Scalable Fair Clustering

Arturs Backurs

Piotr Indyk

Krzysztof Onak

Baruch Schieber

Ali Vakilian

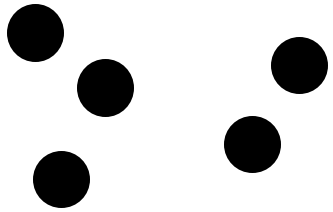
Tal Wagner

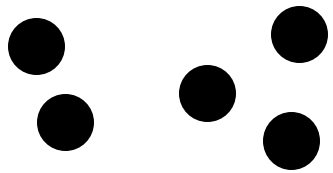
Fair Clustering

- Algorithmic Fairness
- Common Unsupervised Learning Task
- Further Implication: e.g., feature engineering

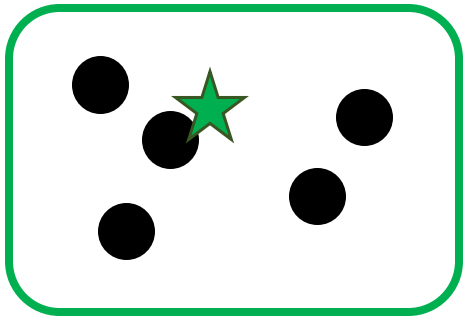
©Petrina Chan/The Tufts Daily

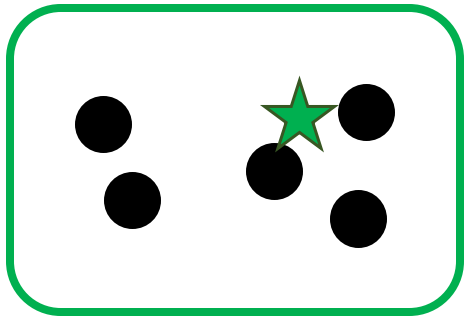
Problem Definition



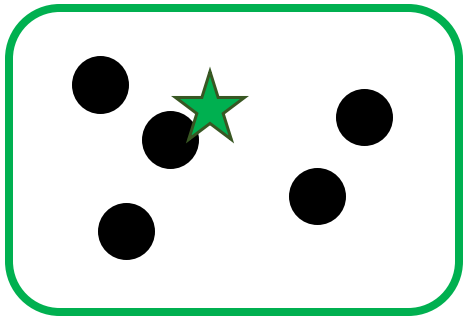


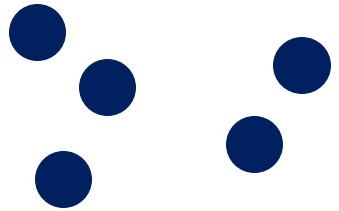
Problem Definition

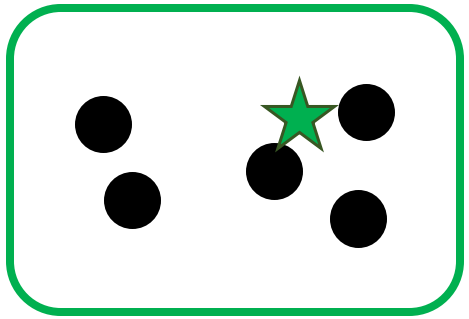


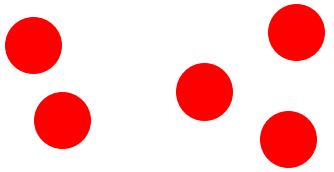


Problem Definition

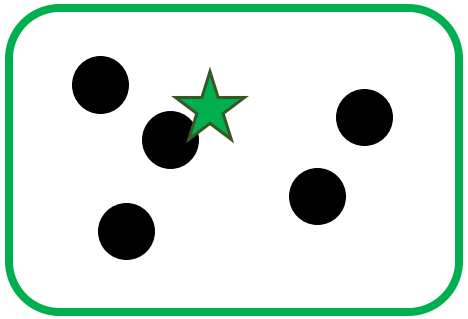


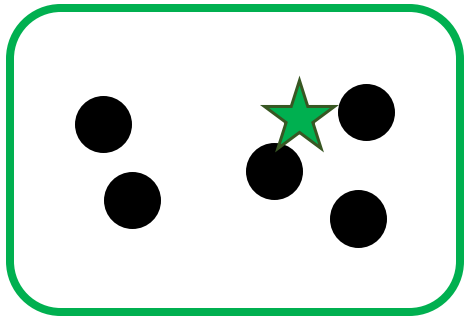


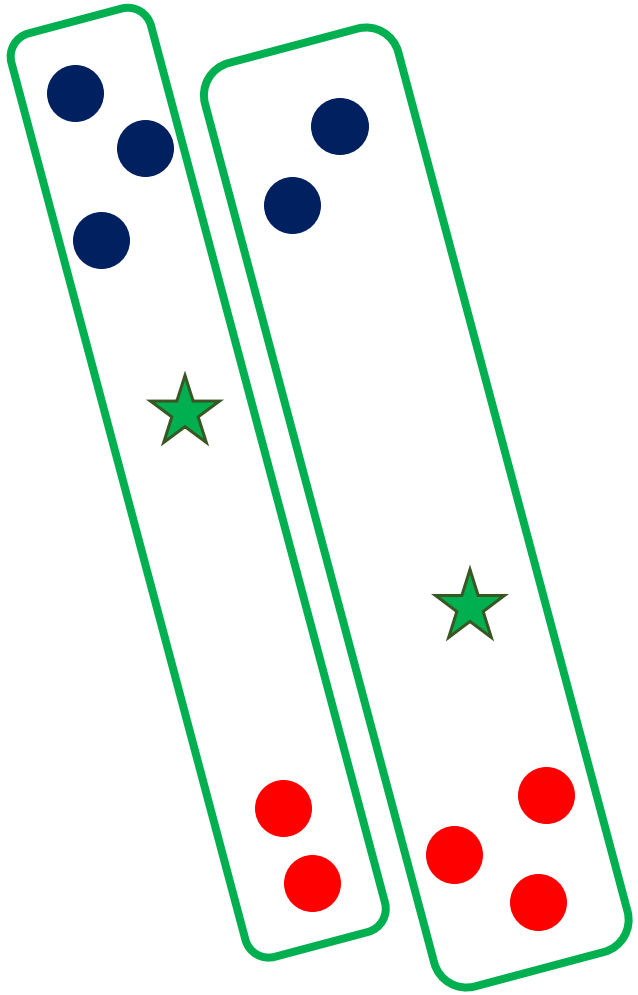




Problem Definition



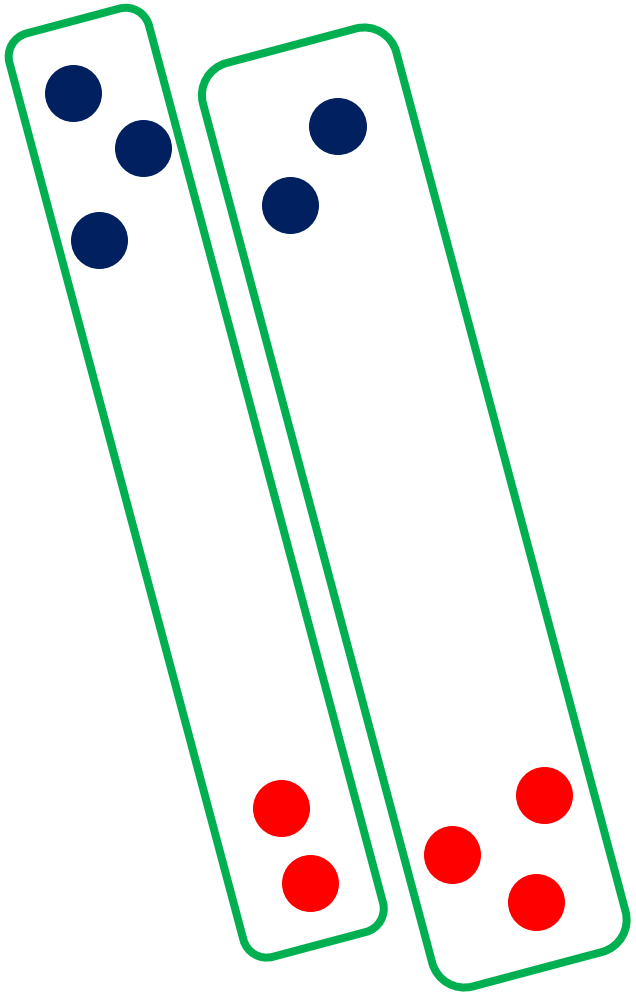




Problem Definition

- Collection of n points P in $\mathbb{R}^d$
- Each point is colored either **red** or **blue**
- Each cluster S has to be *(r,b)-balanced*

$$\frac{b}{r} \leq \frac{\#\ \textbf{red}\ \text{points in}\ S}{\#\ \textbf{blue}\ \text{points in}\ S} \leq \frac{r}{b}$$

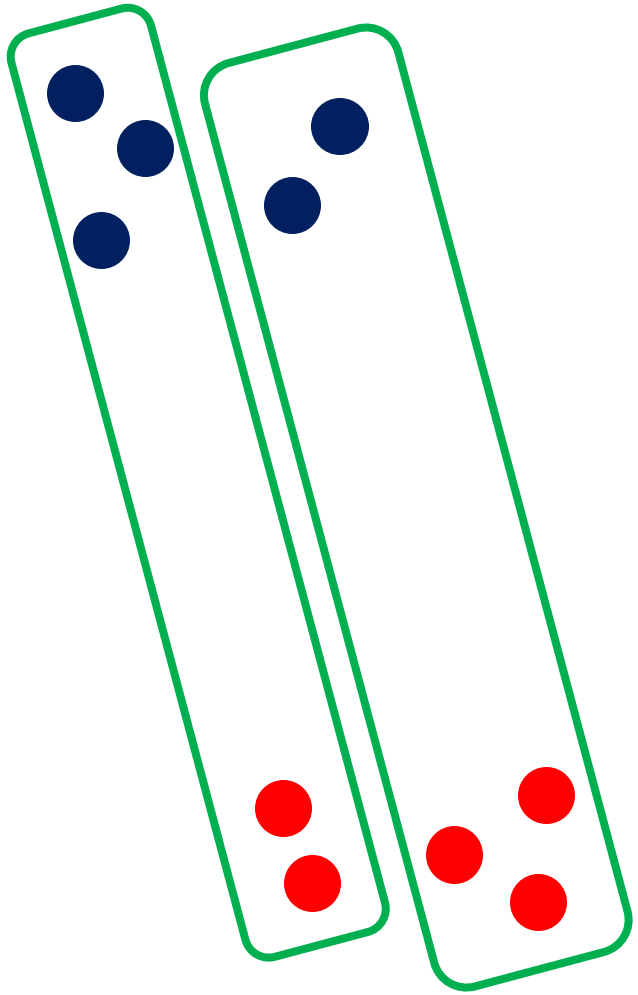


Problem Definition

- Collection of n points P in $\mathbb{R}^d$
- Each point is colored either **red** or **blue**
- Each cluster S has to be *(r,b)-balanced*

$$\frac{b}{r} \leq \frac{\#\ \textbf{red}\ \text{points in}\ S}{\#\ \textbf{blue}\ \text{points in}\ S} \leq \frac{r}{b}$$

$(3, 2)$-balanced



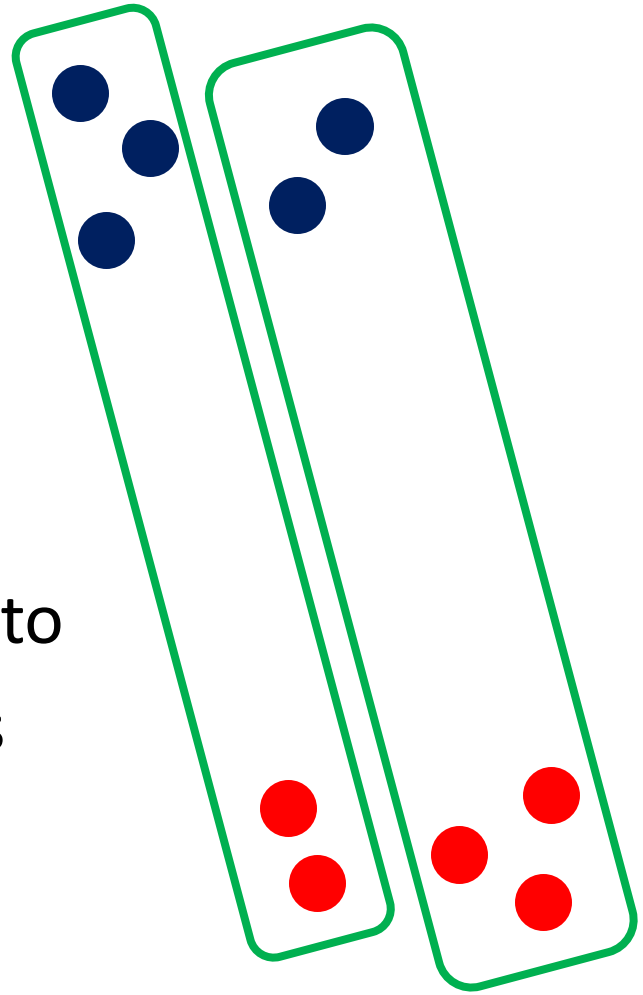
Problem Definition

- Collection of n points P in $\mathbb{R}^d$
- Each point is colored either **red** or **blue**
- Each cluster S has to be *(r,b)-balanced*

$$\frac{b}{r} \leq \frac{\#\ \textbf{red}\ \text{points in}\ S}{\#\ \textbf{blue}\ \text{points in}\ S} \leq \frac{r}{b}$$

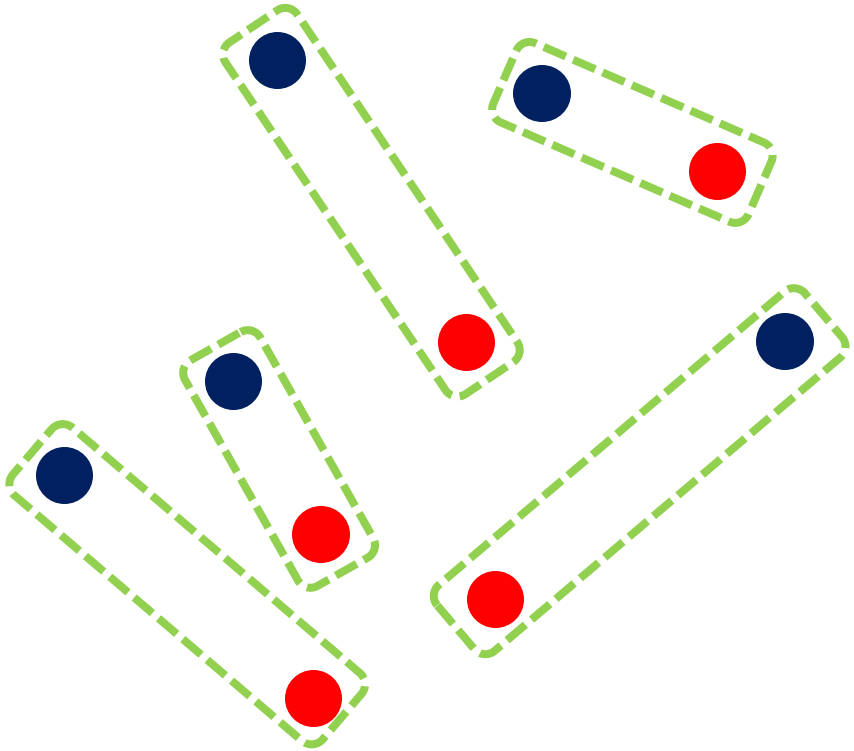
(r,b)-Fair *k*-median: Find $\boldsymbol{k}$ **centers** that partition P into ***(r,b)-balanced*** **clusters** s.t. **average distance** of points to their centers is **minimized**.

$(3, 2)$-balanced



Fair Clustering Through Fairlets [Chierichetti et al]

- **Fairlets**: minimal sets that satisfy the (r,b)-balance requirement



Fair Clustering Through Fairlets [Chierichetti et al]

- **Fairlets**: minimal sets that satisfy the (r,b)-balance requirement

Outline of Algorithm [Chierichetti et al, NeurIPS'17]

I. Compute an approximately optimal fairlet decomposition **α-approx**

II. Cluster the centers of fairlets into k groups **β-approx**

Theorem. The proposed algorithm is $O(\alpha + \beta)$-approximation

Fair Clustering Through Fairlets [Chierichetti et al]

- **Fairlets**: minimal sets that satisfy the (r,b)-balance requirement

Outline of Algorithm [Chierichetti et al, NeurIPS'17]

I. Compute an approximately optimal fairlet decomposition — α**-approx**

II. Cluster the centers of fairlets into k groups — β**-approx**

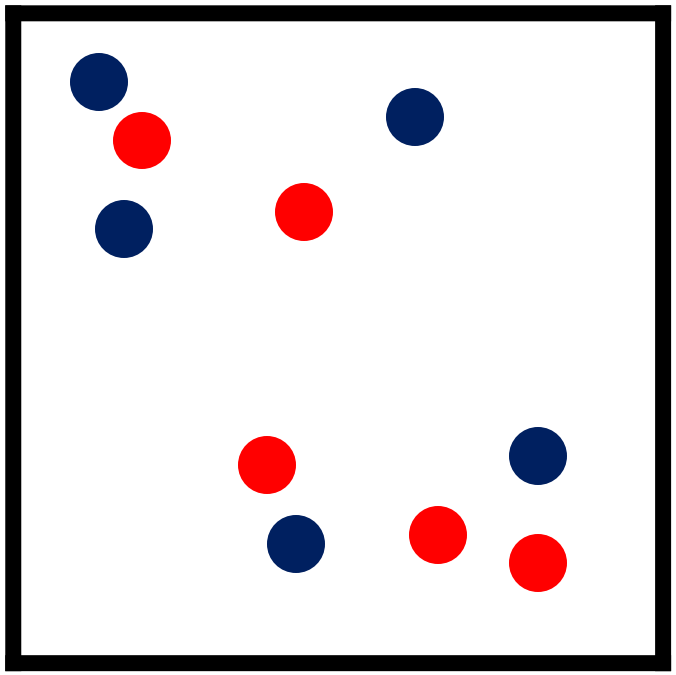
Theorem. The proposed algorithm is $O(\alpha + \beta)$-approximation

Limitations: **1)** Quadratic runtime in step **I**

2) Only works for $(t, 1)$-balanced (t is an integer)

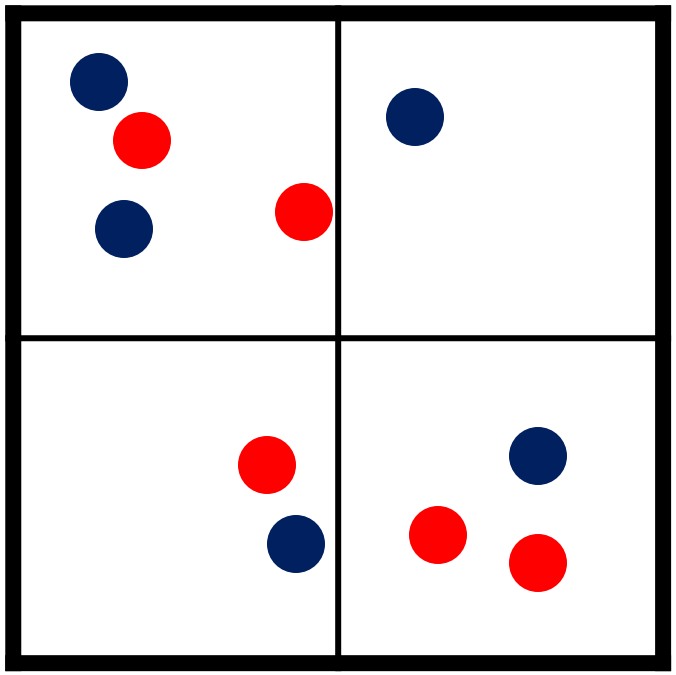
Near Linear Time Fairlet Decomposition

1. HST-embedding of points



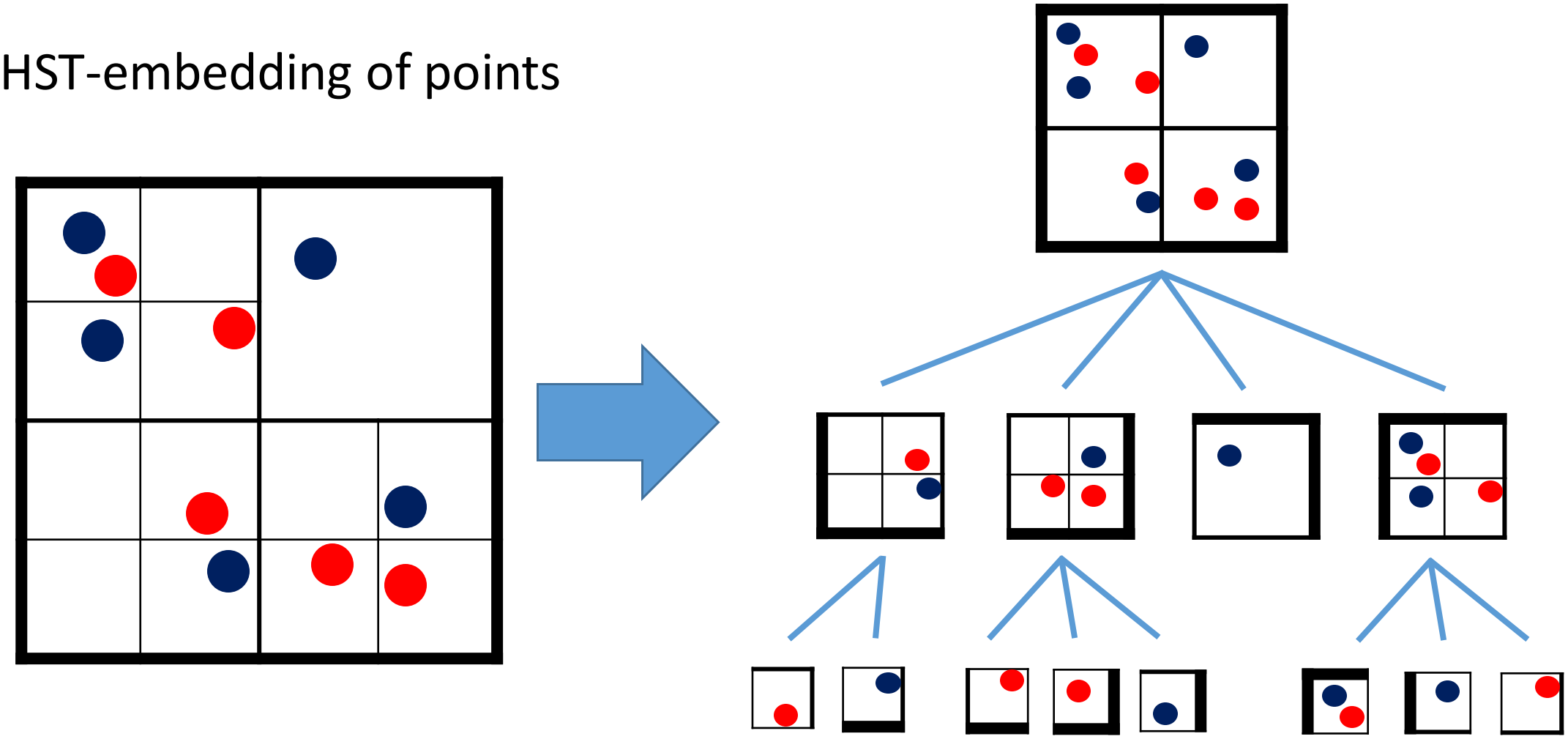
Near Linear Time Fairlet Decomposition

1. HST-embedding of points



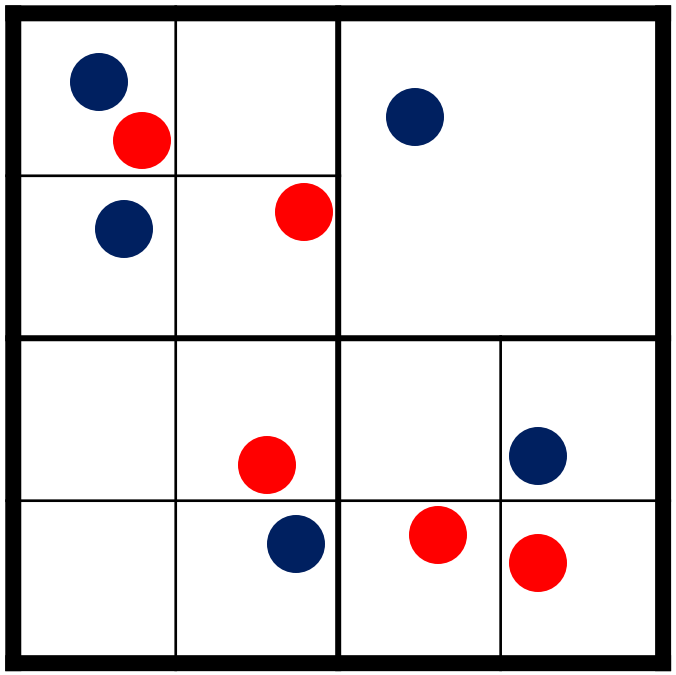
Near Linear Time Fairlet Decomposition

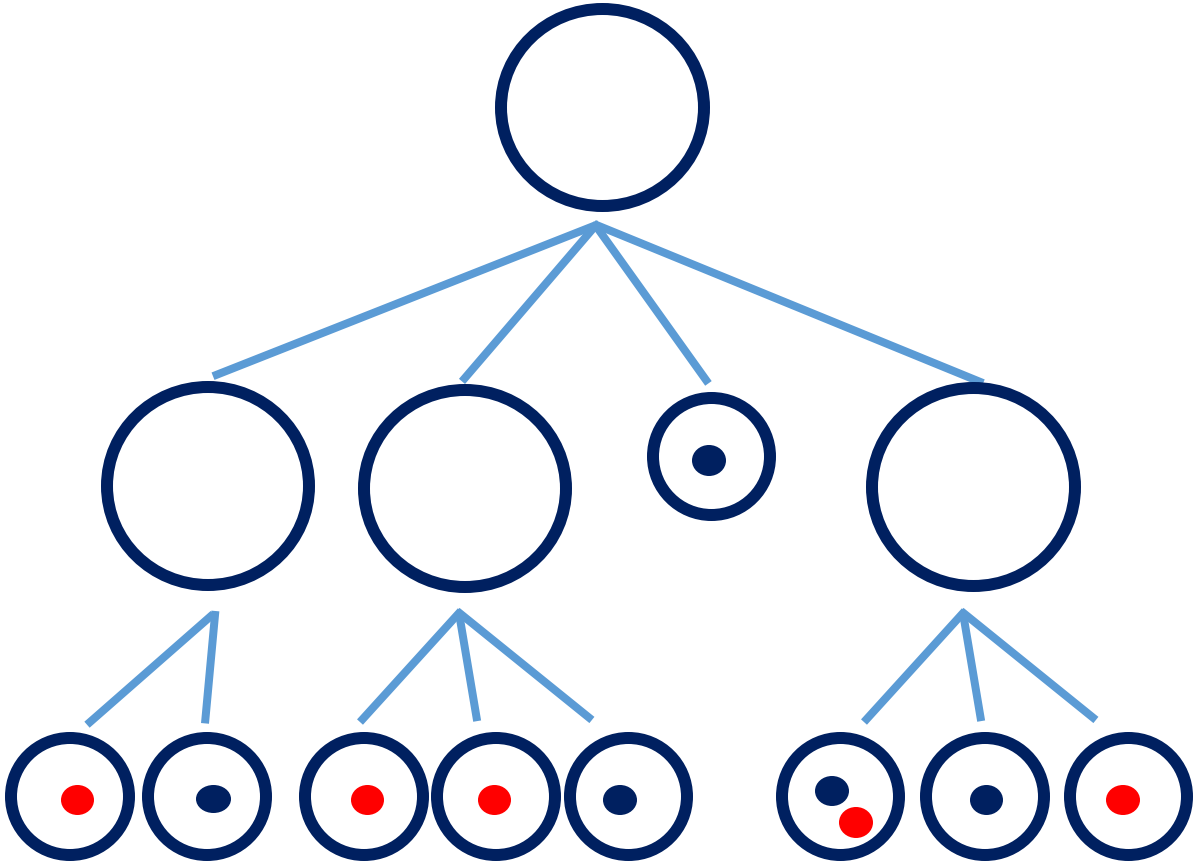
1. HST-embedding of points



Near Linear Time Fairlet Decomposition

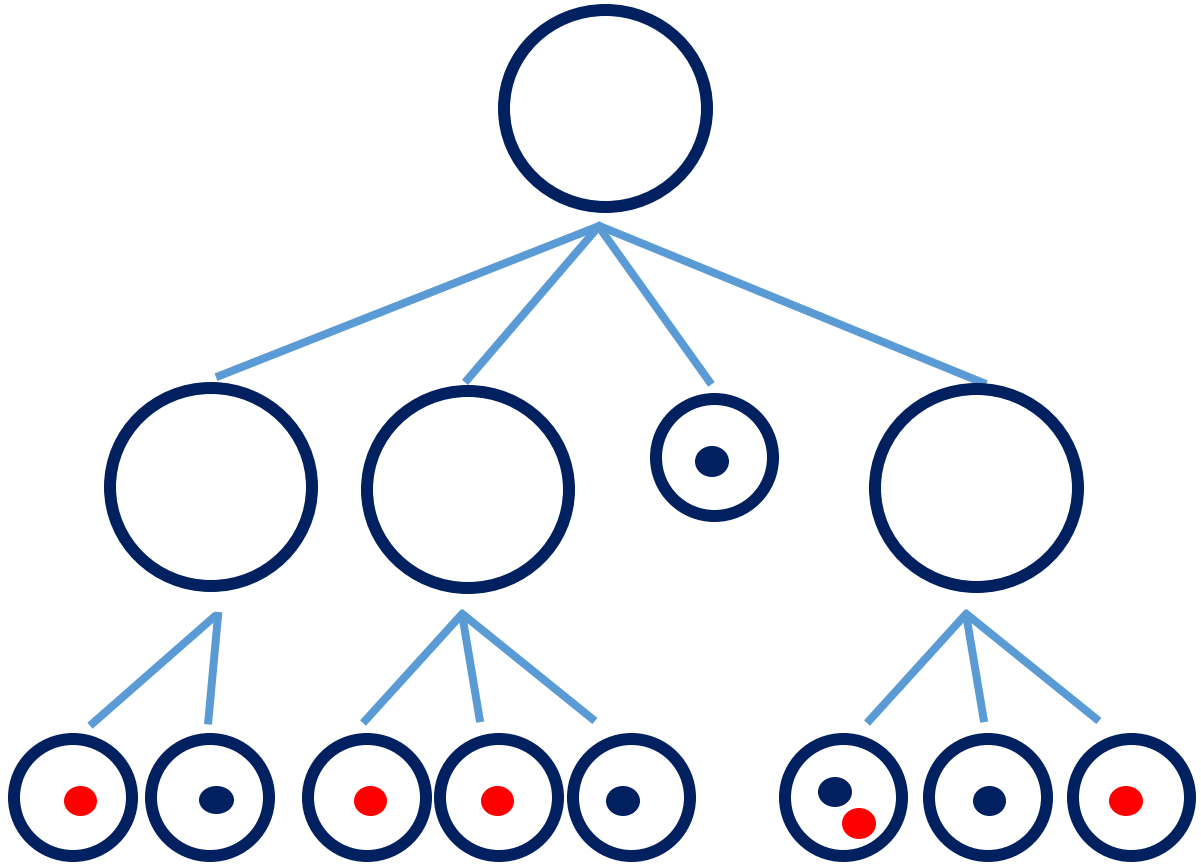
1. HST-embedding of points





Near Linear Time Fairlet Decomposition

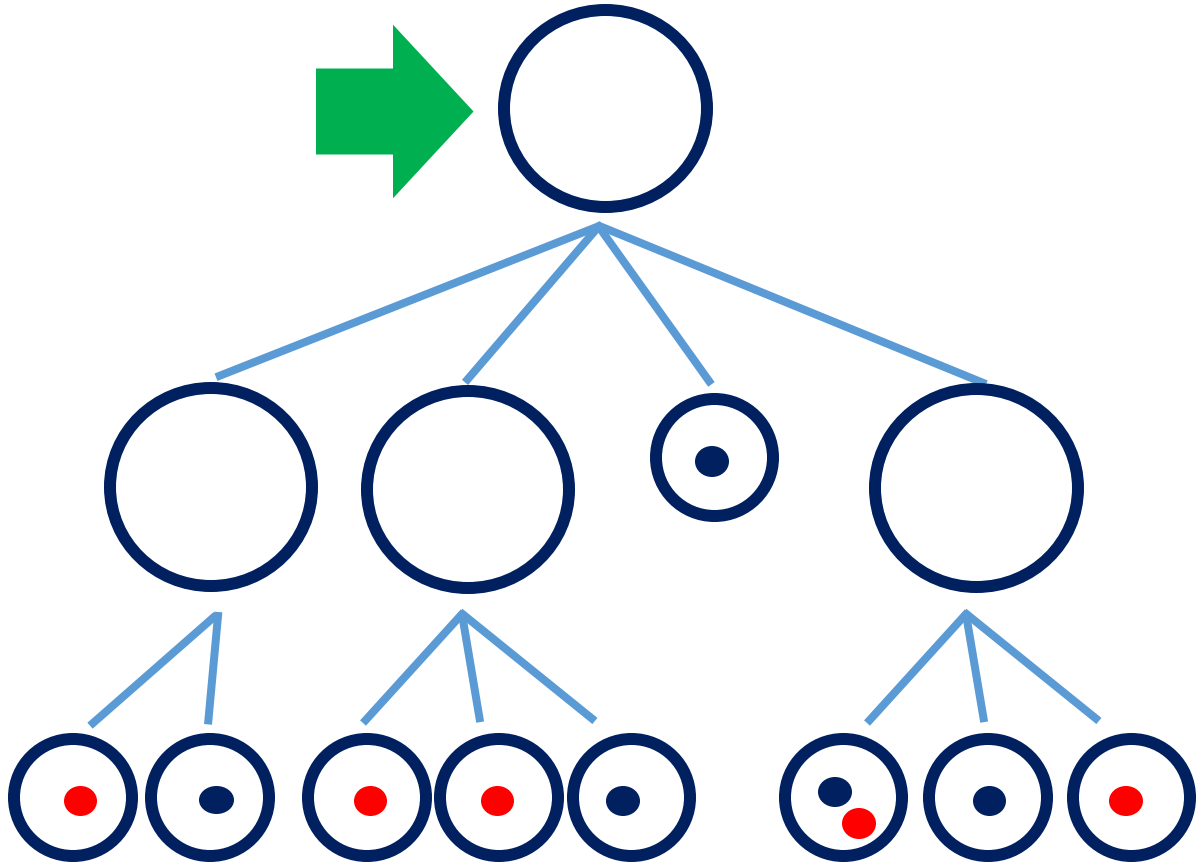
1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction



Near Linear Time Fairlet Decomposition

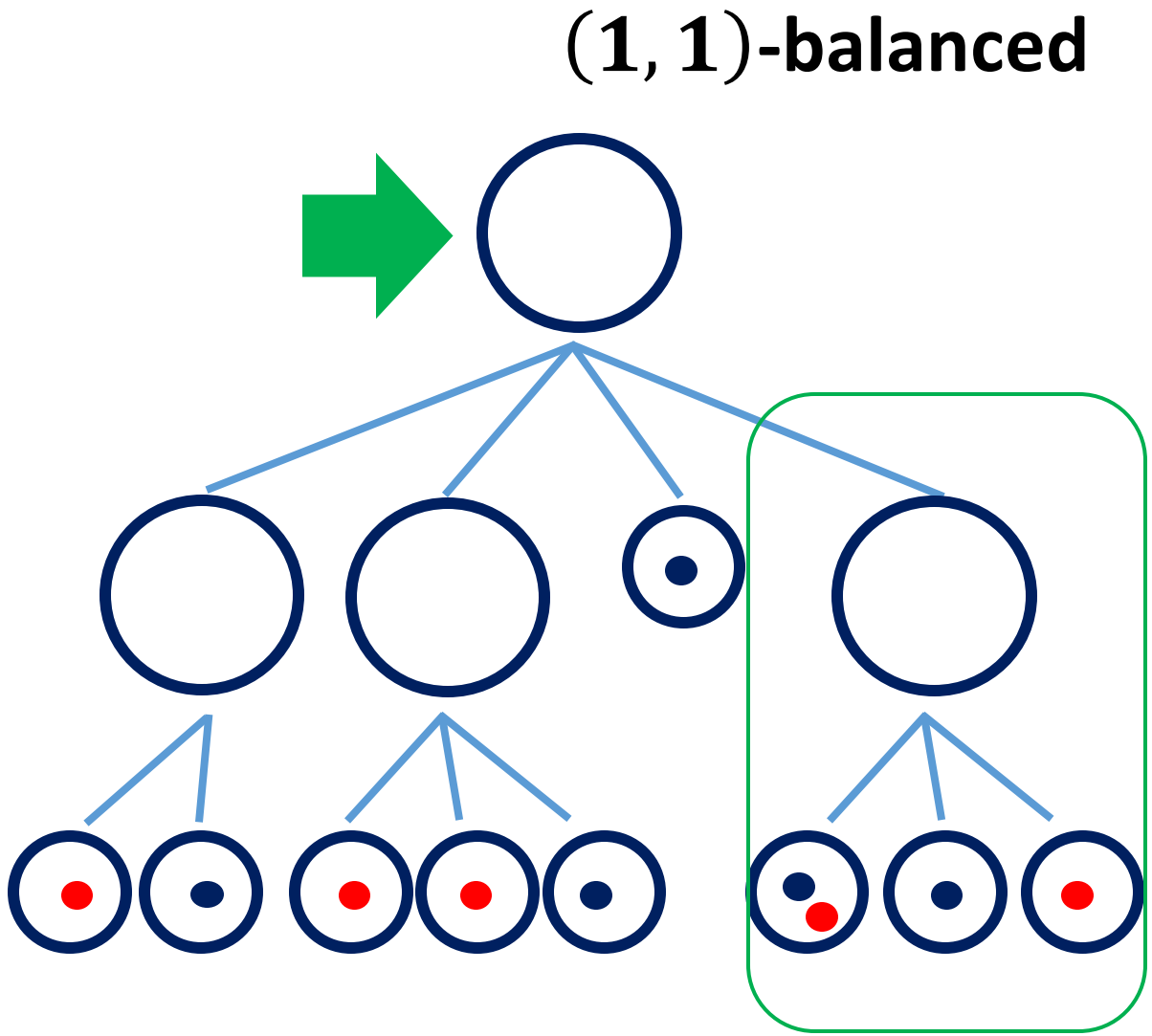
1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction

$(1, 1)$-balanced



Near Linear Time Fairlet Decomposition

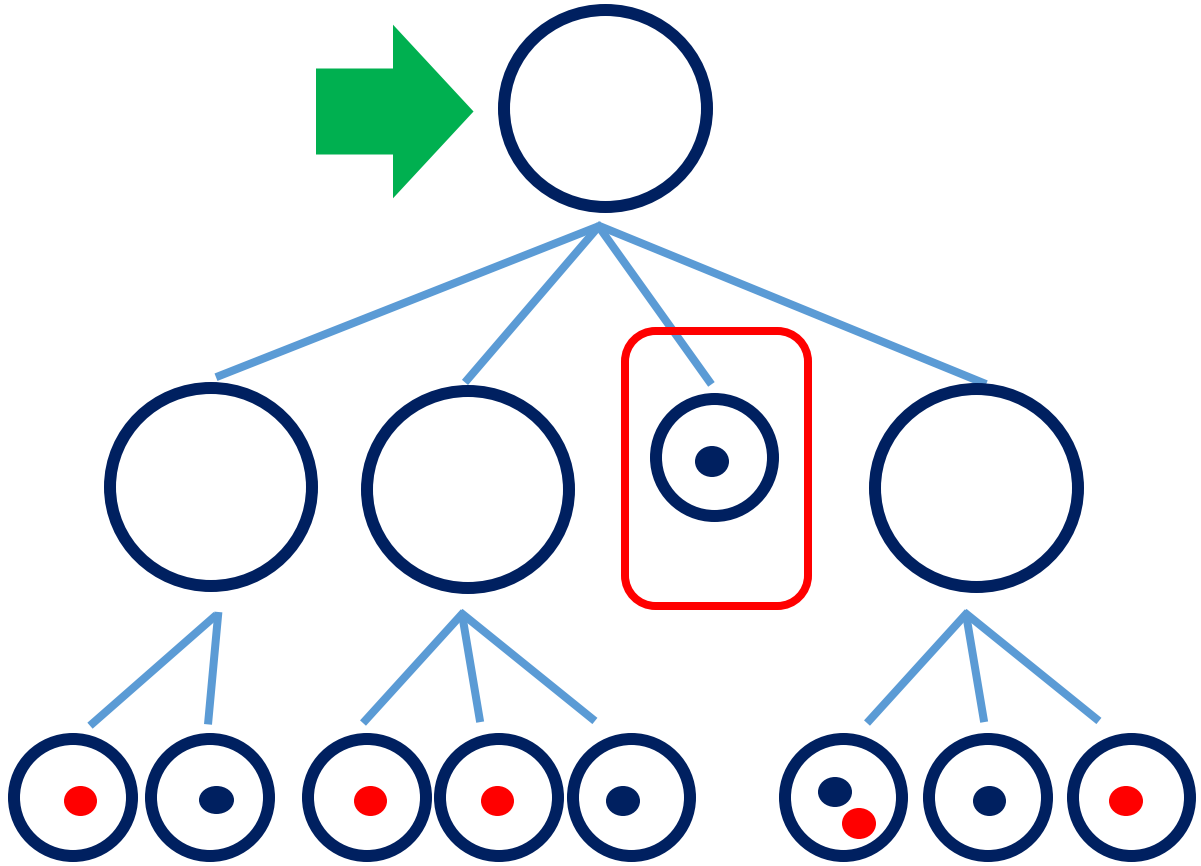
1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction



Near Linear Time Fairlet Decomposition

1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction

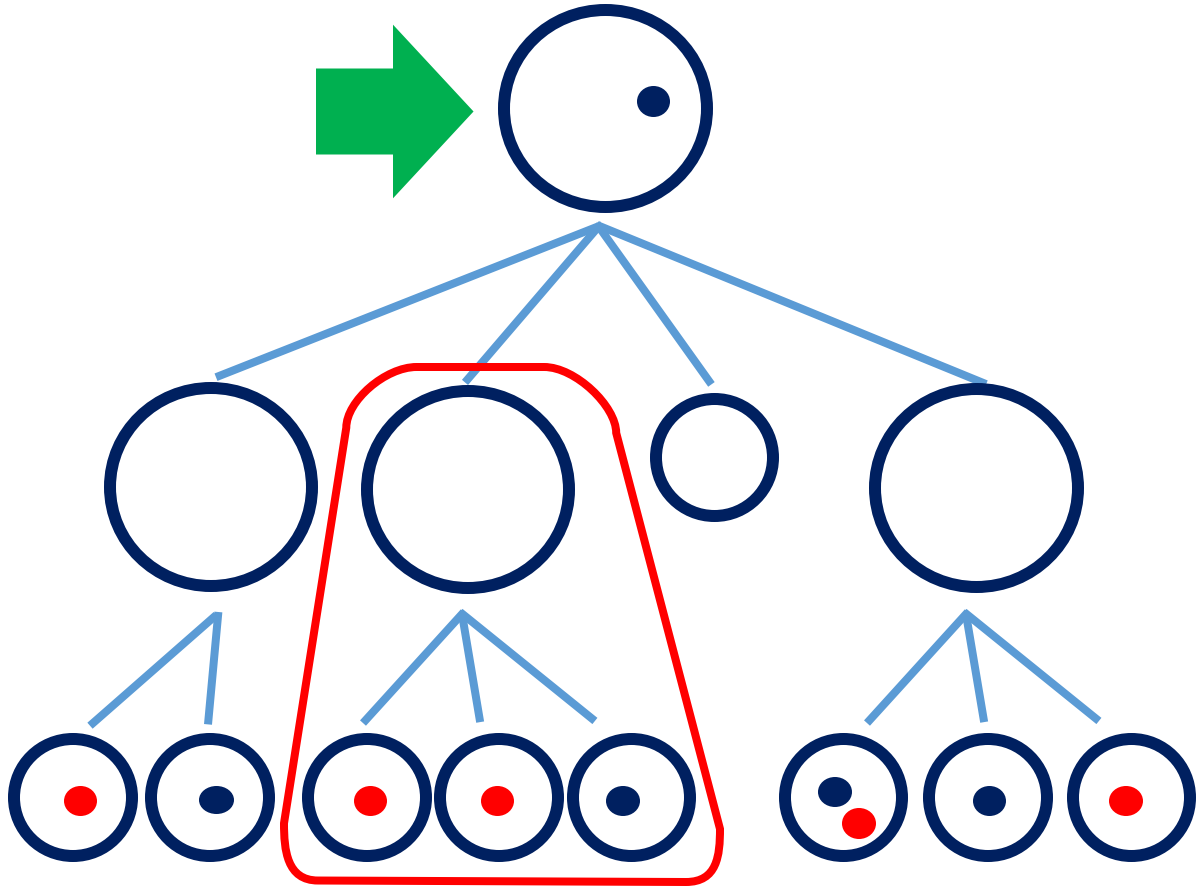
$(1, 1)$-balanced



Near Linear Time Fairlet Decomposition

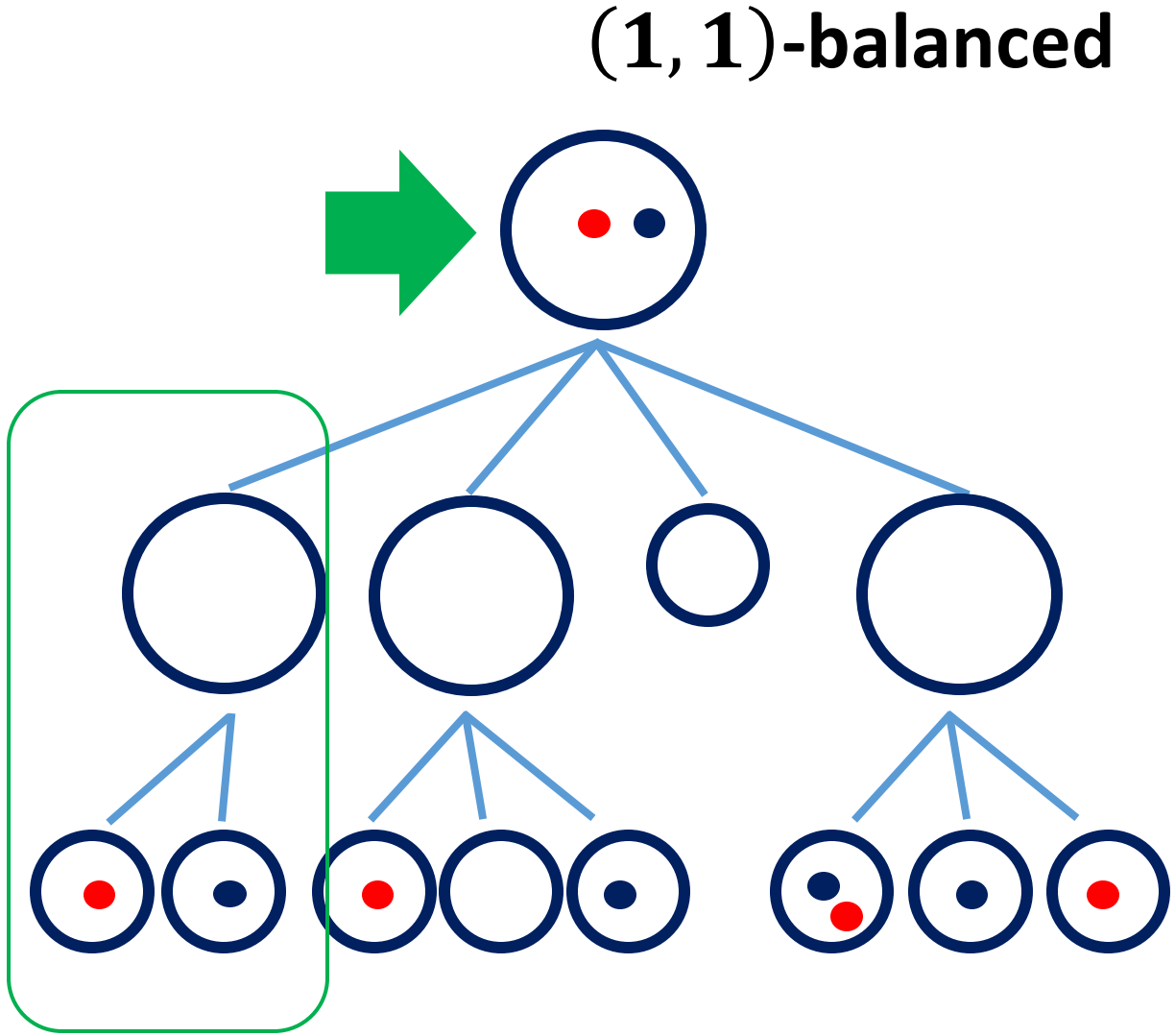
1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction

$(1, 1)$-balanced



Near Linear Time Fairlet Decomposition

1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction



Near Linear Time Fairlet Decomposition

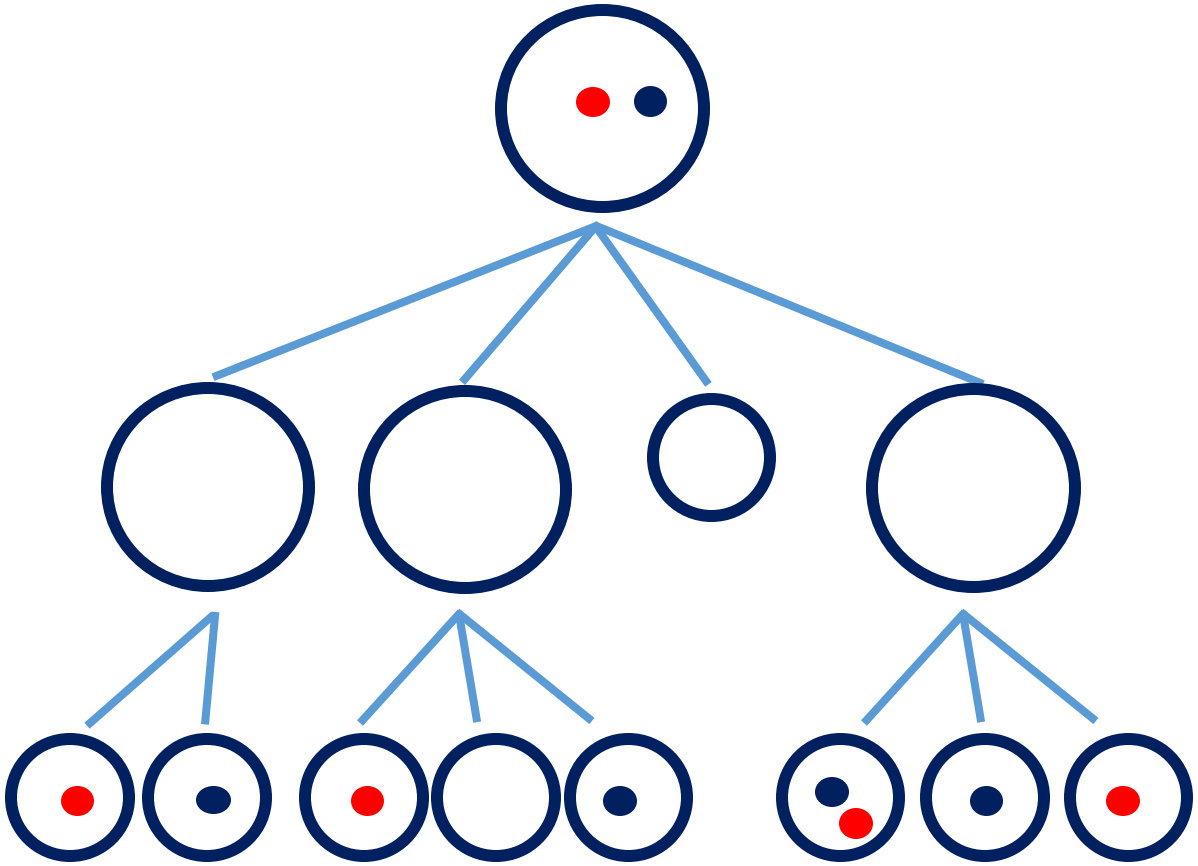
1. HST-embedding of points
2. Top-down traversal + greedy fairlet construction

Theorem. $O(d \cdot \log n)$**-approx.**

for fairlet-decomposition in $O(d \cdot n \cdot \log n)$ **time**

I. Runs in near-linear time

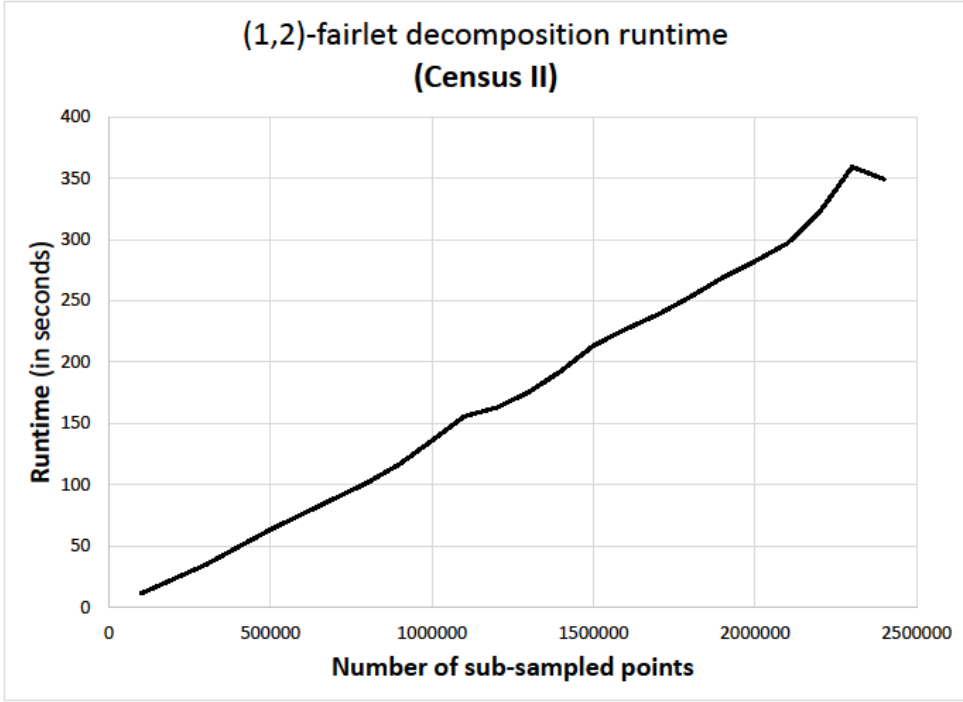
II. Works for all values of (r,b)



Empirical Results

Dataset	Balance	Fairlet Decomposition Cost	
		Previous Work*	Ours
Diabetes	0.8	~ 9836	2971
Bank	0.5	$\sim 5.46 \times 10^5$	5.24×10^5
Census	0.5	$\sim 3.59 \times 10^7$	2.41×10^7

*(Chierichetti et al., NeurIPS 2017)

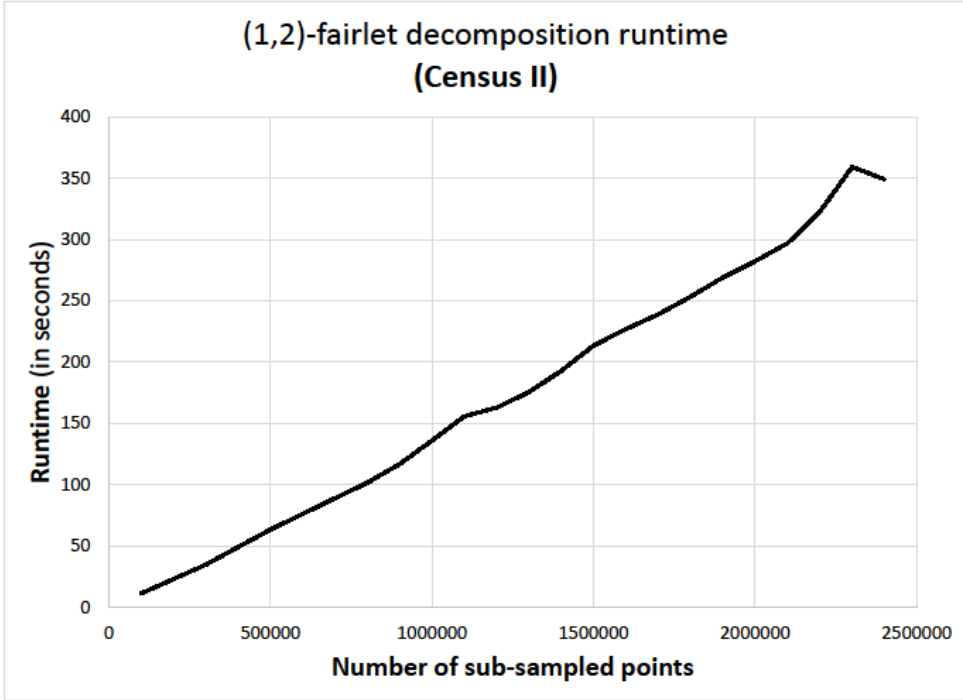


Runtime scales **almost linearly** in the number of points while the **empirical quality** is as good as (Chierichetti et al., 2017)

Empirical Results

Dataset	Balance	Fairlet Decomposition Cost	
		Previous Work*	Ours
Diabetes	0.8	~ 9836	2971
Bank	0.5	$\sim 5.46 \times 10^5$	5.24×10^5
Census	0.5	$\sim 3.59 \times 10^7$	2.41×10^7

*(Chierichetti et al., NeurIPS 2017)



Runtime scales **almost linearly** in the number of points while the **empirical quality** is as good as (Chierichetti et al., 2017)

Thank You!

Poster: @6:30 pm - Pacific Ballroom #84