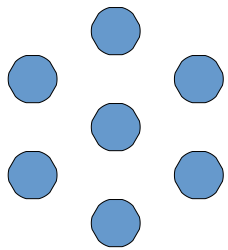
Computer search in pure mathematics

Dániel Varga

Rényi Institute



Structure of the talk

- Introducing the work of the Rényi AI research group in 10 minutes.
- Presenting *"Piercing the chessboard"* by Gergely Ambrus, Imre Bárány, Péter Frankl, DV. SIDMA 2023.
- Giving a rough overview of *"The density of planar sets avoiding unit distances"* by Gergely Ambrus, Adrián Csiszárik, Máté Matolcsi, DV, Pál Zsámboki. 2022. Under review.

Agenda of the talk

- During the presentation of the two geometry results, I will take every opportunity to highlight situations where we were aided by computer search and simulations.
- I am giving away the punchline: The talk will conclude with a call for collaboration: visit us if you feel that such an approach could contribute to the resolution to your favourite open problems.

Introducing the Rényi AI research group

- 12-16 people, depending on how we count part timers.
- Jakab Buda, Adrián Csiszárik, Domonkos Czifra, Botond Forrai, Diego González-Sánchez, Melinda Kiss, Péter Kőrösi-Szabó, Gábor Kovács, Ákos Matszangosz, Márton Muntag, Gergely Papp, Balázs Szegedy, Dávid Terjék, Dániel Varga, Zsolt Zombori, Pál Zsámboki.
- Hiring made possible by large 5 years long grant MILAB, but also other grants.
- We are very grateful to Dezső Miklós and Márta Szomolányi for their project management efforts.
- Most of us work on fundamental research published at theory-focused AI conferences. More than half of us work on instantly utilized applied research.

Artificial neural network

Nowadays, by artificial neural network we mean a huge formula $f_\Theta : \mathbb{R}^n \to \mathbb{R}^m$ defining a continuous function that is almost everywhere differentiable with respect to the parameters Θ.
For example,

$$f_\Theta(x) = W_2 \max(W_1 x + b_1, 0) + b_2,$$

where $\Theta = (W_1, b_1, W_2, b_2)$,

and $x \in \mathbb{R}^n, W_1 \in \mathbb{R}^{k \times n}, b_1 \in \mathbb{R}^k, W_2 \in \mathbb{R}^{m \times k}, b_2 \in \mathbb{R}^m$.

Gradient descent

$$f_\Theta : \mathbb{R}^n \to \mathbb{R}^m$$

$$\Theta_{i+1} := \Theta_i - \lambda \nabla_\Theta \left(\frac{1}{2} \| f_{\Theta_i}(x) - y \|_2^2 \right)$$

where

λ is the learning rate,

(x, y) is an input-output pair.

The idea is that after presenting many input-output pairs, the model f_{Θ_T} does not only memorize them, but also learn to *generalize* to unseen values of x.

A more elaborate example powering ChatGPT

```python
import numpy as np

def gelu(x):
    return 0.5 * x * (1 + np.tanh(np.sqrt(2 / np.pi) * (x + 0.044715 * x**3)))

def softmax(x):
    exp_x = np.exp(x - np.max(x, axis=-1, keepdims=True))
    return exp_x / np.sum(exp_x, axis=-1, keepdims=True)

def layer_norm(x, g, b, eps: float = 1e-5):
    mean = np.mean(x, axis=-1, keepdims=True)
    variance = np.var(x, axis=-1, keepdims=True)
    return g * (x - mean) / np.sqrt(variance + eps) + b

def linear(x, w, b):
    return x @ w + b

def ffn(x, c_fc, c_proj):
    return linear(gelu(linear(x, **c_fc)), **c_proj)

def attention(q, k, v, mask):
    return softmax(q @ k.T / np.sqrt(q.shape[-1]) + mask) @ v

def mha(x, c_attn, c_proj, n_head):
    x = linear(x, **c_attn)
    qkv_heads = list(map(lambda x: np.split(x, n_head, axis=-1), np.split(x, 3, axis=-1)))
    causal_mask = (1 - np.tri(x.shape[0], dtype=x.dtype)) * -1e10
    out_heads = [attention(q, k, v, causal_mask) for q, k, v in zip(*qkv_heads)]
    x = linear(np.hstack(out_heads), **c_proj)
    return x

def transformer_block(x, mlp, attn, ln_1, ln_2, n_head):
    x = x + mha(layer_norm(x, **ln_1), **attn, n_head=n_head)
    x = x + ffn(layer_norm(x, **ln_2), **mlp)
    return x

```

Scaling

- As can be seen, ChatGPT is not that much more complex than our toy example $f_\Theta(x) = W_2 \max(W_1 x + b_1, 0) + b_2$,.
- But it is ridiculously large, with more than 10^{11} real-valued parameters.
- And it can effectively utilize its parameters when predicting each of the more than 10^{12} words it encountered during training.
- Yes, this is puzzling.

Artificial neural networks

- Modern Machine Learning is dominated by artificial neural networks.
- These are surprisingly simple mathematical formulae with a ridiculously large number of real parameters.
- They are trained by gradient descent.

Research area: understanding gradient descent

- Dávid Terjék and Diego González-Sánchez employ functional analysis and convex analysis to understand how neural networks learn.
- Contradicting traditional statistical learning theory, increasing the parameter count helps these models, apparently because gradient descent has a preference for parameter settings that generalize.

Research area: understanding latent representations

- A neural network is a huge formula defining a calculation. Partial calculation results are real vectors called *latent representations*.
- These are the concepts that the network autonomously discovers when learning to solve a task.
- One of our recent papers (Csiszárik et al, NeurIPS 2022) empirically demonstrates an interesting universality property: In some sense, latent representations are unique up to affine transformations.

Research area: automated theorem proving

- Using neural networks to guide a proof search is a natural and fruitful approach to automated theorem proving.
- Zsolt Zombori collaborates with the top researchers of this field.

Research area: computer search in pure mathematics

- Much of the current talk will be dedicated to this topic.

Applied research

Efforts led by Péter Kőrösi-Szabó. Several collaborations in a diverse range of sectors. Examples:

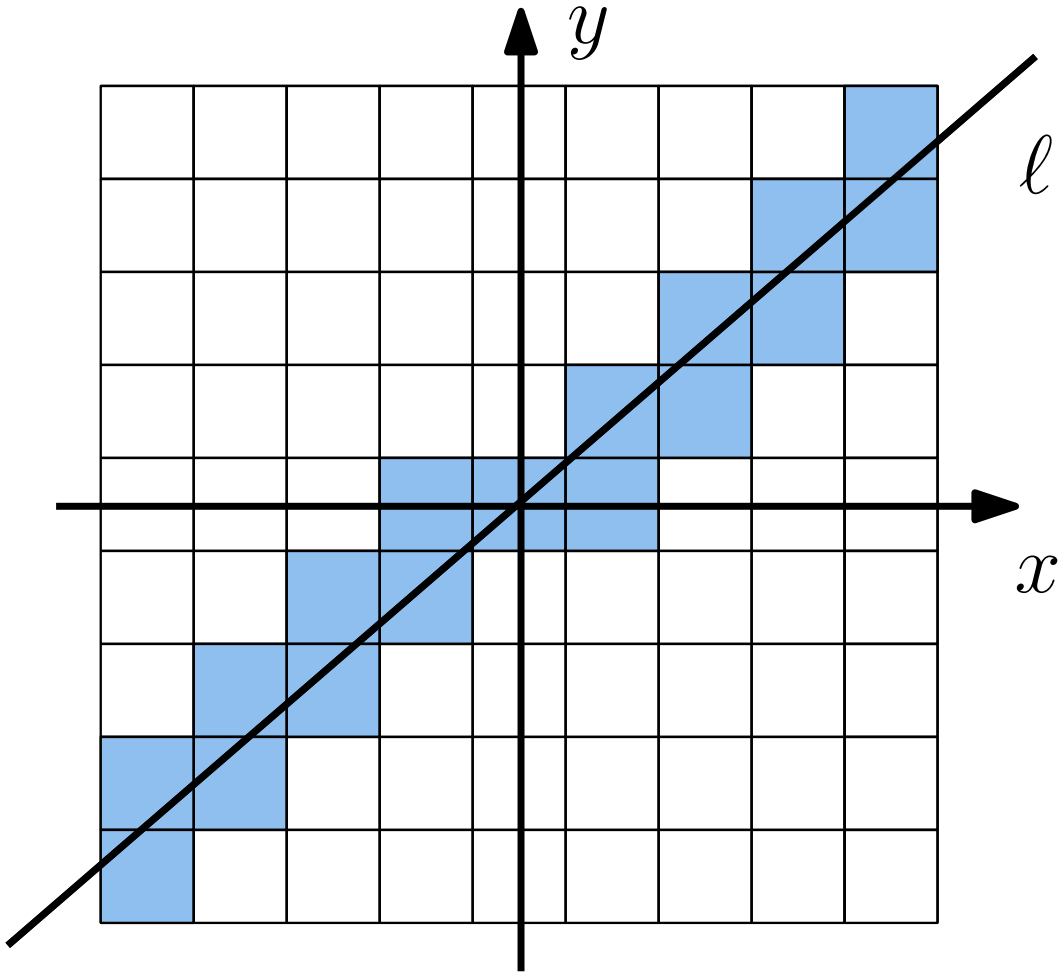
- Alteo (power sector): predicting electricity demand and solar energy production.
- Hungarian State Treasury: structuring and indexing millions of scanned but unstructured documents.
- SOTE: predicting health status based on fetal electrocardiogram.

Feel free to ask us!

- Collectively, our team has significant expertise in all sorts of questions about AI: technical, practical, theoretical, philosophical.
- Since the recent AI breakthroughs, more and more Rényi researchers find us with their questions.
- We are more than happy to help!

Slicing the chessboard

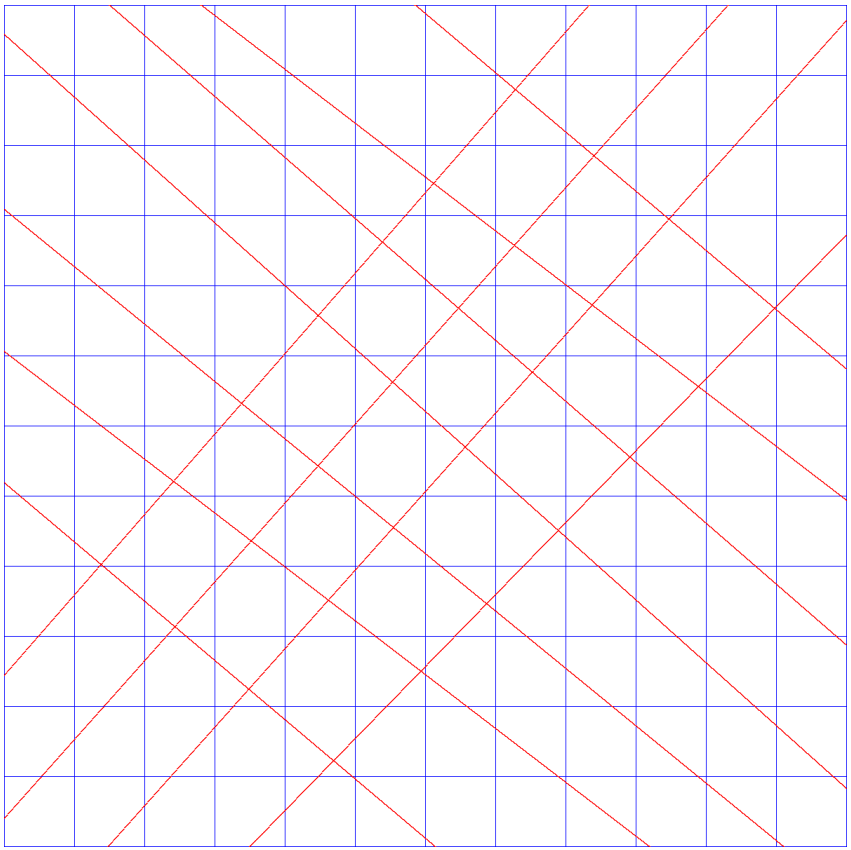
Let us denote by p_n the number of lines needed to intersect the interiors of all n^2 cells of the $n \times n$ square grid.

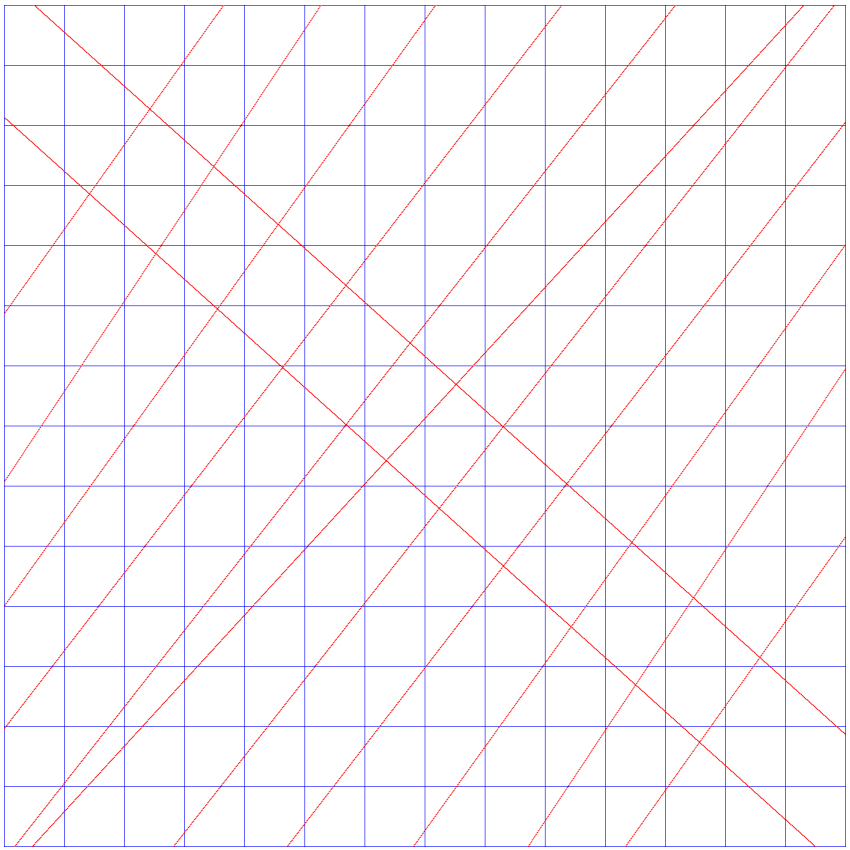


A set cover problem

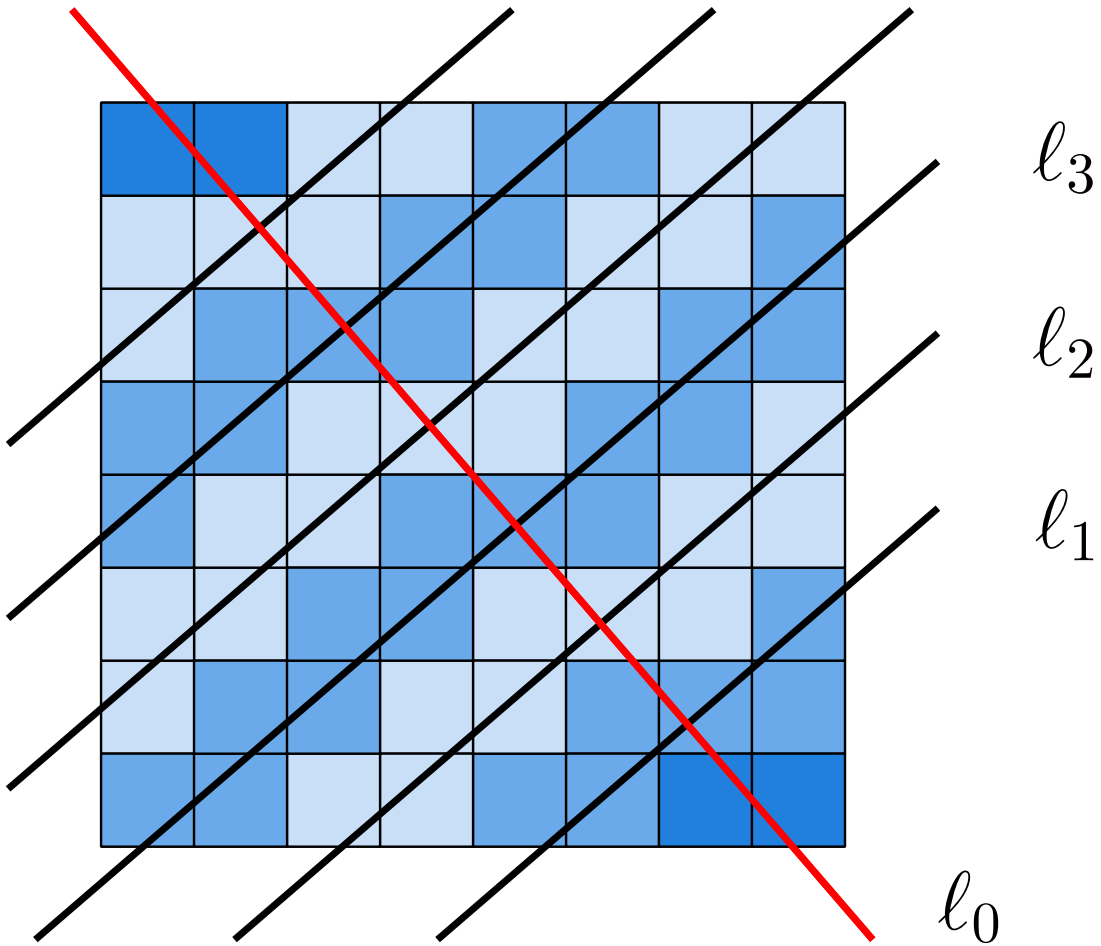
- This is a set cover problem, a case of 0-1 integer programming problems.
- Modern integer programming solvers can give optimal solutions for quite large values of n.

Computer solutions





A visual proof of the $p_n \leq n - 1$ upper bound



(B)

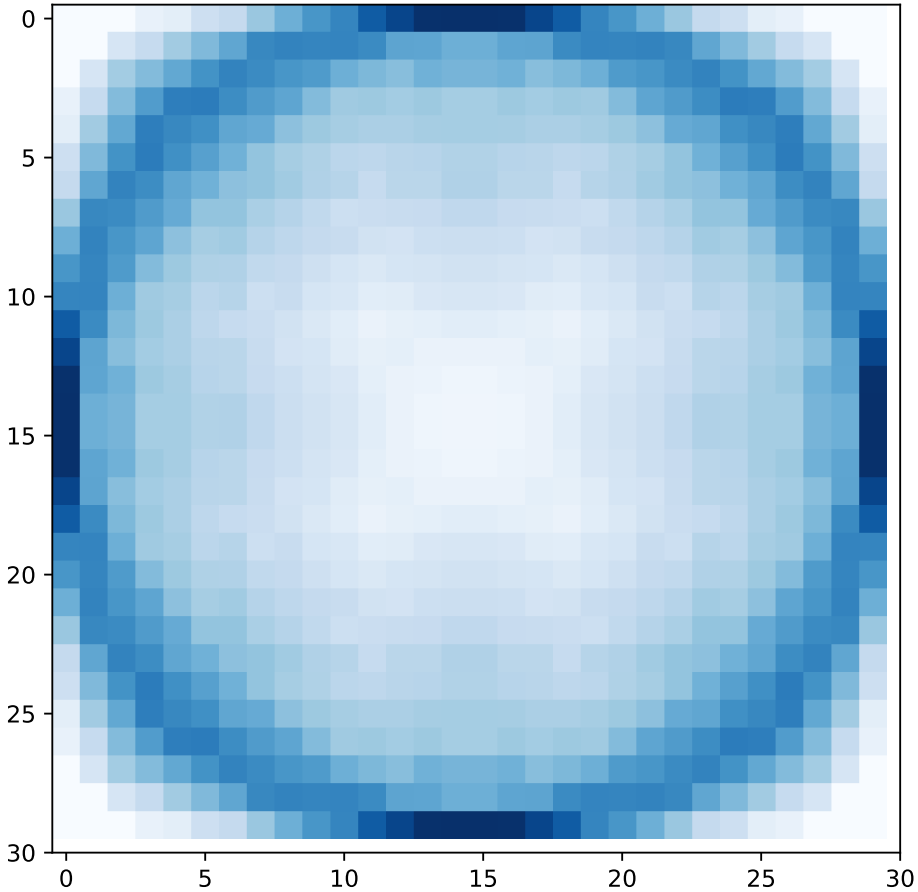
Lower bounds via the dual linear program

Theorem

Let $W \in \mathbb{R}^{n \times n}, W_{ij} \geq 0$ be a weighing of the grid cells such that for any straight line, the intersected cells' weights sum to at most 1. Then $p_n \geq \sum_{i,j} W_{ij}$.

Solving the LP for specific values of n

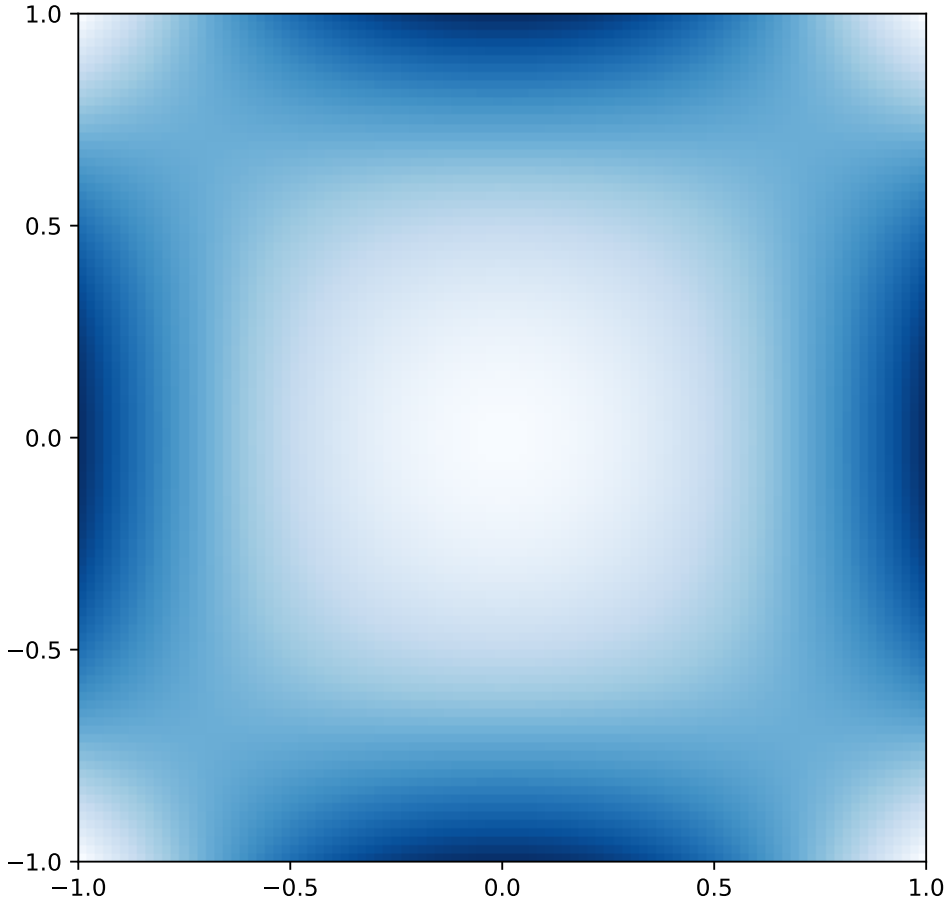
Optimal weighting of the 30×30 grid cells, as found by the dual linear program. It yields a lower bound of $p_n \geq 0.7205n$ for $n = 30$.



A nice lower bound inspired by the above

$$\mu_1(x, y) = \frac{3}{4}(x^2 - 2x^2y^2 + y^2)$$

There's a quite clean analytical proof that this weighing of the cells leads to a lower bound of $p_n \geq (\frac{2}{3} - \varepsilon)n$, for any ε and sufficiently large n.

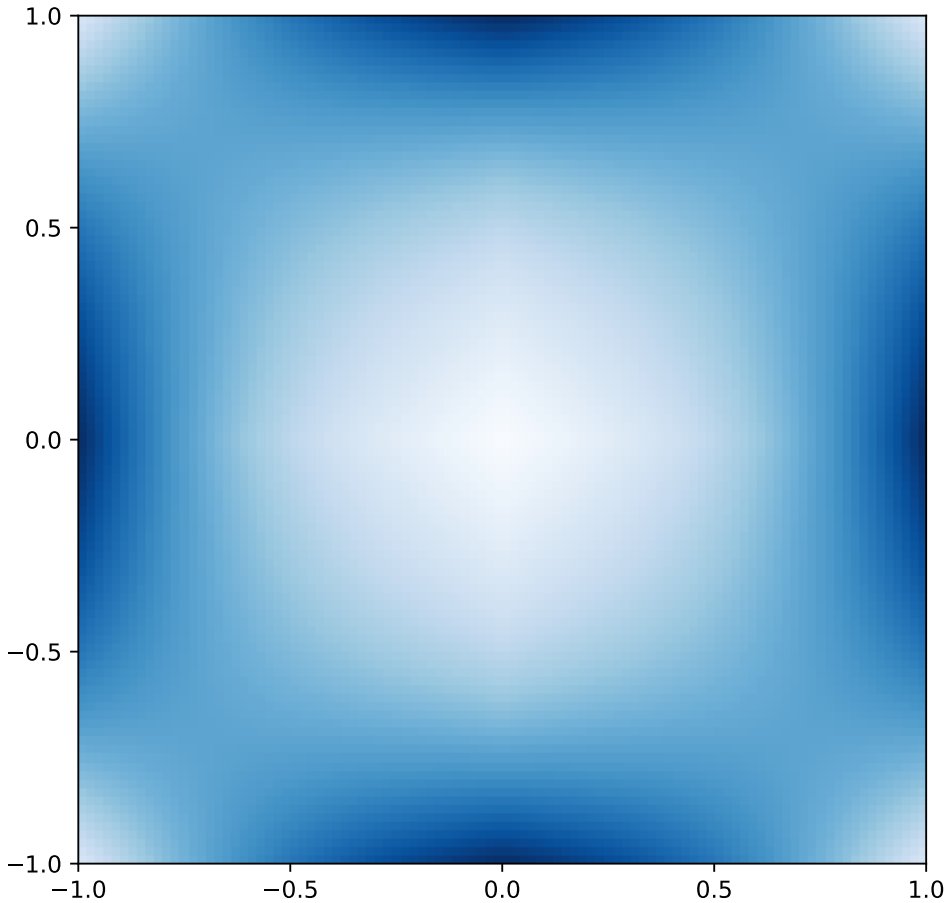


Sharper lower bound inspired by the above

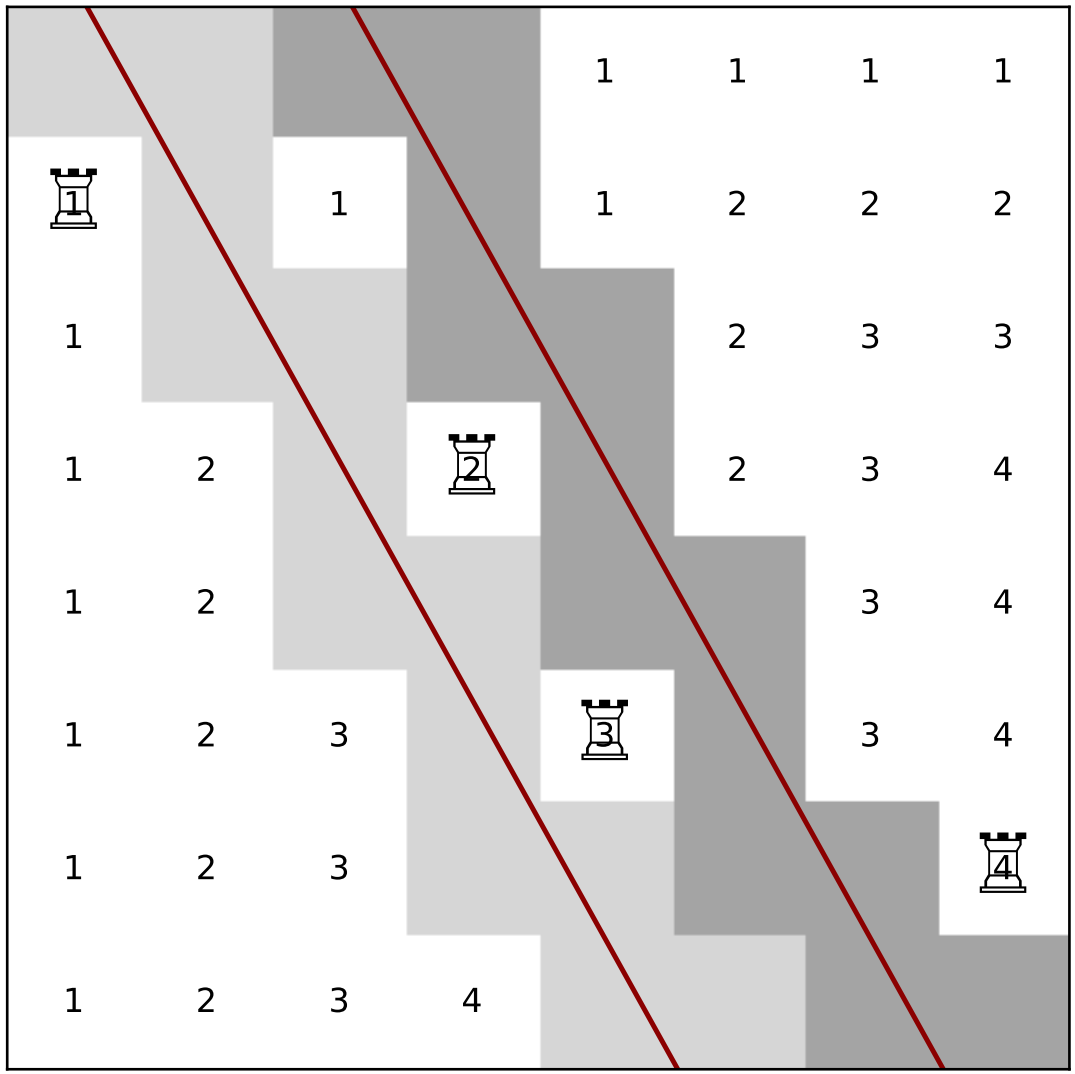
$$\mu_2(x,y) =$$

$$\begin{aligned} &0.3(|x|+|y|) \\ &+0.43(|x|^3+|y|^3) \\ &-0.585(|x|^3|y|+|y|^3|x|) \\ &-0.16x^2y^2 \end{aligned}$$

The above formula was found via gradient-based optimization of parameters, and gives $p_n \geq 0.7n$, for sufficiently large n.



A counterexample to a nice conjecture



The density of planar sets avoiding unit distances

- https://arxiv.org/abs/2207.14179
- Work by Gergely Ambrus, Adrián Csiszárik, Máté Matolcsi, Dániel Varga and Pál Zsámboki.
- Builds on a list of results by Székely, de Oliveira Filho, Ruzsa, Keleti, Matolcsi, Ambrus.

Unit distance avoiding set, Unit distance graph

- We stay on the Euclidean plane.
- A set is called unit distance avoiding, if there are no two elements unit distance away.
- A subset of the plane is called a unit distance graph (UDG) if we interpret it as a graph where two vertices are connected if and only if they are unit distance away.

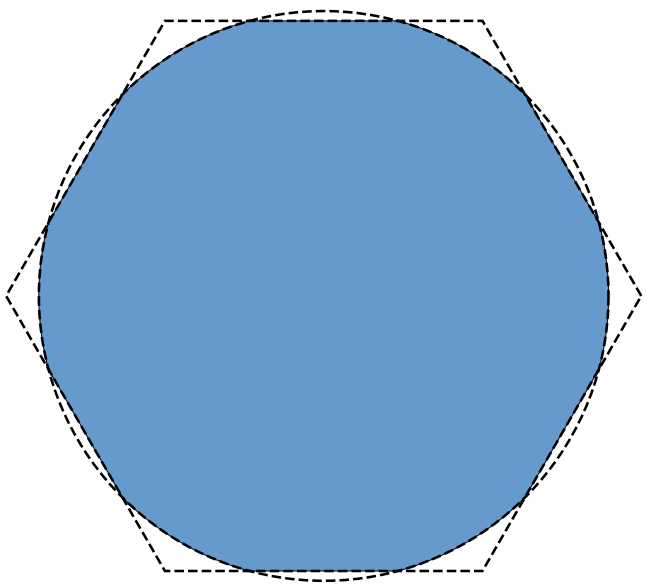
Moser's question: $m_1(\mathbb{R}^2) = ?$

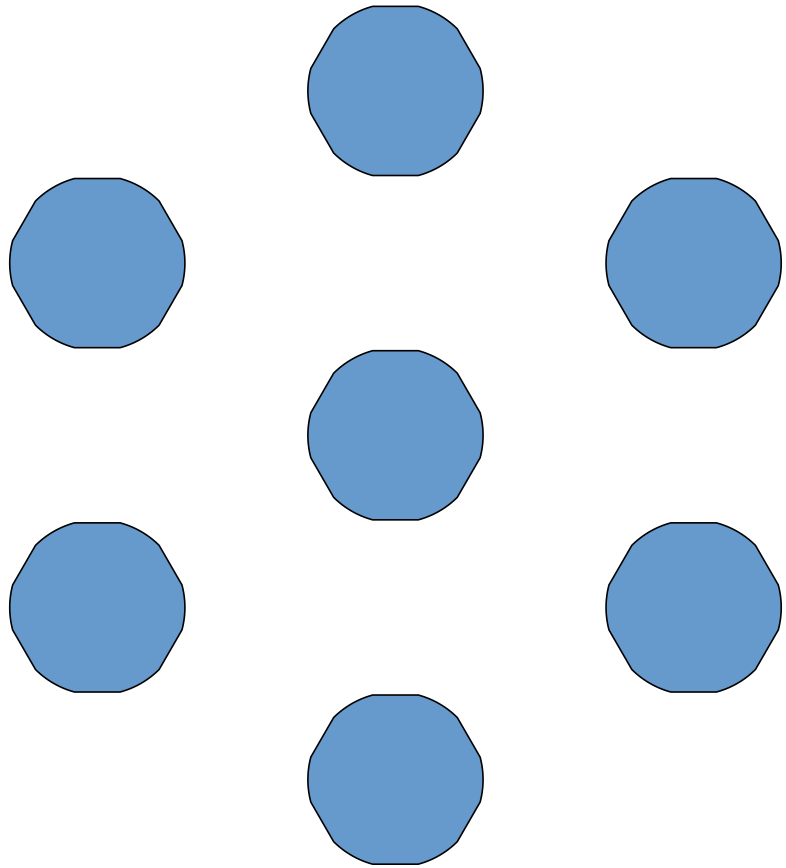
- In 1966, Leo Moser asked for the highest density measurable unit distance avoiding subset of the plane.
- More exactly, he defined $m_1(\mathbb{R}^2)$ to be the supremum of the upper densities of unit-distance free, measurable sets in $\mathbb{R}^2$, and asked about its value.
- When talking about densities on the plane, measure theoretic complications will arise. For brevity, we will ignore all of them today, and promise that they can be precisely dealt with, thanks to the fact that a high density unit distance avoiding set can be turned into a periodic high density unit distance avoiding set.

High density set, first attempt: $m_1(\mathbb{R}^2) > 0.2267$

- Let's take a unit diameter open disk. It has no unit distances, because any distance is smaller than 1.
- Let's pack the disks in a triangular grid such that no two disks are as close as 1. This leads to a density ≈ 0.2267.
- In 1967, H. T. Croft observed that this can be slightly improved by slicing off pieces of the disks.

Croft's tortoise: $m_1(\mathbb{R}^2) > 0.2293$





Erdős's conjecture: $m_1(\mathbb{R}^2) < 0.25$

- The Croft construction has not been improved since its introduction in 1967, and with the benefit of hindsight, it is natural to conjecture that it is indeed optimal. (We have tried and failed to improve it ourselves.)
- Paul Erdős was apparently not so sure about the Croft construction's optimality, so he formulated a weaker conjecture:

Erdős's Conjecture (1985)

$m_1(\mathbb{R}^2) < 1/4$. That is, the supremum of the upper densities of unit-distance free, measurable sets in $\mathbb{R}^2$ is less than $1/4$.

Warm up: $m_1(\mathbb{R}^2) \leq 1/3 \approx 0.3333$

Notation we will use through the talk:

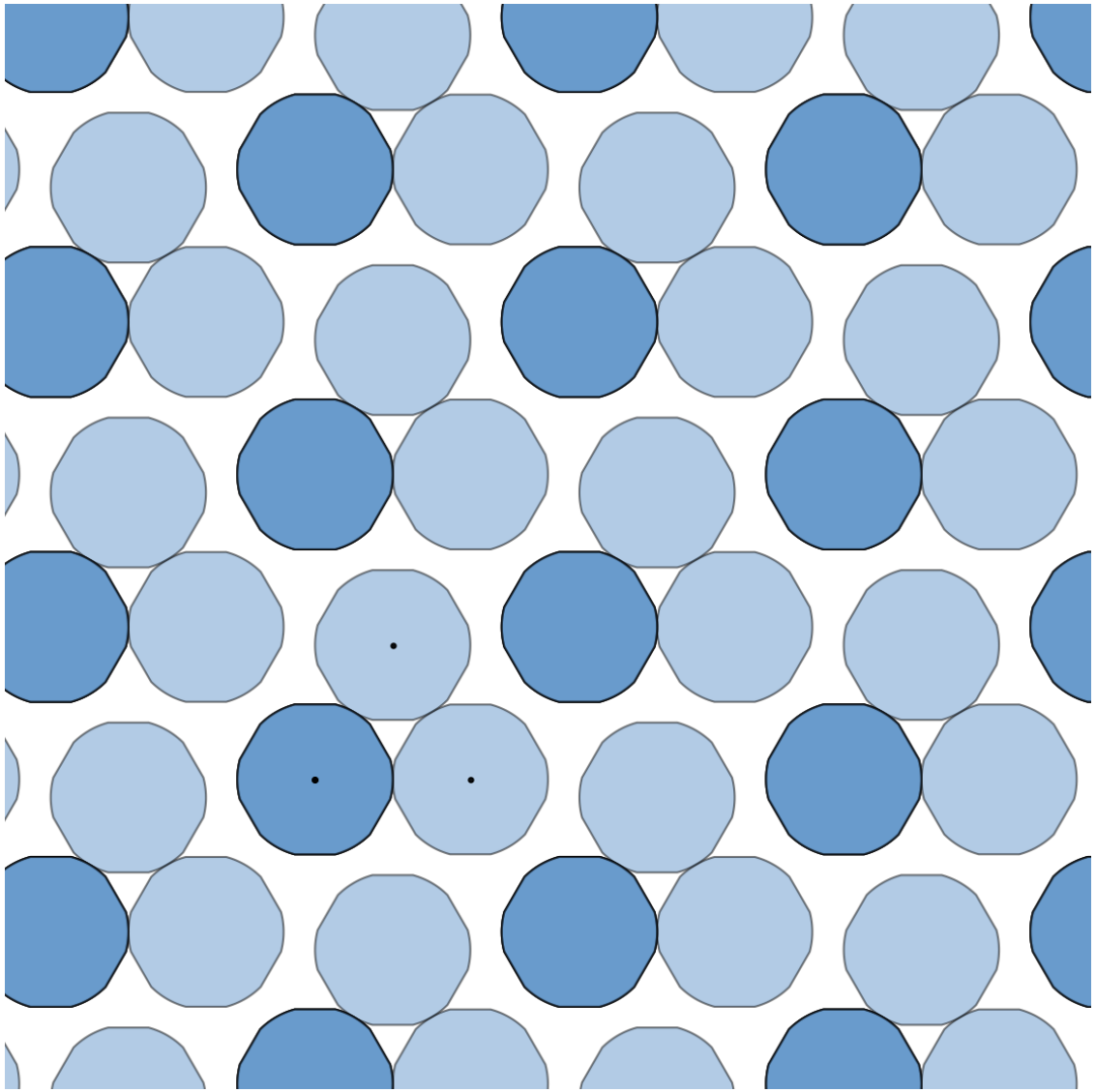
- Let A be a unit distance avoiding set with a supposedly high $\delta(A)$ density.
- Let G be a UDG on finite vertex set $X = \{x_1, \ldots, x_n\}$.
- Let A_i be $A + x_i$, a translated version of A.

Trivial, but fundamental idea

If $|x| = 1$, then $A \cap (A + x) = \emptyset$.
If $|x_i - x_j| = 1$, then $A_i \cap A_j = \emptyset$.

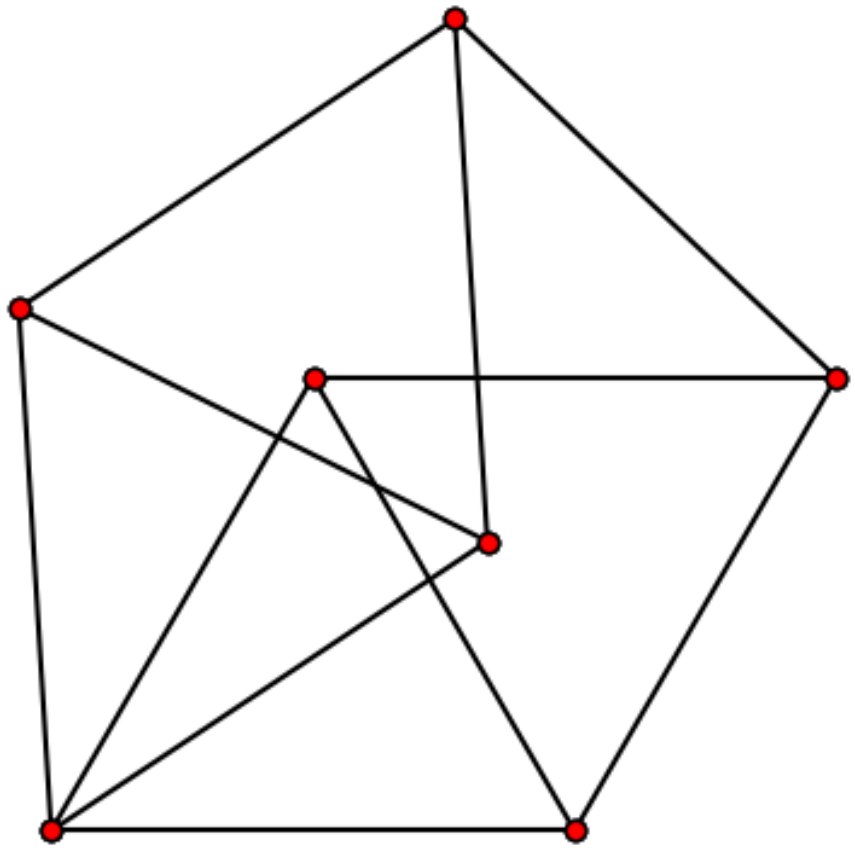
Let G now be the regular unit triangle. A_1, A_2, and A_3 are pairwise disjoint, hence their densities cannot exceed $1/3$.

Warm up: $m_1(\mathbb{R}^2) \leq 1/3 \approx 0.3333$



Moser spindle

- Brothers Leo and William Moser found this 7-vertex graph in 1961.
- Its chromatic number is 4.
- Its largest independent set is of size 2.



Improvement: $m_1(\mathbb{R}^2) \leq 2/7 \approx 0.2857$

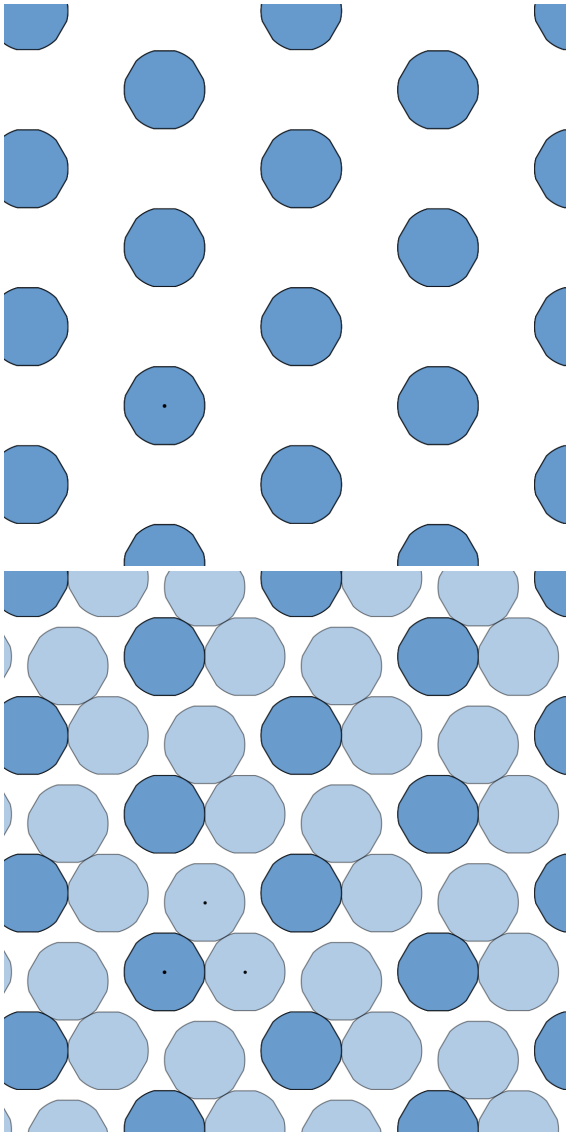
- If some point $x \in \mathbb{R}^2$ is covered by translated sets $A_i : i \in I$, then $\{x_i : i \in I\}$ must be an independent set of G.
- Let us take G to be the Moser spindle. It has 7 vertices.
- It has independence number 2, so no point of the plane is covered more than two-fold with the 7 translated versions of A.
- Hence, the density of A cannot exceed $2/7$.

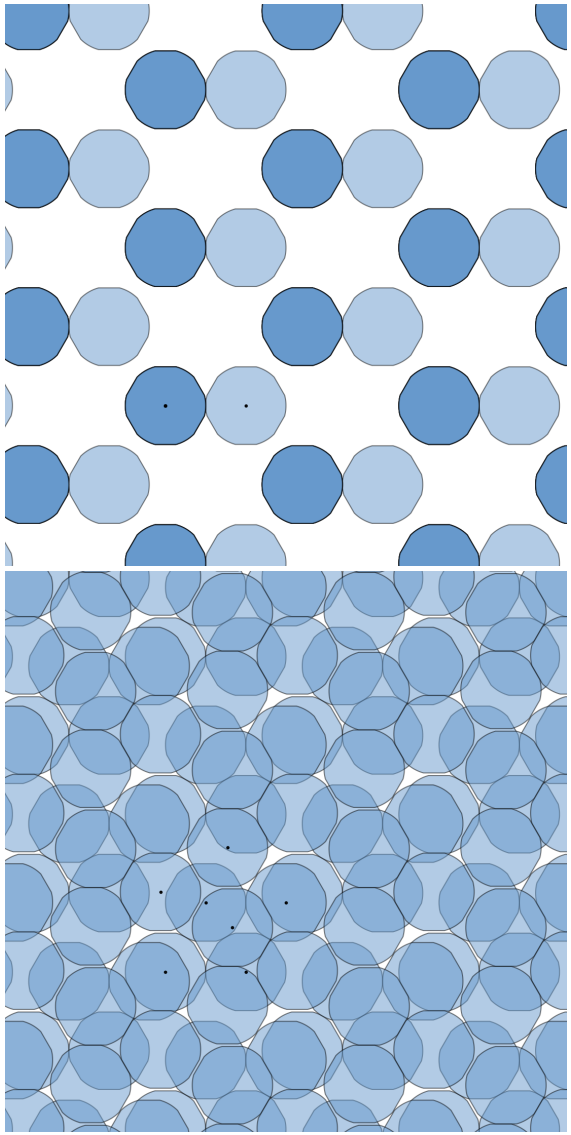
Theorem

For any unit distance graph G,

$$m_1(\mathbb{R}^2) \leq \alpha(G)/|V(G)|.$$

Moser-Croft animation, click here





Bounds via independence ratio

Many tried to settle Erdős's conjecture by trying to present a UDG with $\alpha(G)/|V(G)| < 1/4$, but none succeeded:

- 0.2857 (Moser 1966)
- 0.2813 (Fisher and Ullman 1997)
- 0.2763 (Cranston and Rabern 2015)
- 0.2565 (Bellitto, Pecher, and Sedillot 2018)
- 0.2518 (Parts 2019, unverified)
- 0.2506 (Parts 2020, unverified, 1057 vertices)

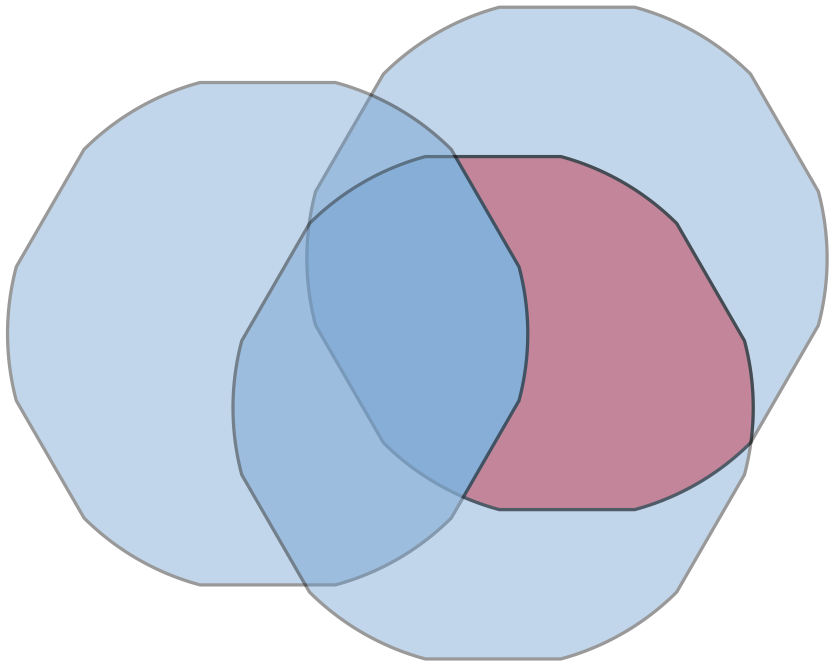
We conjecture that $1/4$ is an inherent limit of this approach.

Rough outline of our proof

- For a given unit distance graph G, we consider the A_i translated versions of A, as before.
- We write up all sorts of linear inequalities for the densities of the intersections of these sets.
- Some extra linear inequalities are implied by the fact that the pairwise intersection densities obey a certain positive definiteness property. This part is crucial for breaking the barrier.
- We use linear programming to check what bound do our linear inequalities imply for the density of A.
- We run a huge computed-aided search for unit distance graphs that give a good bound.
- We settle Erdős's conjecture by presenting a specific 24-vertex graph G_{24} that gives a bound that is better than $1/4$.

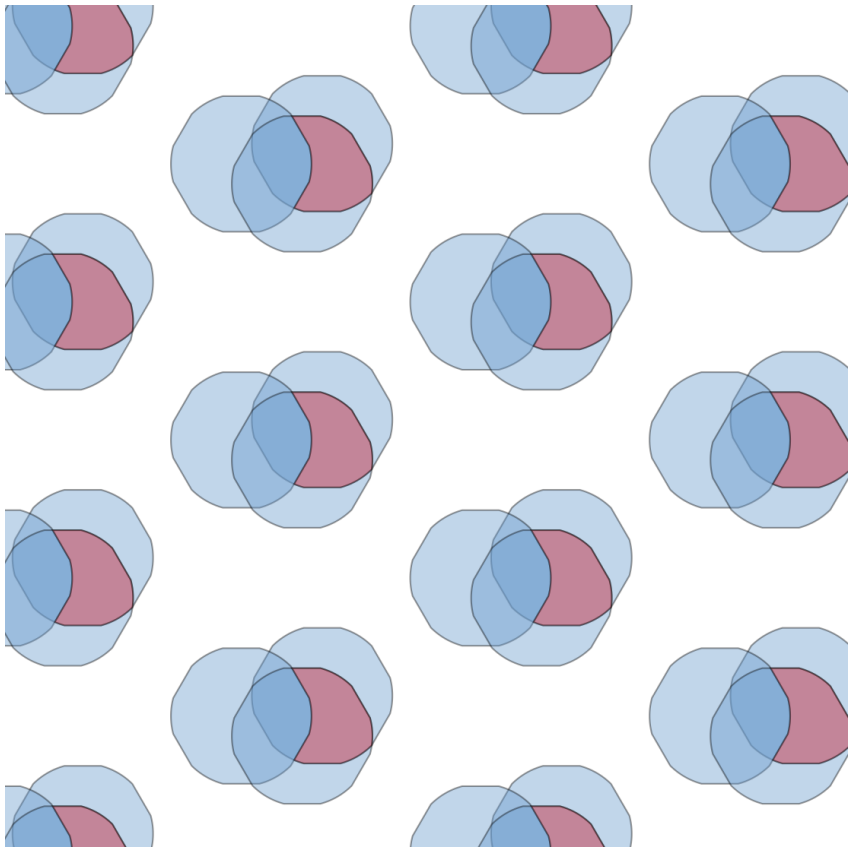
Atoms

Intuitively, what we call atoms are the cells of a Venn diagram. Our proof will proceed by writing up inequalities on their densities. The sets are $\{A_1, \ldots, A_n\}$, the translated versions of A, as before.



Atoms

Intuitively, what we call atoms are the cells of a Venn diagram. Our proof will proceed by writing up inequalities on their densities. The sets are $\{A_1, \ldots, A_n\}$, the translated versions of A, as before.



Atom densities: formal definition

For a set $Y \subset \mathbb{R}^2$ and $\nu \in \{\pm 1\}$ introduce the notation

$$Y^\nu = \begin{cases} Y, \text{ if } \nu = +1 \\ Y^c, \text{ if } \nu = -1 \end{cases} \tag{1}$$

where $Y^c = \mathbb{R}^2 \setminus Y$. Let $\sigma(n) = \{\pm 1\}^n$.
Now, let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^2$, and for each $\varepsilon \in \sigma(n)$, set

$$a_X(\varepsilon) = \delta\Big(\bigcap_{i=1}^{n} (A + x_i)^{\varepsilon_i} \Big). \tag{2}$$

Constraints on the atom densities

(ieP) $a_X(\varepsilon) \geq 0$ for each $\varepsilon \in \sigma(n)$.

(ieI) $a_X(\varepsilon) = 0$ for each $\varepsilon \in \sigma(n)$ such that $\{x_i : i \in [n], \varepsilon_i = +1\}$ contains two points at unit distance.

(ieT) $\sum_{\varepsilon \in \sigma(n)} a_X(\varepsilon) = 1$.

(ie1) $\sum_{\varepsilon \in \sigma(n;i)} a_X(\varepsilon) = \delta(A)$ for every $i \in [n]$.

For a given X, our bound is the largest value of $\delta(A)$ that is consistent with these constraints. This can be found via linear programming.

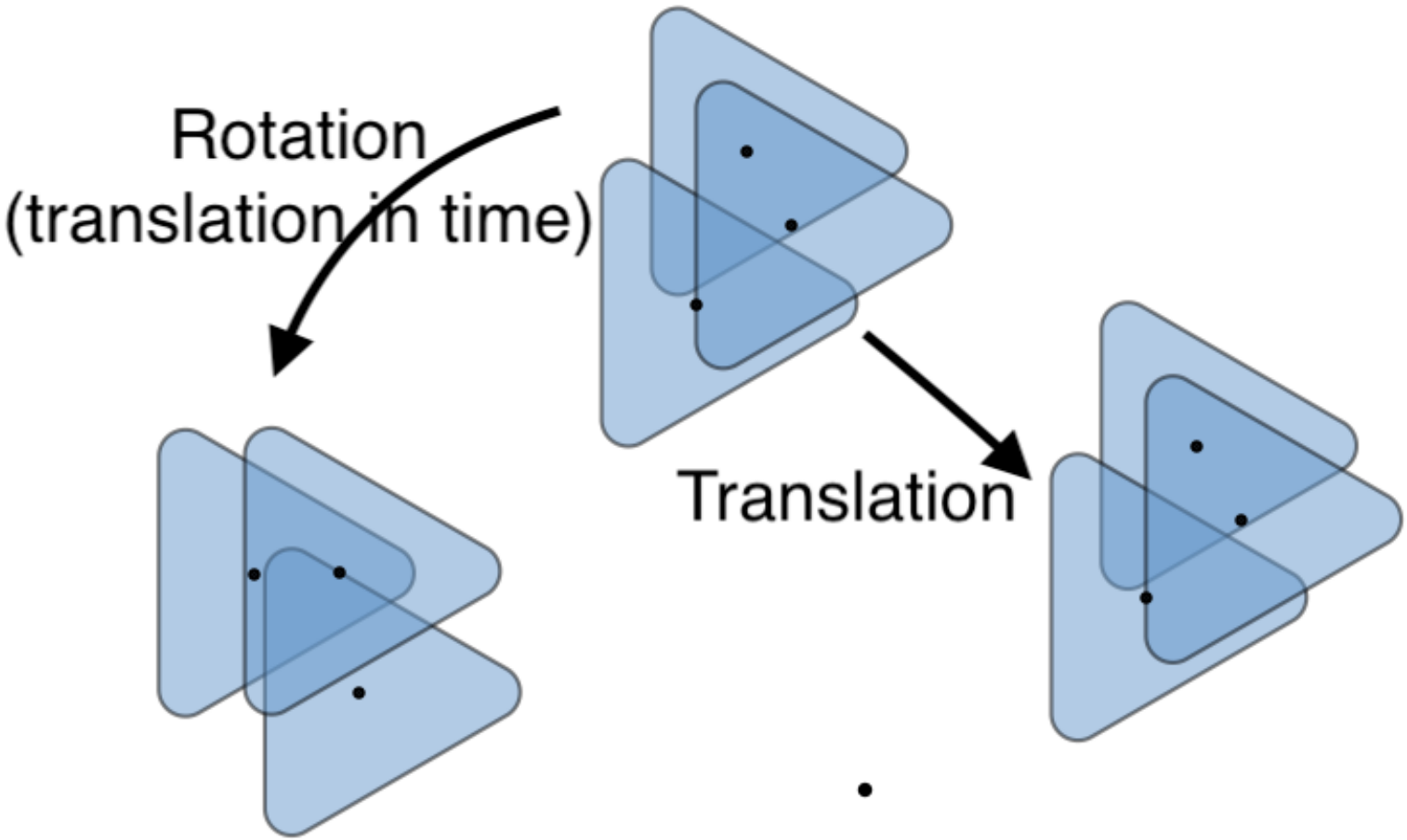
Radialization

Let μ denote the Haar probability measure on $O(2)$, and introduce the notation

$$\mathring{a}_X(\varepsilon) = \int_{O(2)} \delta\Big(\bigcap_{i=1}^{n} (A + \varphi(x_i))^{\varepsilon_i} \Big) \, \mathrm{d}\mu(\varphi)$$

for $\varepsilon \in \{+1, -1\}^n$.
Importantly, all our previous equations and inequalities remain true after radialization, that is, writing $\mathring{a}_X(\varepsilon)$ in place of $a_X(\varepsilon)$.

Radialization animation, click here



GFCN

As the above hopefully illustrates, for any two congruent finite sets $Y, Z \subset \mathbb{R}^2$,

$$\int_{O(2)} \delta\Big(\bigcap_{y \in Y} (A + \varphi(y)) \Big) \, \mathrm{d}\mu(\varphi) = \int_{O(2)} \delta\Big(\bigcap_{z \in Z} (A + \varphi(z)) \Big) \, \mathrm{d}\mu(\varphi)$$

Expressed in terms of atoms we obtain:

$\sum_{T \supseteq Y} \mathring{a}_X(T) = \sum_{S \supseteq Z} \mathring{a}_X(S)$

We call these equations the congruence constraints.

Autocorrelation function

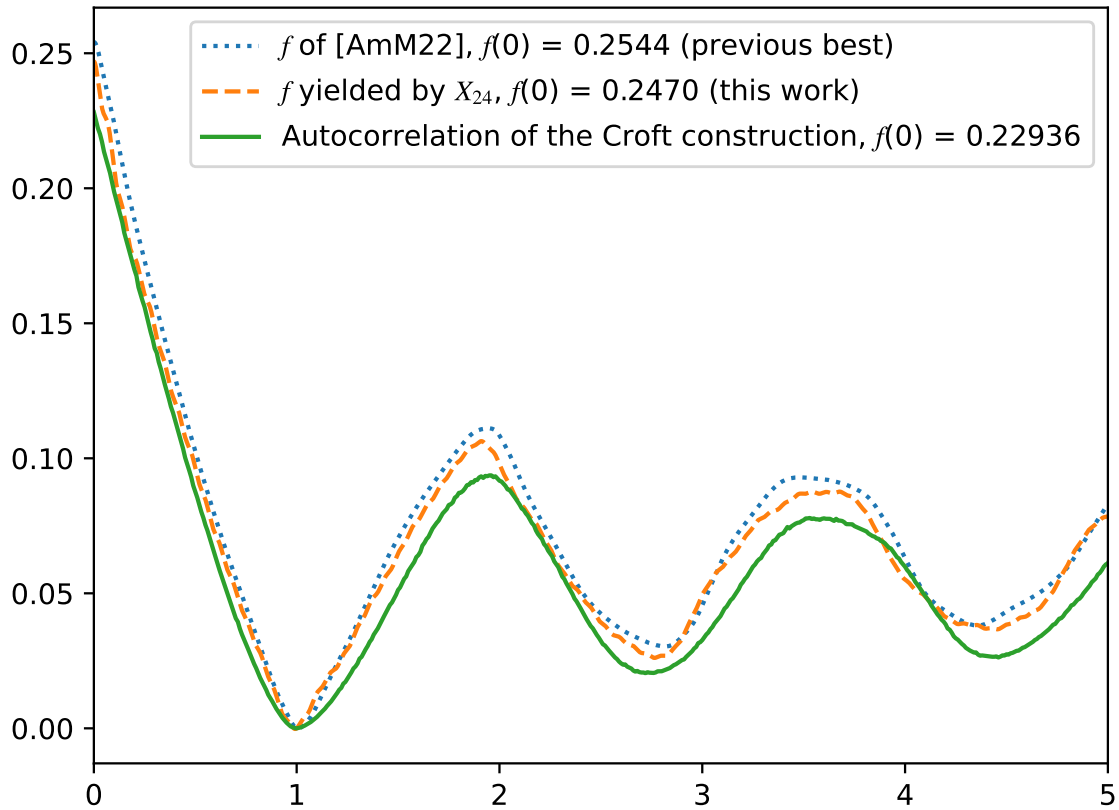
The autocorrelation function of A is defined by

$$f(x) = \int_{O(2)} \delta\Big(A \cap (A + \varphi(x))\Big)\, \mathrm{d}\mu(\varphi)$$

Intuitively, it is the probability that randomly dropping a length x interval on the plane, both its endpoints will be in A. Note the special values $f(0) = \delta(A)$ and $f(1) = 0$.
If some distance x appears in our UDG, we can write $f(x)$ up as a sum of some atom densities.

Autocorrelation function

The green line is the autocorrelation function of the Croft construction.



Positive definiteness

Let Ω_2 be the Bessel function of the first kind with parameter 0.

Theorem

The autocorrelation function $f(x)$ is positive definite. This implies that it can be written in the form

$$f(x) = \sum_{t \geq 0} \kappa(t)\Omega_2(tx), \kappa(t) \geq 0.$$

The final LP

Maximize $\sum_{t\geq 0}\kappa(t)$ subject to

(CP) $\kappa(t)\geq 0$ for every $t\geq 0$

(IEP) $\mathring{a}_X(\varepsilon)\geq 0$ for each $\varepsilon\in\sigma(n)$

(C0) $\sum_{t\geq 0}\kappa(t)\Omega_2(t)=0$

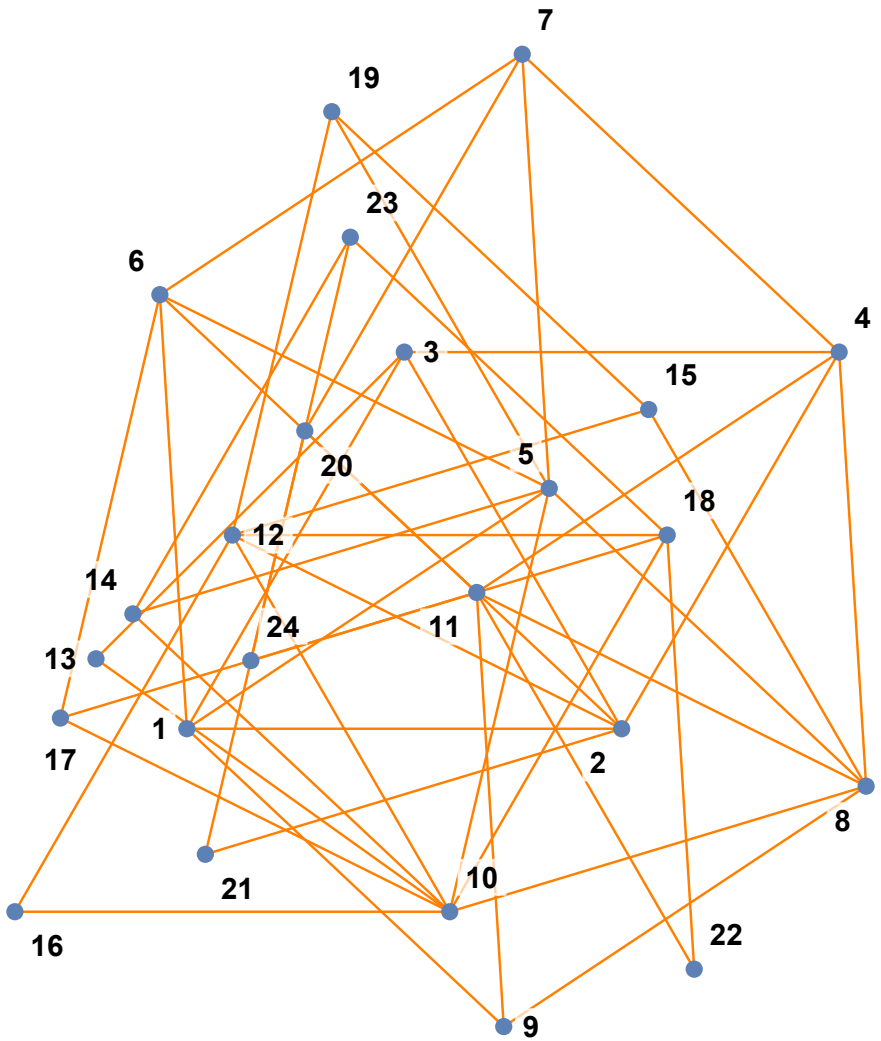
(IET) $\sum_{\varepsilon\in\sigma(n)}\mathring{a}_X(\varepsilon)=1$

(IE1) $\sum_{t\geq 0}\kappa(t)-\sum_{\varepsilon\in\sigma(n;i)}\mathring{a}_X(\varepsilon)=0$ for every $i\in[n]$

(IE2) $\sum_{t\geq 0}\kappa(t)\Omega_2(t|x_i-x_j|)-\sum_{\varepsilon\in\sigma(n;i,j)}\mathring{a}_X(\varepsilon)=0$ for $i\neq j$

(IEC) $\sum_{\varepsilon\in\sigma(n;I)}\mathring{a}_X(\varepsilon)-\sum_{\varepsilon\in\sigma(n;J)}\mathring{a}_X(\varepsilon)=0$ for every $I\cong J$.

The result of the search: G_{24}



G_{24} in symbolic form

$$x_1 = 0$$
$$x_2 = 1$$
$$x_3 = \frac{1}{2} + \frac{\sqrt{3}i}{2}$$
$$x_4 = \frac{3}{2} + \frac{\sqrt{3}i}{2}$$
$$x_5 = \frac{5}{6} + \frac{\sqrt{11}i}{6}$$
$$\dots$$
$$x_{23} = -\frac{\sqrt{33}}{6} + \frac{4}{3} + \frac{\sqrt{11}i}{6} + \frac{\sqrt{3}i}{3}$$
$$x_{24} = -\frac{\sqrt{33}}{4} + \frac{19}{12} - \frac{\sqrt{11}i}{12} + \frac{\sqrt{3}i}{4}$$

The result of the search: $m_1(\mathbb{R}^2) \leq 0.247$

Theorem

The graph G_{24} is a witness to the fact that $m_1(\mathbb{R}^2) \leq 0.247$, settling Erdős's Conjecture.

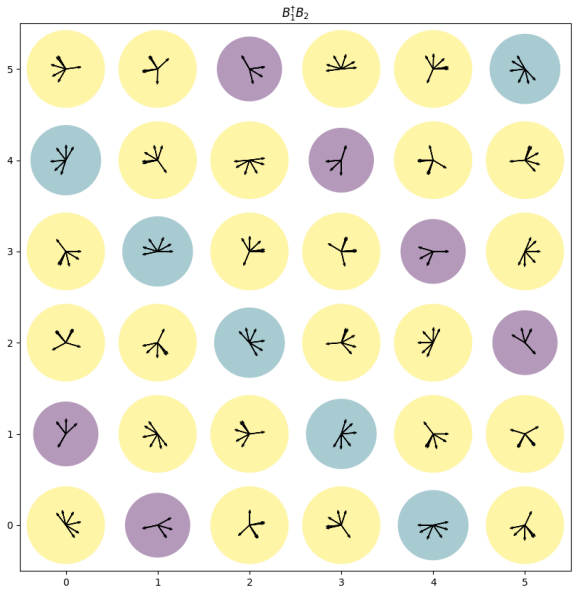
The linear program defined by G_{24} has 22321 atom variables and 12000 Fourier variables. It has 24 (IE1) constrains, 227 (IE2) constraints connecting the Fourier variables to the atom variables, and 2122 (IEC) congruence constraints. Amusingly, $\chi_f(G_{24}) = 7/2$.

Ongoing work

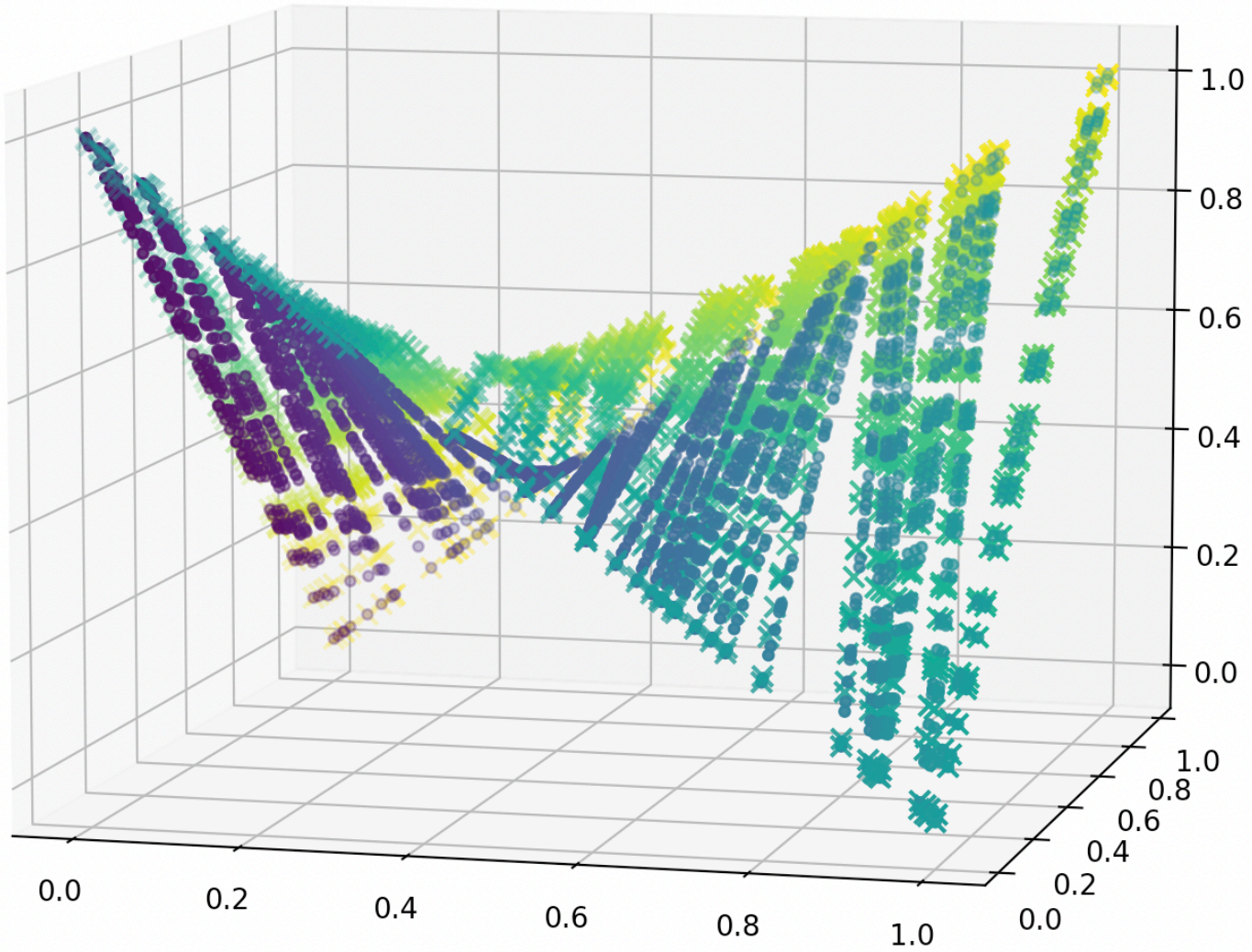
- Upper bounding the fraction of the plane that can be 4-colored (that is, covered by 4 unit distance avoiding sets).
- Attempting to prove that the measurable chromatic number of the plane is at least 6.
- A theoretical conjecture: $FCN(\mathbb{R}^2) = GFCN(\mathbb{R}^2) = 4$. (The Fourier apparatus is necessary for breaking the $1/4$ barrier.)

Mutually unbiased bases in composite dimensions

- Ongoing work with Máté Matolcsi, Ákos Matszangosz and Mihály Weiner.
- Here we use computer search to build a large collection of mathematical objects of interest (MUB triplets, Hadamard cubes).
- We then uncover regularities among them that we can formally prove.

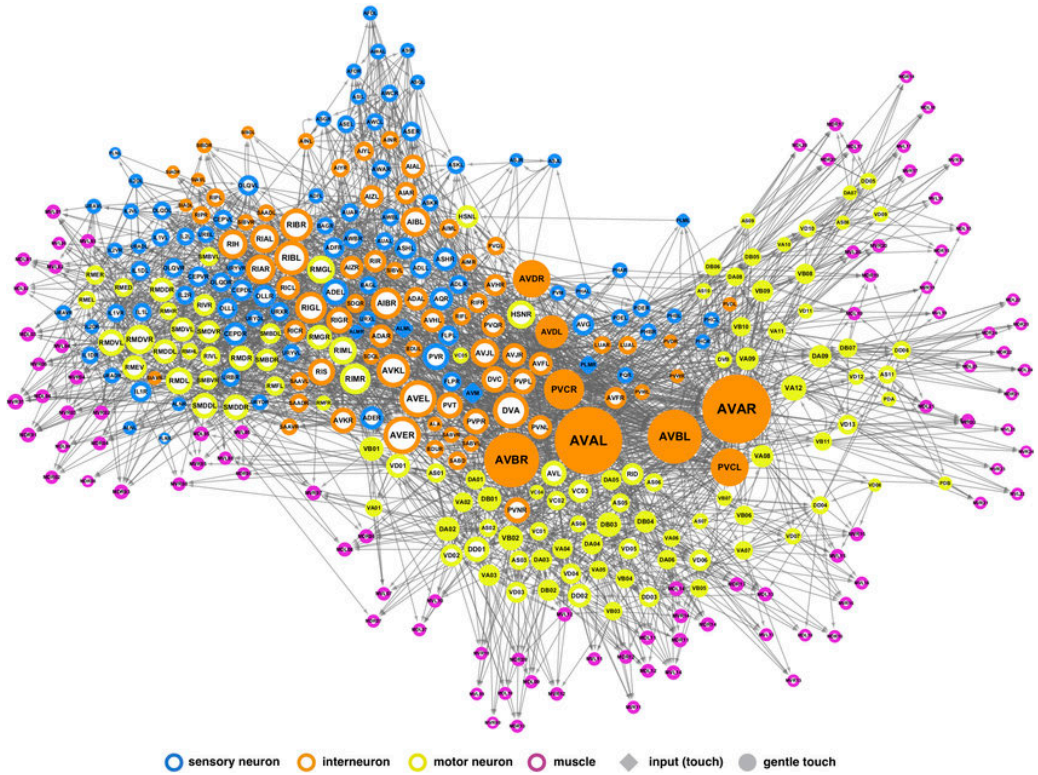


On a question of Imre Bárány



On a question of Albert-László Barabási

- The neural connectome of a C. Elegans (302 neurons) can be constructed as a union of 272 bipartite graphs.
- It cannot be constructed as a union of 271 bipartite graphs.



Conclusion

Let's talk if you feel that such an approach could help solving your favorite open problem!