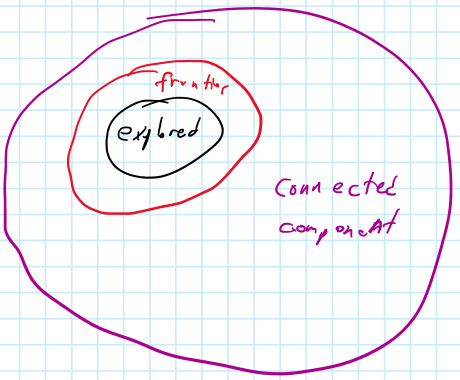
16. Component sizes

Thursday, October 14, 2021 3:52 PM

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



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 by 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 node (finding all neighbors)

Modified BFS

Normally, process stops when connected component is found

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

The modified BFS has the property that the prob. a node is undiscovered after i steps is $(1-p)^i$.

For a graph $G = (n, \frac{d}{n})$, $p = \frac{d}{n}$

Define: $|\text{frontier}'| = |\text{discovered}| - |\text{explored}|$

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

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

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

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

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 size(connected component) by using actual vs. expected frontier size

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

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

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

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

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

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 $d \neq 1$, $d - 1 - \ln d > 0$.

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 unlikely to terminate before 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

For i near $n\theta$, can approximate binomial via Gaussian, which falls off exponentially with the square of the distance from mean

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

Thus, to have a non-vanishing prob, $k \le \sqrt{n}$, so the giant component π 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 size)

Consider a depth first search (DFS)

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

Starting state: $E = \emptyset$, $F = \emptyset$, $U = V$. Treat $F = [v_1, \ldots, v_k]$ 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 w.p. p)
      Remove u from U, Push u onto 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 indicator r.v. corresponding to the ith edge query.

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

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

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

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

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

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

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

$$|E| + |F| = \sum_{i=1}^{t} I_i$$

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

w.h.p. $\sum_{i=1}^{t} I_i \ge \frac{\varepsilon n}{2} + \frac{\varepsilon^2 n}{3}$ (By Chernoff–Hoeffding)

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

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

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

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

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

Thus, $|E||U| \ge \left(\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}\right)\left(n - \frac{\varepsilon n}{2} - \frac{3\varepsilon^2 n}{10} - \frac{\varepsilon n}{30}\right) > \frac{\varepsilon n^2}{2}$

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

Note that frontier is a connected component.