High-dimensional random geometric graphs

Gábor Lugosi

ICREA and Pompeu Fabra University, Barcleona

joint work with

Luc Devroye (McGill University, Montréal)

András György (SZTAKI, Budapest)

Frederic Udina (Pompeu Fabra University, Barcelona),

motivation: alien detection

n satellites search for signs of extraterrestrial invasion.

motivation: alien detection

$\mathbf{n}$ satellites search for signs of extraterrestrial invasion.
Satellite $\mathbf{i}$ receives $\mathbf{Z_i = (Z_{i,1}, \ldots, Z_{i,d})}$.

motivation: alien detection

$\mathbf{n}$ satellites search for signs of extraterrestrial invasion.
Satellite $\mathbf{i}$ receives $\mathbf{Z_i = (Z_{i,1}, \ldots, Z_{i,d})}$.

If there is no signal, all $\mathbf{Z_{i,t}}$ are i.i.d. standard normal.

motivation: alien detection

$\mathbf{n}$ satellites search for signs of extraterrestrial invasion.
Satellite $\mathbf{i}$ receives $\mathbf{Z_i = (Z_{i,1}, \ldots, Z_{i,d})}$.

If there is no signal, all $\mathbf{Z_{i,t}}$ are i.i.d. standard normal.

Alternatively, a small subset $\mathbf{S \subset \{1, \ldots, n\}}$ of satellites receives a common signal embedded in noise:

$$\mathbf{Z_{i,t}} = \begin{cases} \mathbf{N_{i,t}} & \text{if } \mathbf{i \notin S} \\ \mathbf{(N_{i,t} + Y_t)/\sqrt{1+\sigma^2}} & \text{if } \mathbf{i \in S} \end{cases}$$

where the $\mathbf{N_{i,t}}$ are i.i.d. standard normal and the $\mathbf{Y_t}$ are independent normal $\mathbf{(0, \sigma^2)}$.

random correlation graph

For the testing problem, it is natural to calculate pairwise correlations

$$\frac{(\mathbf{Z_i}, \mathbf{Z_i})}{\|\mathbf{Z_i}\| \cdot \|\mathbf{Z_j}\|}$$

and define a graph by connecting $\mathbf{i}$ and $\mathbf{j}$ if the correlation is large enough.

Under the null hypothesis, the $\mathbf{X_i} = \mathbf{Z_i}/\|\mathbf{Z_i}\|$ are uniformly distributed on the sphere and we have a random geometric graph in $\mathbb{R}^{\mathbf{d}}$.

random geometric graph

Given $\mathbf{n}$ i.i.d. points in $\mathbb{R}^{\mathbf{d}}$, connect two with an edge if their distance is $\leq \mathbf{r}$.

Well understood if $\mathbf{d}$ is fixed and $\mathbf{n} \to \infty$.

We are interested in the behavior of the graph when the dimension is large.

random geometric graph

Model: Let $\mathbf{X_1, \dots, X_n}$ be independent vectors, uniform on $\mathbf{S_{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}}$.

For a given $\mathbf{p \in (0, 1)}$, we define the random geometric graph $\mathbf{\overline{G}(n, d, p)}$

Vertex set $\mathbf{V = \{1, \dots, n\}}$.

$\mathbf{i}$ and $\mathbf{j}$ are connected by an edge if an only if

$$\mathbf{(X_i, X_j) \geq t_{p,d}}$$

where $\mathbf{t_{p,d}}$ is such that

$$\mathbb{P}\left\{\mathbf{(X_i, X_j) \geq t_{p,d}}\right\} = \mathbf{p}\ .$$

Equivalently, $\mathbf{i \sim j}$ if and only if $\|\mathbf{X_i - X_j}\| \leq \sqrt{\mathbf{2(1 - t_{p,d})}}$.

edge probability

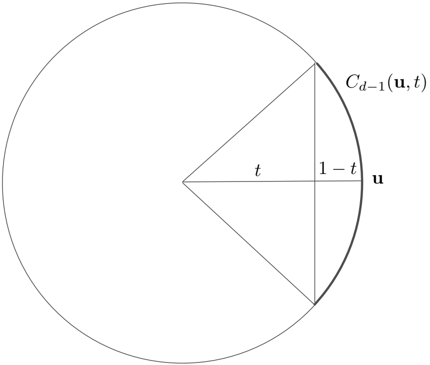
For $\mathbf{p = 1/2}$, $\mathbf{t_{p,d} = 0}$.

edge probability

For $\mathbf{p = 1/2}$, $\mathbf{t_{p,d} = 0}$.

Let $\boldsymbol{\mu_{d-1}}$ be the uniform probability measure over $\mathbf{S_{d-1}}$.
For $\mathbf{u} \in \mathbf{S_{d-1}}$ and $\mathbf{0 \le t \le 1}$, a spherical cap of height $\mathbf{1 - t}$ around $\mathbf{u}$ is

$$\mathbf{C_{d-1}(u, t) = \{x \in \mathbb{R}^d : x \in S_{d-1}, (x, u) \ge t\}}$$



edge probability

$\mathbf{p} = \mu_{\mathbf{d}-1}(\mathbf{C}_{\mathbf{d}-1}(\mathbf{e}, \mathbf{t}_{\mathbf{p},\mathbf{d}}))$ is the normalized surface area of a spherical cap of height $\mathbf{1} - \mathbf{t}_{\mathbf{p},\mathbf{d}}$.

edge probability

$\mathbf{p = \mu_{d-1}(C_{d-1}(e, t_{p,d}))}$ is the normalized surface area of a spherical cap of height $\mathbf{1 - t_{p,d}}$.

It is useful to represent

$$\mathbf{X} = \frac{\mathbf{Z}}{\|\mathbf{Z}\|}$$

with $\mathbf{Z} \in \mathbb{R}^{\mathbf{d}}$ standard normal.

edge probability

$\mathbf{p = \mu_{d-1}(C_{d-1}(e, t_{p,d}))}$ is the normalized surface area of a spherical cap of height $\mathbf{1 - t_{p,d}}$.

It is useful to represent

$$\mathbf{X = \frac{Z}{\|Z\|}}$$

with $\mathbf{Z} \in \mathbb{R}^{\mathbf{d}}$ standard normal.

Clearly, $\mathbb{E}\|\mathbf{Z}\|^2 = \mathbf{d}$.

edge probability

$\mathbf{p} = \mu_{\mathbf{d}-1}(\mathbf{C}_{\mathbf{d}-1}(\mathbf{e}, \mathbf{t}_{\mathbf{p},\mathbf{d}}))$ is the normalized surface area of a spherical cap of height $\mathbf{1} - \mathbf{t}_{\mathbf{p},\mathbf{d}}$.

It is useful to represent

$$\mathbf{X} = \frac{\mathbf{Z}}{\|\mathbf{Z}\|}$$

with $\mathbf{Z} \in \mathbb{R}^{\mathbf{d}}$ standard normal.

Clearly, $\mathbb{E}\|\mathbf{Z}\|^2 = \mathbf{d}$.

Since $\|\mathbf{Z}\|$ is a Lipschitz function of $\mathbf{Z}$, $\mathbf{var}(\|\mathbf{Z}\|) \leq \mathbf{1}$.

edge probability

$\mathbf{p = \mu_{d-1}(C_{d-1}(e, t_{p,d}))}$ is the normalized surface area of a spherical cap of height $\mathbf{1 - t_{p,d}}$.

It is useful to represent

$$\mathbf{X} = \frac{\mathbf{Z}}{\|\mathbf{Z}\|}$$

with $\mathbf{Z} \in \mathbb{R}^{\mathbf{d}}$ standard normal.

Clearly, $\mathbb{E}\|\mathbf{Z}\|^2 = \mathbf{d}$.

Since $\|\mathbf{Z}\|$ is a Lipschitz function of $\mathbf{Z}$, $\mathbf{var}(\|\mathbf{Z}\|) \leq \mathbf{1}$.

In particular, $\|\mathbf{Z}\|/\sqrt{\mathbf{d}} \to \mathbf{1}$ in probability.

This implies $\mathbf{X_1}\sqrt{\mathbf{d}}$ is approximately standard normal.

edge probability

Consequence: for any $\mathbf{s > 0}$,

$$\mu_{d-1}(C_{d-1}(e, s/\sqrt{d})) = \mathbb{P}\{X_1 > s/\sqrt{d}\} \to 1 - \Phi(s)$$

as $\mathbf{d \to \infty}$.

edge probability

Consequence: for any $\mathbf{s > 0}$,

$$\mu_{d-1}(C_{d-1}(e, s/\sqrt{d})) = \mathbb{P}\{X_1 > s/\sqrt{d}\} \to 1 - \Phi(s)$$

as $\mathbf{d \to \infty}$.

For any fixed $\mathbf{p \in (0, 1)}$,

$$\lim_{d\to\infty} t_{p,d}\sqrt{d} = \Phi^{-1}(1 - p)\ .$$

very large dimension

$\mathbf{G(n,p)}$ denotes the Erdős-Rényi random graph. ($\mathbf{n}$ vertices, egdes are present independently, with probability $\mathbf{p}$.)

Total variation distance between two random graphs $\mathbf{G}$ and $\mathbf{G'}$:

$$d_{TV}(G, G') = \max_{\mathcal{G}} |\mathbb{P}\{G \in \mathcal{G}\} - \mathbb{P}\{G' \in \mathcal{G}\}|$$

where the maximum is over all $\mathbf{2^{\binom{n}{2}}}$ sets of graphs over $\mathbf{n}$ vertices.

very large dimension

$\mathbf{G(n,p)}$ denotes the Erdős-Rényi random graph. ($\mathbf{n}$ vertices, egdes are present independently, with probability $\mathbf{p}$.)

Total variation distance between two random graphs $\mathbf{G}$ and $\mathbf{G'}$:

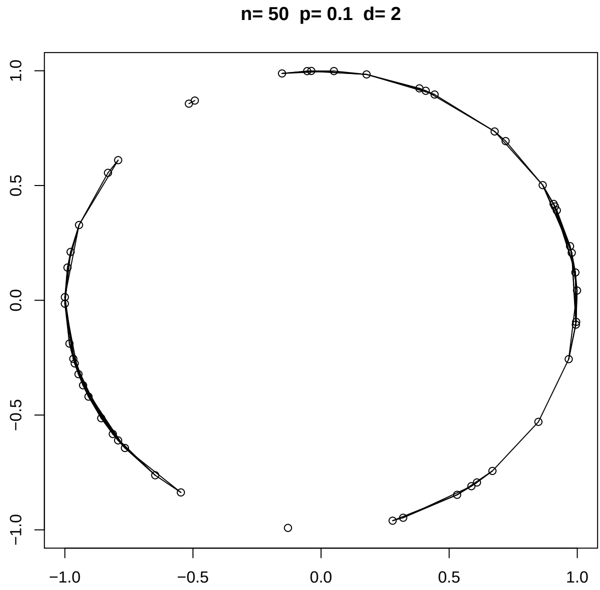
$$d_{TV}(G, G') = \max_{\mathcal{G}} |\mathbb{P}\{G \in \mathcal{G}\} - \mathbb{P}\{G' \in \mathcal{G}\}|$$

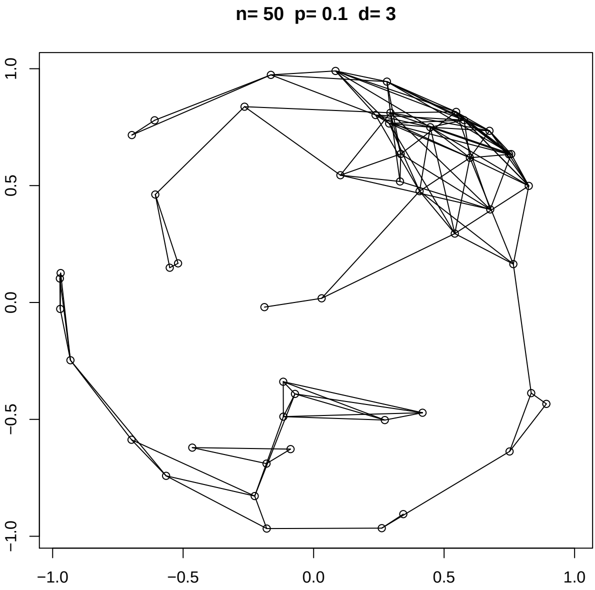
where the maximum is over all $2^{\binom{n}{2}}$ sets of graphs over $\mathbf{n}$ vertices.

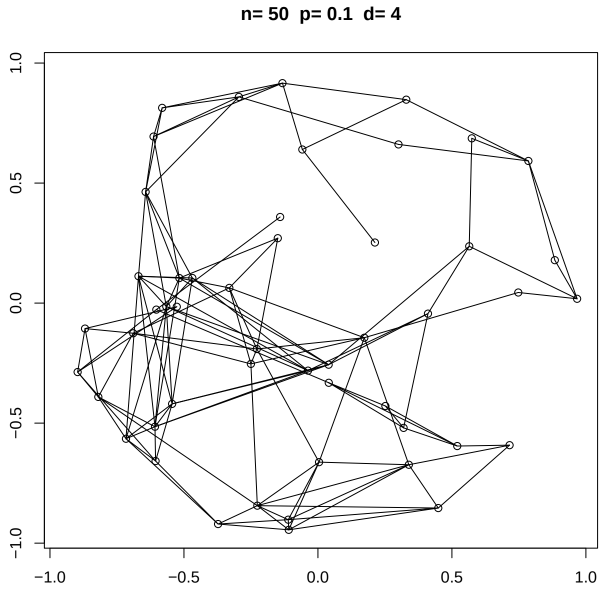
THEOREM. Fix $\mathbf{n}$ and $\mathbf{p}$. Then

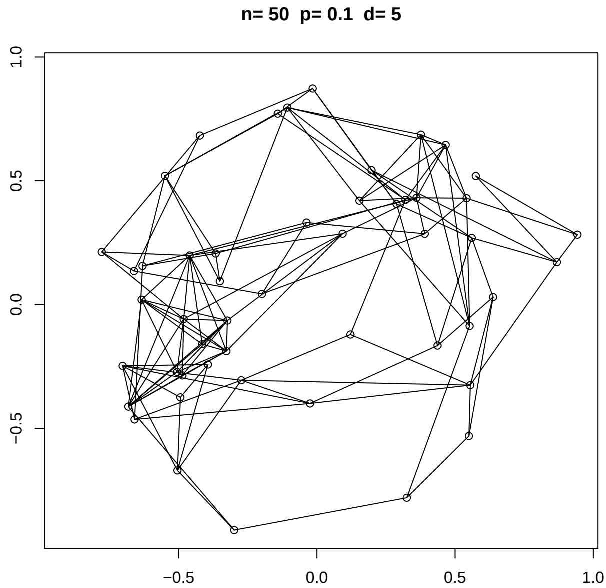
$$\lim_{d\to\infty} d_{TV}(\overline{G}(n,d,p), G(n,p)) = 0 \ .$$

Follows from a multivariate central limit theorem.


n= 50 p= 0.1 d= 2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

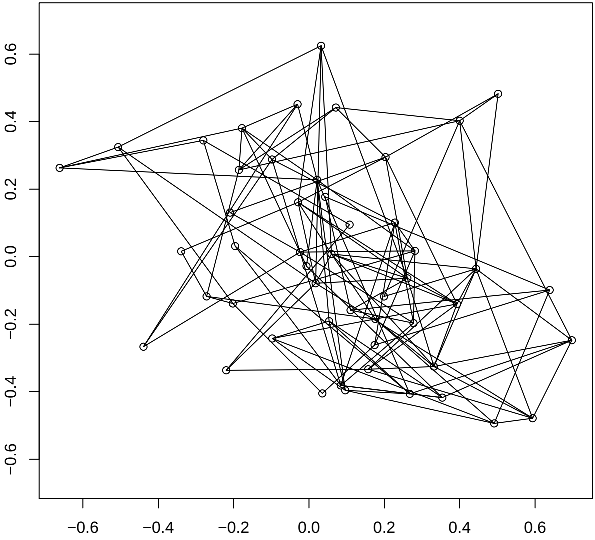

n= 50 p= 0.1 d= 3
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

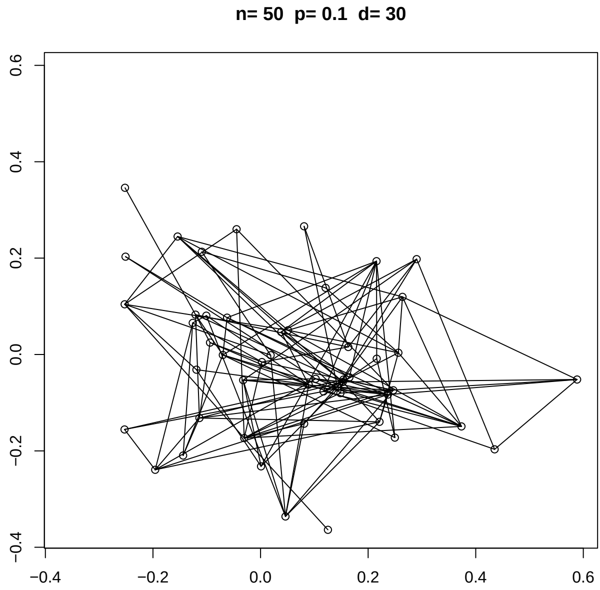

n= 50 p= 0.1 d= 4
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0



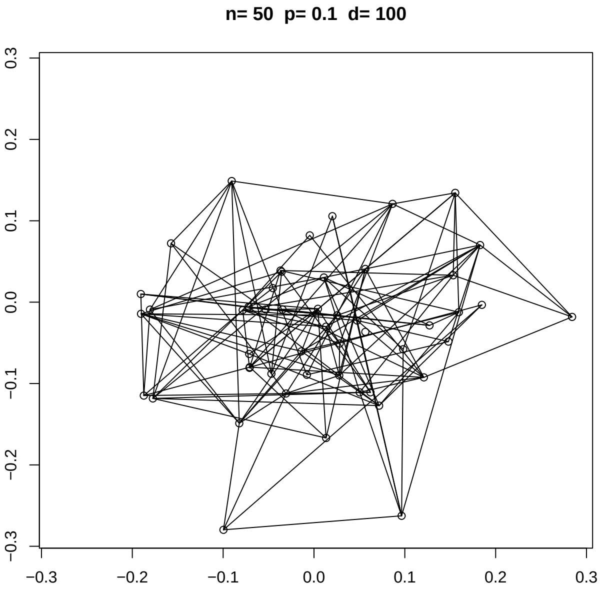
n= 50 p= 0.1 d= 5
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0

n= 50 p= 0.1 d= 10

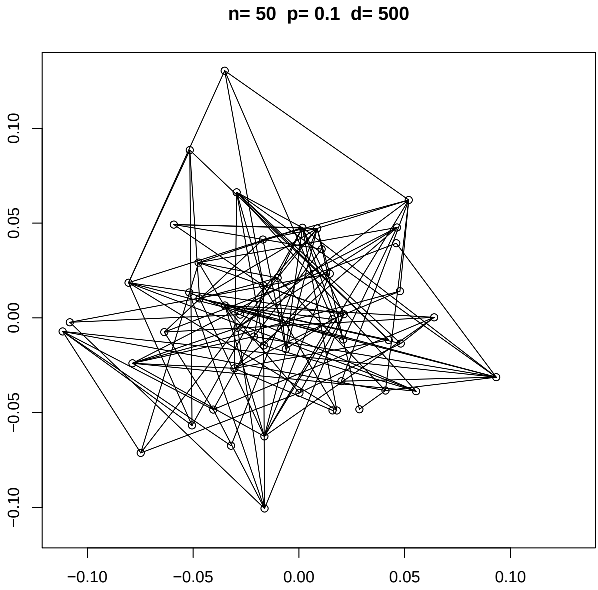




n= 50 p= 0.1 d= 30
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6



n= 50 p= 0.1 d= 100
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3


n= 50 p= 0.1 d= 500
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10

clique number of $\overline{\mathbf{G}}(\mathbf{n}, \mathbf{d}, \mathbf{p})$

For fixed $\mathbf{d}$ and $\mathbf{p}$, the clique number $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p})$ grows linearly with $\mathbf{n}$.
For $\mathbf{d} = \infty$ the behavior is very different:

$$\omega(\mathbf{n}, \infty, \mathbf{p}) = 2 \log_{1/\mathbf{p}} \mathbf{n} - 2 \log_{1/\mathbf{p}} \log_{1/\mathbf{p}} \mathbf{n} + \mathbf{O}(1).$$

How fast does $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p})$ approach the clique number of $\mathbf{G}(\mathbf{n}, \mathbf{p})$?

How large does $\mathbf{d}$ need to be for similar behavior?

clique number bounds

$\mathbf{p}$ is fixed, $\mathbf{n}$ grows.

if $\mathbf{d} \sim$ const., then $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p}) = \Omega_\mathbf{p}(\mathbf{n})$

if $\mathbf{d} \to \infty$, then $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p}) = o_\mathbf{p}(\mathbf{n})$

if $\mathbf{d} = o(\log \mathbf{n})$, then $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p}) \geq \mathbf{n}^{1-o_\mathbf{p}(1)}$

if $\mathbf{d} \sim \log^2 \mathbf{n}$, then $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p}) = O_\mathbf{p}(\log^3 \mathbf{n})$

if $\mathbf{d} \gg \log^3 \mathbf{n}$, then $\omega(\mathbf{n}, \mathbf{d}, \mathbf{p}) = (2 + o_\mathbf{p}(1)) \log_{1/\mathbf{p}} \mathbf{n}$

if $\mathbf{d} \sim \log^5 \mathbf{n}$, then
$\omega(\mathbf{n}, \mathbf{d}, \mathbf{p}) = 2 \log_{1/\mathbf{p}} \mathbf{n} - 2 \log_{1/\mathbf{p}} \log_{1/\mathbf{p}} \mathbf{n} + O_\mathbf{p}(1)$

proof ideas

first three statements are easy (from area estimate of a spherical cap)

proof ideas

first three statements are easy (from area estimate of a spherical cap)

third follows from Jung's theorem and Vapnik-Chervonenkis inequality

proof ideas

first three statements are easy (from area estimate of a spherical cap)

third follows from Jung's theorem and Vapnik-Chervonenkis inequality

Jung's theorem (1901): For every set $\mathbf{A} \subset \mathbb{R}^{\mathbf{d}}$ of diameter at most $\mathbf{1}$ there exists a closed ball of radius $\sqrt{\mathbf{d}/(\mathbf{2}(\mathbf{d}+\mathbf{1}))}$ that includes $\mathbf{A}$.

proof ideas

first three statements are easy (from area estimate of a spherical cap)

third follows from Jung's theorem and Vapnik-Chervonenkis inequality

Jung's theorem (1901): For every set $\mathbf{A} \subset \mathbb{R}^{\mathbf{d}}$ of diameter at most $\mathbf{1}$ there exists a closed ball of radius $\sqrt{\mathbf{d}/(\mathbf{2}(\mathbf{d}+\mathbf{1}))}$ that includes $\mathbf{A}$.

last two statements are the main result.

upper bound for $\mathbf{p = 1/2}$

$\mathbf{N_k}$ is the number of cliques of size $\mathbf{k}$. For $\mathbf{G(n, p)}$,

$$\mathbb{E}\mathbf{N_k} = \binom{n}{k} 2^{-\binom{k}{2}}$$

Let $\delta > 0$ and $\mathbf{K} > 2$. If

$$d \geq \frac{K^3}{\delta^2} \ ,$$

then, for $1 \leq k \leq K$,

$$\mathbb{E}\mathbf{N_k}(n, d, 1/2) \leq \binom{n}{k} \Phi(\delta)^{\frac{(k-1)(k-2)}{2}} \ .$$

upper bound for $p = 1/2$

N_k is the number of cliques of size k. For $G(n, p)$,

$$\mathbb{E}N_k = \binom{n}{k} 2^{-\binom{k}{2}}$$

Let $\delta > 0$ and $K > 2$. If

$$d \geq \frac{K^3}{\delta^2} \, ,$$

then, for $1 \leq k \leq K$,

$$\mathbb{E}N_k(n, d, 1/2) \leq \binom{n}{k} \Phi(\delta)^{\frac{(k-1)(k-2)}{2}} \, .$$

Follows from an inductive argument, using approximate orthogonality of $X_1, \ldots, X_n$.

clique number estimates

The upper bounds for $\omega(n, d, p)$ follow from the first moment method:

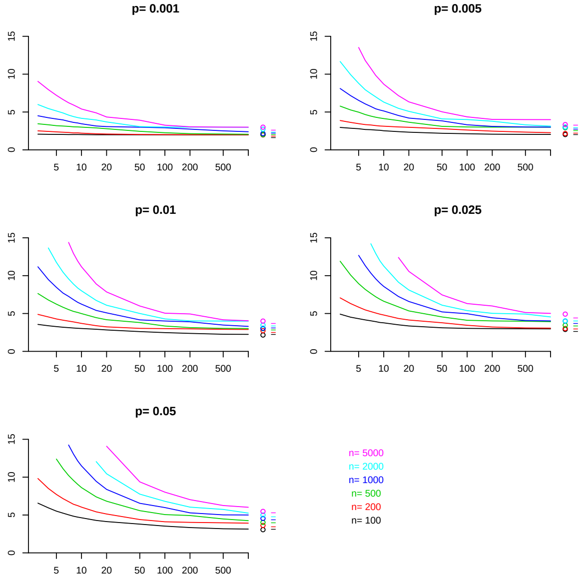
$$\mathbb{P}\{\omega(n, d, 1/2) \geq k\} = \mathbb{P}\{N_k \geq 1\} \leq \mathbb{E}N_k ,$$

Lower bounds for $\omega(n, d, p)$ follow from the second moment method.
First we prove a similar lower bound for $\mathbb{E}N_k$ and then show

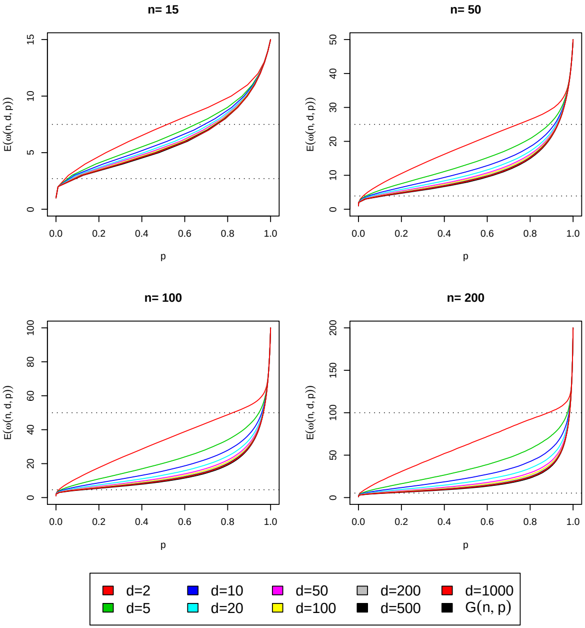
$$\frac{\mathrm{var}(N_k)}{(\mathbb{E}N_k)^2} \to 0$$

for the relevant values of k.

$\omega(\mathbf{n}, \mathbf{d}, \mathbf{p})$ as a function of $\mathbf{d}$



$\omega(\mathbf{n}, \mathbf{d}, \mathbf{p})$ as a function of $\mathbf{p}$



testing hidden dependencies

$\mathbf{Z_i = (Z_{i,1}, \ldots, Z_{i,d}), i = 1, \ldots, n}$.

Null hypothesis: all $\mathbf{Z_{i,t}}$ are i.i.d. standard normal.

testing hidden dependencies

$\mathbf{Z_i = (Z_{i,1}, \ldots, Z_{i,d}), i = 1, \ldots, n}$.

Null hypothesis: all $\mathbf{Z_{i,t}}$ are i.i.d. standard normal.

Alternative hypothesis: $\exists \mathbf{S} \subset \{\mathbf{1, \ldots, n}\}$ with $|\mathbf{S}| \geq \mathbf{m}$ such that

$$\mathbf{Z_{i,t}} = \begin{cases} \mathbf{N_{i,t}} & \text{if } \mathbf{i} \notin \mathbf{S} \\ (\mathbf{N_{i,t} + Y_t})/\sqrt{\mathbf{1} + \sigma^2} & \text{if } \mathbf{i} \in \mathbf{S} \end{cases}$$

where the $\mathbf{N_{i,t}}$ are i.i.d. standard normal and the $\mathbf{Y_t}$ are independent normal $(\mathbf{0}, \sigma^2)$.

test

Define $\mathbf{X_i = Z_i/\|Z_i\|}$ and form the graph $\overline{\mathbf{G}}\mathbf{(n, d, 1/2)}$.
accept the null hypothesis if and only if $\omega\mathbf{(n, d, 1/2) \leq 3\log_2 n}$.

test

Define $\mathbf{X_i = Z_i/\|Z_i\|}$ and form the graph $\overline{\mathbf{G}}\mathbf{(n, d, 1/2)}$.
accept the null hypothesis if and only if $\omega\mathbf{(n, d, 1/2) \leq 3\log_2 n}$.

$\exists \mathbf{C}, \epsilon_\mathbf{n} \rightarrow \mathbf{0}$ such that if

$$\mathbf{d \geq C \max\left(\frac{\ln m}{\sigma^4}, \log_2^3 n\right) \quad \text{and} \quad m > 3\log_2 n}$$

then the test errs with probability $< \epsilon_\mathbf{n}$ under both the null and alternative hypotheses.

test

Define $\mathbf{X_i = Z_i/\|Z_i\|}$ and form the graph $\overline{\mathbf{G}}\mathbf{(n, d, 1/2)}$.
accept the null hypothesis if and only if $\omega\mathbf{(n, d, 1/2) \leq 3 \log_2 n}$.

$\exists \mathbf{C}, \epsilon_\mathbf{n} \to \mathbf{0}$ such that if

$$\mathbf{d \geq C \max\left(\frac{\ln m}{\sigma^4}, \log_2^3 n\right) \quad \text{and} \quad m > 3 \log_2 n}$$

then the test errs with probability $< \epsilon_\mathbf{n}$ under both the null and alternative hypotheses.

This test is computationally very expensive.

questions

sharper bounds for the value of $\mathbf{d}$?

conjecture: $\mathbb{E}\omega(\mathbf{n}, \mathbf{d}, \mathbf{p})$ is nonincreasing in $\mathbf{d}$ for fixed $\mathbf{n}, \mathbf{p}$.

when does two-point concentration kick in?

connectivity threshold? giant component?

a computationally efficient test? (related to hidden clique problem of Alon, Krivelevich, and Sudakov).