6. Graph theory

Monday, February 10, 2020 6:20 PM

Erdős-Renyi random graphs

Let $G(n,p)$ be a graph valued random variable, where n is the number of vertices and $p \in [0,1]$ is the probability any two vertices are linked by an edge, independent of all other edges.

We will use this model to study certain types of phase transition behavior.

Thm 8.1 Let v_1 be a vertex of the random graph $G(n,p)$. Let $\alpha \in (0, \sqrt{np})$. Then

$$\text{Prob}(|np - \deg(v)| \geq \alpha\sqrt{np}) \leq 3e^{-\alpha^2/8}.$$

proof. $\deg(v_1) = \sum_{i=2}^{n} \mathbb{1}_{1,i}$, where $\mathbb{1}_{i,j}$ is an indicator variable denoting the presence of edge (i,j).

Then the proof follows from an application of Chernoff bounds (12.6),

$$\text{Prob}\left(|\deg(v_1) - (n-1)p| \geq c(n-1)p\right) \leq 3e^{-(n-1)pc^2/8}$$

after using the approximation $n-1 \approx n$ for large n.

Let $c = \frac{\alpha}{\sqrt{(n-1)p}}$.

Corollary 8.2 Suppose ε is a positive constant. If $p \geq \frac{9\ln n}{n\varepsilon^2}$, then almost surely every vertex has degree in range $[(1-\varepsilon)np, (1+\varepsilon)np]$.

proof. Let $\alpha = \varepsilon\sqrt{np}$ in the above theorem. Then the probability an individual vertex fails to fall in the range is $\leq 3e^{-\varepsilon^2 np/8}$.

By a union bound, the probability that some vertex falls outside the range is $\leq 3ne^{-\varepsilon^2 np/8}$, so when $p \geq \frac{9\ln n}{n\varepsilon^2}$,

$$\leq 3ne^{-9\ln n/8} = 3n \cdot n^{-\frac{9}{8}} = 3n^{-\frac{1}{8}} = o(1).$$

i.e. if $p = \Omega\left(\frac{\log n}{n}\right)$, then with vanishing probability, all vertexes have tightly constrained degree.

Claim: $G(n, \frac{d}{n})$ has in expectation $\approx \frac{d^3}{6}$ triangles.

Moral justification of independence from n:
As n increases, the number of triples grows with n^3.
But each pair has $\frac{d}{n}$ prob. of having an edge, which decreases with n.
In fact, the chance that all three edges exist changes with $\frac{d^3}{n^3}$,
which exactly cancels out the growth in triplets.

proof. Let Δ_{ijk} be the indicator variable for a triangle i,j,k existing.
Then

proof. Let Δ_{ijk} be the indicator variable for a triangle i, j, k existing.

Then

$$\mathbb{E}(\#\text{triangles}) = \mathbb{E}\Big(\sum_{ijk} \Delta_{ijk}\Big) = \sum_{ijk} \mathbb{E}(\Delta_{ijk}) = \binom{n}{3}\Big(\frac{d}{n}\Big)^3 = \frac{n(n-1)(n-2)}{6}\cdot\frac{d^3}{n^3}$$

linearity of expectation doesn't depend on independence

$$\approx \frac{d^3}{6}.$$

$\boxed{/\!/}$

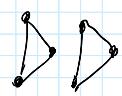
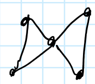
But this only says the expected number of triangles. Let's try to show that with probability bounded away from 0, there is at least one triangle in $G(n, \frac{d}{n})$.

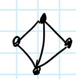
i.e. we need to rule out occasionally having a graph with all the triangles, and normally having none.

Let $x = \#\text{triangles}$, $x = \sum_{ijk} \Delta_{ijk}$. We will use a second moment argument.

$$\mathbb{E}(x^2) = \mathbb{E}\Big(\sum_{ijk}\Delta_{ijk}\Big)^2 = \mathbb{E}\Big(\sum_{\substack{i,j,k\\ i',j',k'}} \Delta_{ijk}\,\Delta_{i'j'k'}\Big)$$

Split the sum into three parts

$S_1 = \{i, j, k, i', j', k' \mid \Delta_{ijk}$ and $\Delta_{i'j'k'}$ share no edges$\}$  or 

$S_2 = \{i, j, k, i', j', k' \mid \Delta_{ijk}$ and $\Delta_{i'j'k'}$ share exactly 1 edge$\}$ 

$S_3 = \{i, j, k, i', j', k' \mid \Delta_{ijk} = \Delta_{i'j'k'}\}$ 

$$\mathbb{E}\Big(\sum_{S_1}\Delta_{ijk}\Delta_{i'j'k'}\Big) = \sum_{S_1}\underbrace{\mathbb{E}(\Delta_{ijk})\,\mathbb{E}(\Delta_{i'j'k'})}_{\text{independent because no edges shared}} \le \Big(\sum_{\substack{\text{all}\\ ijk}}\mathbb{E}(\Delta_{ijk})\Big)\Big(\sum_{\substack{\text{all}\\ i'j'k'}}\mathbb{E}(\Delta_{i'j'k'})\Big) = (\mathbb{E}x)^2$$

$$\mathbb{E}\Big(\sum_{S_2}\Delta_{ijk}\Delta_{i'j'k'}\Big) = \underbrace{\binom{n}{4}}_{\text{ways to choose 4 vertices}}\underbrace{\binom{4}{2}}_{\text{ways to choose 2 vertices lacking an edge}}\cdot\underbrace{p^5}_{\text{chance remaining 5 edges are present}} \approx \frac{n^4}{24}\cdot 6\cdot p^5 = \frac{1}{4}n^4p^5 = \frac{1}{4}n^4\cdot\frac{d^5}{n^5} = \frac{1}{4}\cdot\frac{d^5}{n} = o(1)$$

$$\mathbb{E}\Big(\sum_{S_3}\Delta_{ijk}\Delta_{i'j'k'}\Big) = \mathbb{E}\Big(\sum_{S_3}\Delta_{ijk}\Big) = \mathbb{E}x.$$

since triangles identical.

$\Rightarrow \mathbb{E}(x^2) \leq (\mathbb{E}x)^2 + \mathbb{E}x + o(1)$.

$\Rightarrow \text{Var}(x) = \mathbb{E}(x^2) - (\mathbb{E}x)^2 \leq \mathbb{E}x + o(1)$.

Then $\text{Prob}(x = 0) \leq \text{Prob}(|x - \mathbb{E}x| \geq \mathbb{E}x)$.

By Chebychev,

$$\text{Prob}(x=0) \leq \frac{\text{Var}(x)}{(\mathbb{E}x)^2} \leq \frac{\mathbb{E}x + o(1)}{(\mathbb{E}x)^2} \leq \frac{6}{d^3} + o(1).$$

Thus, if $d > \sqrt[3]{6} \approx 1.8$, $\text{Prob}(x=0) < 1$, so $G(n, \frac{d}{n})$ has a triangle with non-zero probability.

For $d < \sqrt[3]{6}$, $\mathbb{E}x = \frac{d^3}{6} < 1$, so there aren't very many triangles to go around.

Intuitively, need many vertices with deg ≥ 2 to have triangles.

Phase transitions

Def. If $\exists\, p(n)$ s.t. when $\lim_{n\to\infty} \frac{p_1(n)}{p(n)} = 0$, $G(n, p_1(n))$ lacks a property (almost surely),

when $\lim_{n\to\infty} \frac{p_2(n)}{p(n)} = \infty$, $G(n, p_2(n))$ has a property (almost surely),

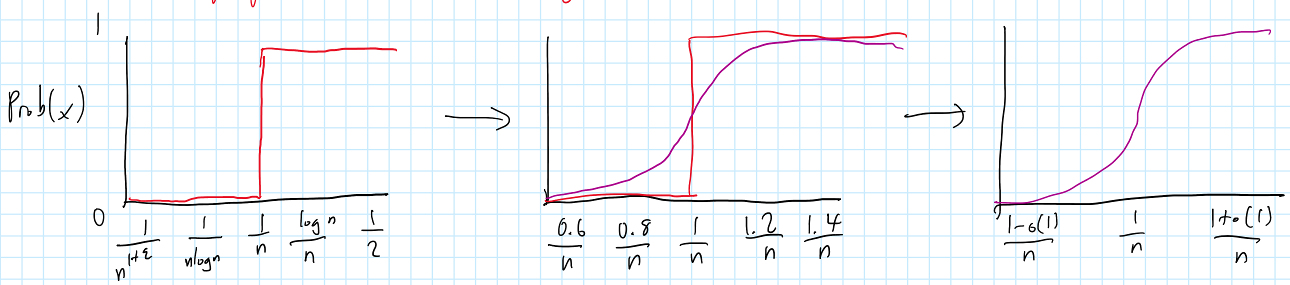
then a phase transition for the property occurs at threshold $p(n)$.

Def. If for $cp(n)$, $c<1$, $G(n, cp(n))$ lacks a property almost surely,

$c > 1$, $G(n, cp(n))$ has a property almost surely,

then $p(n)$ is a sharp threshold.

As the average degree increases in an Erdös-Renyi graph, structural properties suddenly change.



asymptotic phase transition at $\frac{1}{n}$.

smoother when zoomed in, unless sharp,

If previously sharp, can zoom in even more,

Phase transitions of Erdös-Renyi graph

Probability	Behavior
$p = o\left(\frac{1}{n}\right)$	Forest of trees, component size $O(\log n)$
$p = \frac{d}{n}$, $d<1$	Some cycles, component size $O(\log n)$
$p = \frac{d}{n}$, $d=1$	Components of size $O(n^{\frac{2}{3}})$
$p = \frac{d}{n}$, $d>1$	Giant component + $O(\log n)$ components
$p = \frac{1}{2} \cdot \frac{\ln n}{n}$	Giant component + isolated vertices
$p = \frac{\ln n}{n}$	No isolated vertices, Appearance of Hamiltonian circuit. Diameter $O(\log n)$
$p = \sqrt{\frac{2 \ln n}{n}}$	Diameter 2.
$p = \frac{1}{2}$	Clique of size $(2-\varepsilon)\ln n$.

How do we go about proving these properties?

We use so-called moment methods.

Thm First-moment method

Let $x(n)$ denote the number of occurences of an item in a random graph. If $\mathbb{E}\,x(n) \to 0$ as $n \to \infty$, then a random graph almost surely has no occurences of the item.

proof. Markov's inequality. x is non-negative.

$$\text{Prob}(x \geq a) \leq \frac{1}{a}\mathbb{E}\,x, \text{ so } \text{Prob}(x(n) \geq 1) \leq \mathbb{E}\,x(n) \to 0 \text{ as } n \to \infty.$$

Thm Second moment method

Let $x(n)$ be a random variable with $\mathbb{E}\,x > 0$.

If $\text{Var}(x) = o\left((\mathbb{E}\,x)^2\right)$, then $x > 0$ almost surely.

Recall $f(x) = o(g(x))$ iff $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$.

proof. $\text{Prob}(x \leq 0) \leq \text{Prob}(|x - \mathbb{E}\,x| \geq \mathbb{E}\,x)$.

By Chebyshev, $\leq \frac{\text{Var}(x)}{(\mathbb{E}\,x)^2} \to 0$.

(Used in our proof of # triangles)

Corollary Let x be a r.v. with $\mathbb{E}x > 0$. If $\mathbb{E}(x^2) \le (\mathbb{E}x)^2(1+o(1))$, then $x > 0$ almost surely.

Harder to use second moment methods because it can be hard to compute variance without independence. (i.e. $\mathbb{E}xy \neq \mathbb{E}x\,\mathbb{E}y$.

In looking for a phase transition, almost always the transition in probability of an item occurring occurs when the expected number of items transitions.

Thm: The property that $G(n,p)$ has diameter 2 has a sharp threshold at $p = \sqrt{2}\cdot\sqrt{\frac{\ln n}{n}}$.

(i.e. If $p = c\sqrt{\frac{\ln n}{n}}$ for $c<\sqrt{2}$, then the diameter is almost surely >2
for $c>\sqrt{2}$, " " " " " " ≤ 2.)

proof. If G has diameter greater than 2, then $\exists$ non-adjacent vertices i and j s.t. no other vertex is adj to both i and j. Call such a pair "bad".

Let the indicator $I_{ij} = 1$ iff the pair (i,j) is bad.

Let $x = \sum_{i<j} I_{ij}$, the number of bad vertices.

$$\mathbb{E}x = \underbrace{\binom{n}{2}}_{\text{number of pairs } i<j}\underbrace{(1-p)}_{\text{prob } i \text{ and } j \text{ not adjacent}}\underbrace{(1-p^2)}_{\text{prob a vertex is not adjacent to both } i \,\&\, j}{}^{n-2}$$

$n-2$ — number of other vertices.

Setting $p = c\sqrt{\frac{\ln n}{n}}$,
$$\mathbb{E}x \cong \frac{n^2}{2}\left(1-c\sqrt{\frac{\ln n}{n}}\right)\left(1-c^2\cdot\frac{\ln n}{n}\right)^n$$
$$\cong \frac{n^2}{2}\exp\left(-c^2\cdot\ln n\right) \cong \frac{1}{2}n^{2-c^2}.$$

For $c>\sqrt{2}$, $\lim_{n\to\infty}\mathbb{E}x = 0$. By the first moment method, $G(n,p)$ almost surely has no bad pair and hence, has diameter 2.

Now consider $c<\sqrt{2}$, where $\lim_{n\to\infty}\mathbb{E}x = \infty$. We will use the 2nd moment method.

$$\mathbb{E}(x^2) = \mathbb{E}\left(\sum_{i<j} I_{ij}\right)^2 = \mathbb{E}\left(\sum_{i<j} I_{ij}\cdot\sum_{k<l} I_{kl}\right) = \mathbb{E}\left(\sum_{\substack{i<j\\k<l}} I_{ij}I_{kl}\right) = \sum_{\substack{i<j\\k<l}}\mathbb{E}(I_{ij}I_{kl}).$$

Use the same idea as with triangles to partition the set.

$$= \sum_{\substack{i<j\\k<\ell\\ \text{all } i,j,k,\ell \text{ distinct}}} \mathbb{E}(I_{ij} I_{k\ell}) + \sum_{\substack{\{i,j,k\}\\ i<j \\ \text{all } i,j,k \text{ distinct}}} \mathbb{E}(I_{ij} I_{ik}) + \sum_{i<j} \mathbb{E}(I_{ij}^2)$$

When all 4 vertices are distinct, there must be two bad pairs (i,j) and (k,ℓ) for $I_{ij} I_{k\ell} = 1$. Then $\forall u \notin \{i,j,k,\ell\}$, at least one of (i,u), (j,u) is absent

" " " " (k,u), (ℓ,u) is absent.

The probability of both absences is $(1-p^2)^2$

So $\mathbb{E}(I_{ij} I_{k\ell}) \leq (1-p^2)^{2(n-4)} \leq \left(1 - c^2 \frac{\ln n}{n}\right)^{2n} (1+o(1)) \leq n^{-2c^2}(1+o(1))$

$$\Rightarrow \sum_{\substack{i<j\\k<\ell\\ \text{all distinct}}} \mathbb{E}(I_{ij} I_{k\ell}) \leq \frac{1}{4} n^{4-2c^2}(1+o(1))$$

(because $< \frac{1}{4}$ of n^4 4-tuples have $i<j$ $j<k$)

When only 3 distinct vertices, if $I_{ij} I_{ik} = 1$, then $\forall u \notin \{i,j,k\}$, either there is no edge between i and u, or there is an edge (i,u) and both (j,u) and (k,u) are absent.

This probability is $1-p+p(1-p)^2 = 1-2p^2+p^3 \simeq 1-2p^2$. (for one u)

Thus, $\mathbb{E}(I_{ij} I_{ik}) \simeq (1-2p^2)^{n-3} \simeq \exp(-2p^2(n-3)) \simeq \exp(-2c^2 \ln n) \simeq n^{-2c^2}$

$$\Rightarrow \sum_{\substack{\{i,j,k\}\\ i<j}} \mathbb{E}(I_{ij} I_{ik}) \leq n^3 \cdot n^{-2c^2} = n^{3-2c^2}$$

When only 2 distinct vertices $\sum_{i<j} \mathbb{E}(I_{ij}^2) = \mathbb{E}X \simeq \frac{1}{2} n^{2-c^2}$

Together $\mathbb{E}(X^2) \leq \frac{1}{4} n^{4-2c^2} + n^{3-2c^2} + \frac{1}{2} n^{2-c^2} = \frac{1}{4} n^{4-2c^2}\left(1 + 4n^{-1} + 2n^{c^2-2}\right)$

If $c < \sqrt{2}$, $\mathbb{E}(X^2) \leq (\mathbb{E}X)^2 (1+o(1))$.

By a second moment argument, the graph almost surely has at least one bad pair, so the diameter is > 2. $\square$

Thm 8.6 The disappearance of isolated verties in $G(n,p)$ has a sharp threshold of $\frac{\ln n}{n}$.

proof. Let x be the number of isolated vertices.

Then $\mathbb{E}x = n(1-p)^{n-1}$

Let $p = c\frac{\ln n}{n}$. Then $\lim_{n\to\infty}\mathbb{E}x = \lim_{n\to\infty} n\left(1-\frac{c\ln n}{n}\right)^n = \lim_{n\to\infty} n e^{-c\ln n} = \lim_{n\to\infty} n^{1-c}$.

If $c>1$, $\mathbb{E}x \to 0$, so by a first moment argument, almost all graphs have no isolated vertices.

If $c<1$, $\mathbb{E}x \to \infty$, so need a second moment argument.

Assume $c<1$. Let $x = I_1 + \cdots + I_n$, where I_i is the indicator variable for vertex i being isolated.

Then $$\mathbb{E}(x^2) = \sum_{i=1}^{n}\mathbb{E}(I_i^2) + \underbrace{2\sum_{i<j}\mathbb{E}(I_iI_j)}_{\text{because isolation is not independent.}}$$

$$= \mathbb{E}x + n(n-1)\mathbb{E}(I_1I_2)$$
$$= \mathbb{E}x + n(n-1)(1-p)^{n-1}(1-p)^{n-2}$$
$$= \mathbb{E}x + n(n-1)(1-p)^{2(n-1)-1}$$

Then $$\frac{\mathbb{E}(x^2)}{(\mathbb{E}(x))^2} = \frac{n(1-p)^{n-1} + n(n-1)(1-p)^{2(n-1)-1}}{n^2(1-p)^{2(n-1)}} = \frac{1}{n(1-p)^{n-1}} + \left(1-\frac{1}{n}\right)\frac{1}{1-p}$$

For $p = c\frac{\ln n}{n}$ with $c<1$, $\lim_{n\to\infty}\mathbb{E}x = \infty$ and

$$\lim_{n\to\infty}\frac{\mathbb{E}(x^2)}{(\mathbb{E}x)^2} = \lim_{n\to\infty}\left[\underbrace{\frac{1}{n^{1-c}}}_{\to 0} + \left(1-\frac{1}{n}\right)\cdot\frac{1}{1-c\frac{\ln n}{n}}\right] = \lim_{n\to\infty}\left(\frac{1}{1-c\frac{\ln n}{n}}\right) = 1 + o(1).$$

By second moment argument, almost all graphs have isolated vertices for $c<1$. $\blacksquare$

Component sizes

Consider a breadth-first-search (BFS) on a graph.

i.e. explore all neighbors of a starting node, all neighbors of the neighbors, i.e. explore all neighbors of a starting node, all neighbors of the neighbors, and so on recursively.

Discovered but unexplored vertices are the frontier.

When the frontier is 0, the entire connected component has been explored.

But we can imagine generating edges only when we need them.

Define a step as the full exploration of a single vertex.

Further, define a red vertex whenever the BFS finishes, so we can keep on exploring all the components.

This modified BFS has the property that the probability a node is unexplored after i steps is $(1-p)^i$. For a graph $G(n, \frac{d}{n})$, $p = \frac{d}{n}$.

Define the size of the frontier as the number of discovered vertices minus the number of explored vertices.

In a true BFS, this is non-negative but the red vertices can cause this number to be negative.

Let F_i be the size of the frontier at step i.

Then for large n, $\mathbb{E} F_i = n\underbrace{\left(1-(1-p)^i\right)}_{\text{discovered vertices}} - \underbrace{i}_{\substack{\text{explored}\\ \text{vertices}}} \simeq n\left(1-e^{-pi}\right) - i = n\left(1-e^{-\frac{d}{n}i}\right) - i$

Then the normalized frontier size $\frac{\mathbb{E} F_i}{n} = 1 - e^{-\frac{d}{n}i} - \frac{i}{n}$.

Let $x = \frac{i}{n}$ be the normalized # of steps.

Then $f(x) = 1 - e^{-dx} - x$ is the normalized expected size of the frontier.

If $d > 1$, $f(0) = 0$ and $f'(0) = d - 1 > 0$, so f is increasing at 0.

But $f(1) = -e^{-d} < 0$, so for some value $0 < \theta < 1$, $f(\theta) = 0$. (If $d = 2$, $\theta = 0.7968$)

For $d > 1$, $\mathbb{E} F_{i+1} - \mathbb{E} F_i \approx (d-1)i$ for small i.

(because each new node adds $d-1$ new neighbors to the frontier).

We want to understand $\mathbb{P}(F_i = 0)$ for $i < n$, as the first such i marks the size of the first connected component.

For small i, $\mathbb{P}(\text{vertex discovered}) = 1 - \left(1 - \frac{d}{n}\right)^i \simeq \frac{id}{n}$.

For small i, $\mathbb{P}(\text{vertex discovered}) = 1 - \left(1 - \frac{d}{n}\right)^i \approx \frac{id}{n}$.

And the number of discovered vertices $\text{binom}\left(n, \frac{id}{n}\right) \approx \text{Poisson}(id)$

So $\mathbb{P}(k \text{ vertices discovered by step } i) \approx e^{-di} \cdot \frac{(di)^k}{k!}$.

We need exactly i vertices discovered by step i, so probability

$$\approx e^{-di} \cdot \frac{(di)^i}{i!} \approx e^{-di} \frac{d^i i^i}{i^i} e^i = e^{-(d-1)i} d^i = e^{-(d-1-\ln d)i}.$$

For $d \neq 1$, $d - 1 - \ln d > 0$ (by calculus)

Thus, the probability drops off exponentially with i.

Termination probability for $i > c \ln n$ for sufficiently large c is thus $o\left(\frac{1}{n}\right)$.

So it is unlikely to terminate before the Poisson approximation fails, if it is already $\Omega(\ln n)$.

On the other hand, for i near $n\theta$, $\mathbb{E}F_{i+1} - \mathbb{E}F_i = \alpha |i - n\theta|$ for some proportion α.

There are only $|i - n\theta|$ vertices left in expectation to explore, and each step explores those with prob. proportional to remaining.

For i near $n\theta$, can approximate binomial via Gaussian, which falls off exponentially with the square of the distance from the mean. $\left(e^{-\frac{k^2}{\sigma^2}}\right)$ $\sigma^2 \sim n$

Thus to have a non-vanishing prob., $k \leq \sqrt{n}$. So the giant component is in the range $[n\theta - \sqrt{n}, n\theta + \sqrt{n}]$.