The web graph & beyond - universal properties of networks.

...see .pptx for 1st part of the lecture...

Now, lets start to look at structural properties of the web graph.

Q: What properties should we look at?

A: Size (how many nodes?) → last time
Density (how many edges?)
Connectivity
Degrees
Diameter
Clustering (communities)
⋮

We'll now go through 4 of these in detail

① Connectivity

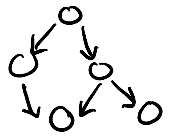
This is one of the most basic properties of a graph...

Q: Why might we care about it for the web graph?

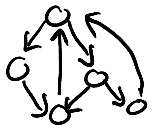
A: One reason is because we depend on crawlers to learn the graph, and this tells us something about how effective crawlers are.

Before we look at data from the web graph, let's review some important results about connectivity in directed graphs.

def: G is a directed acyclic graph (DAG) iff $\forall$ nodes $u \& v$, $u \neq v$, if $\exists$ a path from $u \to v$ then $\nexists$ a path from $v \to u$

e.g.  is a DAG

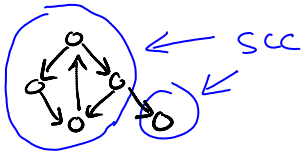
def: G is a strongly connected graph iff $\forall\ u, v$ $\exists$ a path from $u \to v$ & $v \to u$.

e.g.  is strongly connected.

def: A strongly connected component (scc) is a set of nodes S st:

(i) $\forall\ u, v \in S$ $\exists$ a path from $u \to v$ & $v \to u$

(ii) There is not T st $S \subseteq T$ & T satisfies (i).



Def: A weakly connected component (wcc) is a set of nodes S s.t.
(i) $\forall u, v \in S$ $\exists$ a path either from $u \to v$ or $v \to u$
(ii) There is not a T s.t. $S \subseteq T$ & T satisfies (i)

Now, lets make a few observations about these definitions.

Observation 1:

A DAG has no SCCs larger than 1 node.

Observation 2:

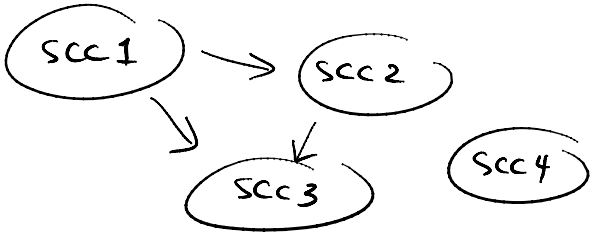
We can partition any graph into a set of SCCs
i.e. every node is in exactly 1 SCC.

Q: Who can give me a proof of Observation 2?

Pf:
- Every node is in at least 1 SCC since alone it is strongly connected.
- No node can be in 2 SCCs, since, if it were, then the combination of the two would also be strongly connected, which violates part (ii) of the definition.

Q: What can we say about the graph connecting the SCCs?

i.e.



A: It must be a DAG!

Q: Who can give me a proof of this?

Q: Who can give me a proof of this?

A: If it wasn't, then ∃ paths st $SCC_i \rightleftarrows SCC_j$, which means the merger of these 2 is also a SCC, which is a contradiction of the fact that SCC_i & SCC_j are SCCs.

Key Consequence:

a natural way to think about connectivity is to characterize the SCCs and the DAG connecting them.

Now, lets look at the connectivity of the web graph.

Q: How do we gather info on the web graph?

A: ... need to do a really big crawl...
Basically pick a bunch of starting points and do Breadth First Searches at each one.

you'll do this on HW2!

★ But, the web graph isn't static!
... the link structure changes constantly...

Many difficult issues in doing a crawl, but we'll ignore them for now.

I'll show you data from Broder et al. [2000], which is one of the largest statistical studies of the web in the literature.

They used an AltaVista crawl, which had 203 million URLs & 1466 million links

had 203 million URLs & 1466 million links

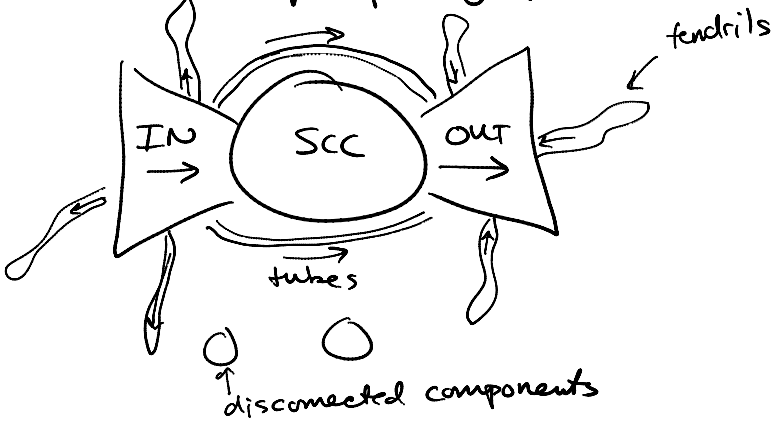
Note the computational issues in dealing with such a large data set.

Q: What do you expect?
How big will the largest SCC be?

On to the data

In the directed graph:

- The largest SCC is ≈30% of the nodes (56,463,993 nodes)
- The second largest is ≈150000 nodes
  → much smaller than the giant component
- the largest weakly connected comp ≈ 90%
- A "map" of the graph is:



SCC ≈ 30%
IN ≈ 20%
OUT ≈ 20%
the rest ≈ 30%

(see paper for the real #s)

Q: Can you give a "story" explaining this figure?

A: One interpretation:

IN → new pages, yet to be discovered and linked to

OUT → insular corporate pages

SCC → the "core"

---

(2) Diameter

Now that we understand a lot about connectivity, the next natural question is how long are the paths?

This is what the <u>diameter</u> tells us about.

There are 2 common notions:

1) maximal distance between any two (connected) nodes
2) the average distance between any two (connected) nodes

<u>Remarks</u>:

- both are sometimes called "diameter". I'll try to use maximal diameter & average distance to distinguish.
- It's typically only useful to talk about the diameter of connected components... so that things are finite. If "connected" is dropped, the diameter of a disconnected graph is ∞.

<u>Q:</u> what do you expect the diameter to be?

<u>A1</u>: The web graph is big and has directed edges, which make paths long... so the diameter should be big → 100-500?

<u>A2</u>: 6?... because of "6 degrees of

A2: 6½ - Because of "6 degrees of separation"

Here are the #s, again from Broder et al

A1: 75% of the time there is no directed path

A2: Max Diameter of SCC ≥ 28
Max Diameter of graph ≥ 503
probably more like 900.

A3: Avg Diameter of directed graph ≈ 16.2
if edges are treated as undirected ≈ 6.8

⇓

"6 degrees of separation"

Q: How has the diameter changed over time? (Of course it grows, but how quickly?)

A: Diameter ≈ log(# nodes)

→ see .pptx for a figure

Big Question: Why do such short paths exist?

→ we'll talk about this a lot more later in the course.

---

③ Degree

We'll now move to perhaps the 2nd most important characteristic - degree

Q: Why do we care about degree?

A: Many reasons...
- Gives us more insight into connectivity
- ★ Tells us about the importance of nodes.

- Tells us about the "density" of the graph

Note: Since we're talking about directed graphs, there are two notions of degree: in-degree & out-degree

☆ What we'll be looking at is the "degree distribution"

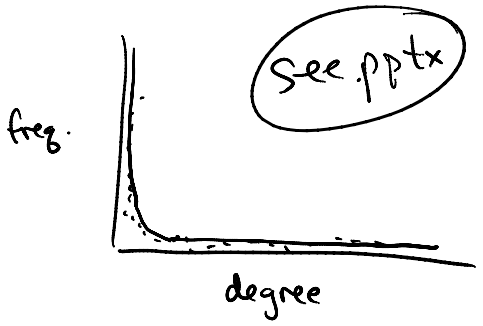
i.e. the # of nodes that have a particular degree

Let's jump right to some data from the web graph...

Q: What do you expect the distribution to be?

A: Maybe a truncated Normal/Gaussian?

... because of the Central Limit Thm everything with a large # of samples is normal, right?



standard scale hides all useful information

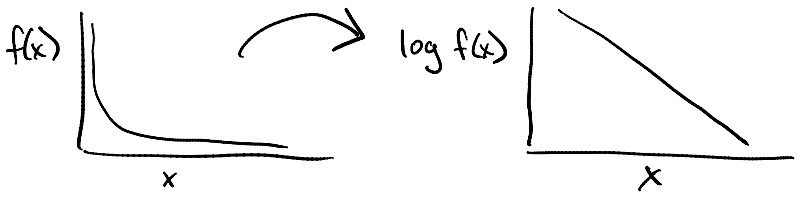
↓

the trick is to change the scale so that the points are linear

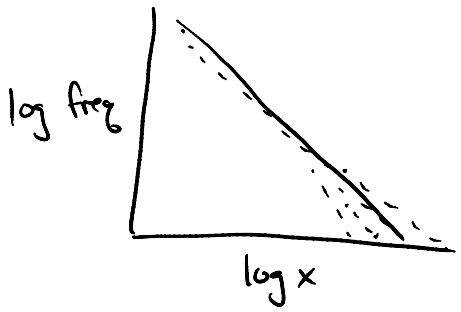
⇒ which tells you the form of the pdf of the distribution.

e.g. (Exponential Distribution)

pdf: $f(x) = e^{-\lambda x}$ ⟶ $\log f(x) = -\lambda x$



In the case of the web data, the right shift is to a log-log scale.



see .pptx

(this plot looks the same for both in-degree & out-degree)

The takeaway is simple:
The degree distribution seems to be linear on a log-log scale.

letting $f(x)$ be the p.d.f., this means $f(x) \approx cx^{-\alpha}$

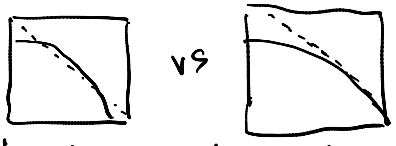
since $\underbrace{\log f(x)}_{\text{y-axis}} = \log c + (-\alpha)\underbrace{\log x}_{\text{x-axis}}$

This is called a "Pareto Distribution" or a "power-law distribution"

It is an example of a heavy-tailed distribution ... and we'll talk about these a lot more later in the course.

2 important remarks for later

- Is it really linear?
- Slope ⇒ parameter α ... how do you fit the line?

e.g.  vs

we'll talk about these issues later...

For now, let's just get a little insight for this by comparing with the Normal Distribution (Gaussian) ... since this is what we expected.

→ A few key differences

1) Heavy-tail ⇒ large variance.

$$\lim_{x \to \infty} \frac{Pr(P>x)}{Pr(N>x)} = \infty$$

⇒ · huge difference between max/min samples

e.g. heights ~ Normal & $\frac{max}{min} \approx 5$

city sizes ~ Pareto & $\frac{max}{min} \approx 150{,}000$

2) No typical value for Pareto
Normal is typically around its mean

☆ We'll look much more deeply at heavy-tails later in the course

Big Question: Why isn't the degree Distribution ≈ Normal?

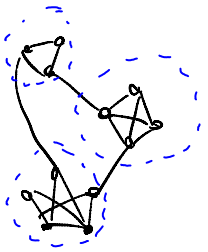
↳ we'll answer this soon.

---

④ Clustering

You probably haven't studied Clustering of graphs before... but the idea is to try to understand whether clusters/communities of nodes are likely, and how big they tend to be.

e.g.

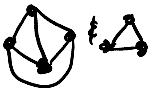


This can be very important for search engines since you want the results to be "varied" and not all from the same cluster.

e.g. if you search for "tigers" you want some results about animals, some about sports, etc. and you don't want them all to be from the same site.

You can think of clustering as a notion of "cliquoishness".

[Remember: A clique is a set of nodes where all the nodes are neighbors

e.g. ]

A node is "highly clustered" if it is much more likely than random for cliques to exist. ⟺ highly clustered a.k.a. "homophilly"

Q: Should we expect the web graph to be highly clustered?

A: Yes!

A: Yes!

If A links to B & C, its likely that B & C are related and thus have a link between them.

"Clustering" is a somewhat vague term, and so there are a lot of different definitions that are used.

Here are 2 of the most common definitions, they both rely on looking for the existence of triangles

you've working with these in your HW

def The overall clustering coefficient of an undirected graph G is

$$Cl(G) = \frac{3 \times (\#\text{ of } \Delta\text{s in the graph})}{\#\text{ of connected triples}}$$

a triple centered at i is an unordered pair of verts connected to i

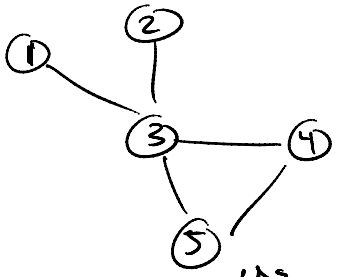
→ so the % of time when a node has 2 neighbors, they are connected.

def The average clustering coefficient of an undirected graph G is

$$Cl^{avg}(G) = \frac{1}{n}\sum_i Cl_i(G)$$

where $Cl_i(G) = \frac{\#\text{ of } \Delta\text{s centered at } i}{\#\text{ of triples centered at } i}$

e.g. $G =$



$$Cl(G) = \frac{3 \times 1 \leftarrow \#\text{ of } \Delta s}{6 + 1 + 1} = \frac{3}{8}$$

↑ triples at 3 ↑ at 5 ↑ at 4

$Cl_3(G) = \frac{1}{6}$ — 1 Δ among 6 neighbor triples

$Cl_5(G) = Cl_4(G) = 1$

$Cl_1(G) = Cl_2(G) = 0$

$\Rightarrow \quad Cl^{avg}(G) = 13/30$

These 2 notions seem very similar, but they can be very different.

On your HW you will give an example where $Cl(G) = 0$ & $Cl^{AVG}(G) = 1$ (which is the max possible difference!)

Now to the web:

Depending on the data set, the web has:

$Cl^{avg}(web) \approx 0.2 - 0.5$

Q: Is this large?

A: Depends on the # of edges in the graph...

Clustering coefficient is meaningless in isolation.

For a graph w/ the same density as the web

the clustering coefficient would be

$\approx .0001 - .001$

if there was no correlation between edges

So the web is "highly clustered"

Big Question: How can we identify large clusters/communities?

→ we'll return to this if we have time.

Now we've gone through 4 key properties and started to understand the structure of the web graph

The take aways are that the web graph:

1) Has a giant weakly connected component.
2) Has a small diameter
3) Has a heavy-tailed degree distribution
4) is highly clustered.

You might think that these properties are things that make the web graph "special"

But, it turns out that these 4 properties are "universal" in the sense that nearly all "networks" that people study exhibit them...

For the rest of the lecture I'll try to convince you of the "universality" of these properties...

[... the rest of the lecture is on .pptx]