Clustering with Fair-Center Representation

Parameterized Approximation Algorithms and Heuristics

Ameet Gadekar
16 Nov 2022

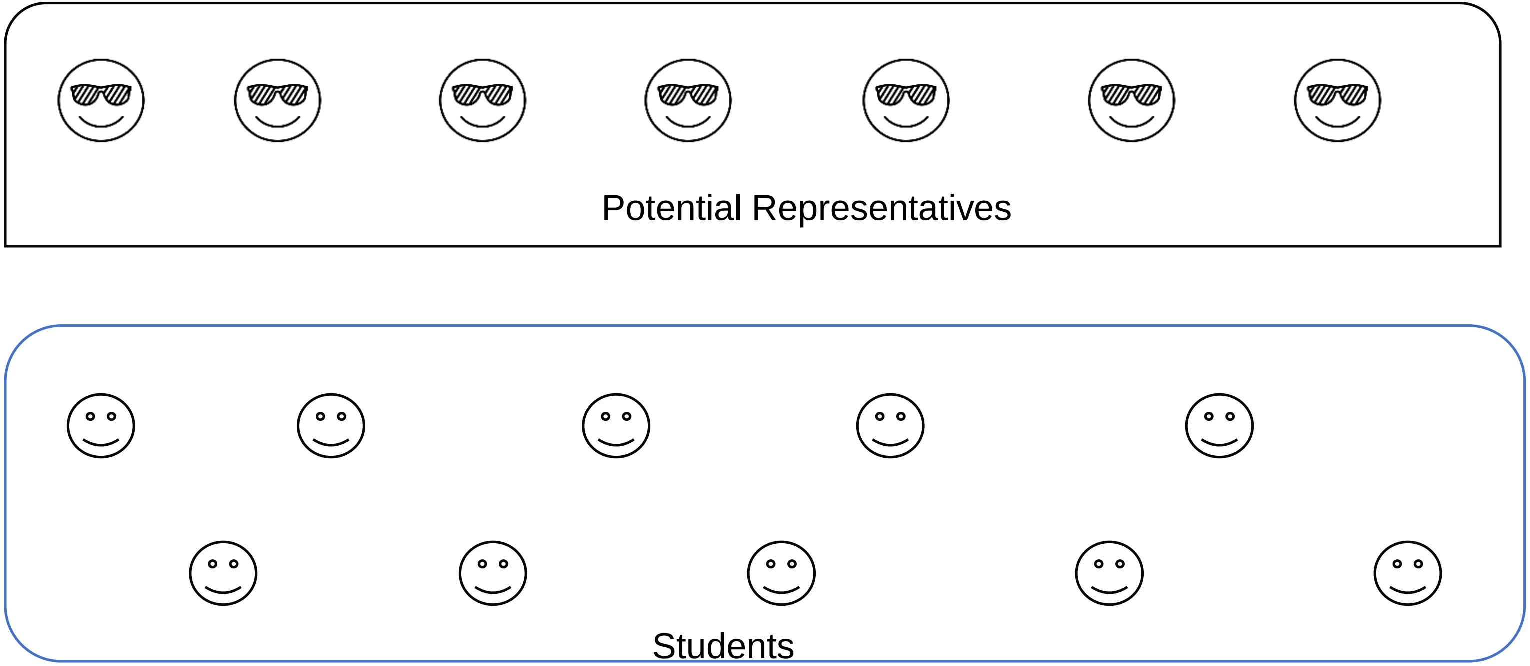
Joint work with

Suhas Thejaswi, Aalto
Bruno Ordozgoiti, QMUL
Michal Osadnik, Aalto

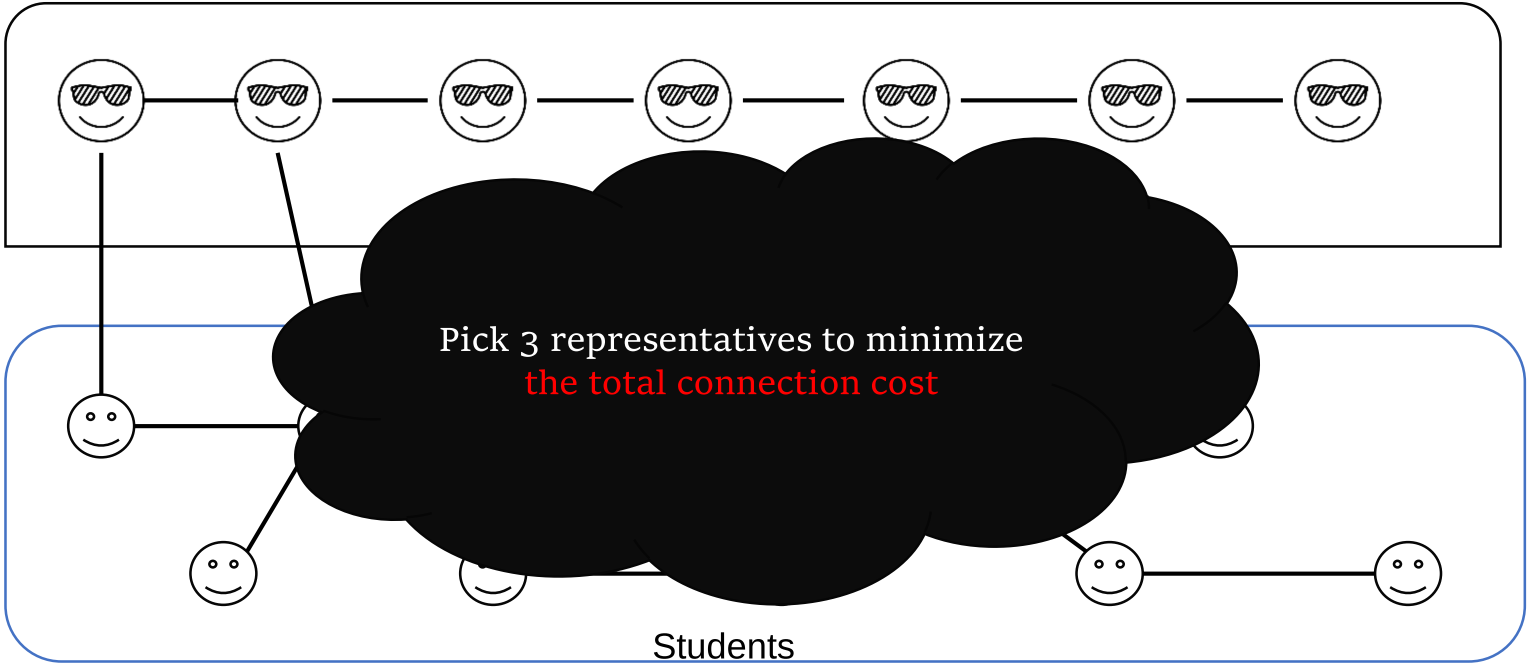
Aalto University
School of Science

Appeared at KDD'22

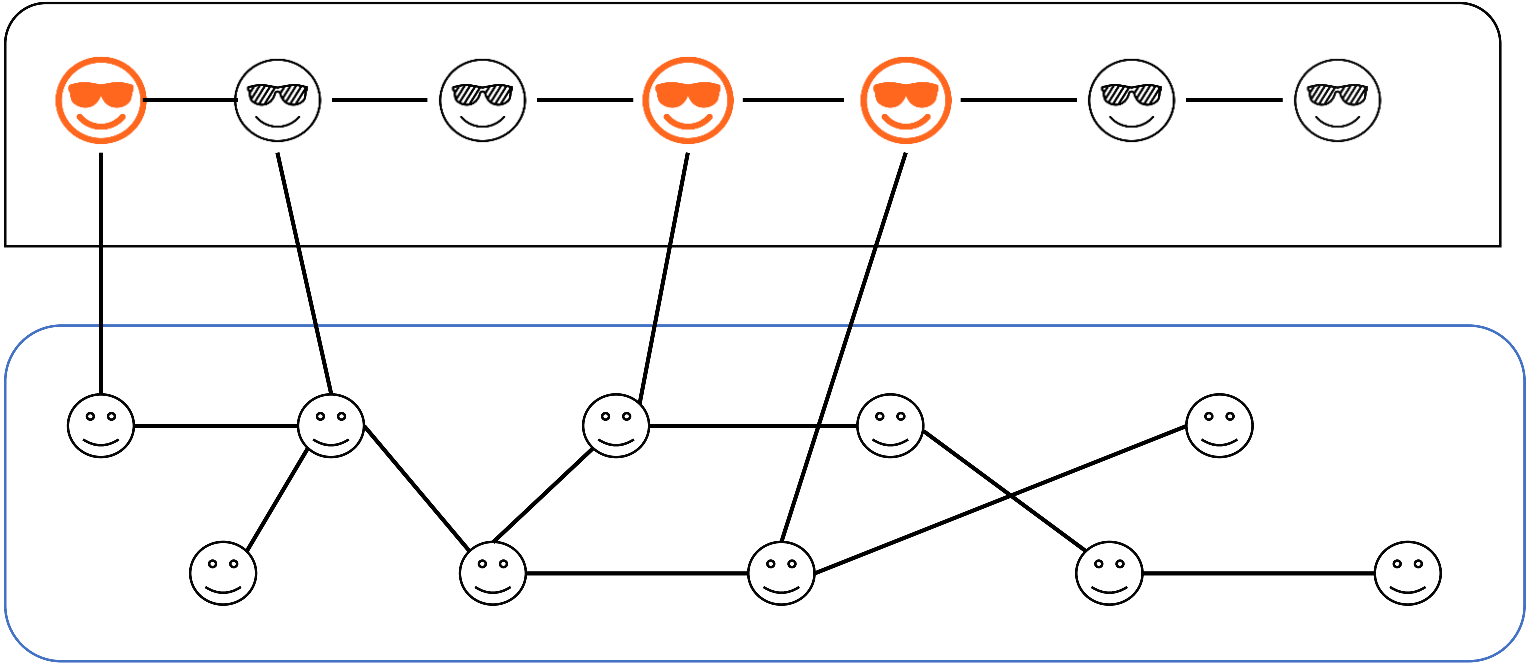
Committee Selection



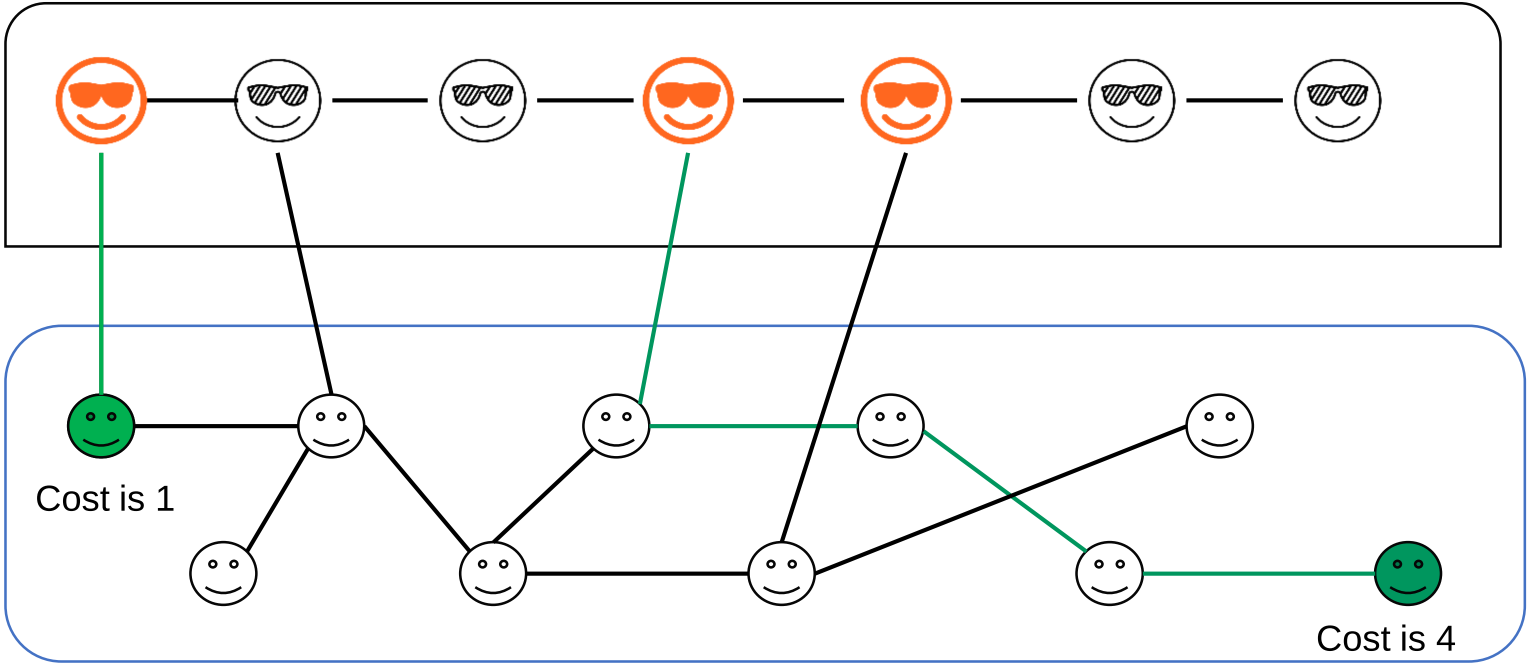
Committee Selection



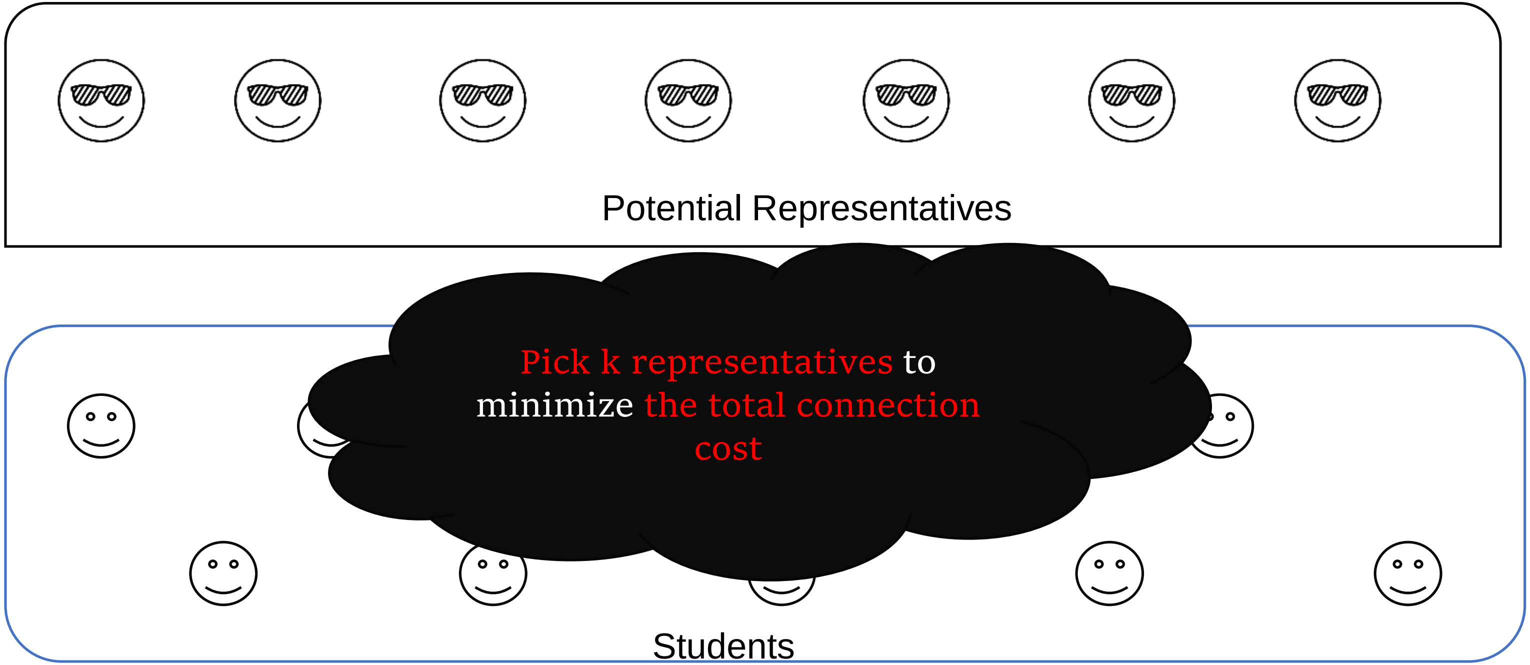
Committee Selection



Committee Selection

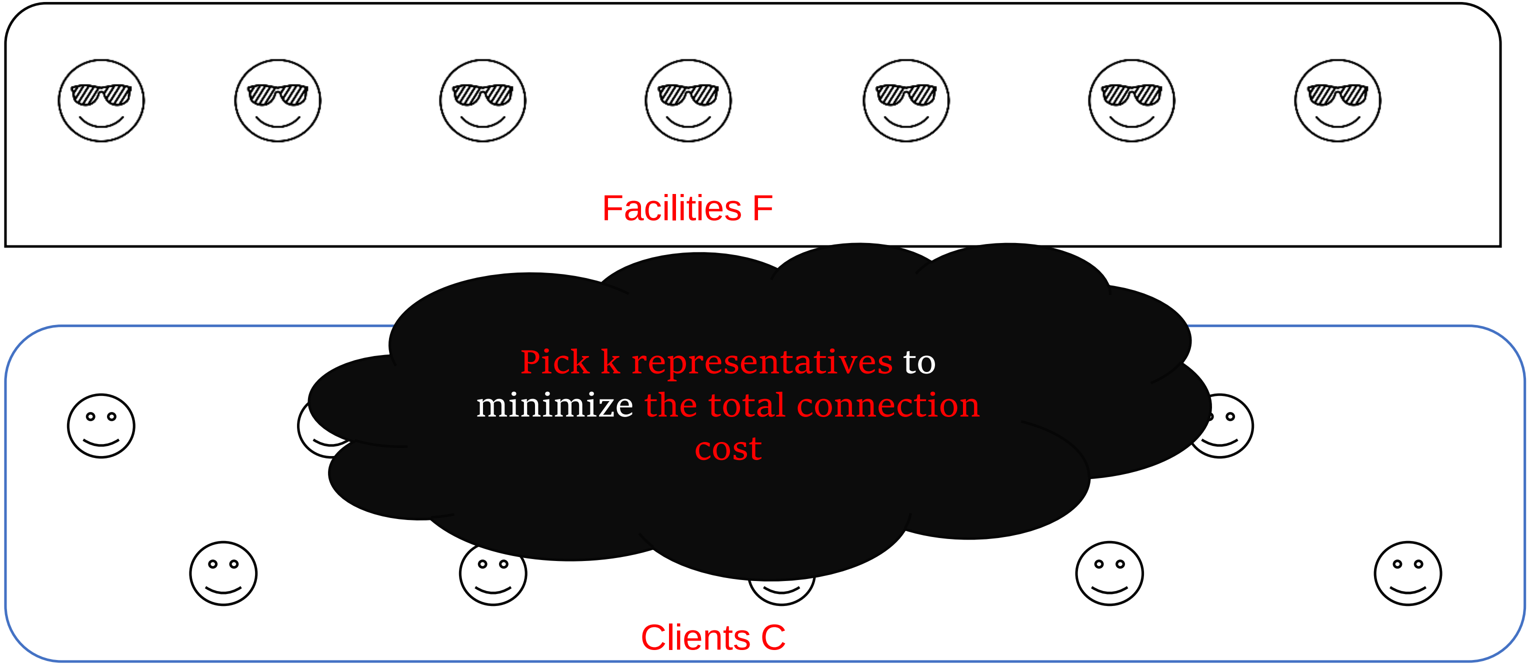


Committee Selection

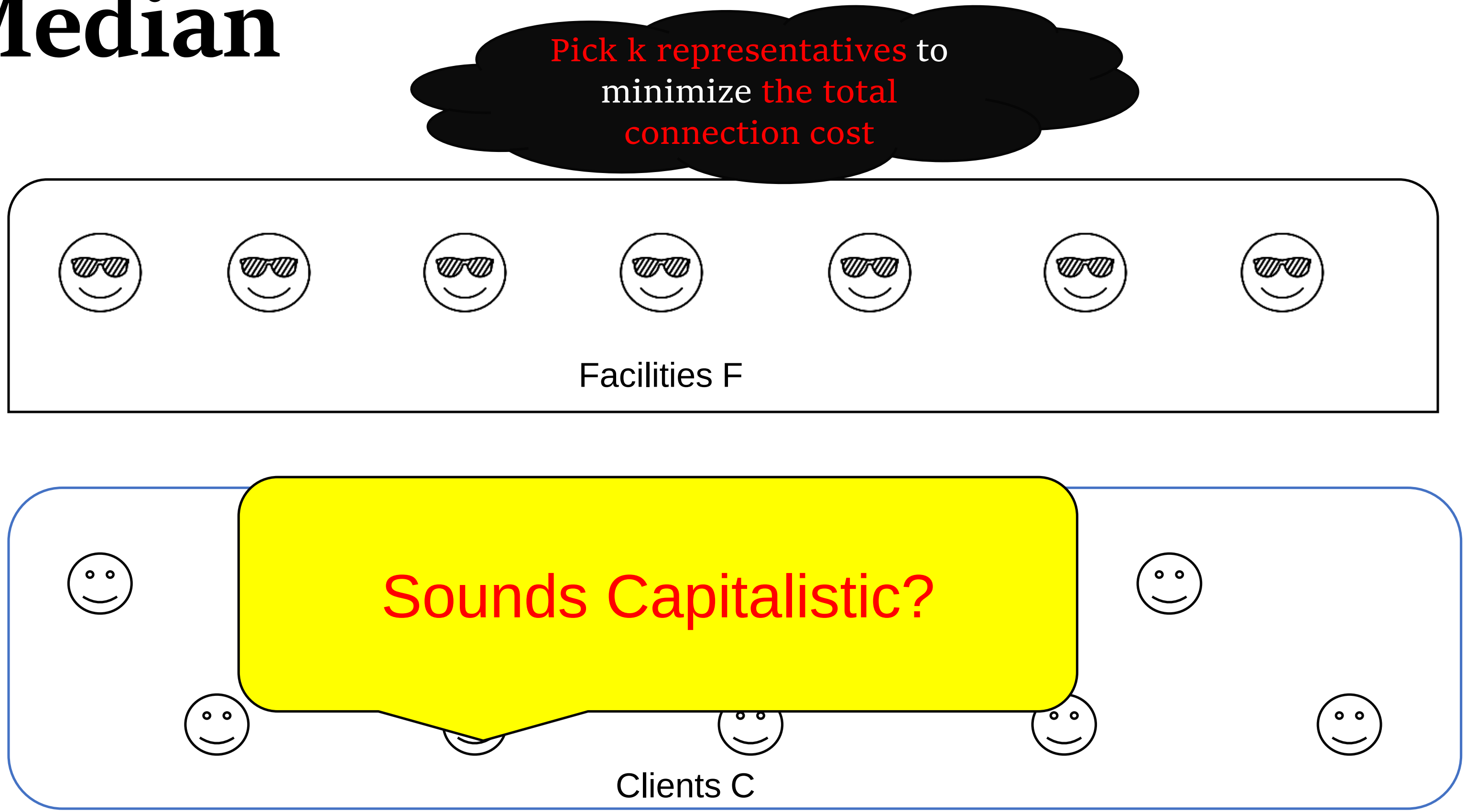


Aalto University
School of Science

k-Median

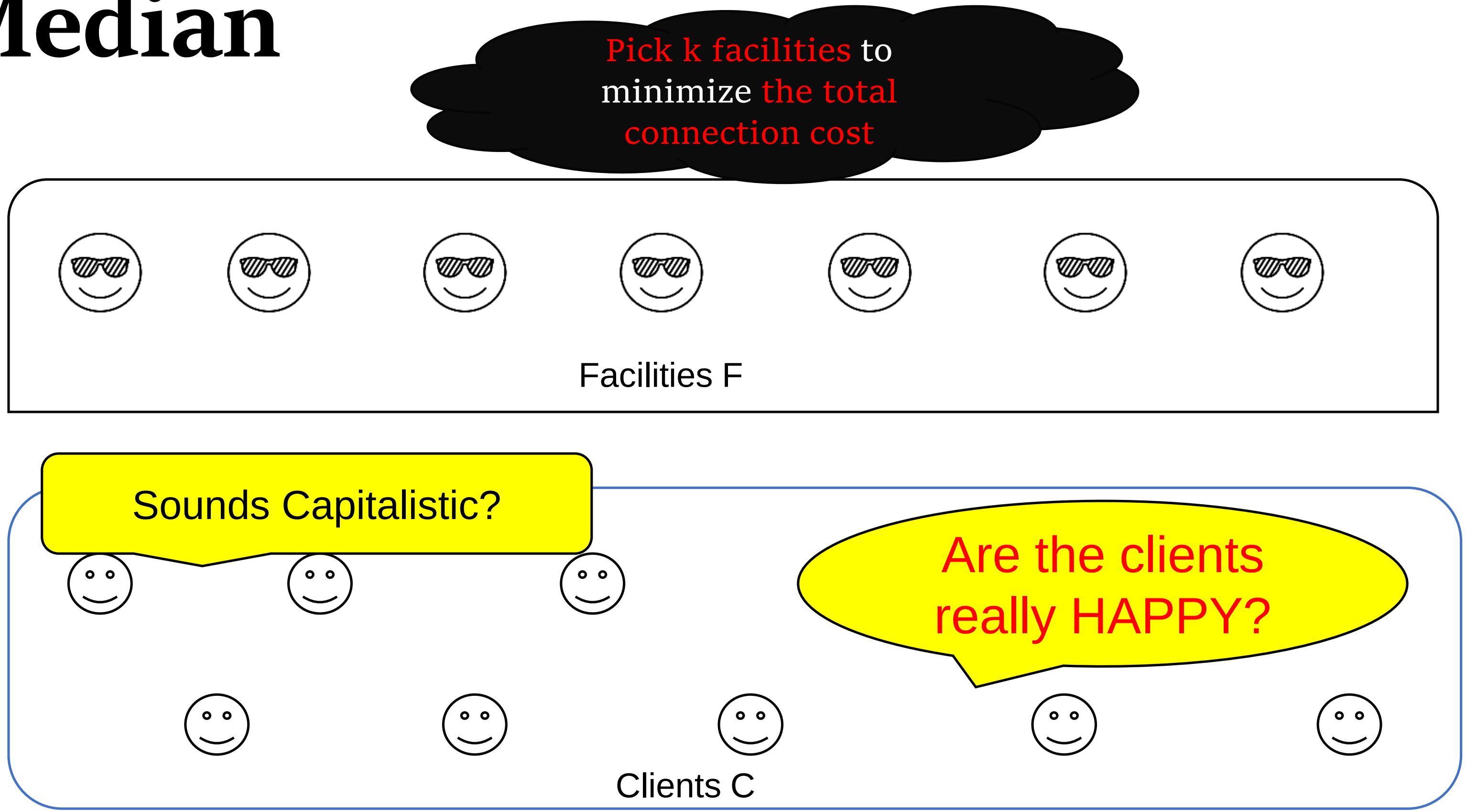

Facilities F
Pick k representatives to
minimize the total connection
cost
Clients C

k-Median



Aalto University
School of Science

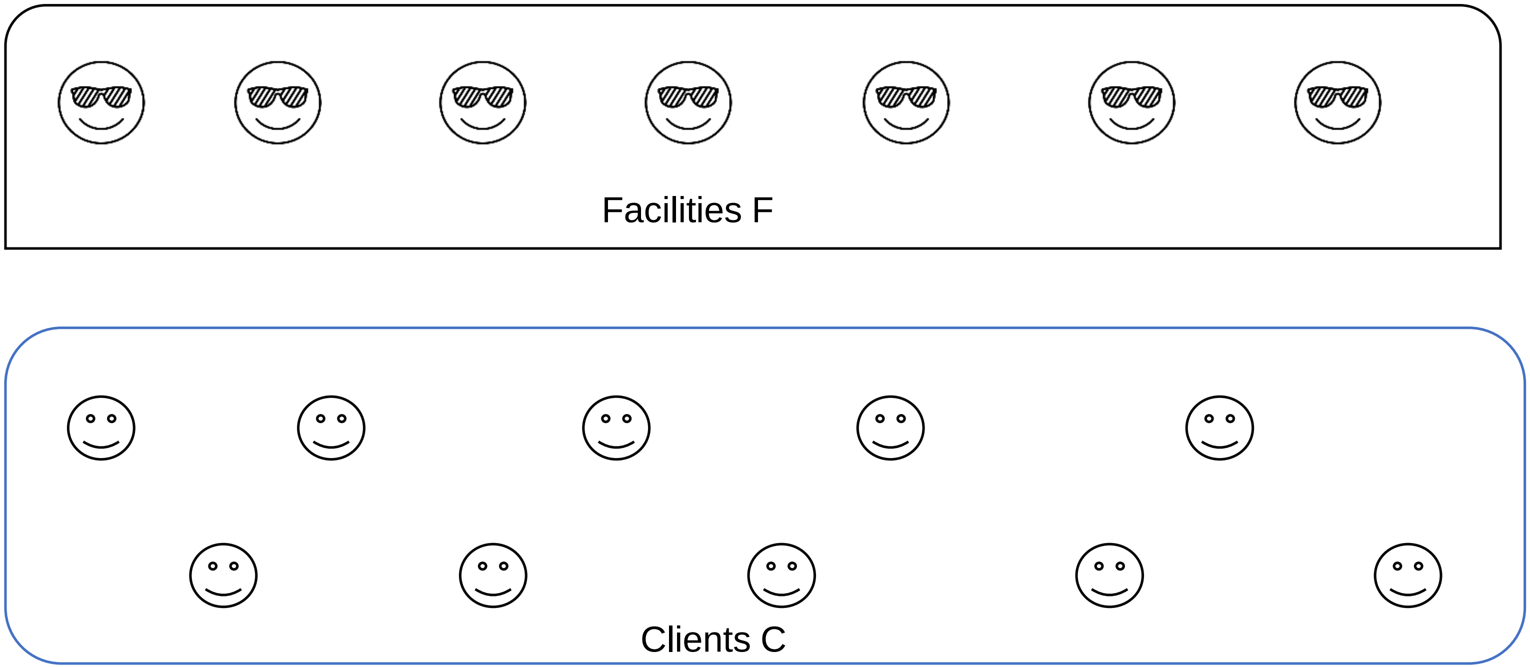
k-Median



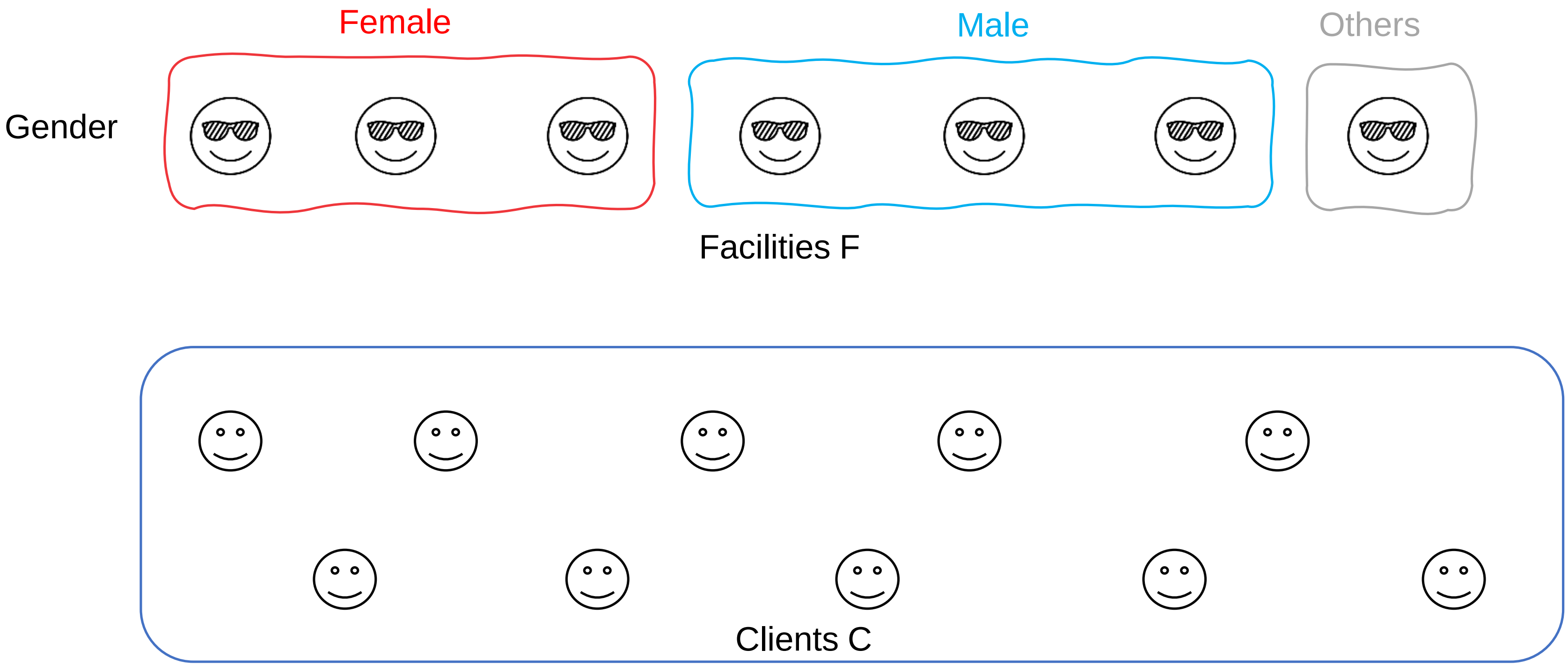
Pick k facilities to minimize the total connection cost
Facilities F
Sounds Capitalistic?
Are the clients really HAPPY?
Clients C

Fair Clustering –
Diversity aware clustering

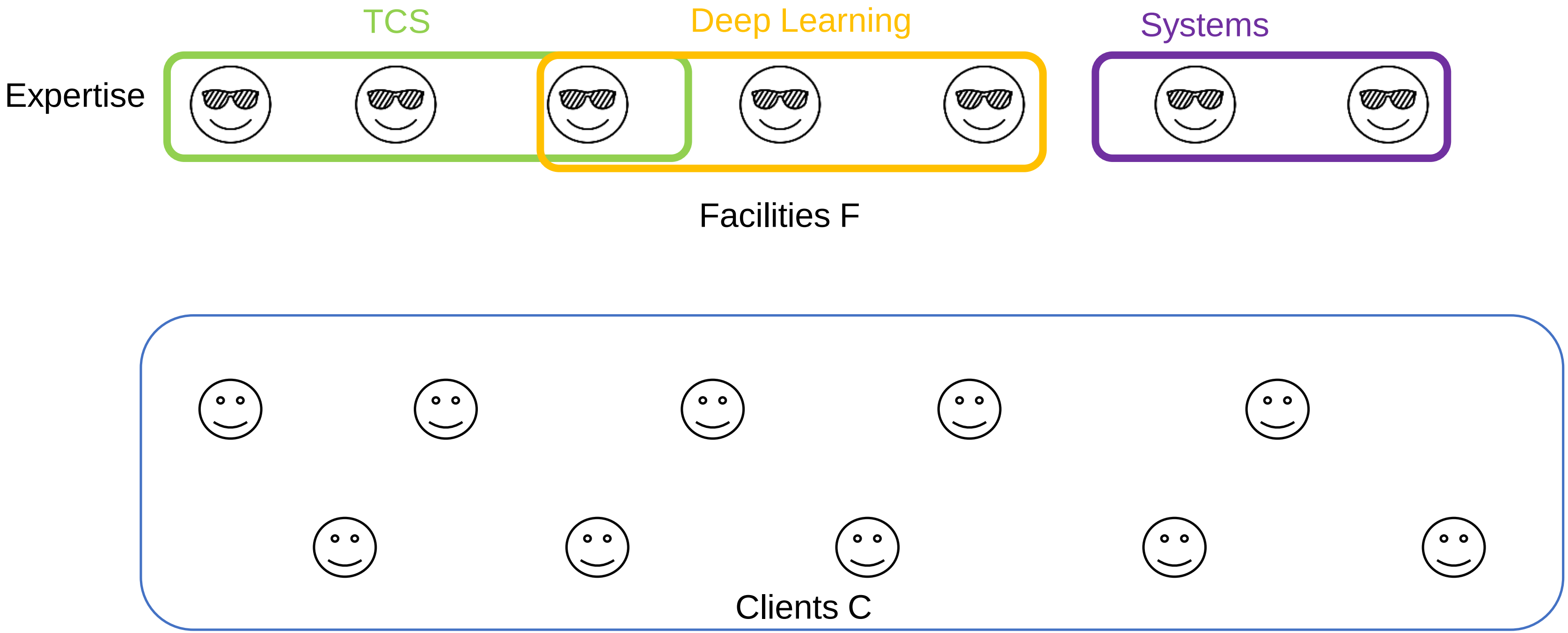
k-Median



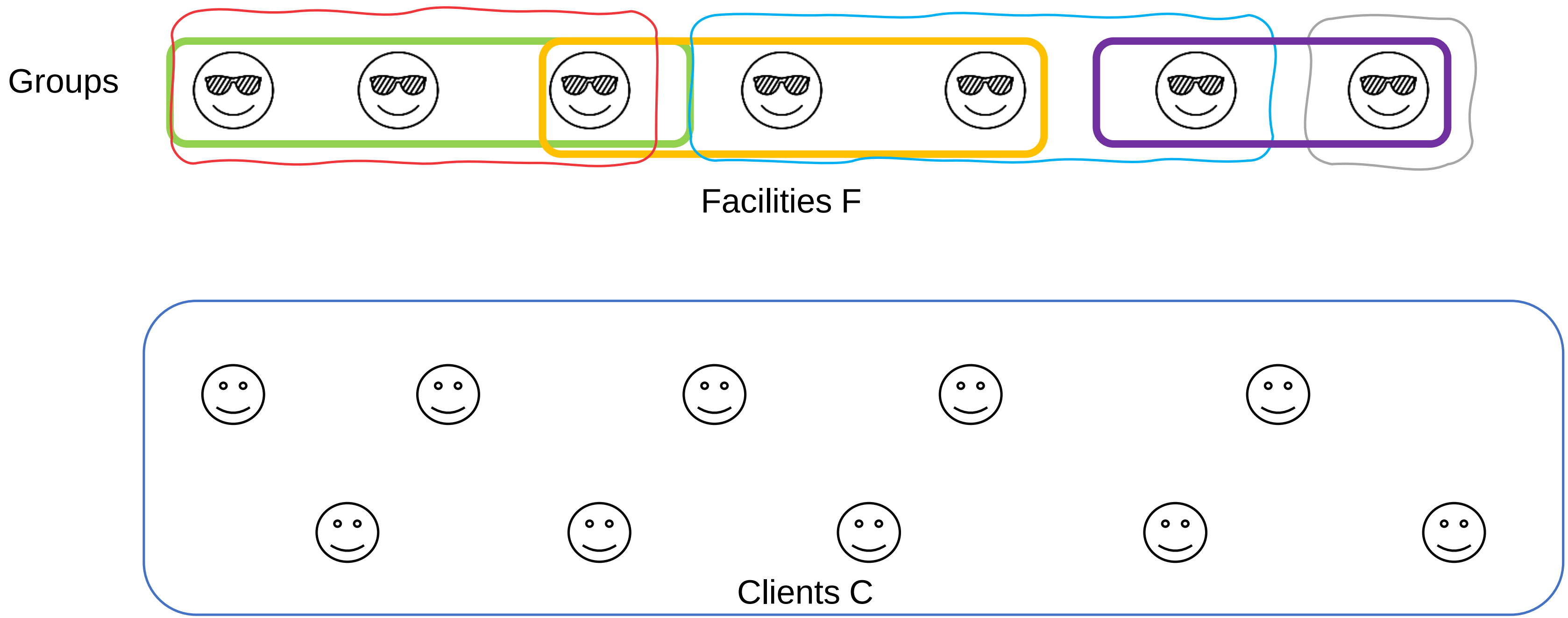
Diversity aware k-Median



Diversity aware k-Median



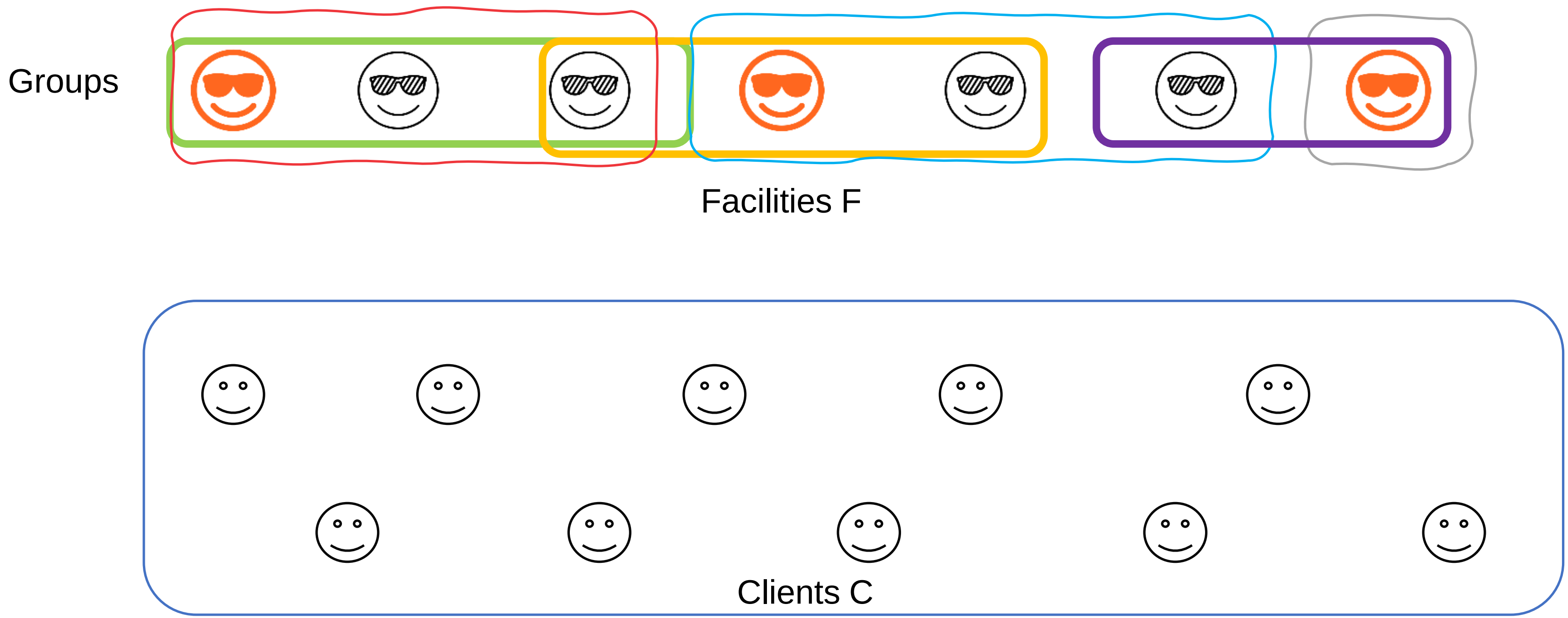
Diversity aware k-Median



Diversity aware k-Median

Aalto University
School of Science

Diversity aware k-Median



Diversity aware k-Median

- Set of Facilities F
- Set of Clients C
- Distance function d
- Groups $(G_1, \cdots, G_t)$ over F, i.e., $G_i \subseteq F$
- Diversity constraints
  $[a_i, b_i]$ for each G_i

Goal:

k-Median

zed subset X of F with minimum total connection cost that respects diversity constraints.

$$min_X \sum_{c \in C} d(c, X)$$

s.t. $a_i \leq |G_i \cap X| \leq b_i$ for $i \in [t]$

$$|X| = k$$

Literature

- **Avoid over-representation**
  - Well studied problem
  - Red-blue median problem
    [HKK ESA'10, Algorithimica'12]
  - Matroid Median problem
    [KKNSS SODA'11, CLLW IPCO'13, Swamy ACM Trans.' 16]
  - Constant factor approximation algorithms

- **Avoid under-representation**
  - Recently defined and studied
    [TOG ECML-PKDD'21]
  - Computationally very different than its counter-part

Our results – Price for Diversity

- **Trivial algorithm $\boldsymbol{O(|F|^k)}$**
  - best to hope for!
    (unless SETH fails)
- **Even any approximation in time $\boldsymbol{O(|F|^{k-\epsilon})}$ is ruled out!**
  - Captures Dominating Set

- **What if we allow additional running time?**
  - Say $f(k, t)poly(|F|)$?
- **Unfortunately, the problem is hard even when for**

  $$\boldsymbol{f(k, t)poly(|F|)}$$

Our results – Best Algorithms

- What if we want to approximate in time $f(k,t)poly(|F|)$, for some f?

We can find $(1+\frac{2}{e}+\epsilon)$-approximation for Diversity aware k-median in randomized time $f(k,t,\epsilon)poly(|F|)$.

Our results – Best Algorithms

- What if we want to approximate in time $f(k,t)poly(|F|)$, for some f?

We can find 1.74 -approximation for Diversity aware k-median in randomized time $f(k,t,\epsilon)poly(|F|)$.

$$f(k,t,\epsilon) = \left(\frac{2^t}{\epsilon}\right)^{O(k)}$$

Our results – Best Algorithms

- What if we want to approximate in time $f(k,t)poly(|F|)$, for some f?

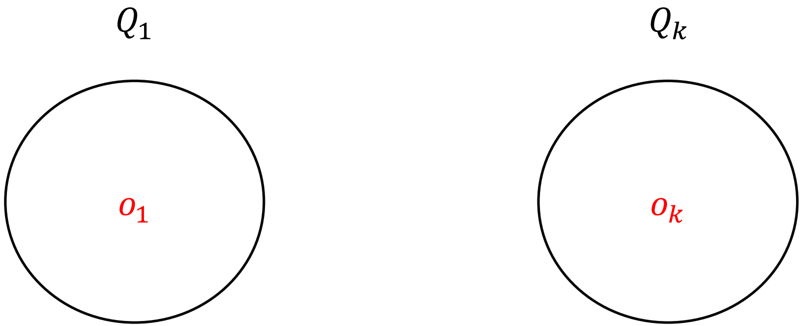
We can find $(1+\frac{2}{e}+\epsilon)$-approximation for Diversity aware k-median in randomized time $f(k,t,\epsilon)poly(|F|)$.

The approximation factor is tight*.

*Assuming Gap-ETH.

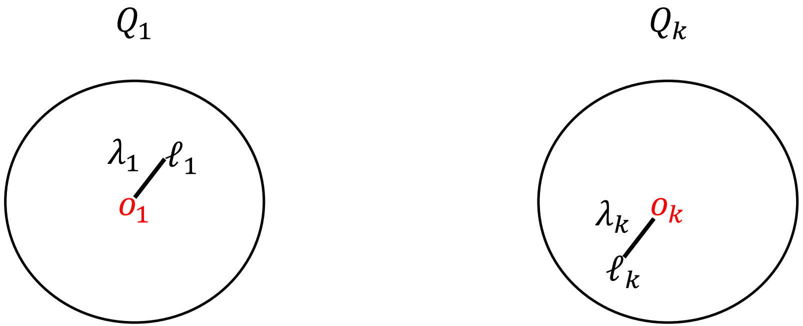
Overview of the algorithm

- Suppose the groups are disjoint…
- Consider some optimal solution $O = (o_1, \cdots, o_k)$
- Let $(Q_1, \cdots, Q_k)$ be the clusters due to O
- How do we identify these clusters?



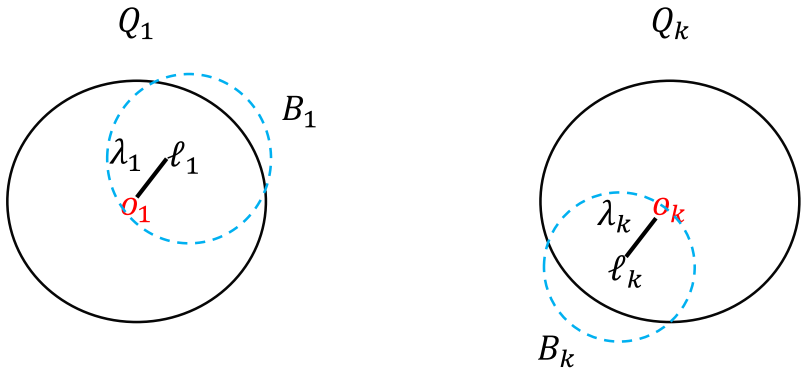
Overview of the algorithm

- Suppose $|C|$ is small.
- Then, we can identify each Q_i by a closest client ℓ_i to o_i
- Let $\lambda_i := d(o_i, \ell_i)$



Overview of the algorithm

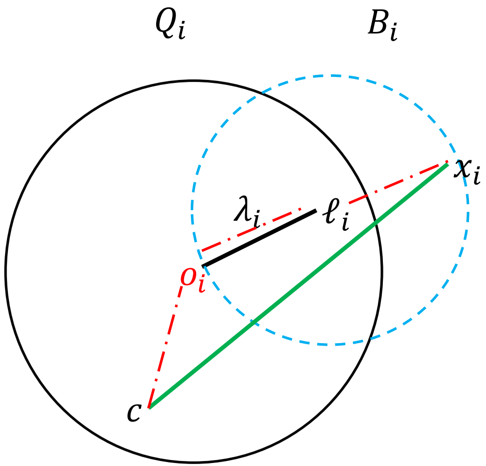
- Then, if we know (ℓ_i, λ_i), then we can consider the ball B_i at ℓ_i of radius λ_i



Overview of the algorithm

- Then, if we know (ℓ_i, λ_i), then we can consider the ball B_i at ℓ_i of radius λ_i
- We know that $o_i \in B_i$
- For $c \in Q_i$,for any facility $x_i \in B_i$
  $$d(c, x_i) \leq 3\, d(c, o_i)$$
- Hence, for $X = (x_1, \cdots, x_k)$

$$\sum_c d(c, X) \leq 3 \sum_c d(c, O)$$



Overview of the algorithm

- **How do we handle diversity constraints?**
  - Smart way of picking facilities from B_is
- **How do we find (ℓ_i, λ_i)?**

  $(k/\epsilon)^{O(k)} poly(|F|)$ time

  - Use client coresets to reduce the size to roughly $O(k \log |C|)$
  - Discretize the distances
- **How do we improve the approximation factor?**
  - Using more clever approach – submodular optimization

Overview of the algorithm

- **How do we handle diversity constraints?**
  - Smart way of picking facilities from B_is

Infact, with more ideas, we can solve the general version when the groups are intersecting, resulting in time $\left(\frac{2^t}{\epsilon}\right)^{O(k)} poly(|F|)$

- **How do we improve the approximation factor?**
  - Using more clever approach – submodular optimization

Aalto University
School of Science

Other results

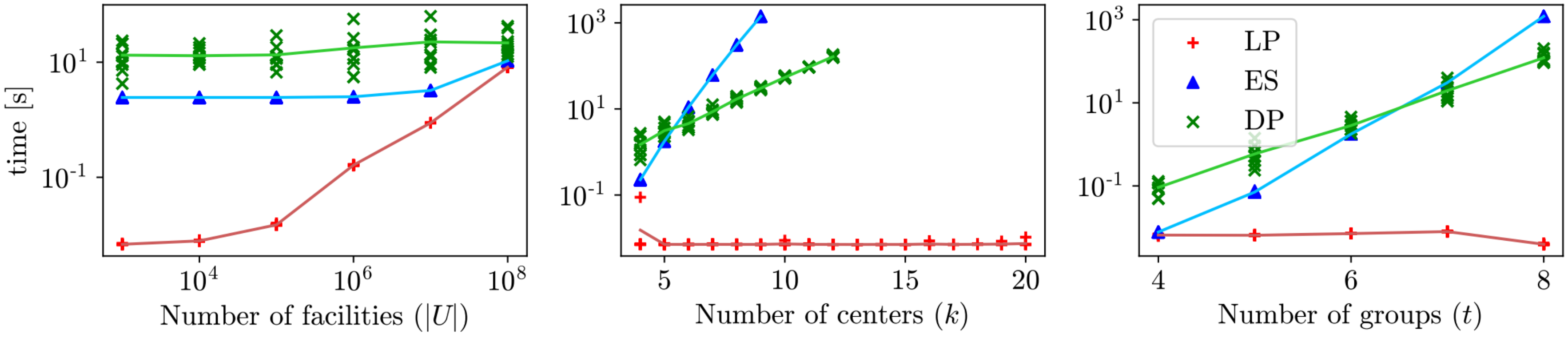
- **Algorithm extends to objectives other than k-Median**

- **Fast algorithm for bicriteria solution**
    - based on a dynamic program for the feasibility problem

- **Local search based heuristics**

- **LP based heuristics**

Experiments — scalability



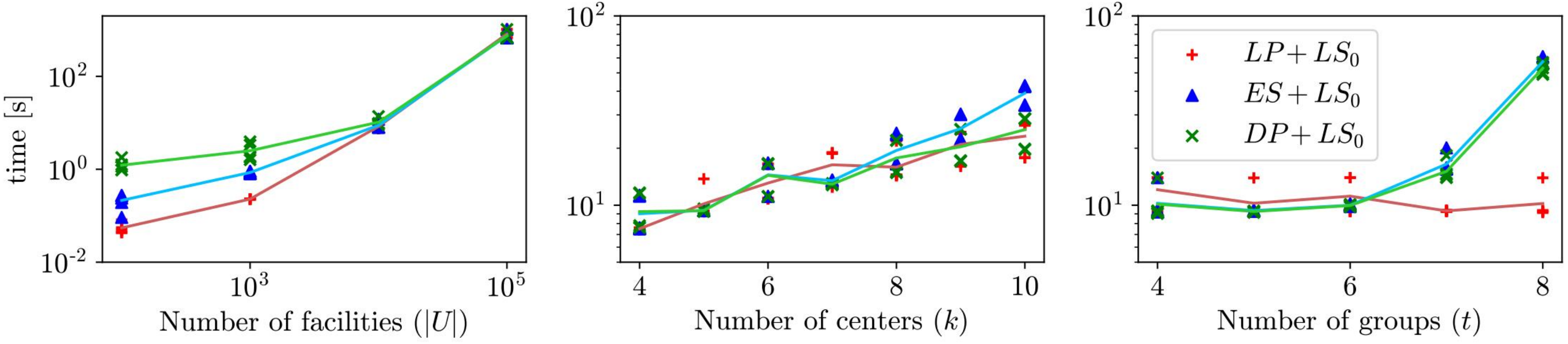
Scalability of algorithms for finding a feasible constraint pattern.

- Synthetic data
- Desktop configuration

- LP : Linear program
- ES : Exhaustive search
- DP : Dynamic program

Aalto University
School of Science

Experiments — scalability



Scalability of bicriteria algorithms

- Synthetic data
- Desktop configuration

- LS_0 : Local search on k-Median
- LP : Linear program
- ES : Exhaustive search
- DP : Dynamic program

Experiments — real data set

Table 2: Experiments on datasets with $k = 6$, $t = 4$ and $\vec{r} = \{3, 3, 2, 1\}$.

				Bicriteria approximation $(2k, \alpha)$									Heuristics (k)				FPT (k, t, ϵ)	
			LS_0	LS_0 + LP			LS_0 + ES			LS_0 + DP			LP + LS_1		ES + LS_1		$(3 + \epsilon)$-apx	
Dataset	$\|U\|$	D	time	time	ζ^*	k^*	time	ζ^*	k^*	time	ζ^*	k^*	time	ζ^*	time	ζ^*	time	ζ^*
switzerland	123	14	0.05	0.14	0.92	10	0.05	0.92	10	0.09	0.92	10	0.35	1.08	0.16	1.08	16 841.32	2.82
hepatitis	155	20	0.07	0.07	0.94	11	0.07	0.95	10	0.11	0.95	10	0.39	1.07	0.27	1.07	18 922.51	1.81
va	200	14	0.06	0.06	0.95	11	0.06	0.95	11	0.10	0.98	9	0.20	1.27	0.01	1.27	14 855.96	1.76
hungarian	294	14	0.14	0.14	0.95	10	0.14	0.96	9	0.17	0.98	8	0.74	1.02	4.00	1.01	-	-
heart-failure	299	13	0.18	0.19	0.93	11	0.19	0.95	9	0.22	0.95	9	0.71	1.05	3.72	1.05	-	-
cleveland	303	14	0.09	0.10	0.93	10	0.10	0.99	9	0.13	0.99	8	0.47	1.07	1.33	1.05	-	-
student-mat	395	33	0.24	0.25	0.96	12	0.25	0.97	12	0.28	0.99	8	0.36	1.05	0.32	1.05	-	-
house-votes-84	435	17	0.16	0.16	0.97	10	0.16	0.98	9	0.19	0.98	9	0.71	1.17	3.20	1.11	-	-
student-por	649	33	0.50	0.51	0.98	10	0.50	0.98	10	0.53	0.99	9	0.49	1.02	0.52	1.02	-	-
drug-consumption	1884	32	2.58	2.69	0.98	12	2.68	0.98	12	2.72	0.99	8	0.49	1.08	0.41	1.07	-	-
bank	4521	17	8.56	8.72	0.97	10	8.71	0.99	10	8.76	0.98	9	1.41	1.10	2.07	1.10	-	-
nursery	12960	9	40.21	40.48	0.99	10	40.66	0.99	10	40.43	0.99	9	22.38	1.14	43.20	1.14	-	-
vehicle-coupon	12684	26	51.87	51.34	0.98	12	50.88	0.98	12	50.98	0.99	8	8.59	1.12	16.43	1.12	-	-
credit-card	30000	25	928.77	945.56	0.99	12	939.98	0.99	12	941.07	1.00	8	9.18	1.18	18.89	1.18	-	-
dutch-census	32561	15	376.73	384.15	0.97	12	390.82	0.98	12	385.36	0.99	8	76.34	1.40	151.18	1.32	-	-
bank-full	45211	17	934.14	958.79	0.97	11	958.86	0.98	11	948.85	0.97	10	103.57	1.10	202.73	1.10	-	-
diabetes	101 766	50	15 896.14	-	-	-	-	-	-	-	-	-	829.96	1.07	1 503.05	1.01	-	-

Thank you

- **Appeared at KDD'22**
- **Selected for ACM Showcase on Kudos:**

  https://www.growkudos.com/publications/10.1145%252F3534678.3539487/reader
- **Source Code:**

  github.com/suhastheju/diversity-aware-clustering
- **Image credits:** Midjourney

Aalto University
School of Science