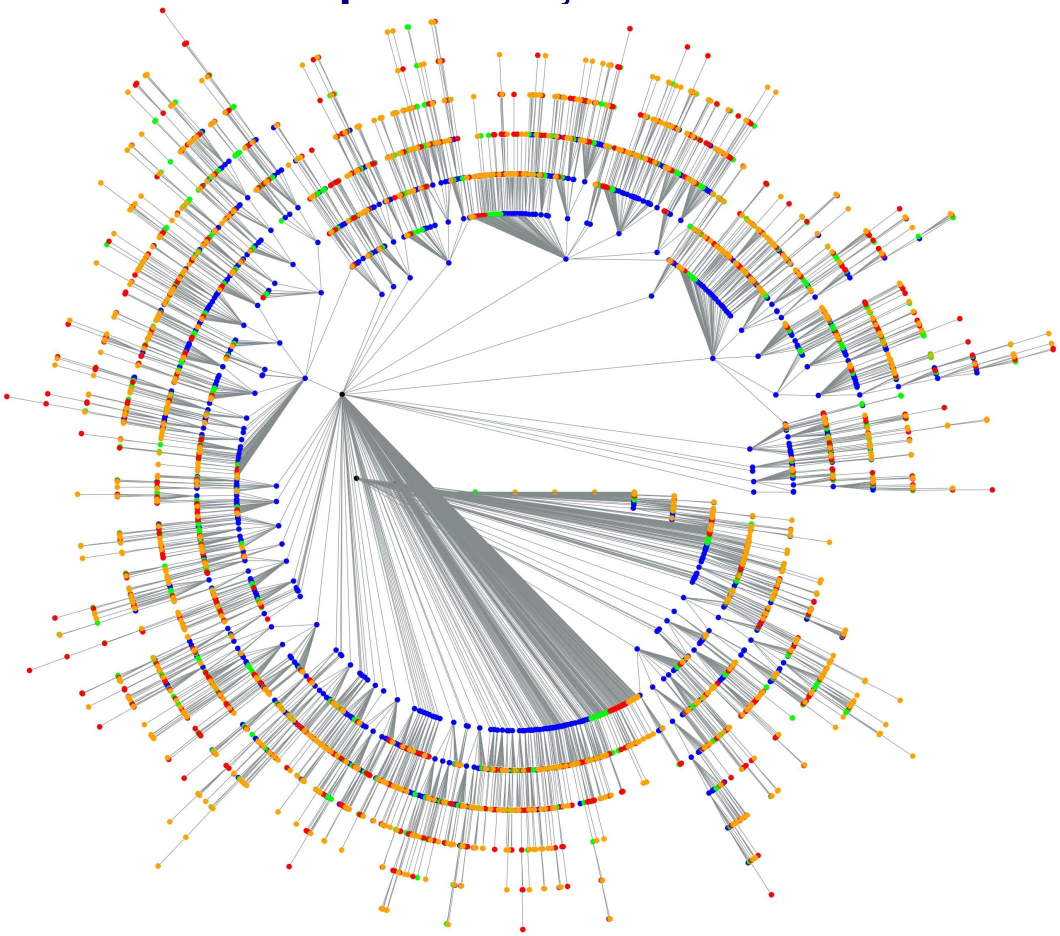
ECS 253 / MAE 253, Lecture 3

April 10, 2023



"Preferential Attachment, Network Growth, Master Equations"

Announcements

- Two tracks to the class:

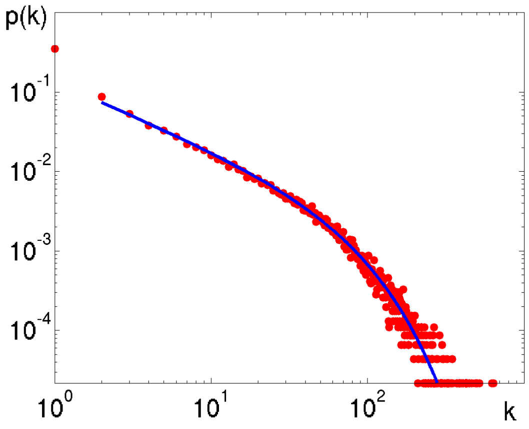
  Track A: (1) Common homeworks (e.g. HW1.pdf, HW2.pdf) and (2) HW1a.pdf, HW2a.pdf etc. (HW1a, posted by Thurs.)

  Track B: (1) Common homeworks (e.g. HW1.pdf, HW2.pdf) and (2) HW1b.pdf, HW2b.pdf etc.

- Project
  - Teams of 5-6 people ideal
  - Negative results are OK
  - Ideally aim to have a result for a journal or conference
  - HW1a is the starting point
  - Today: time to pitch your idea

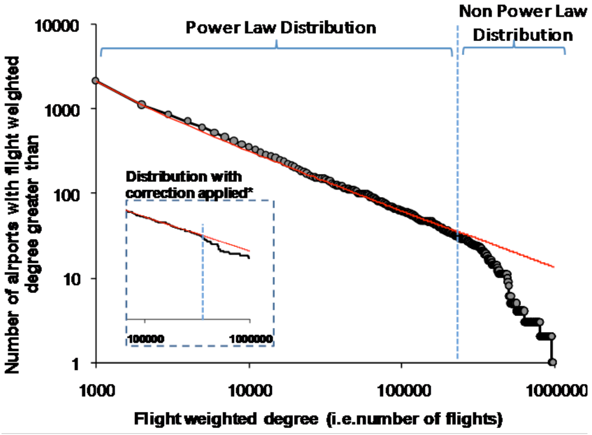
Back to basics of networks

Recall: broad scale degree distribution



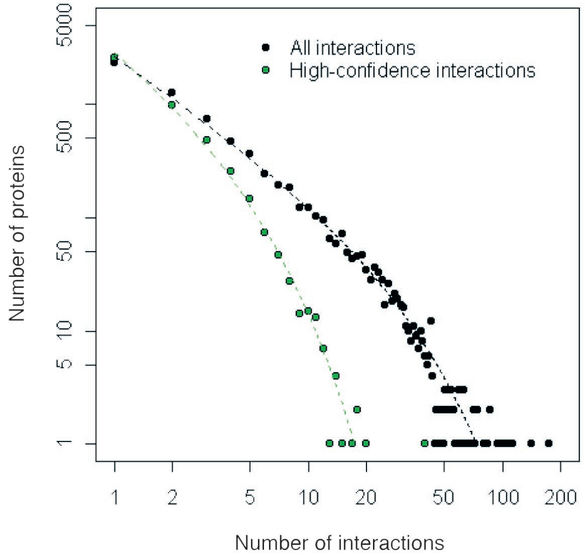
Social contacts

Szendröi and Csányi



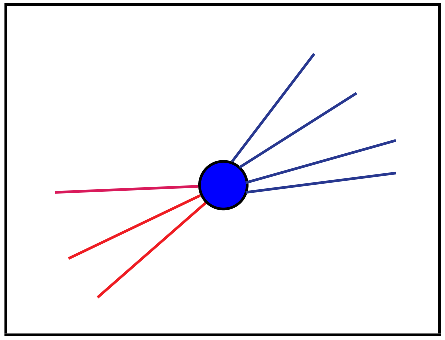
Airport traffic

Bounova 2009



Protein interactions

Giot et al Science 2003



(node degree)

Approximating broad scale by a Power Law
Properties of a power law PDF

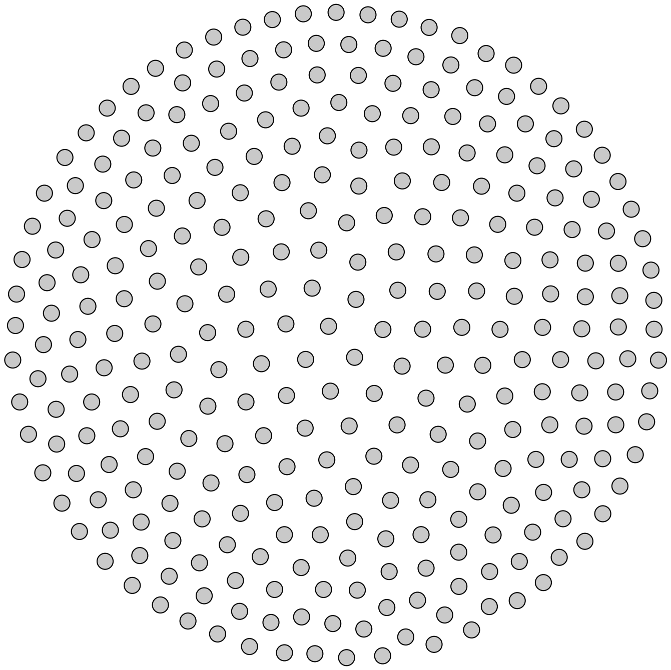
(PDF = probability density function)

$$\boxed{p_k = Ak^{-\gamma}}$$

- 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 are finite.

Recall: The "classic" random graph, $G(N, p)$
(A Classic 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.

– Undirected, so $N(N-1)/2$ total edges possible.

- Each possible edge added with probability p.

- Expected number of edges

$$\langle m \rangle = pN(N-1)/2$$

What does the resulting graph look like?

(Typical member of the ensemble)

Kinetic theory equivalent — e.g., HW1b

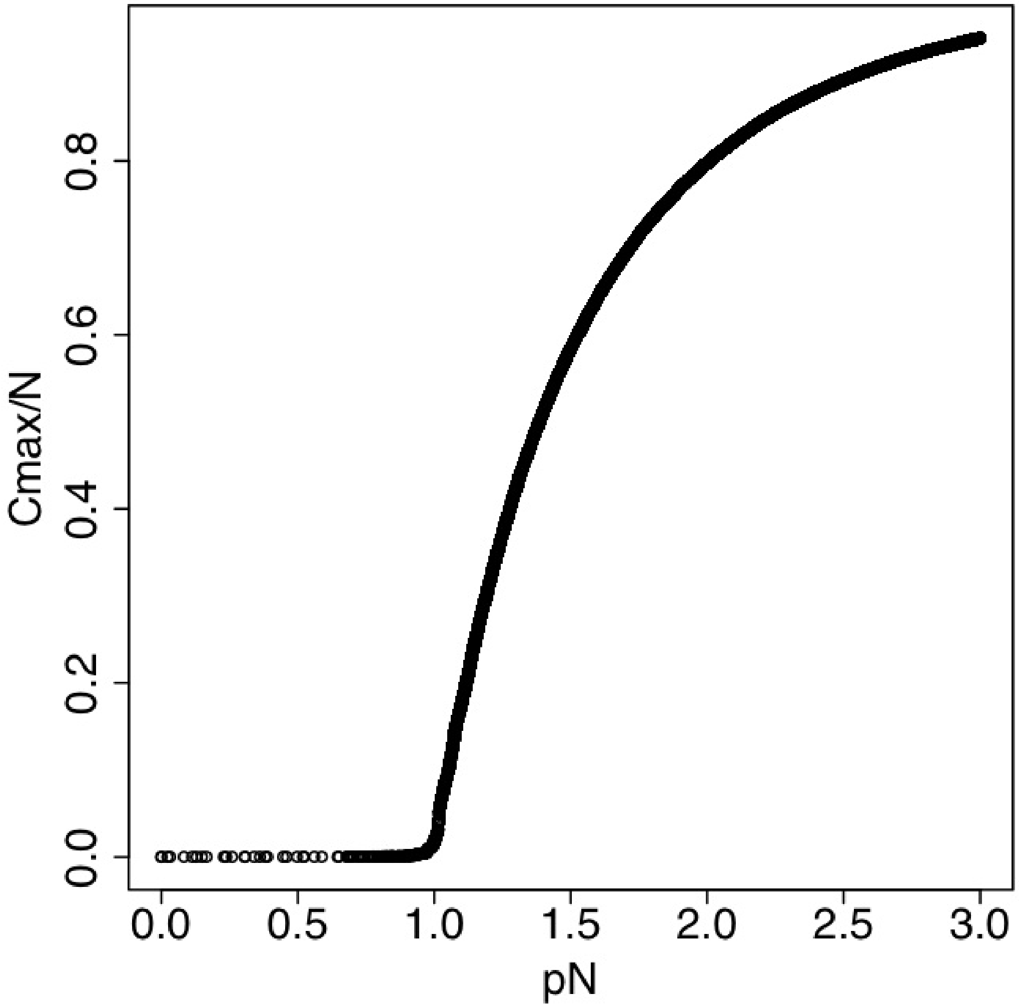
“Kinetic theory of random graphs: From paths to cycles”, E. Ben-Naim and P. L. Krapivsky, Phys. Rev. E 71, 026129 (2005).

- Add random edges one-at-a-time. After t edges, probability p of any edge is $p = 2t/N(N-1)$

- Allows a mapping between p and t.

- Kinetic theory allows us to interpret this as a dynamical process, as seen in remainder of lecture.

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 $t_c = 1$.

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!

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.
- 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:

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

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

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

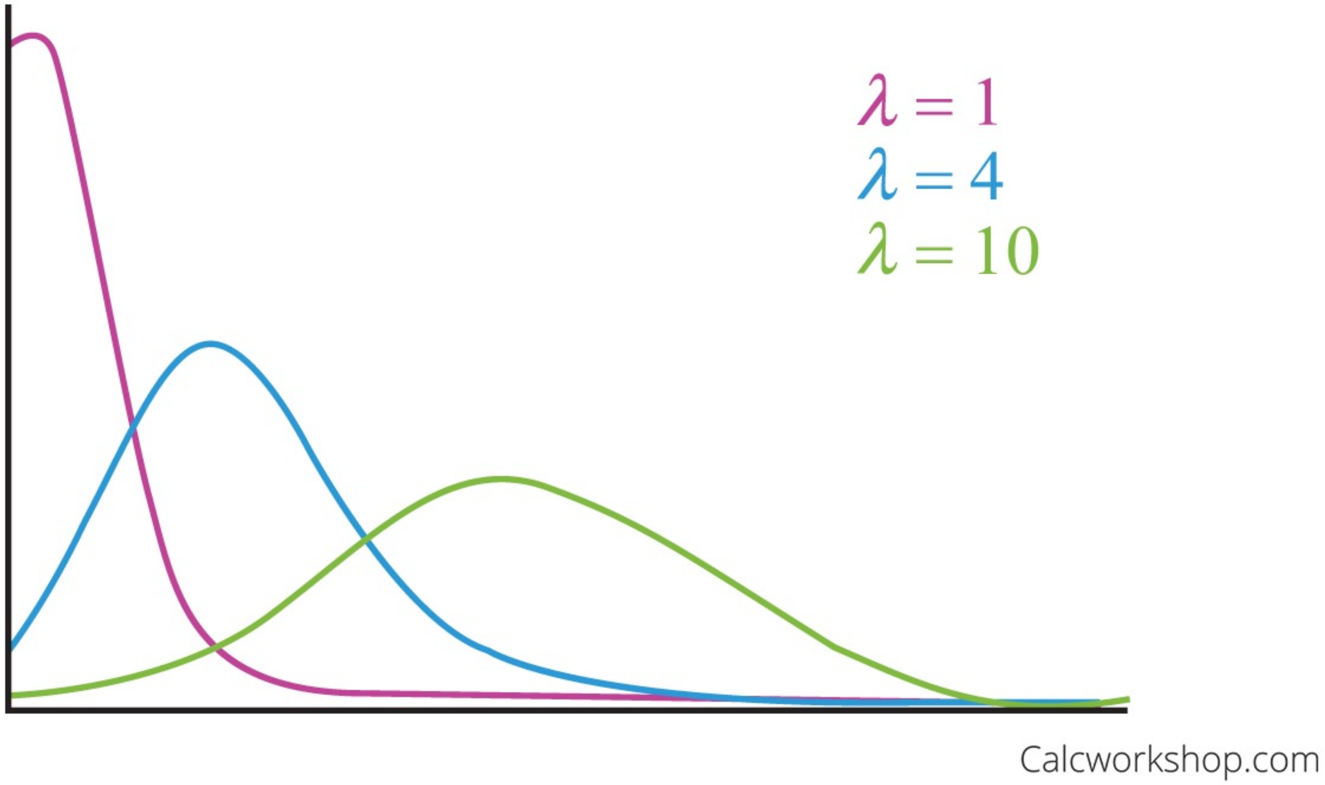
- Substituting $p = z/n$ in the second line:

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

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

Poisson Distribution

Convention is that average is λ.
(We use z to relate to the literature.)



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, $\boxed{l \approx \log n / \log z}$

- 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.

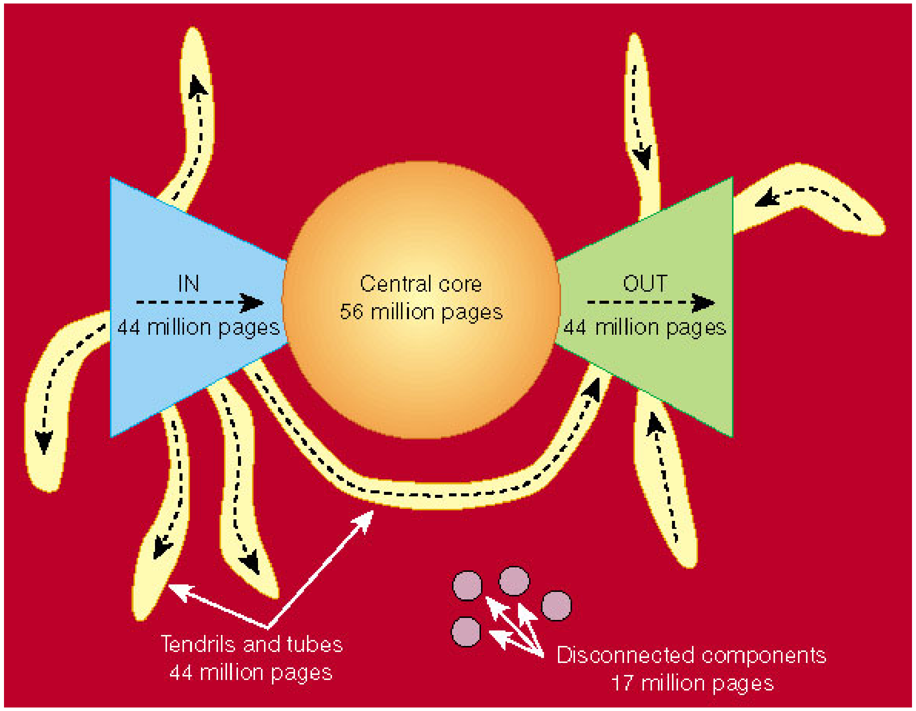
4. Clustering coefficient; none.

So, 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! Many 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?

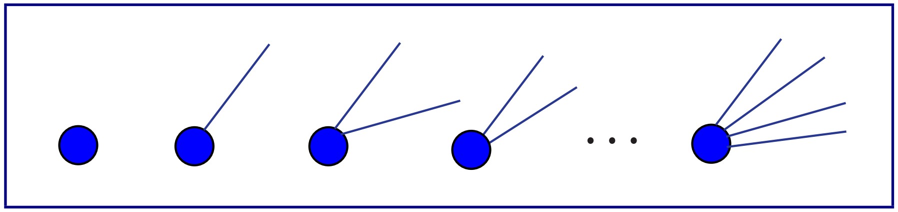
- Much of our basic intuition comes from random graphs.
- Phase transition and the existence of the giant component. Even if not a giant component, many systems have a dominate component much larger than all others.



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

Generalized random graph – later advanced HW accommodate any degree sequence

The configuration model,
Bollobás (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.
- The **degree sequence** is the explicit set of n values for the exact degree, k_i, of vertex i. It is generated by sampling in some unbiased way from p_k.
- 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).
- Emergence of a giant component when expected number of second neighbors greater than expected number of first neighbors.

“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).

Back to power laws
Power laws in social systems

- Popularity of web pages and web search terms: $N_k \sim k^{-1}$

- Rank of city sizes ("Zipf's Law"): $N_k \sim k^{-1}$

- Pareto. In 1906, Pareto made the now famous observation that twenty percent of the population owned eighty percent of the property in Italy, later generalised by Joseph M. Juran and others into the so-called Pareto principle (also termed the 80-20 rule) and generalised further to the concept of a Pareto distribution.

- Usually explained in social systems by "the rich get richer" (preferential attachment).

Known Mechanisms for Power Laws

- Phase transitions (e.g., power law behavior at the critical point, e.g., the distribution of component sizes (see HW1b).)
- Random multiplicative processes (fragmentation)
- Combination of exponentials (e.g. word frequencies)
- Preferential attachment / Proportional attachment (Polya 1923, Yule 1925, Zipf 1949, Simon 1955, Price 1976, Barabási and Albert 1999)

Attractiveness (rate of growth) is proportional to size,

$$\frac{ds}{dt} \propto s$$

Origins of preferential attachment

- 1923 — Polya, urn models.
- 1925 — Yule, explain genetic diversity.
- 1949 — Zipf, distribution of city sizes (*1/f*).
- 1955 — Simon, distribution of wealth in economies. (“The rich get richer”).
- [Interesting note, in sociology this is referred to as the *Matthew effect* after the biblical edict, “For to every one that hath shall be given ... ” (Matthew 25:29)]

Preferential attachment in networks

D. J. de S. Price: “Cumulative advantage”

- D. J. de S. Price, “Networks of scientific papers” *Science*, 1965. First observation of power laws in a network context. Studied paper co-citation network.

- D. J. de S. Price, “A general theory of bibliometric and other cumulative advantage processes” *J. Amer. Soc. Info. Sci.*, 1976.

Cumulative advantage seemed like a natural explanation for paper citations:

The rate at which a paper gains citations is proportional to the number it already has. (Probability to learn of a paper proportional to number of references it currently has).

Preferential attachment in networks, continued

- Cumulative advantage did not gain traction at the time. But was rediscovered some decades later by Barabási and Albert , in the now famous paper (over 30,000 citations c.f. Google Scholar):

- "Emergence of Scaling in Random Networks", *Science* **286**, 1999.

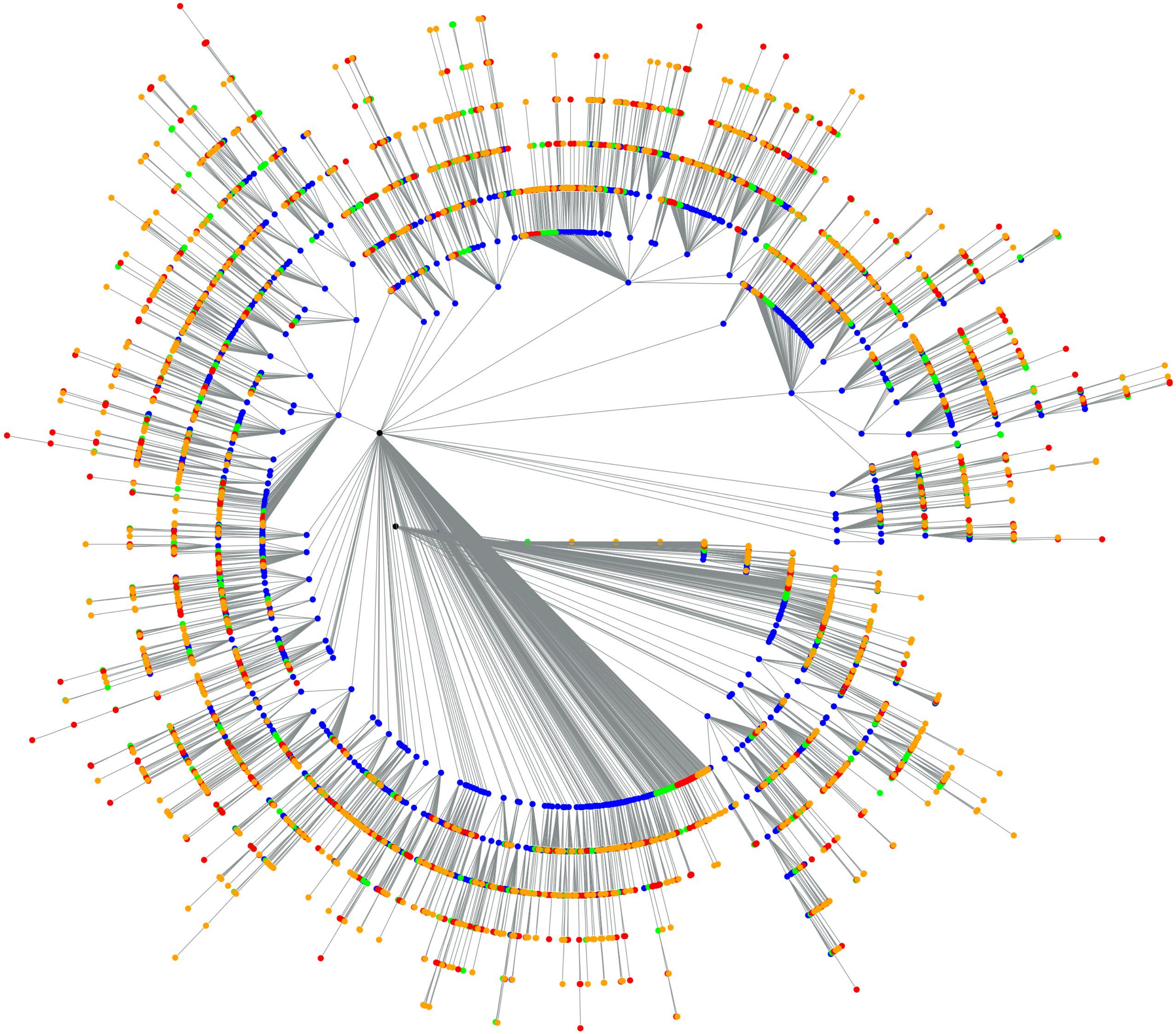
- They coined the term "preferential attachment" to describe the phenomena.

(de S. Price's work resurfaced after BA became widely reknown.)

The Barabási and Albert model

- A discrete time network evolution process, relating the graph $G(t+1)$ to $G(t)$.
- Start at t=0 with a single isolated node.
- At each discrete time step, a new node arrives.
- This new node makes m edges to already existing nodes. (Why m edges? i.e., what happens if $m = 1$?)
- The likelihood of a new edge to connect to an existing node j is proportional to the degree of node j, denoted d_j.
- We are interested in the limit of large graph size, $n \to \infty$.

Visualizing a PA graph ($m = 1$) at $n = 5000$



Probabilistic treatment (kinetic theory)

- Start at $t = 0$ with one isolated node (or a small core set).
  – At time t the total number of nodes added $n = t$.
  – At time t the total number of edges added is mt.

- Let $d_j(t)$ denote the degree of node j at time t.

- Probability an edge added at $t + 1$ connects to node j:

$$Pr(t + 1 \rightarrow j) = d_j(t) / \sum_j d_j(t).$$

- Normalization constant easy (but time dependent):
  $\sum_j d_j(t) = 2mt$

  (Each node 1 through t, contributes m edges.)
  (Each edge augments the degree of two nodes.)

Network evolution
Process on the degree sequence

- The probability a new edge connects to a particular node j of degree k at time t:

  $d_j(t)/\sum_j d_j(t) = k/2mt$

- Also, when a node of degree k gains an attachment, it becomes a node of degree $k+1$.

- When the new node arrives, it increases by one the number of nodes of degree m.

The “Master Equation” / “rate eqns” / “kinetic theory”
Evolution of the typical graph

(Let $n_{k,t} \equiv$ expected number of nodes of degree k at time t,
and $n_t \equiv$ total number of nodes added by time t: Note $n_t = t$)

For each arriving link:

- For $k > m$: $\quad n_{k,t+1} = n_{k,t} + \frac{(k-1)}{2mt}\, n_{k-1,t} - \frac{k}{2mt}\, n_{k,t}$
- For $k = m$: $\quad n_{m,t+1} = n_{m,t} - \frac{m}{2mt}\, n_{m,t}$

The “Master Equation” / “rate eqns” / “kinetic theory”
Evolution of the typical graph

For each arriving link (from last page):

- For $k > m$: $\quad n_{k,t+1} = n_{k,t} + \frac{(k-1)}{2mt}\, n_{k-1,t} - \frac{k}{2mt}\, n_{k,t}$
- For $k = m$: $\quad n_{m,t+1} = n_{m,t} - \frac{m}{2mt}\, n_{m,t}$

Each new node contributes m links (and one new node). Assuming $n \to \infty$ there are no multi-edges:

- For $k > m$: $\quad n_{k,t+1} = n_{k,t} + \frac{m(k-1)}{2mt}\, n_{k-1,t} - \frac{mk}{2mt}\, n_{k,t}$
- For $k = m$: $\quad n_{m,t+1} = n_{m,t} + 1 - \frac{m^2}{2mt}\, n_{m,t}$

Translating from number of nodes $n_{k,t}$ to probabilities $p_{k,t}$

$$p_{k,t} = n_{k,t}/n(t) = n_{k,t}/t$$
$$\to n_{k,t} = t\, p_{k,t}$$

For each arriving node, after m edges added:

- For $k > m$: $(t+1)\, p_{k,t+1} = t\, p_{k,t} + \frac{(k-1)}{2}\, p_{k-1,t} - \frac{k}{2}\, p_{k,t}$

- For $k = m$: $(t+1)\, p_{m,t+1} = t\, p_{m,t} + 1 - \frac{m}{2}\, p_{m,t}$

Steady-state distribution

We want to consider the final, steady-state: $\mathbf{p_{k,t} = p_k}$.

- For $k > m$: $(t+1)\, p_k = t\, p_k + \frac{(k-1)}{2}\, p_{k-1} - \frac{k}{2}\, p_k$
- For $k = m$: $(t+1)\, p_m = t\, p_m + 1 - \frac{m}{2}\, p_m$

Rearranging and solving for p_k:

- For $k > m$: $p_k = \frac{(k-1)}{(k+2)}\, p_{k-1}$
- For $k = m$: $p_m = \frac{2}{(m+2)}$

Recursing p_k to p_m

$$p_k = \frac{(k-1)(k-2)\cdots(m)}{(k+2)(k+1)\cdots(m+3)} \cdot p_m = \frac{m(m+1)(m+2)}{(k+2)(k+1)k} \cdot \frac{2}{(m+2)}$$

$$\boxed{p_k = \frac{2m(m+1)}{(k+2)(k+1)k}}$$

For $k \gg 1$

$$\boxed{p_k \sim k^{-3}}$$

For more on master equations

- “Rate Equations Approach for Growing Networks”, P. L. Krapivsky, and S. Redner, invited contribution to the *Proceedings of the XVIII Sitges Conference on “Statistical Mechanics of Complex Networks”*.
- *Dynamical Processes on Complex Networks*, Barratt, Barthelemy, Vespignani

Applications to cluster aggregation (e.g. Erdos-Renyi)

- “Kinetic theory of random graphs: From paths to cycles”, E. Ben-Naim and P. L. Krapivsky, Phys. Rev. E 71, 026129 (2005).
- “Local cluster aggregation models of explosive percolation”, R. M. D’Souza and M. Mitzenmacher, Physical Review Letters, 104, 195702, 2010.

Possible topic areas, 1

- Transportation networks and flows; multi-modal transportation
- Open source software – e.g., social and technological networks in github
- Machine learning – e.g., bring network connectivity into binary classifiers
- Power grid modeling
- Opinion dynamics / social unrest / multiplex opinion dynamics
- Ranking in networks; especially temporal, multilayered, higher-order
- Multilayered and temporal macaque monkey networks
- Shocks and tipping points
- Metrics for multilayered, temporal, or higher-order networks
- Co-author and citation networks

Possible topic areas, 2

- Math network of theorems and proofs
- Control of complex networks
- Neuroscience
- Food networks
- Recommendation systems
- Biological networks
- Terrorist networks
- See also class homepage “Projects” tab

Your ideas?