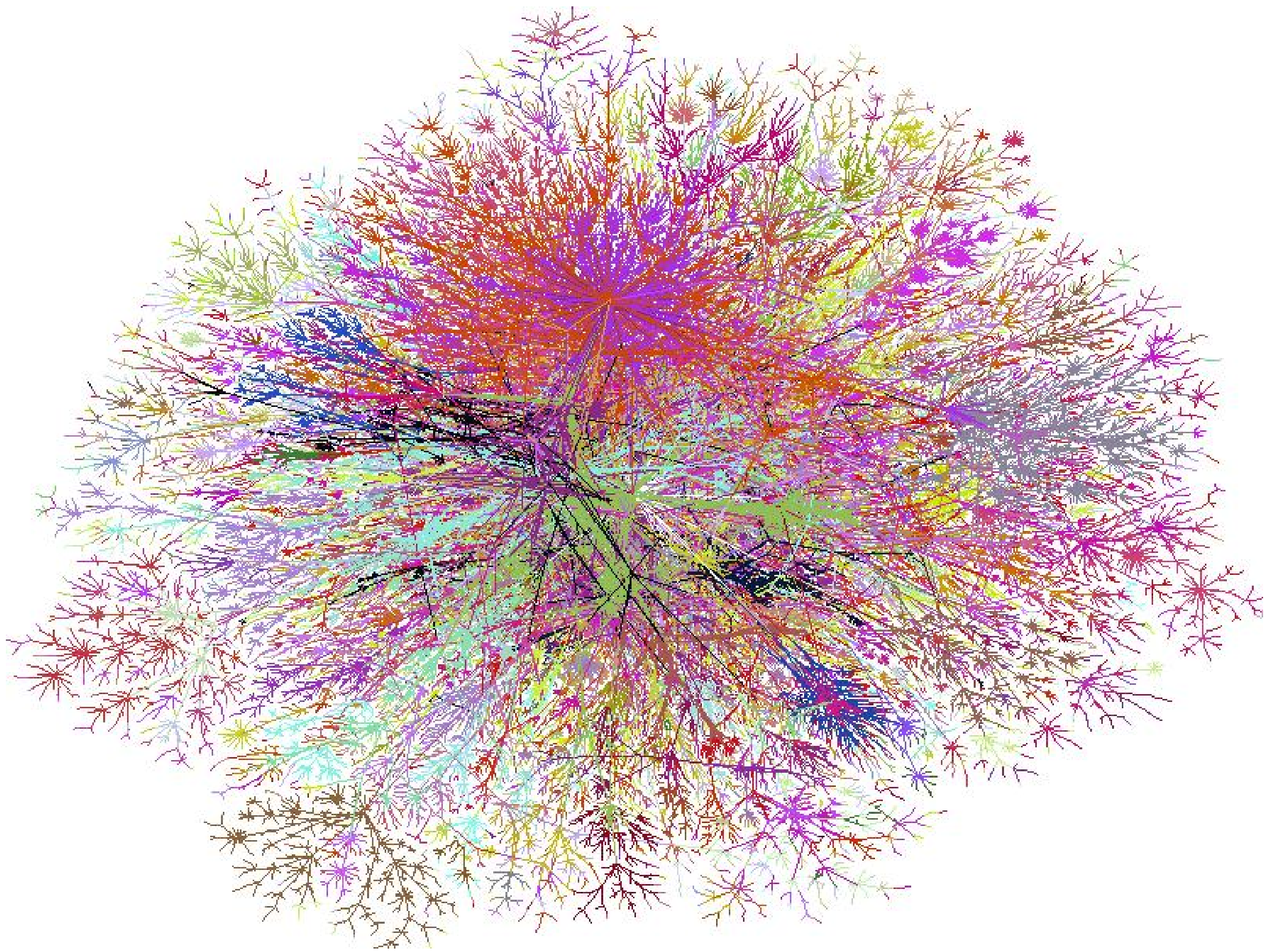
ECS 253 / MAE 253, Lecture 4

April 12, 2018



"Power laws and network robustness"

Announcements

- 24 people turned in project ideas.
  (Check the canvas gradebook that your entry was received.)

- I assume the remaining 36 people will choose “Track B” advanced HWs.

Our ideas

- Macaque monkey networks (Marton Posfai)
  - Community structure
  - Hierarchical structure
  - Temporal patterns and motifs

- Power grids (Martin Rohden)
  - Modeling cascading failures
  - Comparison of artificial and real-world power grids

Our ideas, cont.

- Social systems and influence (Keith Burghardt)
  - Distinguishing bots versus humans in online networks
  - Complex contagion: optimal topologies for spreading of cascades
  - How does network structure influence “group think”?
  - How do aggregate country-level networks differ from individual social networks and why?
- Ecological networks and mutualistic interactions (Weiran Cai)
  - Niche competition in networks (both ecology and business)
  - Ecological analogies of the break-down of the New York City Garment Industry
  - Disrupting terrorist networks

Your ideas

- Networks in Github
  - Farhana Sarker, Vaishnavi Karanam
  - Jingxian Liao
- Macaque monkey networks
  - Josephine Hubbard, Meredith Lutz, Niklas Braun
- Social change, political movements, protest networks
  - Calvin Koon-Stack
  - Priscilla Jean Pierre
- Transportation networks
  - Heidi Schweizer: interplay of transportation and food pricing (particularly freight and grains from 2014)
  - Justin Perona, Baotuan Nguyen

Your ideas, cont.

- Power grids
  - Armando Casillas
  - Erin Musabandesu
- Food networks; ontologies and connections to disease
  - Tarini Naravane (collaboration with Barabasi Lab at Northeastern)
- Recommendation systems
  - Ishita Sharma
  - Kenan Nalhant (might switch to Track B)
- Network resilience
  - Molly O'Connor

Your ideas, cont.

- Machine learning You would need to connect this with networks!
  – Liyang Zhong, Zheng Zhang, Quan Zou
  – Trevor Chan (applications to biology; optimal gene regulation)
- Phylogenetic networks and branching processes
  – Kevin Hundnall
- Mutli-threaded web crawler (Needs to be about networks; so ranking is appropriate)
  – Jinhua Cao
- Terrorist networks; mutualistic and parasitic relations with nation states
  – Grayson Gordan

Finalizing project groups

GOAL: Align course projects with Postdoctoral Scholar interests.

- Decision on Track A or Track B by next Tues (April 17)
- Track A: Finalize topic area by next Tues (April 17)
- Track A: Finalize the teams by next Thurs (April 19)
  This will be HW1a assignment.

Now back to network science content.

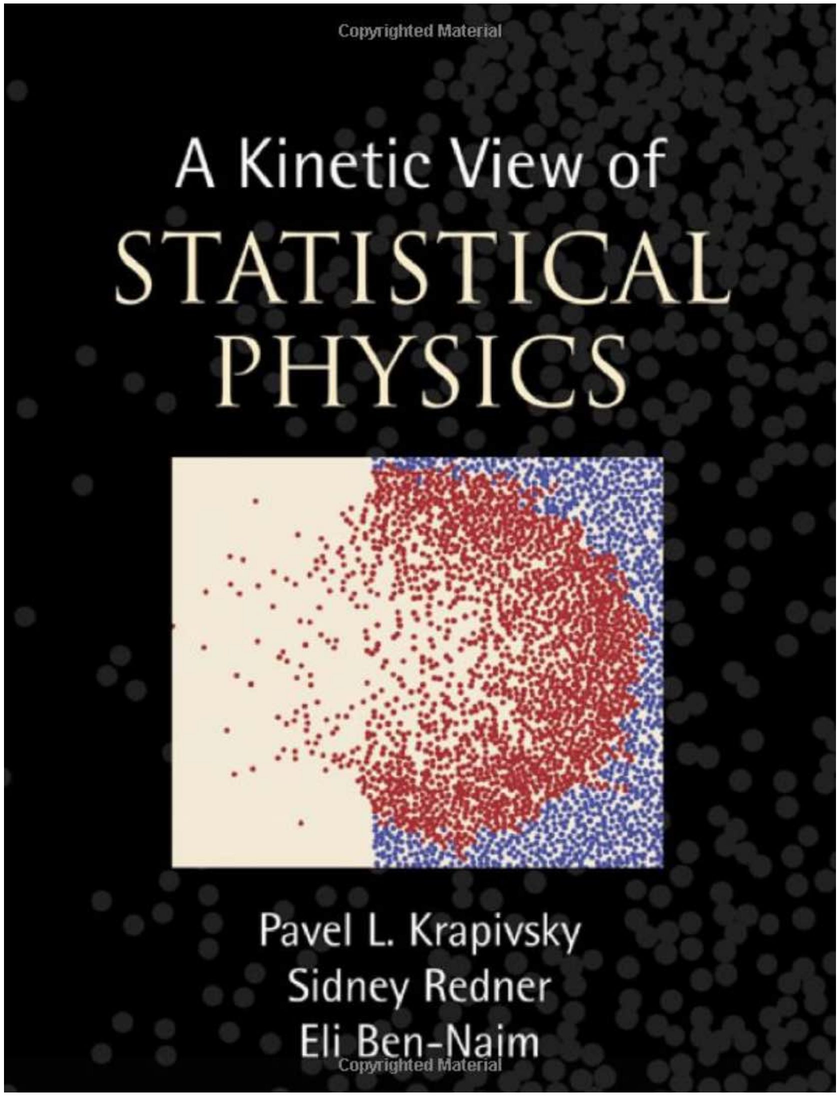
Network models studied so far

- Erdős-Rényi random graphs, $G(N, p)$
  – Initialized with N isolated nodes
  – Edges arrive in discrete time process with uniform prob.
  – Poisson degree distribution
  – No clustering
  – Emergence of a giant component

- Preferential attachment graphs
  – Initialized with one (or a small set) of seed nodes
  – Nodes arrive and attach with m edges choosing "parent" with prob proportional to parent's degree.
  – Power law deg dist with $\gamma = 3$
  – Clustering tuned by setting m
  – Fully connected network by construction

Summary of kinetic theory / rate eqn approach

- A stochastic, discrete time process for an evolving graph $G(t) \to G(t+1)$.
- **Assumption 1: Study the average ("mean-field") random graph in limit $N \to \infty$.**
- Let $n_{k,t}$ denote the *expected (i.e. average)* number of nodes of degree k at time t into the process. (So $n_{k,t}$ 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.
- Note $p_{k,t} = n_{k,t}/n_t$ and rewrite in terms of probabilities. (Note you can formulate the equation in terms of probabilities from beginning).
- **Assumption 2: Assume steady state** $p_{k,t} \to p_k$.
- Solve for a recurrence relation for the p_k's. For PA, $p_k = k^{-3}$ for large k.
- Need to show **concentration** (Assump 1) and **convergence** (Assump 2)



Cambridge Univ Press, 2010.

Barabási-Albert model: "Preferential attachment"

- A network growth model, starting from a small number m_0 of seed nodes.
- Each discrete time step a new node arrives and adds m edges to the graph.
- Each new edge connects to a node of degree k with probability $d_k / \sum_k d_k$.

The definition of the Barabási-Albert model leaves many mathematical details open:

- It does not specify the precise initial configuration of the first m_0 nodes.
- It does not specify whether the m links assigned to a new node are added one by one, or simultaneously. (We assume simultaneously and analyze the process for large n. In the limit $n \to \infty$, the likelihood of multi-edges approaches zero.)

PA via "rate eqns" / "kinetic theory"

Evolution of the typical (mean-field) graph

– Let $n_{k,t}$ denote the *expected number* of nodes of degree k at time t.
– Thus $p_{k,t} = n_{k,t}/n_t$.

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

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

PA analyzed via rate equation approach , solving for p_k

By definition, $p_{k,t} = n_{k,t}/n_t$.

Rewriting and assuming steady-state, that $p_{k,t} \to p_k$, yields:

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

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

For $k \gg 1$

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

Further mathematical details

PA analyzed via rate equation approach

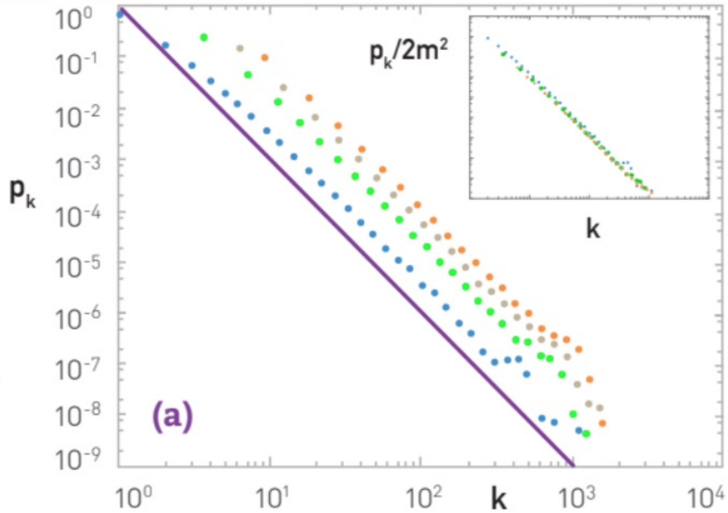
- Krapivsky, Redner, Leyvraz, PRL 2000
- Dorogovtsev, Mendes, Samukhin, PRL 2000

Proof of PA (including concentration and convergence)

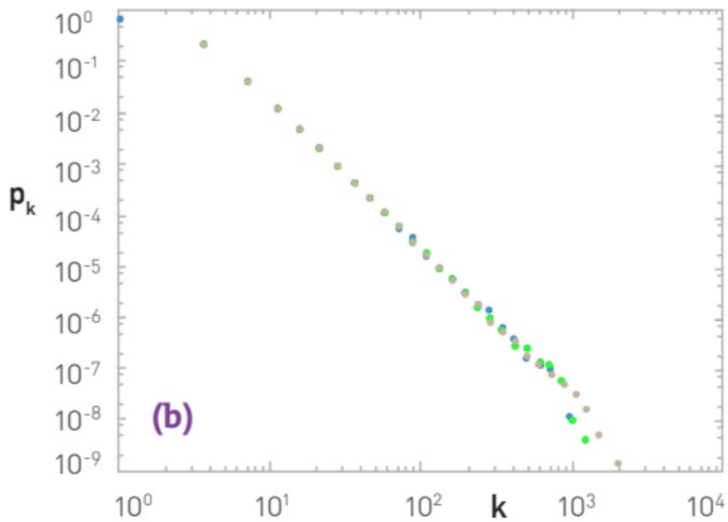
- Bollobas, O. Riordan, J. Spencer, and G. Tusnady, The degree sequence of a scale-free random process, Random Struc. Alg. 18(3), 279-290, 2001

NUMERICAL SIMULATION OF THE BA MODEL

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



(a) We generated networks with N=100,000 and m_0=m=1 (blue), 3 (green), 5 (grey), and 7 (orange). The fact that the curves are parallel to each other indicates that γ is independent of m and m_0. The slope of the purple line is -3, corresponding to the predicted degree exponent γ=3. Inset: (5.11) predicts p_k~$2m^2$, hence $p_k/2m^2$ should be independent of m. Indeed, by plotting $p_k/2m^2$ vs. k, the data points shown in the main plot collapse into a single curve.



(b) The Barabási-Albert model predicts that p_k is independent of N. To test this we plot p_k for N = 50,000 (blue), 100,000 (green), and 200,000 (grey), with m_0=m=3. The obtained p_k are practically indistinguishable, indicating that the degree distribution is stationary, i.e. independent of time and system size.

Barabasi, Network Science book

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

- **Edge-arrival PA graph**
  - K.-I. Goh, B. Kahng, D. Kim, *Phys. Rev. Lett.* **87**, (2001).
  - W. Aiello, F. Chung, and L. Lu. "A random graph model for power law graphs." Experimental Mathematics 10.1 (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.

– Yields $p_k = Ak^{-\gamma}$ with $\gamma = \mu = -1/(\gamma - 1)$.

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

* "Chung-Lu" model used extensively to generate graphs.

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

- **Erdős-Rényi-like process with node arrival**

  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 $\delta \geq 1/8$.

  – Infinite order phase transition. (Kosterlitz Thouless transition.)

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

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

Preferential Attachment and "Scale-free networks"

Why a power law is "scale-free"

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

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

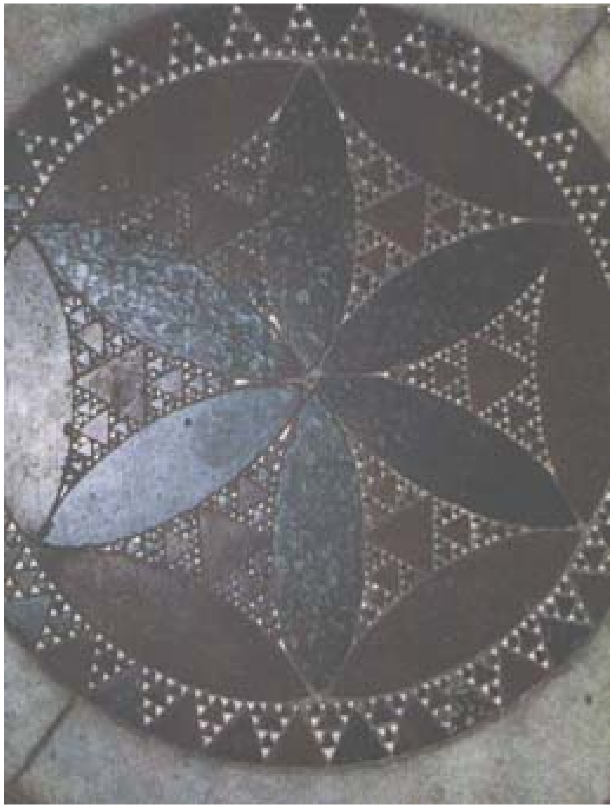
$\boxed{\frac{p(bk)}{p(k)} = b^{-\gamma}}$ regardless of k.

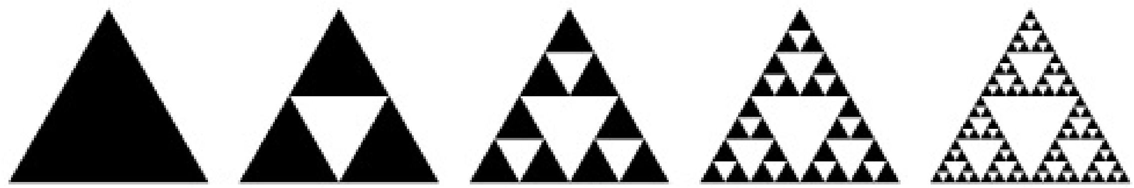
In contrast consider: $p(k) = A \exp(-k)$.

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”**

  Power Law Random Graph (PLRG) is a more precise term

  Yet **“Scale free network”** now used pervasively: e.g., Wikipedia: “a network whose degree distribution follows a power law, at least asymptotically. ”

Power laws in real-world networks?

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.

The controversy continues

- Arxiv posting from Jan 2018, Anna D. Broido, Aaron Clauset "Scale-free networks are rare"
- Quanta Magazine, Feb 15, 2018, "Scant Evidence of Power Laws Found in Real-World Networks"
- Quanta article is carried by *The Atlantic*
- Barabasi response: https://www.barabasilab.com/post/love-is-all-you-need

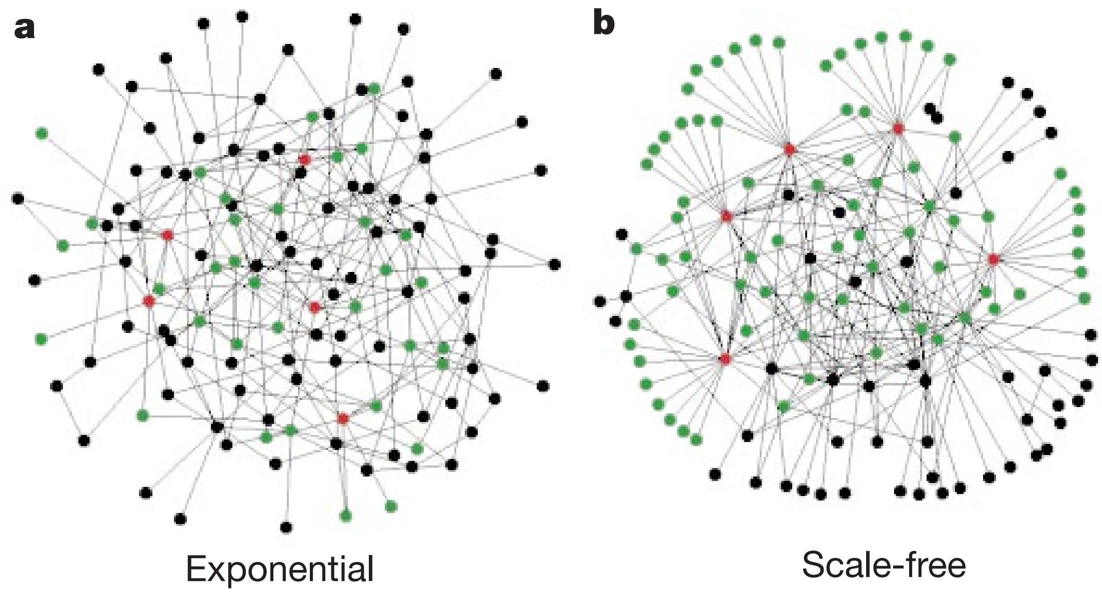
In part, the implications of Power Law Random Graphs have consequences on robustness and vulnerability as we see next.

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.

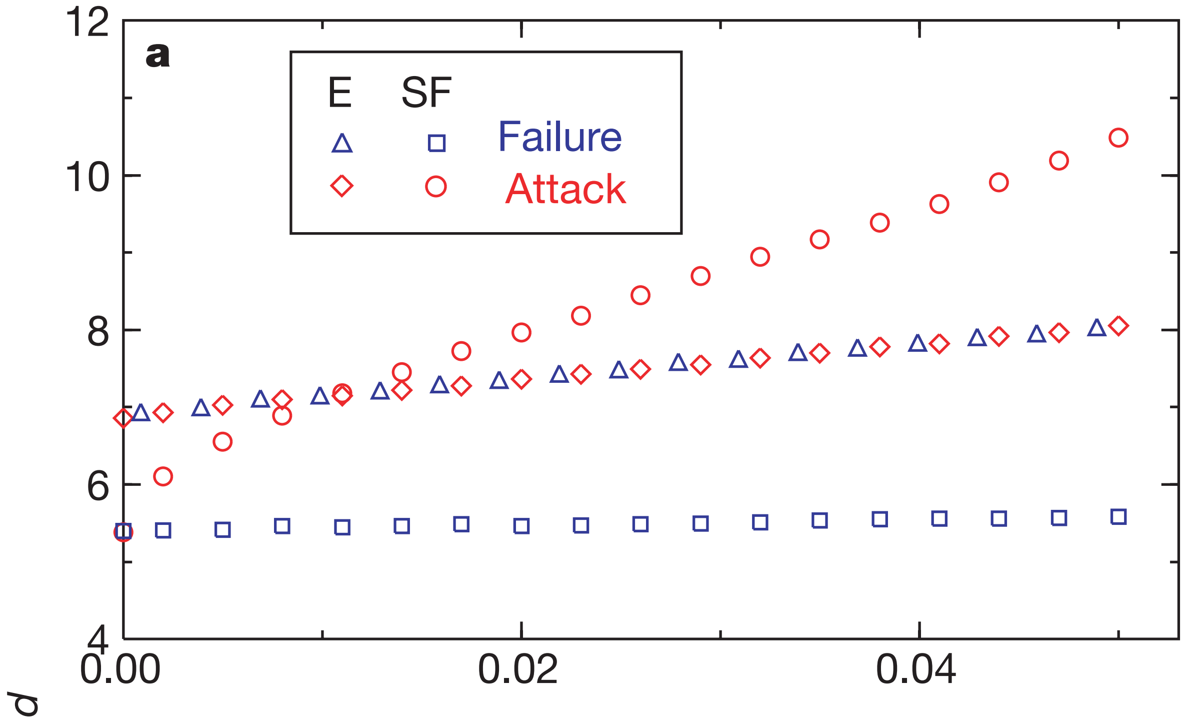
Albert, Jeong and Barabasi, "Error and attack tolerance of complex networks", Nature, **406** (27) 2000.



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

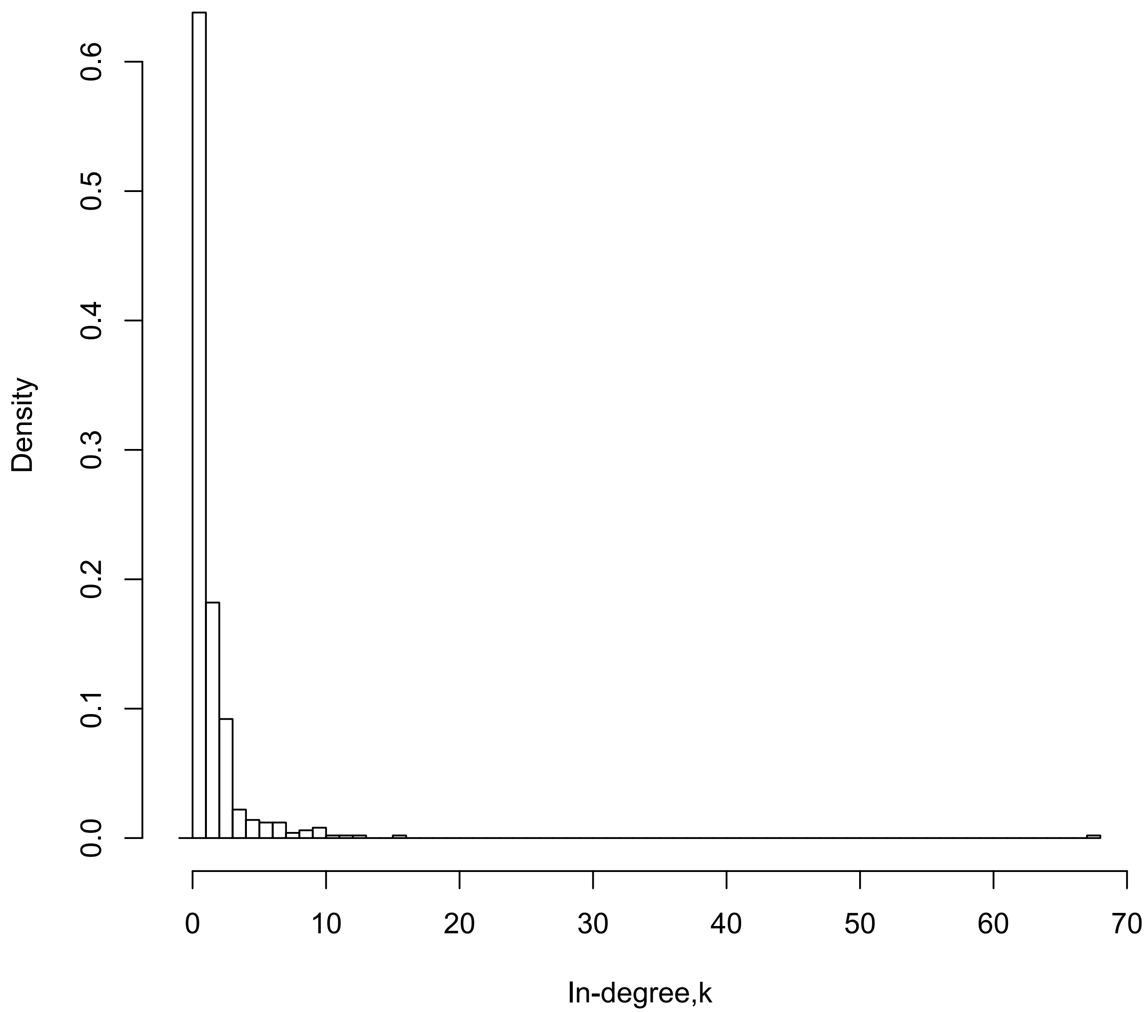
- Exp has 27% of green nodes, SF has 60%.
- PLRG: Connectivity extremely robust to random failure.
- PLRG: Connectivity extremely fragile to targeted attack (removal of highest degree nodes).

Exponential vs scale-free: Robustness



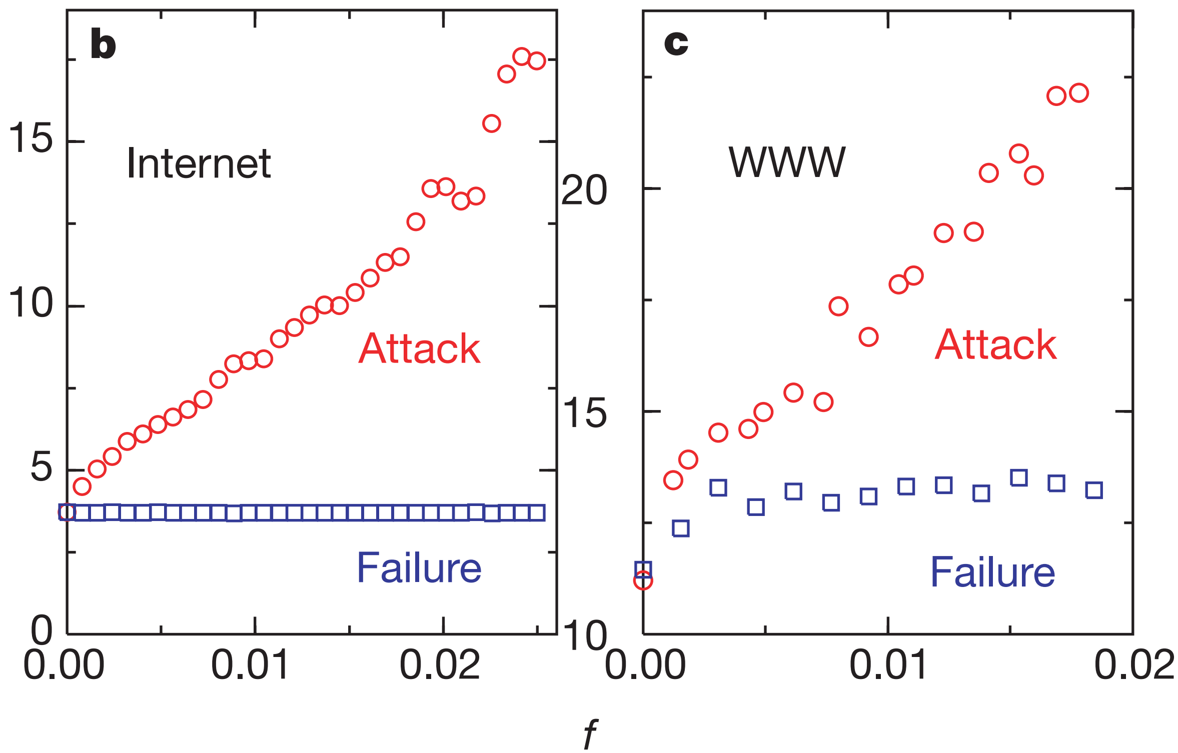
- (Remember, bigger diameter is worse.)
- SF are extremely robust to random failure (blue squares). Remove fraction of nodes at random, and no change in diameter.
- SF are very fragile to targeted attack (removal of highest degree nodes).

Histogram of a typical PA run
Degree distribution (Here N=500)



- Choosing node at random overwhelmingly leads to low degree node

Degree-targeted removal on real sample topologies



- Used the topological map of the Internet, containing 6,209 nodes and 12,200 links $< k >= 3.4$), collected (in 1999 or 2000) by the National Laboratory for Applied Network Research http://moat.nlanr.net/Routing/rawdata/

- World-Wide Web data measured on a sample containing 325,729 nodes and 1,498,353 links, such that $< k >= 4.59$.

Albert, Jeong and Barabasi, Nature, 406 (27) 2000

"The Achilles Heel of the Internet"

- "How robust is the Internet?" Yuhai Tu, *Nature* (New and Views) **406** (27) 2000.

- "Scientists spot Achilles heel of the Internet", CNN, July 26, 2000.

Percolation theory to show the similar results follow in an analytic mathematical formulation

- R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, “Resilience of the Internet to Random Breakdowns”, *Phys. Rev. Lett.* 85, 4626 (2000).

- Callaway, Duncan S.; M. E. J. Newman, S. H. Strogatz and D. J. Watts, “Network Robustness and Fragility: Percolation on Random Graphs”. *Phys. Rev. Lett.* 85: 546871 (2000).

- $\langle k \rangle$ finite, but $\langle k^2 \rangle \to \infty$ for PLRG with $2 < \gamma < 3$, the cornerstone for the arguments.

Results from Callaway et al
Robustness to random removal

- Degree dist, $p_k \sim k^{-\gamma} e^{-k/C}$ (power law with cutoff w $C \to \infty$).
- Let q be probability that a vertex is "active"/"infected".
  For simplicity assume independent of k.
- Then $p_k q$ is probability of having degree k and being infected.
- Calculate $\langle s \rangle$, the mean cluster size of infected nodes. Find (via generating functions ... details later in the course) that

$$\langle s \rangle = q + \frac{q^2 \langle k \rangle}{1 - (q \langle k^2 \rangle / \langle k \rangle)}$$

- $\langle s \rangle \to \infty$ when denominator $1 - q \left\langle k^2 \right\rangle / \langle k \rangle = 0$, i.e.,

$\boxed{q_c = \frac{\langle k \rangle}{\langle k^2 \rangle}}$ Infinite cluster even if probability $\to$ 0, when $p_k \sim k^{-\gamma}$ for $2 < \gamma < 3$) .

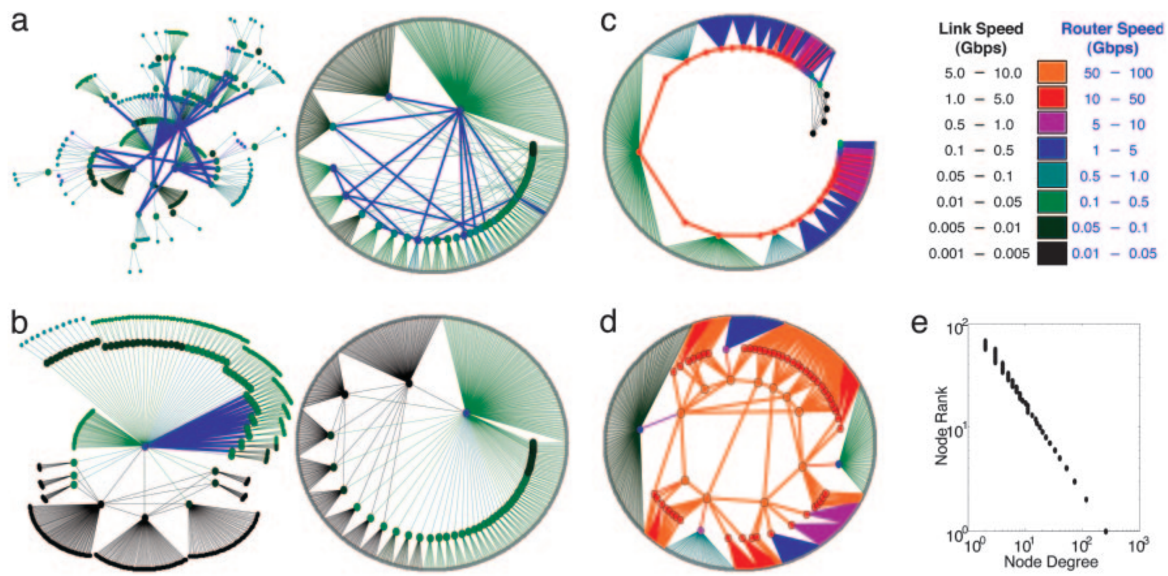
Does the ensemble of random graphs really model engineered or biological systems?

(Is the Internet a random scale-free graph?)

Random vs engineered vs evolved (e.g. biological) systems

- REDUNDANCY!!! a key principle in engineering (and evolution?).

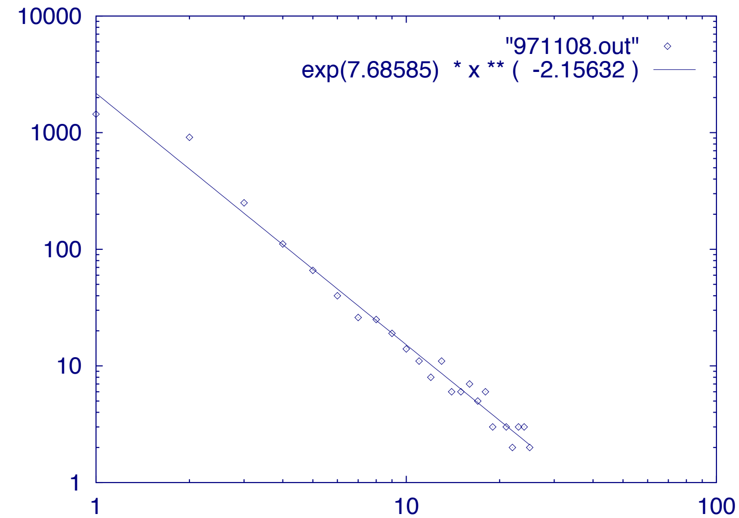
- The ‘robust yet fragile’ nature of the Internet
  Doyle, Alderson, Li, Low, Roughan, Shalunov, Tanaka, Willinger, PNAS **102** (4) 2005.



- Degree distribution is not the whole story.

Wikipedia entry on "scale-free networks"

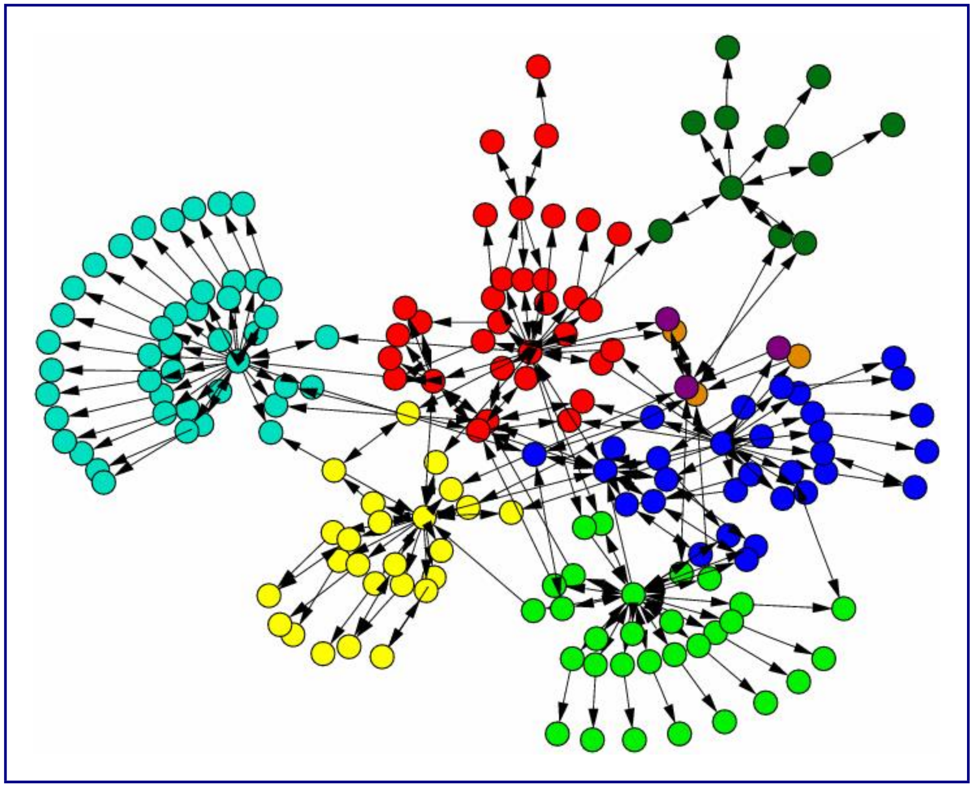
- Good discussion of the history and controversy
  - Faloutsos SIGCOMM 1999 paper on power law in Internet based on trace route sampling.

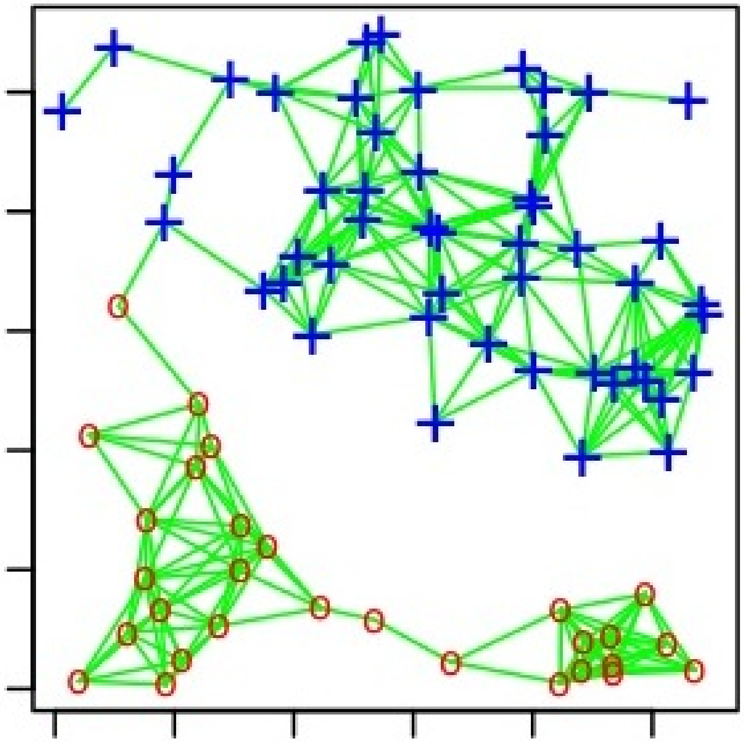


(a) Int-11-97

  - Although many real-world networks are thought to be scale-free, the evidence often remains inconclusive, primarily due to the developing awareness of more rigorous data analysis techniques.

Effectively breaking up different networks





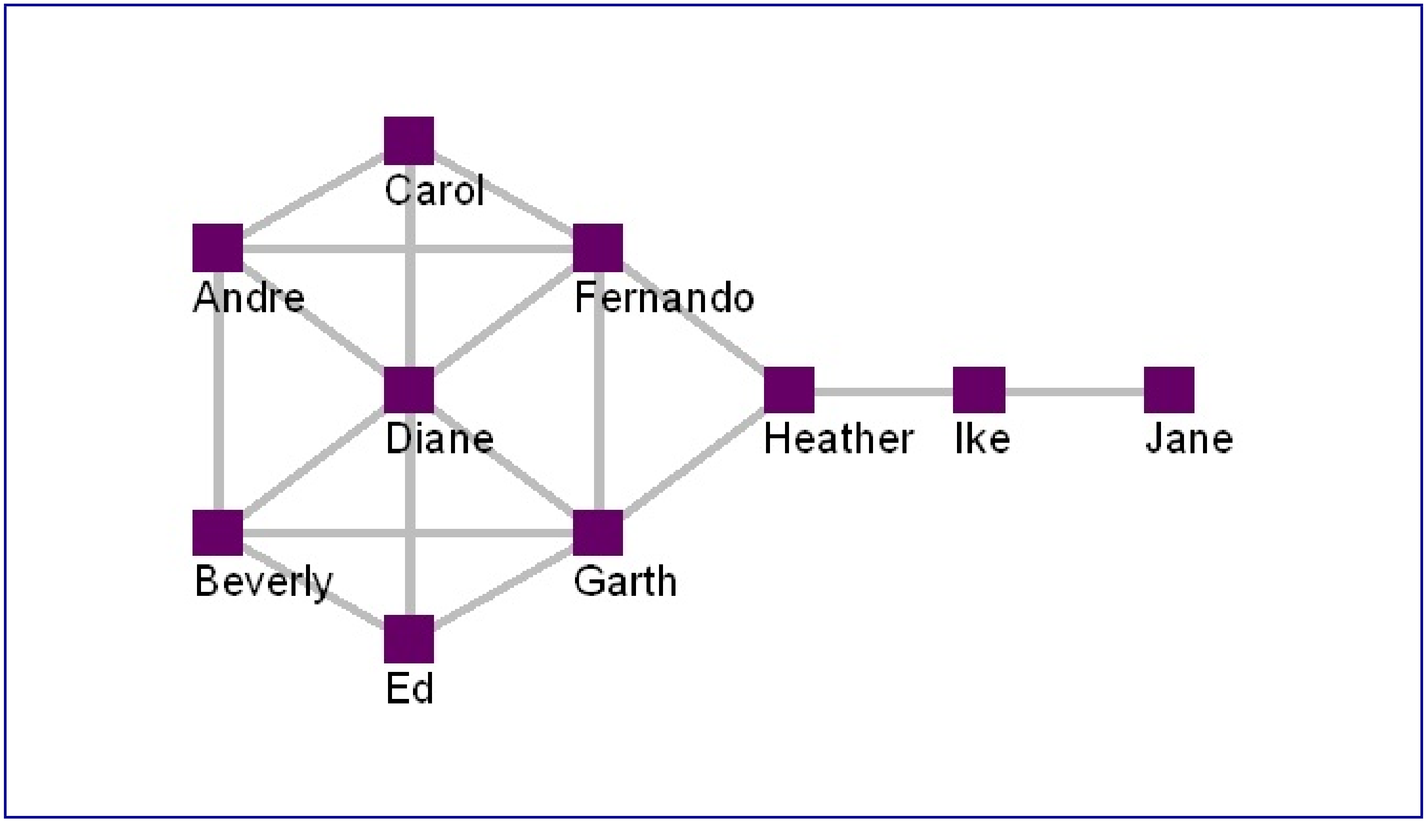
What other types of nodes play key roles?

Other types of important nodes

A classic example from Social Network Analysis (SNA)

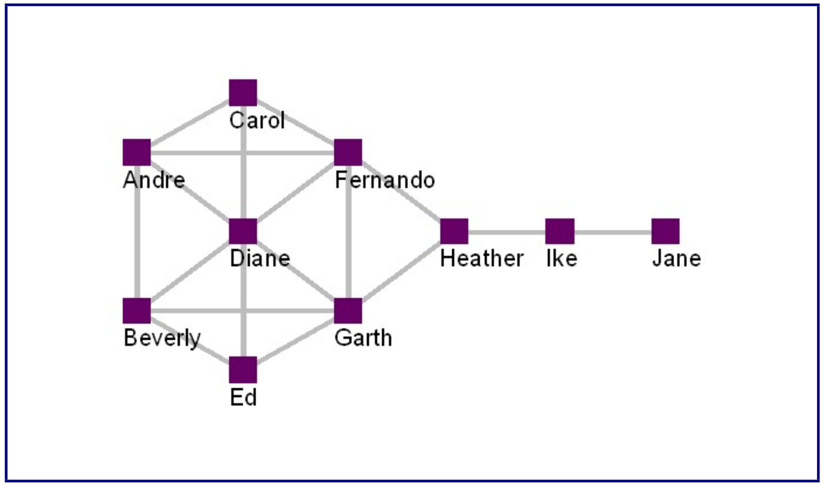
[http://www.fsu.edu/∼spap/water/network/intro.htm]

The "Kite Network"



Who is important and why?

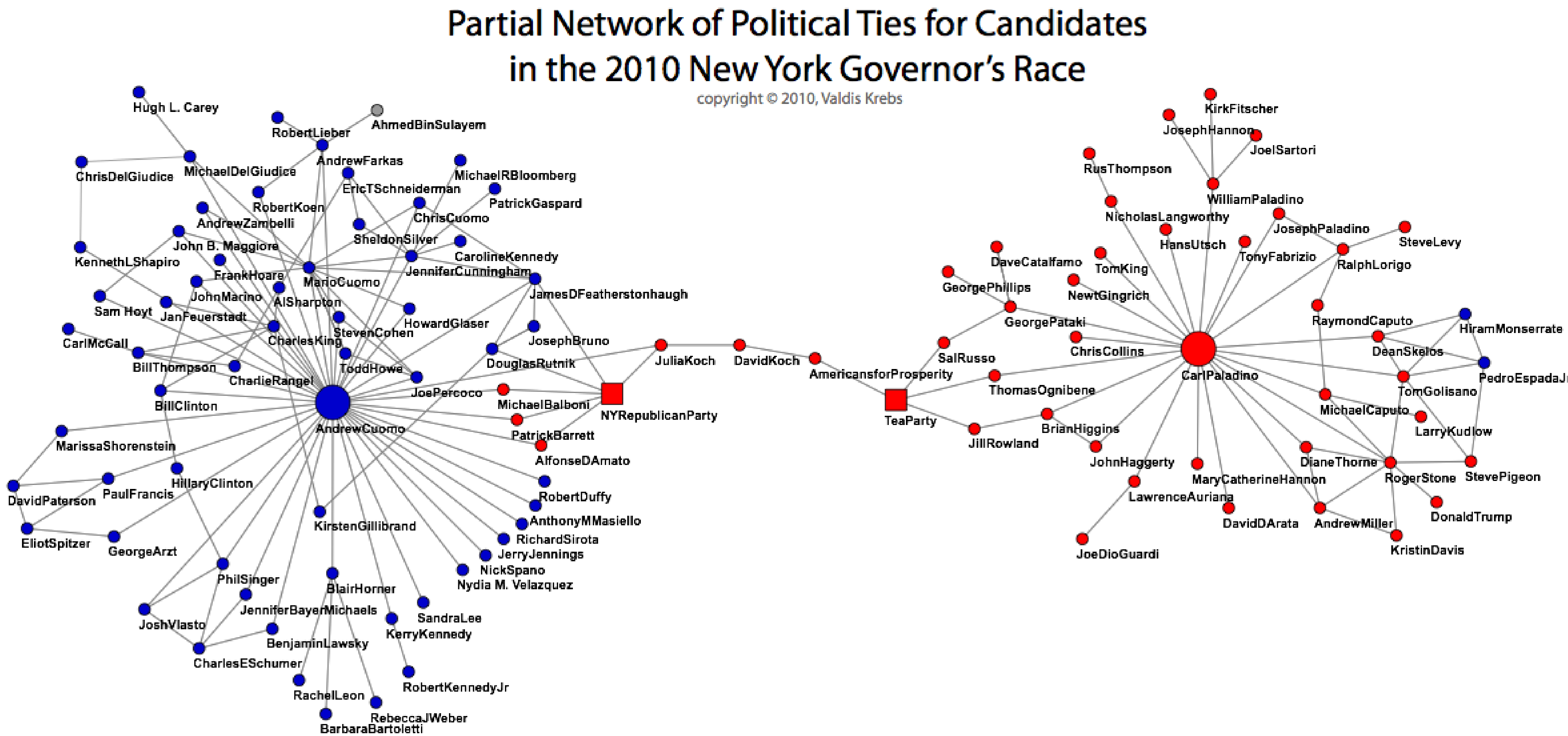
The Kite Network



- Degree – Diane looks important (a "hub").
- Betweenness – Heather looks important (a "connector"/"broker").
- Closeness – Fernando and Garth can access anyone via a short path.
- Boundary spanners – as Fernando, Garth, and Heather are well-positioned to be "innovators".
- Peripheral Players – Ike and Jane may be an important resources for fresh information.

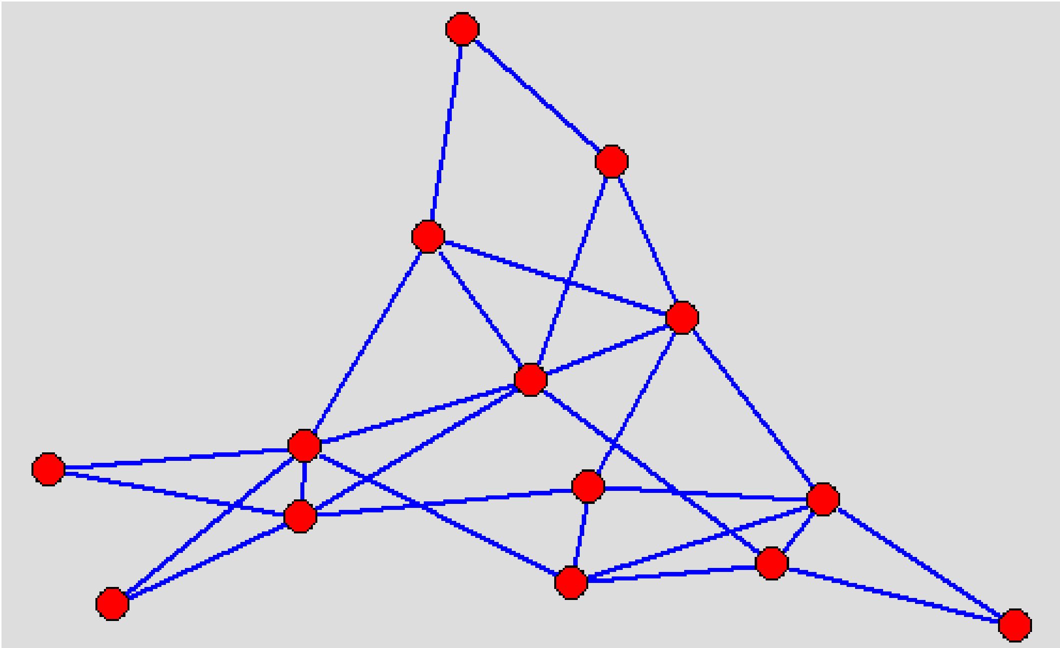
A contemporary social network

(Taken from http://www.thenetworkthinkers.com/)



Betweenness Centrality

[Freeman, L. C. "A set of measures of centrality based on betweenness." *Sociometry* **40** 1977]



A measure of how many shortest paths between all other vertices pass through a given vertex.

Betweenness (formal definition)

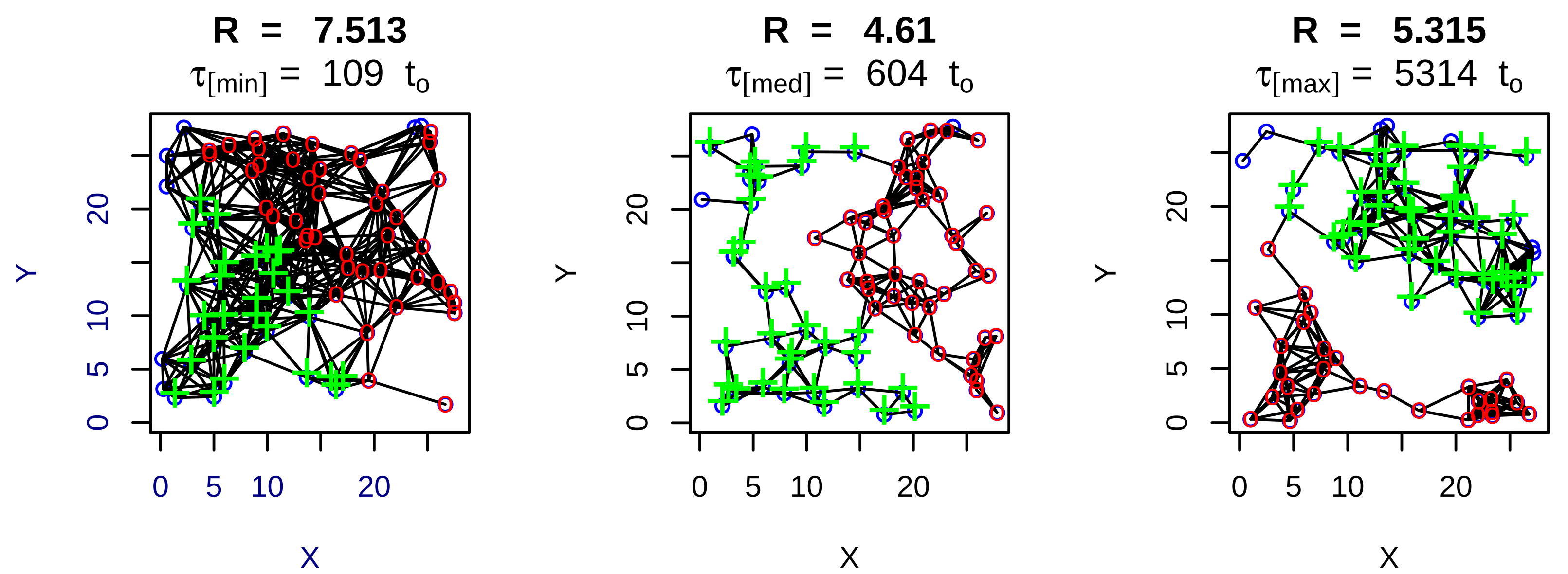
For a given vertex i:

$$\boxed{B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}}$$

- Where σ_{st} is the number of shortest geodesic paths between s and t.
- And $\sigma_{st}(i)$ are the number of those passing through vertex i.

(Calculating shortest paths efficiently ... http://en.wikipedia.org/wiki/Dijkstra's_algorithm)

Betweenness and eigenvalues (bottlenecks)



- Bottlenecks have large betweenness values.
- In social networks betweenness is a measure of a nodes "centrality" and importance (could be a proxy for influence).
- In a road network, high betweenness could indicate where alternate routes are needed.
- Also a measure of the resilience of a network (next page).

Targeted attack by different metrics

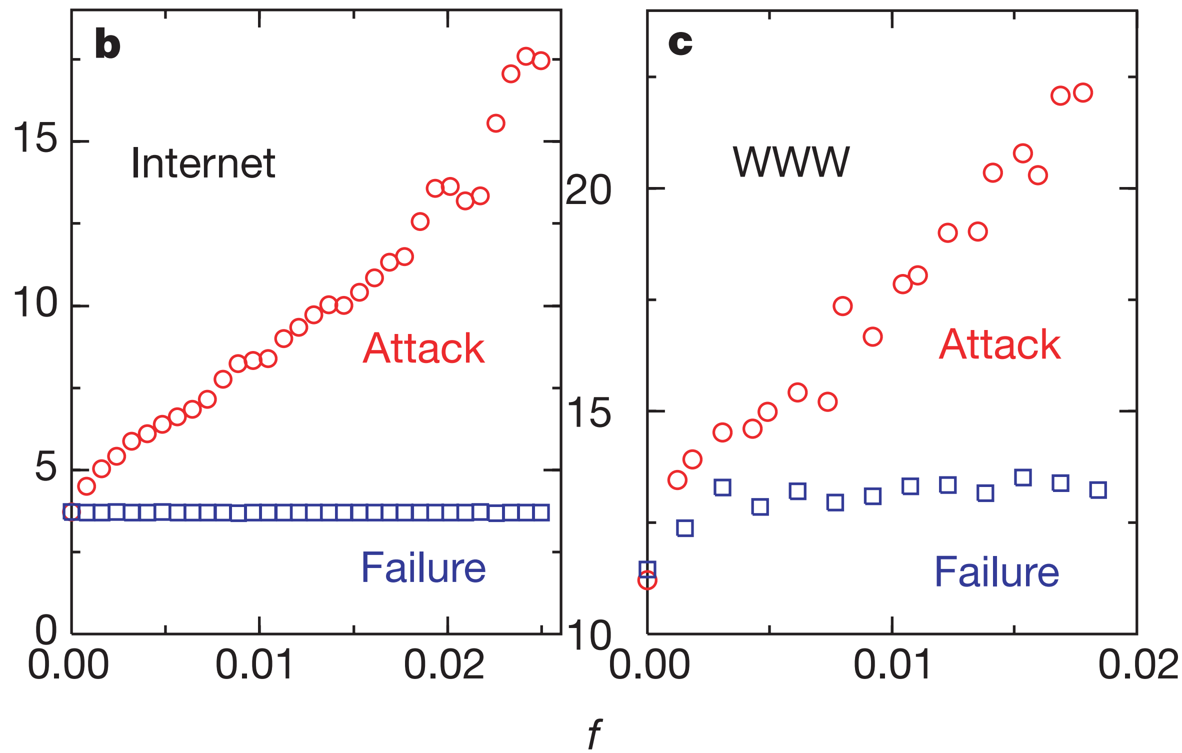
Holme P, Kim BJ, Yoon CN, Han SK (2002) "Attack vulnerability of complex networks". *Phys. Rev. E* **65**:056109

- Degree centrality
- Betweeness centrality

Typically (but not always) high degree are high betweeness.

High betweeness the more effective strategy to break up a network's connectivity.

But back to Albert, Jeong and Barabasi



So why did Albert, Jeong and Barabasi find that their sample of the internet topology was vulnerable to degree targeted attack?

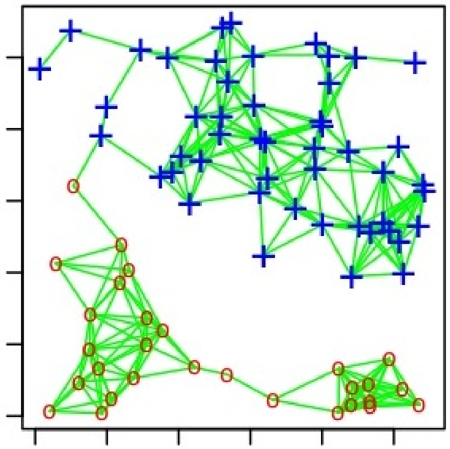
How to measure the structure of the Internet?

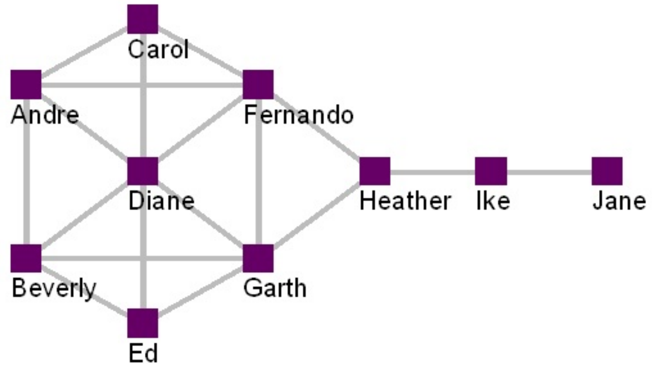
The focus of the next lecture (Lecture 5)

Summary

- **"Error and attack tolerance of complex networks"**
  Random networks with power law degree distribution show:
  – Fragility to degree-targeted removal
  – Robustness to random node removal
  (This is in the context of keeping the full network connected.)

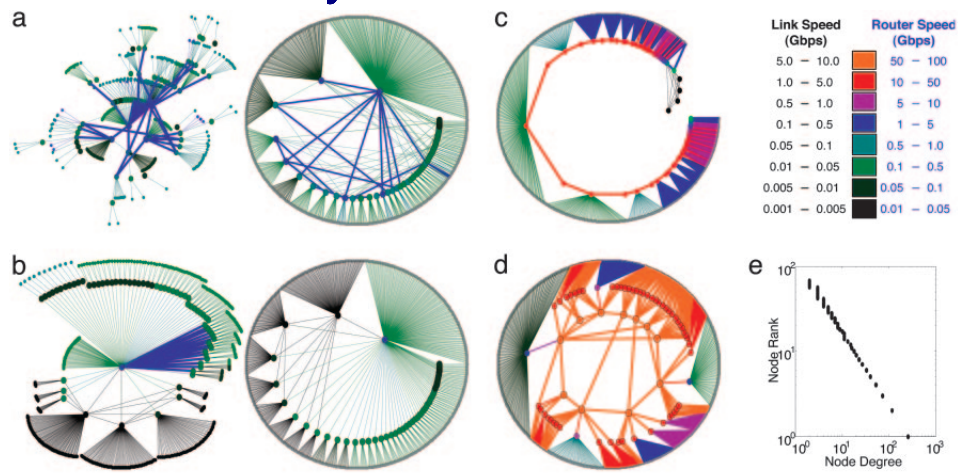
- **Important nodes beyond degree**
  – Betweeness centrality (shortest paths)
  (Are their local ways to detect this?)
  – Boundary spanners / peripheral players / weak-ties





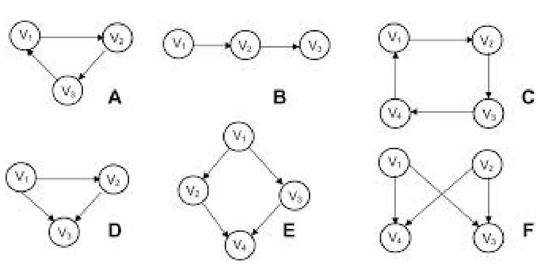
Structure beyond degree distribution

- Power law degree distribution actually a weak constraint on network structure:

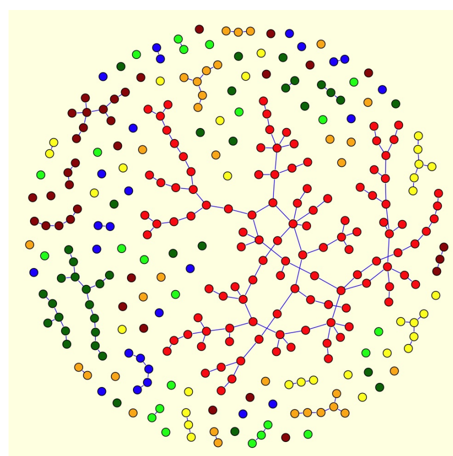


- Additional properties include:

Motifs



Components



Communities