Sparse random graphs with many triangles

Suman Chakrborty

TU/e and UL

s.chakraborty1@tue.nl

September 7, 2021

Joint ongoing work with Remco van der Hofstad and Frank den Hollander.

Overview

Sparse Erdős-Rényi random graphs

We will consider $G_{n,p}$

- Undirected, loopless graph on n vertices.
- Each edge is present with probability $p = \frac{\lambda}{n}$ independently of each other.

Some properties of these graphs...

- Number of edges is approximately λn with high probability (probability going to one as n goes to infinity).
- 'On an average' the graph looks like a tree.
- Small number of triangles: approx $Poi\left(\frac{\lambda^3}{6}\right)$.

Some practical facts

Many networks/graphs observed in real-life have the following properties:

- They are sparse, that is number of edges and number of vertices are of the same order.
- The vertices are exchangeable.
- Typically contains many triangles (linear in number of vertices).

This suggests that networks that are frequently observed in real-life, are mathematically rare. So our first question of interest:

How rare are the networks that we see very often?

Formulation

Let T be the number of triangles in $G_{n,p}$, and k_n be a sequence of positive real numbers going to infinity, we will answer two questions first:

- for $p = \frac{\lambda}{n}$ sharp upper and lower bounds of
$$\mathbb{P}(T \geq k_n).$$
- for $p = \frac{\lambda}{n}$ how the random graphs 'look like' conditional on the event $\{T \geq k_n\}$.

before we state our results..

Why isn't it a classical LDP problem (use Varadhan's hammer etc.)? What is the state of the art?

Difficulty and the literature

Recall that for i.i.d random variables $X_1, X_2, \ldots, X_n$, establishing LDP (upper bound) of $Z_n := \frac{1}{n}\sum_{i=1}^n X_i$ is often achieved by computing

$$\mathbb{P}(Z_n \geq t) \leq \exp(-n\theta t)\,\mathbb{E}\left(\exp\left(\theta\sum_{i=1}^n X_i\right)\right) = \exp(-n\theta t)\left(\mathbb{E}\left(\exp\left(\theta X_1\right)\right)\right)^n$$

The fact that Z_n is just a linear sum of independent random variables helped!! Now, let $A = (a_{ij})$ be the adjacency matrix of $G_{n,p}$, then

$$6T(G) = \sum_{i,j,k=1}^n a_{ij}a_{jk}a_{ki},$$

here a_{ij}'s are i.i.d Bernoulli(p) random variables but unfortunately the function is far from linear and there are peculiar dependence present among the terms.

Difficulty and the literature

In the last 7 years this problem (and other related problems) received enormous attention. But (unfortunately) all studies are restricted to the range $\frac{\log n}{n} << p << 1$

- Approaches using classical concentration inequalities were summarized in the article titled 'The infamous upper tail' by Janson and Ruciński 2002.
- Chatterjee-Dembo started the theory of Nonlinear LDP in 2016. They proved LDP of T for some ranges of p.
- It got further developed by Augeri 2018, Cook-Dembo 2020.
- Harel, Mousset, and Samotij 2019 settled the LDP problem of T for all $\frac{\log n}{n} << p << 1$.

First Main Result: a rough statement

Our first main result roughly states the following: for $p_n = \frac{\lambda}{n}$, $k_n \to \infty$ (faster than $(\log n)^{2/3}$), and a constant $A > 0$,

$$\log \mathbb{P}(T \geq Ak_n) \approx -\frac{1}{2}[6Ak_n]^{2/3} \log(1/p_n).$$

In fact, we have obtained much more..the precise statement gives some control over the error terms as well. We will need some notation before giving the precise statement. For $n \in \mathbb{N}$, $p \in (0,1)$ and $A, k > 0$, define

$$\Phi_{n,p,k}(A) := \min\{e_G \log(1/p) \colon \; G \subseteq K_n, \, \mathbb{E}_G(T) \geq Ak\},$$

where e_G is the number of edges in G, $\mathbb{P}_G(.)$ is the law of $G_{n,p}$ conditional on the event that it contains a given graph G. Also write

$$\varepsilon_n = k_n^{-2/3} \log(1/p_n).$$

First Main Result

Theorem (C, Hofstad, Hollander 2021+)

(Upper bound) Suppose that $\lim_{n\to\infty} \varepsilon_n = 0$. Then, for every sequence of positive reals $(w_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} w_n = 0$, and such that $w_n k_n \geq \binom{n}{3}(p_n)^C$ for n large enough and some constant $C > 0$,

$$-\log \mathbb{P}(T \geq A k_n) \leq (1+\varepsilon_n)\Phi_{n,p_n,k_n}(A + w_n) \text{ for } n \text{ large enough.}$$

(Lower bound) For every $\delta > 0$ and every sequence of positive reals $(w_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} w_n = 0$, $\lim_{n\to\infty} w_n^2 k_n^{1/3} = \infty$ and $\lim_{n\to\infty}(\log w_n)/\log(1/p_n) = 0$,

$$-\log \mathbb{P}(T \geq A k_n) \geq (1 - \widehat{\Psi}_n)\Phi_{n,p_n,k_n}(A - 2w_n) + C(\delta) \text{ for } n \text{ large enough,}$$

where

$$\widehat{\Psi}_n = \frac{C'}{w_n^2 k_n^{1/3}} + \frac{\log[C' w_n^{-5} \log^3(1/p_n)]}{\log(1/p_n)} \text{ for some constant } C' > 0.$$

Solution of the variational problem

The following theorem provides asymptotically sharp upper and lower bounds on $\Phi_{n,p_n,k_n}(A)$ under a milder condition than those needed in the main theorem

Theorem (C, Hofstad, Hollander 2021)

For every sequence of positive reals $(w_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} w_n = 0$ and $\lim_{n\to\infty} w_n k_n^{1/3} = \infty$,

$$\frac{1}{2}[6A(1-w_n)k_n]^{2/3}\log(1/p_n) \leq \Phi_{n,p_n,k_n}(A) \leq \frac{1}{2}[6A(1+w_n)k_n]^{2/3}\log(1/p_n)$$

for n large enough.

Some Intuitions

Recall the rough form of our theorem

$$\log \mathbb{P}(T \geq Ak_n) \approx \frac{1}{2}[6Ak_n]^{2/3} \log(1/p_n).$$

There could be many ways to create more triangles in the random graph, what is the 'cheapest way' in our setting? An easy lower bound is the following:

$$\mathbb{P}(T \geq Ak_n) \geq \mathbb{P}(1, 2, \ldots, (6Ak_n)^{1/3} \text{ is a clique}) = p_n^{\binom{(6Ak_n)^{1/3}}{2}}.$$

For $p_n = \frac{\lambda}{n}$ this approximately becomes

$$\left(\frac{\lambda}{n}\right)^{\frac{(6Ak_n)^{2/3}}{2}} \approx \exp\left(-\log(1/p_n)\frac{(6Ak_n)^{2/3}}{2}\right).$$

It turns out that this is indeed 'almost cheapest' way to create a large number of triangles.

Structure Theorem

For $\delta > 0$, G is called an δ-clique if G has minimum degree $(1 - 4\delta^{1/2})(2e_G)^{1/2}$. Here is the structural result:

Theorem (C, Hofstad, Hollander 2021)

Under the conditions of Upper-bound of the main theorem, for every $\delta > 0$ there exists a constant $C > 0$ such that

$$\mathbb{P}\left(G_{n,p_n} \text{ contains no } C(\widehat{\Psi}_n + w_n)\text{-clique with } e_G \text{ edges} \mid T \geq Ak_n \right) \leq \delta,$$

for n large enough, $e_G = \frac{1}{2}[6(A - 2w_n)k_n(1 - Ck_n^{-1/3})]^{2/3}$.

In other words the graph will contain a $C(\widehat{\Psi}_n + w_n)$-clique with high probability.

Takeaway

- Good we have obtained sharp upper and lower bounds of the upper tail $\mathbb{P}(T \geq Ak_n)$.
- Good we have obtained structural information about the graph conditional on $\{T \geq Ak_n\}$.
- Bad unfortunately, the structure theorem shows that although the event $\{T \geq Ak_n\}$ is very rare, it only affects a tiny part of the graph (called localisation).
- Bad In fact we have been able to show the graph looks like a tree on an average (more precise results are under investigation).

Can we formulate an event which inherently affects the local structure of the graph?

Number of vertices that are part of some triangle

Consider $V_T(G)$, the number of vertices in G that are part of some triangle. Few things to note:

- In a sparse random graphs only a small number of vertices are part of the triangle (dominated by $3Poi\left(\frac{\lambda^3}{6}\right)$).
- If the graph is conditioned on the event, say, $\{V_T(G) \geq 0.3n\}$, then the 'local structure' of the graph will be different.
- The problem is that we do not know how to express $V_T(G)$ as a polynomial of the entries of the adjacency matrix. Therefore neither the classical techniques nor the new theory of non-linear large deviations seem to be applicable in this case.
- We have not found any literature on this problem.

Second main theorem

Theorem (C, Hofstad, Hollander 2021)

Let $(k_n)_{n\in\mathbb{N}}$ be such that $\lim_{n\to\infty} \frac{k_n}{\log n} = \infty$. Then

$$\mathbb{P}(V_T(G_{n,p_n}) \geq k_n) \leq n \exp\left(-\tfrac{1}{3}k_n \log(\tfrac{1}{3}k_n) + Ck_n\right)$$

for some constant $C > 0$.

Theorem (C, Hofstad, Hollander 2021)

Let $(k_n)_{n\in\mathbb{N}}$ be such that $\lim_{n\to\infty} k_n = \infty$ and $k_n \leq (1-\delta)n$ for a constant $\delta \in (0,1)$. Then

$$\mathbb{P}(V_T(G_{n,p_n}) \geq k_n) \geq \exp\left(-\tfrac{1}{3}k_n \log(\tfrac{1}{3}k_n) - Ck_n\right)$$

for some constant $C > 0$.

So what is the intuition here?

- Here we have the following ansatz: the 'cheapest way' in which the event $\{V_T(G_{n,p_n}) \geq k_n\}$ can occur is that there are $k_n/3$ vertex disjoint triangles.
- The lower bound will follow by calculating probability of this event.
- For the upper bound, we developed some new techniques. In particular, we proved a graph decomposition lemma that enables us to precisely count the number of ways $\{V_T(G_{n,p_n}) \geq k_n\}$ can occur. This might be of independent interest.

Discussions

- Conditional on $\{T \geq k_n\}$, can we prove finer structural results?
- Conditional on $\{V_T \geq k_n\}$ can we determine the local-weak limit of the graph? Does that even exists?
- Can we leverage our result on V_T to prove a corresponding Laplace principle? (we are hoping to investigate this in the near future)
- What about extending the results to other subgraphs?

THANK YOU

Questions??