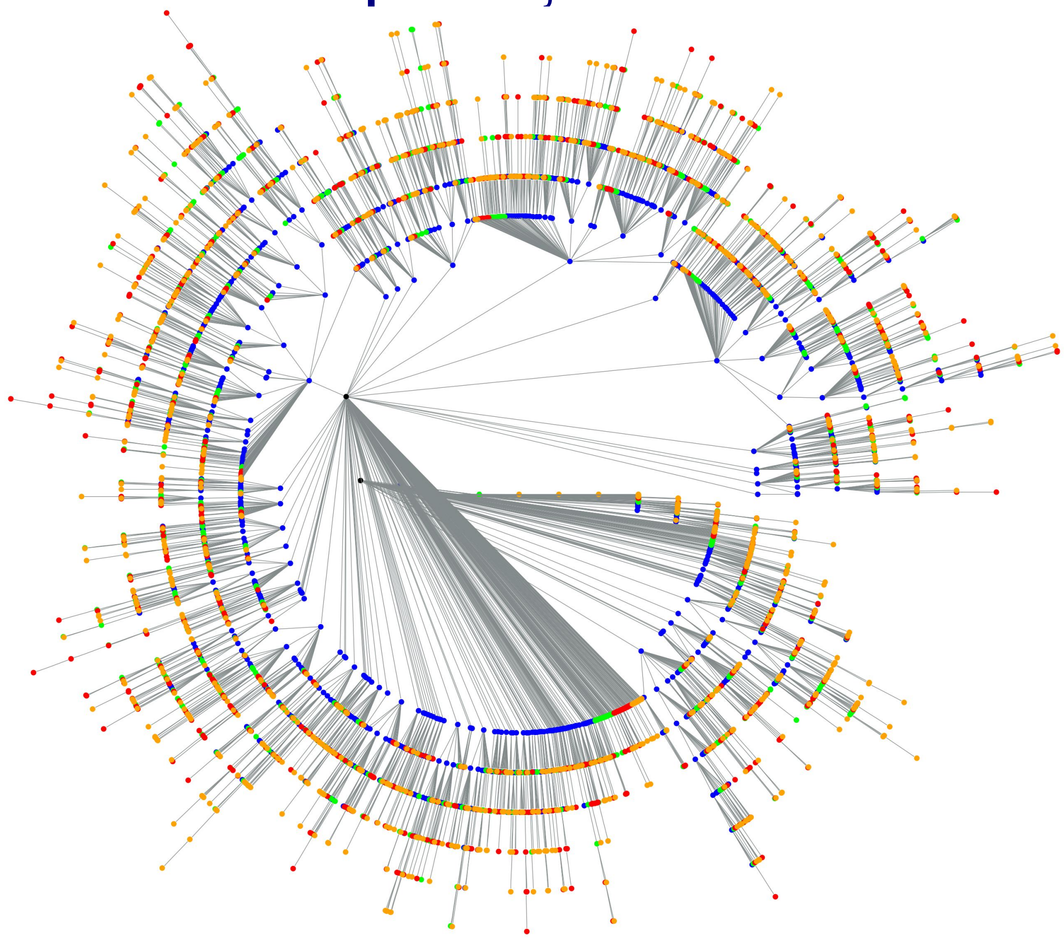
ECS 253 / MAE 253, Lecture 3

April 5, 2016



"Preferential Attachment, Network Growth, Master Equations"

Announcements

- Today, lecture 4:10-5:30pm.
- Then, work in project ideas and team formation with the TAs 5:30-6pm.
- HW1 to be posted (by Thurs) — includes project pitch or mathematical modeling
- Project
  - Teams of 4-5 people ideal
  - Negative results are OK
  - Ideally aim to have a result for a journal of conference

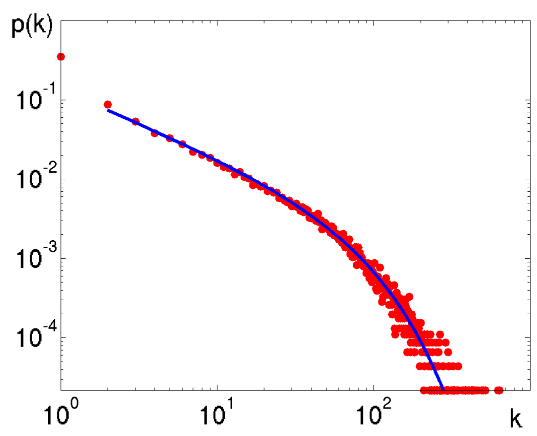
Project pitch

- One-paragraph ($\sim$200 words) describing your idea. Submitted via Smartsite and shared with the class.

- Skill sets to merge:
  Domain specific questions / Methods / Data sets

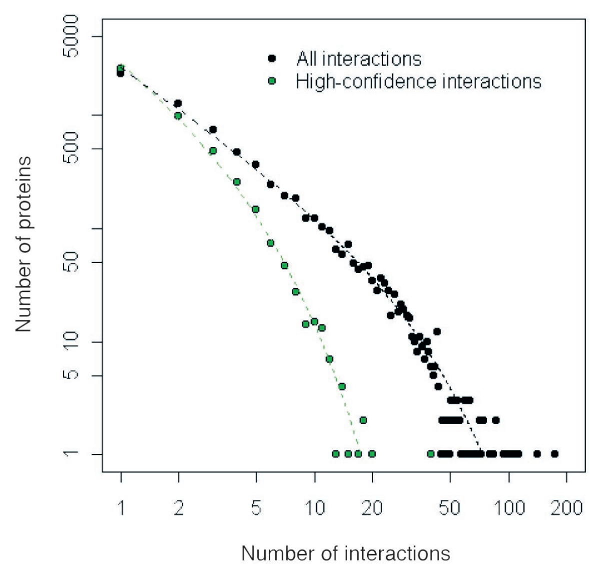
- We have a lot of ideas:
  - Ranking in networks
  - Ising models (esp on hierarchical lattices)
  - Multilayer networks
  - Game theory
  - Shocks and tipping points

Back to basics of networks
Recall: broad scale degree distribution



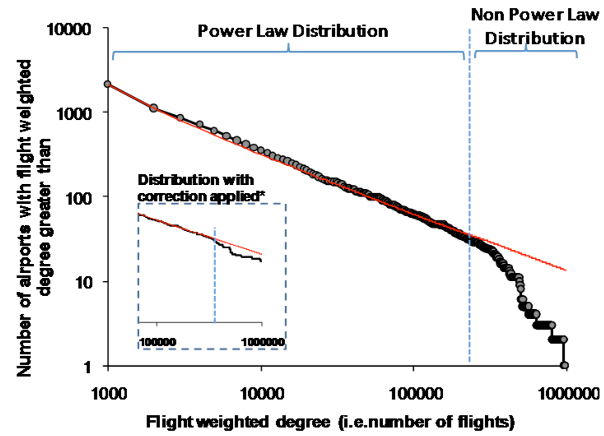
Social contacts

Szendröi and Csányi



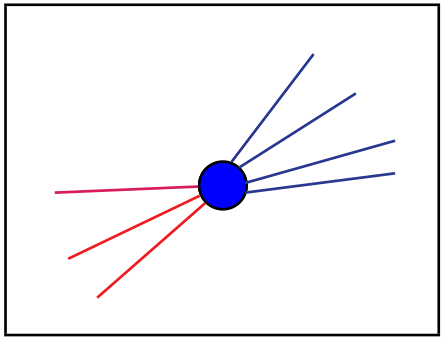
Protein interactions

Giot et al Science 2003



Airport traffic

Bounova 2009



(node degree)

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

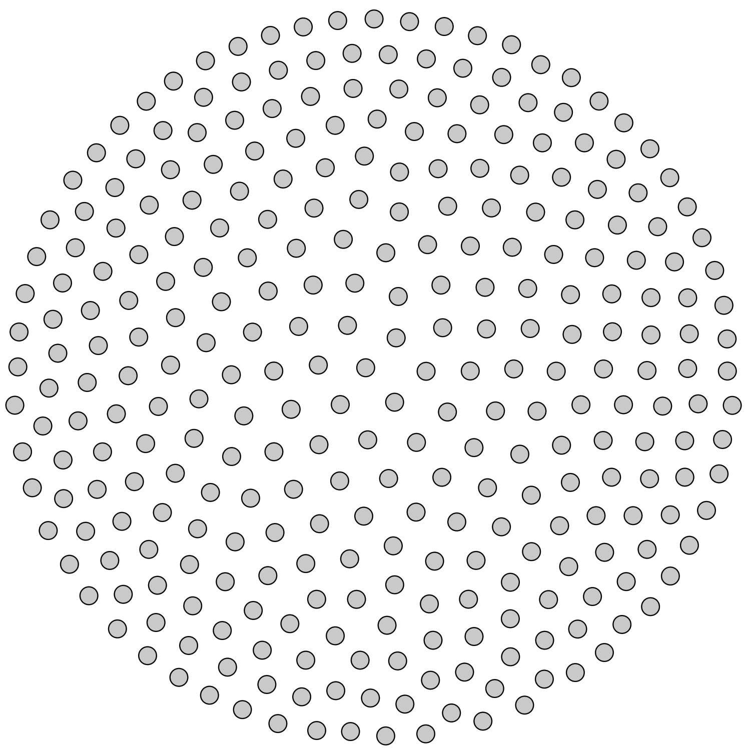
(PDF = probability density function)

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

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



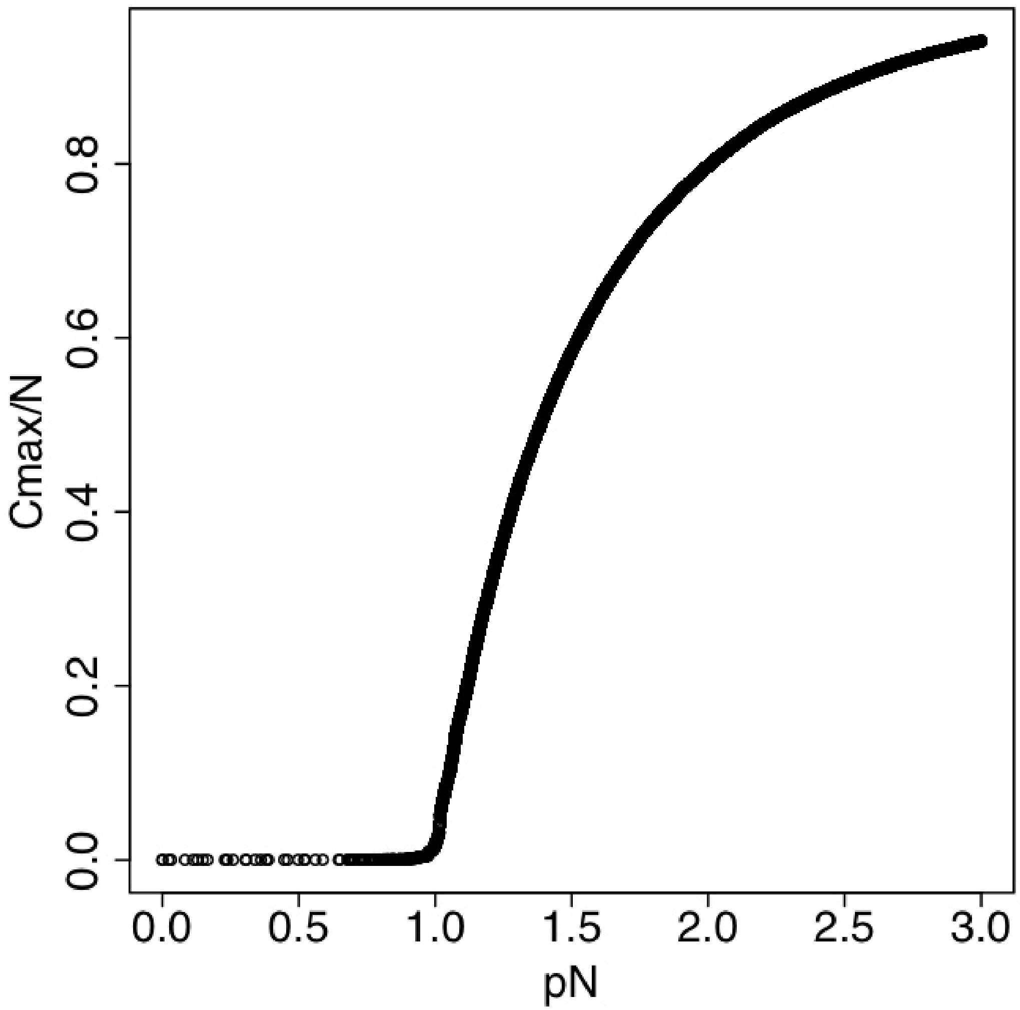
- Start with N isolated vertices.
- Add random edges one-at-a-time.

  $N(N-1)/2$ 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)

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 $t_c = 1$.

Degree distribution of $G(n, p)$

- 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:

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

Bi-nomial converges to Poisson:

$$\lim_{n\to\infty} p_k = z^k e^{-z}/k!$$

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.

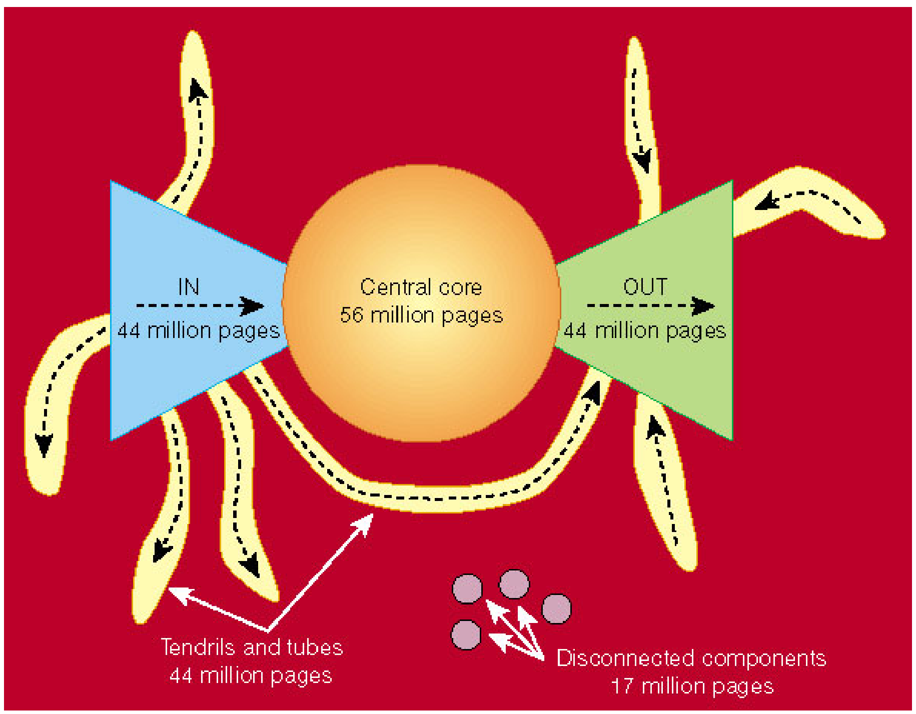
2. Poisson degree distribution: NO! Many real networks have much broader distributions.

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?

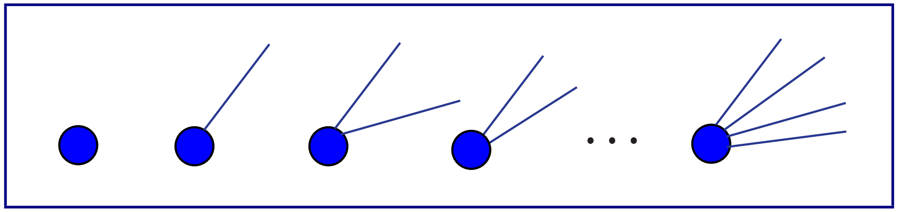
- Much of our basic intuition comes from random graphs.
- Phase transition and the existence of the giant component. Even if not a giant component, many systems have a dominate component much larger than all others.



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

Generalized random graph – advanced HW accommodate any degree sequence

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).

- Emergence of a giant component when expected number of second neighbors greater than expected number of first neighbors.

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 (singularities)
- 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 24,000 citations c.f. Google Scholar):

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

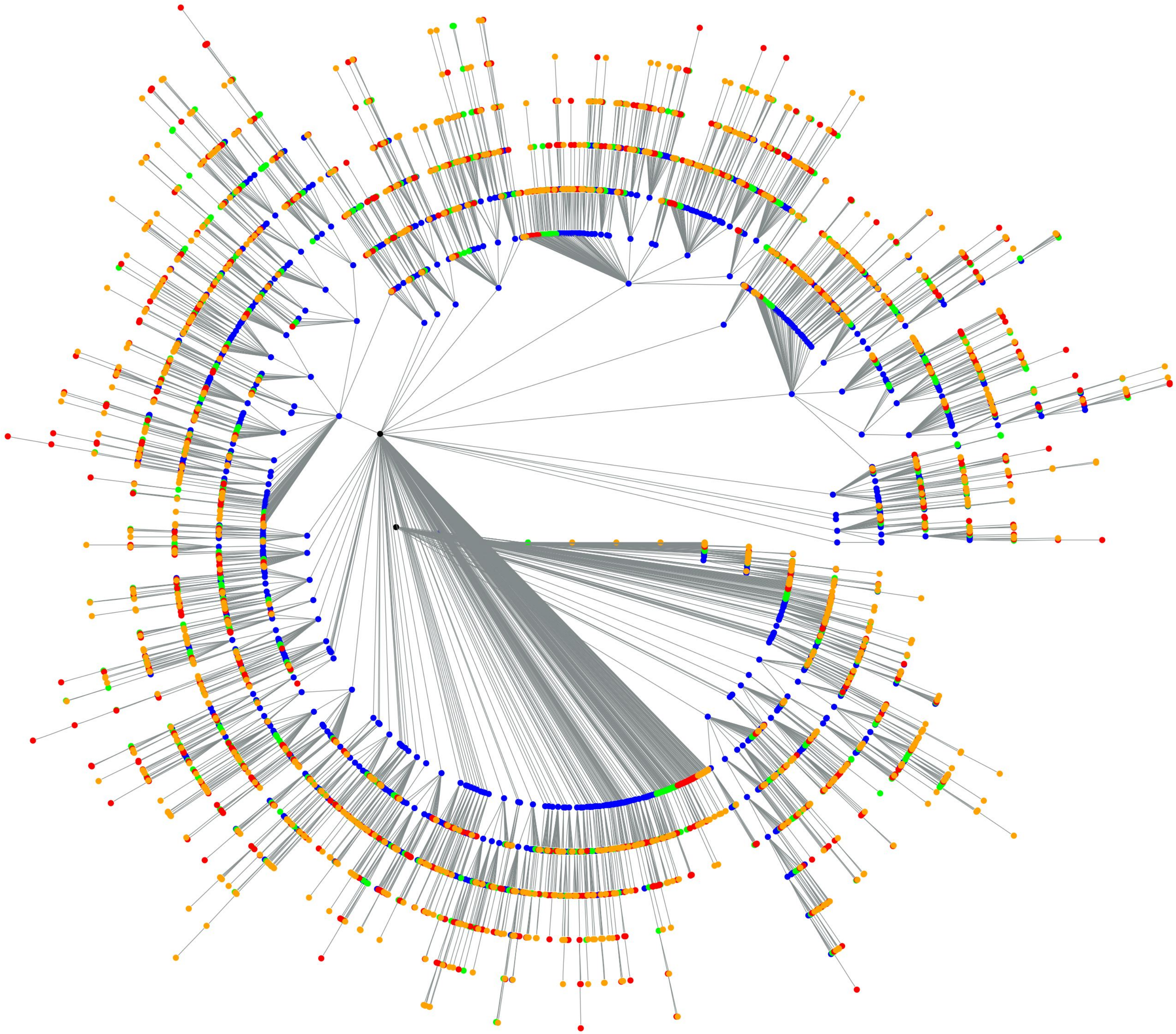
- They coined the term “preferential attachment” to describe the phenomena.

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

The Barabási and Albert model

- A discrete time process.
- Start with single isolated node.
- At each time step, a new node arrives.
- This node makes m connections to already existing nodes. (Why m edges?)
- We are interested in the limit of large graph size.

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



Probabilistic treatment

- Start at $t = 0$ with one isolated node.
  At time t the total number of nodes $n = t$.

- Probability incoming node attaches to node j:

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

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

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

- Prob connect to particular node of degree k at time t:
  $q_{k,t} = k/2mt$

  (Using $q_{k,t}$ since reserving $p_{k,t}$ for degree distribution.)

Network evolution
Process on the degree sequence

- As shown $q_{k,t} = k/2mt$.

- 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"

(Let $n_{k,t} \equiv$ expected number of nodes of degree k at time t, and $n_t \equiv$ total number of nodes at 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 $k = m$: $\quad n_{m,t+1} = n_{m,t} + 1 - \frac{m}{2mt}\, n_{m,t}$

And each arriving node contributes m links:

- 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 $p_{k,t}$

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

$$\rightarrow n_{k,t} = t\, p_{k,t}$$

For each arriving node, after m edges added:

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

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

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

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

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

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.

Did we *prove* the behavior of the degree distribution?

Details glossed over

1. Proof of ***convergence*** to steady-state (i.e. prove $p_{k,t} \to p_k$)

2. Proof of ***concentration*** (Need to show fluctuations in each realization small, so that the average n_k describes well most realizations of the process).

   – For this model, we can use the second-moment method (show that the effect of one different choice at time t dies out exponentially in time).

- see: B. Bollobás, O. Riordan, J. Spencer, and G. Tusnady, "The degree sequence of a scale-free random process", *Random Structures and Algorithms* **18**(3), 279-290, 2001.

Issues

- Whether there are really true power-laws in networks? (Usually requires huge systems, and no constraints on resources).

- Only get $\gamma = 3$!

- Note only the old nodes are capable of attaining high degree.

Generalizations of Pref. Attach.

- Vary steps of P.A. with steps of *random* attachment.

  Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) *Phys Rev Lett* 85.

  Achieves $2 < \gamma < 3$.

- Consider *non-linear* P.A., where prob of attaching to node of degree $k \sim (d_k)^b$.

  "Organization of growing random networks", P. L. Krapivsky and S. Redner, *Phys. Rev. E* 63, 066123 (2001).

  – Sublinear ($b < 1$); deg dist decays faster than power law.

  – Superlinear ($b > 1$): one node emerges as the center of a "star"-like topology.

Alternatives to PA that yield $p_k \sim k^{-\gamma}$

- Copying models
  - WWW:
    - “The web as a graph: measurements, models, and methods”, J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. S. Tomkins, *Proceedings of the 5th annual international conference on Computing and combinatorics*, 1999.
    - “Stochastic models for the Web graph”, R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41st IEEE Symp. on Foundations of Computer Science, pages 5765, 2000.*
  - Biology (Duplication-Mutation-Complementation)
    - “Modeling of Protein Interaction Networks”, Alexei Vázquez, Alessandro Flammini, Amos Maritan, Alessandro Vespignani, *Complexus* Vol. 1, No. 1, 2003

- Optimization models (trade-off between tree-metic and space-metric)
  - Fabrikant-Koutsoupias-Koutsoupias (2002).
  - D’Souza-Borgs-Chayes-Berger-Kleinberg (2007).

Other approaches

- “Winners don’t take all: Characterizing the competition for links on the web”, D M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, C. Lee Giles, *PNAS* 99 (2002).
- First mover advantage
- Second mover advantage

Further reading:

(All refs available on "references" tab of course web page)

PA model of network growth

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

- B. Bollobás, O. Riordan, J. Spencer, and G. Tusnady, "The degree sequence of a scale-free random process", *Random Structures and Algorithms* **18**(3), 279-290, 2001.

- Newman Review, pages 30-35.

- Durrett Book, Chapter 4.

Further reading, cont.

Fitting power laws to data

- Newman Review, pages 12-13.
- M. Mitzenmacher, “A Brief History of Generative Models for Power Law and Lognormal Distributions”, *Internet Mathematics* **1** (2), 226-251, 2003.
- A. Clauset, C. R. Shalizi and M.E.J. Newman, “Power-law distributions in empirical data”, *SIAM review*, 2009.

Difference between ER and PA is not due to edge versus node arrival

- **Node-arrival "Erdős-Rényi graph"**

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

  – At each discrete time step a new node arrives, and with probability δ a new randomly selected edge arrives.

  – Emergence of giant component only if $\delta \geq 1/8$.

  – (That "giant" is finite even as $n \to \infty$).

  – Positive degree-degree correlations (higher degree by virtue of age).

- **Edge-arrival PA graph**

  K.-I. Goh, B. Kahng, D. Kim, *Phys. Rev. Lett.* **87**, 278701 (2001).
  F. Chung and L. Lu, *Annals of Combinatorics* **6**, 125 (2002).

  – Initialized with N isolated nodes, labeled $i \in \{1, 2, ..., N\}$, where each node i has a weight $w_i = (i + i_0 - 1)^{-\mu}$.

  – Two vertices (i, j) selected with probability $w_i / \sum_k w_k$ and $w_j / \sum_k w_k$ respectively and connected by an edge.

  – (Master eqn analysis: Lee, Goh, Kahng and Kim, Nucl. Phys. B 696, 351 (2004).)

Summary of “Master eqn” / rate eqn approach

- Let $n_{k,t}$ denote the *expected (i.e. average)* number of nodes of degree k expected at time t into the process. (Note $n_{k,t}$ is a real number, not an integer.)

- Write $n_{k,t+1}$ in terms of the $n_{k,t}$'s, accounting for the rates at which node degree is expected to change.

- Translate from $n_{k,t}$ to $p_{k,t} = n_{k,t}/n_t$, which is equal to $n_{k,t}/t$ for PA.

- Assume $p_{k,t} \to p_k$.

- Solve for a recurrence relation for the p_k's.

- Need to prove *convergence and concentration*

Preferential Attachment and "Scale-free networks"

Why a power law is "scale-free"

- Power law for "x", means "scale-free" in x:

$$p(bx) = (bx)^{-\gamma} = b^{-\gamma} p(x)$$

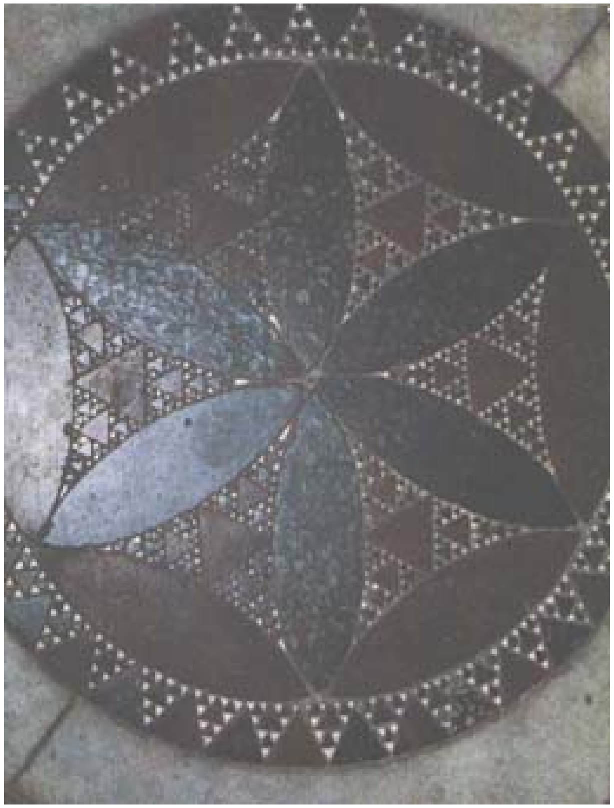
$\boxed{\frac{p(bk)}{p(k)} = b^{-\gamma}}$ regardless of k.

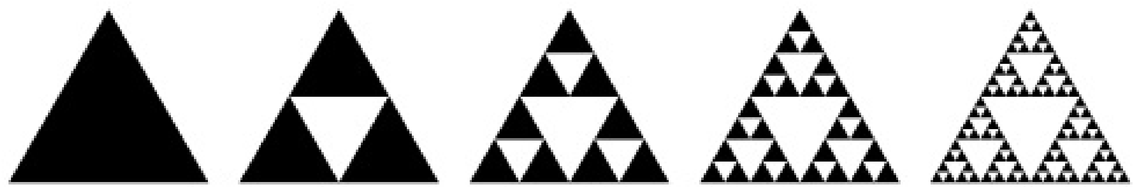
In contrast consider: $p(k) = A\exp(-k)$.

So $p(bk) = A\exp(-bk)$.

$\boxed{\frac{p(bk)}{p(k)} = \exp[-k(b-1)]}$ dependent on k

Self-similar/scale-free fractal structures





Sierpinski Sieve/Gasket/Fractal, $N \sim r^d$.

When r doubles, N triples: $3 = 2^d$

$d = \log N / \log r = \log 3 / \log 2$

Power law degree distribution $\neq$ "scale-free network"

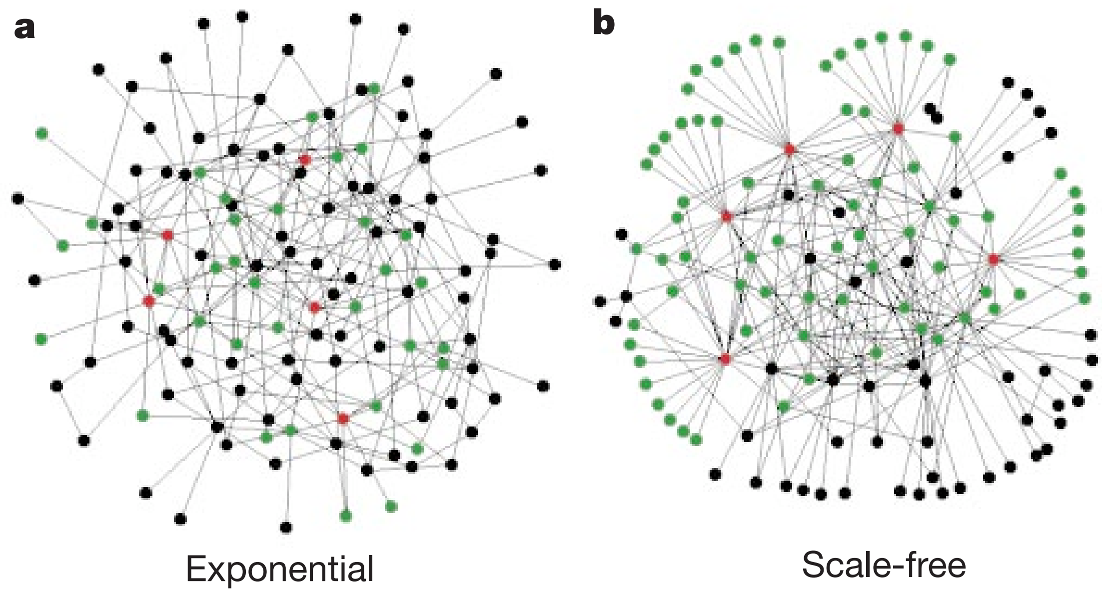
- Power law for "x", means "scale-free" in x.
- BUT only for that aspect, "x". May have a lot of different structures at different scales.
- **More precise: "network with scale-free degree distribution"**

Robustness of a network

- Robustness/Resilience: A network should be able to absorb disturbance, undergo change and essentially maintain its functionality despite failure of individual components of the network.

- Often studied as maintaining connectivity despite node and edge deletion.

Albert, Jeong and Barabasi, “Error and attack tolerance of complex networks”, Nature, **406** (27) 2000.



N=130, E=215, Red five highest degree nodes; Green their neighbors.

- Exp has 27% of green nodes, SF has 60%.
- PLRG: Connectivity extremely robust to random failure.
- PLRG: Connectivity extremely fragile to targeted attack (removal of highest degree nodes).