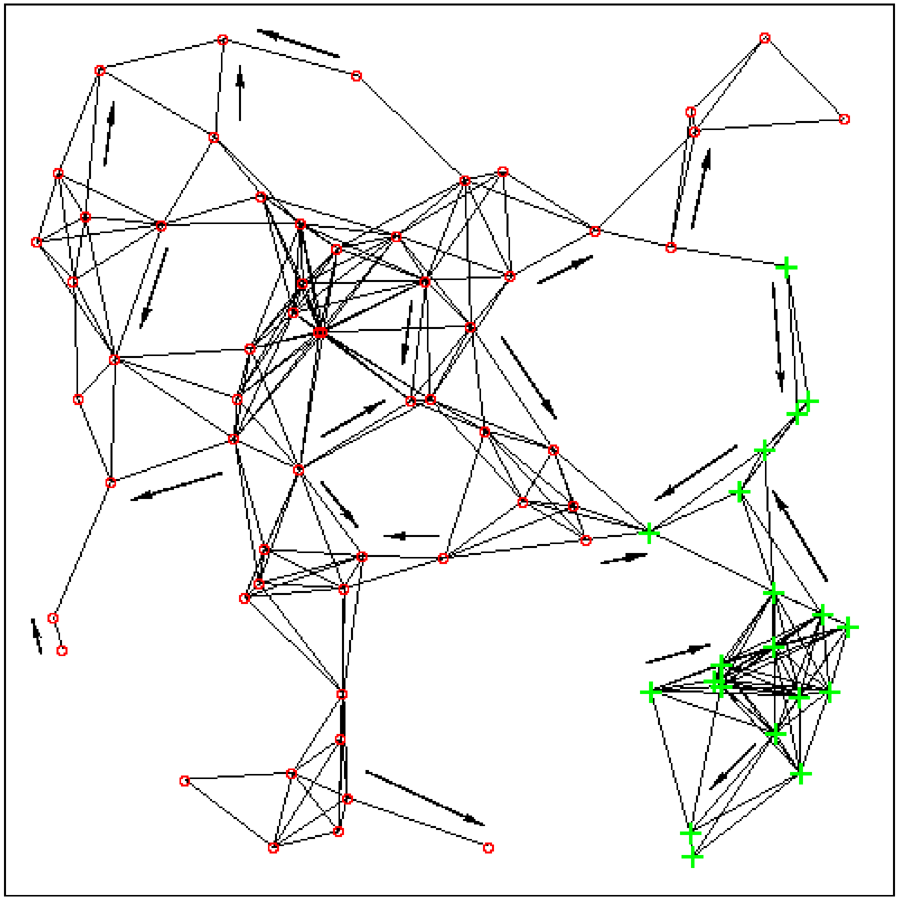
MAE 298, Lecture 3

April 6, 2006



"Network Growth Models"

Recall: 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.

Properties (1) and (3) are in-line with real-world networks, but not properties (2) and (4).

Degree distribution

- A large number of real-world networks, from an extensive range of applications, have “heavy-tailed” degree distributions.

- Also can be considered “broad-scale”.

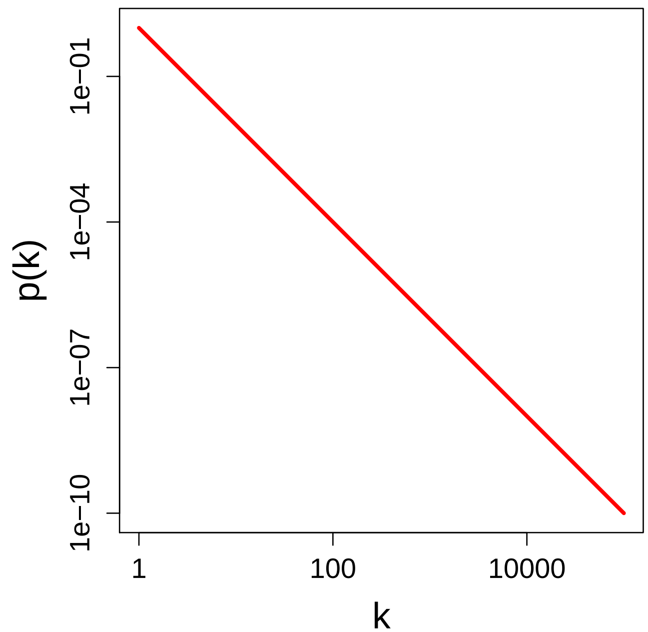
- The simplest example of such a distribution is a power law.

What is a power law?

(Also called a “Pareto Distribution” in statistics).

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

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



Properties of a power law distribution

See chalk board discussion. Essentially:

$$\boxed{\int x^n = \frac{1}{n+1}\, x^{n+1}}$$

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 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.

Power laws in the real world

Confusion

- Power law
- Log normal
- Weibull

All three 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): $P_k = \sum_{l=k}^{\infty} p_l$.

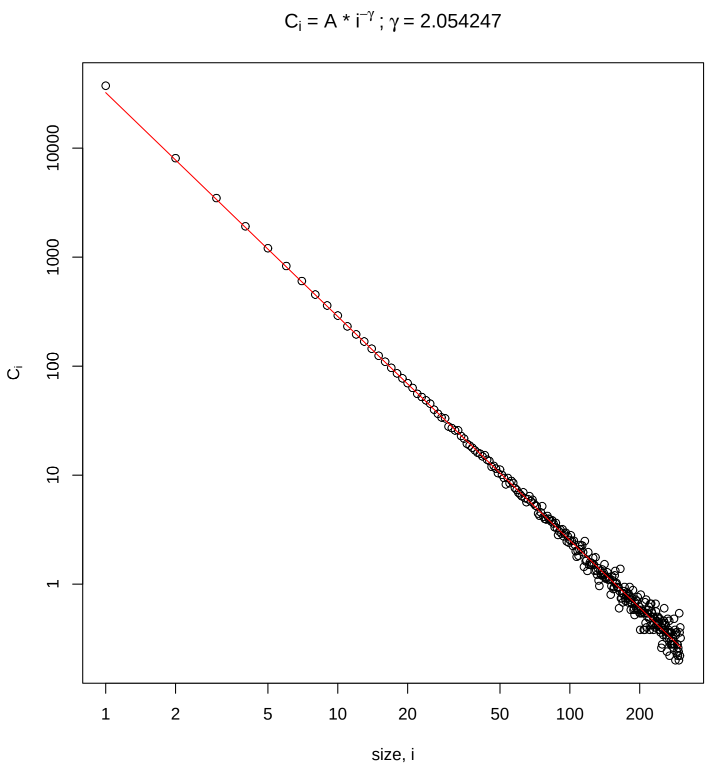
For more details see:

- Newman Review, pages 12-13.
- Mitzenmacher Review (reference given at end).

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, $\gamma = 2.5$)



Power laws in social systems

- Popularity of web pages: $N_k \sim k^{-1}$
- 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 is proportional to size,

$$\frac{dP(s)}{dt} \propto s$$

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 (about 1000 citations in SCI) paper:

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

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

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.

Probability

- Probability incoming node attaches to node j:

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

- Probability incoming node attaches to any node of of degree k:

  (# nodes of degree k)/(# nodes) x (degree of that node)/(degree sum over all nodes) =

$$\boxed{\frac{kp_k}{\sum_k d_k} = \frac{kp_k}{2mn}}$$

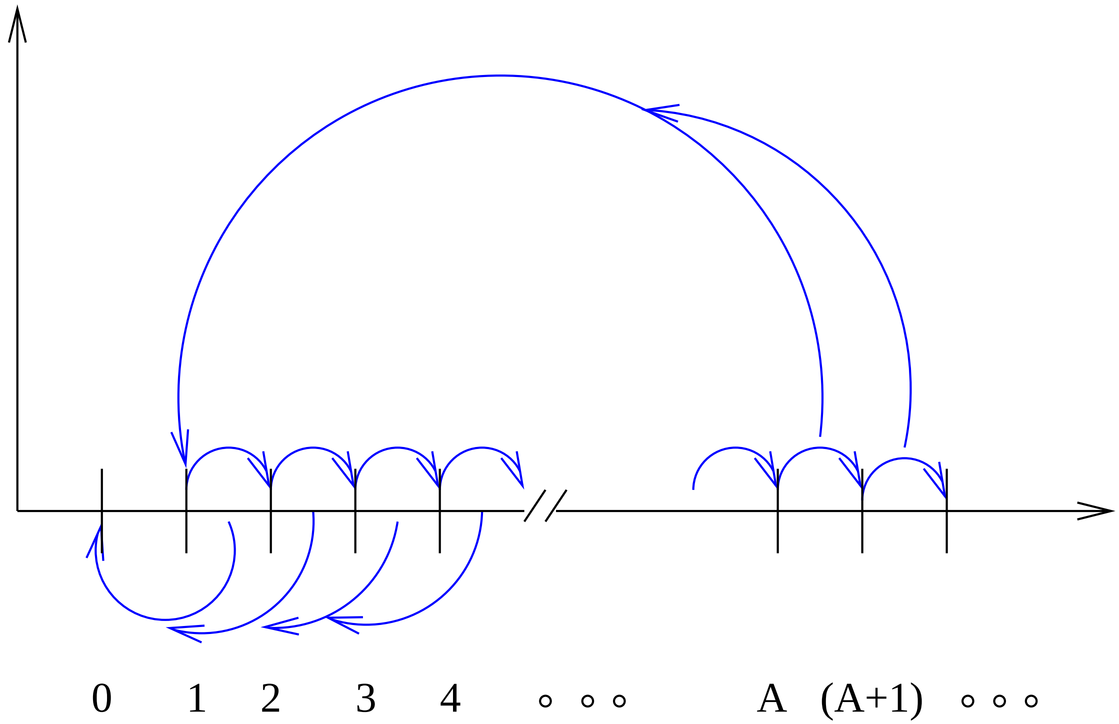
Network evolution
Process on the degree sequence

- Note that p_k will change in time!
  So we show denote this explicitly: $p_{k,t}$

- 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.

Markov flow



Process on the degree sequence, cont.

(Let $n_{k,t} \equiv$ 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 $k = m$: $\quad n_{m,t+1} = n_{m,t} + 1 - \frac{m}{2mt}\, n_{m,t}$

But each arriving node contributes m links:

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

Translating back to probabilities

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

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

Rearranging and solving for p_k:

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

Recursing 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 $k \gg 1$

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

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

Details glossed over

1. Proof of *convergence* to steady-state

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).

Issues

- Whether there are really true power-laws in networks? (Usually requires huge systems, and no constraints on resources).
- Only get $\gamma = 3$!

Generalizations of Pref. Attach.

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

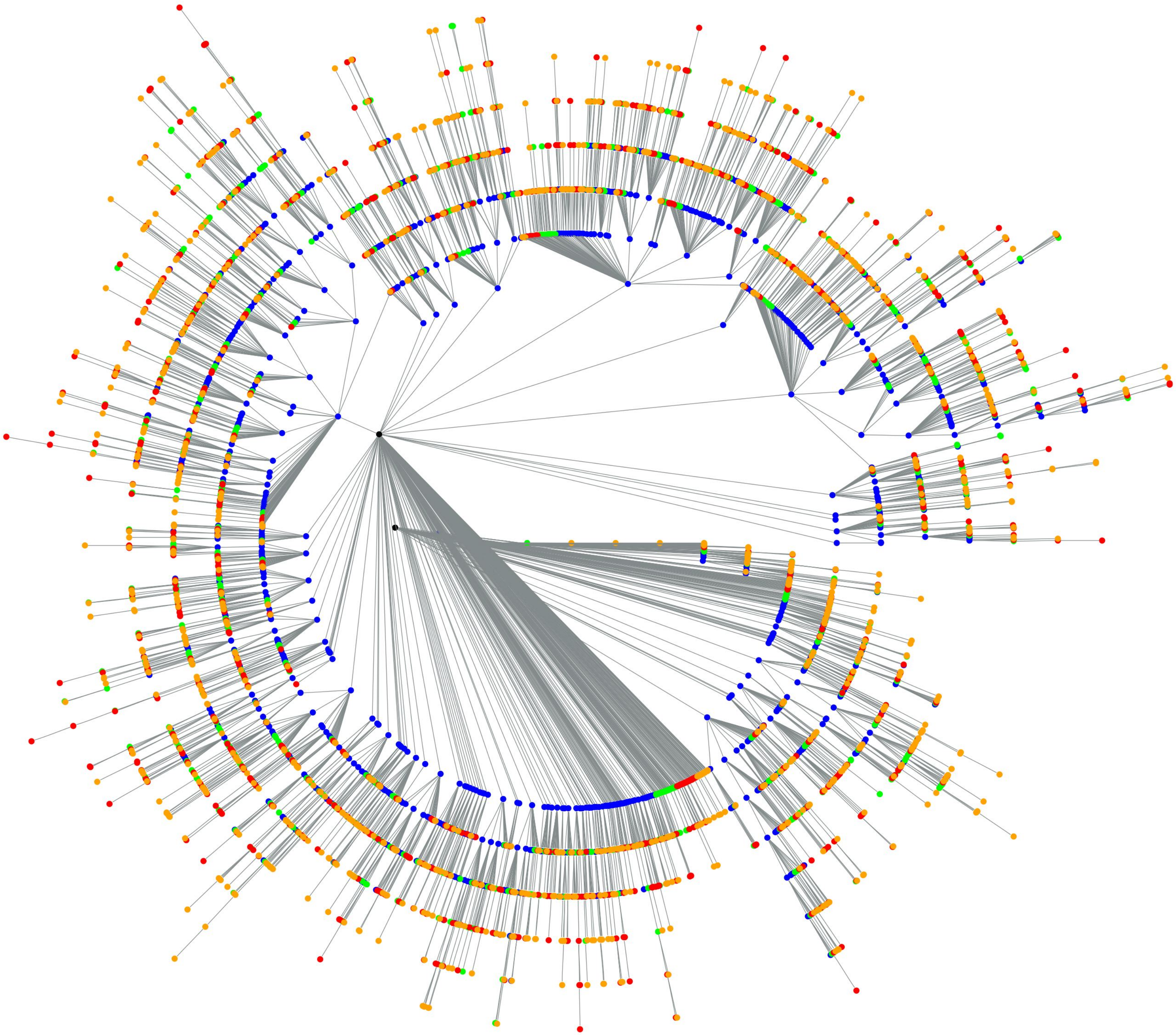
Simulating PA

Basic code for simulating PA with $m = 1$ using R:

- runPA $\leftarrow$ function(N=100)

```
{
    # outLink[i] is the parent of i
    outLink ← numeric(N)
    # numlinks[i] is number total-links (in and out) for node i
    numLinks ← numeric(N)+1
    for(i in 2:N)
    {
        p ← sample(c(1:(i-1)),size=1,prob=numLinks[1:(i-1)])
        outLink[i] ← p
        numLinks[p] ← numLinks[p]+1
    }
    return(list(outLink, numLinks))
}
```

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



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.