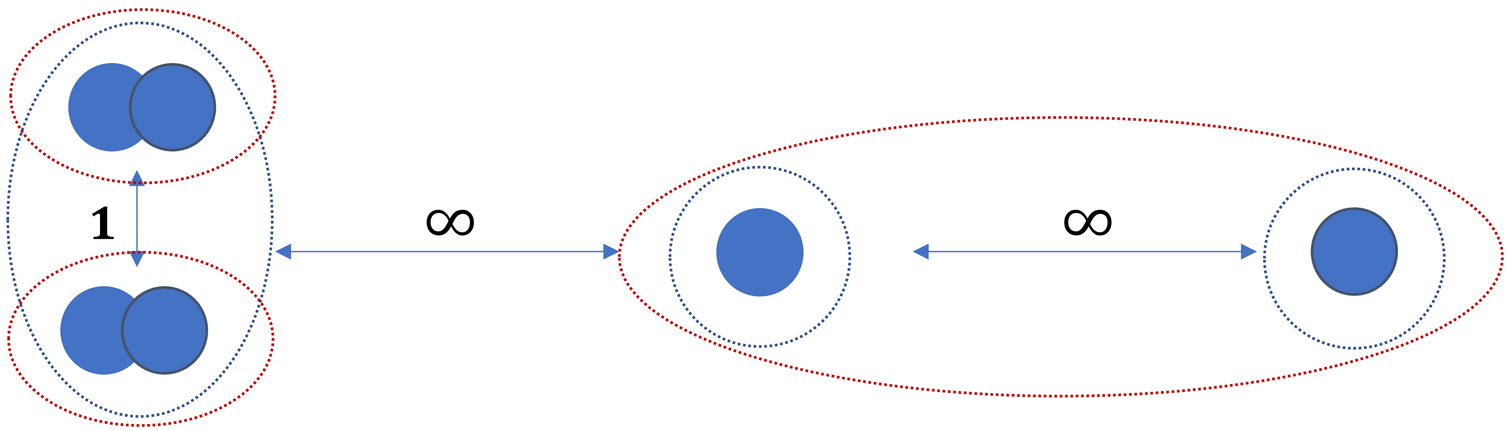
- Theorem: Greedy Capture returns a clustering solution that is JR, 2-IF and in the $1+\sqrt{2}$-core .

- Theorem [Kellerhals and Peters '24]: Any clustering solution that satisfies JR, it also satisfies 2-IF and is in the $1+\sqrt{2}$-core .

- Theorem [Kellerhals and Peters '24]:
  - ❑ Any clustering solution that satisfies α-IF, it is also in the $2 \cdot \alpha$-core
  - ❑ Any clustering solution that is in the α-core, it also satisfies $(1+\alpha)$-IF

Core, JR and IF vs k-means, k-median, k-center

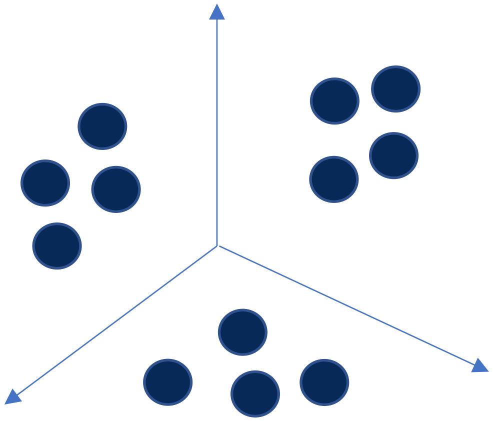
$k = 3$



Non-Centroid Clustering

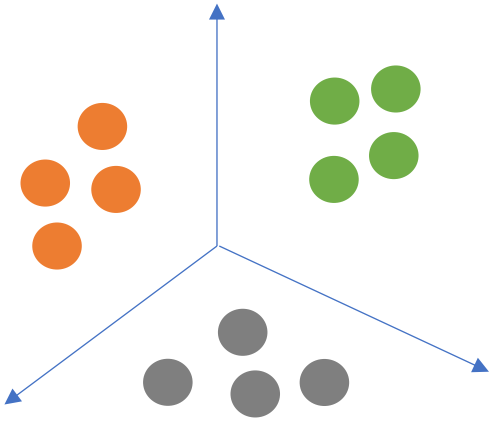
Non-Centroid Clustering

- Input:
  - Set N of n data points



Non-Centroid Clustering

- Input:
  - Set N of n data points
- Output:
  - Partition the individuals into k clusters, $C = \{C_1, \dots, C_k\}$



- Goal: Similar individuals are assigned to the same cluster

Core in Non-Centroid Clustering

- Definition in Committee Selection: W is in the core if
  - For all $S \subseteq N$ and $T \subseteq M$
  - If $|S| \geq |T| \cdot n/k$ (**large**)
  - Then, $u_i(W) \geq u_i(S)$ for some $i \in S$

- Definition in Non-Centroid Clustering: C is in the α-core, with $\alpha \geq 1$, if
  - For all $S \subseteq N$
  - If $|S| \geq n/k$ (**large**)
  - Then, $\ell_i\big(C(i)\big) \leq \alpha \cdot \ell_i(S)$ for some $i \in S$

- **Average Loss:** For each $S \subseteq N$, $\ell_i(S) = \frac{1}{|S|} \sum_{i' \in S} d(i, i')$

- **Maximum Loss:** For each $S \subseteq N$, $\ell_i(S) = \max_{i' \in S} d(i, i')$

Core in Non-Centroid Clustering

- Theorem [Caragiannis et al. ‘24]:

- Average Cost
    - A variation of Greedy Capture returns a clustering solution in the $O(n/k)$-core
    - For $\alpha < 1.3$, a clustering solution in the α-core is not guaranteed to exist

  - Maximum Cost
    - A variation of Greedy Capture returns a clustering solution in the 2-core

- **Open Questions:**
  - Average Cost: Does a clustering solution in the $O(1)$-core always exist?
  - Maximum Cost: Does a clustering solution in the core always exist?

FJR in Non-Centroid Clustering

- Definition in Committee Selection: W satisfies FJR if
  - For all $S \subseteq N, T \subseteq M$ and $\ell, \beta \in \{1, \dots, k\}$
  - If $|S| \geq |T| \cdot {}^{n}/_{k}$ (large) and $u_i(T) \geq \beta, \forall i \in S$ (cohesive)
  - Then, $u_i(W) \geq \beta$ for some $i \in S$

- Definition in Non-Centroid Clustering: C satisfies α-FJR, with $\alpha \geq 1$, if
  - For all $S \subseteq N$ and $\beta \in \mathbb{R}$
  - If $|S| \geq n/k$ (**large**) and $\ell_i(S) \leq \beta, \forall i \in S$ (cohesive)
  - Then, $\ell_i(C(i)) \leq \alpha \cdot \beta$ for some $i \in S$

- **Average Loss:** For each $S \subseteq N$, $\ell_i(S) = \frac{1}{|S|} \sum_{i' \in S} d(i, i')$

- **Maximum Loss:** For each $S \subseteq N$, $\ell_i(S) = \max_{i' \in S} d(i, i')$

FJR in Non-Centroid Clustering

Greedy Cohesive Algorithm

1. $N' \leftarrow N$
2. $j \leftarrow 0$
3. *While* $|N'| \geq n/k$
4. $\quad j \leftarrow j + 1$
5. $\quad C_j \leftarrow argmin_{S \subseteq N' : |S| \geq n/k} \max_{i \in S} \ell_i(S)$
6. $\quad N' \leftarrow N' \setminus S$
7. *If* $|N'| \geq 0$
8. $\quad j \leftarrow j + 1$
9. $\quad C_j \leftarrow N'$
10. *Return* $\{C_1, \ldots, C_j\}$

FJR in Non-Centroid Clustering

- Theorem [Caragiannis et al. '24]: Greedy Cohesive Algorithms returns a clustering solution that is FJR
- Proof:
- Let C be the solution that Greedy Capture returns
- Suppose for contradiction that there exists $S \subseteq N$, with $|S| \geq n/k$ such that
  - $\ell_i(S) \leq \beta, \forall i \in S$ (cohesive)
  - $\ell_i(C(i)) > \beta$ for all $i \in S$
- Let i^* be the first agent in S that was assigned to a cluster C_j
- Then we have that $\max_{i \in C_j} \ell_i(C_j) \geq \ell_{i^*}(C_j) > \beta$
- But then the algorithm would choose S instead of C_j

FJR in Non-Centroid Clustering

- Theorem [Caragiannis et al. ‘24]:
  - Greedy Cohesive Algorithm returns a clustering solution that satisfies FJR
  - Average Cost
    - Greedy Capture returns a clustering solution that satisfies 4-FJR
  - Maximum Cost
    - Greedy Capture returns a clustering solution that satisfies 2-FJR

- **Open Question:**
  - Can we efficiently find a solution that satisfies FJR?

Classic Objectives

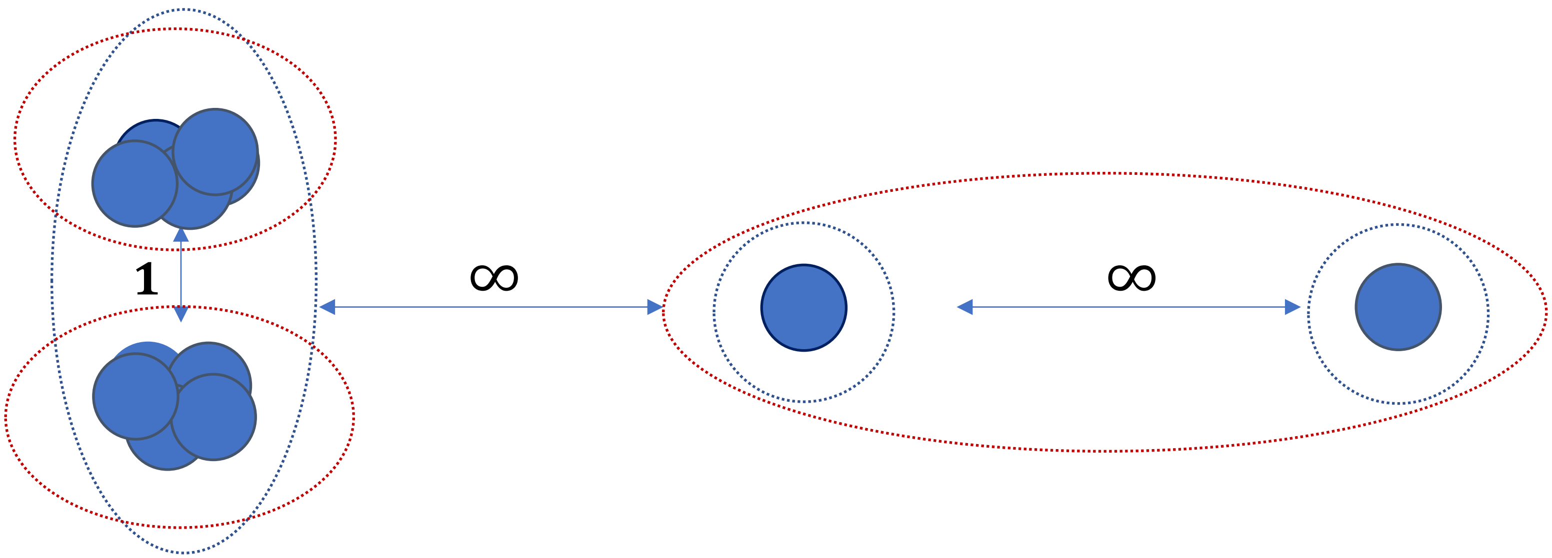
- k-median: Minimizes the within-cluster sum of distances
  - $\min_{C} \sum_{j\in[k]} \frac{1}{|C_j|} \sum_{i,i'\in C_j} d\,(i,i')$

- k-means: Minimizes the within-cluster of the square of the distances
  - $\min_{C} \sum_{j\in[k]} \frac{1}{|C_j|} \sum_{i,i'\in C_j} d^2(i,i')$

- k-center: Minimizes the maximum distance
  - $\min_{\substack{C:\\ |C|\leq k}} \max_{i\in N} d(i, C(i))$

Proportional Fairness vs Classic Objectives

$k = 3$



Envy-Freeness

[Ahmadi, Awasthi, Khuller, Kleindessner, Morgenstern, Sukprasert, Vakilian, 2022]

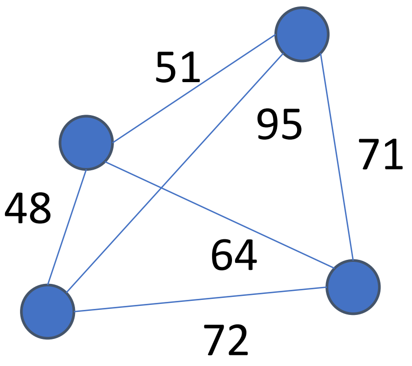
[Aamand, Chen, Liu, Silwal, Sukprasert, Vakilian, Zhang, 2023]

- α-Envy-freeness: For each $i \in N$ and $j \in [k]$ with $i \notin C_j$, either $C(i) = \{i\}$ or

$$\frac{1}{|C(i)| - 1} \sum_{i' \in C(i)} d(i, i') \leq \frac{1}{|C_j|} \sum_{i' \in C_j} d(i, i')$$

- Theorem: A envy-free clustering does not always exist
- Proof:

k=2



Envy-Freeness

[Ahmadi, Awasthi, Khuller, Kleindessner, Morgenstern, Sukprasert, Vakilian, 2022]

[Aamand, Chen, Liu, Silwal, Sukprasert, Vakilian, Zhang, 2023]

- α-Envy-freeness: For each $i \in N$ and $j \in [k]$ with $i \notin C_j$, either $C(i) = \{i\}$ or

$$\frac{1}{|C(i)| - 1} \sum_{i' \in C(i)} d(i, i') \leq \frac{1}{|C_j|} \sum_{i' \in C_j} d(i, i')$$

- Theorem: Deciding if there exists an envy-free solution is an NP-hard problem

Envy-Freeness

[Ahmadi, Awasthi, Khuller, Kleindessner, Morgenstern, Sukprasert, Vakilian, 2022]

[Aamand, Chen, Liu, Silwal, Sukprasert, Vakilian, Zhang, 2023]

- α-Envy-freeness: For each $i \in N$ and $j \in [k]$ with $i \notin C_j$, either $C(i) = \{i\}$ or

$$\frac{1}{|C(i)|-1}\sum_{i' \in C(i)} d(i,i') \leq \frac{\alpha}{|C_j|}\sum_{i' \in C_j} d(i,i')$$

- Theorem: An $O(1)$-envy-free clustering always does (and can be computed efficiently)

Envy-Freeness

[Ahmadi, Awasthi, Khuller, Kleindessner, Morgenstern, Sukprasert, Vakilian, 2022]

[Aamand, Chen, Liu, Silwal, Sukprasert, Vakilian, Zhang, 2023]

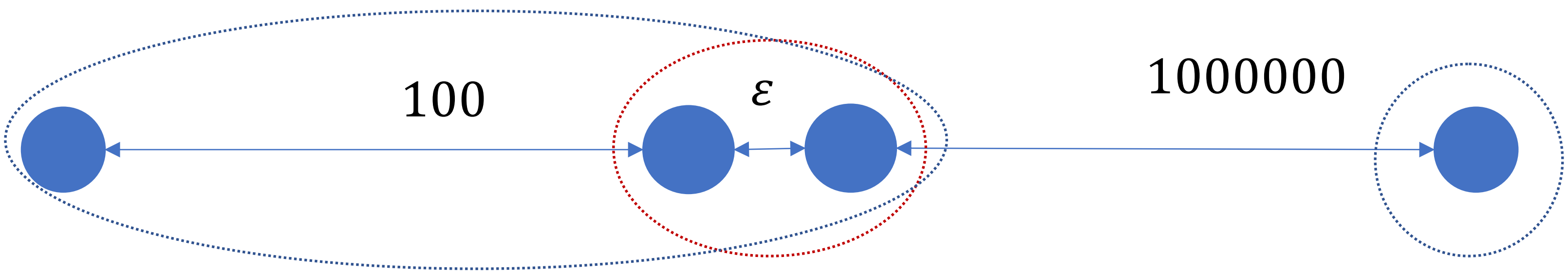
- α-Envy-freeness: For each $i \in N$ and $j \in [k]$ with $i \notin C_j$, either $C(i) = \{i\}$ or

$$\ell_i(C(i) \setminus \{i\}) \leq \ell_i(C(j))$$

- **Average Loss:** For each $S \subseteq N$, $\ell_i(S) = \frac{1}{|S|} \sum_{i' \in S} d(i, i')$
- **Maximum Loss:** For each $S \subseteq N$, $\ell_i(S) = \max_{i' \in S} d(i, i')$
- **Minimum Loss:** For each $S \subseteq N$, $\ell_i(S) = \min_{i' \in S} d(i, i')$

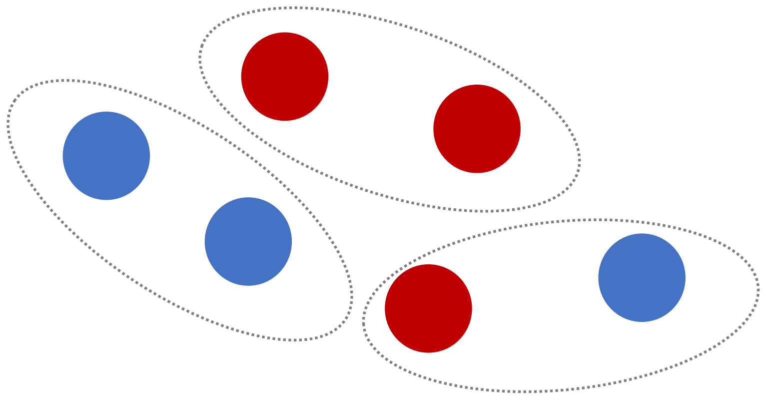
Core vs Envy-Freeness

$k = 2$



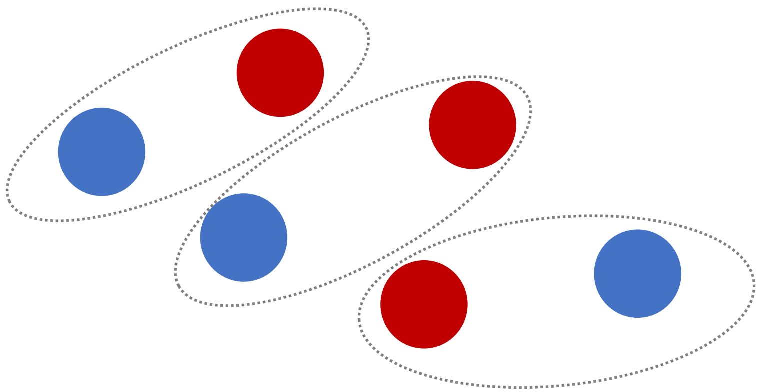
Demographic Fairness [Chierichetti et al. 2017]

- Demographic Groups:
  - There is predefined set of protected groups (e.g. race or gender)
  - Each individual/data point belongs to one group
  - Disparate Impact in ML: The impact of a system across protected groups
  - Disparate Impact in Clustering: The impact on a group is measured by how many individuals of that group are in each cluster



Demographic Fairness [Chierichetti et al. 2017]

- Demographic Groups:
  - There is predefined set of protected groups (e.g. race or gender)
  - Each individual/data point belongs to one group
  - Disparate Impact in ML: The impact of a system across protected groups
  - Disparate Impact in Clustering: The impact on a group is measured by how many individuals of that group are in each cluster



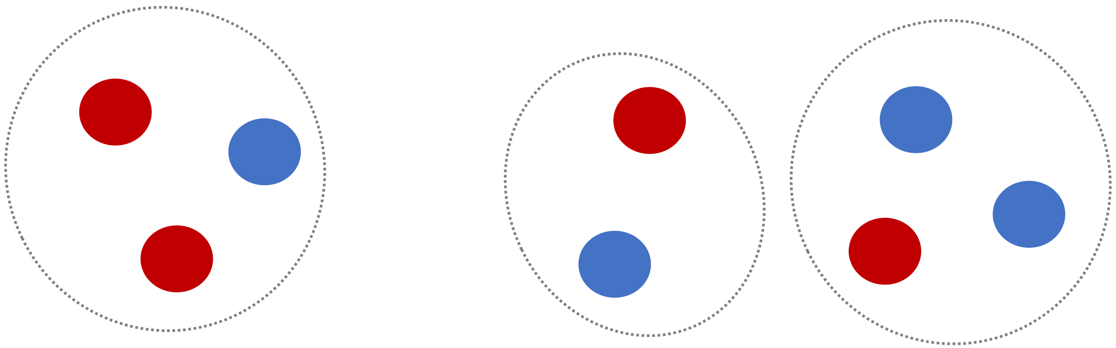
Balancedness [Chierichetti et al. 2017]

- Let $G_1, \dots, G_t$ be the protected groups
- Let $C = \{C_1, \dots, C_k\}$ be a clustering solution
- The balancedness in each cluster C_j is measured as:

$$balance(C_j) = \min_{i \neq i' \in [t]} \frac{|G_i \cap C_j|}{|G_{i'} \cap C_j|}$$

- The balancedness of a clustering solution $C = \{C_1, \dots, C_t\}$ is measured as:

$$balance(C) = \min_{j \in [k]} balance(C_j)$$



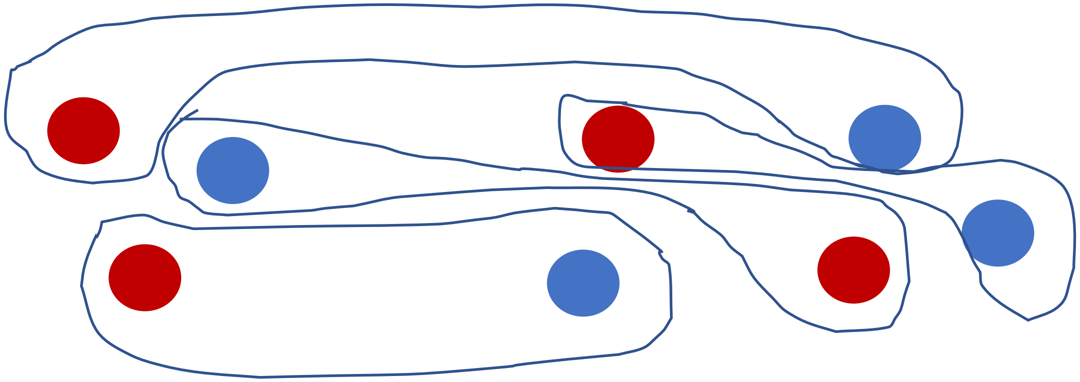
Balancedness [Chierichetti et al. 2017]

- Let $G_1, \dots, G_t$ be the protected groups
- Let $C = \{C_1, \dots, C_k\}$ be a clustering solution
- The balancedness in each cluster C_j is measured as:

$$balance(C_j) = \min_{i \neq i' \in [t]} \frac{|G_i \cap C_j|}{|G_{i'} \cap C_j|}$$

- The balancedness of a clustering solution $C = \{C_1, \dots, C_t\}$ is measured as:

$$balance(C) = \min_{j \in [k]} balance(C_j)$$



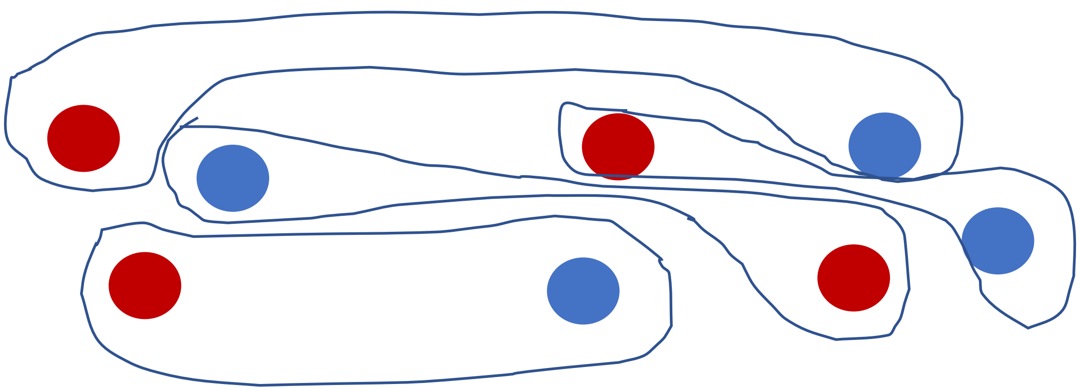
Balancedness [Chierichetti et al. 2017]

- Let $G_1, \dots, G_t$ be the protected groups
- Let $C = \{C_1, \dots, C_k\}$ be a clustering solution
- The balancedness in each cluster C_j is measured as:

$$balance(C_j) = \min_{i \neq i' \in [t]} \frac{|G_i \cap C_j|}{|G_{i'} \cap C_j|}$$

- The balancedness of a clustering solution $C = \{C_1, \dots, C_t\}$ is measured as:

$$balance(C) = \min_{j \in [k]} balance(C_j)$$



Balancedness [Chierichetti et al. 2017]

- Theorem:
  - k-center: 4 $-$approximation with balance $1/t$
  - k-median:$2+\sqrt{3}$ $-$approximation with balance 1
  - k-median:t $+ 2 + \sqrt{3}$ $-$approximation with balance $1/t$

Bounded Representation [Bercea et al. 2019]

- Let $G_1, \dots, G_t$ be the protected groups
- Let $C = \{C_1, \dots, C_k\}$ be a clustering solution
- For (α, β)- bounded representation we require that

$$\alpha \leq \left|G_i \cap C_j\right| \leq \beta, \quad \forall i \in [t] \text{ and } \forall j \in [k]$$

- Standard objectives such as k-center, k-median and k-means are maximized subject to (α, β)- bounded representation constraints

Socially Fair Clustering [Makarychev et al. 2021]

- Let $G_1, \dots, G_t$ be the protected groups (not necessarily disjoint)
- Measure the ℓ_p −loss for each group, i.e.

  Loss(G_j)= $\sum_{i \in G_j} d(i, C)^p$
- Goal: Minimize the maximum loss over all the t groups

- Theorem: There exists a polynomial time algorithm that finds a $O(e^{O(p)} \frac{\log t}{\operatorname{loglog} t})$-approximation to the socially fair ℓ_p clustering problem

Fair Range [Hotegni et al. 2023]

- Let $G_1, \ldots, G_t$ be the protected groups
- Fair Range: $\alpha_j \leq \left|C \cap G_j\right| \leq \beta_j$
- Theorem: There exists a polynomial time algorithm that finds constant approximation ℓ_p clustering problem with fair range for any $p \in [1, \mathrm{in}f]$