The missing log in large deviations for triangle counts

Sourav Chatterjee

(Courant Institute)

The setup

- Let $G(n,p)$ be the Erdős-Rényi random graph: n vertices, edge probability p, independent edges.
- Given a graph H, let X_H denote the number of copies of H in $G(n,p)$.
- The distribution of X_H has been studied extensively:
  - First results by Erdős and Rényi (1960).
  - In the very sparse case, $\mathbb{P}(X_H > 0)$ studied by Bollobás (1981).
  - Necessary and sufficient condition for the asymptotic normality of X_H obtained by Ruciński (1988).
  - Sharp large deviation inequalities for $\mathbb{P}(X_H \leq (1-\epsilon)\mathbb{E}(X_H))$ obtained by Janson, Łuczak and Ruciński (2000).
  - One question that remained open for a long time was the issue of sharp large deviations for the 'upper tail', i.e.

$$\mathbb{P}(X_H \geq (1+\epsilon)\mathbb{E}(X_H)).$$

The upper tail problem

- Almost completely intractable until the year 2001, when the first general exponential tail bound was obtained by Van Vu, and soon after by Janson and Ruciński (2002). Far from optimal.
- Major breakthrough in 2004: Kim and Vu showed that if H is a triangle, then for any $\epsilon > 0$, there is a constant $C(\epsilon) > 0$ such that whenever $p > C(\epsilon)^{-1} n^{-1} \log n$,

$$e^{-C(\epsilon)^{-1} n^2 p^2 \log(1/p)} \leq \mathbb{P}(X_H \geq (1+\epsilon)\mathbb{E}(X_H)) \leq e^{-C(\epsilon) n^2 p^2}.$$

- At the same time, Janson et. al. (2004) proved a similar result for general H, with a difference of $\log(1/p)$ between the upper and lower bounds.

Why should we care?

- Well-known open problem in the theory of random graphs.
- Janson and Ruciński, in their 2002 paper 'The infamous upper tail', try out an exhaustive array of techniques:
  - Azuma-Hoeffding type inequalities.
  - Talagrand's concentration inequalities.
  - Kim and Vu's polynomial method.
  - A large variety of combinatorial methods.
- Kim and Vu (2004) implement a sophisticated version of their polynomial method.
- Boucheron, Lugosi and Massart (2004) apply their striking new concentration inequalities to this question.
- Thus: there is no concentration inequality that gives the right answer, even for triangle counts where it is simply a third degree polynomial of independent random variables.
- So, naturally interesting as a purely probabilistic problem or a problem in concentration of measure even if you don't care about the Erdős-Rényi graph.

Resolution of the conjecture for triangle counts

Theorem (C., 2010)

Let T be the number of triangles in an Erdős-Rényi graph $G(n,p)$. For each $\epsilon > 0$ there is a sufficiently small positive constant $C(\epsilon)$ such that whenever $C(\epsilon)^{-1} n^{-1} \log n \le p \le C(\epsilon)$, we have

$$e^{-C(\epsilon)^{-1} n^2 p^2 \log(1/p)} \le \mathbb{P}(T \ge (1+\epsilon)\mathbb{E}(T)) \le e^{-C(\epsilon) n^2 p^2 \log(1/p)}.$$

Successive localization

- Fix n, p and ϵ.
- Let $G = G(n,p)$, $T =$ number of triangles in G.
- Call an edge in G 'good' if there are less than $\epsilon np/\log(1/p)$ triangles containing the edge.
- Call a vertex 'good' if it has less than $7np$ neighbors in G.
- Let

$T' :=$ #△s with all good edges.

$T_0 :=$ #△s with at least one bad edge, but all good vertices.

$T_1 :=$ #△s with one bad vertex and two good vertices.

$T_2 :=$ #△s with two bad vertices and one good vertex.

$T_3 :=$ #△s with all bad vertices.

Then

$$T \le T' + T_0 + T_1 + T_2 + T_3.$$

First step

- For each $1 \le i < j < k \le n$ let

$$Y_{ijk} := 1\{ijk \text{ is a } \triangle\}.$$

- Then each Y_{ijk} is a Bernoulli r.v. with mean p^3, and $T = \sum Y_{ijk}$.
- Y_{ijk} are not all independent. But Y_{ijk} and $Y_{i'j'k'}$ are independent if ijk and $i'j'k'$ do not share two or more vertices. This is a form of local dependence.

Concentration under local dependence: a first attempt

- Let $(X_i)_{i \in F}$ be a finite collection of random variables taking value in $[0,1]$.
- For each i, there is a set $N_i \subseteq F$ such that X_i is independent of $(X_j)_{j \notin N_i}$. N_i will be called the neighborhood of i.
- Let $S := \sum X_i$, and $S_i := \sum_{j \notin N_i} X_j$.
- Suppose $|N_i| \le a$ for all i.
- Note that: (1) $S - S_i \le a$. (2) X_i, S_i are independent. (3) $S_i \le S$.
- Then for any $\theta > 0$,

$$\mathbb{E}(Se^{\theta S}) = \sum_{i \in F} \mathbb{E}(X_i e^{\theta S}) \overset{(1)}{\le} \sum_{i \in F} \mathbb{E}(X_i e^{\theta a + \theta S_i}) \overset{(2)}{=}$$

$$e^{\theta a} \sum_{i \in F} \mathbb{E}(X_i)\mathbb{E}(e^{\theta S_i}) \overset{(3)}{\le} e^{\theta a}\mathbb{E}(e^{\theta S}) \sum_{i \in F} \mathbb{E}(X_i).$$

- Thus, if $m(\theta) := \mathbb{E}(e^{\theta S})$ and $\lambda := \mathbb{E}(S)$, then

$$m'(\theta) \le \lambda e^{\theta a} m(\theta).$$

Remarks

- This gives, for all $t \geq \lambda$,

$$\mathbb{P}(S \geq t) \leq \exp\Big(-\frac{t}{a}\Big(\log\frac{t}{\lambda} - 1 + \frac{\lambda}{t}\Big)\Big).$$

  (Recall: $S = \sum X_i$, $\lambda = \mathbb{E}(S)$, $a =$ size of dependency nbhd.)
- Reminiscent of the method of dependency graphs (sub-area of Stein's method) in normal approximation.
- Can be shown to be reasonably sharp by considering a sum of independent random variables, each repeated a times.
- Consider $Y_{ijk} = 1\{ijk \text{ is a } \triangle\}$. Here size of dependency neighborhood is $a =$ number of triplets $i'j'k'$ that share at least two vertices with $ijk = 3n - 8$.
- From the above tail inequality, we get

$$\mathbb{P}(T \geq (1+\epsilon)\mathbb{E}(T)) \leq e^{-C(\epsilon)\frac{n^3p^3}{a}} = e^{-C(\epsilon)n^2p^3}.$$

  Not even close to the right answer! Reason: the size of the dependency nbhd $\gg$ typical number of $\triangle$s in the nbhd.

Towards a more powerful concentration inequality under local dependence

- Recall: an edge is good if it is contained in $< r$ $\triangle$s, where $r = \epsilon np/\log(1/p)$. Call a $\triangle$ 'good' if it has all good edges.
- Let $X_{ijk} = 1\{ijk \text{ is a good } \triangle\}$. Recall: $Y_{ijk} = 1\{ijk \text{ is a } \triangle\}$.
- Recall: $T' = \sum X_{ijk} = \#$ good $\triangle$s, $T = \sum Y_{ijk} = \#\triangle$s.
- Fix ijk. Let

$$T'' := \# \text{ good } \triangle\text{s that do not contain the edges } ij, jk, ki,$$

$$T''' := \# \text{ good } \triangle\text{s that do not contain the edges } ij, jk, ki,$$

if we force the edges ij, jk, ki to be present.

- Note: (1) $T''' \le T'' \le T'$. (2) If $X_{ijk} = 1$, then $T''' = T''$. (3) If $X_{ijk} = 1$, $T' - T'' \le 3r$. (4) Y_{ijk}, T''' are independent. (5) $X_{ijk} \le Y_{ijk}$.
- Then $\mathbb{E}(X_{ijk}e^{\theta T'}) \overset{(3)}{\le} e^{3\theta r}\mathbb{E}(X_{ijk}e^{\theta T''}) \overset{(2)}{=} e^{3\theta r}\mathbb{E}(X_{ijk}e^{\theta T'''}) \overset{(5)}{\le} e^{3\theta r}\mathbb{E}(Y_{ijk}e^{\theta T'''}) \overset{(4)}{=} e^{3\theta r}\mathbb{E}(Y_{ijk})\mathbb{E}(e^{\theta T'''}) \overset{(1)}{\le} e^{3\theta r}p^3\mathbb{E}(e^{\theta T'})$.

Tail bound for number of good triangles

- Thus, for each ijk,

$$\mathbb{E}(X_{ijk}e^{\theta T'}) \le e^{3\theta r}p^3\mathbb{E}(e^{\theta T'}).$$

  Here $X_{ijk} = 1\{ijk \text{ is a good } \triangle\}$, $r = \epsilon np/\log(1/p)$ and $T' = \#$ good $\triangle$s.
- Thus, if $m(\theta) := \mathbb{E}(e^{\theta T'})$ then

$$m'(\theta) = \mathbb{E}(T'e^{\theta T'}) \le \lambda e^{3\theta r}m(\theta),$$

  where $\lambda = \binom{n}{3}p^3 = \mathbb{E}(T)$.
- This gives the correct tail bound for T'. So, what did we do?
- We restricted ourselves to 'good' triangles, which have the effect of diminishing the size of the dependency neighborhood (which is r in this case) to the right size.
- The above method can be generalized.

Concentration inequality under local dependence

Theorem

Let $(X_i)_{i\in F}$, $(X_i')_{i\in F}$, $(X_{j(i)})_{i,j\in F}$ be nonnegative random variables with finite m.g.f. and a be a constant, such that for all i,

(a) $X_i \le X_i'$,

(b) X_i' *and* $\sum_{j\in F} X_{j(i)}$ *are independent,*

(c) $\sum_{j\in F} X_{j(i)} \le \sum_{j\in F} X_j$, *and*

(d) *whenever* $X_i > 0$,

$$\sum_{j\in F} X_j \le a + \sum_{j\in F} X_{j(i)}.$$

Let $\lambda := \sum_{i\in F} \mathbb{E}(X_i')$. *Then for any* $t \ge \lambda$,

$$\mathbb{P}\Big(\sum_{i\in F} X_i \ge t\Big) \le \exp\Big(-\frac{t}{a}\Big(\log\frac{t}{\lambda} - 1 + \frac{\lambda}{t}\Big)\Big).$$

Controlling bad triangles

- Recall: For $r = 0, 1, 2, 3$, $T_r = \#\triangle$s with at least one bad edge, r bad vertices, and $3 - r$ good vertices.
- With some extra effort, the concentration inequality works for T_1 and T_2.
- T_3 is different: (1) The set of bad vertices is of size $\leq np\log(1/p)$ with probability at least $1 - \exp(-Cn^2p^2\log(1/p))$. (2) Any subgraph with $\geq \epsilon n^3p^3$ triangles must have at least $c(\epsilon)n^2p^2$ edges. (3) The probability that there exists a set of size $\leq np\log(1/p)$ with $\geq c(\epsilon)n^2p^2$ edges is bounded by $\exp(-C(\epsilon)n^2p^2\log(1/p))$. Combining, this gives

$$\mathbb{P}(T_3 \geq \epsilon n^3p^3) \leq e^{-C(\epsilon)n^2p^2\log(1/p)}.$$

Controlling T_0

- Recall: $T_0 = \#\triangle$s with at least one bad edge and all good vertices.
- Call a set of edges 'fermionic' if no two share a common vertex.
- For any set of edges F, let $t(F) := \sum_{e \in F} t(e)$, where $t(e) = \#\triangle$s containing e.
- Define two events:

$$E_1 := \{\exists \text{ a fermionic set of bad edges of size } > np\log(1/p)\}.$$

$$E_2 := \{\exists \text{ a fermionic set } F \text{ of bad edges of size } \le np\log(1/p) \text{ with } t(F) > \epsilon n^2 p^2\}.$$

- The concentration inequality is used to show that both $\mathbb{P}(E_1)$ and $\mathbb{P}(E_2)$ are bounded by $\exp(-C(\epsilon)n^2p^2\log(1/p))$.
- Claim: If E_1 and E_2 are false, then $T_0 \le 15\epsilon n^3 p^3$.

Proof of claim

- Suppose E_1 and E_2 are false. Let B' be the set of those bad edges, both of whose endpoints are good vertices.
- Call two elements of B' 'adjacent' if they share a common vertex.
- Since the degree of an endpoint of any element of B' is $< 7np$, the maximum vertex degree of the adjacency graph on B' is $< 14np$.
- Thus, there is a coloring of B' with $\leq 14np + 1 \leq 15np$ colors such that no two adjacent edges in B' receive the same color.
- For each color x, let F_x denote the subset of B' that receives the color x.
- Now take the color x that maximizes $t(F_x)$. Then

$$t(F_x) \geq \frac{t(B')}{\text{number of colors}} \geq \frac{t(B')}{15np}.$$

- $E_1^c \implies |F_x| \leq np\log(1/p)$; $E_2^c \implies t(F_x) \leq \epsilon n^2 p^2$. Thus, $T_0 \leq t(B') \leq 15\epsilon n^3 p^3$. This completes the proof.