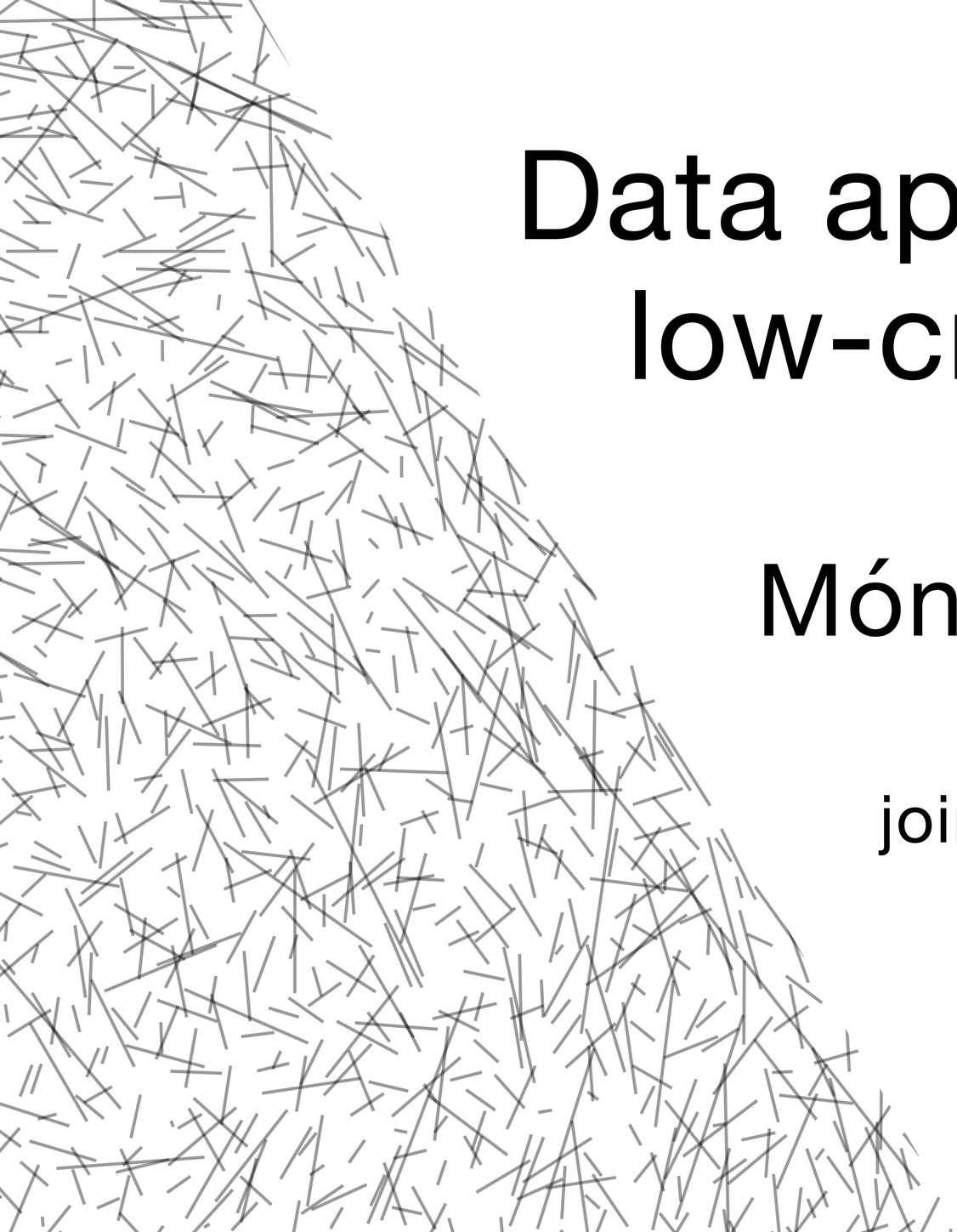


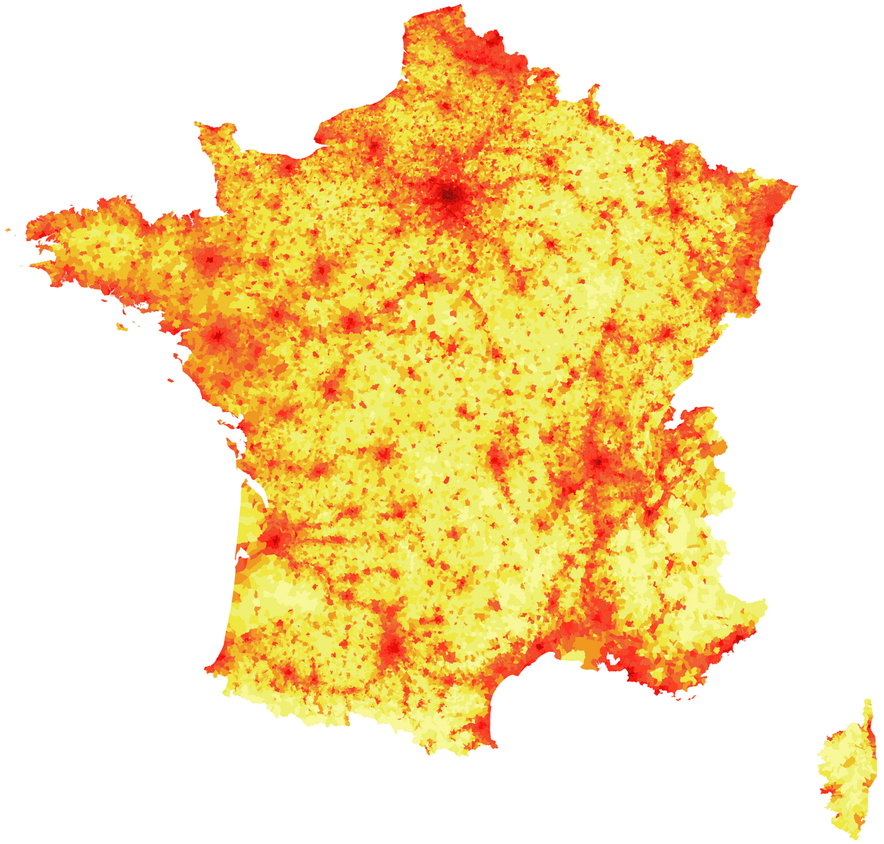
Data approximation using low-crossing matchings

Mónika Csikós

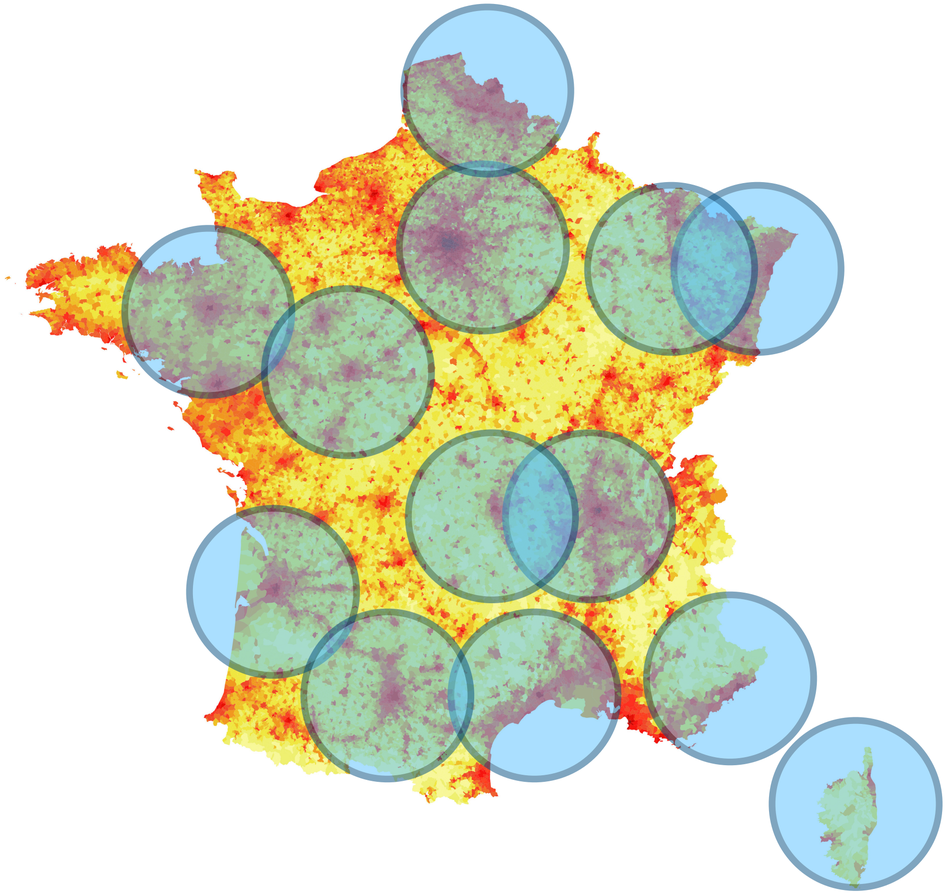
joint work with Nabil Mustafa

INSTITUT
DE RECHERCHE
EN INFORMATIQUE
FONDAMENTALE

How many people live within 100 km from a university?

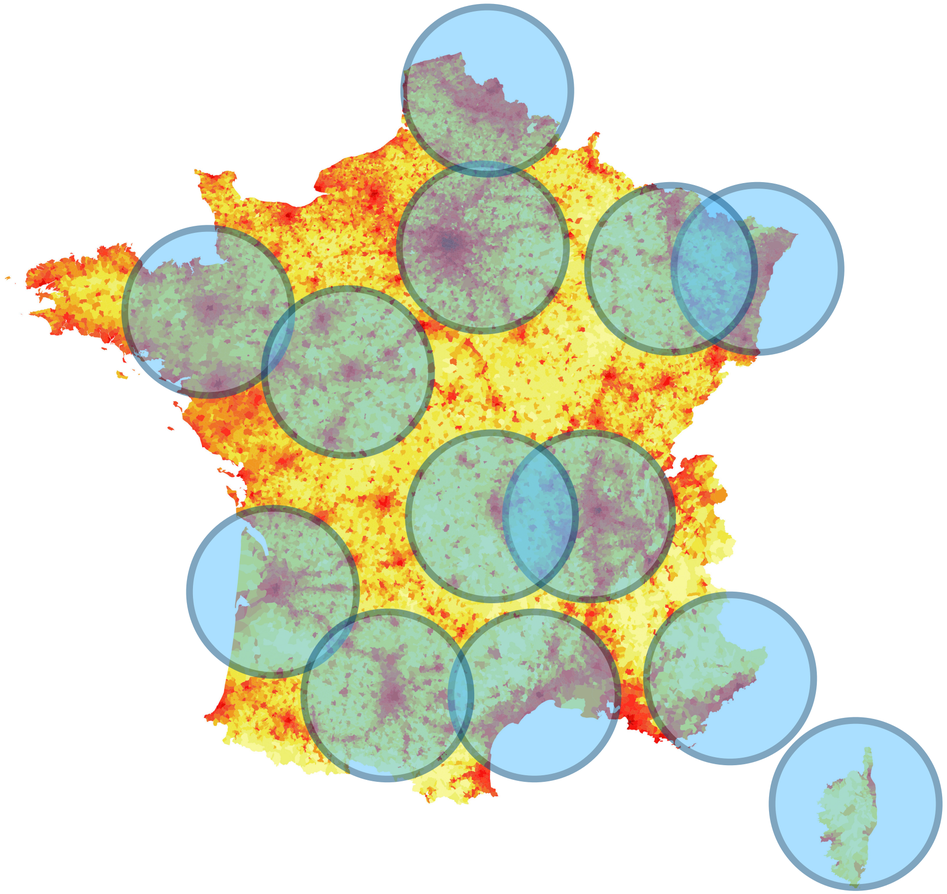


How many people live within 100 km from a university?



65M calls

How many people live within 100 km from a university?

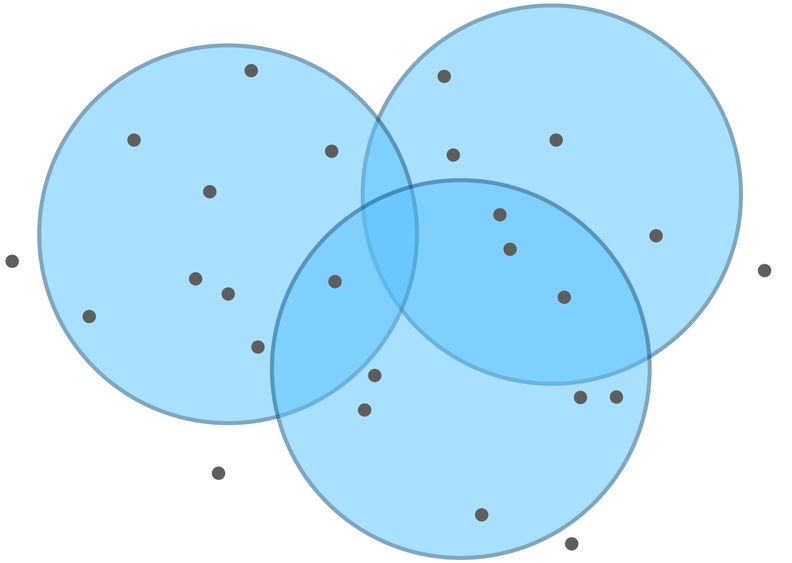


65M calls 32.5M calls

Combinatorial data approximation

Model with a set system: $(X, \mathcal{S})$

n elements m subsets of X



Combinatorial data approximation

Model with a set system: $(X, \mathcal{S})$

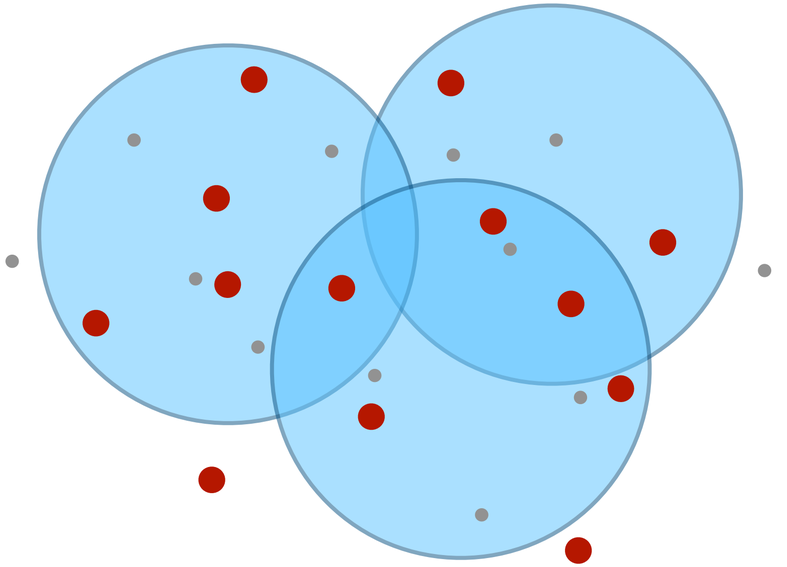
n elements m subsets of X

Goal: Pick half of the elements in X such that from each $S \in \mathcal{S}$,

'half' of the points in S are present in A.

Formally, $A \subset X$ of size $\frac{n}{2}$ that minimizes the discrepancy

$$\mathrm{disc}_{\mathcal{S}}(A) = \max_{S \in \mathcal{S}} \Big| \, |S| - 2\,|A \cap S| \, \Big|$$



Random sampling

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

Random subset $A \subseteq X$ s.t. $x \in A$ with probability $\frac{1}{2}$: $\mathbb{E}|A \cap S| = \frac{|S|}{2}$

Random sampling

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

Random subset $A \subseteq X$ s.t. $x \in A$ with probability $\frac{1}{2}$: $\mathbb{E}\,|A \cap S| = \frac{|S|}{2}$

Chernoff-Hoeffding + union bound:

$$\mathbb{P}\left[\max_{S \in \mathcal{S}} \left|2\,|A \cap S| - |S|\,\right| > \delta\right] \leq 2m \cdot \exp\left(-\frac{\delta^2}{2n}\right)$$

$$\Longrightarrow \operatorname{disc}_{\mathcal{S}}(A) = O\left(\sqrt{n \ln m}\right)$$

Random sampling

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

Recursively sample $A_1 \subseteq X$, $A_2 \subseteq A_1, A_3 \subseteq A_2, \ldots$, always halving the size

Random sampling

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

Recursively sample $A_1 \subseteq X$, $A_2 \subseteq A_1, A_3 \subseteq A_2, \ldots$, always halving the size

$$|A_1| = \frac{n}{2} \qquad 2\,|A_1 \cap S| = |S| \pm O\left(\sqrt{n \ln m}\right)$$

$$|A_2| = \frac{n}{4} \qquad 4\,|A_2 \cap S| = |S| \pm O\left(\sqrt{2n \ln m}\right)$$

$$\vdots$$

$$|A_i| = \frac{n}{2^i} \qquad 2^i\,|A_i \cap S| = |S| \pm O\left(\sqrt{2^i n \ln m}\right)$$

Epsilon-approximations

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

ε-approximation problem:

Given $\varepsilon \in (0,1)$, find smallest $A \subseteq X$ such that $\max_{S \in \mathcal{S}} \left| \frac{|S|}{|X|} - \frac{|A \cap S|}{|A|} \right| \leq \varepsilon$

Uniform sampling gives

$$|A_i| = \frac{n}{2^i} \qquad 2^i |A_i \cap S| = |S| \pm \underbrace{O\left(\sqrt{2^i n \ln m}\right)}_{\leq \varepsilon n} \implies i \leq \log \frac{\varepsilon^2 n}{\ln m}$$

$\implies$ A uniform sample of $O\left(\frac{\ln m}{\varepsilon^2}\right)$ points is an ε-approximation with constant probability

Uniform sampling guarantees

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

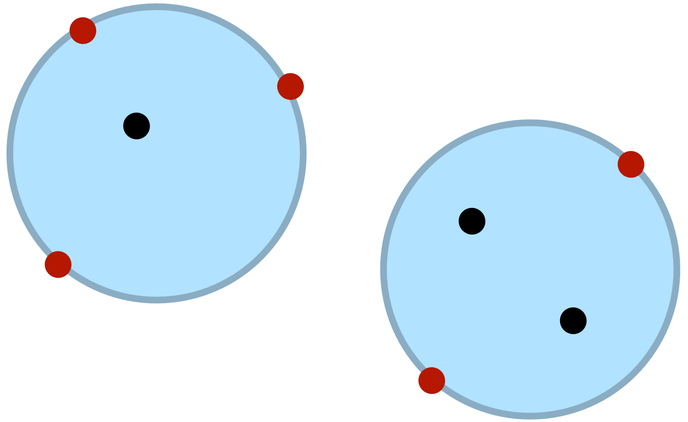
	Arbitrary
Discrepancy error	$\sqrt{n \ln m}$
ε-approximation size	$\frac{\ln m}{\varepsilon^2}$

Structured set systems

set system $(X, \mathcal{S})$

X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk

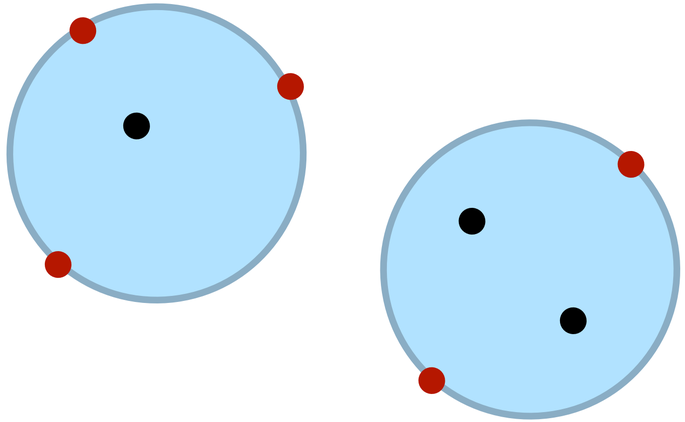


Structured set systems

set system $(X, \mathcal{S})$

X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk $\implies \left|\mathcal{S}|_Y\right| < 2^{|Y|}$ even for $|Y| = 4$



$\mathcal{S}|_Y = \{Y \cap S \ : \ S \in \mathcal{S}\}$

Structured set systems

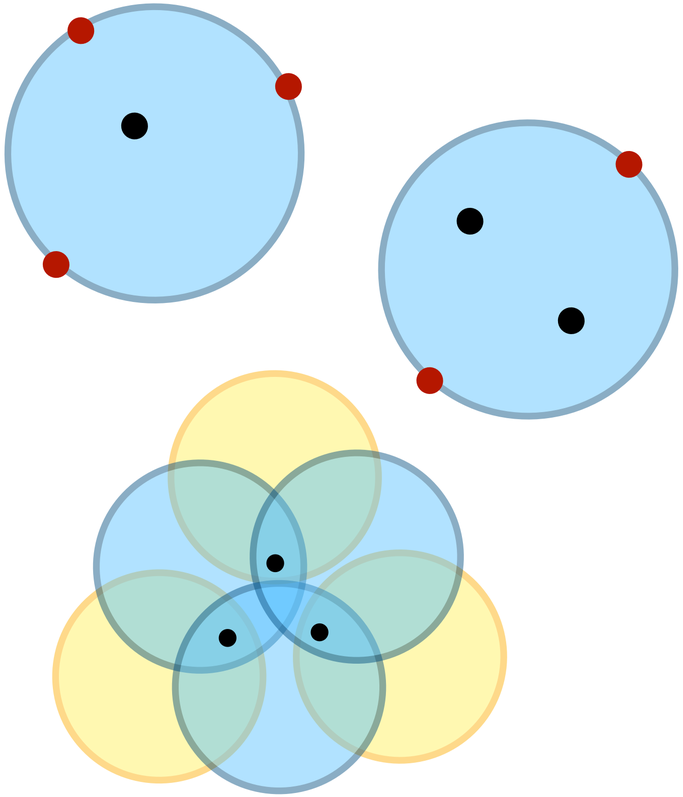
set system $(X, \mathcal{S})$

X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk $\implies \left|\mathcal{S}|_Y\right| < 2^{|Y|}$ even for $|Y| = 4$

but $\left|\mathcal{S}|_Y\right| = 2^3$ possible for $|Y| = 3$

$\mathcal{S}|_Y = \{Y \cap S \ : \ S \in \mathcal{S}\}$



Structured set systems

set system $(X, \mathcal{S})$

X : set of n points in $\mathbb{R}^2$

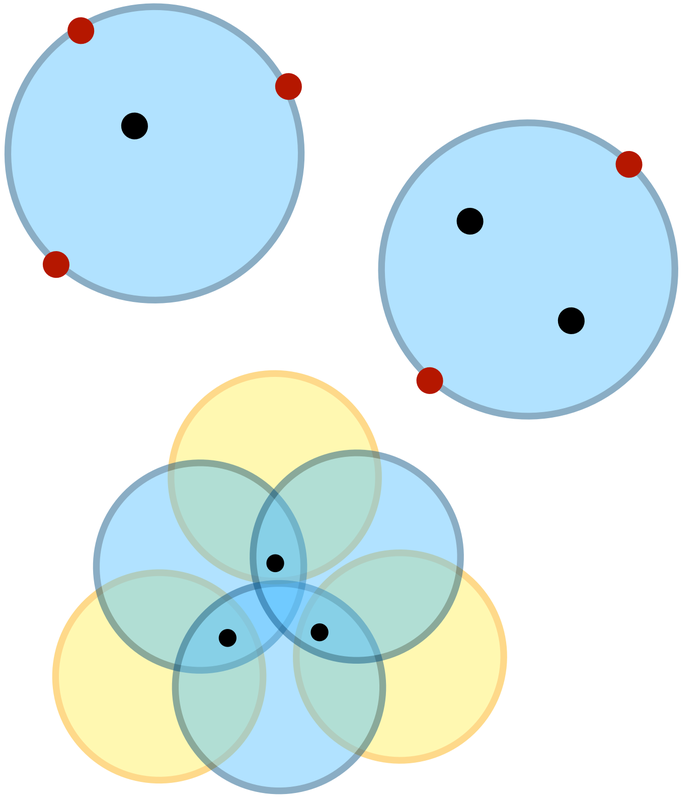
each set in $\mathcal{S}$ is defined by a disk $\Longrightarrow \left|\mathcal{S}|_Y\right| < 2^{|Y|}$ even for $|Y| = 4$

but $\left|\mathcal{S}|_Y\right| = 2^3$ possible for $|Y| = 3$

$\mathcal{S}|_Y = \{Y \cap S \ : \ S \in \mathcal{S}\}$

$\Longrightarrow$ VC-dimension of $(X, \mathcal{S})$ is 3.

The VC-dimension of $(X, \mathcal{S})$ is the size of the largest $Y \subseteq X$ s.t. $\left|\mathcal{S}|_Y\right| = 2^{|Y|}$.



Structured set systems

set system $(X, \mathcal{S})$

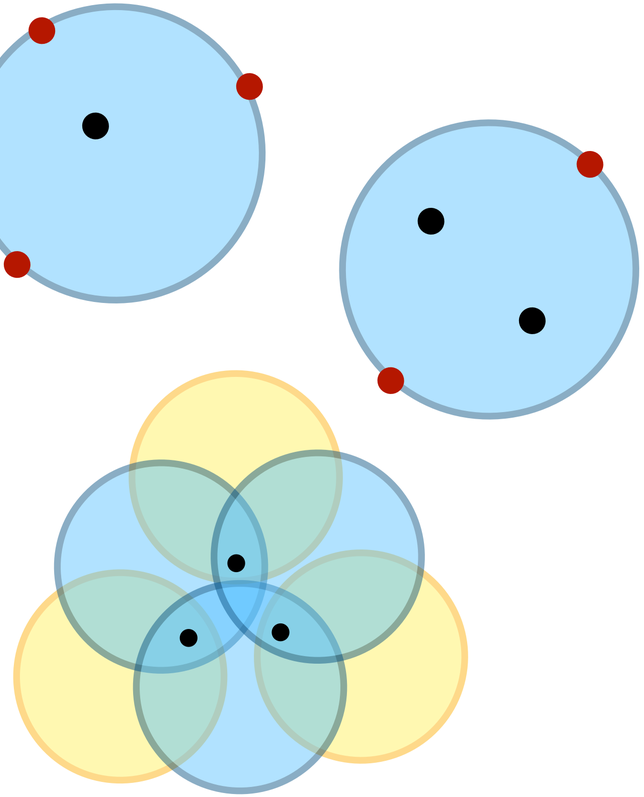
X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk $\Longrightarrow \left|\mathcal{S}|_Y\right| < 2^{|Y|}$ even for $|Y| = 4$

but $\left|\mathcal{S}|_Y\right| = 2^3$ possible for $|Y| = 3$

$\mathcal{S}|_Y = \{Y \cap S \ : \ S \in \mathcal{S}\}$

$\Longrightarrow$ VC-dimension of $(X, \mathcal{S})$ is 3.



The VC-dimension of $(X, \mathcal{S})$ is the size of the largest $Y \subseteq X$ s.t. $\left|\mathcal{S}|_Y\right| = 2^{|Y|}$.

$$\mathrm{d_{VC}}(X, \mathcal{S}) = d \quad \Longrightarrow \quad \left|\mathcal{S}|_Y\right| = O\left(|Y|^d\right)$$

(Vapnik-Chervonenkis 1971, Sauer, Shelah 1972)

Uniform sampling guarantees

set system $(X, \mathcal{S})$, $n = |X|$, $m = |\mathcal{S}|$

	Arbitrary	VC-dim $\leq d$
Discrepancy error	$\sqrt{n \ln m}$	$\sqrt{dn \ln n}$
ε-approximation size	$\frac{\ln m}{\varepsilon^2}$	$\frac{d}{\varepsilon^2}$

(Li-Long-Srinivasan 2001)

Non-uniform sampling

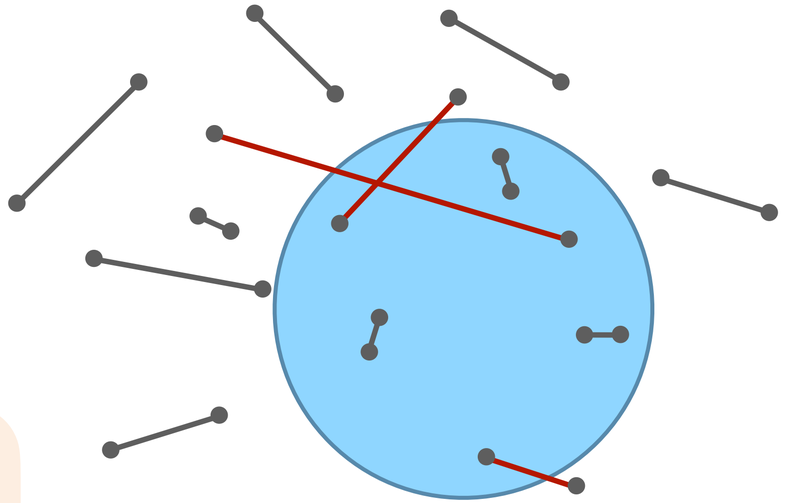
X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk

Non-uniform sampling

X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk



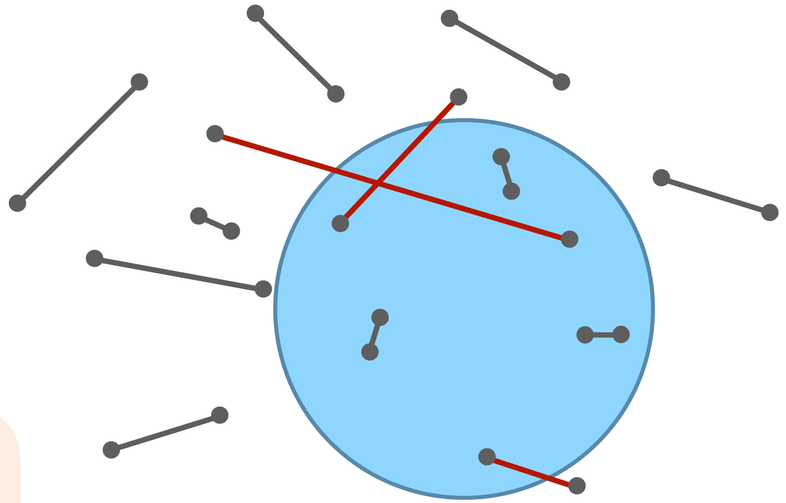
Theorem (Welzl 1988, Chazelle-Welzl 1989)

X has a perfect matching such that any disk crosses $O\left(\sqrt{n}\right)$ edges.

Non-uniform sampling

X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk



Theorem (Welzl 1988, Chazelle-Welzl 1989)

X has a perfect matching such that any disk crosses $O\left(\sqrt{n}\right)$ edges.

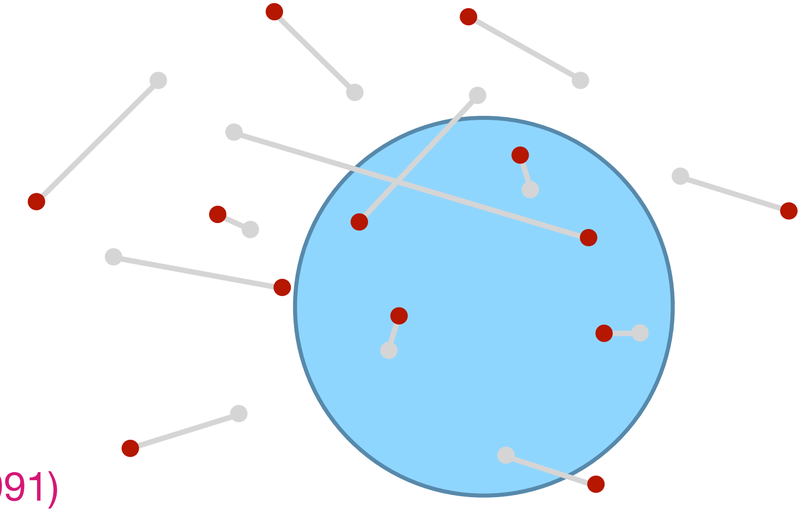
Idea: Take one end-point of each matching edge randomly

$\Longrightarrow$ only $O\left(\sqrt{n}\right)$ random variables for each disk

$$\mathbb{P}\left[\max_{S\in\mathcal{S}} \left|2\,|A\cap S| - |S|\right| > \delta\right] \leq m\cdot\exp\left(-\frac{\delta^2}{2\sqrt{n}}\right)$$

$\Longrightarrow$ $\operatorname{disc}_{\mathcal{S}}(A) = O\left(n^{1/4}\sqrt{\ln n}\right)$ with constant probability

(Matoušek-Welzl-Wernish 1991)



Non-uniform sampling

X : set of n points in $\mathbb{R}^2$

each set in $\mathcal{S}$ is defined by a disk

Theorem (Welzl 1988, Chazelle-Welzl 1989)

X has a perfect matching such that any disk crosses $O\left(\sqrt{n}\right)$ edges.

Idea: Take one end-point of each matching edge randomly

$\Longrightarrow$ only $O\left(\sqrt{n}\right)$ random variables for each disk

$$\mathbb{P}\left[\max_{S\in\mathcal{S}} \left|2\,|A\cap S| - |S|\right| > \delta\right] \leq m\cdot\exp\left(-\frac{\delta^2}{2\sqrt{n}}\right)$$

$\Longrightarrow$ $\mathrm{disc}_{\mathcal{S}}(A) = O\left(n^{1/4}\sqrt{\ln n}\right)$ with constant probability

(Matoušek-Welzl-Wernish 1991)

X : n points in $\mathbb{R}^d$

$\mathcal{S}$: balls in $\mathbb{R}^d$

$O\left(n^{1-1/d}\right)$

$O\left(\sqrt{dn^{1-1/d}\ln n}\right)$

Approximation bounds

X : n points in $\mathbb{R}^d$, $\mathcal{S}$: subsets induced by balls

	Uniform sampling		Structural properties + non-uniform sampling
	Arbitrary	VC-dim $\leq d$	
Discrepancy error	$\sqrt{n \ln m}$	$\sqrt{dn \ln n}$	$\sqrt{dn^{1-1/d} \ln n}$
ε-approximation size	$\frac{\ln m}{\varepsilon^2}$	$\frac{d}{\varepsilon^2}$	$\left(\frac{d \ln \frac{1}{\varepsilon}}{\varepsilon^2}\right)^{1-1/d}$

Approximation bounds

X : n points in $\mathbb{R}^d$, $\mathcal{S}$: subsets induced by balls

		Uniform sampling	
	Arbitrary	VC-dim $\leq d$	Structural properties + non-uniform sampling
Discrepancy error	$\sqrt{n \ln m}$	$\sqrt{dn \ln n}$	$\sqrt{dn^{1-1/d} \ln n}$
ε-approximation size	$\frac{\ln m}{\varepsilon^2}$	$\frac{d}{\varepsilon^2}$	$\left(\frac{d \ln \frac{1}{\varepsilon}}{\varepsilon^2}\right)^{1-1/d}$

Need a matching to sample

Iterative reweighing method

Maintain weights on balls

1. initially, each ball has weight $= 1$

2. for $i = 1,\ldots, n/2$
   - select an edge e_i crossing balls of minimum total weight
   - double the weight of each balls crossing e_i
   - remove all edges incident to e_i

Return $e_1, \ldots, e_{n/2}$

(Chazelle-Welzl 1989)

Iterative reweighing method

Maintain weights on balls

1. initially, each ball has weight $= 1$

2. for $i = 1, \ldots, n/2$
   - select an edge e_i crossing balls of minimum total weight — $O(n^2 m)$ time
   - double the weight of each balls crossing e_i
   - remove all edges incident to e_i

Return $e_1, \ldots, e_{n/2}$

(Chazelle-Welzl 1989)

Iterative reweighing method

Maintain weights on balls

1. initially, each ball has weight $= 1$

2. for $i = 1,\ldots, n/2$
   - select an edge e_i crossing balls of minimum total weight
   - double the weight of each balls crossing e_i
   - remove all edges incident to e_i

Return $e_1, \ldots, e_{n/2}$

$O(n^2m)$ time

sampling according to edge weights

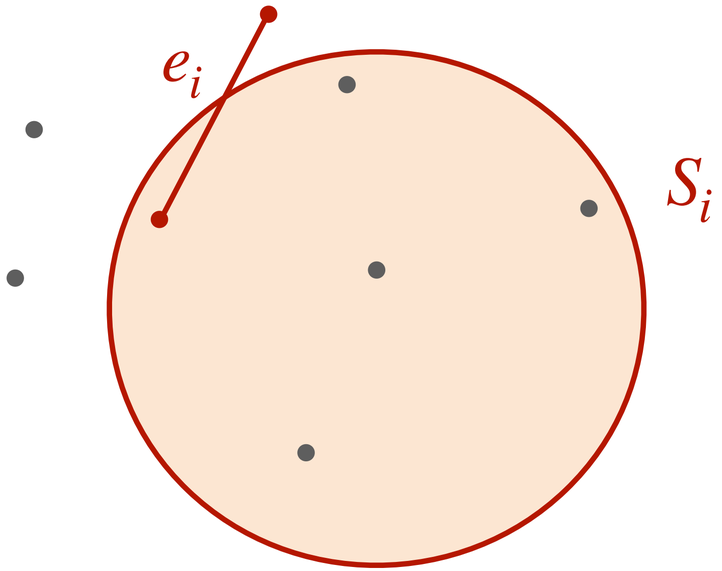
(Chazelle-Welzl 1989)

Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2.
   - sample an edge e_i according to the current edge-weights
   - sample a ball S_i according to the current range-weights
   - double the weight of each ball crossing e_i
   - halve the weight of each edge crossing S_i



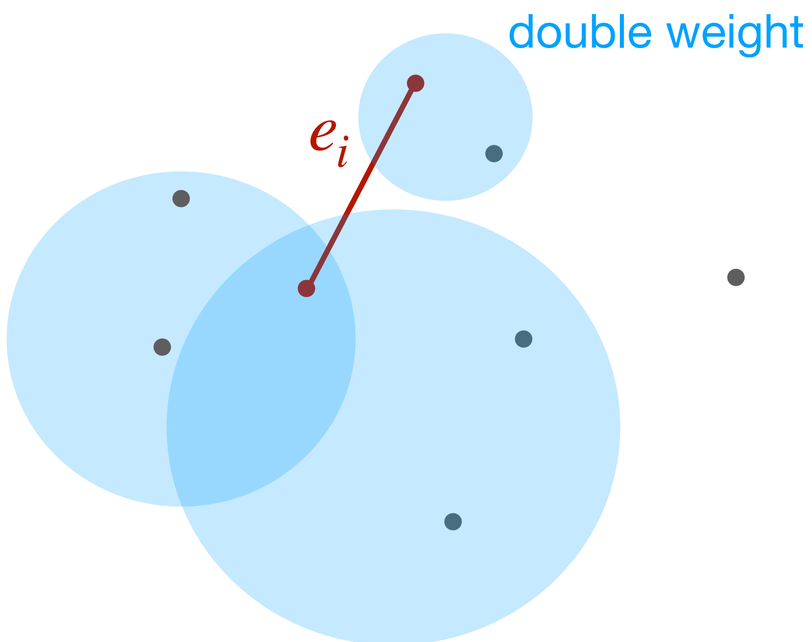
Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2.

- sample an edge e_i according to the current edge-weights
- sample a ball S_i according to the current range-weights
- double the weight of each ball crossing e_i
- halve the weight of each edge crossing S_i



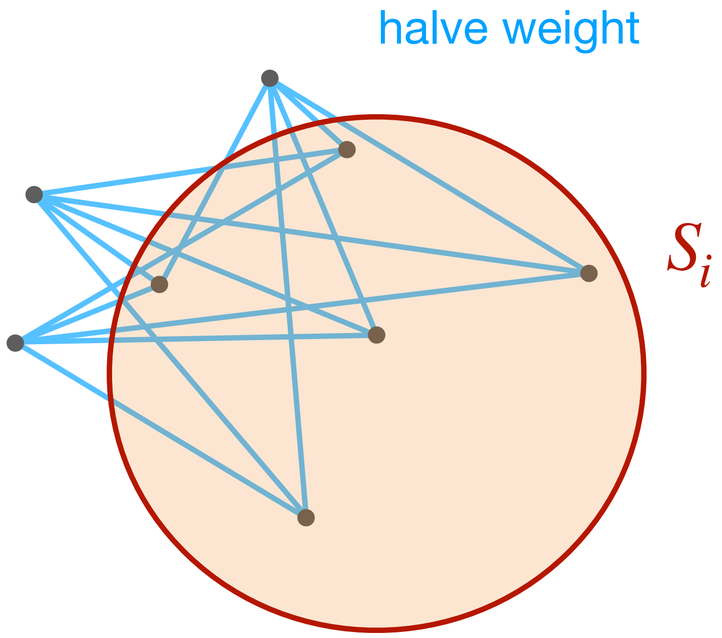
Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2.

- sample an edge e_i according to the current edge-weights
- sample a ball S_i according to the current range-weights
- double the weight of each ball crossing e_i
- halve the weight of each edge crossing S_i



Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2.
   - sample an edge e_i according to the current edge-weights
   - sample a ball S_i according to the current range-weights
   - double the weight of each ball crossing e_i
   - halve the weight of each edge crossing S_i
   - add e_i to the matching and remove all edges incident to e_i

Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2. for $i = 1, \ldots, n/4$
   - sample an edge e_i according to the current edge-weights
   - sample a ball S_i according to the current range-weights
   - double the weight of each ball crossing e_i
   - halve the weight of each edge crossing S_i
   - add e_i to the matching and remove all edges incident to e_i

3. recurse on uncovered points

$$\mathbb{E} \text{ crossing number of } e_1, \ldots, e_{n/4} \lesssim n^{1-1/d} + \ln m$$

Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2. for $i = 1, \ldots, n/4$
   - sample an edge e_i according to the current edge-weights
   - sample a ball S_i according to the current range-weights
   - double the weight of each ball crossing e_i
   - halve the weight of each edge crossing S_i
   - add e_i to the matching and remove all edges incident to e_i

$O(m + n^2)$ time

3. recurse on uncovered points

$$\mathbb{E} \text{ crossing number of } e_1, \ldots, e_{n/4} \lesssim n^{1-1/d} + \ln m$$

Reweighing on edges and balls

Maintain weights on edges and balls

1. initially, each ball and edge has weight $= 1$

2. for $i = 1, \ldots, n/4$
   - sample an edge e_i according to the current edge-weights
   - sample a ball S_i according to the current range-weights
   - double the weight of each ball crossing e_i
   - halve the weight of each edge crossing S_i
   - add e_i to the matching and remove all edges incident to e_i

3. recurse on uncovered points

Random update with probability $\mathbf{p}$

$$\mathbb{E} \text{ crossing number of } e_1, \ldots, e_{n/4} \lesssim n^{1-1/d} + \frac{\ln m}{\mathbf{p}}$$

Our method

Maintain weights on edges and balls

0. sample $E_1, \ldots, E_{n/4} \subset E$ and $\mathcal{S}_1, \ldots, \mathcal{S}_{n/4} \subset \mathcal{S}$ with $\mathbb{E}\left[|\mathcal{S}_i|\right] = \frac{m \ln m}{n^{1-1/d}}$ and $\mathbb{E}\left[|E_i|\right] = n^{1+1/d} \ln m$

1. initially, each range and edge has weight $= 1$

2. for $i = 1, \ldots, n/4$
   - sample an edge e_i according to the current edge-weights
   - sample a ball S_i according to the current range-weights
   - double the weight of each ball in $\mathcal{S}_i$ crossing e_i
   - halve the weight of each edge in E_i crossing S_i
   - add e_i to the matching and remove all edges incident to e_i

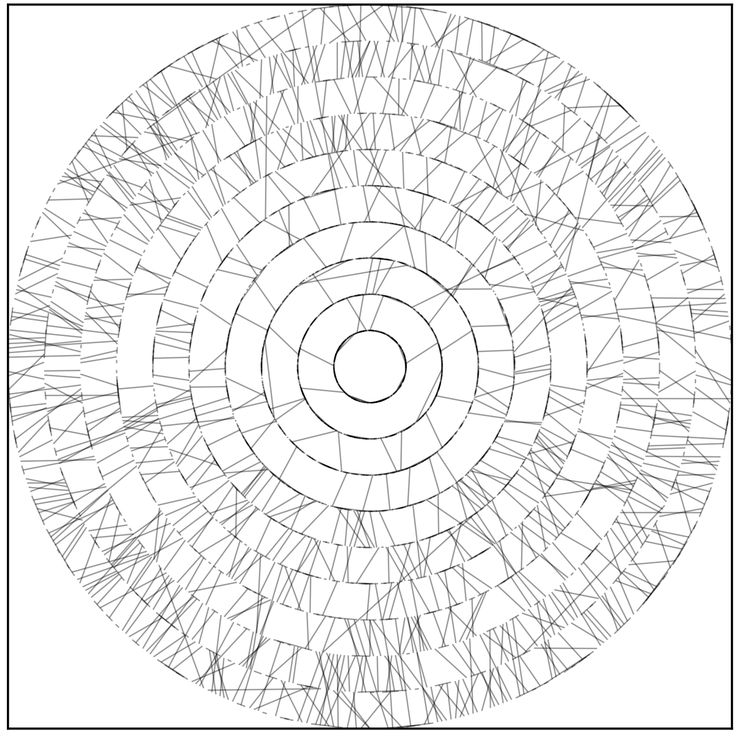
3. recurse on uncovered points

Matching with crossing number $O\left(n^{1-1/d}\right)$ in time $\tilde{O}\left(mn^{1/d} + n^{2+1/d}\right)$

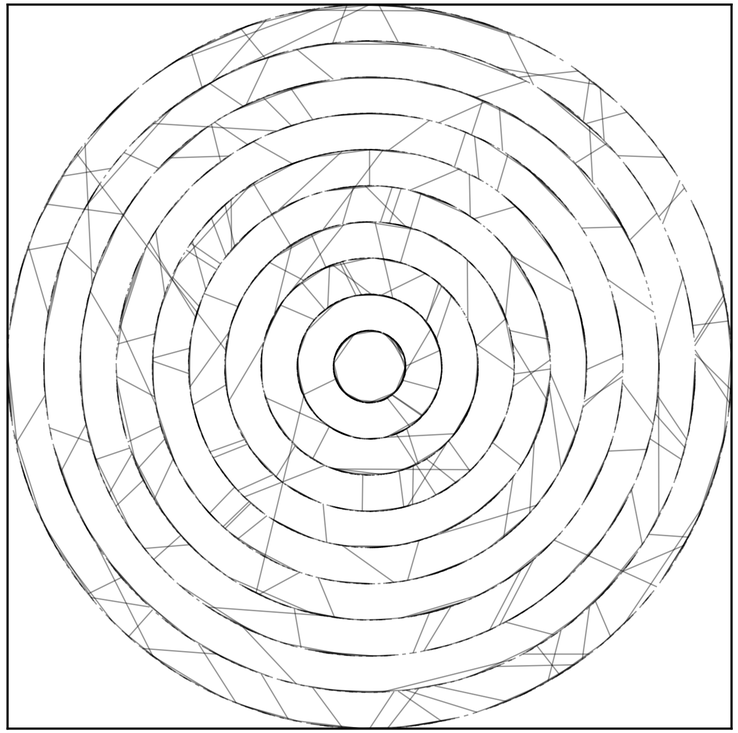
Matchings — $\mathbb{R}^2$

5000 points on 10 co-centric circles

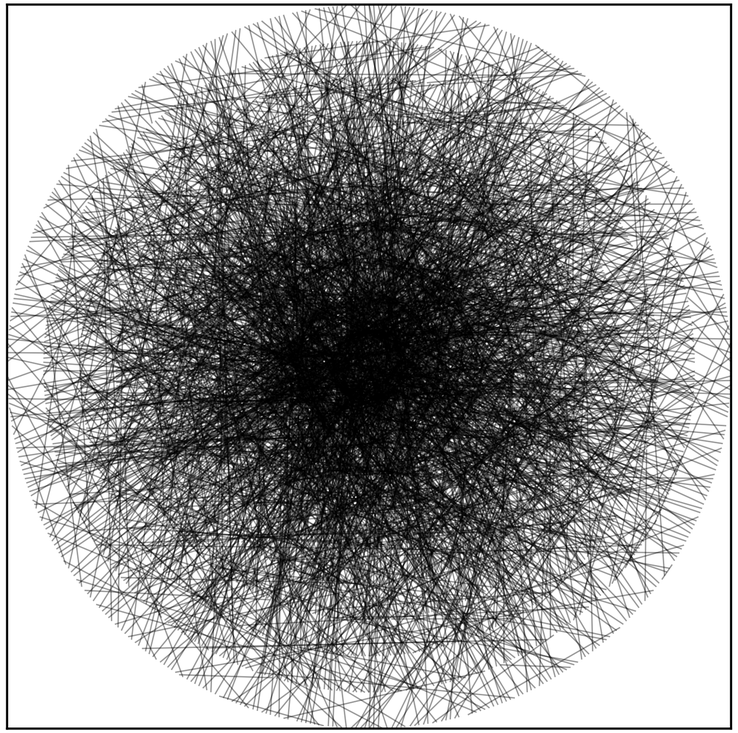
Our method, Ranges: half-planes



Our method, Ranges: disks



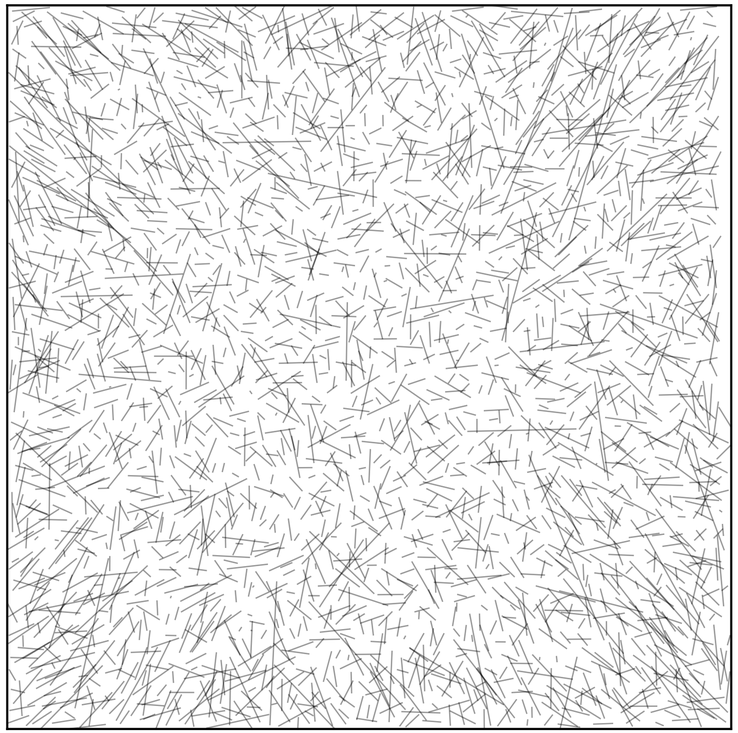
Random Matching



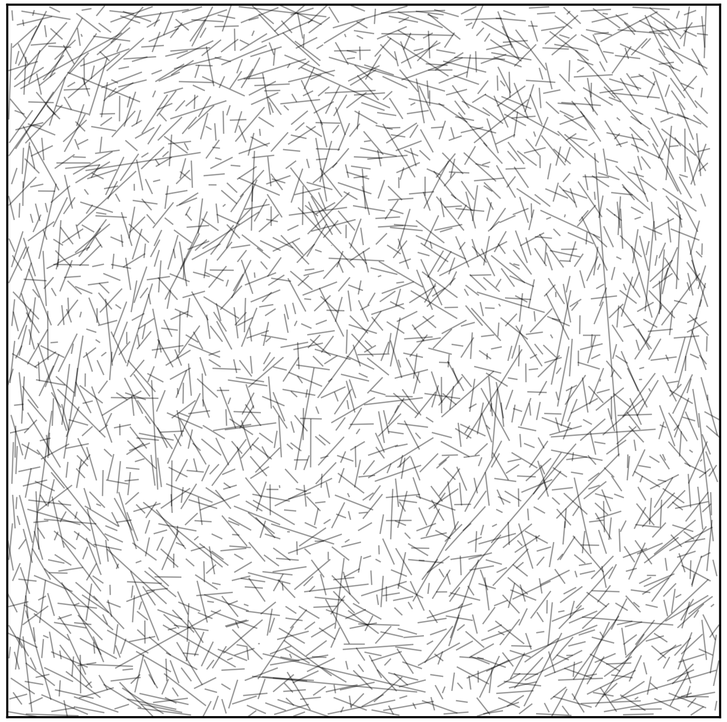
Matchings — $\mathbb{R}^2$

5000 points picked uniformly at random from 5000 grid cells

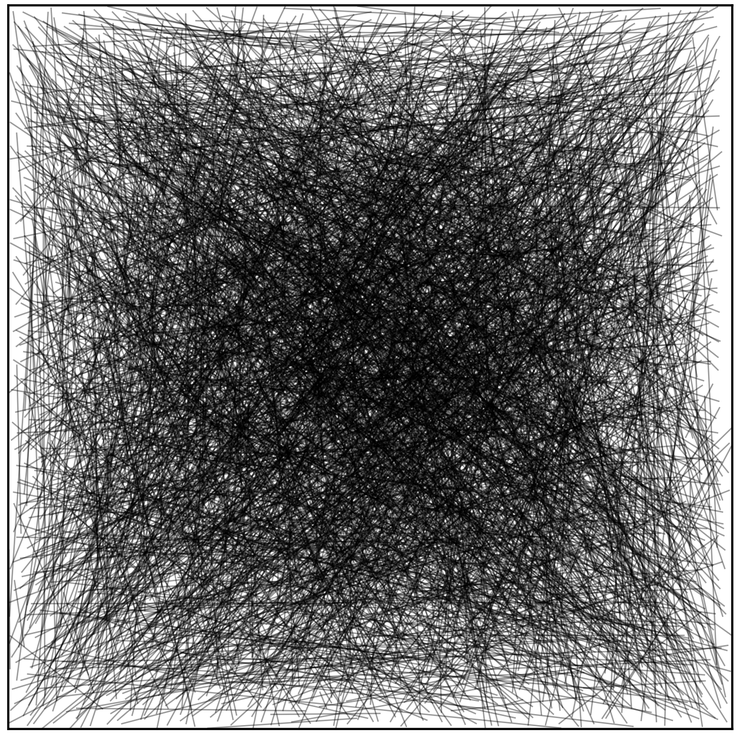
Our method, Ranges: half-planes



Our method, Ranges: disks



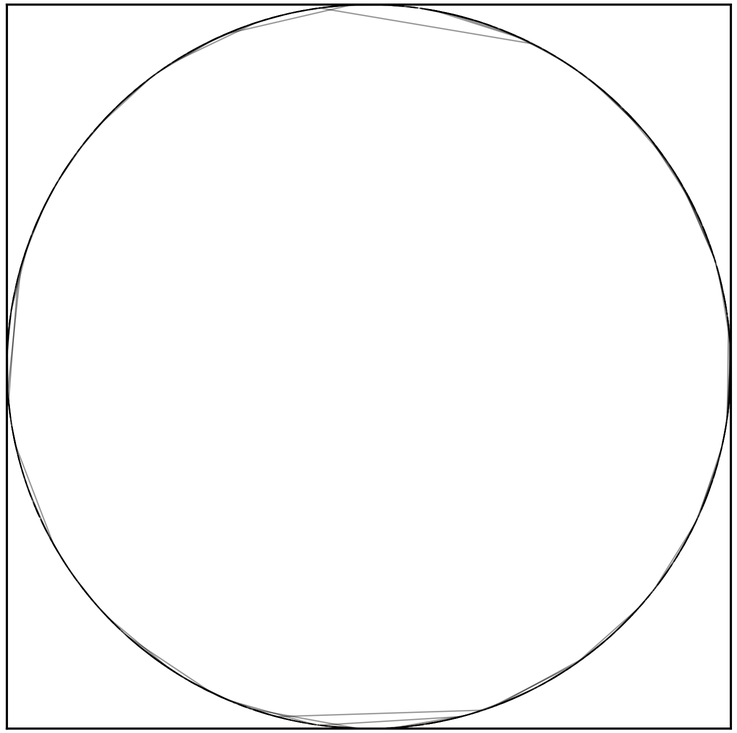
Random Matching



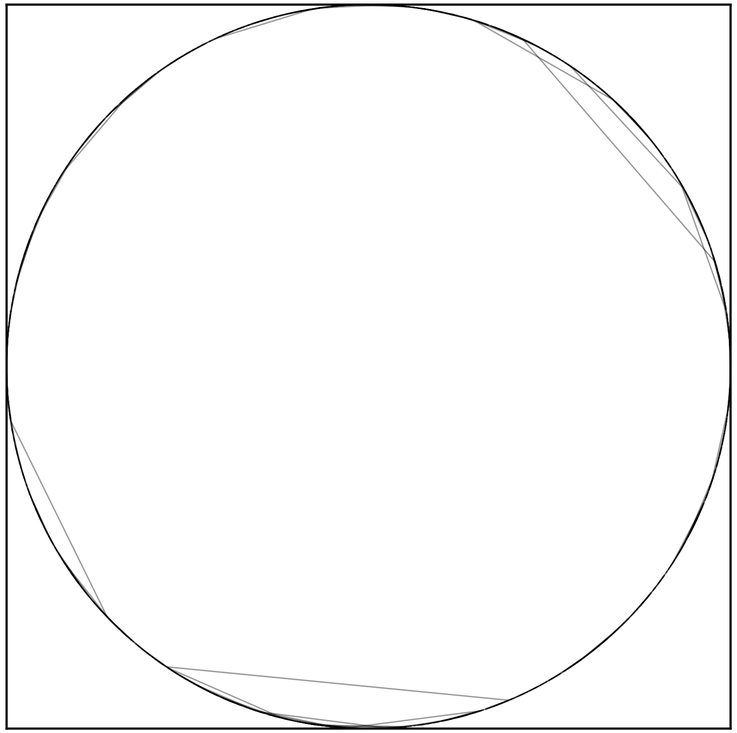
Matchings — $\mathbb{R}^2$

5000 points picked uniformly at random from 5000 grid cells

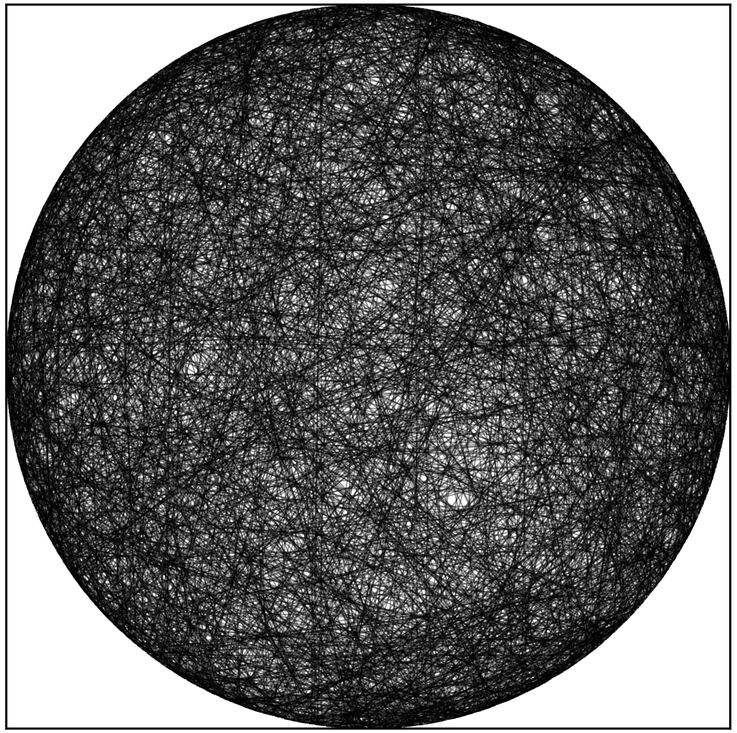
Our method, Ranges: half-planes



Our method, Ranges: disks

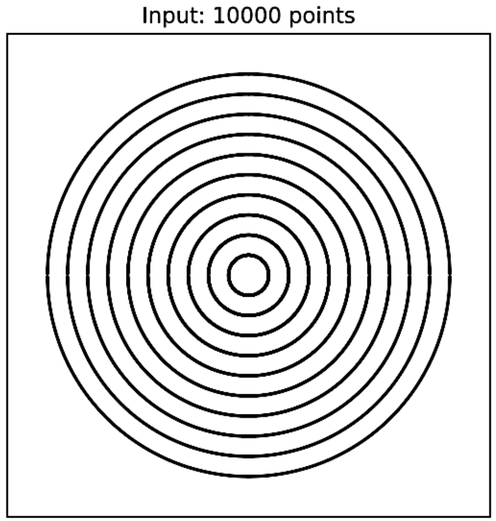


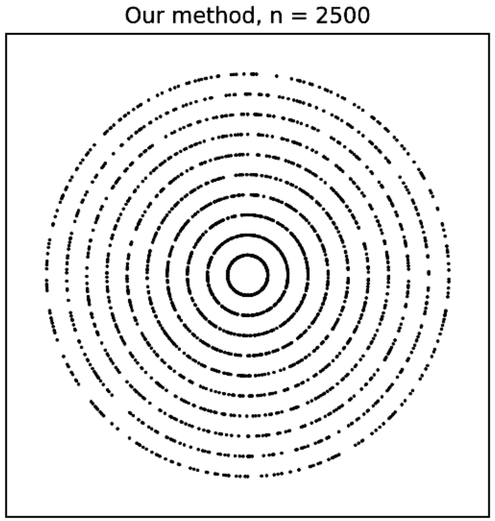
Random Matching

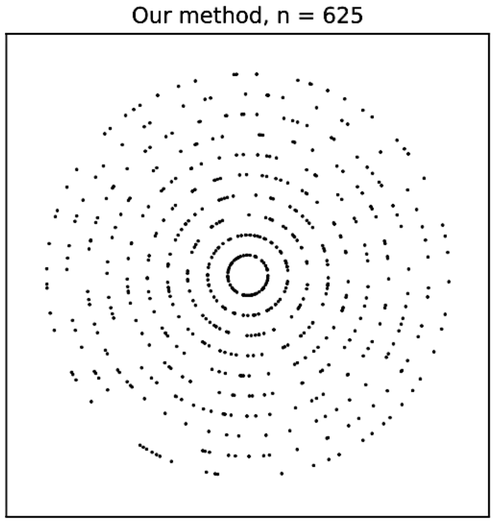


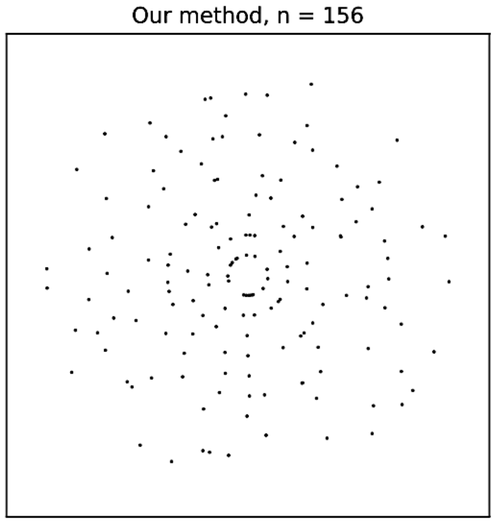
Approximations — $\mathbb{R}^2$

Our method









Random sampling