Fair Clustering for Data Summarization

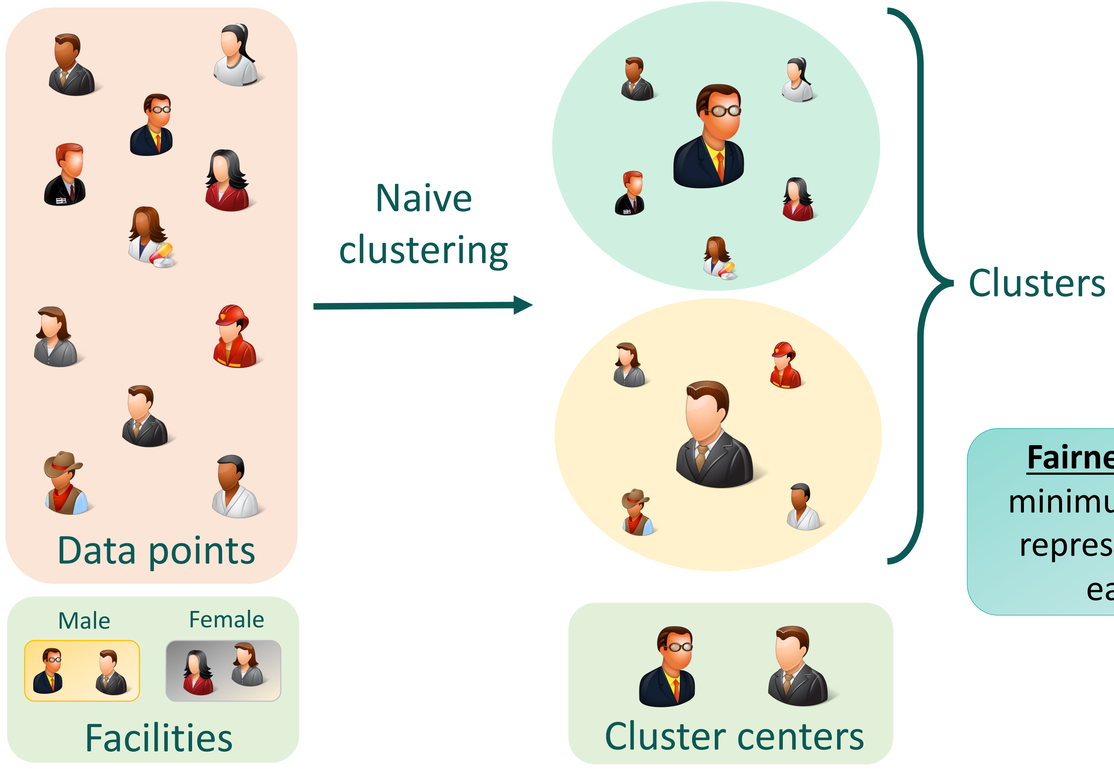
Ameet Gadekar Aristides Gionis **Suhas Thejaswi**

KTH VETENSKAP OCH KONST

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

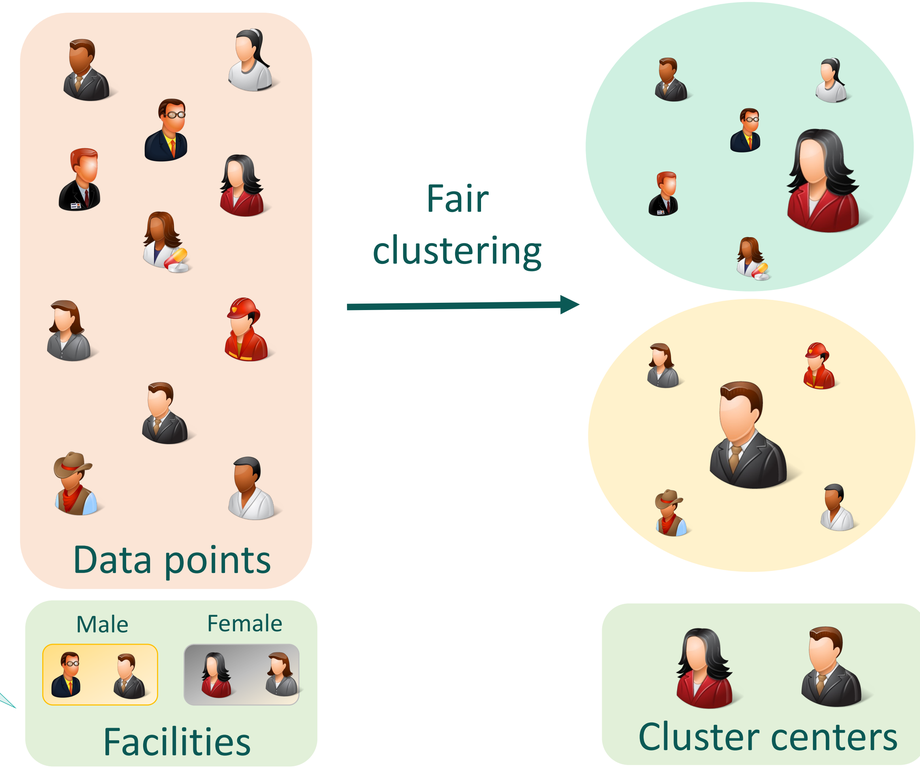
CISPA

The k-supplier problem



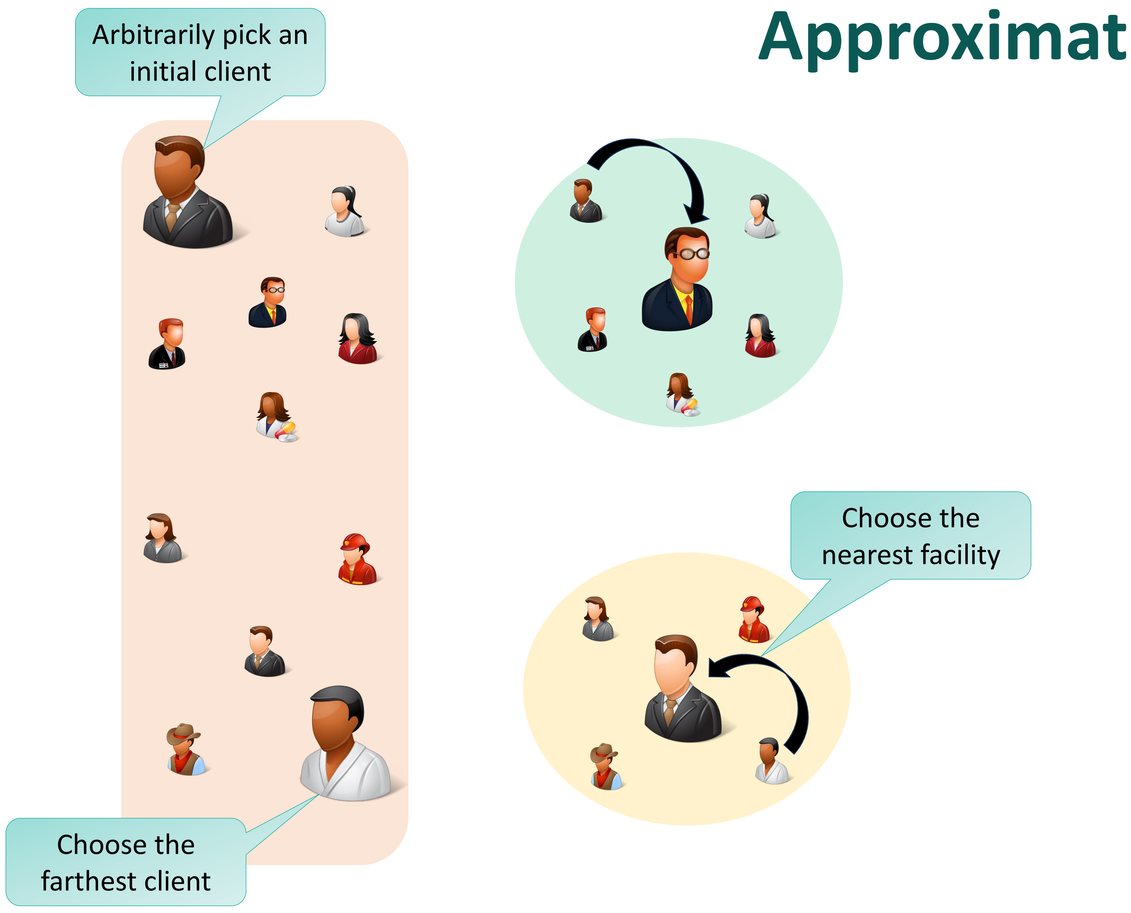
Cluster the data points by choosing $k = 2$ representative centers / facilities to minimize the maximum distance from any point to its closest representative

The fair k-supplier problem



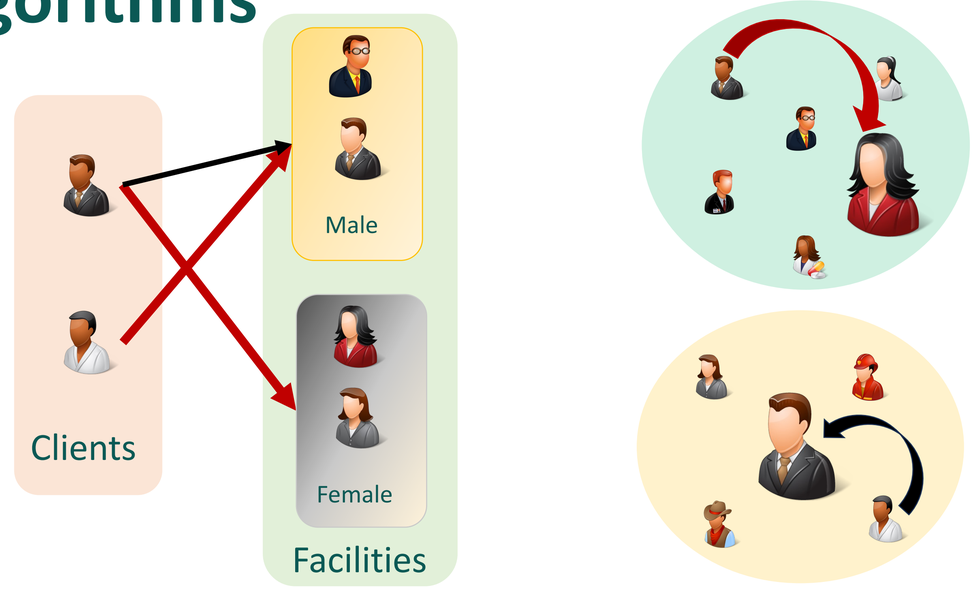
The goal remains to minimize the maximum distance, but with fairness constraints: each group must contribute a minimum (and a maximum) number of cluster centers

Approximation algorithms



A greedy 3-approximation for the (unfair) k-supplier problem

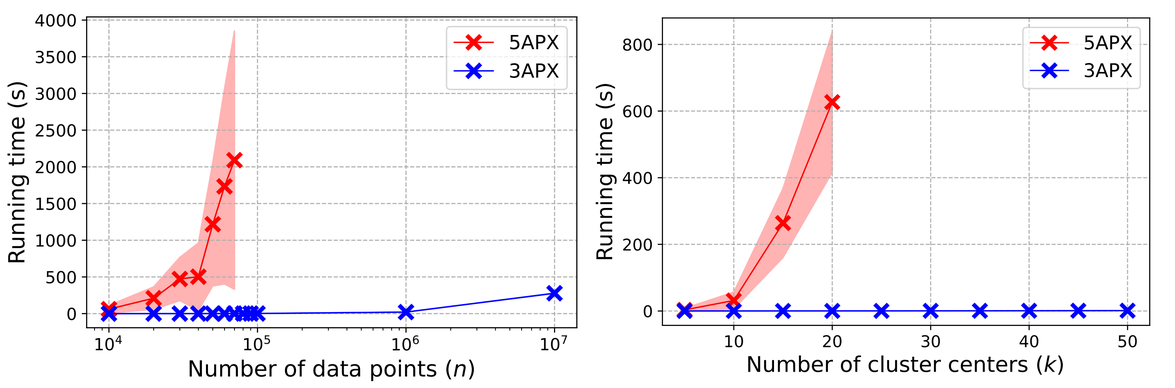
- Arbitrarily choose an initial client, then iteratively select the farthest client from the current set for k rounds
- Choose the nearest facility to each selected client



A 3-approximation for the fair k-supplier problem using matching framework

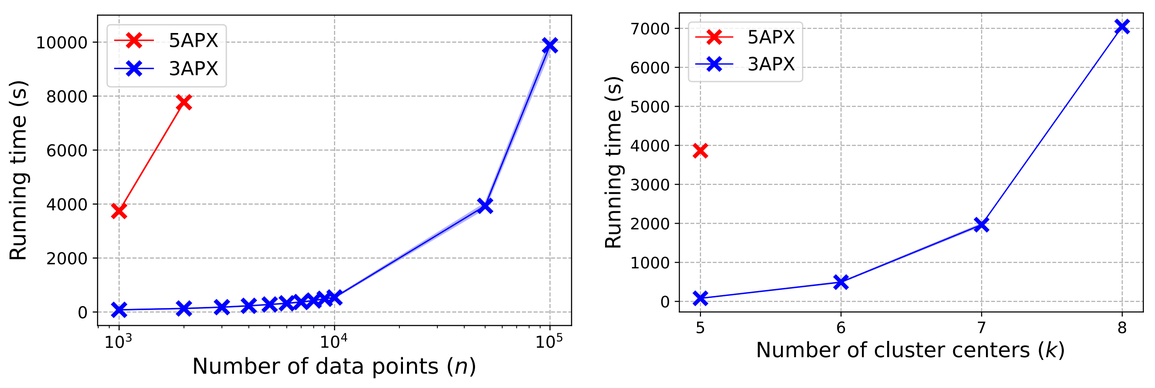
- Arbitrarily choose an initial client, then iteratively select the farthest client from the current set for k rounds
- Guess the optimal radius using binary search (over at most nk values)
- Build a bipartite graph: connect a client to each group within the guessed radius
- Find a maximum matching and select one facility per matched group within the radius
- This yields a 3-approximation in $O((kn + k^2\sqrt{k})\log n \log k)$ time
- For intersecting groups, reduce the problem to multiple instances with disjoint groups, incurring an overhead of $O(2^{tk})$
- n is number of data points and t is number of groups

Experimental results



Salability of our 3-approximation algorithm (3-APX) vs. Chen et al.'s (TCS, 2024) 5-approximation (5-APX) for fair k-supplier with disjoint groups

- ***Left plot:*** Runtime scaling with increasing number of data points n, for $k = 10$ and 5 disjoint groups
- ***Right plot:*** Runtime scaling with increasing number of cluster centres k, for a fixed dataset of $n = 10\,000$ points
- ***Fairness constraint:*** select an equal number of cluster centres from each group



Salability of our 3-approximation algorithm (3-APX) vs. Thejaswi et al.'s (ArXiv, 2024) 5-approximation (5-APX) for fair k-supplier with intersecting groups

- ***Left plot:*** Runtime scaling with increasing number of data points n, for $k = 5$ and 5 intersecting groups. Fairness constraint: choose at least 2 cluster centres from each group
- ***Right plot:*** Runtime scaling with increasing number of cluster centres k, for a fixed dataset of $n = 1000$ points. Fairness constraint: choose $\frac{2k}{t}$ cluster centres from each group