Isomorphisms between dense random graphs

Lutz Warnke

UC San Diego

Joint work with *Erlang Surya* and *Emily Zhu* (UC San Diego)

Context

Fundamental Problem

Is an induced copy of F (or a large part of F) contained in G?

- Variant of 'Subgraph Containment Problem'
- Relevant in Applications: Pattern Recognition, Computer vision, etc
- Many heuristic algorithms (NP-complete)

Today

Random variants of this problem: F and G independent random graphs

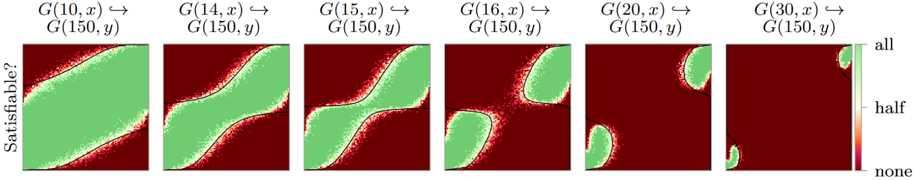
- When does induced copy of G_{n,p_1} appear in G_{N,p_2}? How many copies?
- Size of largest common induced subgraph of G_{N,p_1} and G_{N,p_2}?
- Difficult benchmark problem for algorithms

Part I: Why induced containment of G_{n,p_1} in G_{N,p_2}?

C. McCreesh, P. Prosser, C. Solnon, and J. Trimble (2018)

Deciding $G_{n,p_1} \sqsubseteq G_{N,p_2}$ is difficult benchmark problem for algorithms

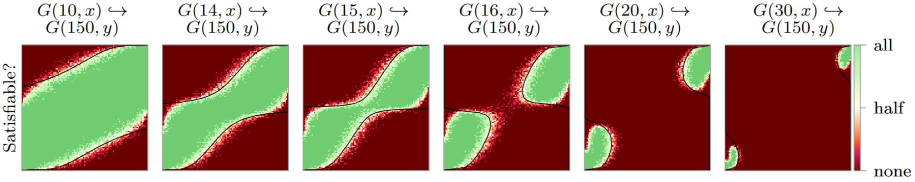
Empirically discovered interesting phase transition diagram:



Interest in Combinatorics and Probability

- **Knuth**: asked for mathematical explanation
- **Chatterjee–Diaconis**: explained middle-points $p_1 = p_2 = 1/2$
- **This talk**: we explain all $(p_1, p_2) \in (0,1)^2$

When induced copy appears: previous work (uniform case)



We write $H \sqsubseteq G$ if G contains an induced copy of H

Chatterjee-Diaconis (2021)

$$\lim_{N\to\infty} \mathbb{P}\left(G_{n,1/2} \sqsubseteq G_{N,1/2}\right) = \begin{cases} 1 & \text{if } n \leq 2\log_2 N + 1 - \varepsilon_N \\ 0 & \text{if } n \geq 2\log_2 N + 1 + \varepsilon_N \end{cases}$$

- Proof uses first and second moment method:
  - $X =$ Number of induced copies of $G_{n,1/2}$ in $G_{N,1/2}$
- Does not extend to $G_{n,p_1} \sqsubseteq G_{N,p_2}$ when $p_2 \neq 1/2$:
  - Second moment method fails due to large variance: $\operatorname{Var} X \gg (\mathbb{E}X)^2$

When induced copy appears: new result (general case)

Appearance of induced copy of G_{n,p_1} in G_{N,p_2} (Surya-W.-Zhu, 2023+)

Let $p_1, p_2 \in (0,1)$ be constants. Define $a := 1/\left(p_2^{p_1}(1-p_2)^{1-p_1}\right)$. Then

- Uniform case: if $p_2 = 1/2$, then $a = 2$ and

$$\lim_{N\to\infty} \mathbb{P}\left(G_{n,p_1} \sqsubseteq G_{N,p_2}\right) = \begin{cases} 1 & \text{if } n \leq 2\log_a N + 1 - \varepsilon_N, \\ 0 & \text{if } n \geq 2\log_a N + 1 + \varepsilon_N. \end{cases}$$

- Nonuniform case: if $p_2 \neq 1/2$, then

$$\lim_{N\to\infty} \mathbb{P}\left(G_{n,p_1} \sqsubseteq G_{N,p_2}\right) = \begin{cases} 1 & \text{if } n - (2\log_a N + 1) \to -\infty, \\ f(c) & \text{if } n - (2\log_a N + 1) \to c, \\ 0 & \text{if } n - (2\log_a N + 1) \to \infty, \end{cases}$$

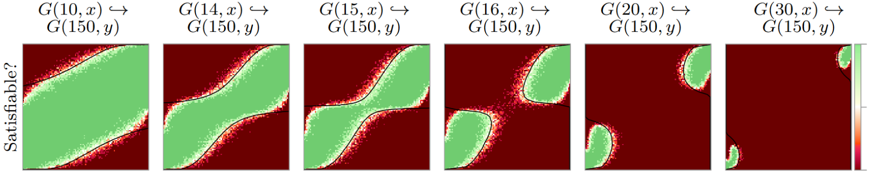
where $f(c) := \mathbb{P}\left(\mathrm{N}(0,\sigma^2) \geq c\right)$ with $\sigma = \sigma(p_1, p_2)$

- Sharpness of phase transition differs for $p_2 = 1/2$ and $p_2 \neq 1/2$

When induced copy appears: new result (remarks)

Remarks

- Confirms simulation based predictions:



- Answers question of Chatterjee-Diaconis
- Difference to size of largest clique in G_{N,p_2} (differs by additive $\Theta(\log\log N)$ due to size of automorphism group)
- Deviation in edge-count $e(G_{n,p_1})$ causes large variance when $p_2 \neq 1/2$ (responsible for different 'sharpness' when $p_2 = 1/2$ and $p_2 \neq 1/2$)

Proof overview $p_2 \neq 1/2$: number of edges of G_{n,p_1} matters

For pseudorandom property $\mathcal{P}$ (controls automorphisms of subgraphs etc):

$$\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2}) \approx \sum_{H \in \mathcal{P}} \mathbb{P}(G_{n,p_1} = H)\mathbb{P}(H \sqsubseteq G_{N,p_2})$$

If $n = 2\log_a N + 1 + c$ and H has $e(H) = p_1\binom{n}{2} + \delta n$ edges, then

$$\mathbb{E}X_H = (N)_n \cdot p_2^{e(H)}(1-p_2)^{\binom{n}{2}-e(H)} \approx \left[\left(\frac{p_2}{1-p_2}\right)^{\delta} a^{-c}\right]^n$$

so edge-deviation δn determines whether $\mathbb{E}X_H \to \infty$, which via second moment method (work!) implies $\mathbb{P}(H \sqsubseteq G_{N,p_2}) \to 1$. CLT then gives

$$\begin{aligned}\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2}) &\approx \sum_{H \in \mathcal{P}} \mathbb{P}(G_{n,p_1} = H)\mathbb{1}_{\{e(H) \geq p_1\binom{n}{2} + \delta_c n\}} \\ &\approx \mathbb{P}\Big(e(G_{n,p_1}) \geq p_1\binom{n}{2} + \delta_c n\Big) \approx f(c)\end{aligned}$$

How many copies: Asymptotic distribution

$X =$ Number of induced copies of G_{n,p_1} in G_{N,p_2}

Uniform case: Asymptotically Poisson

If $p_2 = 1/2$ and $n \geq 2\log_a N - 1 + \varepsilon_N$, then $\mathrm{d_{TV}}(X, \mathrm{Po}(\mu)) \to 0$.

By Stein-Chen method and pseudorandomness

Nonuniform case: 'squashed' log-normal

If $p_2 \neq 1/2$ and $n - (2\log_a N - 1) \to c$, then

$$\frac{\log(1+X)}{\log N} \xrightarrow{\mathrm{d}} \mathrm{SN}(-c, \sigma^2)$$

for a 'squashed' log-normal distribution $\mathrm{SN}(\mu, \sigma^2)$ with $\sigma = \sigma(p_1, p_2)$, i.e., with cumulative distribution function $F(x) := \mathbb{1}_{\{x \geq 0\}} \mathbb{P}(\mathrm{N}(\mu, \sigma^2) \leq x)$.

By second moment method and conditioning on number of edges $e(G_{n,p_1})$

Proof ingredient: Pseudorandom Properties

In Second Moment Calculation we restrict to pseudorandom H:

- Every large induced subgraph of H has trivial automorphism group
- Edges in every large subgraph of H are 'super-concentrated'

Difference between $G_{n,m}$ and $G_{n,p}$ matters

Edges of uniform $G_{n,m}$ are '*more concentrated*' than of binomial $G_{n,p}$

Example: for all vertex-subsets $S \subseteq [n]$, writing $p = m/\binom{n}{2}$ we have

$$\left| e(G_{n,m}[S]) - \binom{|S|}{2} p \right| \le n^{2/3}(n - |S|),$$

while for sets S of size $|S| = n - o(n^{1/3})$ we expect that

$$\left| e(G_{n,p}[S]) - \binom{|S|}{2} p \right| \ge \Omega\Big(|S| \sqrt{p(1-p)} \Big) = \Theta(n) \gg n^{2/3}(n - |S|)$$

Part II: Another induced containment variant

So far: when does induced copy of G_{n,p_1} appear in G_{N,p_2}?

Now: largest part of G_{n,p_2} that appears as induced copy of G_{N,p_2}

Size of largest (#vertex) common induced subgraph of G_{N,p_1} and G_{N,p_2}?

- Considered by Chatterjee–Diaconis in uniform case $p_1 = p_2 = 1/2$: motivated by fact that two infinite Rado graphs $G_{\infty,1/2}$ are isomorphic
- Natural question (should have been asked 30+ years ago!)

Two point concentration: largest common induced subgr.

I_N = size of largest common induced subgraph of G_{N,p_1} and G_{N,p_2}

Chatterjee-Diaconis (2021): uniform case

For $p_1 = p_2 = 1/2$, I_N is concentrated on two values around

$$4\log_2 N - 2\log_2\log_2 N - 2\log_2(4/e) + 1$$

Surya-Warnke-Zhu (2023+): general case

For constants $p_1, p_2 \in (0,1)$, I_N is concentrated on two values around

$$\max_{p\in[0,1]} \min\left\{x_N^{(0)}(p),\ x_N^{(1)}(p),\ x_N^{(2)}(p)\right\},$$

where for some b_0, b_1, b_2 depending on p_1, p_2 we have

$$x_N^{(0)}(p) = 4\log_{b_0} N - 2\log_{b_0}\log_{b_0} N - 2\log_{b_0}(4/e) + 1,$$

$$x_N^{(i)}(p) := 2\log_{b_i} N - 2\log_{b_i}\log_{b_i} N - 2\log_{b_i}(2/e) + 1.$$

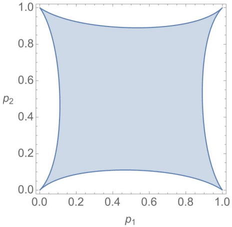
Failure of (naive) first moment prediction

X_n= # of pairs of common induced n-vertex subgraphs of G_{N,p_1} and G_{N,p_2}

First moment prediction (heuristic) for 'correct' vertex-size n

- $\mathbb{E}X_n \ll 1$ implies $\mathbb{P}(X_n = 0) \to 1$
- $\mathbb{E}X_n \gg 1$ implies $\mathbb{P}(X_n \geq 1) \to 1$

- Chatterjee and Diaconis confirmed prediction when $p_1 = p_2 = 1/2$
- We proved that prediction is only true in the following (p_1, p_2) region:

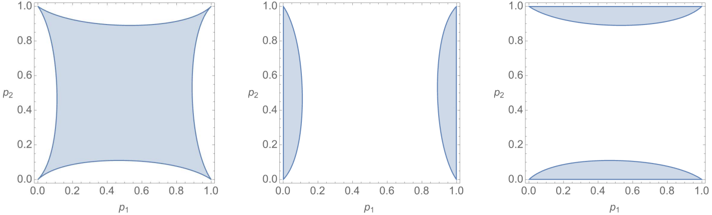


- Outside that region second moment method fails due to large variance

Form of answer: why optimize over three different terms?

Graph H fails to appear in G_{N,p_1} and G_{N,p_2}:

1. expected number of pairs of copies of H in G_{N,p_1} and G_{N,p_2} is $o(1)$
2. expected number of copies of H in G_{N,p_1} is $o(1)$
3. expected number of copies of H in G_{N,p_2} is $o(1)$



(a) case 1 (b) cases 1,2 (c) cases 1,3

Figure: The corresponding conditions determine the ‘optimal’ size n of H

Two point concentration: largest common induced subgr.

I_N = size of largest common induced subgraph of G_{N,p_1} and G_{N,p_2}

Surya-Warnke-Zhu (2023+): general case

For constant $p_1, p_2 \in (0,1)$, I_N is concentrated on two values around

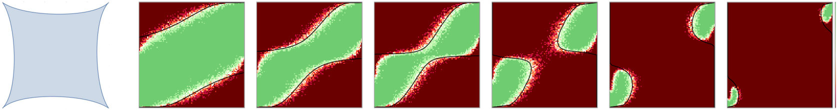
$$\max_{p\in[0,1]} \min\left\{x_N^{(0)}(p),\, x_N^{(1)}(p),\, x_N^{(2)}(p)\right\},$$

where for some b_0, b_1, b_2 depending on p_1, p_2 we have

$$x_N^{(0)}(p) = 4\log_{b_0} N - 2\log_{b_0}\log_{b_0} N - 2\log_{b_0}(4/e) + 1,$$
$$x_N^{(i)}(p) = 2\log_{b_i} N - 2\log_{b_i}\log_{b_i} N - 2\log_{b_i}(2/e) + 1.$$

- The optimization over p takes all possible edge-densities into account.
- Surprising: form of answer changes for constant edge-probability
- Proof uses (fairly technical) first and second moment method

Summary



Questions we answered

- When does induced copy of G_{n,p_1} appear in G_{N,p_2}? How many copies?
- Size of largest common induced subgraph of G_{N,p_1} and G_{N,p_2}?

- Each time vanilla second moment failed due to large variance
- Unusual distribution: squashed lognormal
- Surprising: form of answer changes for constant edge-probabilities

Open Problem

Size of the largest common induced subgraph of G_{N_1,p_1} and G_{N_2,p_2}?

- Complete understanding would unify our results