Now that we've seen that the properties we observed in the web graph hold more broadly, if you're like me, it makes you very curious what is causing these properties to emerge.

This is the question we'll try to answer in the next few lectures:

"Where do these universal properties come from?"

Often research in this area is termed "network science"

As opposed to the engineering question: "How can I exploit these properties when designing systems"
→ This is what we'll move to after we discuss where these properties come from... e.g. in google search.

Random Graphs

As a 1st step towards trying to understand where these properties come from, lets consider the simplest model of forming a network we can think of.

... clearly we want to use randomness in the growth process...

(1) Let's start with the simplest [illegible]

Q: What's the simplest way you can think of to build a "random" graph?
(Let's keep it simple & just use undirected edges.)

A: Add nodes 1 at a time and connect the new node to each other node indep. w/ prob p.

⇕

Build a graph on n nodes by allowing each possible edge to appear indep. w/ prob p.

This is a classical model of a random graph called an Erdos-Renyi Graph and usually denoted by $G_{n,p}$ or $G(n,p)$.

Before we start to study these, let me tell you a little about their history.

- first introduced by Gilbert (of course :) in 1959.
- Soon after Erdos & Renyi introduced them independently, and got the naming credit because they studied them in much more detail.

They were introduced not as a model for "real" networks, but instead as a construct for proving the existence of graphs with certain properties. i.e. if you show that a $G(n,p)$ has property P w/ prob >0 then there must exist graphs with property P.

→ If you've never heard about Erdos, you should look him up! He was a great character, and a brilliant mathematician.

Paul Erdös (1913-1996)

Basically lived as a vagabond, out of his suitcase, traveling between universities & conferences staying with the mathematicians he visited.

Quirks:

- Upon arrival he'd announce "my brain is open"
- Always talked about proofs from "the Book", which is where the most elegant arguements were held.

Lots of important results from graph theory, basically invented the "probabilistic method",

Most prolific mathematician → 1475 papers! 511 colaborators!

↓ which led to the "Erdos Number"

see ppt

Now back to $G(n,p)$. Given this simple model, the question becomes:

"Does $G(n,p)$ have our 4 universal properties?"

If so, it would suggest that they emerge as simple consequences of

many, independent random events
... a similar mechanism as provided by the central limit theorem.

⇒ As you can guess, the explanation won't be so simple. $G(n,p)$ won't have all of our universal properties.

Exploring $G(n,p)$

Q: Can anyone immediately see any of the properties that $G(n,p)$ won't satisfy?

A1: $G(n,p)$ is not highly clustered.

If $i \sim j$ and $i \sim k$ then the prob that $j \sim k$ is still p, independently of the fact that j & k are connected to i.

(You'll prove some results on your HW about the existence of Δ's)

A2: $G(n,p)$ does not have a heavy-tailed degree distribution.

Q: OK then, what is the degree dist.?

A: Each node has $n-1$ possible edges, each of which exists w/ prob p.

(since we don't allow self loops)

⇒ Binomial $(n-1, p)$.

So $\Pr(d(i) = m) = \binom{n-1}{m} p^m (1-p)^{n-m}$

& $E[d(i)] = (n-1)p$

And, the binomial is definitely not heavy-tailed.
You'll show on your HW that this is $\approx$ Poisson, which has an "exponential" tail ($\sim e^{-cx}$)
↑ lighter than a polynomial tail ($\sim x^{-\alpha}$)

OK, we can rule two props out immediately. To understand the other 2, we'll have to do some work.

We'll start by looking at connectivity.

Connectivity of $G(n,p)$.

The 1st thing we realize when we start to think about connectivity is that it clearly depends on n & p!

Q: What are some cases where things are easy?

A1: $p=0$ → graph is n isolated vertices

A2: $p=1$ → graph is a big clique.

A3: $p=\frac{2}{n^2}$ → $E[\#\text{edges}] \approx 1$ for large n

$p=\frac{d}{n^2}$ → $E[\#\text{edges}] \approx d/2$ for large n and all components are small

A4: $p>0$ is constant → if n is large enough, then graph is 1 connected component

w/ diameter 2 (w/ prob 1).

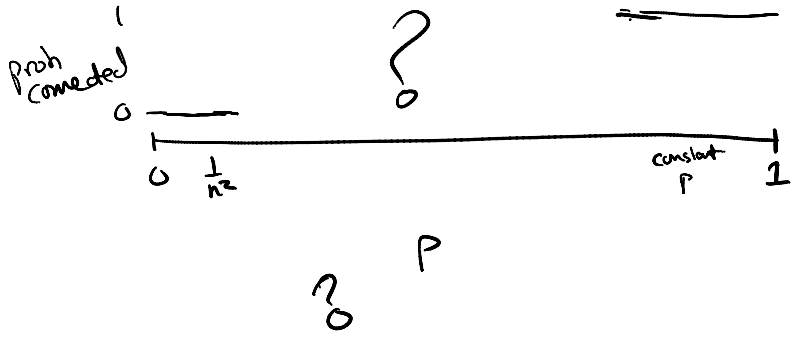
↑ You'll prove this on your HW!

So, when we think about large graphs, if p is a constant, we understand connectivity → if $p>0$ we're connected & have a small diameter.

But, this is clearly an unrealistic setting since $E[\text{degree}] = np \to \infty$ as $n \to \infty$

So, we need a p that is decreasing w/ n, but larger than $\frac{1}{n^2}$ for the model to be "realistic".

In summary:



To start to understand more, let's focus on a first step towards connectivity:

"When do isolated vertices disappear?"

(i.e. "how large does p need to be")

As a starting point, we'll look at the expected # of isolated vertices.

To analyze this, we'll use the same "linearity of expectation" trick you used on HW1 (hopefully).

Let X = # of isolated vertices

$$X_i = \begin{cases} 1 & \text{if vertex } i \text{ is isolated} \\ 0 & \text{else.} \end{cases}$$

so $X = \sum_{i=1}^{n} X_i$

$$\begin{aligned} \therefore E[X] &= n\,E[X_i] \\ &= n \Pr(X_i = 1) \\ &= n(1-p)^{n-1} \end{aligned}$$

Now, let's play with this a bit to understand it better... what we want to understand is how it changes as a function of $p = p(n)$.

☆ The real-world graphs we want to understand are all "huge") so we should look at $n \to \infty$ to understand them:

$$\lim_{n\to\infty} n(1-p)^{n-1} = \lim_{n\to\infty} n(1-p)^n \cdot \underbrace{\left(\frac{1}{1-p}\right)}_{\to 1 \text{ since } p \to 0}$$

$$= \lim_{n\to\infty} ne^{-pn}$$

$\left(1 - \frac{1}{x}\right)^x \to e^{-1}$ as $x \uparrow \infty$

$\Downarrow$

$\left(\left(1 - \frac{1}{1/p}\right)^{1/p}\right)^{np} \to e^{-np}$ as $p \downarrow 0$

(How many people are familiar w/ this approx?)

Suppose $p = p(n) = \frac{c \log n}{n}$ ← log base e.

$$= \lim_{n\to\infty} ne^{-c\log n}$$

$$= \lim_{n\to\infty} n^{1-c}$$

$$= \lim_{n\to\infty} n^{1-c}$$

$$= \begin{cases} 0 & \text{if } c>1 \\ \infty & \text{if } c<1 \end{cases}$$

So, we see that $E[X]$ has a sharp threshold where the $\mathbb{E}[\#\text{ of isolated verts}]$ goes immediately from 0 to ∞.

But, we'd love to know something stronger
→ that almost all graphs have 0 or ∞ isolated vertices.

Q: why don't we know that yet?

A: eg. There may be a few graphs with ∞ many isolated vertices while most have hardly any.

⇒ we need to additionally prove that the distribution of X is "concentrated" around $E[X]$

To do this we'll use Markov's & Chebychev's (which you proved & used on your HW).

- Case 1 (easy)
  If $p > \frac{\log n}{n}$ isolated vertices disappear w/ prob 1

  Let $\varepsilon > 0$: $\Pr(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon} = 0$ by Markov's.

  $\Rightarrow \Pr(X=0) \to 1$ as $n \to \infty$.

- Case 2 (harder)
  If $p < \frac{\log n}{n}$ we can't simply use Markov's

Instead we need to use the
"2nd moment method"
(i.e. apply Chebycher's)

To apply Chebycher's, we need to 1st calculate $\mathrm{Var}[X] = E[X^2] - (E[X])^2$.

$$E[X^2] = E\left[\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j\right]$$

(since X_i are 0,1.)

$$= n\,E[\textstyle\sum X_i] + n(n-1)\,E[X_i X_j]$$

$$= E[X] + n(n-1)\,\Pr(X_i = 1 \,\&\, X_j = 1)$$

$$= E[X] + n(n-1)\,\underbrace{(1-p)^{n-1}(1-p)^{n-2}}_{(1-p)^{2n-3}}$$

Now, by Chebycher's, we have:

$$\Pr\left(X \leq \tfrac{E[X]}{2}\right) \leq \Pr\left(|X - E[X]| \geq \tfrac{E[X]}{2}\right)$$

(which $\to \infty$)

$$\leq 4\frac{\mathrm{Var}[X]}{(E[X])^2}$$

$$= \frac{4E[X^2]}{(E[X])^2} - 4$$

$$= \frac{4E[X]}{(E[X])^2} + \frac{4n(n-1)(1-p)^{2n-3}}{(E[X])^2} - 4$$

(w/ $p < \frac{\log n}{n}$; $E[X] = n(1-p)^{n-1}$)

So, we have:

$$\lim_{n \to \infty} \Pr\left(X \leq \tfrac{E[X]}{2}\right) \leq 0 + \frac{4}{1-p} - 4$$

$$= 0$$

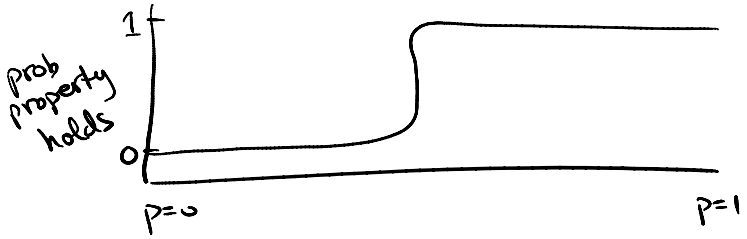
$\Rightarrow$ w/prob 1 there exist lots of isolated verticies.

Okay, after a long calculation, we've now got something interesting:

if $\begin{cases} p > \frac{\log n}{n} \to \text{no isolated verts w/prob 1} \\ p < \frac{\log n}{n} \to \text{lots of isolated verts w/prob 1.} \end{cases}$

Apart from its usefulness for understanding connectivity, this shows us an example of a sharp "threshold" for the existence of a property.

→ It turns out that almost all properties of random graphs have thresholds like this where they hold w/ prob 0 up to some $p(n)$ and w prob 1 above that $p(n)$.



☆ You'll prove that there is a similar threshold for the existence of Δ's on your HW.

The existence of thresholds like this is one of the fascinating things about Erdos-Renyi graphs.

In fact, there are deep results called 0-1 laws that say that "all" properties of random graphs have thresholds where

the property either holds w/prob 0 or 1 as $n \to \infty$.

Now that we've understood when isolated vertices disappear, we want to understand how much larger $p(n)$ should be to guarantee connectivity.

Q: How much bigger do you expect?

A: Not much, since isolated vertices are probably the last things that become connected.

⇒ It turns out that the graph becomes connected exactly at the same p where isolated vertices disappear!

Now, here's the complete summary of the evolution of connectivity in $G(n,p)$:

$p = \frac{d}{n^2}$: $E[\text{edges}] = d/2$ & components are all of size 1 or 2 w.p.1.

$p \ll \frac{1}{n}$: graph consists of only trees of size $O(\log n)$

$p = \frac{d}{n}, d<1$: There exist a constant # of cycles. Components are still of size $O(\log n)$

$p = \frac{1}{n}$: A giant component emerges of size $\Theta(n^{2/3})$. It is w.p.1 a tree.

$p = \frac{d}{n}, d>1$: A giant component of size $\Theta(n)$ emerges.

$p = \frac{1}{4}\frac{\log n}{n}$: Giant component swallows everything but a constant # of isolated verticies

$p = \frac{\log n}{n}$: All isolated verticies disappear & the graph becomes connected

↑ It may be surprising that these happen at the same time!

Q: How does this relate to our 4 universal properties of networks?

A: The fact that a giant component exists and that all other components are small aligns very nicely with what we saw in the web graph and other networks.

So → it's reasonable to imagine that the connectivity properties we discussed emerge naturally from a sequence of independent random connections!

☆ Interestingly, the diameter of these graphs is also "small" (you will prove 1 result of this form on your HW).

Conclusion

This simplistic, first-cut at a model of networks actually did pretty well.

It gave us insight into

It gave us insight into
2 of the 4 properties
we wanted to understand!

In the next few lectures, we'll focus on the other 2 properties & see if we can understand what causes them to emerge.

Next time: Heavy-tails.

Thanks!