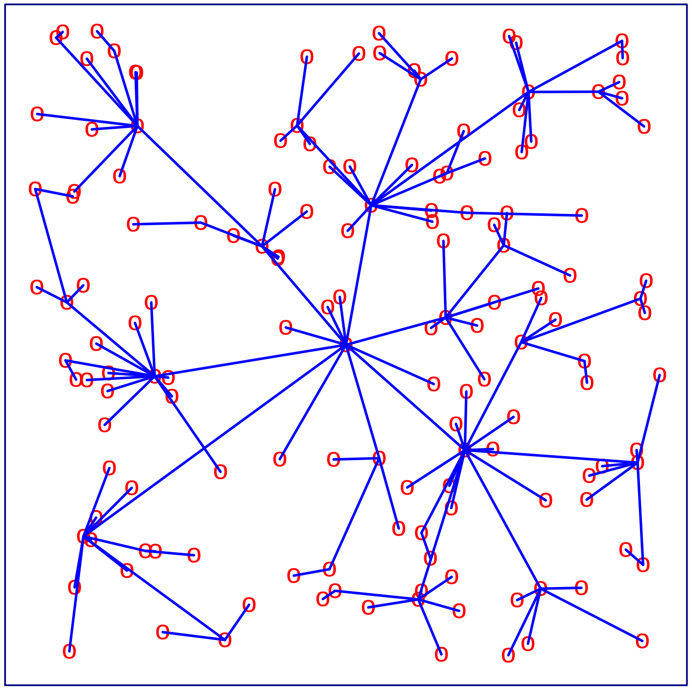
ECS 253 / MAE 253, Lecture 2

Mar 31, 2016

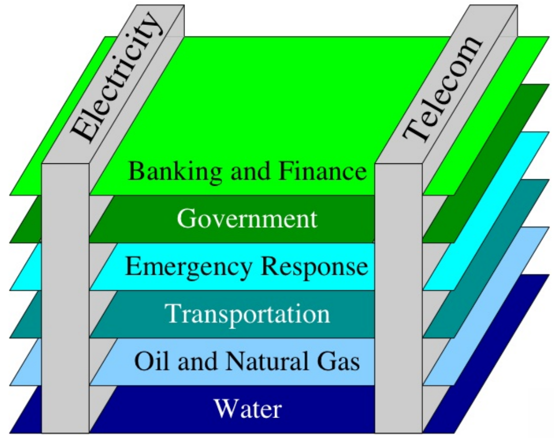


"Power laws, Random graphs, phase transitions"

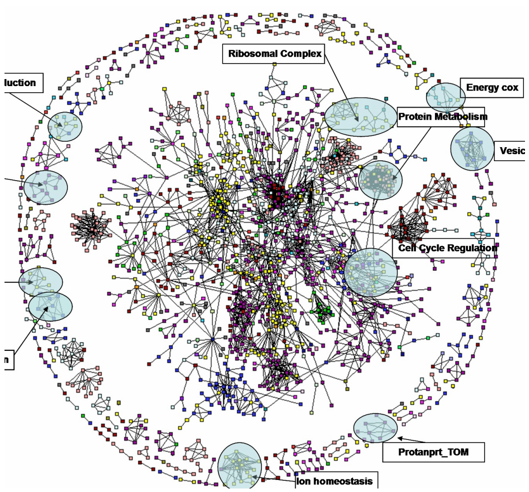
Announcements

- Room change: 176 Chem
- All students on the waitlist WILL be added.

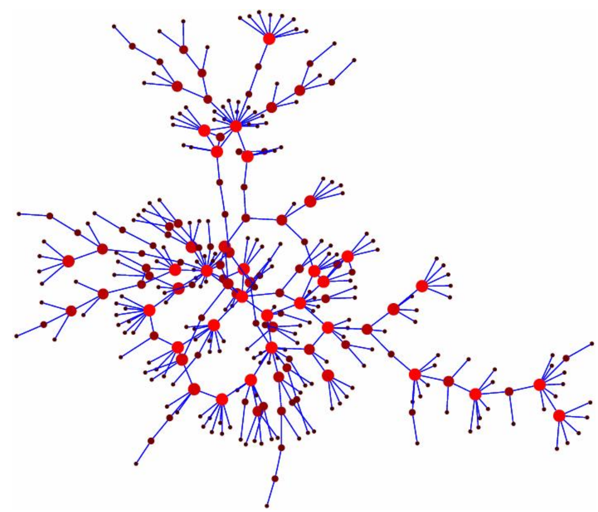
Complex networks are ubiquitous:



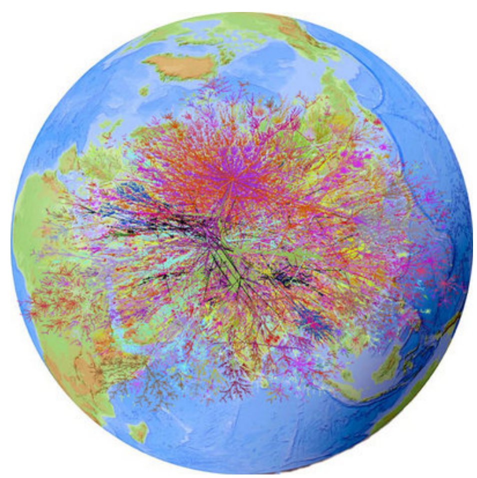
Critical Infrastructure



Biological & Ecological networks



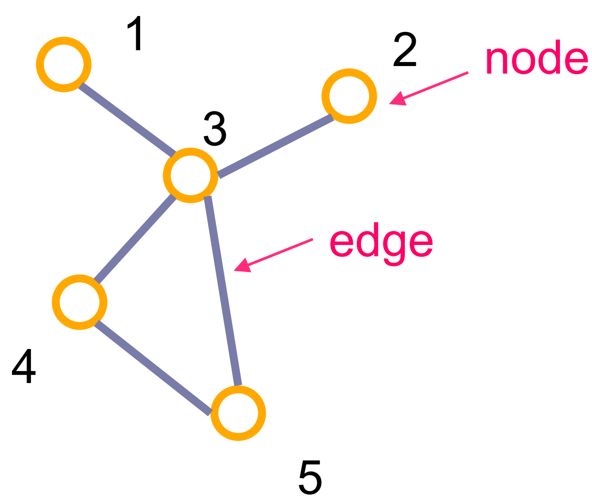
Social networks: Economics & Epidemics



Information and Communication technology

What are networks?

- Networks are collections of points joined by lines.



"Network" ≡ "Graph"

points	lines	
vertices	edges, arcs	math
nodes	links	computer science
sites	bonds	physics
actors	ties, relations	sociology

Slide from Adamic's course

What is a Network?

- Topology (i.e., structure: nodes/vertices and edges/links)
  Measures of topology

- Activity (i.e., function, processes on networks, dynamics of nodes and edges)

Networks: Physical, Biological, Social, Technological

- **Geometric** versus **virtual** (Internet versus WWW).
- **Natural** /spontaneously arising versus **engineered** /built.
- **Directed** versus **undirected** edges.
- Each network may **optimize** something unique.
- Identifying **similarities** and fundamental **differences** can guide future design/understanding.
- Interplay of **topology** and **function** ?
- Unifying features: – **Broad heterogeneity in node degree.** – **Small Worlds** (Diameter $\sim \log(N)$).

Subtle details of edges

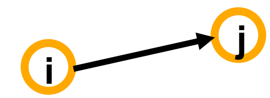
Network elements: edges

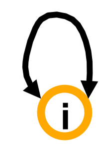
- Directed (also called arcs)
  - A -> B (E_{BA})
    - A likes B, A gave a gift to B, A is B's child
- Undirected
  - A <-> B or A – B
    - A and B like each other
    - A and B are siblings
    - A and B are co-authors
- Edge attributes
  - weight (e.g. frequency of communication)
  - ranking (best friend, second best friend…)
  - type (friend, relative, co-worker)
  - properties depending on the structure of the rest of the graph: e.g. betweenness
- Multiedge: multiple edges between two pair of nodes
- Self-edge: from a node to itself

Slide from Adamic's course

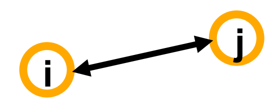
Adjacency matrices

- Representing edges (who is adjacent to whom) as a matrix
  - A_{ij} = 1 if node i has an edge to node j
    = 0 if node i does not have an edge to j



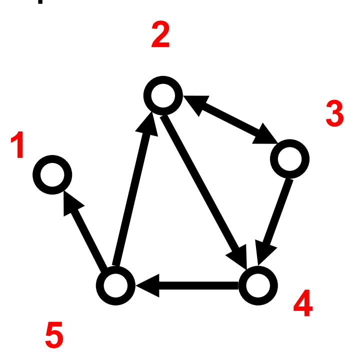


  - A_{ii} = 0 unless the network has self-loops
    - If self-loop, A_{ii}=1



  - $A_{ij} = A_{ji}$ if the network is undirected,
    or if i and j share a reciprocated edge

Example:

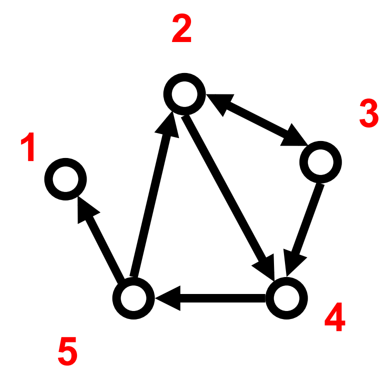


$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Slide from Gunes course, UNR

Adjacency lists

- Edge list
  - 2 3
  - 2 4
  - 3 2
  - 3 4
  - 4 5
  - 5 2
  - 5 1



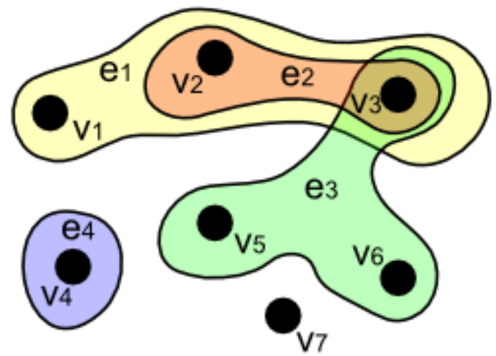
- Adjacency list
  - is easier to work with if network is
    - large
    - sparse
  - quickly retrieve all neighbors for a node
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2

Slide from Gunes course, UNR

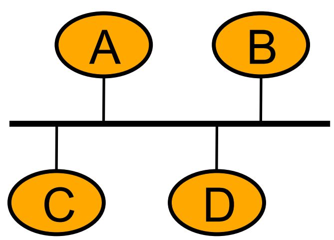
HyperGraphs

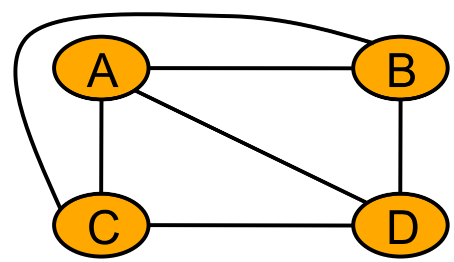
- Edges join more than two nodes at a time (*hyperEdge*)

- Affliation networks



- Examples
  - Families
  - Subnetworks



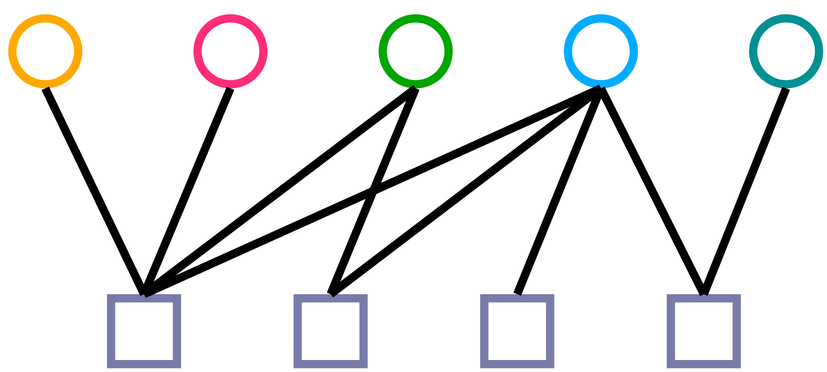


Can be transformed to a *bipartite network*

Slide from Gunes course, UNR

Bipartite (two-mode) networks

- edges occur only between two groups of nodes, not within those groups
- for example, we may have individuals and *events*
  - directors and boards of directors
  - customers and the items they purchase
  - metabolites and the reactions they participate in

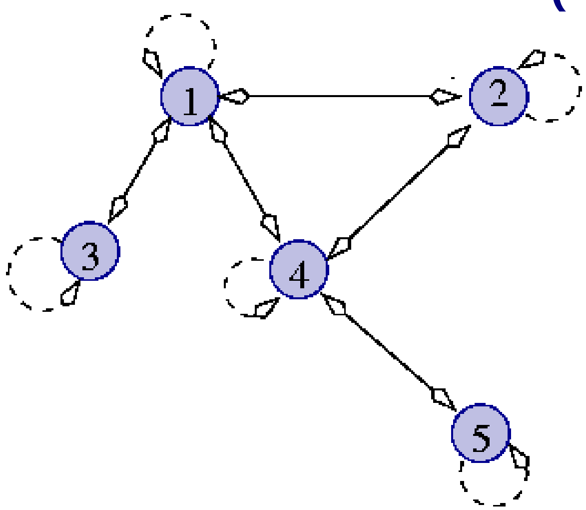


Slide from Gunes course, UNR

NETWORK TOPOLOGY; simple edges

Binary connectivity matrix, M:

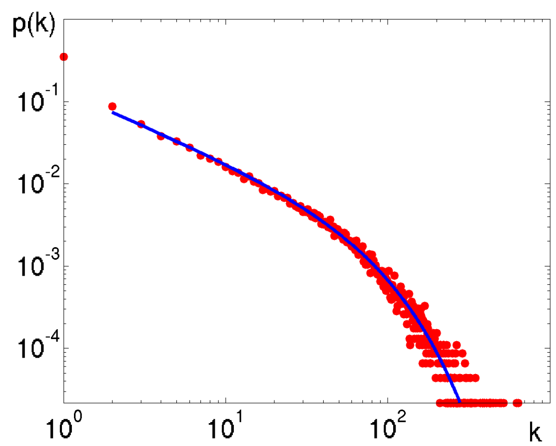
$$M_{ij} = \begin{cases} 1 \text{ if edge exists between } \; i \text{ and } j \\ 0 \text{ otherwise.} \end{cases}$$



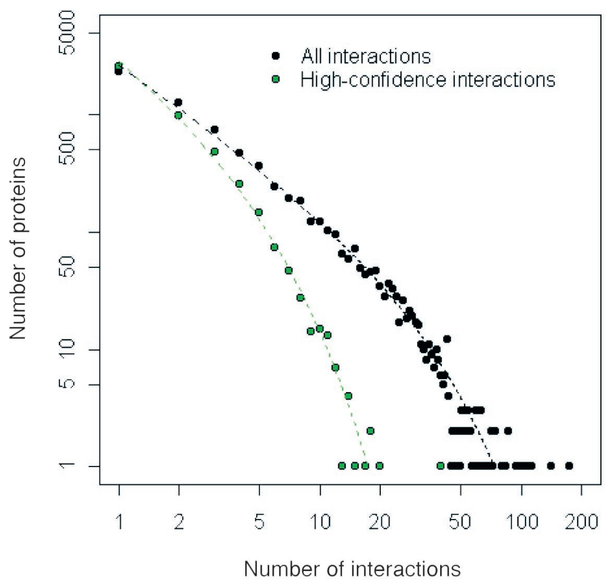
$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} = M$$

Node **degree** is number of links.

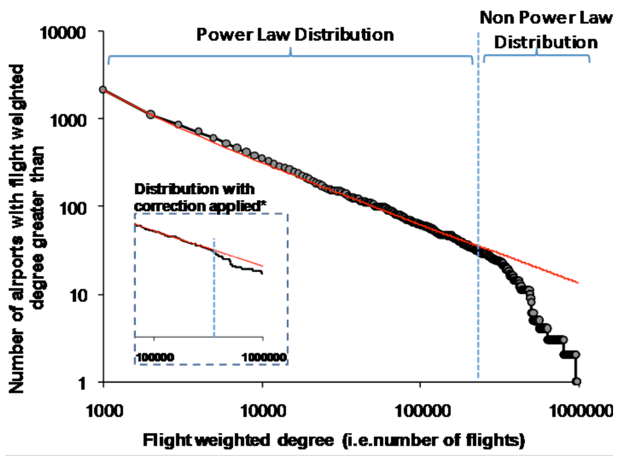
Commonly observed: broad scale degree distribution



Social contacts
Szendröi and Csányi



Protein interactions
Giot et al Science 2003



Airport traffic
Bounova 2009

- A few hubs, dominated by leaves
- Small data sets, power laws vs log normal, stretched-exponential, etc...
- Exceptions: Power grids? Router-level Internet?

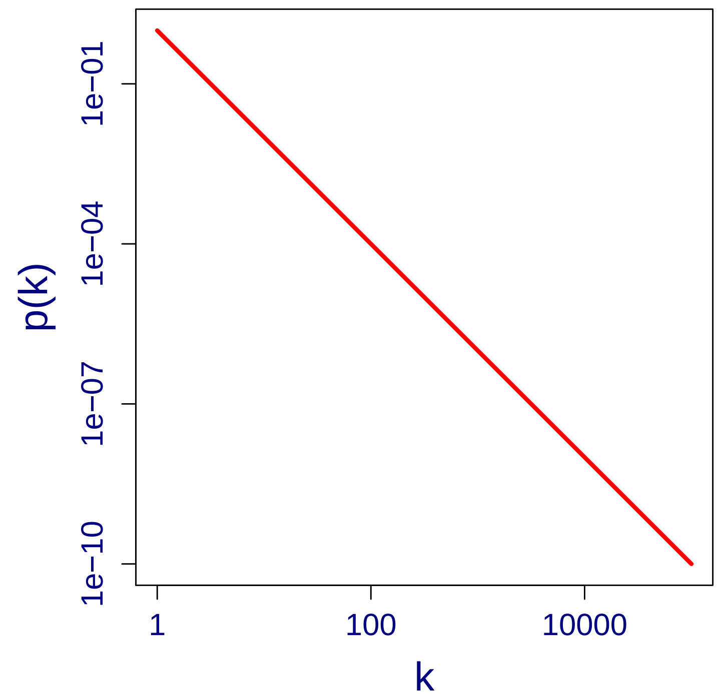
Degree distribution

- Often observe **"heavy-tailed"** / **"broad-scale"** degree distributions.

- The simplest example of such a distribution is a power law (Pareto distribution).

$$p_k \sim k^{-\gamma}$$

$$\ln p_k \sim -\gamma \ln k$$

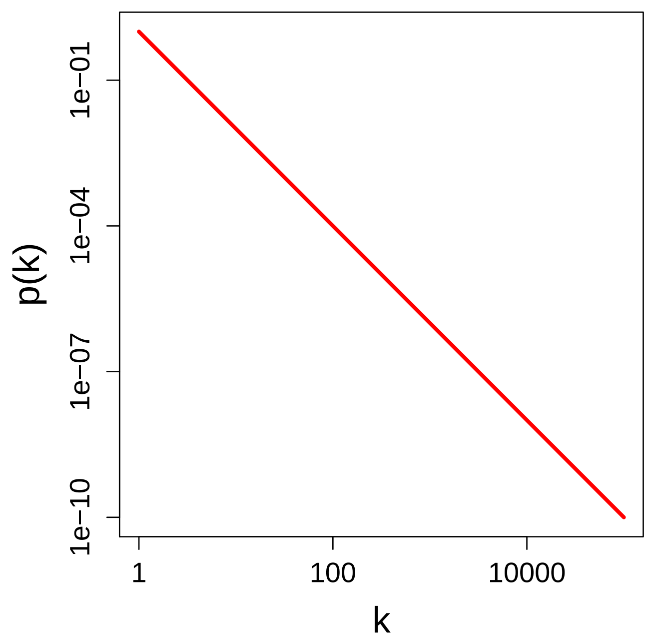


What is a power law?

(Also called a "Pareto Distribution" in statistics).

$$p_k \sim k^{-\gamma}$$

$$\ln p_k \sim -\gamma \ln k$$

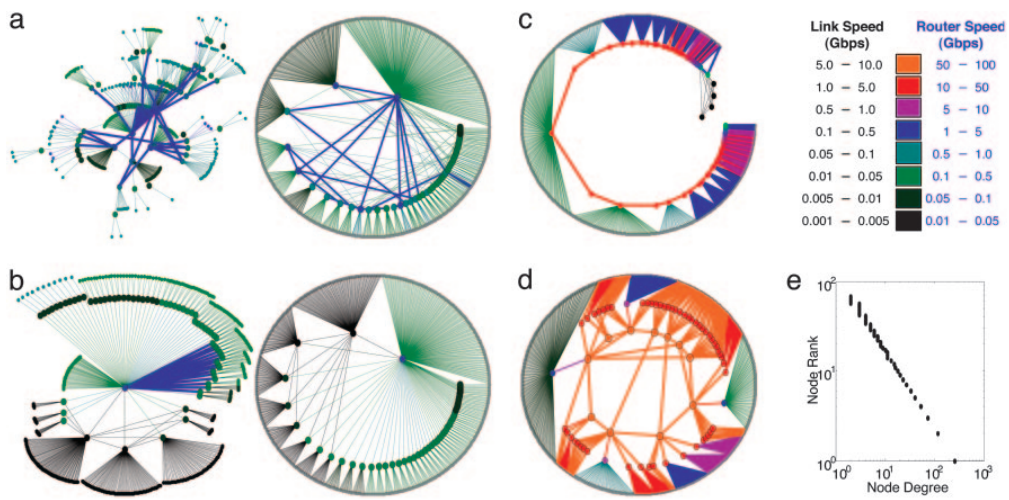


Many network growth models produce power law degree distribution (we will study some of these)

- Preferential attachment
- Copying models (WWW, biological networks, ...)
- Optimization models

Degree distribution misses other structure.

- Doyle, et. al.,
  PNAS **102** (4)2005.



Power law probability distributions: $p_k = Ak^{-\gamma}$ **with** $\gamma > 0$

- $0 \leq p_k \leq 1 \quad \forall k$ which are valid degrees (typically $k \in \mathbb{Z}^+$).
- Must be properly normalized:

$$\sum_{k=1}^{\infty} p_k = \sum_{k=1}^{\infty} \frac{A}{k^\gamma} = 1$$

- Approximating discrete sum by integral:

$$1 = \int_{k=1}^{\infty} \frac{A}{k^\gamma} = -\left(\frac{A}{\gamma - 1}\right) \frac{1}{k^{(\gamma-1)}}\bigg|_{k=1}^{\infty}$$

$$= \left(\frac{A}{1-\gamma}\right)\left(\frac{1}{\infty^{(\gamma-1)}} - 1\right)$$

- Finite requirement means $\gamma > 1$, in which case $A = (\gamma - 1)$.

The first two moments (mean and variance)

- Mean degree:

$$\langle k \rangle = \sum_{k=1}^{\infty} k p_k \approx \int_{k=1}^{\infty} k p_k dk$$

Diverges (i.e., $\langle k \rangle \to \infty$) if $\gamma \leq 2$.

- Second moment:

$$\left\langle k^2 \right\rangle = \sum_{k=1}^{\infty} k^2 p_k \approx \int_{k=1}^{\infty} k^2 p_k dk$$

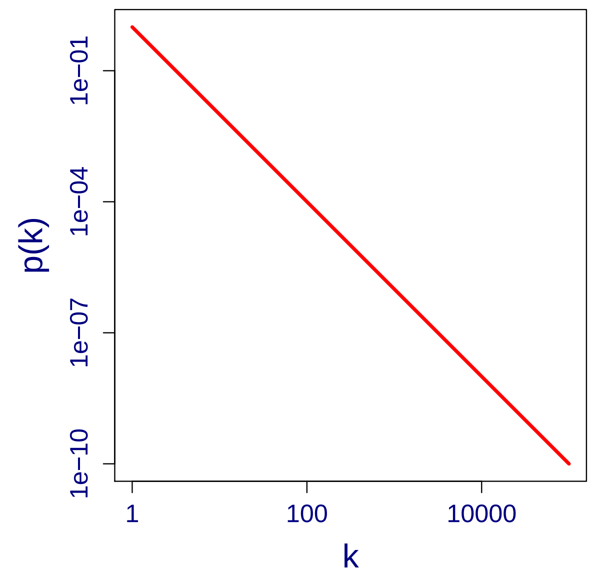
Diverges (i.e., $\left\langle k^2 \right\rangle \to \infty$) if $\gamma \leq 3$.

- Variance $= \left\langle k^2 \right\rangle - \langle k \rangle^2$, likewise diverges if $\gamma \leq 3$.

Properties of a power law PDF (Summary)

(PDF = probability density function)

- To be a properly defined probability distribution need $\gamma > 1$.
- For $1 < \gamma \leq 2$, both the average $\langle k \rangle$ and variance σ^2 are infinite!
- For $2 < \gamma \leq 3$, average $\langle k \rangle$ is finite, but variance σ^2 is infinite!
- For $\gamma > 3$, both average and variance finite.



Power laws in the real world

Confusion

- Power law
- Log normal
- Weibull
- Stretched exponential

All of these distributions can look the same! (Especially when we are dealing with finite data sets — not enough data to get good statistics).

How to deal with real data

- Can adjust bin size: increase exponentially with degree.
- Consider the Cumulative PDF (the CDF): $P_k = \sum_{l=k}^{\infty} p_l$.

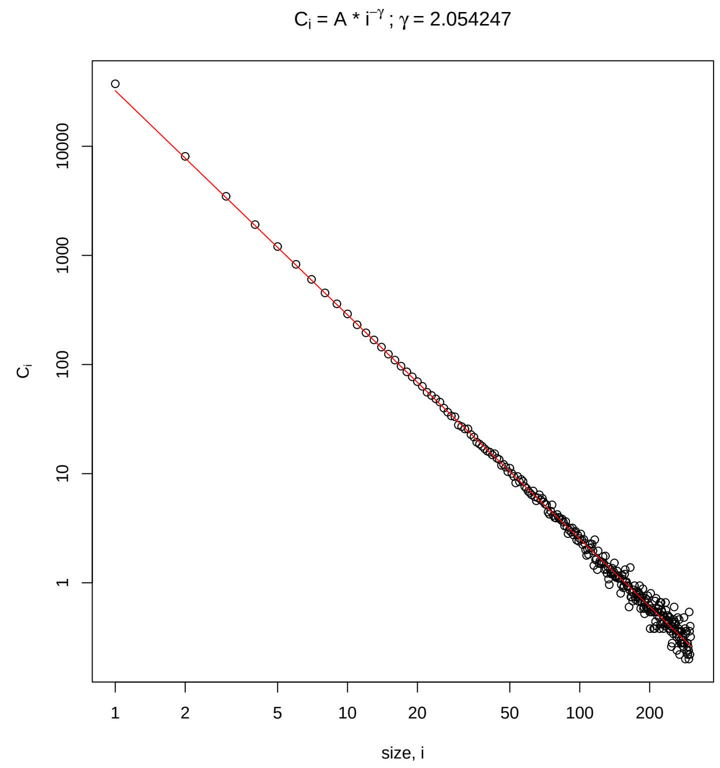
Good reviews:

- Aaron Clauset, Cosma R. Shalizi, M. E. J. Newman. “Power-Law Distributions in Empirical Data”, *SIAM Review*, Vol. 51, No. 4. (2009), pp. 661-703.
- A Brief History of Generative Models for Power Law and Lognormal Distributions Michael Mitzenmacher, *Internet Math*. Vol 1 (2003), 226-251.

But definitively observed in many systems

- Signature of a system at the "critical point" of a phase transition.

- Random graphs at critical point; component sizes: $N_k \sim k^{-5/2}$ (Note, $\gamma = 2.5$)



- Social systems (more next time)

The origins of network theory: Random graphs

What does a “typical” graph with n vertices and m edges look like?

- P. Erdös and A. Rényi, “On random graphs”, *Publ. Math. Debrecen.* **6**, 1959.

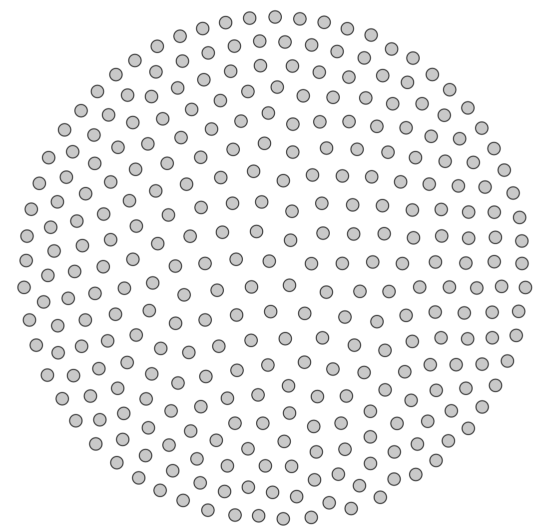
- P. Erdös and A. Rényi, “On the evolution of random graphs”, *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 1960.

- E. N. Gilbert, “Random graphs”, *Annals of Mathematical Statistics* **30**, 1959.

Papers which started the field of graph theory.

Erdös-Rényi random graphs

- Consider a *labelled* graph. Each vertex has a label ranging from $[1, 2, 3, \cdots n]$, for a set of n vertices. (This will make counting and analysis easier.)



- Let E denote the total number of edges possible:

$$E = \binom{N}{2} = \frac{N!}{2!(N-2)!} = \frac{N(N-1)}{2}$$

(If directed edges, we would not divide by 2).

Two formulations

- 1) $\mathcal{G}(n,p)$: The *ensemble* of graphs constructed by putting in edges with probability p, independent of one another. (An edge is present with probability p and absent with probability $[1-p]$.)

  Let $G(n,p)$ denote a random realization of $\mathcal{G}(n,p)$.

- 2) $\mathcal{G}(n,m)$: The ensemble of all graphs with n nodes and exactly m edges.

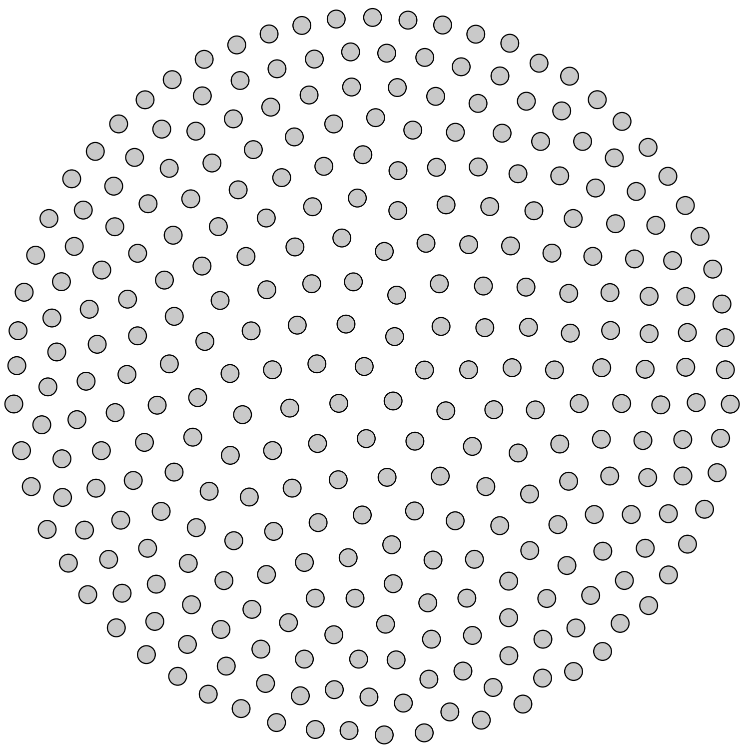
  Let $G(n,m)$ denote a random realization of $\mathcal{G}(n,m)$.

- The two are almost interchangeable with $m = pE$.
  (Recall, E is total number of edges possible).

- We will focus on $G(n,p)$.

The "classic" random graph, $G(N, p)$
(The Null Model)

- P. Erdös and A. Rényi, "On random graphs", *Publ. Math. Debrecen.* 1959.
- P. Erdös and A. Rényi, "On the evolution of random graphs", *Publ. Math. Inst. Hungar. Acad. Sci.* 1960.
- E. N. Gilbert, "Random graphs", *Annals of Mathematical Statistics*, 1959.



- Start with N isolated vertices.
- Add random edges one-at-a-time.
  $N(N-1)/2$ total edges possible.
- After E edges, probability p of any edge is $p = 2E/N(N-1)$

What does the resulting graph look like?

(Typical member of the ensemble)

Explicitly building $G(n,p)$

- Build a realization of $G(n,p)$ by the following graph process:

- Start with n isolated vertices.

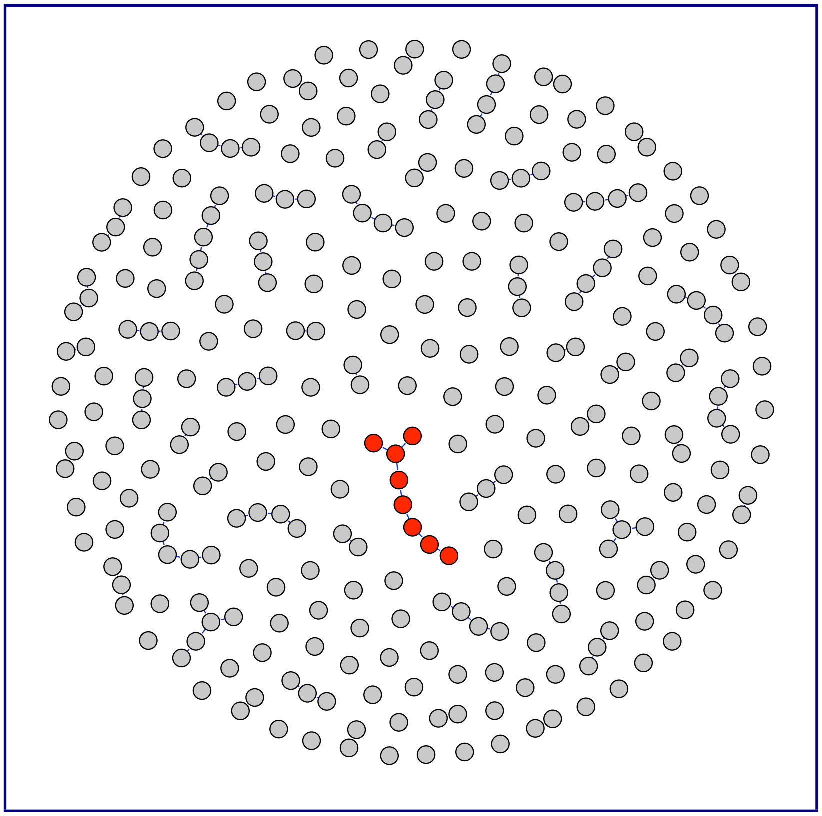
- At each discrete time step, add one edge chosen at random from edges not yet present on the graph.

- At "time" t (i.e., at the addition of t edges), we have built a realization of $G(n,p)$ where $p = t/E$.

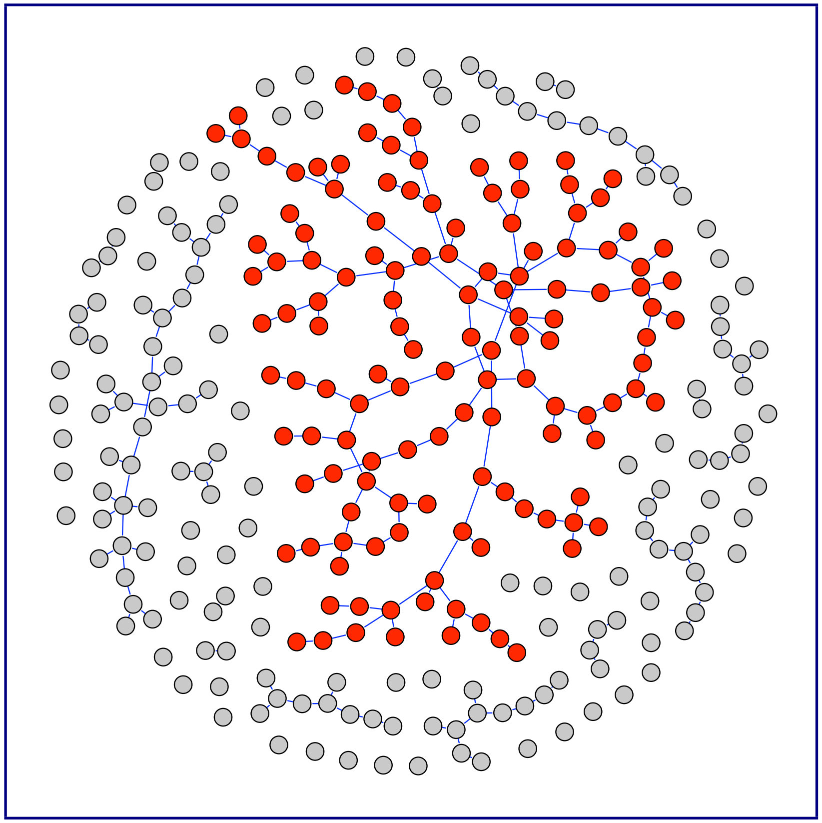
- This is a Markov process (build graph at time $t+1$ from graph at time t).

Ben-Naim, **Krapivsky**, "Kinetic theory of random graphs", *PRE*, 2005.

N=300



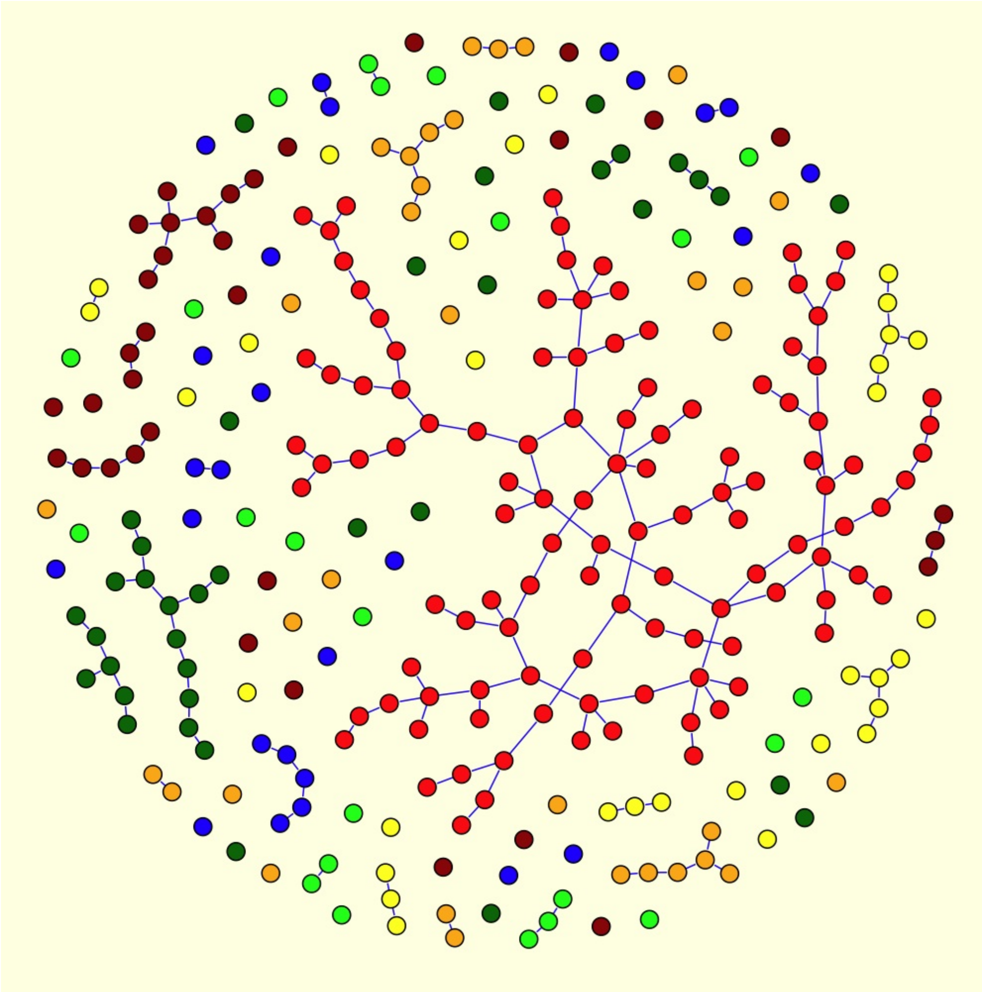
$$p = 1/400 = 0.0025$$



$$p = 1/200 = 0.005$$

Component

A component is a subset of vertices in the graph each of which is reachable from the other by some path through the network.



Behavior for small p

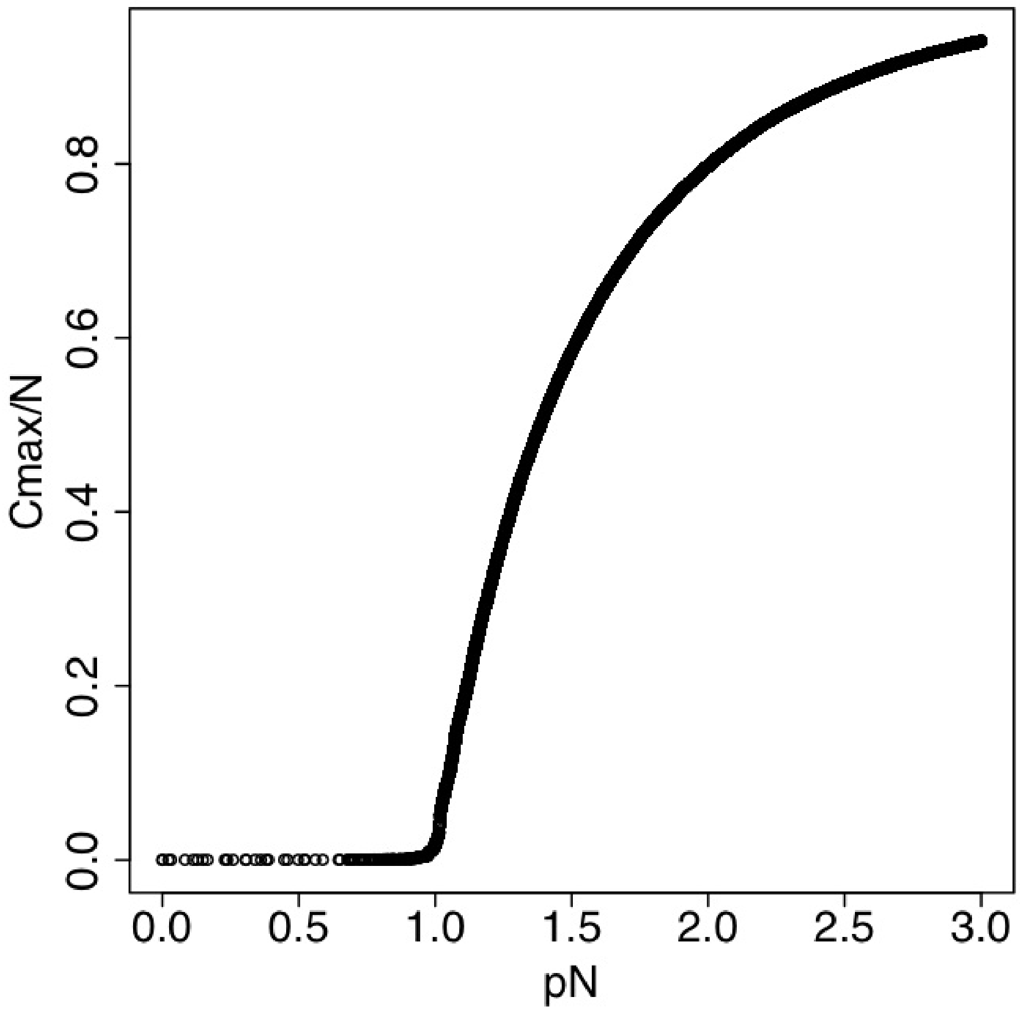
- Consider a realization $G(n, p)$ for $0 < p < 1$ and $n \to \infty$. (A number of interesting properties of random graphs can be proven in this limit).

- Consider the size of the largest component of $G(n, p)$ as a function of p, $C_{max}(p)$.

- For small p, few edges on the graph. Almost all vertices disconnected. The components are small, with size $O(\log n)$, independent of p.

- Keep increasing p (or equivalently t in our model). At $p = 1/n$ (i.e. $t = E/n$), something surprising happens:

Emergence of a "giant component"



- $p_c = 1/N$.
- $p < p_c$, $C_{\max} \sim \log(N)$
- $p > p_c$, $C_{\max} \sim A \cdot N$

(Ave node degree $t = pN$ so $t_c = 1$.)

Branching process (Galton-Watson); "tree"-like at $t_c = 1$.

A Phase Transition!

An abrupt sudden change in one or more physical properties, resulting from a small change in a external control parameter.

Examples from physical systems:

- Magnetization
- Superconductivity
- Liquid/Gas
- Bose-Einstein condensation

Giant component observed in real-world networks

- Formation reminiscent of many real-world networks.
  "Gain critical mass".

- Lower bound on emergence of epidemic outbreak.

- The giant component/Strongly Connected Component used extensively to categorize networks.

Phase transition in connectivity

- Below $p = 1/n$, only small disconnected components.
- Above $p = 1/n$, one large component, which quickly gains more mass. All other components remain sub-linear.
- Note the average node degree, z:

$$\begin{aligned} z &= (2 \times \#edges)/\#vertices \\ &= (2pn(n-1)/2)/n = pn(n-1)/n = (n-1)p \approx np. \end{aligned}$$

(Factor of 2 since each edge contributes degree to two vertices – each end of the edge contributes.)

Recall, expected number of edges, is pn(n-1)/2 .

- At the phase transition, $z = np = 1$. The phase transition occurs when the average vertex degree is one!

Erdős-Rényi, a continuous, second order transition: Mean-field scaling behaviors

- Divergence of susceptibility: $\chi = \frac{\partial m}{\partial h} \sim |T - T_c|^{-\gamma}$
- Random graph "susceptibility" (second moment of the component sizes): $\chi = \sum_{i=1}^{\infty} i^2 n_i$
- For Erdős-Rényi, $\chi \sim |t_c - t|^{-\gamma}$, with $\gamma = 1$.

Power law correlation lengths and response functions $\rightarrow$

Potential EARLY WARNING SIGNALS

(e.g., Scheffer et al. *Nature* 461, 2009)

Is connectivity a good thing?

- Communication, transportation networks
- Spreading of a virus (human or computer)

Random graphs as real-world networks?

- What about degree distribution, clustering, assortativity....?

  – Shown later, Erdos-Renyi yields a Poisson degree distribution, but “configuration” models work around this.

  – Still need null models to match other properties.

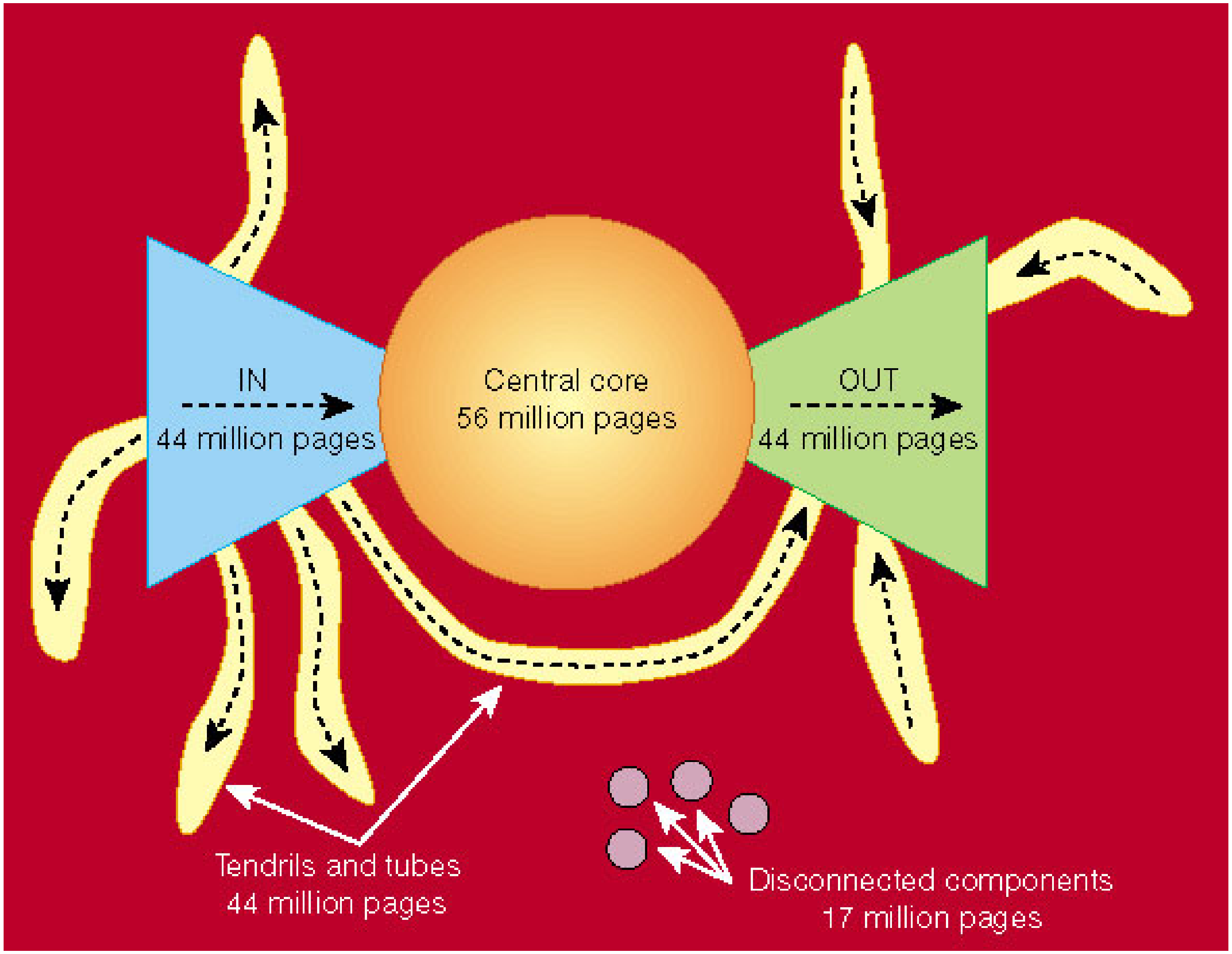
- e.g., “Network Analysis in the Social Sciences”, S. P. Borgatti, A. Mehra, D. J. Brass, G. Labianca, *Science* **323**, 892-895, 2009.

  – Why would a real network look like a random one?

  – Local properties of nodes and edges, not statistics of the network.

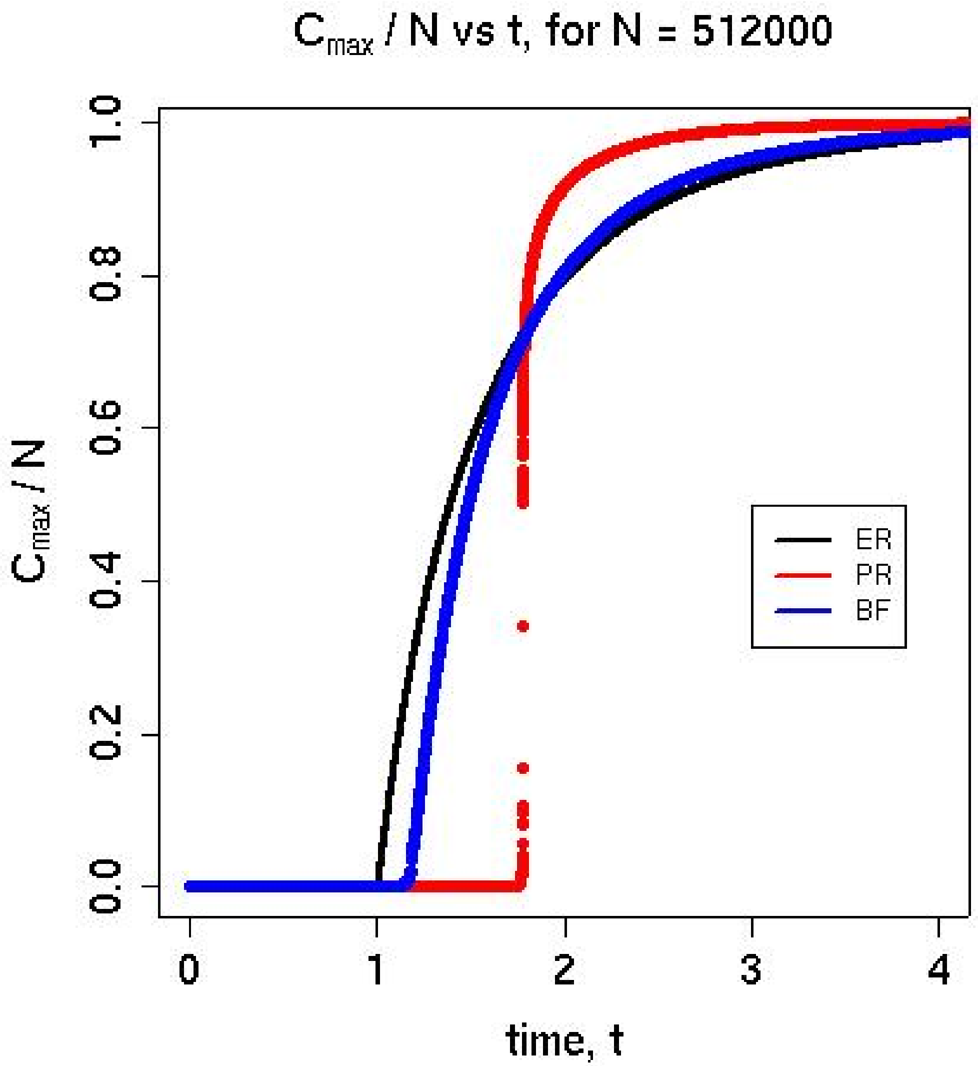
- Developing the correct null models?

The giant component/Strongly Connected Component of the WWW



From “The web is a bow tie” Nature **405**, 113 (11 May 2000)

Algorithms for suppressing the emergence of the Giant Component



e.g. "Explosive percolation", Achlioptas, D'Souza, Spencer, *Science*, 2009.

Back to Erdös-Rényi random graphs

Degree distribution of a graph

- The degree of a node is how many edges connect that node to others.

- If edges are *directed*, a node has a distinct in-degree and out-degree. (Edges in $G(n, p)$ are undirected, so don't have to make that distinction here).

  The degree distribution of the graph is the distribution over all the degrees of all the nodes.

Degree distribution of $G(n, p)$

- Now consider $G(n, p)$ for a fixed value of p and the large n limit.
- The mean degree $z = (n-1)p$ is constant.
- The absence or presence of an edge is independent for all edges.
  - Probability for node i to connect to all other n nodes is p^n.
  - Probability for node i to be isolated is $(1-p)^n$.
  - Probability for a vertex to have degree k follows a binomial distribution:

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}.$$

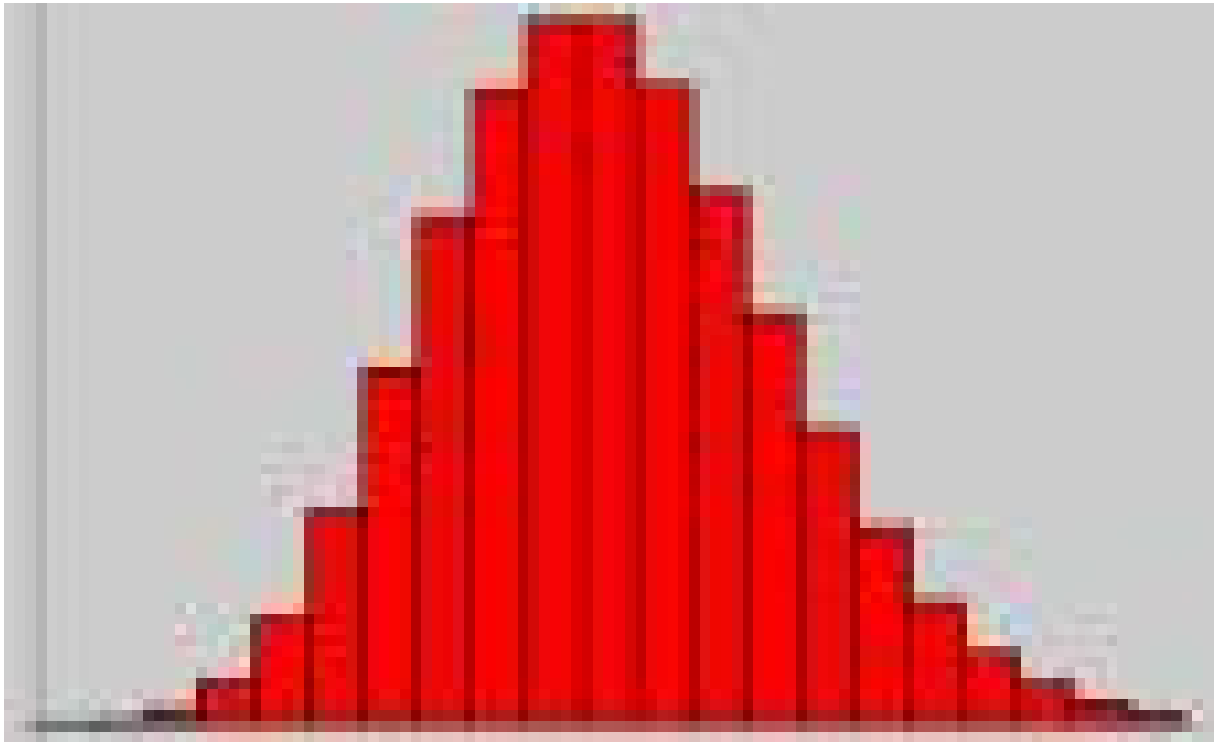
Binomial converges to Poisson as $n \to \infty$

- Recall that $z = (n-1)p = np$ (for large n).

- $$\begin{aligned}\lim_{n\to\infty} p_k &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} (z/n)^k (1-z/n)^{n-k} \\ &= z^k e^{-z}/k!\end{aligned}$$

For more details see for instance: http://en.wikipedia.org/wiki/Poisson_distribution

Poisson Distribution



Diameter

The diameter of a graph is the *maximum* distance between any two connected vertices in the graph.

- Below the phase transition, only tiny components exist. In some sense, the diameter is infinite.
- Above the phase transition, all vertices in the giant component connected to one another by some path.
- The mean number of neighbors a distance l away is z^l. To determine the diameter we want $z^l \approx n$. Thus the typical distance through the network, $l \approx \log n / \log z$.
- This is a small-world network: diameter $d \sim O(\log N)$.

Clustering coefficient

A measure of transitivity: If node A is known to be connected to B and to C, does this make it more likely that B and C are connected?

(i.e., The friends of my friends are my friends)

- In E-R random graphs, all edges created independently, so no clustering coefficient!

Properties of Erdös-Rényi random graphs:

1. Phase transition in connectivity at average node degree, $z = 1$ (i.e., $p = 1/n$).

2. Poisson degree distribution, $p_k = z^k e^{-z}/k!$.

3. Diameter, $d \sim \log N$, a small-world network.

4. Clustering coefficient; none.

How well does $G(n, p)$ model common real-world networks?

1. Phase transtion: Yes! We see the emergence of a giant component in social and in technological systems.

2. Poisson degree distribution: NO! Most real networks have much broader distributions.

3. Small-world diameter: YES! Social systems, subway systems, the Internet, the WWW, biological networks, etc.

4. Clustering coefficient: NO!

Well then, why are random graphs important?

- Much of our basic intuition comes from the study of random graphs.
- Phase transition and the existence of the giant component. Even if not a giant component, many systems have a dominate component much larger than all others.

Generalized random graph

The configuration model (1970's)
Molloy and Reed (1995)

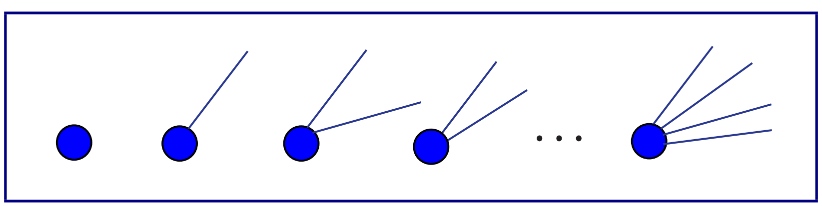
- Specify a degree distribution p_k, such that p_k is the fraction of vertices in the network having degree k.

- We chose an explicit *degree sequence* by sampling in some unbiased way from p_k. And generate the set of n values for k_i, the degree of vertex i.

- Think of attaching k_i "spokes" or "stubs" to each vertex i.

- Choose pairs of "stubs" (from two distinct vertices) at random, and join them. Iterate until done.

- Technical details: self-loops, parallel edges, ... (neglect in $n \to \infty$ limit).



"Are randomly grown graphs really random?"

Callaway, Hopcroft, Kleinberg, Newman, Strogatz. *Phys Rev E* **64** (2001)

- Rather than Erdos-Renyi, add vertices one-by-one.
- At each discrete step, t:
  – a new vertex arrives, and
  – with probability δ a new randomly selected edge is added.
- In large t limit see emergence of giant component as function of δ (giant exists for $\delta \geq 1/8$).
- But size of "giant" is finite (even as $n \to \infty$).
- Positive degree-degree correlations (higher degree by virtue of age).

Summary: Terms introduced today

- Component
- Phase transition
- Degree distribution
- Graph diameter

Further reading on random graphs

- M. E. J. Newman review, pages 20-25. (Heuristic arguments)
- R. Durrett book, Chaps 1 and 2. (Technical proofs)
- B. Bollobás, *Random Graphs*, 2nd Edition, Cambridge U Press, 2001 (the seminal text on the mathematics of random graphs).