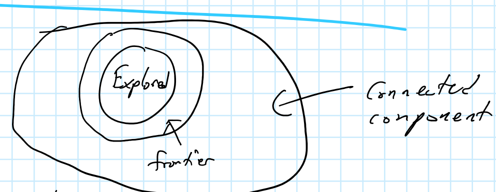
6.5 Random graph theory continued

Thursday, February 13, 2020 3:39 AM

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

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 $d > 1$ $\mathbb{E}F_i - \mathbb{E}F_0 \simeq (d-1)i$ for small i

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

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. $\left(e^{-\frac{k^2}{\sigma^2}}\right)$ $\sigma^2 \sim n$

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

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

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.

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.

where $p=\frac{(1+\varepsilon)}{n}$ 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 vertices.

Starting state: $E=\emptyset$, $F=\emptyset$, $U=V$. Treat $F=[v_1,\dots,v_k]$ as a stack, with v_k as 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 with prob p)

Remove u from U. Push u onto the stack F.

Else,

Pop v_k off F. Add v_k to E.

(i.e. repeat edge queries until one is true or we run out of $u\in U$)

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

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

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

$\le \varepsilon n^2 p$ w.h.p. ($\mathbb{E}$ is $\frac{\varepsilon^2 n p}{2}$)

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

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.

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

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

$$|U|=n-|E|-|F|\ge n-\frac{n}{3}-\frac{\varepsilon^2 n}{30}\ge 1 \quad \text{if } n\ge 2.$$

(so DFS still active)

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

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

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

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

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

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

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

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

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

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

~ 0.008854

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

Thus, there is at least one connected component with at least $\frac{\varepsilon^2 n}{30}$ vertices.

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 all other components have size $O(\log n)$.

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

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

Then $\text{Prob}(|K_1 \cap A| = \omega(\log n) \text{ and } |K_2 \cap A| \cap \omega(\log n)) \geq \frac{\delta}{2}$,
because we can imagine randomly permuting vertex labels, and
both K_1 and K_2 w.h.p. have $\frac{\varepsilon}{4}$ fraction of their nodes in A. (expected $\frac{\varepsilon}{4}$)

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

Thus, if we can show there exists only 1 component that intersects A in $\omega(\log n)$ vertices, we would be done.

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

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 between 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.