Diverse Data Selection under Fairness Constraints

Zafeiria (Iro) Moumoulidou — zmoumoulidou@cs.umass.edu
Andrew McGregor — mcgregor@cs.umass.edu
Alexandra Meliou — ameli@cs.umass.edu

DREAM LAB
DATA SYSTEMS RESEARCH FOR EXPLORATION, ANALYTICS, AND MODELING

The Fair-Swap algorithm: # colors = 2

Universe of points placed on a line: **4 — approx.**

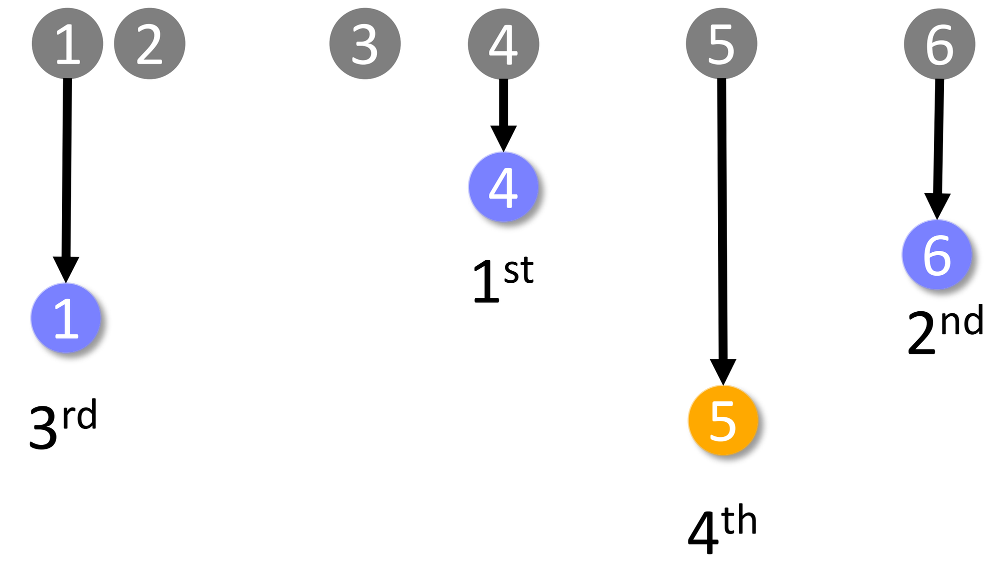
1 2 3 4 5 6

Select a set of 4 points with:

● = 2 # ● = 2

A. Color-blind Phase

1. Use farthest-first traversal heuristic to pick the points.



There is a **+1** blue point and a **-1** orange point is missing. Let's **swap**!

B. Swap Phase

1. Let's **add** an **orange** point!

1 2 4 5 6

We add point 2 because it's the **farthest away** from the other orange point 5!

2. Let's **remove** a **blue** point!

~~1~~ 2 4 5 6

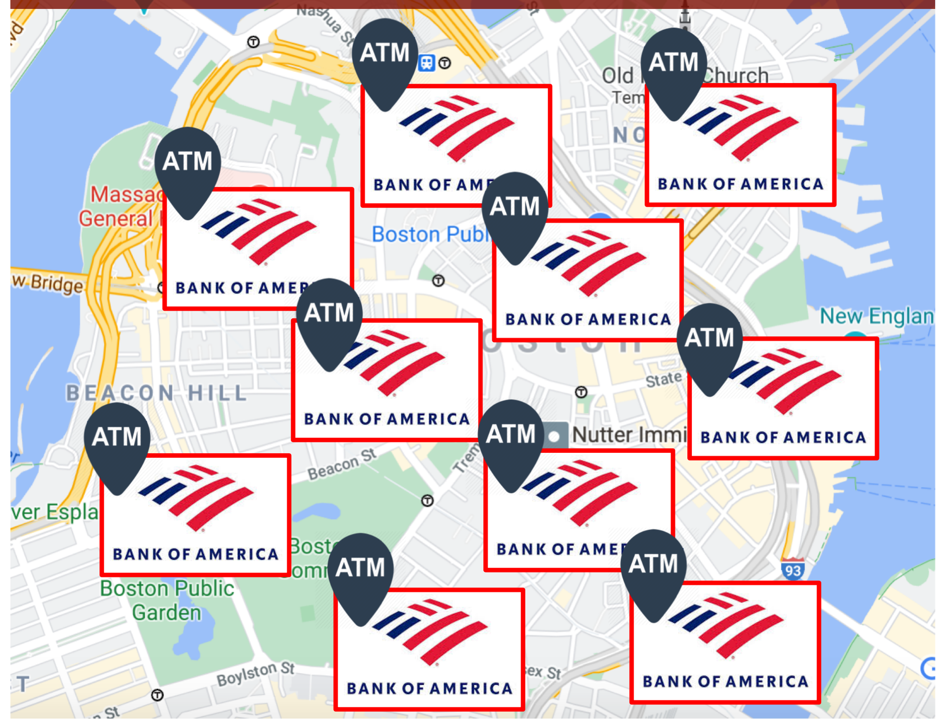
We remove point 1 because it's the **closest** to point 2 we just added!

We are done!

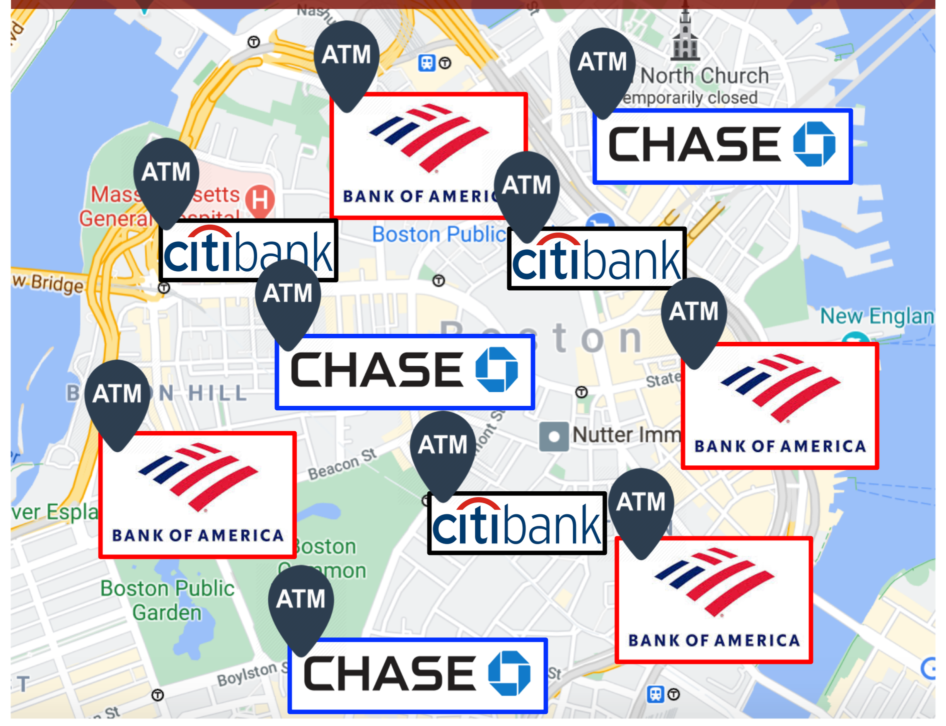
2 4 5 6

Fair and Diverse Data Selection

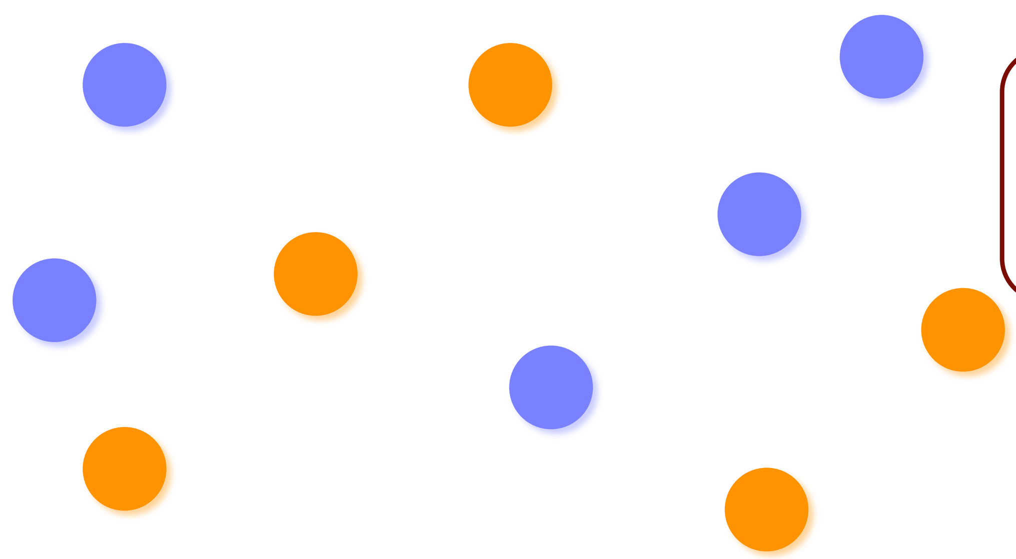
Max-Min Diversification



Fair Max-Min Diversification



The Fair Max-Min diversification problem



Maximize *diversity* subject to *fairness* constraints

Metric space $(\mathcal{U}, d)$

m = # colors

$$\textsc{Fair Max-Min}: \ \underset{\mathcal{S}\subseteq\mathcal{U}}{\text{maximize}} \ \min_{\substack{u,v\in\mathcal{S}\\ u\neq v}} d(u,v)$$

NP-hard

$$\text{subject to } |\mathcal{S}\cap\mathcal{U}_i| = k_i, \ \forall i\in[m]$$

Can we do better?

Moumoulidou et al. [ICDT 2021]	4— approx. (m = 2) (3m-1)— approx. (m ≥ 3)
Addanki et al. [ICDT 2022]	**General Metric Spaces** 2— approx. with expected fairness 6— approx. with (1-ε) fairness **Euclidean Metric Spaces** Exact solution. (1-dimensional spaces) (1+ε)— approx. with (1-ε) fairness *+ Streaming & Distributed Implementations*

m = # colors ε ∈ [0,1]: approx. error parameter

The Fair-Flow algorithm: # colors ≥ 3

Universe of points placed on a line: **(3m-1) — approx.**

1 2 3 4 5 6 7 8 9

Select a set of 4 points with:

● = 1 # ● = 1 # ● = 2

Oracle → d_{FAR}, d_{CLOSE}

A. Step 1

For every color:
Find a maximal set of points that are ≥ d_{FAR} far apart.

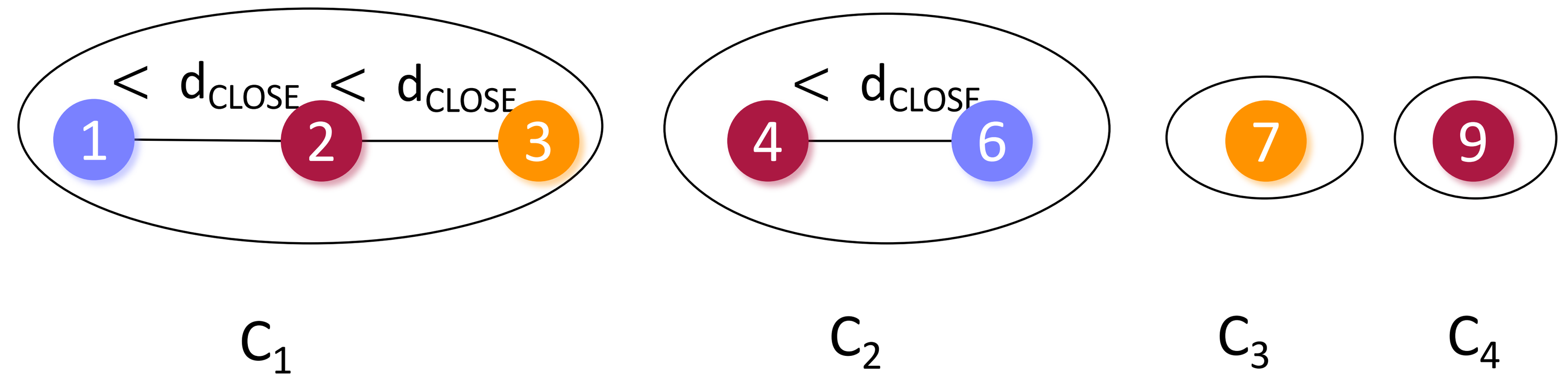
1 ≥ d_{FAR} 6

2 ≥ d_{FAR} 4 ≥ d_{FAR} 9

3 ≥ d_{FAR} 7

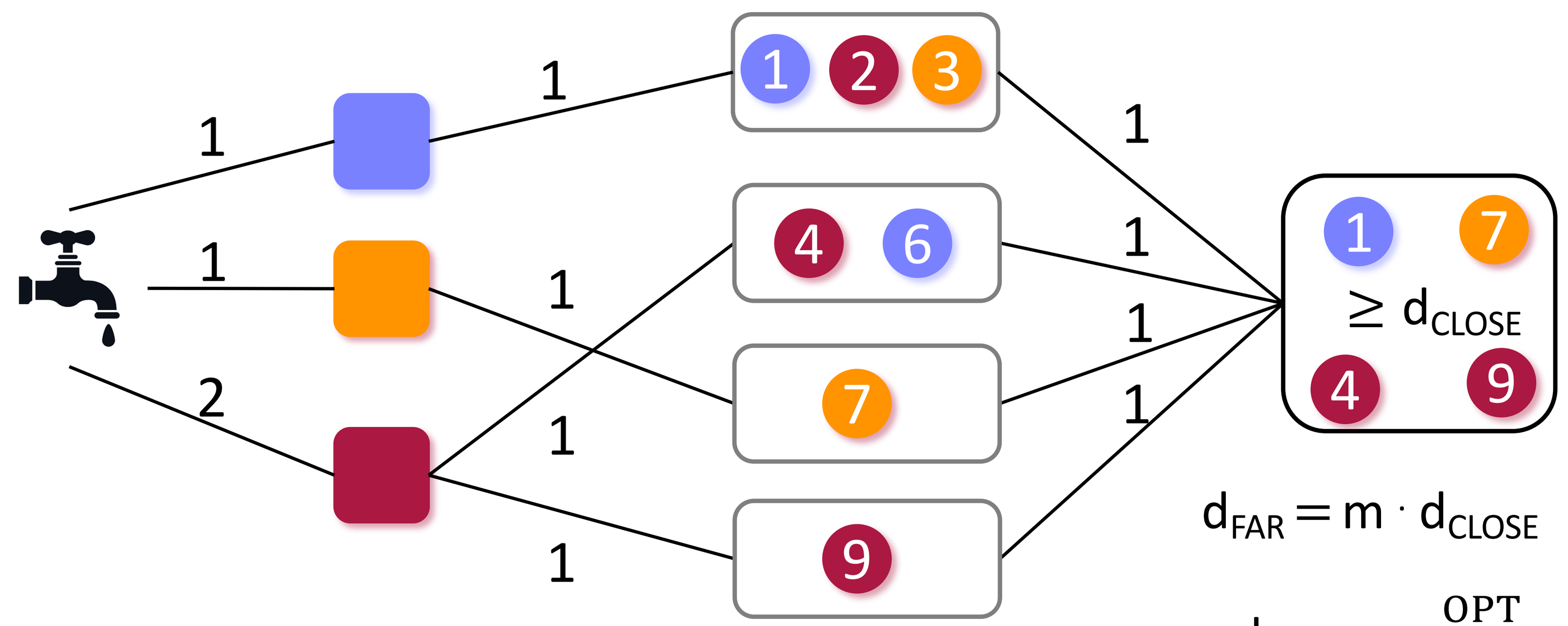
B. Step 2

Connect with an edge any two points < d_{CLOSE} far apart.



C. Step 3

Solve a Max-Flow problem to find a fair and diverse set.



$d_{FAR} = m \cdot d_{CLOSE}$

$d_{CLOSE} = \frac{OPT}{3m-1}$