

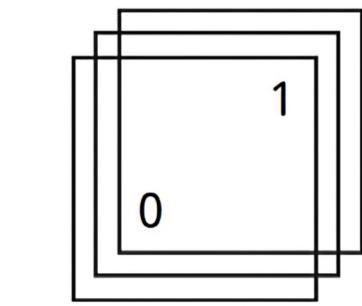


Diverse Data Selection under Fairness Constraints

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:

1 2 3 4 5 6

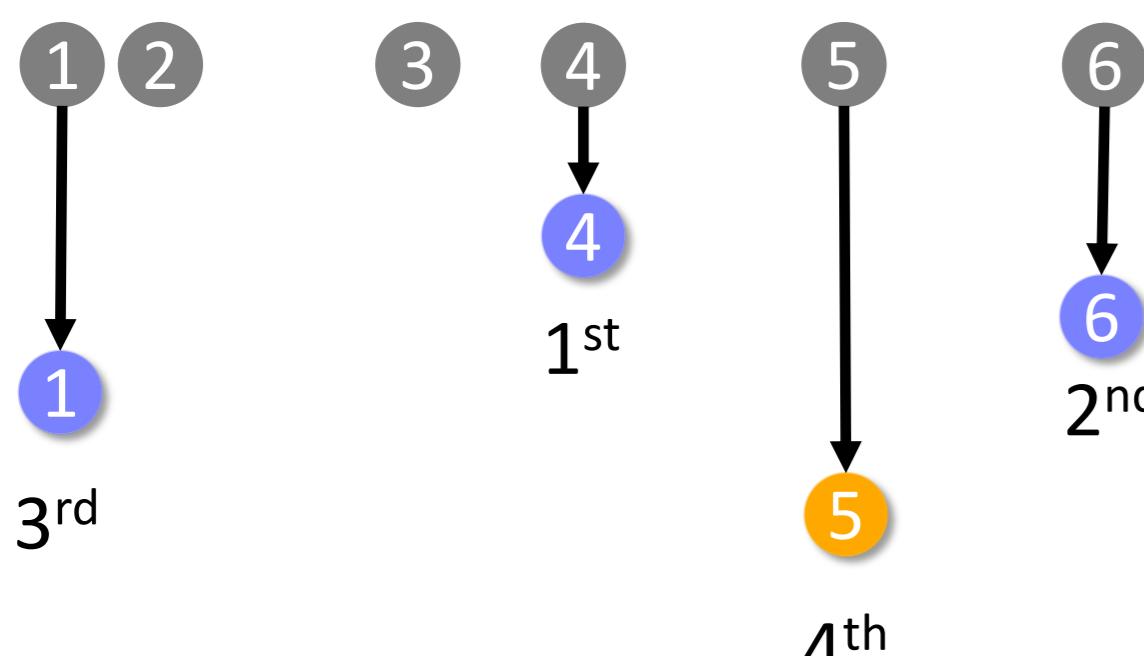
4 – approx.

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!



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

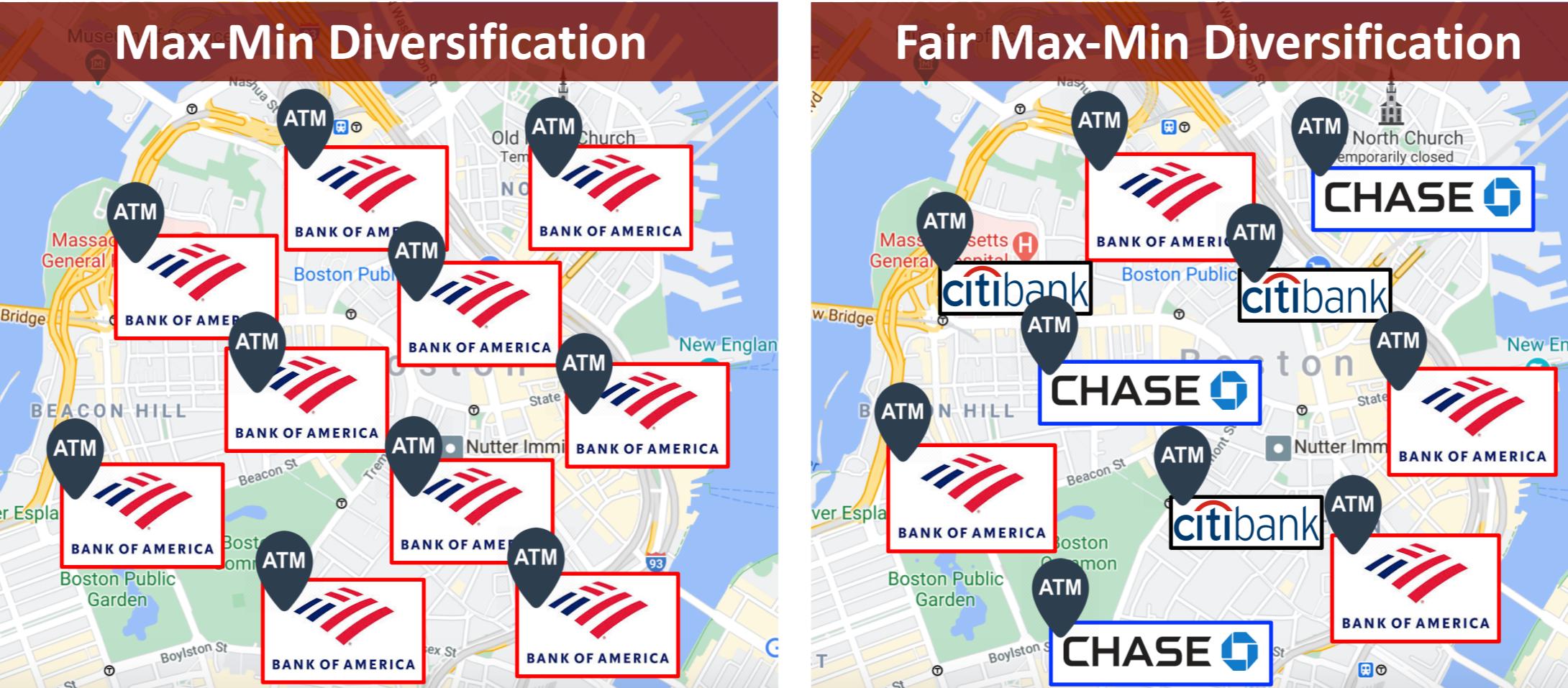
2. Let's remove a blue point!



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

We are done!

Fair and Diverse Data Selection



The Fair Max-Min diversification problem



Can we do better?

Moumoulidou et al.
[ICDT 2021]

4 – approx. ($m = 2$)
($3m-1$) – approx. ($m \geq 3$)

Addanki et al.
[ICDT 2022]

General Metric Spaces
2 – approx. with expected fairness
6 – approx. with $(1-\epsilon)$ fairness
Euclidean Metric Spaces
Exact solution. (1-dimensional spaces)
 $(1+\epsilon)$ – approx. with $(1-\epsilon)$ fairness
+ Streaming & Distributed Implementations

$m = \# \text{ colors}$ $\epsilon \in [0, 1]$: approx. error parameter

The Fair-Flow algorithm: # colors ≥ 3

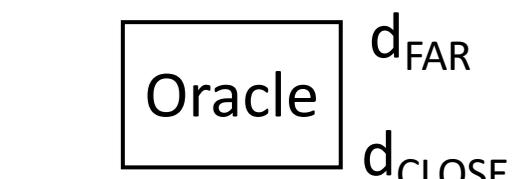
Universe of points placed on a line:

($3m-1$) – approx.

1 2 3 4 5 6 7 8 9

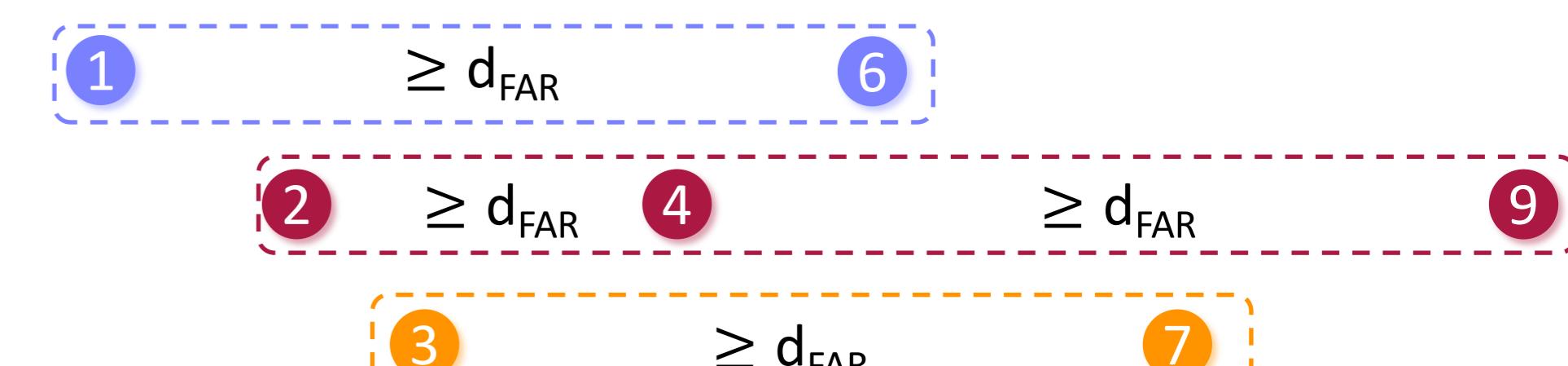
Select a set of 4 points with:

● = 1 # ○ = 1 # ■ = 2



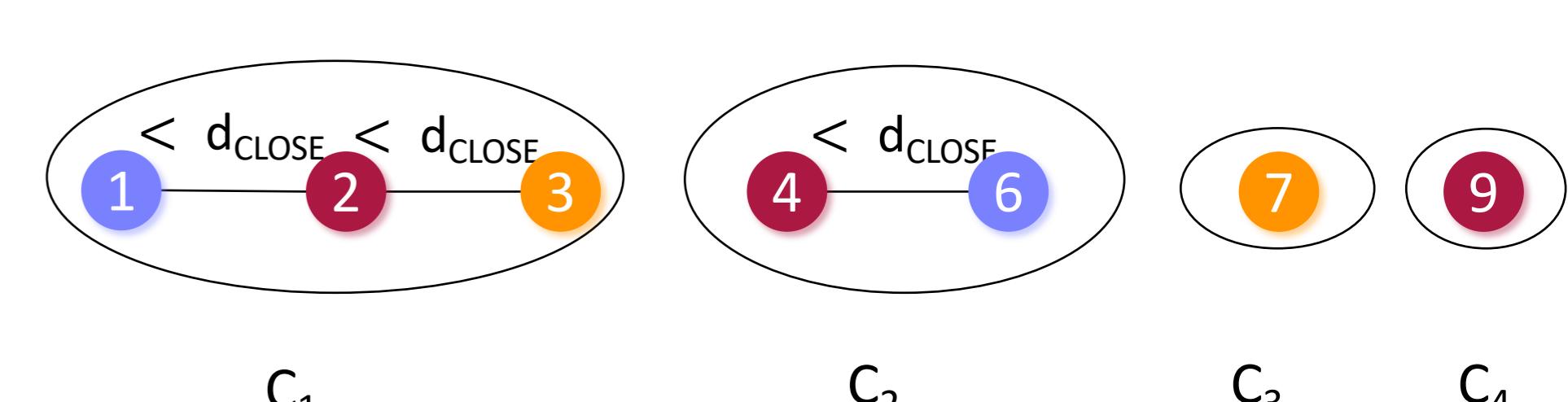
A. Step 1

For every color:
Find a maximal set of points that are $\geq d_{\text{FAR}}$ far apart.



B. Step 2

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



C. Step 3

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

