Large Deviations in Discrete Random Structures

Yufei Zhao

(Oxford → MIT)

June 2017

The upper tail problem

Let X be the number of triangles in the Erdős–Rényi random graph $\mathbf{G}(n,p)$.

$$\mathbb{E}X = \binom{n}{3} p^3$$

Central Limit Theorem (Ruciński ’88): X is asymptotially normal, i.e.,

$$\frac{X - \mathbb{E}X}{\sqrt{\mathrm{Var}\, X}} \to \text{Normal}, \qquad \text{as } n \to \infty, \text{provided } np \to \infty, n(1-p) \to \infty$$

Problem: Estimate $\mathbb{P}(X \geq (1+\delta)\mathbb{E}X)$ (fixed $\delta > 0$)

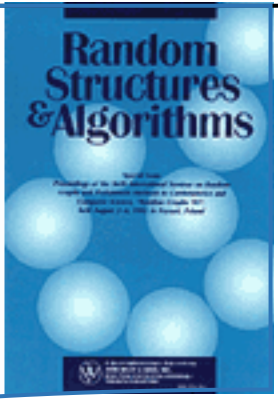
X = # triangles in **G**(n,p). $\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = ?$

[redacted]

The Infamous Upper Tail

Random Structures & Algorithms 2002

Svante Janson, Andrzej Ruciński



Janson, Oleszkiewicz, Rucinski ’04
Bollobás ’81, ’85
Janson, Luczak, Rucinski ’02, ’04
Vu ’01
Kim & Vu ’04
Chatterjee & Dey ’10

Order of
$$\log \mathbb{P}(X \geq (1+\delta)\mathbb{E}X)$$
independently determined by

DeMarco & Kahn ’11
and
Chatterjee ’11

Order of the rate

Theorem (DeMarco—Kahn '11, Chatterjee '11).
Let X denote the number of triangles in $\mathbf{G}(n,p)$.
Fix $\delta > 0$. For $p \gtrsim (\log n)/n$,

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{\Theta_\delta(n^2p^2)}$$

Proof of lower bound:

Force a clique on $m = \Theta_\delta(np)$ vertices

Obtain $\binom{m}{3} \geq (1+\delta)\binom{n}{3}p^3$ triangles

Occurs with probability $p^{\binom{m}{2}} = p^{\Theta_\delta(n^2p^2)}$

Large deviations

- Determine the exact asymptotics of
$$\log \mathbb{P}(X \geq (1+\delta)\mathbb{E}X)$$
- The “main reason” for large deviation?
- Conditioned on $X \geq (1+\delta)\mathbb{E}X$, what does a typical instance look like?

Question (Chatterjee—Varadhan ’11).
Fix $0 < p < q < 1$. Let G be an instance of $\mathbf{G}(n,p)$ conditioned on having at least as many triangles as a typical $\mathbf{G}(n,q)$.
$\underbrace{\text{Is } G \approx \mathbf{G}(n,q) \text{ in cut-distance?}}$

$\boxed{e_G(U) = \binom{|U|}{2}q + o(n^2) \text{ for every } U \subset V.}$

Possibilities:

- **Yes**: more edges, uniformly distributed (replica symmetry)
- **No**: some other non-uniform distribution of edges (symmetry breaking)

Does **G**(n,p), conditioned on having $\geq \binom{n}{3}q^3$ triangles, look like **G**(n,q)?

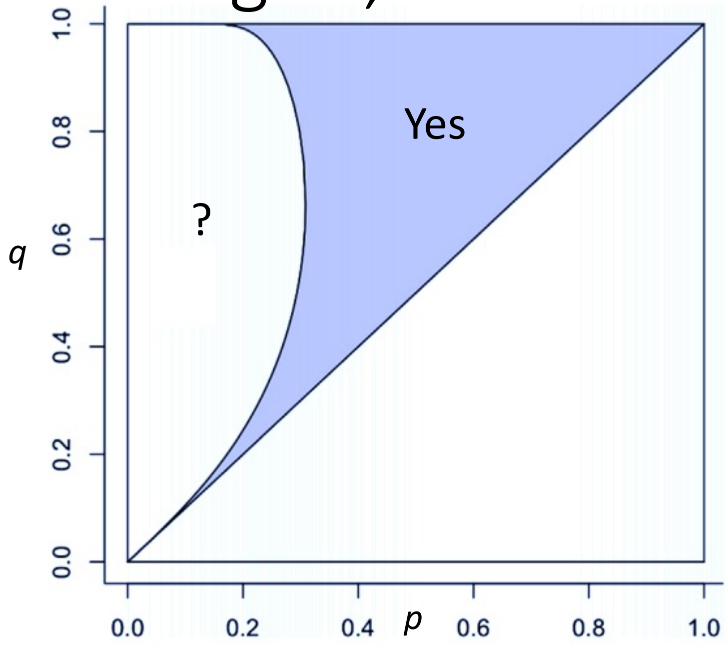
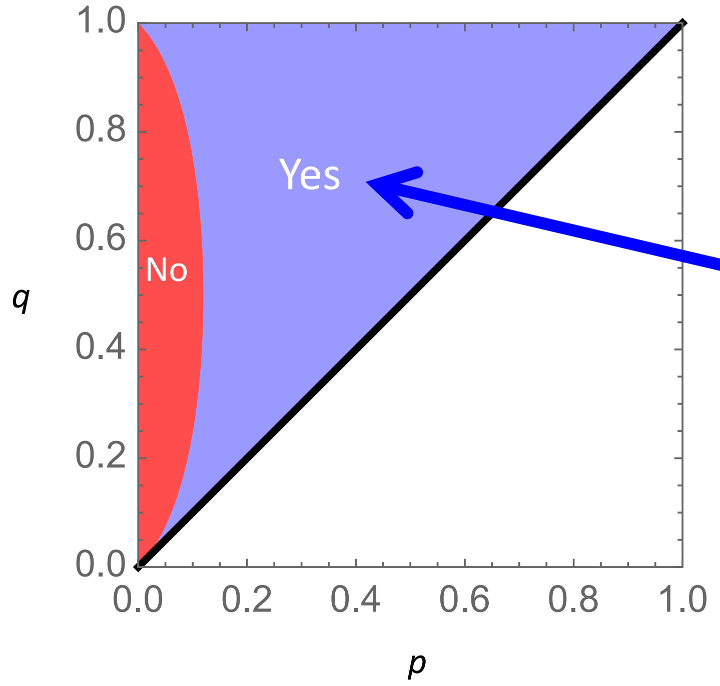


Figure from
[Chatterjee & Dey ’10]

[Chatterjee & Dey ’10]
- Stat. phys. methods
- Stein’s method

[Chatterjee & Varadhan ’11]
- Szemerédi’s regularity lemma
- Graph limits

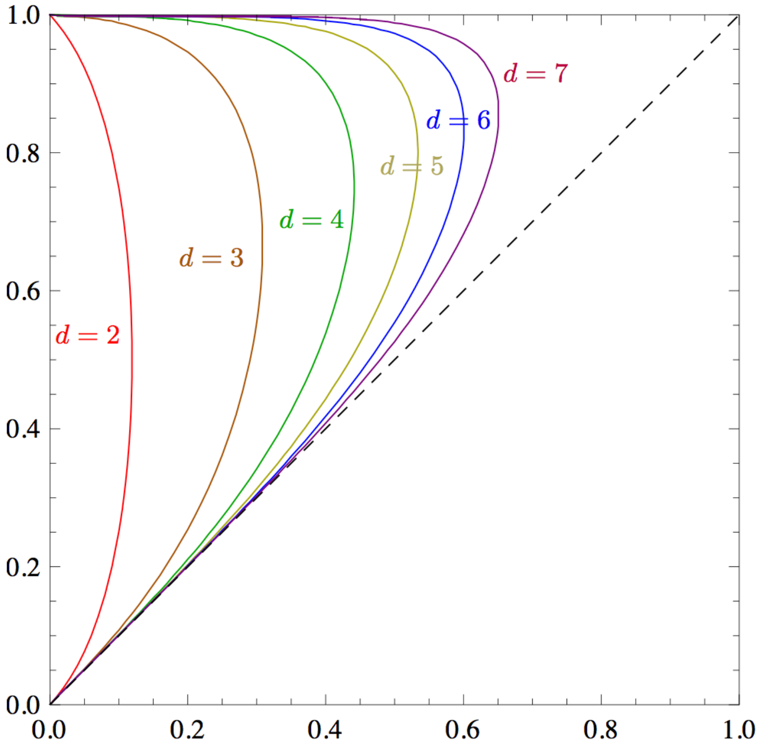
Does **G**(n,p), conditioned on having $\geq \binom{n}{3} q^3$ triangles, look like **G**(n,q)?



Theorem (Lubetzky—Z. '15).
Replica symmetry phase:

$$p \geq \left(1 + (q^{-1} - 1)^{1/(1-2q)}\right)^{-1}$$

Upper tail of *H*-density



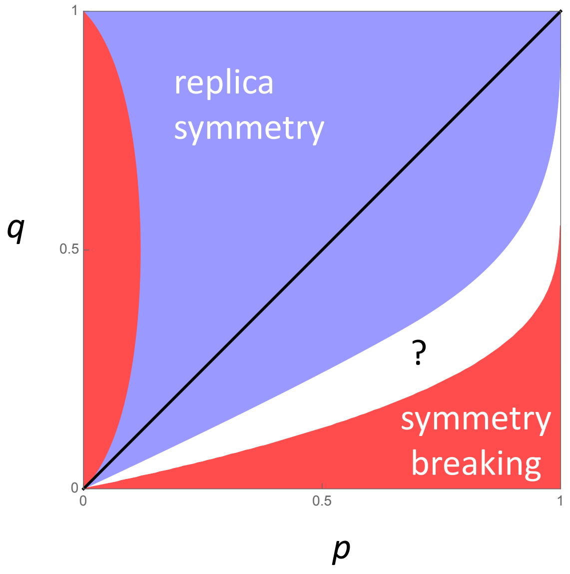
[Lubetzky—Z. ’15] Identified the phase diagram for *H*-density if *H* is *d*-regular.

The phase diagram depends only on *d*.

Also: upper tail large deviation of the top eigenvalue of **G**(n,p).
(Top eigvalue typically ≈ np; what if ≥ nq?)
Same diagram as d = 2

Open: any irregular *H*,
e.g., a path of two edges

Lower tail



[Z. 2017]

$X \leq (1-\delta)\mathbb{E}X$ as $p \to 0$

δ = 0.01 Replica symmetry

δ^* critical ???

δ = 0.99 Symmetry breaking

Review of large deviations

Fixed $0 < p < q < 1$.

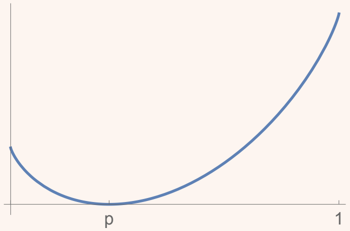
$X \sim \text{Binomial}(n, p)$. $\mathbf{P}(X \geq nq) = ??$

$$\boxed{\log \mathbb{P}(X \geq nq) = -(I_p(q) + o(1))n \qquad \text{as } n \to \infty}$$

"cost of tilting"

Relative entropy (KL divergence):

$$I_p(x) := x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}$$



Triangles in $\mathbf{G}(n,p)$

- The number of triangles in $\mathbf{G}(n,p)$ is a sum of **dependent** indicator random variables

- Heuristics for the natural variational problem

Triangles in G(n,p)

For each pair (i, j) of vertices

- Tilt its probability to some $q_{ij} \geq p$
- Pay $I_p(q_{ij})$ cost in log probability.

Objective: minimize relative entropy cost $\min \sum_{1 \leq i < j \leq n} I_p(q_{ij})$

Constraint: enough triangles $\sum_{1 \leq i < j < k \leq n} q_{ij} q_{ik} q_{jk} \geq \binom{n}{3} q^3$

This actually works! The minimum is asymptotically $-\log \mathbb{P}(X \geq \binom{n}{3} q^3)$

Chatterjee—Varadhan ’11	dense setting: p constant
Chatterjee—Dembo ’16	sparse setting: $p \geq n^{-1/42} \log n$
Eldan ’17+	improved: $p \geq n^{-1/18} \log n$

Compact formulation

$W(x,y) = W(y,x)$

- A graphon is a symmetric measurable function $W: [0,1]^2 \to [0,1]$.

Discrete variational problem

Minimize $\sum_{1\le i<j\le n} I_p(q_{ij})$

Subj to

$$\sum_{1\le i<j<k\le n} q_{ij}q_{ik}q_{jk} \ge \binom{n}{3} q^3$$

Graphon variational problem [Chatterjee—Varadhan]

Minimize $\int_{[0,1]^2} I_p(W(x,y))\, dxdy$

Subj to

$$\int_{[0,1]^3} W(x,y)W(x,z)W(y,z)\, dxdydz \ge q^3$$

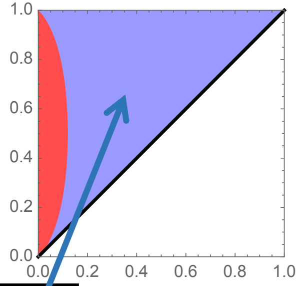
- Due to compactness of the space of graphons under cut metric (Lovasz—Szegedy), the above minimum is always attained
- Though in general we don't know how to solve the variational problem

What do the minimizing graphons represent?

The set of relative entropy minimizing graphons represents the most likely graphs conditioned on the rare event.

Replica symmetry:
Is it minimized (uniquely) by the constant graphon?
If so, the conditioned random graph is close to Erdős–Rényi (in cut distance).

Upper tail phase diagram

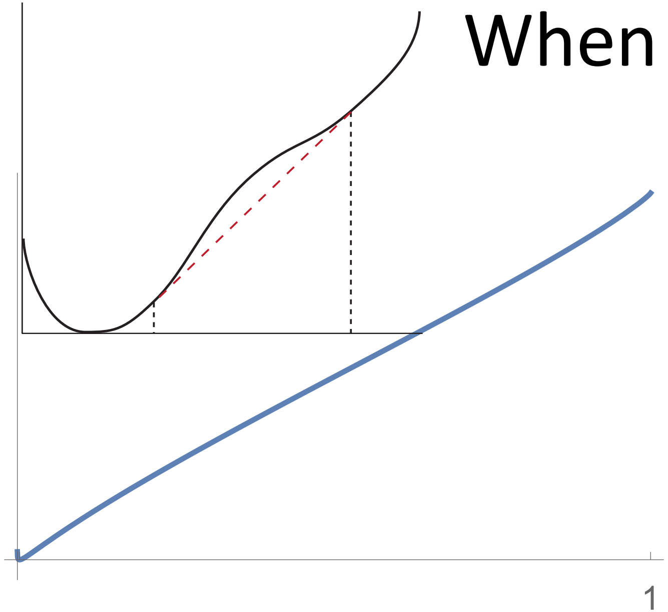


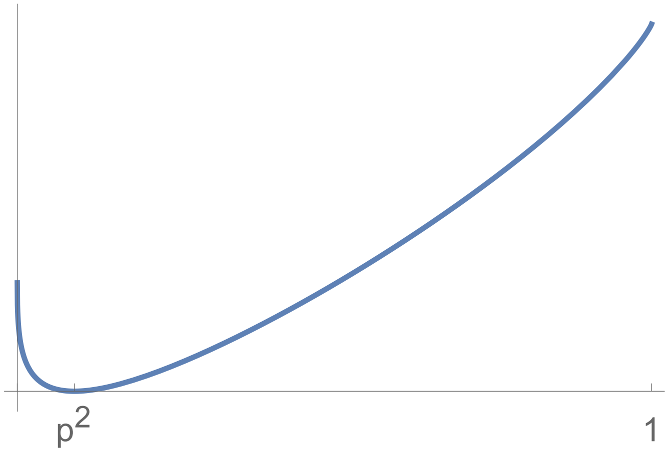
Theorem (Lubetzky—Z. '15)**.**
Let $0 < p < q < 1$. The constant graphon $W \equiv q$ minimizes $\int_{[0,1]^2} I_p\big(W(x,y)\big)\,dxdy$ subject to

$$\int_{[0,1]^3} W(x,y)W(x,z)W(y,z)\,dxdydz \geq q^3$$

if and only if the point $(q^2, I_p(q))$ lies on the convex minorant of $x \mapsto I_p(\sqrt{x})$.

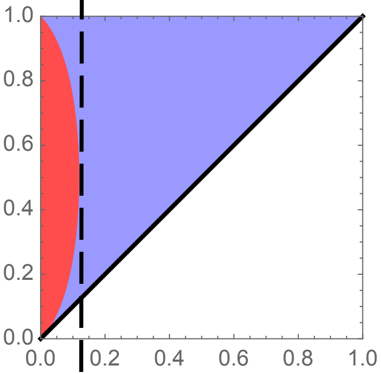
When is $x \mapsto I_p(\sqrt{x})$ convex?





Not convex for

$$p < \frac{1}{1+e^2}$$



Always convex for

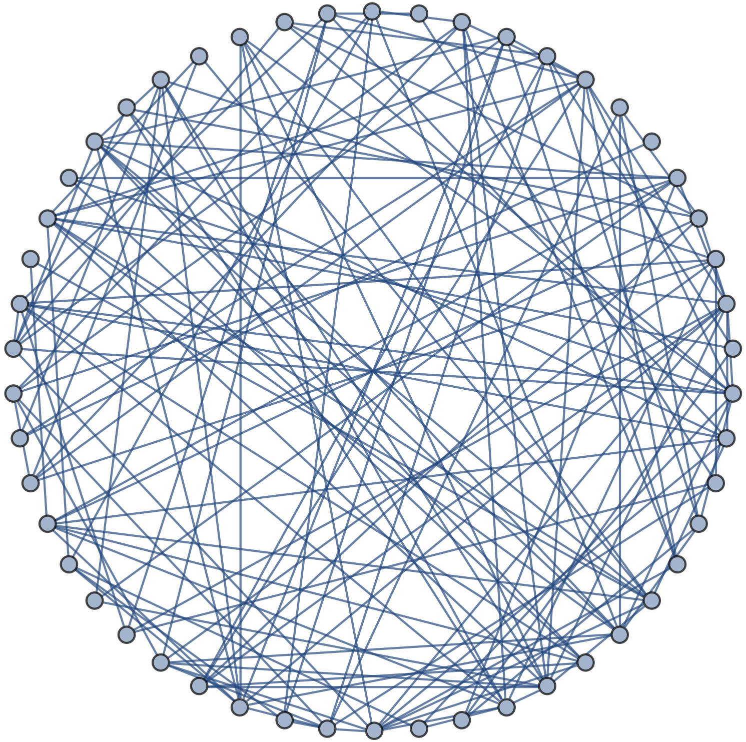
$$p \geq \frac{1}{1+e^2} \approx 0.12$$

Sparse setting

G(n,p)

$p = p_n \to 0$ as $n \to \infty$,

perhaps slowly



Order of the rate

Theorem (DeMarco—Kahn ’11, Chatterjee ’11)**.**
Let *X* denote the number of triangles in **G**(*n*,*p*).
Fix $\delta > 0$. For $p \gtrsim (\log n)/n$,

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{\Theta_\delta(n^2p^2)}$$

Proof of lower bound:

Force a clique on $m = \Theta_\delta(np)$ vertices

Obtain $\binom{m}{3} \geq (1+\delta)\binom{n}{3}p^3$ triangles

Occurs with probability $p^{\binom{m}{2}} = p^{\Theta_\delta(n^2p^2)}$

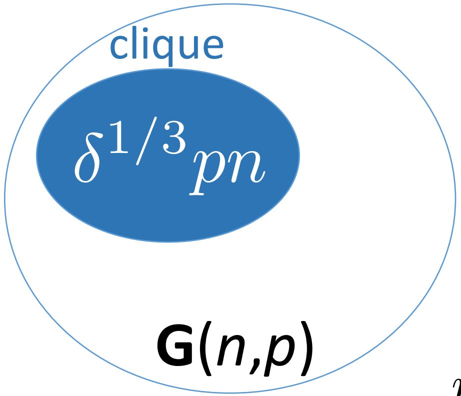
Theorem (Chatterjee—Dembo/Eldan + Lubetzky—Z.).

Let X denote the number of triangles in $\mathbf{G}(n,p)$.

Fix $\delta > 0$. With $p \to 0$ and and $p \geq n^{-1/18} \log n$,

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{(1+o(1))\min\left\{\frac{1}{2}\delta^{2/3}, \frac{1}{3}\delta\right\} p^2 n^2}$$

Proof of lower bound:



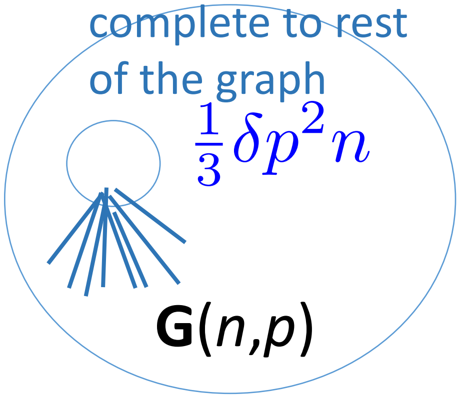
$\sim \delta p^3 \binom{n}{3}$

extra triangles

With probability:

$p^{(1+o(1))\frac{1}{2}\delta^{2/3}p^2n^2}$

Preferred for δ > 27/8



$\sim \delta p^3 \binom{n}{3}$

extra triangles

With probability:

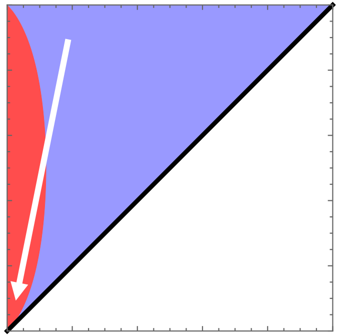
$p^{(1+o(1))\frac{1}{3}\delta p^2n^2}$

Preferred for δ < 27/8

Nonlinear large deviations

Theorem (Chatterjee—Dembo)**.** For sparse random graphs, the large deviation problem reduces to the natural variational problem,
provided that $p > n^{-\alpha}$ for some explicit $\alpha > 0$.

Theorem (Lubetzky—Z.)**.** Provided $p \to 0$ and $p\sqrt{n} \to \infty$, the value of the variational problem for triangles is $\left(1 + o(1)\right) \min\left\{\frac{1}{2}\delta^{2/3}, \frac{1}{3}\delta\right\} p^2 \log \frac{1}{p}$.



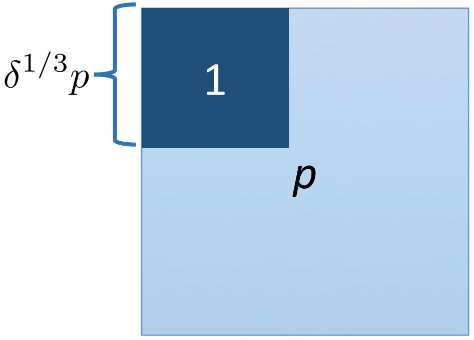
Theorem (Chatterjee—Dembo/Eldan + Lubetzky—Z.).
Let X denote the number of triangles in $\mathbf{G}(n,p)$.
Fix $\delta > 0$. With $p \to 0$ and and $p \geq n^{-1/18} \log n$,

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{(1+o(1))\min\left\{\frac{1}{2}\delta^{2/3}, \frac{1}{3}\delta\right\} p^2 n^2}$$

Similarly for the number of K_t

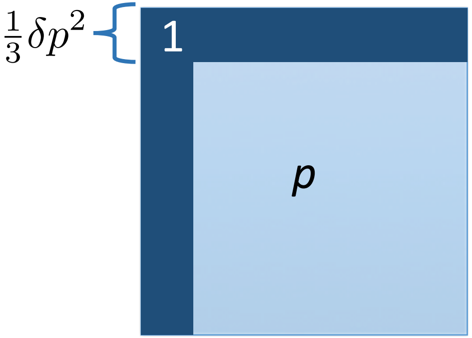
[Bhattacharya, Ganguly, Lubetzky, Z.]
Solution for every H

Proof of lower bound:



$(1+o(1))\delta p^3 \binom{n}{3}$
extra triangles

With probability:
$p^{(1+o(1))\frac{1}{2}\delta^{2/3}p^2n^2}$



$(1+o(1))\delta p^3 \binom{n}{3}$
extra triangles

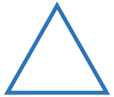
With probability:
$p^{(1+o(1))\frac{1}{3}\delta p^2 n^2}$

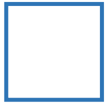
Theorem (Bhattacharya, Ganguly, Lubetzky, Z.).
Fix $\delta > 0$ and a graph H. Let X_H = # copies of H in $\mathbf{G}(n,p)$.
With $p \to 0$ and and $p \geq n^{-1/6e(H)} \log n$,

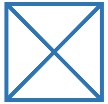
$$\mathbb{P}(X_H \geq (1+\delta)\mathbb{E}X_H) = p^{(c_H(\delta)+o(1))p^\Delta n^2}$$

where Δ = max deg H, and $c_H(\delta) > 0$ is an explicit constant …

For example

For $H = C_3$  $c_H(\delta) = \min\{\frac{1}{2}\delta^{2/3}, \frac{1}{3}\delta\}$

For $H = C_4$  $c_H(\delta) = \min\left\{\frac{1}{2}\delta^{1/2}, -1 + \left(1 + \frac{1}{2}\delta\right)^{1/2}\right\}$

For $H = K_4$  $c_H(\delta) = \min\{\frac{1}{2}\delta^{1/2}, \frac{1}{4}\delta\}$

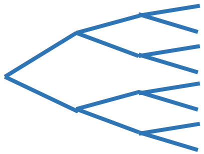
Theorem (Bhattacharya, Ganguly, Lubetzky, Z.).
Fix $\delta > 0$ and a graph H. Let X_H = # copies of H in $\mathbf{G}(n,p)$.
With $p \to 0$ and and $p \geq n^{-1/6e(H)} \log n$,

$$\mathbb{P}(X_H \geq (1+\delta)\mathbb{E}X_H) = p^{(c_H(\delta)+o(1))p^{\Delta}n^2}$$

where Δ = max deg H, and $c_H(\delta) > 0$ is an explicit constant …

For example

For $H = K_{2,3}$  $c_H(\delta) = (1+\delta)^{1/2} - 1$

For $H =$  $c_H(\delta) = -\frac{3}{2} + \frac{1}{2}\sqrt{5 + 4\sqrt{1+\delta}}$

Theorem (Bhattacharya, Ganguly, Lubetzky, Z.).
Fix $\delta > 0$ and a graph H. Let X_H = # copies of H in $\mathbf{G}(n,p)$.
With $p \to 0$ and and $p \geq n^{-1/6e(H)} \log n$,

$$\mathbb{P}(X_H \geq (1+\delta)\mathbb{E}X_H) = p^{(c_H(\delta)+o(1))p^{\Delta}n^2}$$

where Δ = max deg H, and $c_H(\delta) > 0$ is an explicit constant …

Independence polynomial: $P_H(x) \coloneqq \sum_{\text{indep set } I} x^{|I|}$

Let H^* denote the subgraph of H induced by its maximum degree vertices.

Let $\theta > 0$ satisfy $P_{H^*}(\theta) = 1 + \delta$. Then, for a connected graph H,

$$c_H(\delta) = \begin{cases} \min\{\theta, \frac{1}{2}\delta^{2/v(H)}\} & \text{if } H \text{ is regular} \\ \theta & \text{if } H \text{ is irregular} \end{cases}$$

Arithmetic progressions

The exponent on n can be improved to 1/18 for k=3 and 1/48 for k=4.

Theorem (Bhattacharya, Ganguly, Shao, Z.).
Fix k and $\delta > 0$. Let X_k denote the number of k-term arithmetic progressions in a random subset of {1, 2, …, N} where every element is included with probability p. With $p \to 0$ and $p \geq n^{-1/6k(k-1)} \log n$,

$$\mathbb{P}(X_k \geq (1+\delta)\mathbb{E}X_k) = p^{(1+o(1))\sqrt{\delta p^k n^2}}$$

The order in the exponent was determined by Warnke, and holds for all

$$p \gtrsim \left(\frac{\log n}{n}\right)^{1/(k-1)}$$

- Proof of lower bound: plant an interval of length $\sim \sqrt{\delta p^k n^2}$
- Chatterjee—Dembo developed a LDP for 3-term APs.
  We give a LDP for longer APs, building on Eldan's work
- The effort to widen the range of p led to issues in additive combinatorics

Complexity of AP-dual functions

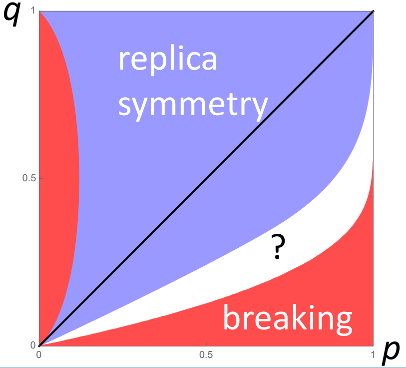
Via discrete Fourier analysis on $\mathbb{Z}/N\mathbb{Z}$: for any $f, g : \mathbb{Z}/N\mathbb{Z} \to [-1, 1]$,

$$\mathbb{E}_{x,y\in\mathbb{Z}/N\mathbb{Z}} f(x)g(x+y)g(x+2y) \le \sup_{r\in\mathbb{Z}/N\mathbb{Z}} \left|\mathbb{E}_{x\in\mathbb{Z}/N\mathbb{Z}} f(x)e^{2\pi i\, rx/N}\right|$$

Conjecture: There is a set Φ of N^C many bounded functions such that for any $f, g : \mathbb{Z}/N\mathbb{Z} \to [-1, 1]$,

$$\mathbb{E}_{x,y\in\mathbb{Z}/N\mathbb{Z}} f(x)g(x+y)g(x+2y)g(x+3y) \le \sup_{\phi\in\Phi} \left|\mathbb{E}_{x\in\mathbb{Z}/N\mathbb{Z}} f(x)\phi(x)\right|$$

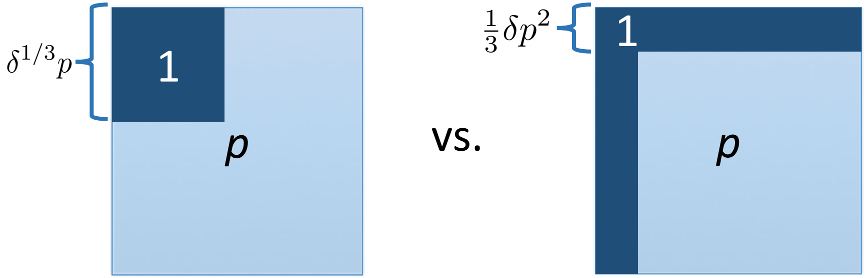
Dense setting:



Problem: improve this!

Sparse setting: $\underbrace{n^{-1/18}}\log n \le p \to 0$

$$\mathbb{P}(X \ge (1+\delta)\mathbb{E}X) = p^{(1+o(1))\min\{\frac{1}{2}\delta^{2/3}, \frac{1}{3}\delta\}\, p^2 n^2}$$



References:

Chatterjee, Varadhan (2011), The large deviation principle for the Erdős-Rényi random graph

Chatterjee, Dembo (2016), Nonlinear large deviations

Eldan (2017+), Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations

----

Lubetzky, Zhao (2015), On replica symmetry of large deviations in random graphs

Lubetzky, Zhao (2017), On the variational problem for upper tails in sparse random graphs

Zhao (2017), On the lower tail variational problem for random graphs

Bhattacharya, Ganguly, Lubetzky, Zhao (2015+), Upper tails and independence polynomials in random graphs

Bhattacharya, Ganguly, Shao, Zhao (2017+), Upper tails for arithmetic progressions in a random set

See Chatterjee's Bulletins AMS survey article and upcoming Saint-Flour monograph on large deviations in random graphs.