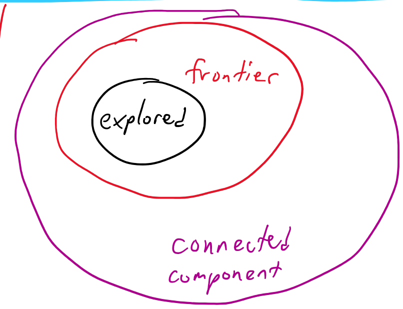
16. Component Sizes for Erdos-Renyi

Tuesday, October 12, 2021 6:38 PM

Component sizes for $G(n, \frac{d}{n})$ $d>1$

Once you get far enough, frontier is too big to stop.



Consider a breadth-first-search (BFS) on a graph.
i.e. explore all neighbors of a starting node,
explore all neighbors of the neighbors,
and so on recursively.

Frontier: discovered but unexplored vertices (discovered means neighbors of explored)

When $|\text{frontier}| = 0$, done exploring an entire connected component.

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

Define a **step** as the full exploration of a single vertex. (finding all neighbors)

Modified BFS:

Normally, the process will stop when a connected component is explored.

Whenever $|\text{frontier}| = 0$, create a new undiscovered red vertex connected to all other vertices w.p. p, which we then explore to reach a new connected comp.

The 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: $\underbrace{|\widehat{\text{frontier}}|}$ $= |\text{discovered}| - |\text{explored}|$

modified and potentially negative because red nodes are explored but not discovered.

Let $F_i = |\text{frontier}'|$ at step i.

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

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 $d=2$, $\theta=0.7968$)

Note: True BFS must be completed by the time $f(\theta)=0$, so we know an upper bound on the expected size of the connected component.

Let's bound the size of the connected component using the actual vs. expected frontier size

Let's first show that we don't stop shortly after $c\ln n$ steps

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)

So we must have giant component or only small components $O(\log n)$

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

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

$\#$ discovered vertices $\approx \text{binom}\left(n, \frac{id}{n}\right) \approx \text{Poisson}(id)$

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

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, probability drops exponentially with i

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

So it is unlikely to terminate before the Poisson approx 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 discovers these 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 mean $\left(e^{-\frac{k^2}{\sigma^2}}\right)$ $\sigma^2\sim n$

$\text{binom}\left(n,\frac{id}{n}\right) \approx \mathcal{N}\left(id, id\left(1-\frac{id}{n}\right)\right)$ $id\left(1-\frac{id}{n}\right)\sim n\theta d(1-\theta d)\sim n$

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

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

Existence of giant component

We just showed that components are either $O(\log n)$ or $\Omega(n)$.

Let's prove that $G(n,p)$ with $p = \frac{1+\varepsilon}{n}$ has a giant component w.h.p.
with $0 < \varepsilon < \frac{1}{8}$ (Note, for larger ε, only increases component sizes)

Consider a depth-first search (DFS)

Let E = fully explored vertices
U = unvisited vertices
F = frontier of visited and still being explored nodes

always look at last added

Starting state: $E = \emptyset$, $F = \emptyset$, $U = V$. Treat $F = [v_1, \ldots, v_k]$ as a stack, with v_k the active vertex.

Repeat until $U = \emptyset$
- If $F = \emptyset$, let $F = [u]$, $u \in U$ arbitrarily chosen.
- Else $(F \neq \emptyset)$
  - If $\exists (v_k, u)$ for $u \in U$ (can generate edges on the fly w.p. p)
    - Remove u from U. Push u onto the stack F. (i.e. repeat edge queries until one is true or we run out of $u \in U$)
  - Else,
    - Pop v_k off F. Add v_k to E.

Lemma 8.7 After $\frac{\varepsilon n^2}{2}$ edge queries, w.h.p. $|E| < \frac{n}{3}$.

proof. If not, at some $t < \frac{\varepsilon n^2}{2}$, $|E| = \frac{n}{3}$.

$|F| \le \sum_{i=1}^{t} I_i$, where I_i is the Bernoulli r.v. corresponding to the ith edge query.

$$\le \varepsilon n^2 p \quad \text{w.h.p.} \qquad \left(\mathbb{E} \text{ is } \frac{\varepsilon n^2 p}{2}\right)$$

$$\le \frac{1}{8} \cdot n^2 \cdot \left(\frac{1+\frac{1}{8}}{n}\right) = \frac{9}{64} \cdot n < \frac{n}{3}.$$

Thus, at time t, $|U| = n - |E| - |F| \ge \frac{n}{3}$.

By construction, there must be no edges between U and E, but that means at least $|E||U| \ge \frac{n^2}{9}$ queries, so $t \ge \frac{n^2}{9}$.

Contradiction, because $t \le \frac{\varepsilon n^2}{2} \le \frac{n^2}{16}$. ✗

Note that F is always a connected component.

Note that F is always a connected component.

Lemma 8.8 After $t = \frac{\varepsilon n^2}{2}$ edge queries, w.h.p. $|F| \geq \frac{\varepsilon^2 n}{30}$.

proof. Suppose $|F| < \frac{\varepsilon^2 n}{30}$. Then

$$|U| = n - |E| - |F| \geq n - \frac{n}{3} - \frac{\varepsilon^2 n}{30} \geq 1 \quad \text{if } n \geq 2. \quad \text{(so DFS still active)}$$

$$|E| + |F| = \sum_{i=1}^{t} I_i \quad \text{(because yes answers to edge queries move from } U \text{ to } F\text{)}$$

$$\mathbb{E}\sum_{i=1}^{t} I_i = \frac{\varepsilon n^2 p}{2} = \frac{(1+\varepsilon)\varepsilon n}{2} = \frac{\varepsilon n}{2} + \frac{\varepsilon^2 n}{2}$$

$$\Rightarrow \text{w.h.p.} \quad \sum_{i=1}^{t} I_i \geq \frac{\varepsilon n}{2} + \frac{\varepsilon^2 n}{\underline{\underline{3}}} \quad \text{(By Chernoff-Hoeffding)}$$

$$\Rightarrow |E| \geq \frac{\varepsilon n}{2} + \frac{\varepsilon^2 n}{3} - \frac{\varepsilon^2 n}{30} = \frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}.$$

Again, $|E||U| \leq \frac{\varepsilon n^2}{2}$.

$$|E|\left(n - |E| - |F|\right) \leq \frac{\varepsilon n^2}{2}$$

In the range of $|E|$ in $\left[\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}, \frac{n}{3}\right]$, for F fixed, $|F| \leq \frac{n}{3}$,

$$\frac{d}{d|E|}|E|\left(n - |E| - |F|\right) = n - 2|E| - |F| \geq 0, \text{ so } |E||U| \text{ increases with } |E|.$$

Thus,
$$|E||U| \geq \underbrace{\left(\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}\right)\left(n - \frac{\varepsilon n}{2} - \frac{3\varepsilon^2 n}{10} - \frac{\varepsilon^2 n}{30}\right)}_{\frac{\varepsilon n^3}{2} - \frac{\varepsilon^2 n^2}{4} - \frac{3\varepsilon^3 n^2}{20} - \frac{\varepsilon^3 n^2}{60} + \frac{3\varepsilon^2 n^2}{10} - \frac{3\varepsilon^3 n^2}{20} - \frac{9\varepsilon^4 n^2}{100} - \frac{\varepsilon^4 n^2}{100}} > \frac{\varepsilon n^2}{2}$$

$$= \frac{\varepsilon n^3}{2} + \varepsilon^3 n^2 \underbrace{\left(\frac{5}{100} - \frac{19}{60}\varepsilon - \frac{1}{10}\varepsilon^2\right)}_{>0 \text{ if } \varepsilon < \frac{1}{8}}$$

~ 0.008884

This is a contradiction, so w.h.p. $|F| \geq \frac{\varepsilon^2 n}{30}$. $\square$

No other large components

Claim: For any $\varepsilon > 0$, $p = \frac{1+\varepsilon}{n}$, w.h.p. there is only one giant component in $G(n,p)$, all other components have size $O(\log n)$.

proof. Suppose $G(n,p)$ has δ prob. of 2 distinct components K_1, K_2 of size $\omega(\log n)$.

Let $A = \{1, 2, \ldots, \frac{\varepsilon n}{2}\}$

Then $\text{Prob}(|K_1 \cap A| = \omega(\log n)$ and $|K_2 \cap A| = \omega(\log n)) \geq \frac{\delta}{2}$

because we can randomly permute vertex labels, and both K_1 and K_2 have $\frac{\varepsilon}{4}$ fraction of their nodes in A. (expected $\frac{\varepsilon}{2}$)

WTS only 1 $\omega(\log n)$ component intersects A in $\omega(\log n)$ vertices.

Let $B = V - A$, $|B| = n(1 - \frac{\varepsilon n}{2})$

B has at least 1 giant component C^*, $|C^*| = \omega(\log n)$.

Let $C_1, C_2, C_3, \ldots$ be $\omega(\log n)$ components within A.

$\forall i$, there are $\omega(n \log n)$ potential edges b/t C_i and C^*.

Thus $\text{Prob}(C_i \text{ not connected to } C^*) \leq (1-p)^{\omega(n \log n)} = \frac{1}{n^{\omega(1)}}$

By union bound, all C_i's are connected to C^* w.h.p.

Thus, only 1 component intersects A in $\omega(\log n)$ vertices

$\Rightarrow$ only 1 large component in A. $\square$