Graph Data

Everything Data

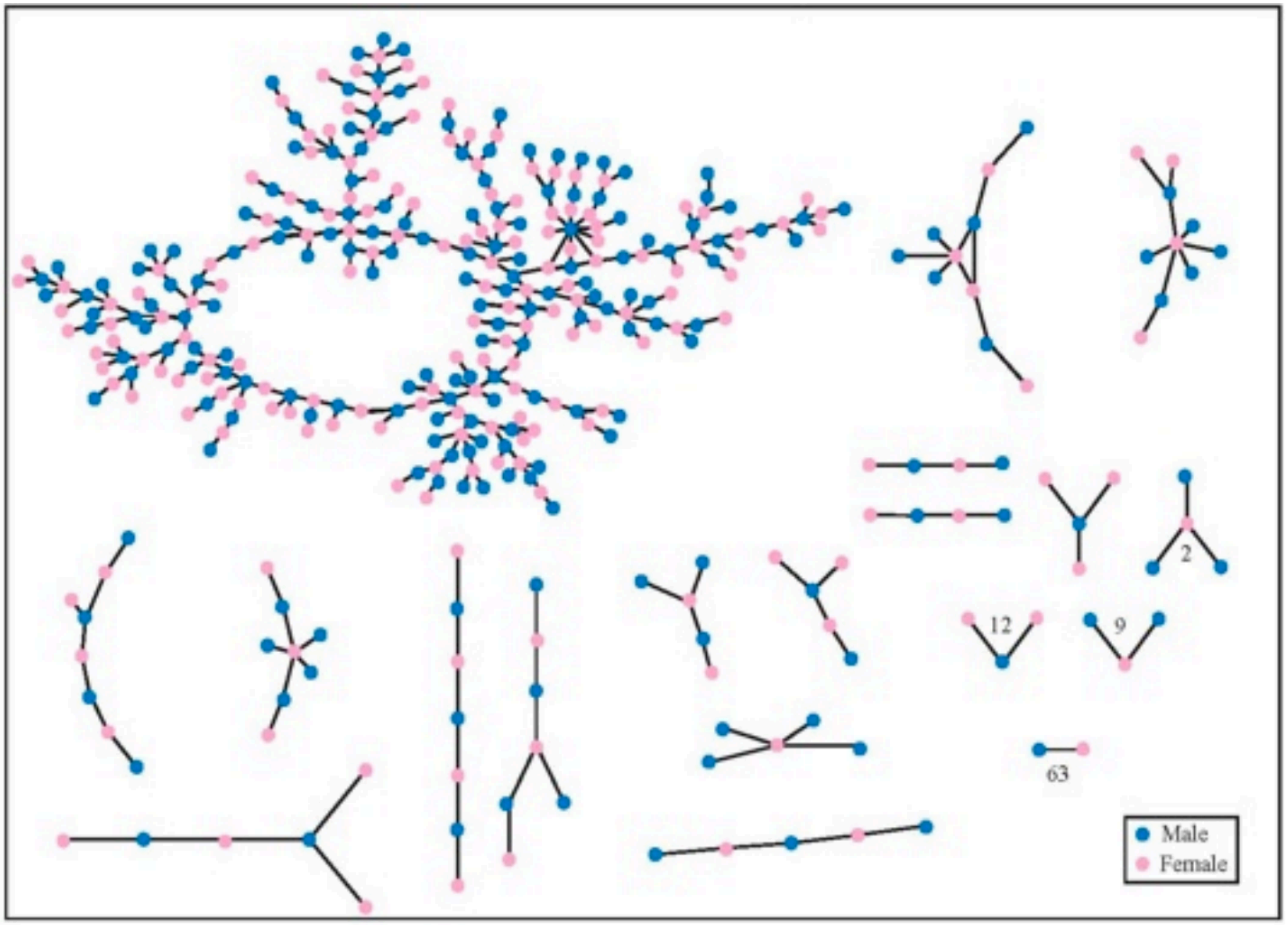
CompSci 216 Spring 2015

DUKE
COMPUTER SCIENCE

Announcements (Wed. Mar. 4)

- **Homework #8** will be posted by tomorrow
  - Due midnight Sunday 3/15
    - Right when the Spring Break ends
- **Proposal feedback** to be emailed during Spring Break
- **Next project milestone**: progress report due Monday 3/30

An example



Nodes: students in a large American high school

Edges (undirected): romantic relationships during a 18-month period being studied

Bearman, Moody, Stovel. *American Journal of Sociology*, 110(1), 2004

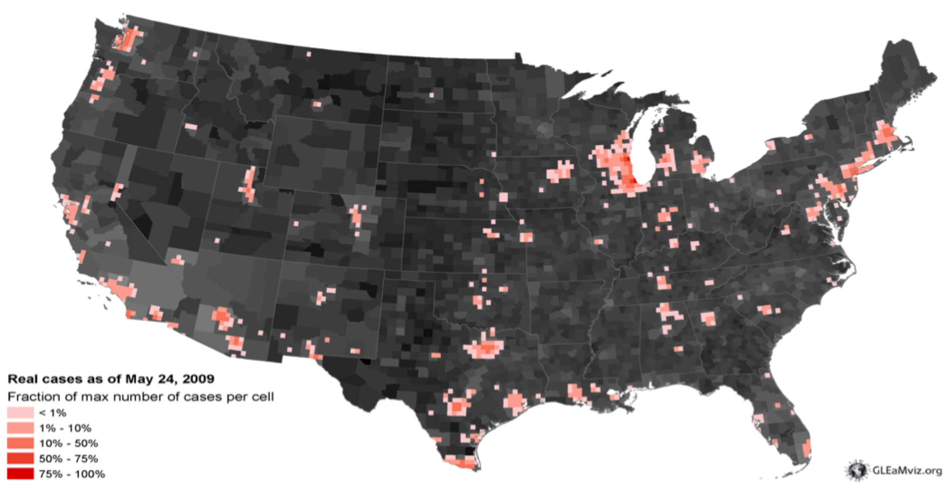
http://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch02.pdf

Global Epidemic and Mobility

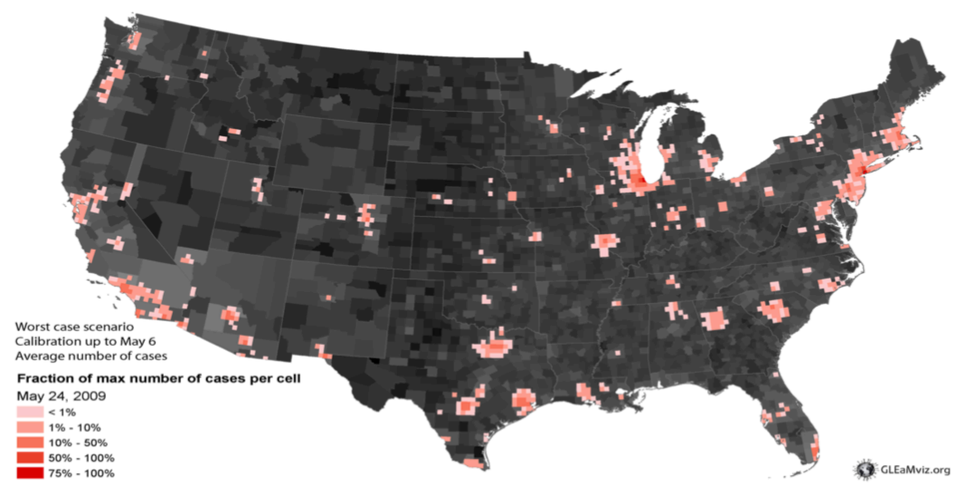
http://www.gleamviz.org/
http://barabasilab.neu.edu/courses/phys5116/content/Class1_NetSci_2012/01_CLASS_2012_Introduction.pdf

Predicting the H1N1 pandemic

Real

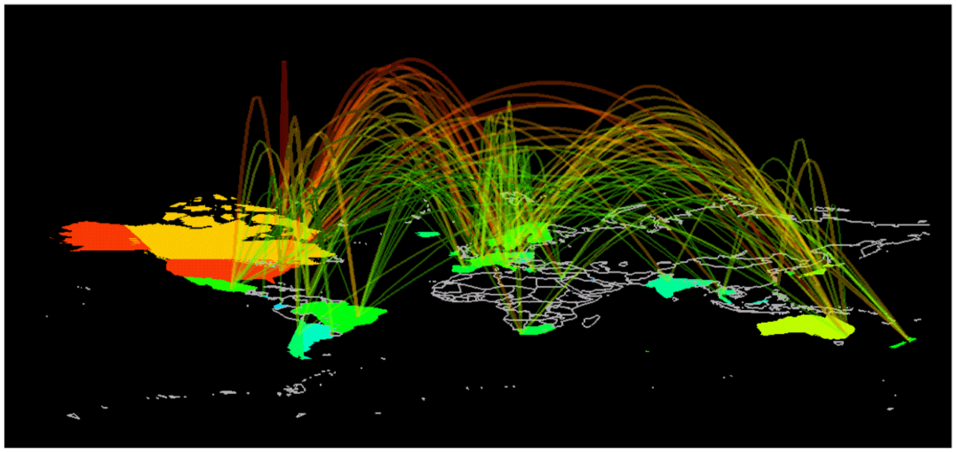


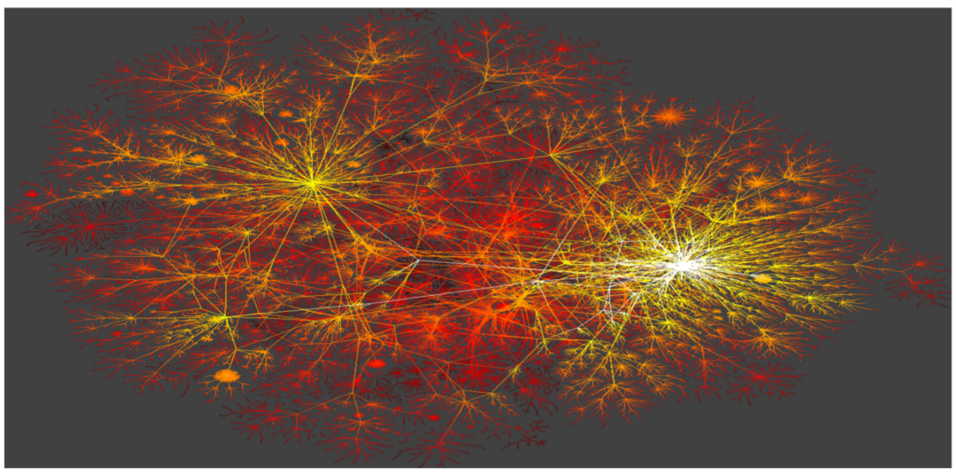
Projected

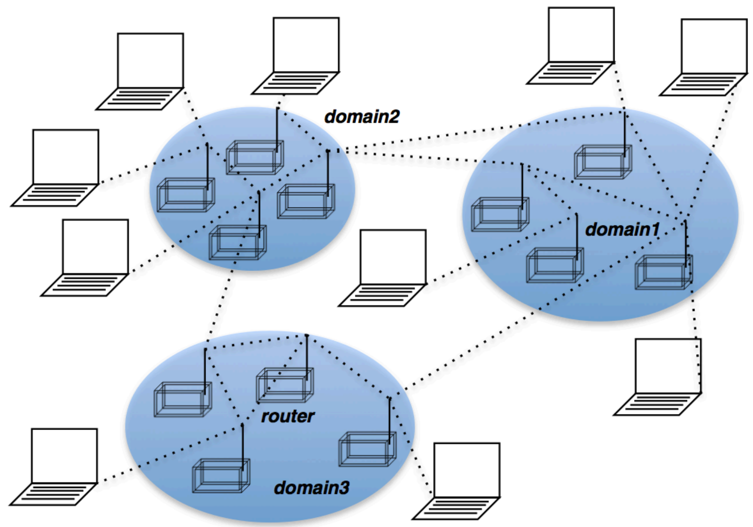


http://www.gleamviz.org/
http://barabasilab.neu.edu/courses/phys5116/content/Class1_NetSci_2012/01_CLASS_2012_Introduction.pdf

The Internet







http://barabasilab.neu.edu/courses/phys5116/content/Class1_NetSci_2012/01_CLASS_2012_Introduction.pdf

Social Graph behind Facebook

Keith Shepherd's "Sunday Best." http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/
http://barabasilab.neu.edu/courses/phys5116/content/Class1_NetSci_2012/01_CLASS_2012_Introduction.pdf

Nodes, edges, and degrees

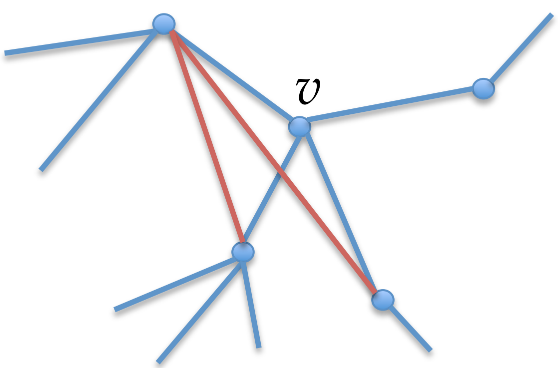
- A graph is specified by (V, E)
  - V is a set of ***nodes*** (or ***vertices***)
  - E is a set of ***edges***, each connecting two nodes
    - Can be ***undirected*** (e.g., friendships) or ***directed*** (e.g., links between Web pages)
- *Degree* of a node v: # edges incident to v
  - For directed graphs, a node has an *in-degree* (# incoming edges) and an ***out-degree*** (# outgoing edges)

Paths and connectivity

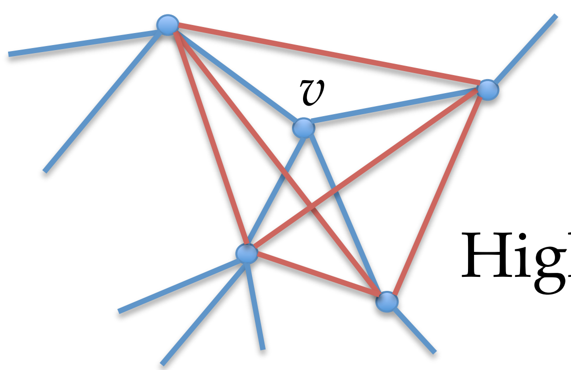
- A ***path*** is a walk (along edges) from one node to another in a graph
  - For directed graphs, edge directions matter
- ***Distance*** from node v_1 to node v_2: length of shortest path from v_1 to v_2 (in # edges)
- A (strongly) ***connected component*** is a subset of nodes such that
  - Every node in the subset has a path to every other
  - This subset is *maximal*; i.e., it is not part of some larger set with the above property

Some important statistics

- Degree distribution
- Distance distribution
  - *Diameter*: maximum distance
- ***Clustering coefficient*** of node v is the probability that two nodes directly linked to v are also directly linked to each other



$C(v) = 2 / (4 \times 3/2) = 1/3$



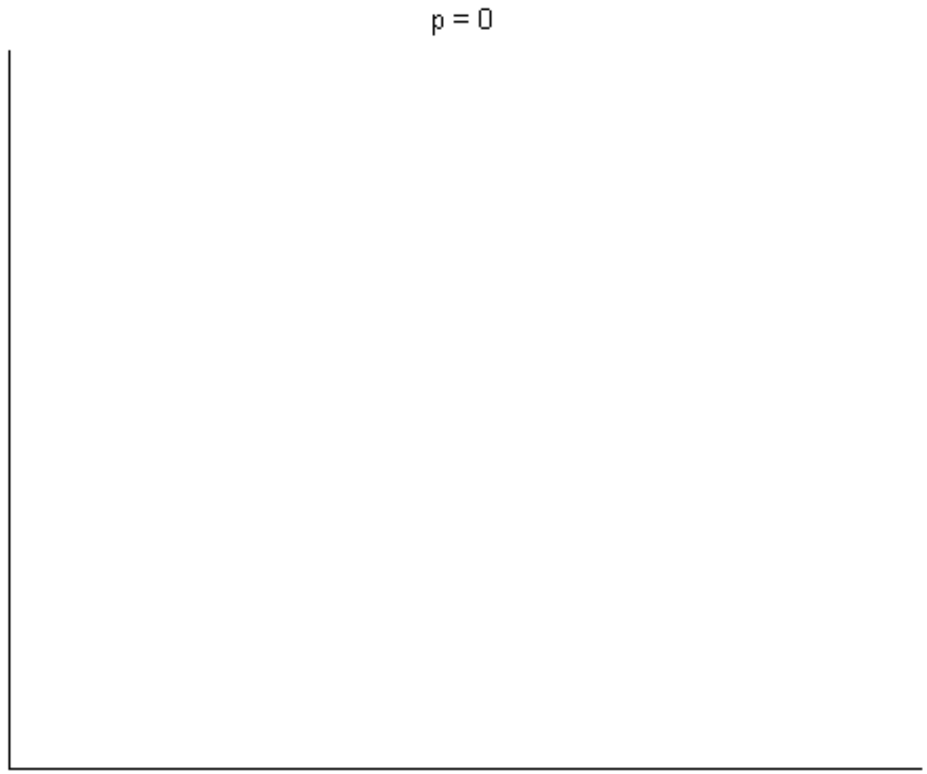
$C(v) = 6 / (4 \times 3/2) = 1$

Between 0 and 1

Higher = more "clustered"

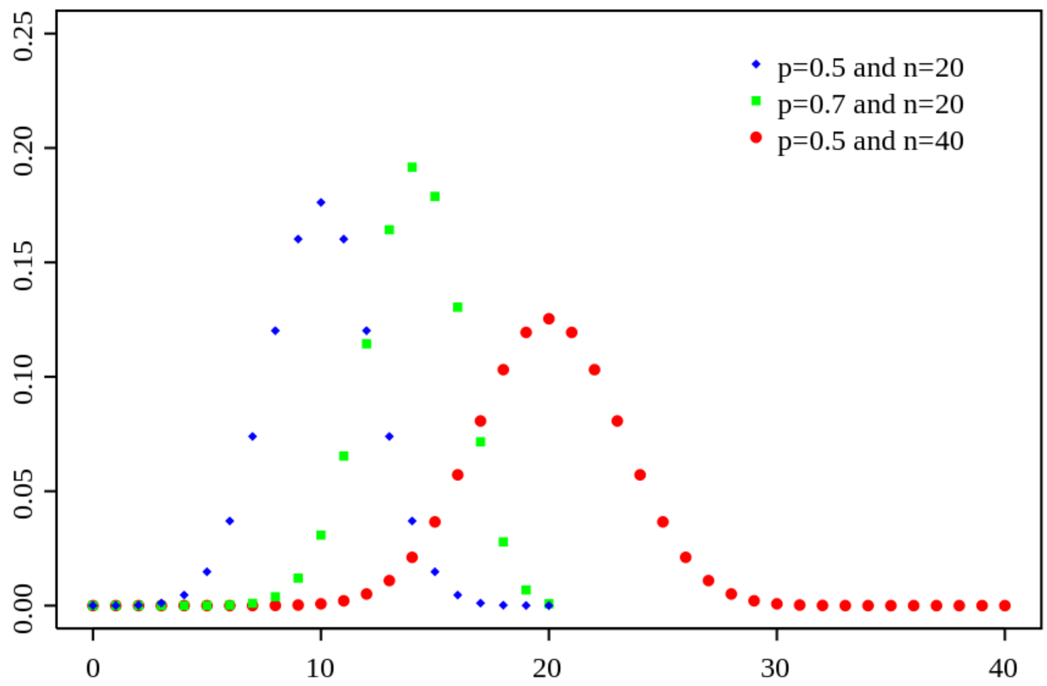
A simple model: random graph

N nodes; draw an edge between each pair by a preset probability *p*



https://www.cs.purdue.edu/homes/dgleich/demos/matlab/random_graphs/erdosreyni.html

Degrees in G_{random}

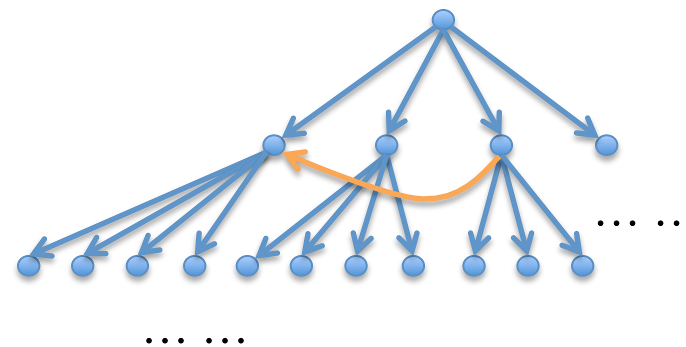


- *Binomial* with mean $\langle k \rangle = Np$
  - Approximated by bell curve

http://en.wikipedia.org/wiki/File:Binomial_distribution_pmf.svg

Distances in G_{random}

- Diameter = $O(\ln N / \ln \langle k \rangle)$
  - Imagine a *breadth-first search*
  - Each hop expands the neighborhood by about $\langle k \rangle$
    - We might get a previously visited node, but the chance is small unless a significant portion of the graph has already been covered
  - The probability of missing a node after enough hops is very small



Clustering coefficients in G_{random}

- Give node v, for each pair of v's neighbors, the probability is p
  - By definition of random graph
- So v's clustering coefficient is $p = \langle k \rangle / N$

Case study: social network

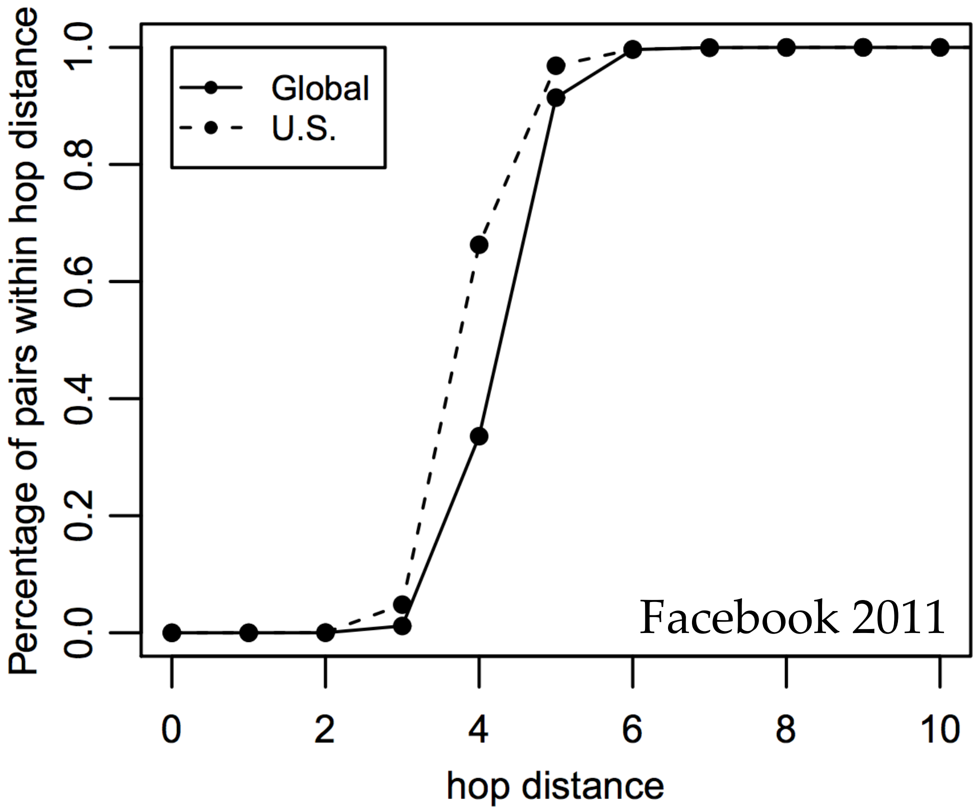
Keith Shepherd's "Sunday Best." http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/
http://barabasilab.neu.edu/courses/phys5116/content/Class1_NetSci_2012/01_CLASS_2012_Introduction.pdf

6 degrees of Kevin Bacon

Everyone is six or fewer steps away from any other person in the world, via a chain of "a friend of a friend"

http://www.hollywoodreporter.com/heat-vision/kevin-bacon-google-six-degrees-369927

Distance distribution, Facebook



Small-world property: most nodes are not directly connected, yet they can be reached from every other by a small number of hops

In this regard, social networks are very similar to random graphs!

But wait a second…

Ugander et al. *The Anatomy of the Facebook Social Graph*, 2011. http://arxiv.org/pdf/1111.4503.pdf

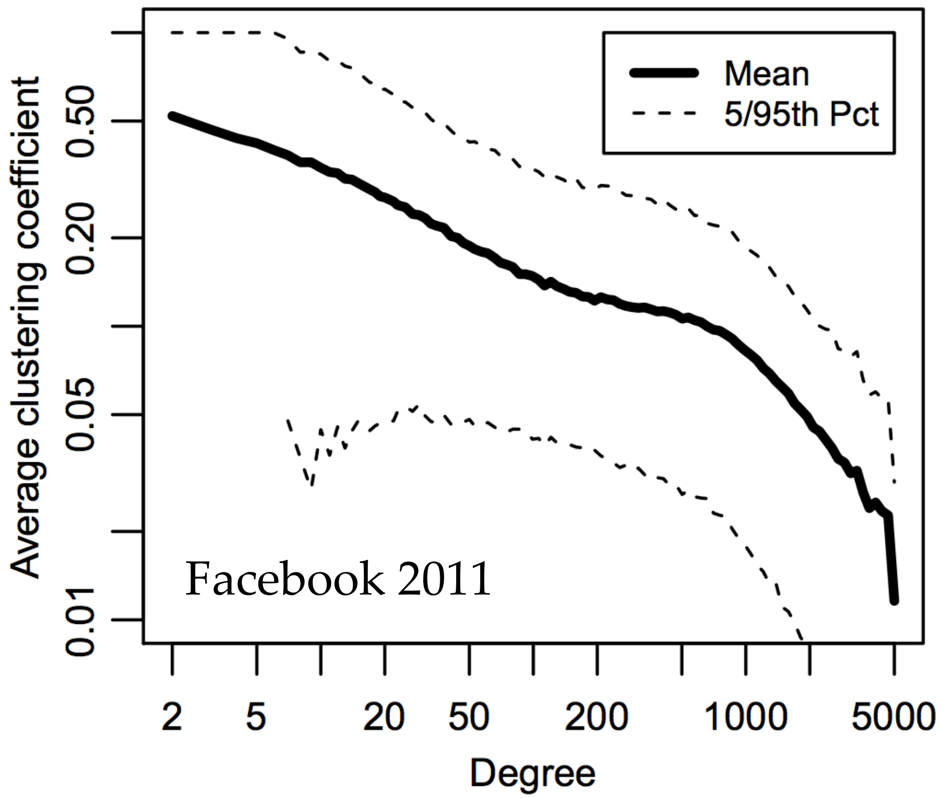
Prevalence of *triadic closure*

If two people have a common friend, then there is a higher chance that they will become friends

- High "clusteredness" as measured by clustering coefficient

http://good-thinking.co.uk/tag/befriend/

Clustering coefficient, Facebook



Really high compared with a random graph!

- $C_{\text{random}} \approx \langle k \rangle / N$
- Average # friends < 200
  - Median: 99
- # nodes: 721 million

In this regard, social networks are very "clustered" and very different from random graph!

Ugander et al. *The Anatomy of the Facebook Social Graph*, 2011. http://arxiv.org/pdf/1111.4503.pdf

Puzzle

How can a social network be so "clustered" yet offer such short distances?

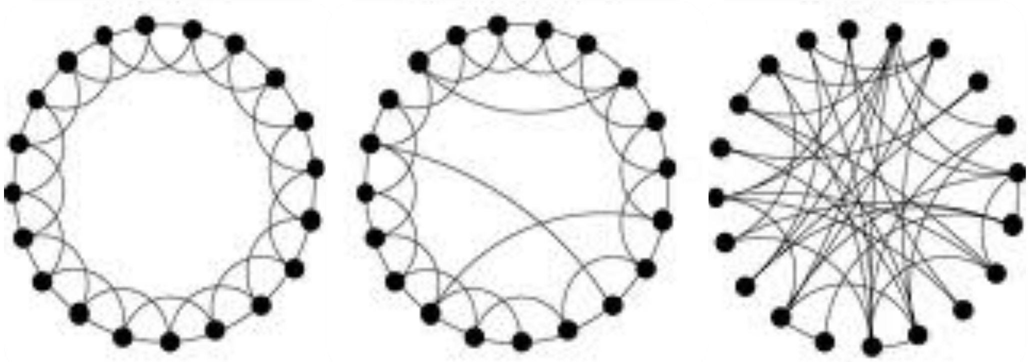
E.g., road networks have "local links"

- Relatively high clustering coefficients
- But not a small world!

http://barabasilab.neu.edu/courses/phys5116/content/Class4_NetSci_2012/04_CLASS_2012_Scale-Free_Property.pdf

Explanation by *Watts-Strogatz*

- Start with a lattice network
- "Rewire" every edge randomly with probability β



$$\langle d \rangle = \frac{N}{2k} \qquad \langle d \rangle = \frac{\ln N}{\ln k}$$

$$C = \frac{3}{4} \qquad C = \frac{k}{N}$$

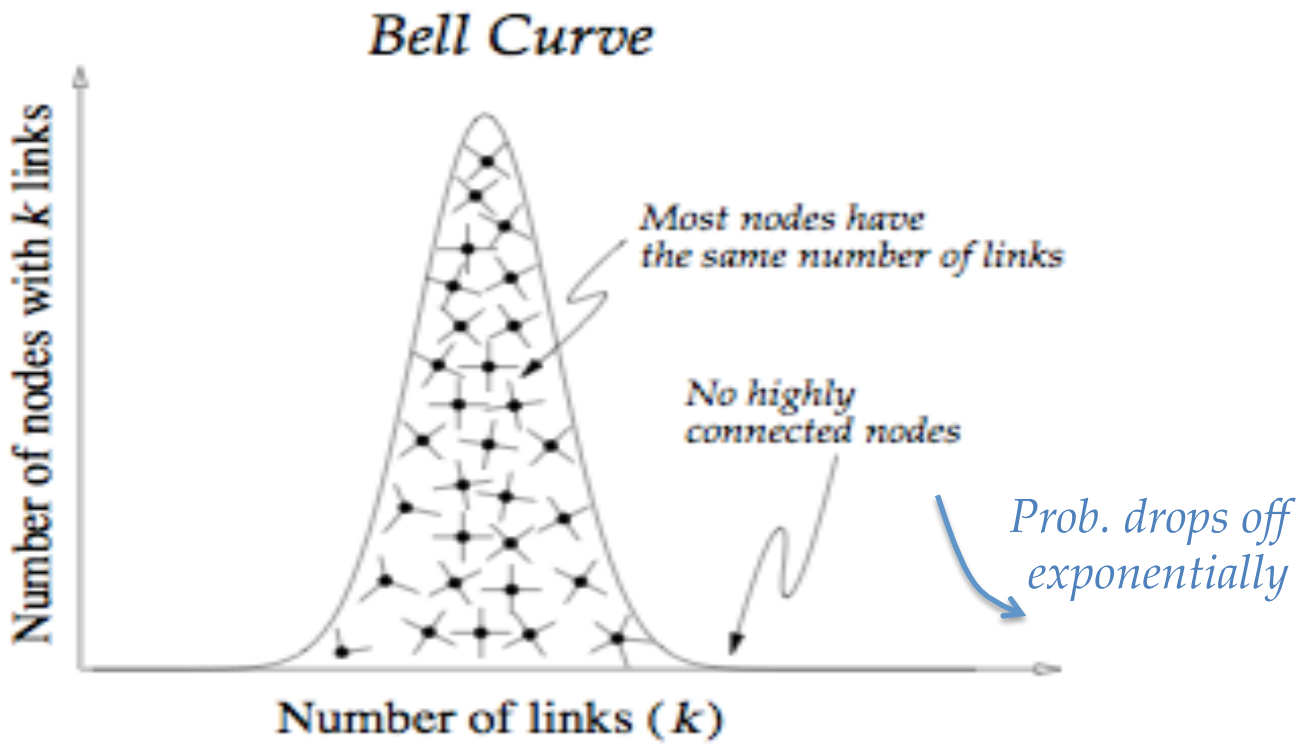
$$0 \xrightarrow{\beta} 1$$

It takes a lot of randomness to ruin "clusteredness," but a very small amount to overcome "locality"

So are social networks Watts-Strogatz?

http://barabasilab.neu.edu/courses/phys5116/content/Class5_NetSci_2012/05_CLASS_2012_The_Small_World.pdf

Degree distribution, random

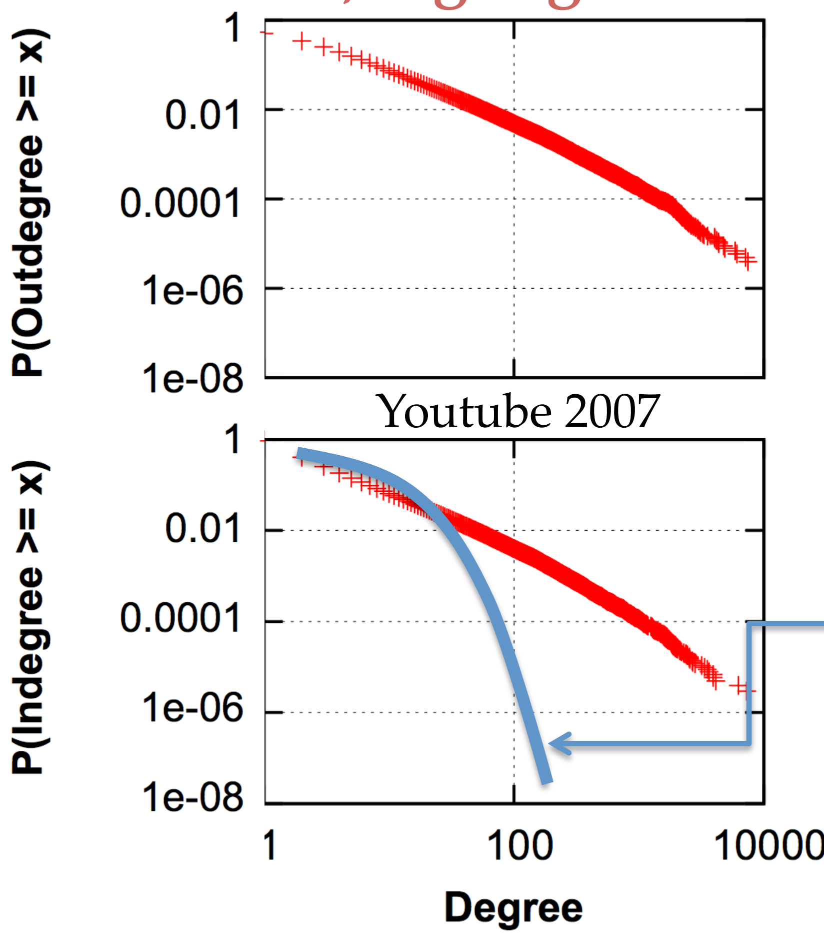


- *Watts-Strogatz* also gives you a bell curve
  - If social networks really behave this way, there will be no individuals who are either immensely popular or extremely recluse

http://barabasilab.neu.edu/courses/phys5116/content/Class4_NetSci_2012/04_CLASS_2012_Scale-Free_Property.pdf

Degree distribution, Youtube*

Cumulative, log-log



Distribution has a "*heavy tail*"; i.e., some individuals (albeit a small number) have a huge number of friends

An exponential tail (e.g., bell curve) would look like this

**We will come back to Facebook later*

Mislove et al. , 2011. "Measurement and analysis of online social networks." *IMC* 2007
http://conferences.sigcomm.org/imc/2007/papers/imc170.pdf

Explanation: rich-get-richer

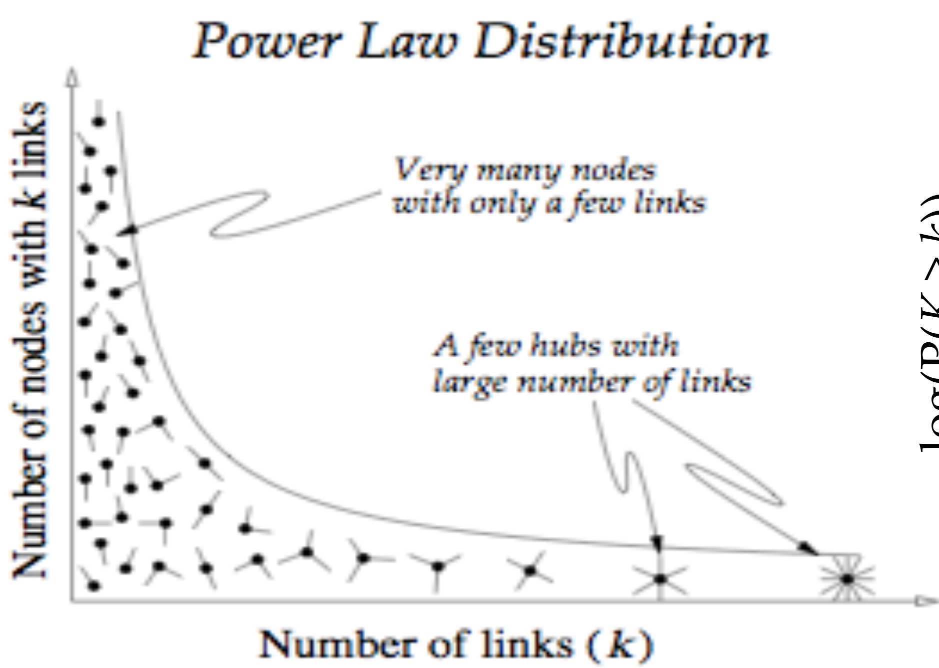
Barabasi-Albert

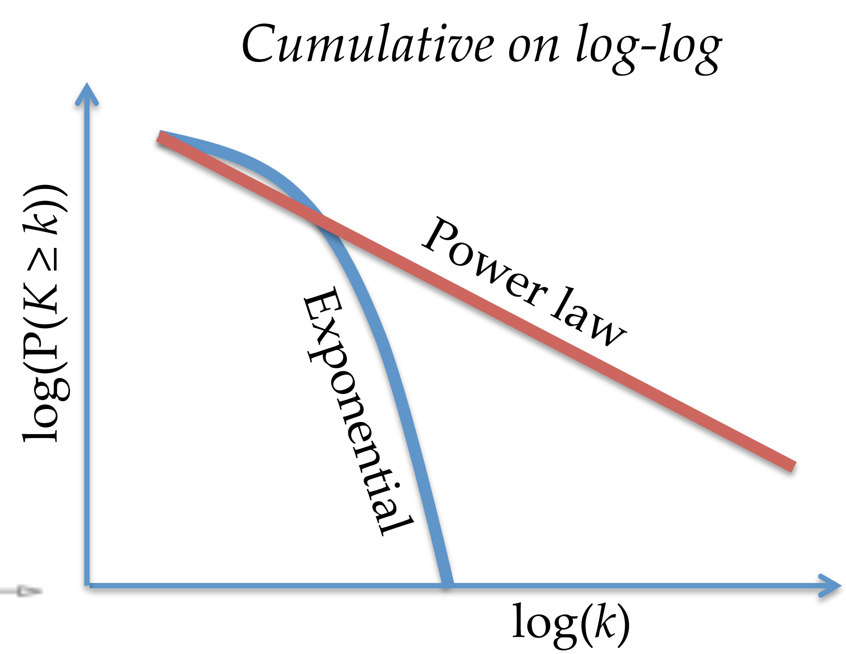
- Start with a (connected) graph of size m_0
- Add new nodes one at a time
  - Each connects to $m \leq m_0$ existing nodes with probability proportional to # existing edges they already have

See http://en.wikipedia.org/wiki/Preferential_attachment for a more general formulation

Implication of *Barabasi-Albert*

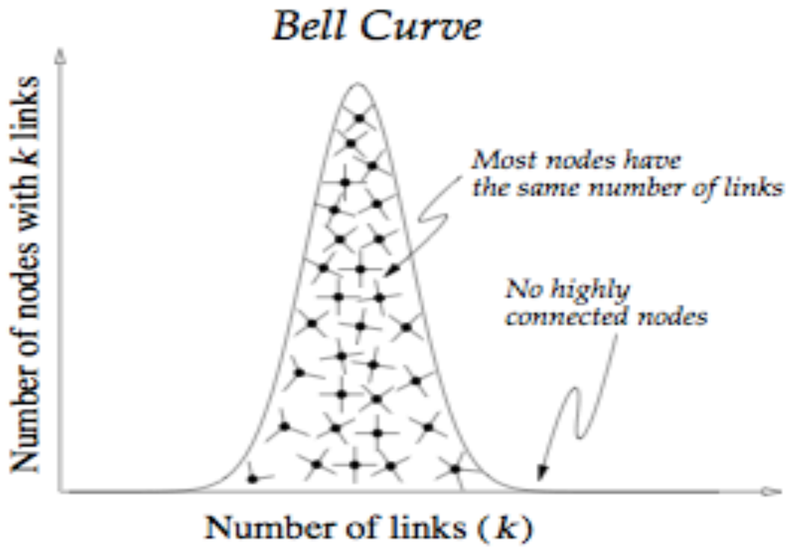
- Degree distribution: ***power law***
  $P(k) \propto k^{-\gamma}$, or equivalently $P(K \geq k) \propto k^{-\gamma+1}$
  – Sometimes such graphs are called ***scale-free***

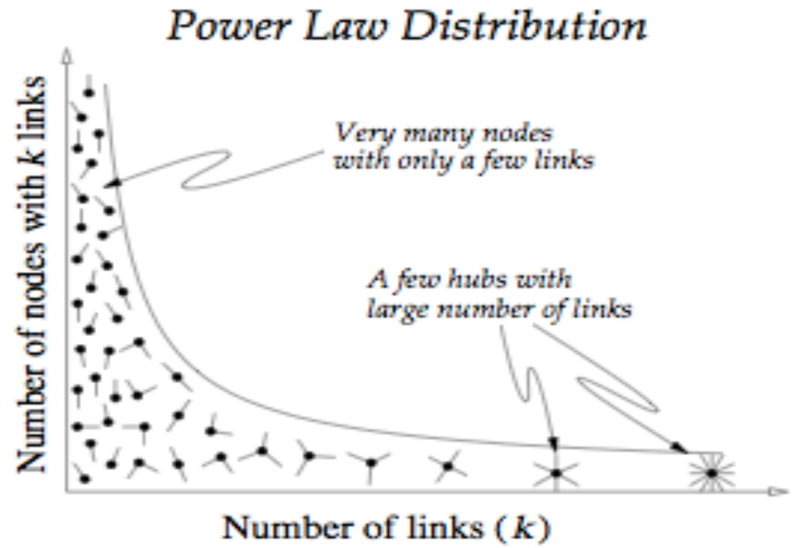


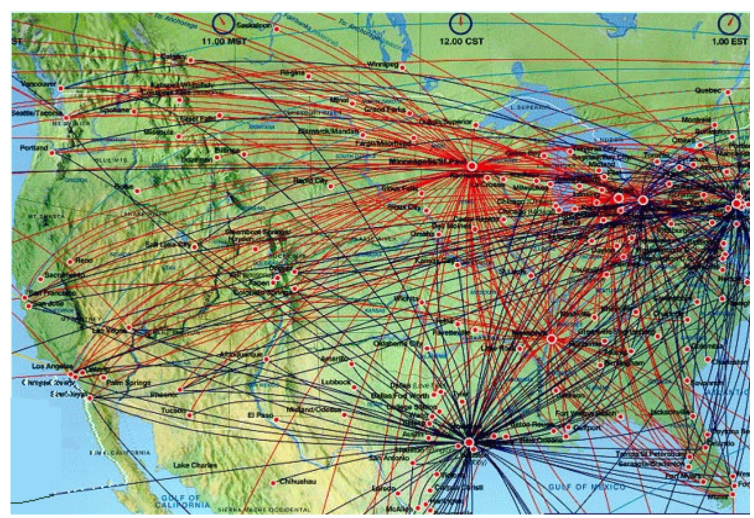


http://barabasilab.neu.edu/courses/phys5116/content/Class4_NetSci_2012/04_CLASS_2012_Scale-Free_Property.pdf

Exponential vs. power law







http://barabasilab.neu.edu/courses/phys5116/content/Class4_NetSci_2012/04_CLASS_2012_Scale-Free_Property.pdf

Other implications of *B-A*

- Average distance $\propto \ln N / (\ln \ln N)$
  - Small world, again
- Empirically, clustering coefficient = $O(N^{-0.75})$
  - Better than random, but still not as clustered as real social networks
- A few highly connected hubs hold network together
  - *Robust* against random node failures, yet
  - *Fragile* again targeted attacks

Note:

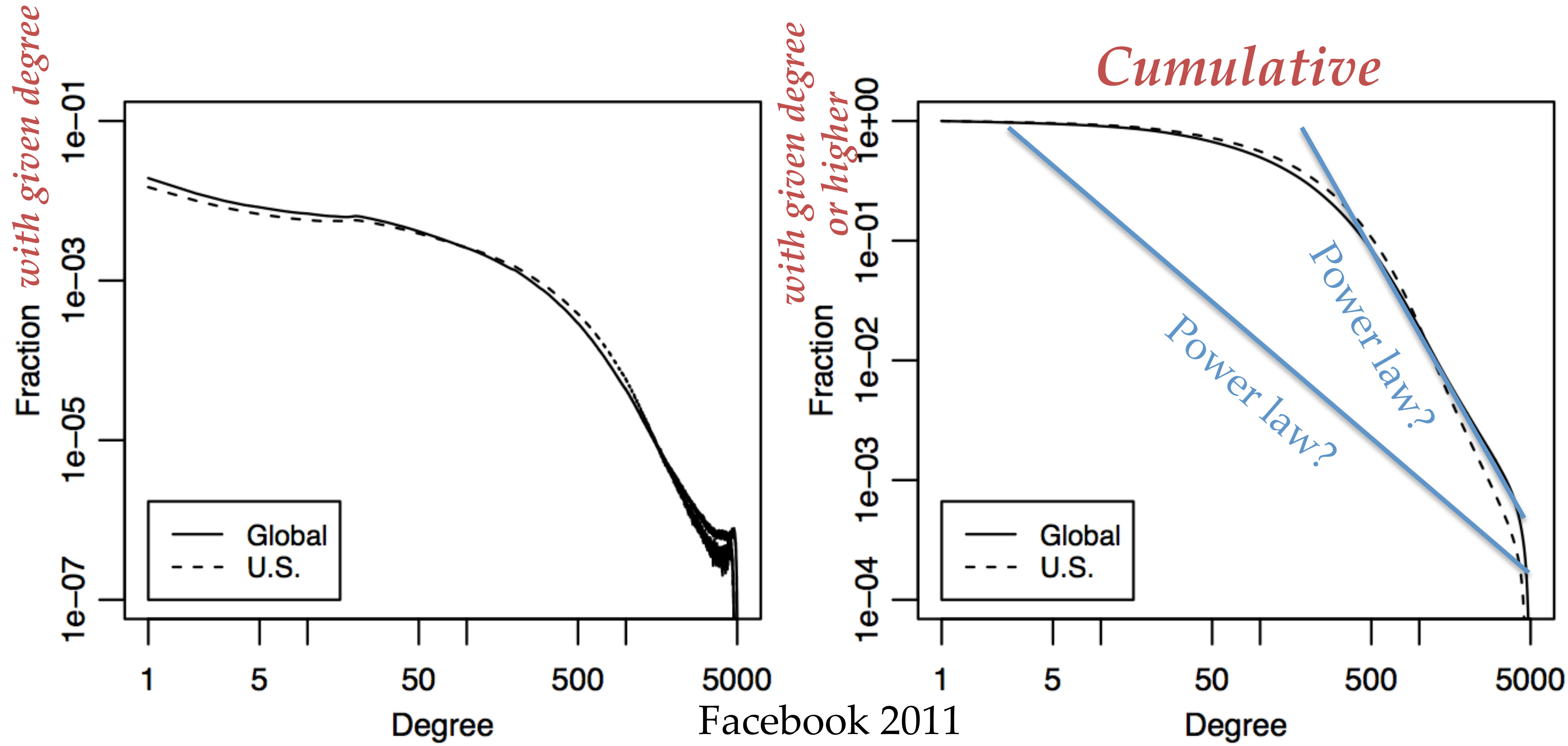
- *Barabasi-Albert* is just one way to get power law graphs
- Implications above don't necessarily follow from having a power law degree distribution
- A good read: Li et al. "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications." *Internet Mathematics*, 2005

Power law observed in real life

- Internet backbone, Web graph, many social networks (including co-authoring and co-acting graphs), protein-protein interaction network, etc.
  - At least for some range of k
- But oftentimes researchers rush to conclusion
- *… validation of power-law claims remains a very active field of research…*

http://en.wikipedia.org/wiki/Power_law#Validating_power_laws

Degree distribution, Facebook



Facebook 2011

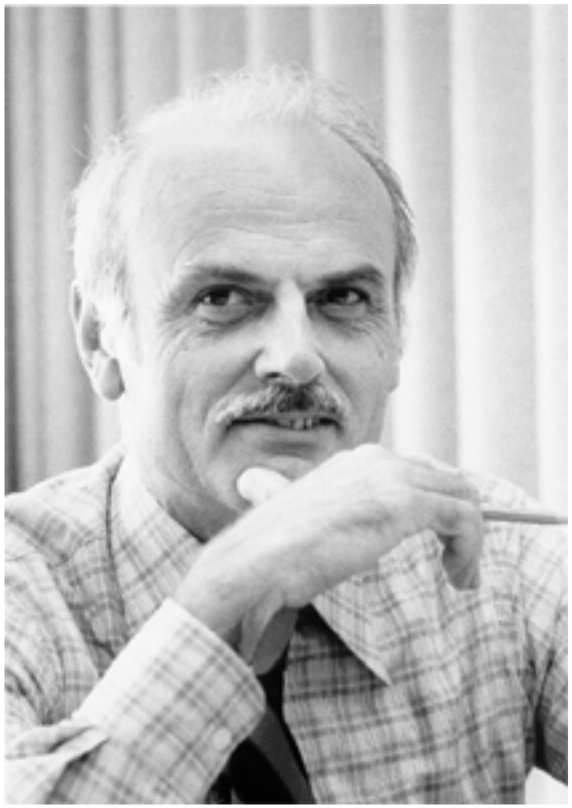
Visual identification is often useful but can sometimes mislead
(If you have to, always use the cumulative distribution)

Ugander et al. *The Anatomy of the Facebook Social Graph*, 2011. http://arxiv.org/pdf/1111.4503.pdf

Recap

- Simple stats to keep in mind when looking at a big graph
  - Degree distribution, distance distribution, clustering coefficient
- Interesting characteristics of some graphs
  - Power law, small world, triadic closure
- "All models are wrong, but some are useful." – George E. P. Box

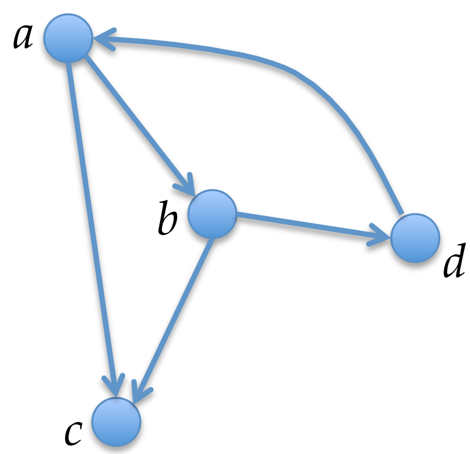
Representing graphs



A database person or mathematician?

http://en.wikipedia.org/wiki/File:Edgar_F_Codd.jpg http://en.wikipedia.org/wiki/File:Carl_Friedrich_Gauss.jpg

Relational representation



Store edges in a table `edge(src, tgt, ...)`

src	tgt
a	*b*
a	*c*
b	*c*
b	*d*
d	*a*

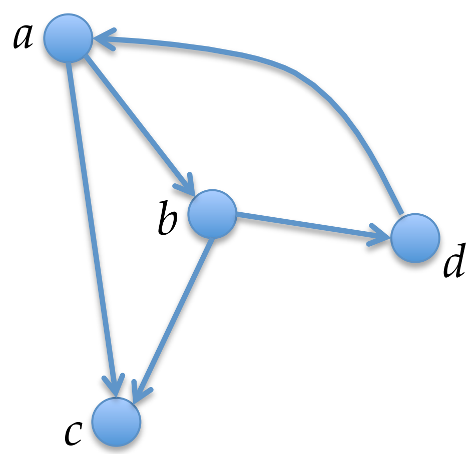
Can also include edge properties, e.g., weight, type, etc. as additional columns in "..."

Optionally, store nodes in a table `node(id, ...)`

id
a
b
c
d

Can include node properties, e.g., name, annotation, etc., as additional columns in "..."

Matrix representation



$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array}\begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} = E$$

- $e_{i,j} = 1$ if $i \rightarrow j$, or 0 otherwise
  - Or use the value to code edge weight
- Remember the mapping between node ids and row/column indexes

2-hop neighbors

Relational:

```
SELECT e1.src, e2.tgt
FROM edge e1, edge e2
WHERE e1.tgt = e2.src;
```

- What's the count of (a, b) in the result?

Matrix:

$E \times E$, or simply E^2

- What's the value of result entry (i, j)?

How about 3-hop, 4-hop, …, n-hop?

Next time, after MapReduce…

- Measures of “centrality” (how important nodes/edges are)
- Scalable graph data processing