Small Random Samples via the Crossing Distance

Metric Graph Theory and Related Topics

NABIL H. MUSTAFA

LIPN, USPN

Uniform Approximations of Set Systems

Approximations of Set Systems

n elements $\qquad$ m sets

set system $(\boldsymbol{X}, \boldsymbol{\mathcal{R}})$

compute an approximation $\boldsymbol{A} \subseteq \boldsymbol{X}$

$$S \in \mathcal{R}$$

$$|\boldsymbol{A}| \qquad \mathbf{E}\big[|\boldsymbol{A} \cap \boldsymbol{S}|\big]$$

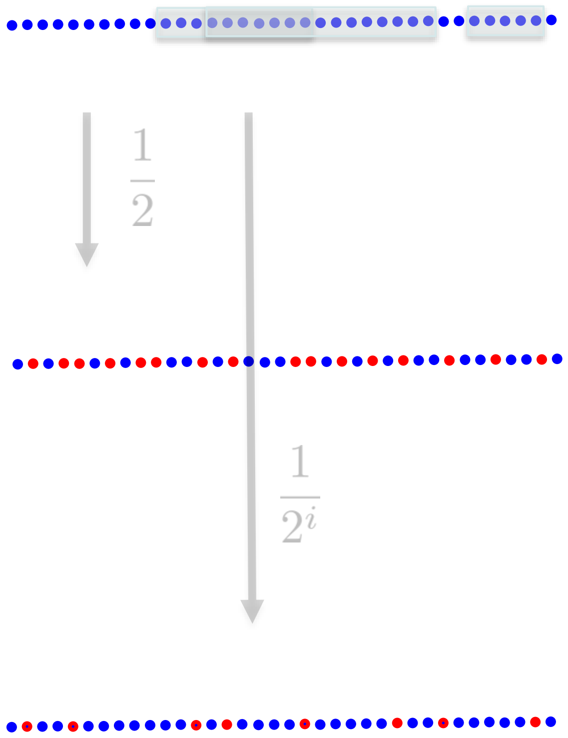
discrepancy

$$\frac{n}{2} \qquad \frac{|S|}{2} \qquad \frac{|S|}{2} \pm \boxed{\texttt{error}}$$

i steps

ϵ-approximation

$$\boxed{\frac{n}{2^i}} \qquad \frac{|S|}{2^i} \qquad \frac{|S|\,|A|}{n} \pm \epsilon\,|A|$$



RANDOM SAMPLING FOR DISCREPANCY

$n \quad m$

set system $(\boldsymbol{X}, \boldsymbol{\mathcal{R}})$

Chernoff-Hoeffding bound:

For any $\Delta > 0$ and $S \in \mathcal{R}$

$$\Pr\left[Y_S \geq \Delta\right] < \exp\left(-\frac{\Delta^2}{2|S|}\right)$$

color each point $\{+1, -1\}$ independently at random

for a fixed $S \in \mathcal{R}$: discrepancy $O\left(\sqrt{|S|}\right)$ with constant probability

probability of discrepancy being at least η for some $S \in \mathcal{R}$

$$< m \cdot \exp\left(-\Theta\left(\frac{\eta^2}{n}\right)\right)$$

$$\leq m \cdot e^{-\ln 2m} \leq \frac{1}{2}$$

discrepancy → $\boldsymbol{\eta = \Theta\left(\sqrt{n \ln m}\right)}$

RANDOM SAMPLING FOR APPROXIMATIONS

set system $\boldsymbol{(\overset{n}{X}, \overset{m}{\mathcal{R}})}$ **discrepancy** $\boldsymbol{O\left(\sqrt{n \ln m}\right)}$

for all $S \in \mathcal{R}$:

$$A_1 \subseteq X$$

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

$$A_2 \subseteq A_1$$

$$|A_2 \cap S| = \frac{|A_1 \cap S|}{2} \pm O\left(\sqrt{\frac{n}{2} \ln m}\right) = \frac{|S|}{4} \pm O\left(\frac{1}{2}\sqrt{n \ln m} + \sqrt{\frac{n}{2} \ln m}\right)$$

after i steps:

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

setting $t = \log \frac{\epsilon^2 n}{\ln m}$ gives

$$|A_t| = O\left(\frac{1}{\epsilon^2} \log m\right) \qquad |A_t \cap S| = \frac{|S|\,|A_t|}{n} \pm \epsilon\,|A_t|$$

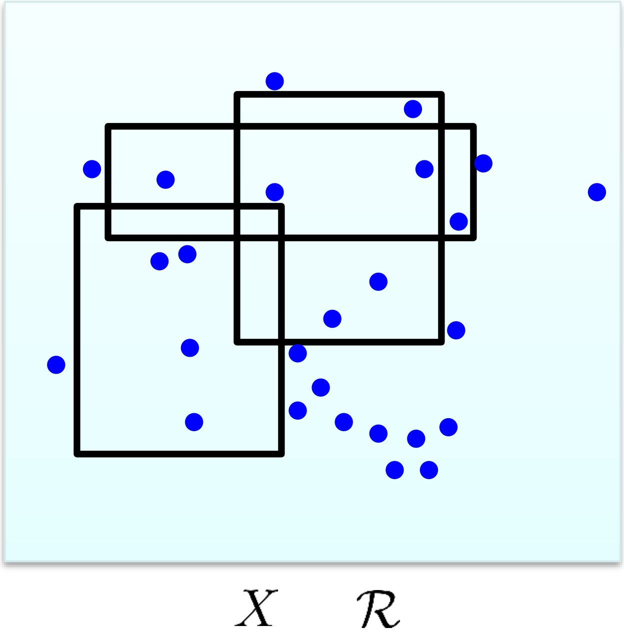
LOCALLY NICE SYSTEMS

Theorem: A uniform random sample of X of size $\Theta\left(\frac{1}{\epsilon^2}\ln m\right)$ is an ϵ-approximation.

'locally polynomial' set systems

total number of sets $|\mathcal{R}|$: $O\left(n^4\right)$

number of subsets on $Y \subseteq X$: $O\left(n^4\right) \rightarrow O\left(|Y|^4\right)$



combinatorially

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

the *projection* of $\mathcal{R}$ onto Y

a constant d such that

$$|\mathcal{R}|_Y| = O\big(|Y|^d\big) \text{ for any } Y \subseteq X$$

[Vapnik-Chervonenkis, 1971]

RANDOM SAMPLING FOR APPROXIMATIONS

set system $\left(\overset{n}{\boldsymbol{X}}, \overset{m = O\left(n^d\right)}{\boldsymbol{\mathcal{R}}}\right)$

discrepancy $\boldsymbol{O\left(\sqrt{n \ln m}\right)}$

for all $S \in \mathcal{R}$:

$$\boxed{\begin{gathered} A_1 \subseteq X \\ |A_1 \cap S| = \frac{|S|}{2} \pm O\left(\sqrt{n \ln \boxed{m}}\right) \boldsymbol{n^d} \end{gathered}}$$

$$\boxed{\begin{gathered} A_2 \subseteq A_1 \\ |A_2 \cap S| = \frac{|A_1 \cap S|}{2} \pm O\left(\sqrt{\frac{n}{2} \ln m}\right) = \frac{|S|}{4} \pm O\left(\frac{1}{2} \overset{\boldsymbol{n^d}}{\sqrt{n \ln \boxed{m}}} + \overset{\boldsymbol{(n/2)^d}}{\sqrt{\frac{n}{2} \ln \boxed{m}}}\right) \end{gathered}}$$

after i steps: $|A_i \cap S| = \frac{|S|}{2^i} \pm O\left(\sqrt{\frac{n}{2^i} \ln \boxed{m}}\right) \boldsymbol{\left(n/2^i\right)^d}$

setting $t = \log \frac{\epsilon^2 n}{\ln \boxed{m}}$ gives

$$|A_t| = O\left(\frac{1}{\epsilon^2} \log \boxed{m}\right)^{\boldsymbol{\left(\frac{1}{\epsilon}\right)^d}}$$

$$\boldsymbol{\left(\frac{1}{\epsilon}\right)^d}$$

$$|A_t \cap S| = \frac{|S|\,|A_t|}{n} \pm \epsilon\,|A_t|$$

[Vapnik-Chervonenkis, 1971]

APPROXIMATION BOUNDS

n elements
m subsets
d dimension

	Random Arbitrary	VC dimension Uniform Sampling
Discrepancy	$\sqrt{n \ln m}$	$\sqrt{dn \ln n}$
Approximations	$\frac{1}{\epsilon^2} \ln m$	$\frac{d}{\epsilon^2} \ln \frac{1}{\epsilon}$ $\boxed{\rightarrow \frac{d}{\epsilon^2}}$

[Talagrand, 1994]

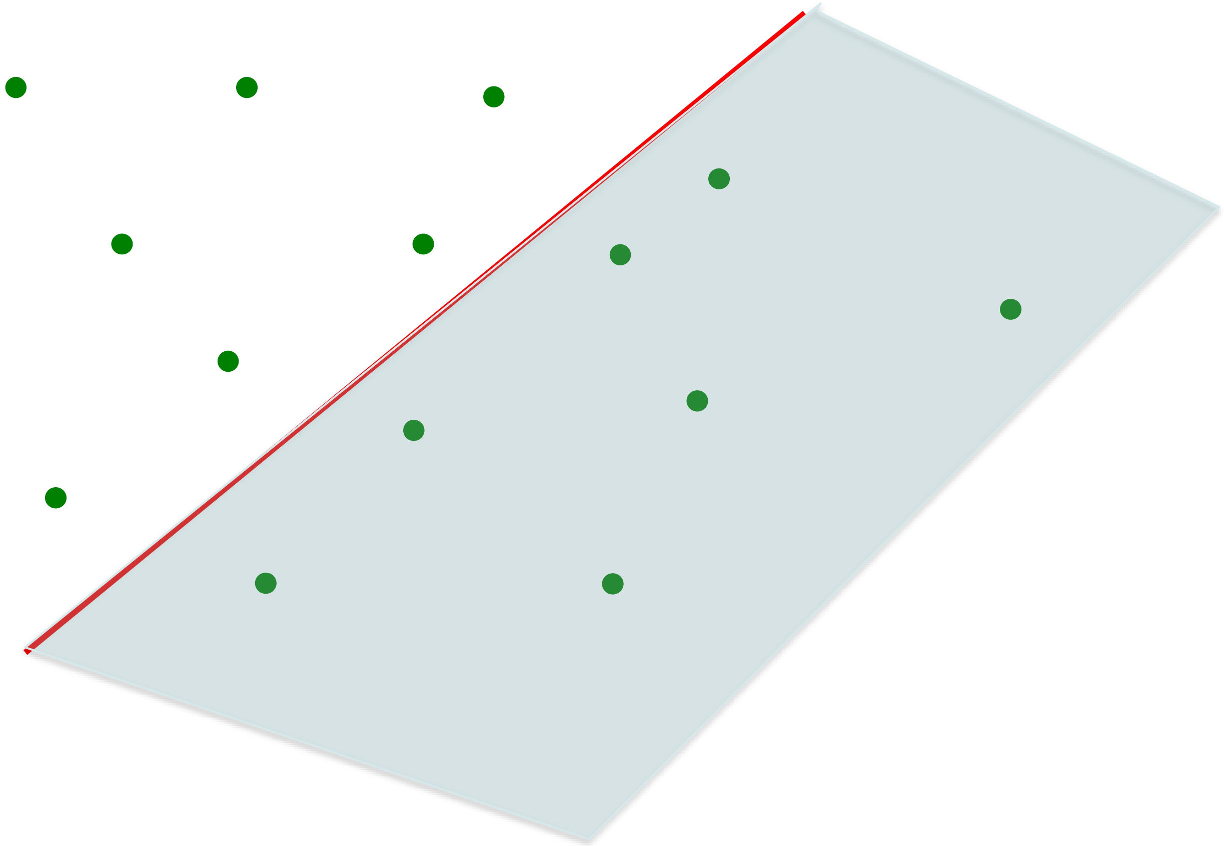
[Li, Long, Srinivasan, 2001]

usefulness: **oblivious uniform approximation**

Non-Uniform Approximations

SET SYSTEM INDUCED BY HALF-SPACES

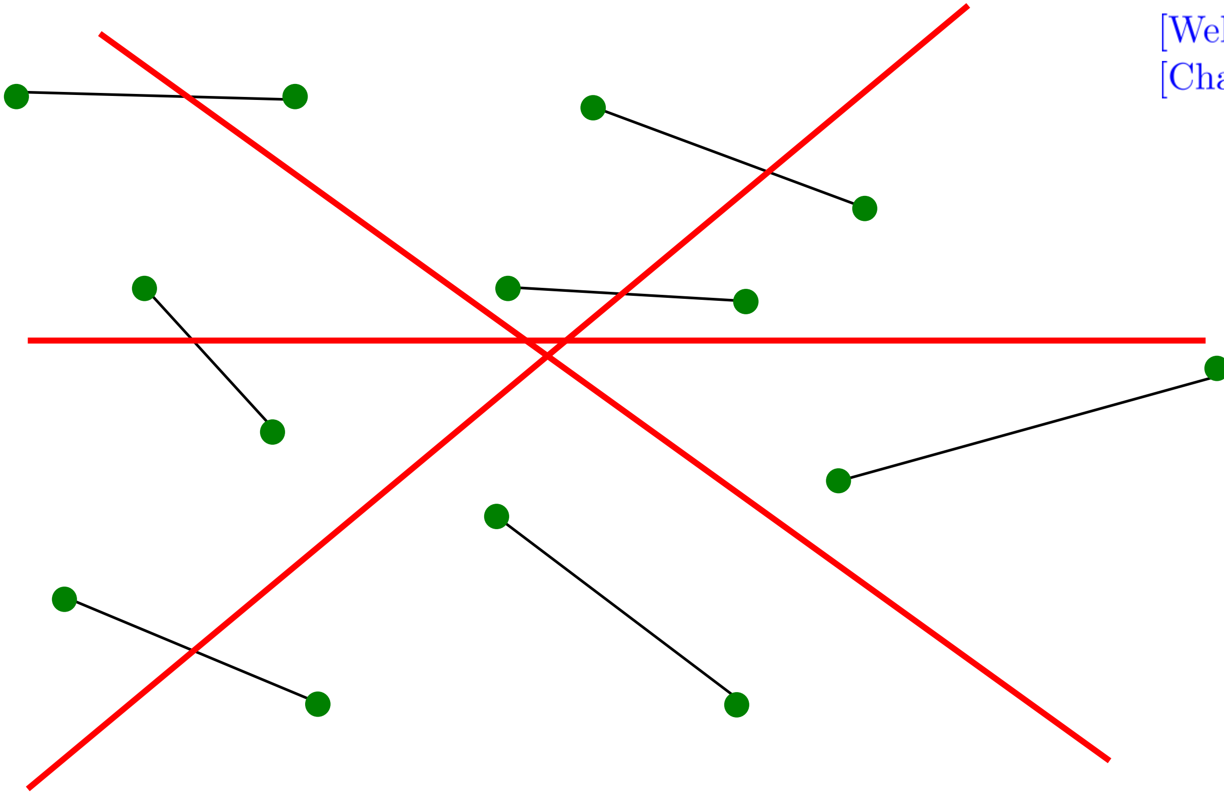
set system induced by half-spaces in $\mathbb{R}^d$



$O\left(\sqrt{n \ln \boxed{n^2}}\right) \rightarrow O\left(n^{1/4}\right)$ **optimal**

SPANNING TREES WITH LOW CROSSING NUMBER

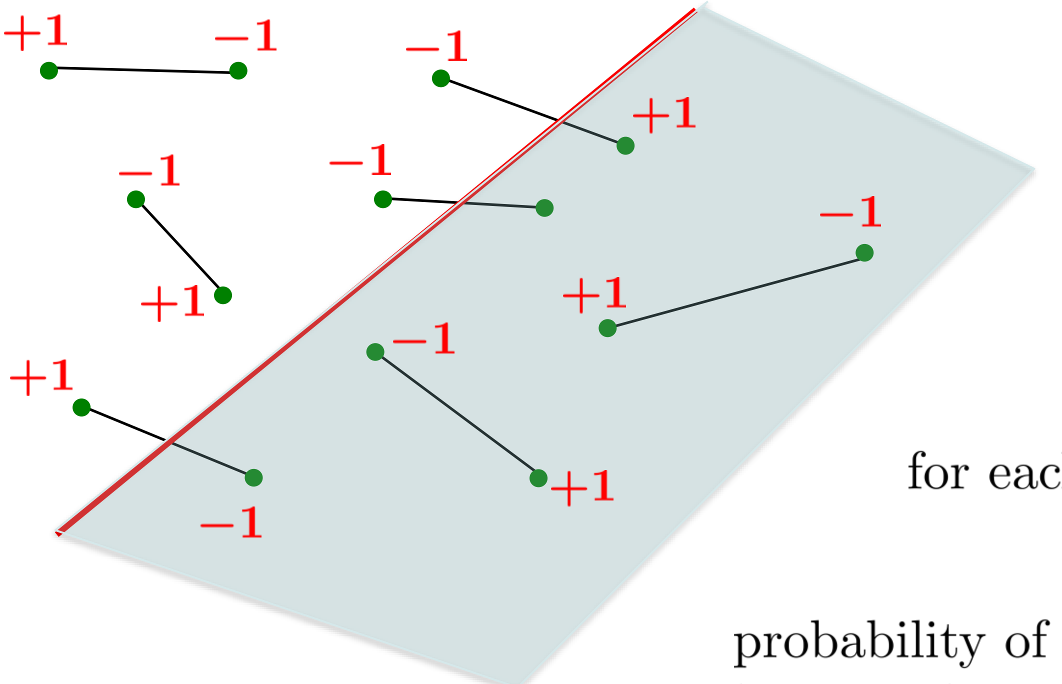
Theorem: Given any set P of n points in $\mathbb{R}^2$, there exists a matching M on P such that any line crosses at most $O(\sqrt{n})$ edges of M.



[Welzl 1988]
[Chazelle-Welzl 1989]

> One almost wouldn't believe that after thousands of years of geometry, it is still possible to discover such pretty theorems about points in the plane.
>
> `J. Matoušek, ICM 1998`

SPANNING TREES WITH LOW CROSSING NUMBER



Chernoff-Hoeffding bound:

For any $\Delta > 0$ and $S \in \mathcal{R}$

$$\Pr\left[Y \geq \Delta\right] < \exp\left(-\frac{\Delta^2}{2|S|}\right)$$

for each half-space, $O\left(\sqrt{n}\right)$ events

probability of discrepancy being at least η

$$\leq m \cdot \exp\left(-\Theta\left(\frac{\eta^2}{\sqrt{n}}\right)\right)$$

$$\leq m \cdot e^{-\ln 2m} \leq \frac{1}{2}$$

$$\eta = \Theta\left(n^{1/4}\sqrt{\ln m}\right)$$

[Matoušek-Welzl-Wernisch 1993]

APPROXIMATION BOUNDS

n elements
m subsets
d dimension

	Random	VC dimension	
		Uniform Sampling	Non-Uniform Sampling + Combinatorics
Discrepancy	$\sqrt{n \ln m}$	$\sqrt{dn \ln n}$	$\boxed{n^{\frac{1}{2}-\frac{1}{2d}}}$
Approximations	$\frac{1}{\epsilon^2} \ln m$	$\frac{d}{\epsilon^2}$	$\boxed{\frac{1}{\epsilon^{2-\frac{2}{d+1}}}}$

Problem: now need to construct matchings to be able to sample

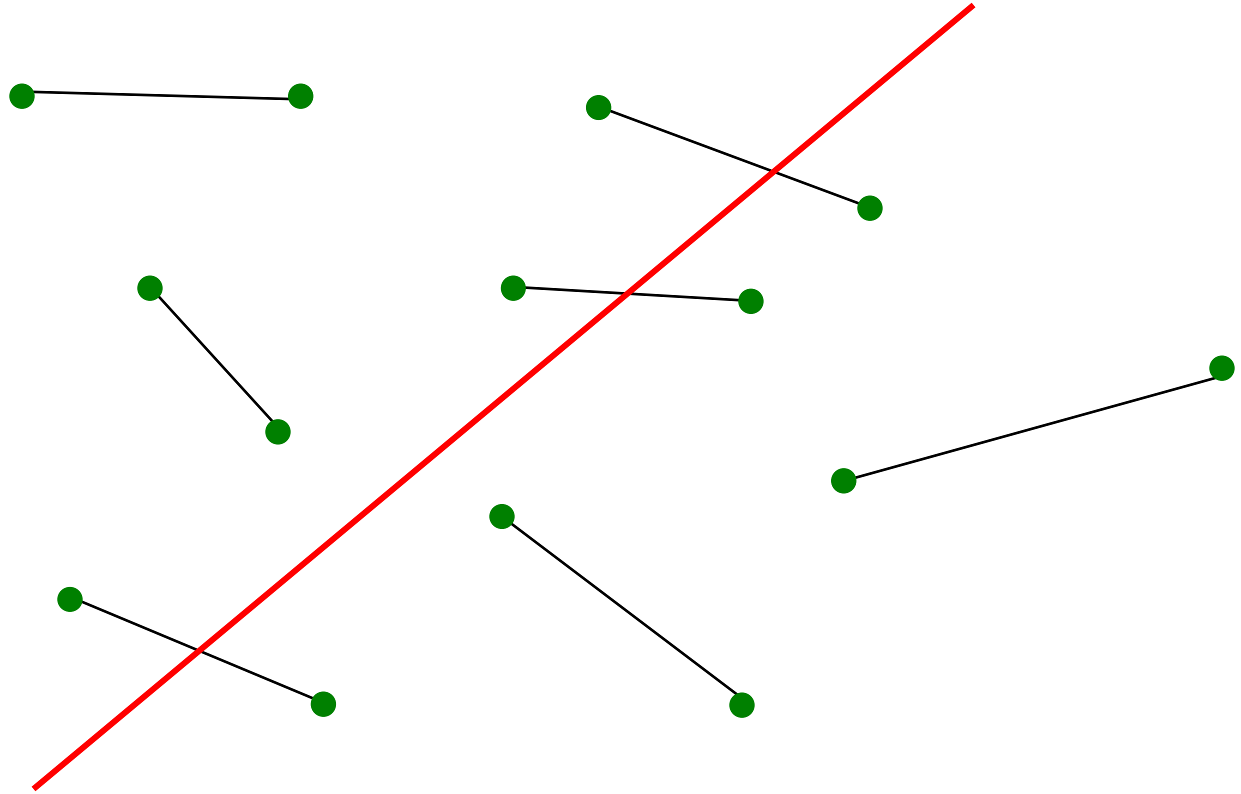
faster constructions imply improved algorithms for constructing small samples

LOW CROSSING MATCHINGS

SPANNING TREES WITH LOW CROSSING NUMBER

Theorem: Given any set P of n points in $\mathbb{R}^2$, there exists a matching M on P such that any line crosses at most $O(\sqrt{n})$ edges of M.

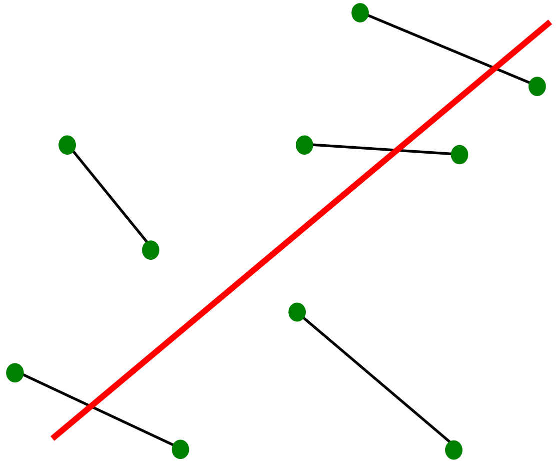
[Welzl 1988]
[Chazelle-Welzl 1989]



More generally: construct matchings for set systems, with a natural generalized notion of crossing

SPANNING TREES WITH LOW CROSSING NUMBER

Given points P, and lines $\mathcal{L}$, let M be a tree on P such that any line in $\mathcal{L}$ crosses at most $O(\sqrt{n})$ edges of M.



$$|P| = n \qquad |\mathcal{L}| = m$$

each line $l \in \mathcal{L}$ crosses at most $\sqrt{n}$ edges of M

summed over all $l \in \mathcal{L}$, there are $m \cdot \sqrt{n}$ crossings between lines and edges

one edge of M must be crossed by at most $\frac{m \cdot \sqrt{n}}{n-1} = O\left(\frac{m}{\sqrt{n}}\right)$ lines in $\mathcal{L}$

Key Property: Given P and $\mathcal{L}$, there exist two points $p_i, p_j \in P$ such that the line segment $\overline{p_i p_j}$ crossed by $O\left(\frac{m}{\sqrt{n}}\right)$ lines in $\mathcal{L}$.

SPANNING TREES WITH LOW CROSSING NUMBER

Key Property: Given X and $\mathcal{S}$, there exist two elements $p_i, p_j \in X$ such that the pair $\{p_i, p_j\}$ is crossed by $\leq \frac{m}{\sqrt{n}}$ sets in $\mathcal{S}$.

$X_p \subseteq \mathcal{S}$: sets of $\mathcal{S}$ containing p

need to show: two sets X_p, X_q with symmetric difference $O\left(\frac{m}{\sqrt{n}}\right)$

if symmetric difference between every pair is $\Omega\left(\Delta\right)$

Packing lemma: number of sets is then $O\left(\left(\frac{m}{\Delta}\right)^2\right)$

[Haussler 1995]

[Ezra 2014]

[M. 2016]

$$O\left(\left(\frac{m}{\Delta}\right)^2\right) < n \qquad \rightarrow \qquad \Delta > \frac{m}{\sqrt{n}}$$

key is structure of the k-distance graph on $(X, \mathcal{S})$

$k = 1$ [Haussler-Littlestone-Warmuth 1990]

$k = 2$ [Chepoi-Labourel-Ratel 2019]

much better understood for partial cubes [Chepoi-Knauer-Philibert 2020]

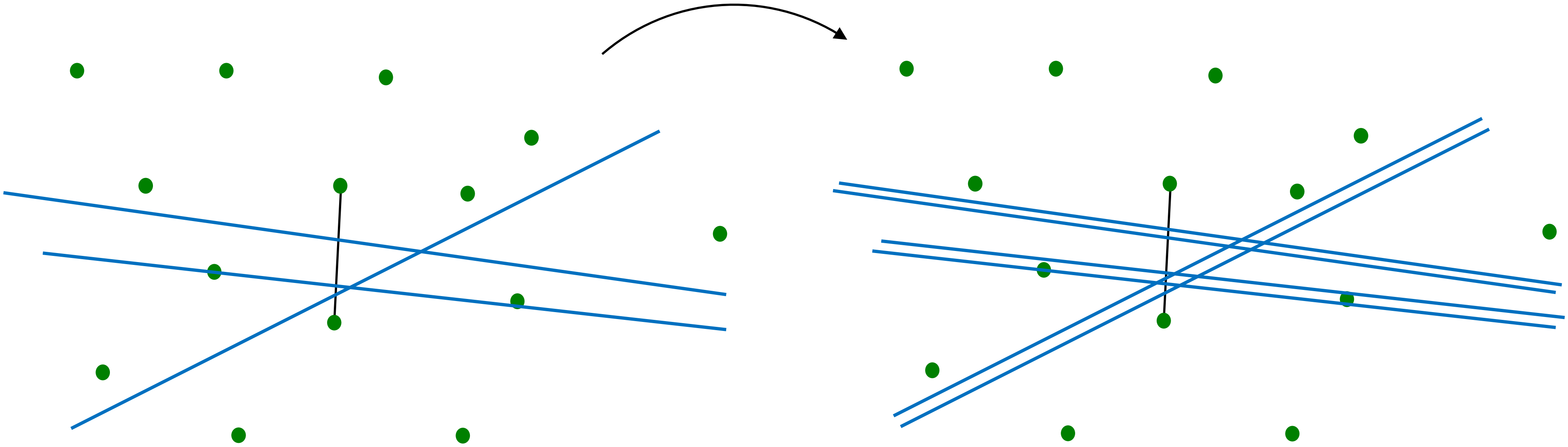
SPANNING TREES WITH LOW CROSSING NUMBER

first idea: construct tree iteratively by adding these edges one by one

given n points and $\mathcal{L}$, find an edge intersecting $\leq m/\sqrt{n}$ lines in $\mathcal{L}$

add to M, remove both endpoints, and iterate on $n-2$ remaining points

second idea: want to discourage a line from intersecting too many edges



at each step, add an edge e, and double the lines that intersect e

iterate on the remaining $n-2$ points with the new larger set of lines $\mathcal{L}'$

very surprising that such a heuristic-y argument works

SPANNING TREES WITH LOW CROSSING NUMBER

Why does it work?

at each step, find an edge intersecting at most $m/\sqrt{n}$ lines

so number of new lines added to $\mathcal{L}$ are at most $m/\sqrt{n}$

at the end, the total number of lines added is 'small'

Now take any line $l \in \mathcal{L}$

each time it intersects an edge, it's weight is doubled

if it intersects k edges, total weight in the end: 2^k

but 2^k is upper-bounded by the total weight, which is small

so k cannot be too large!

SPANNING TREES WITH LOW CROSSING NUMBER

[Welzl 1987]
[Chazelle-Welzl 1989]

	Points	**Lines**	**Edge Intersects**
Step 1	n	m	$m/\sqrt{n}$
Step 2	$n-2$	$m(1+1/\sqrt{n})$	$\frac{m(1+1/\sqrt{n})}{\sqrt{n-2}}$
Step 3	$n-4$	$m(1+1/\sqrt{n})(1+1/\sqrt{n-2})$	$\frac{m(1+1/\sqrt{n})(1+1/\sqrt{n-2})}{\sqrt{n-4}}$

total weight of lines after $n/2$ steps: $m\prod_{i=0}^{n/2}\left(1+1/\sqrt{n-2i}\right)$

if any line intersects k edges, it has weight $\boxed{2^k \leq m\prod_{i=0}^{n/2}\left(1+1/\sqrt{n-2i}\right)}$

solving this gives $k = O(\sqrt{n})$

Problem: how to efficiently find the edge to add at each iteration

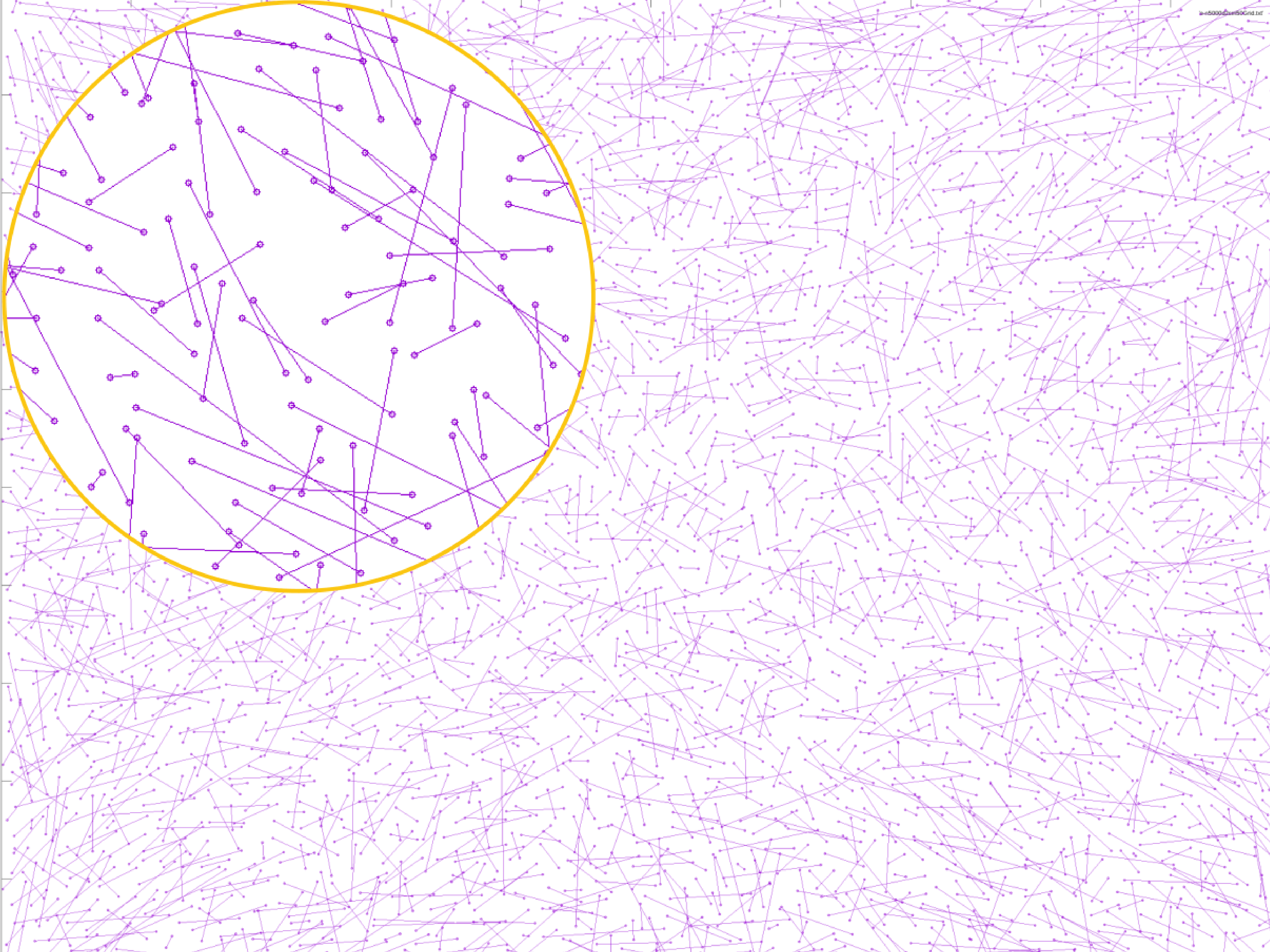
New Algorithm

SPANNING TREES WITH LOW CROSSING NUMBER

[Erdős-Selfrige 1973] [Grigoriadis-Kachiyan 1995] [Koufogiannakis-Young 2014] [Agarwal-Pan 2016]

Algorithm:

```
  set $\omega_1(e) = 1$ for all $e \in E$
  set $\pi_1(S) = 1$ for all $S \in \mathcal{S}$

For $i = 1, \ldots, \frac{n}{4}$:

    $e_i$: sampled from distribution induced by $\omega_i$        $\Pr[e_i = e] = \frac{\omega_i(e)}{\omega_i(E)}$

    $S_i$: sampled from distribution induced by $\pi_i$        $\Pr[S_i = S] = \frac{\pi_i(S)}{\pi_i(\mathcal{S})}$

    double weight of each $S \in \mathcal{S}$ crossing $e_i$

            for $S \in \mathcal{S}$:    $\pi_{i+1}(S) = \pi_i(S)\left(1 + I(e_i, S)\right)$

    halve weight of each $e \in E$ crossing $S_i$

            for $e \in E$:    $\omega_{i+1}(e) = \omega_i(e)\left(1 - \frac{1}{2} I(e, S_i)\right)$

    add $e_i$ to matching

    set weights of all edges incident to the two endpoints of $e_i$ to 0
```



SPANNING TREES WITH LOW CROSSING NUMBER

running time too slow!

[Csikos-M. 2021]

above algorithm improved using structural properties of crossing distance:

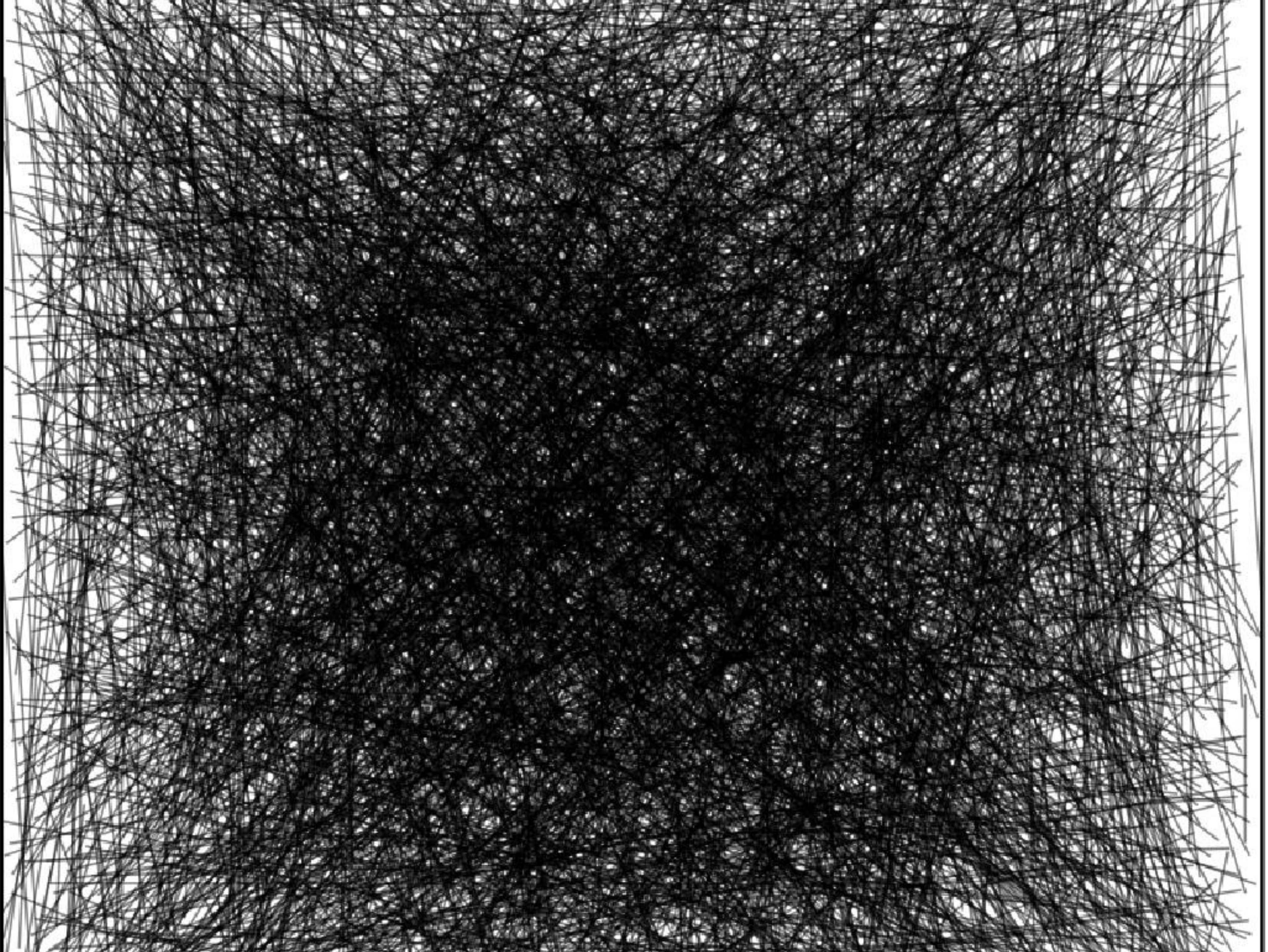
pigeonholing gave a bound on ‘short edge’

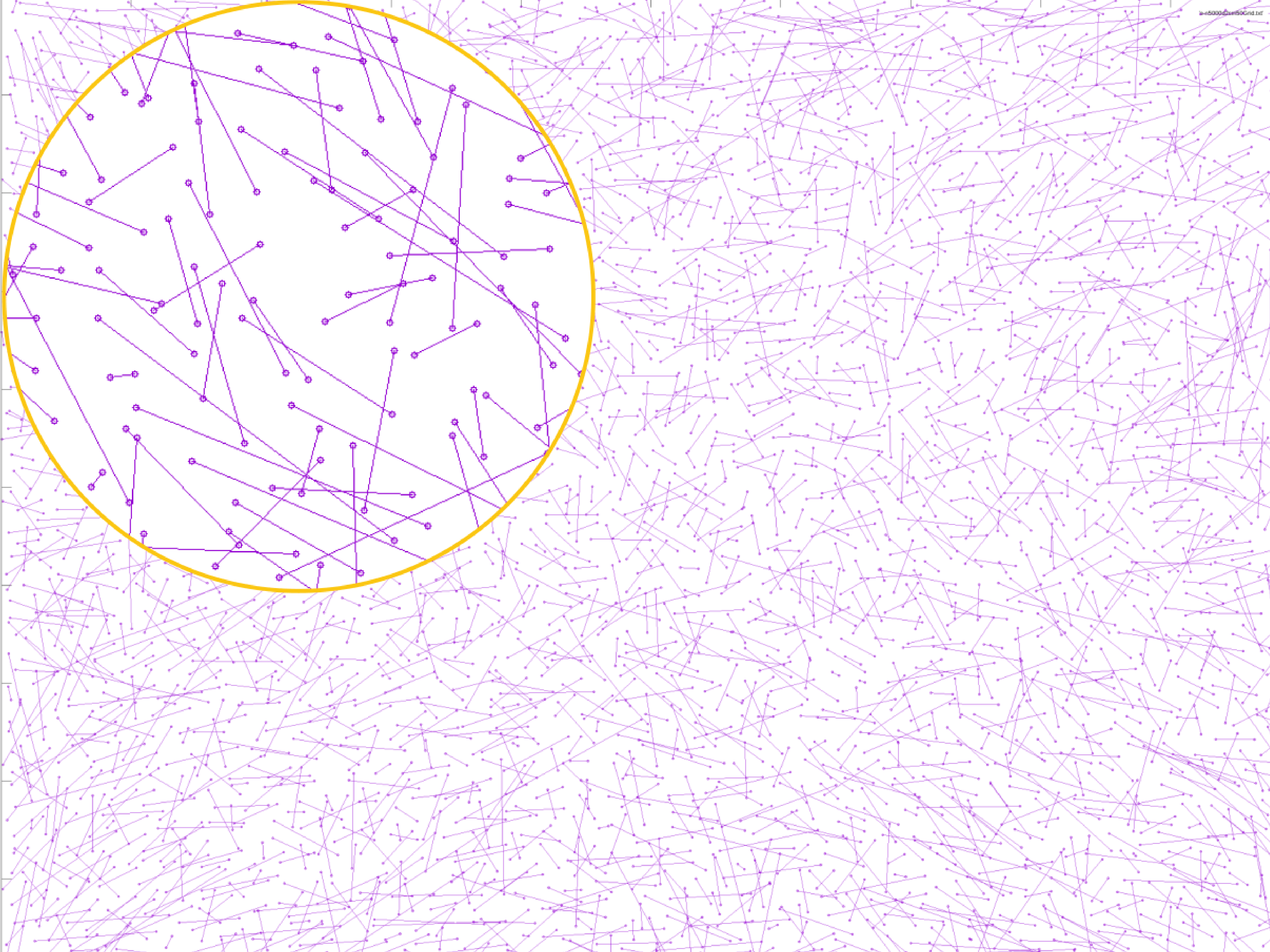
in fact, *many* short edges → existence of a sparse set of edges

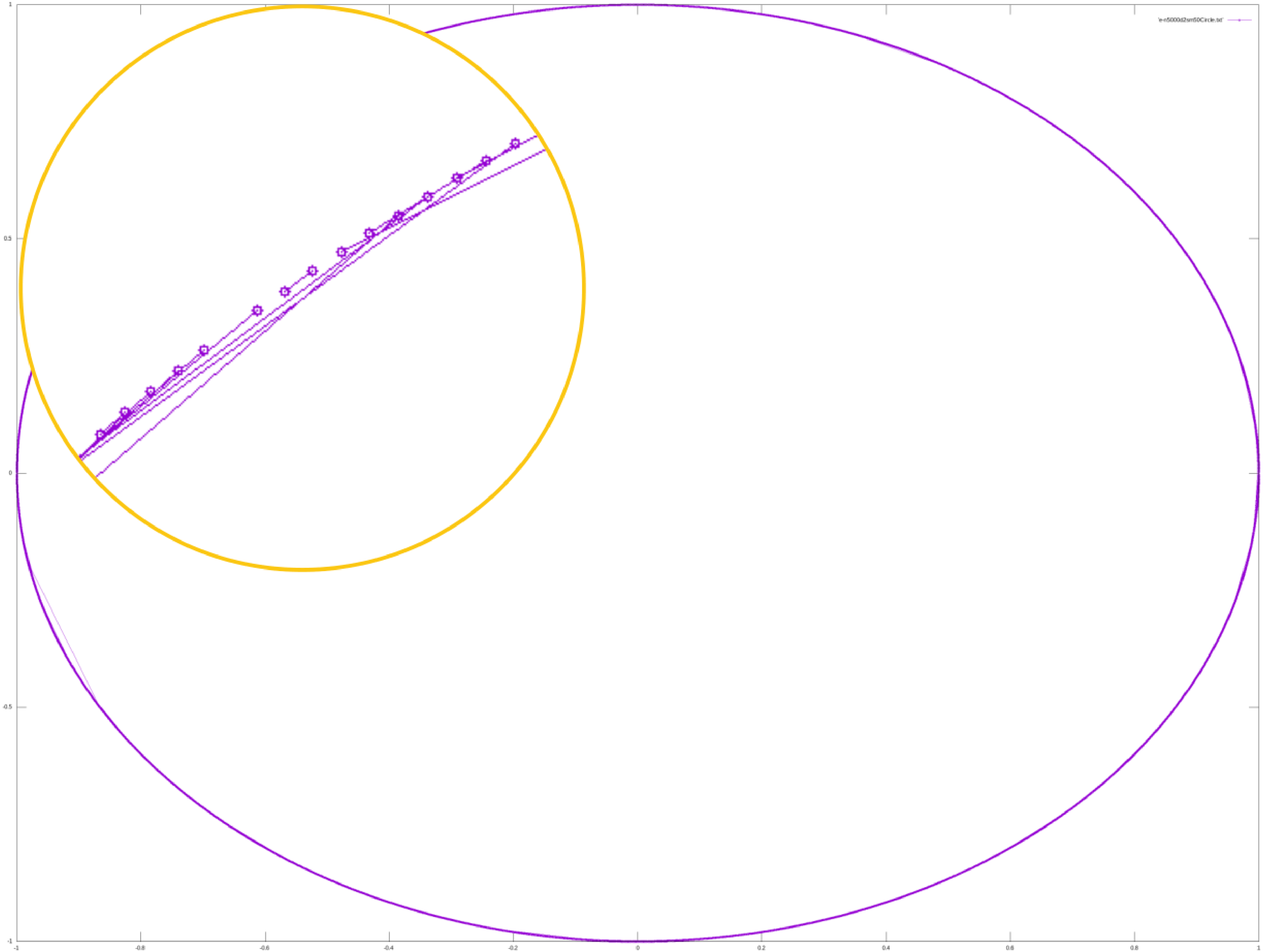
also, able to show that update only a random sample at each step

more involved probabilistic analysis using martingales

simple, easy implementation



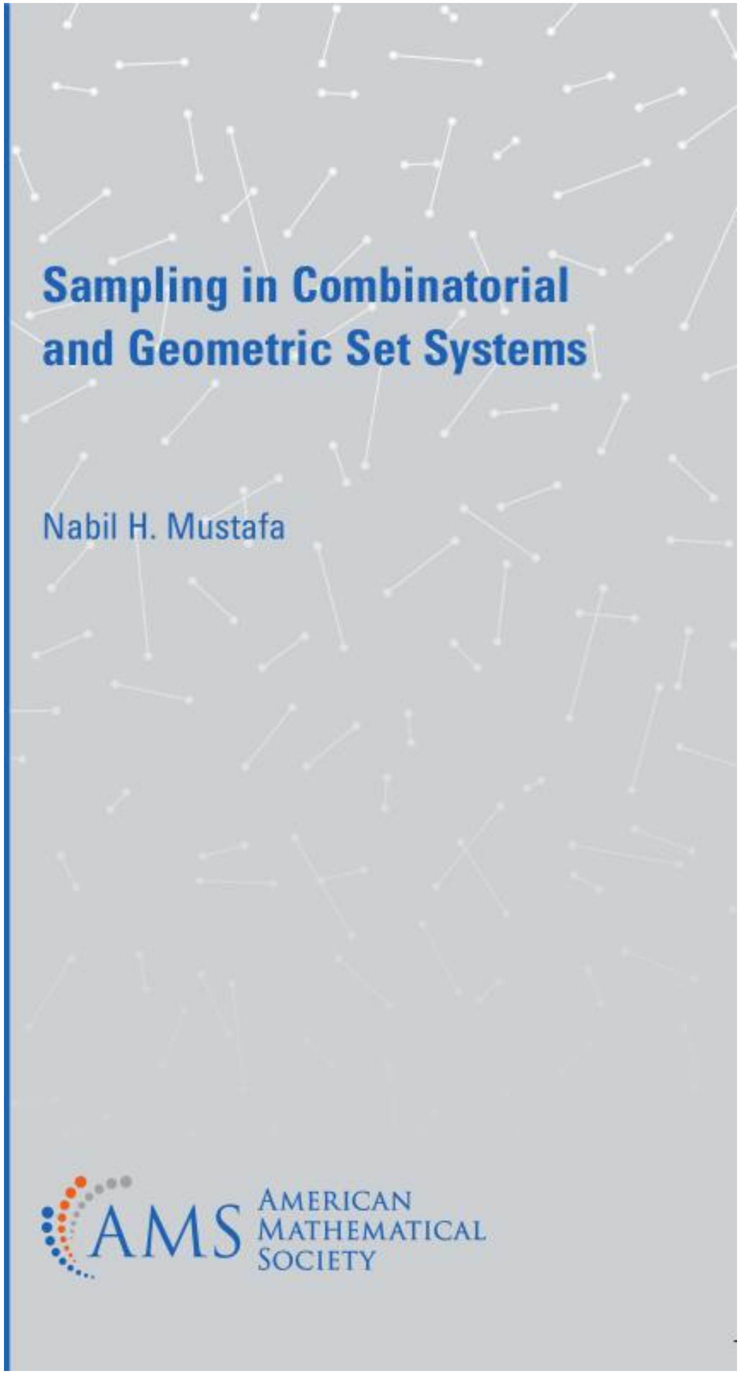



'e-n5000d2sm50Circle.txt'
1
0.5
0
-0.5
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

[Csikos-M. 2021]

	Our crossing number	**Our time**	**Previous crossing number**	**Previous time**
Arbitrary with $\pi^*_{\mathcal{S}}(k) \leq ck^d$	$\left(\frac{5c^{1/d}d}{d-1}+o(1)\right)n^{1-\frac{1}{d}}$	$\tilde{O}\left(mn^{2/d}+n^{2+2/d}\right)$	$O\left(n^{1-1/d}\right)$	$\tilde{O}\left(mn^2\right)$ [Har-Peled 2009] [Chekuri-Quanrud 2018]
Geometric induced by balls in $\mathbb{R}^d$	$\left(6d^2+o(d^2)\right)n^{1-\frac{1}{d}}$	$\tilde{O}\left(dn^{2+2/d}\right)$	$O\left(n^{1-1/d}\right)$	$\tilde{O}\left(n^{3+1/d}\right)$ [Har-Peled 2009] [Chekuri-Quanrud 2018]
Geometric induced by semi-alg. ranges in $\mathbb{R}^d$ (s eq.'s of deg Δ)	$\left(\frac{20e\Delta sd}{d-1}+o(1)\right)n^{1-\frac{1}{d}}$	$\tilde{O}\left(s\Delta^d\left(mn^{2/d}+n^{2+2/d}\right)\right)$	$O\left(10^d s\Delta n^{1-1/d}\right)$	$O\left(n^{O(d^3)}\right)$ [Agarwal-Matousek-Sharir 2013]
Geometric induced by half-spaces in $\mathbb{R}^d$	$\left(6d^2+o(d^2)\right)n^{1-\frac{1}{d}}$	$\tilde{O}\left(dn^{2+2/d}\right)$	$\geq 264d^4 \cdot n^{1-1/d}$	$O\left(d^2n\right)$ [Chan 2012]

Thank you



Sampling in Combinatorial and Geometric Set Systems

Nabil H. Mustafa

AMS American Mathematical Society