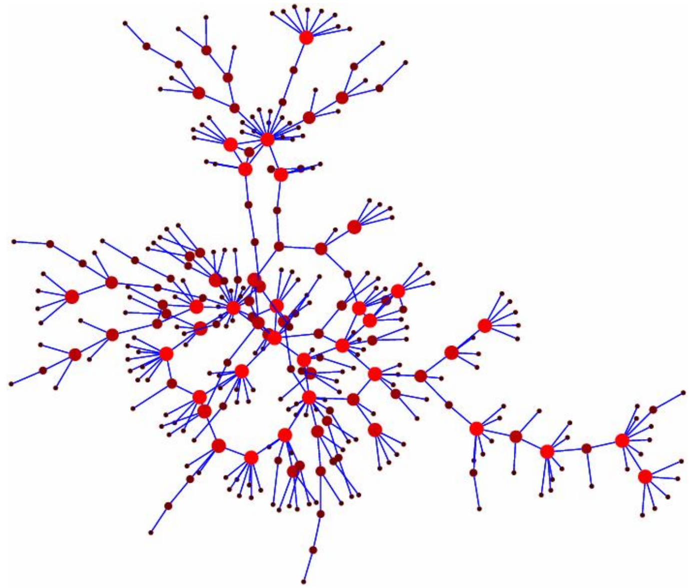
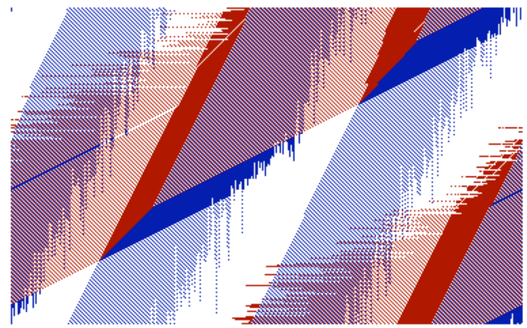
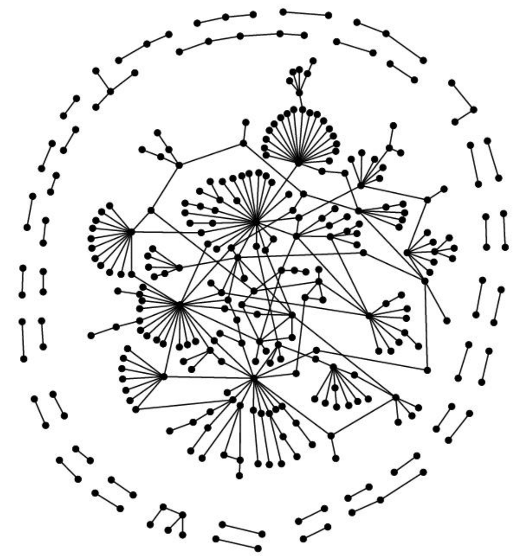
“Understanding networks: Topology, Activity, Phase transitions”

Raissa D’Souza *UC Davis*

Dept of Mechanical and Aeronautical Eng.

Center for Computational Science and Eng.

Raissa's Professional history: i.e., (How did I get here?)

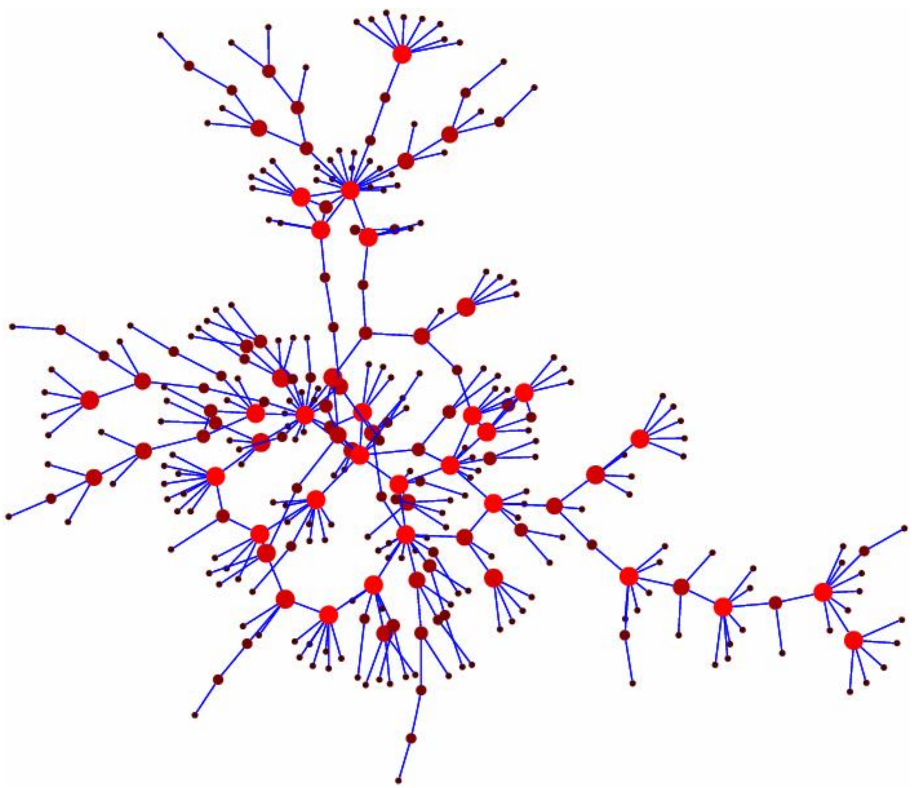
- 1999, PhD, Physics, Massachusetts Inst of Tech (MIT):
  – Joint appointment: Statistical Physics and Lab for Computer Science
- 2000-2002, Postdoctoral Research Fellow, Bell Laboratories:
  – Joint appointment: Fundamental Mathematics and Theoretical Physics Research Groups.
- 2002-2005, Postdoctoral Research Fellow, Microsoft Research:
  – "Theory Group" (Interdisciplinary group in Physics and Theoretical Computer Science)
- 2005-present, Assistant Professor, UC Davis:
  – Dept of Mechanical and Aeronautical Eng., and Center for Computational Science and Eng.

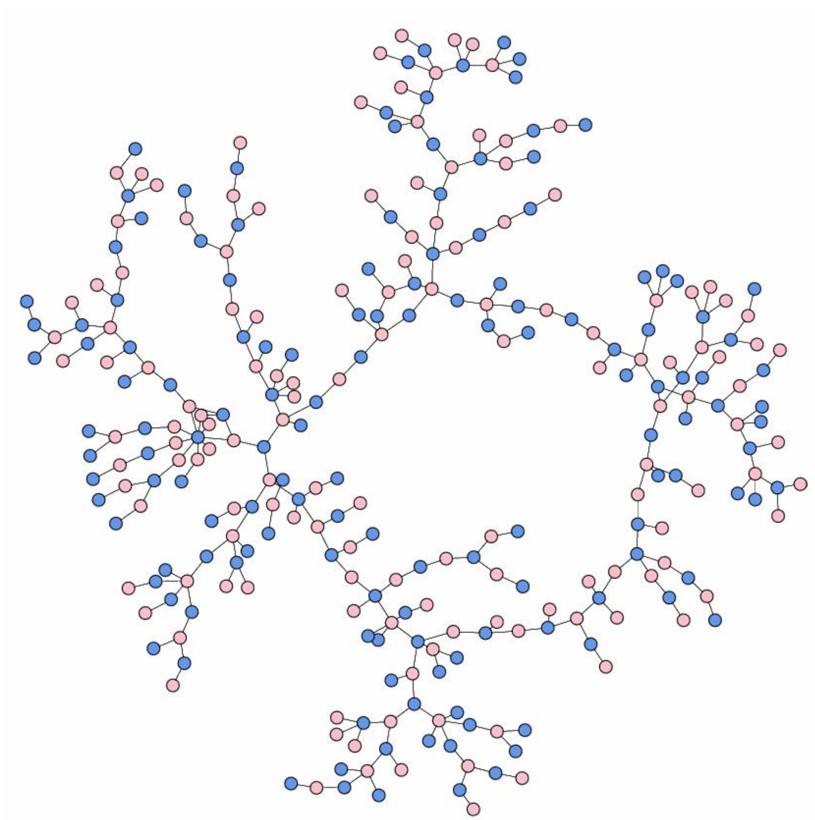
1996 CSSS in Santa Fe with Prof. Dave Feldman!

This week's focus: Networks

- Topology (i.e., structure)
- Activity (i.e., function)
- Phase transitions

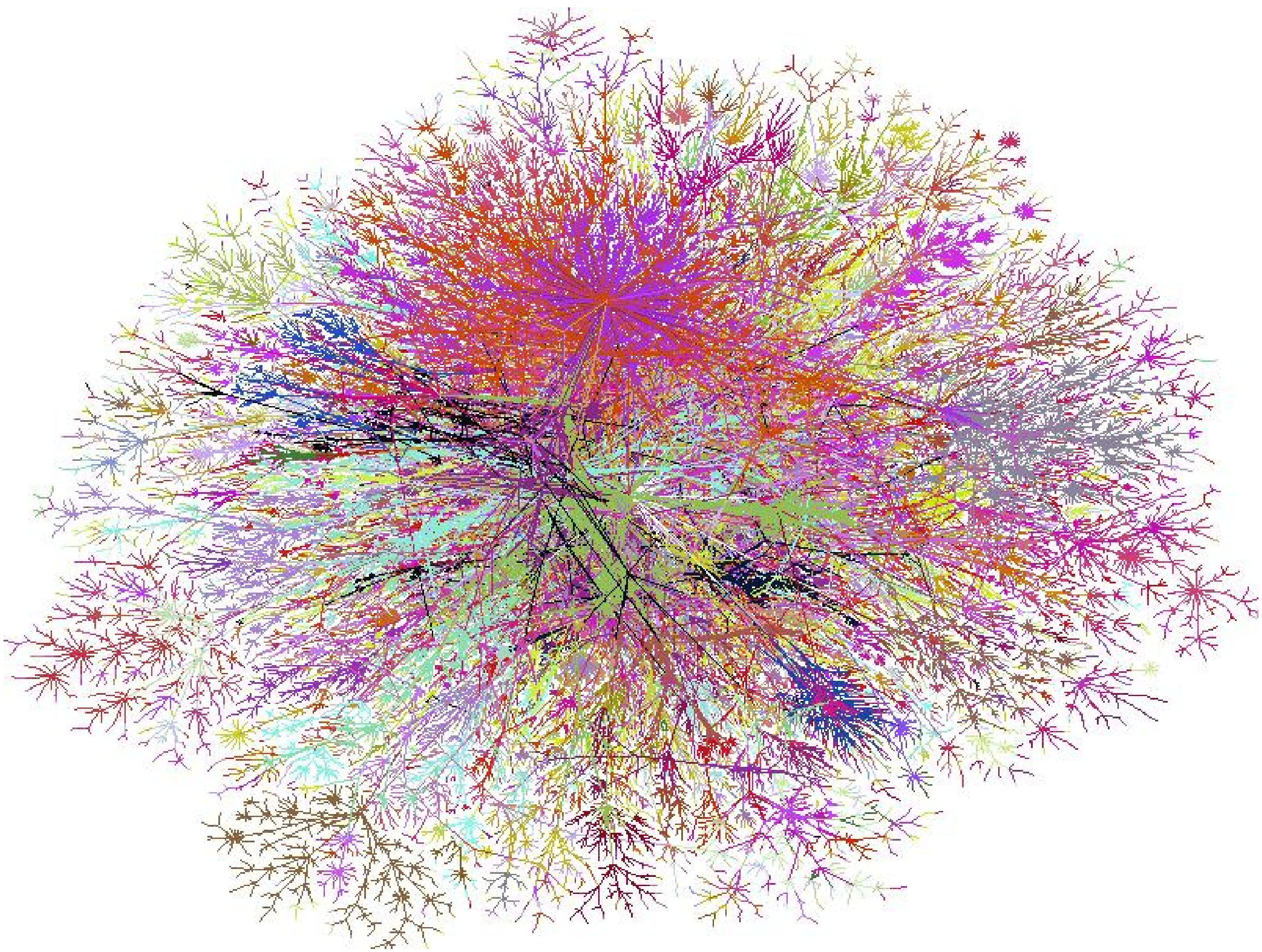
Example social networks
(Immunology; viral marketing)





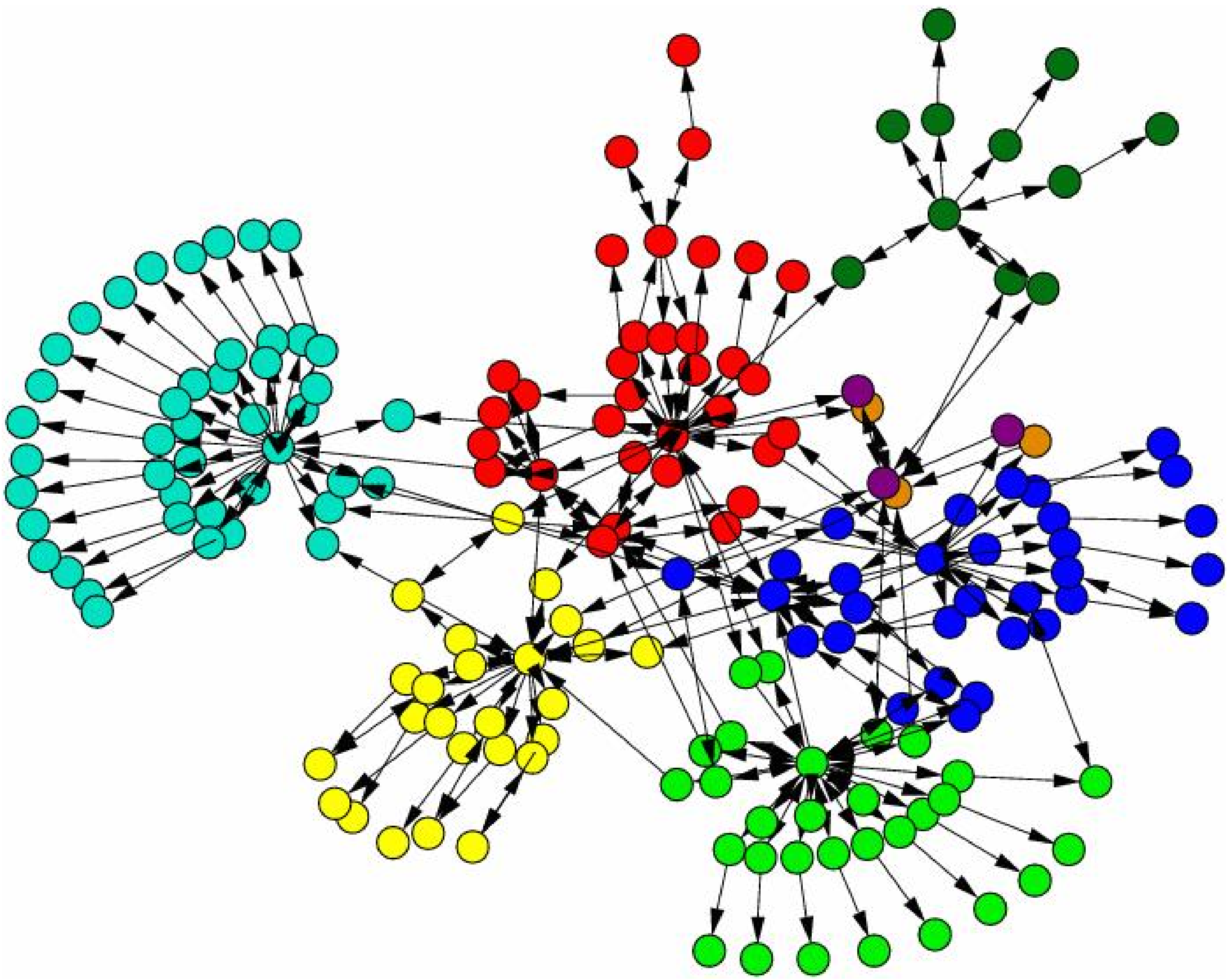
M. E. J. Newman

The Internet
(Robustness to failure; optimizing future growth; testing protocols on sample topologies)



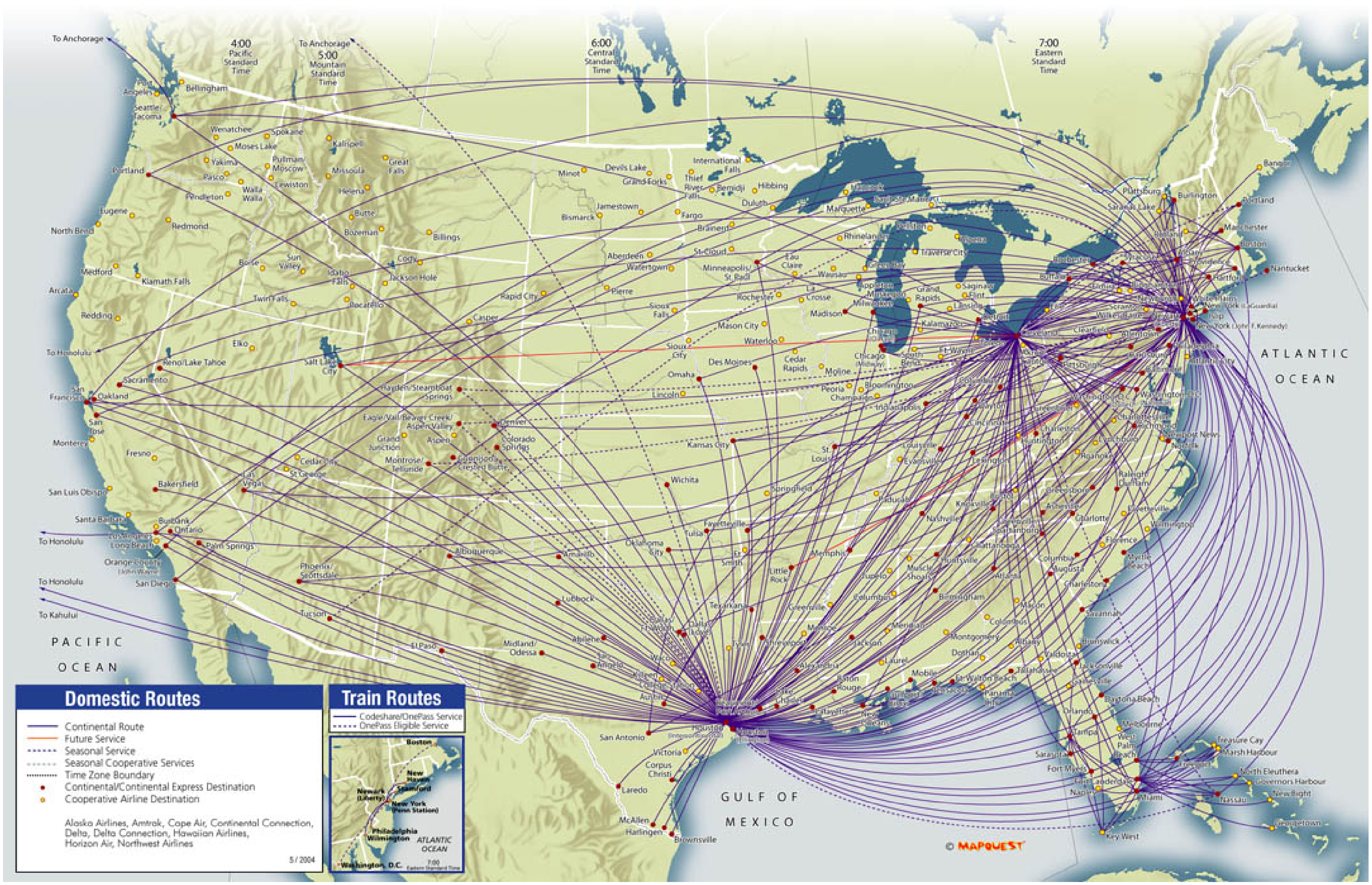
H. Burch and B. Cheswick

A typical web domain
(Web search/organization and growth)



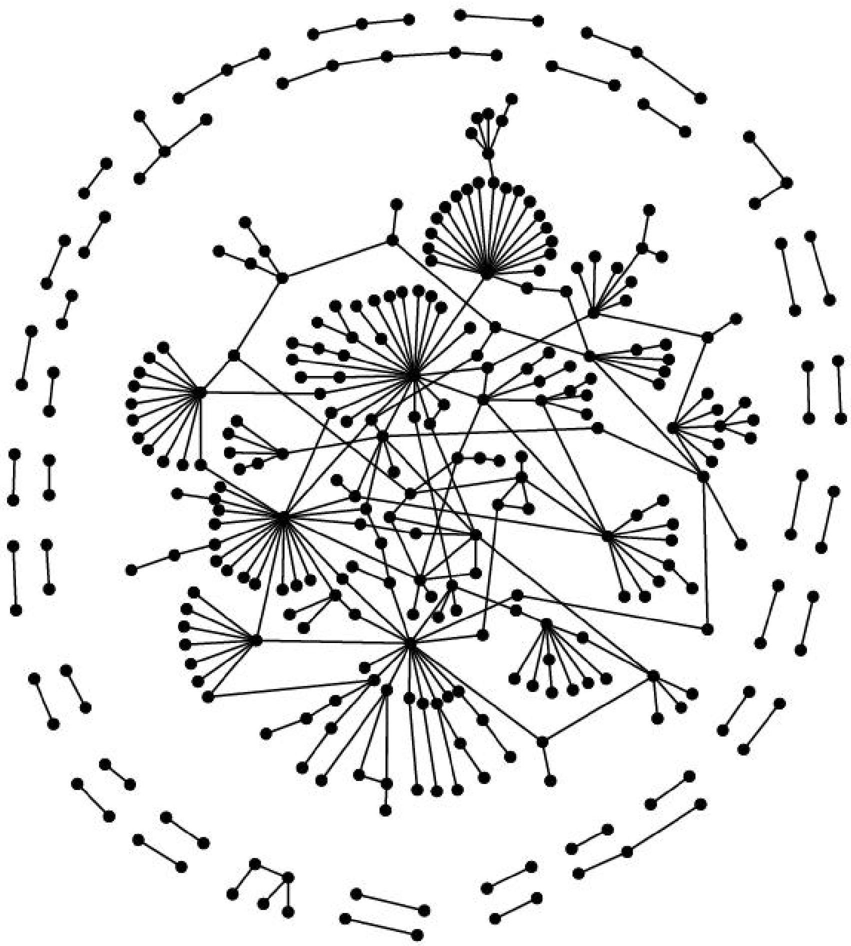
M. E. J. Newman

The airline network
(Optimization; dynamic external demands)



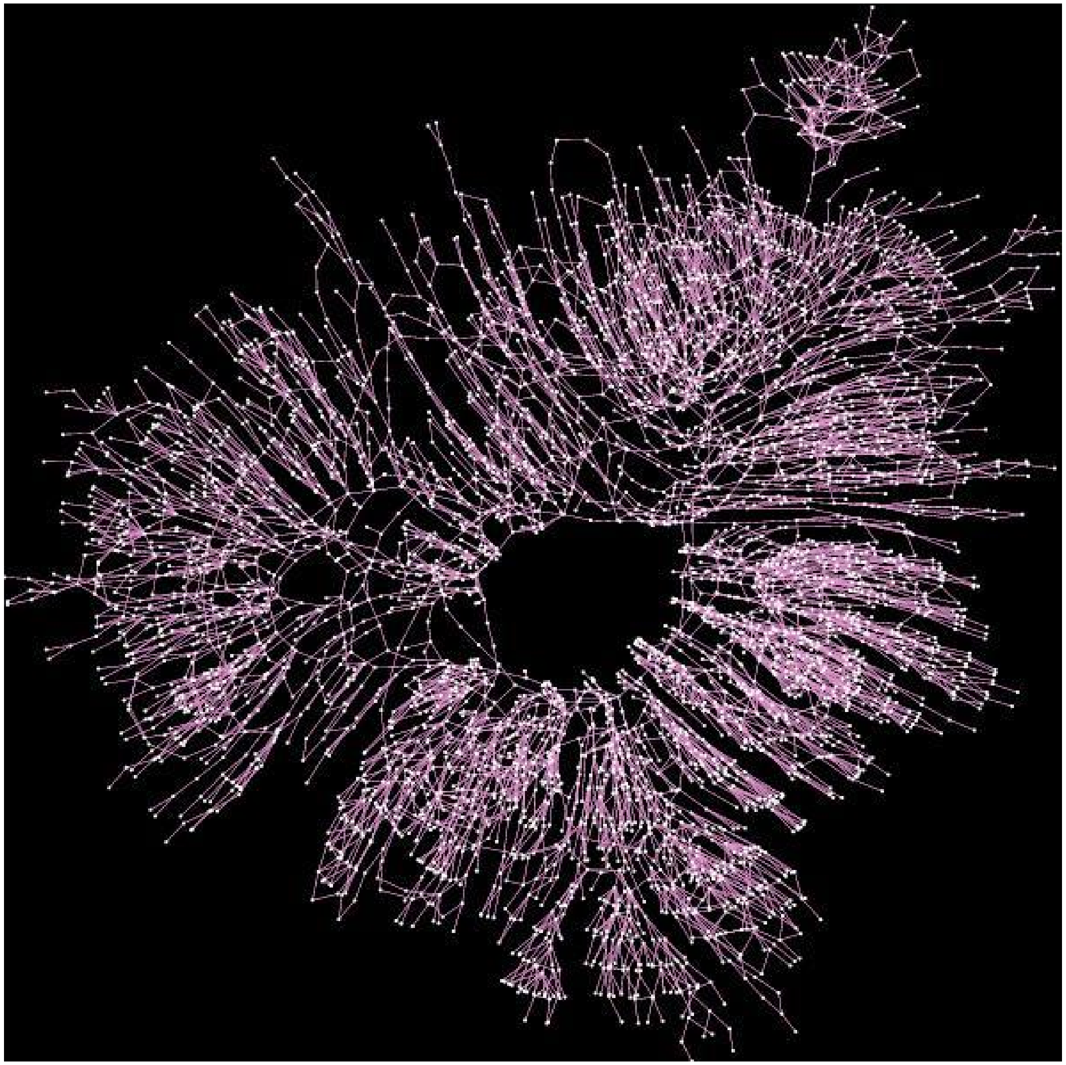
Continental Airlines

Yeast protein signaling network (Control mechanisms in biology)



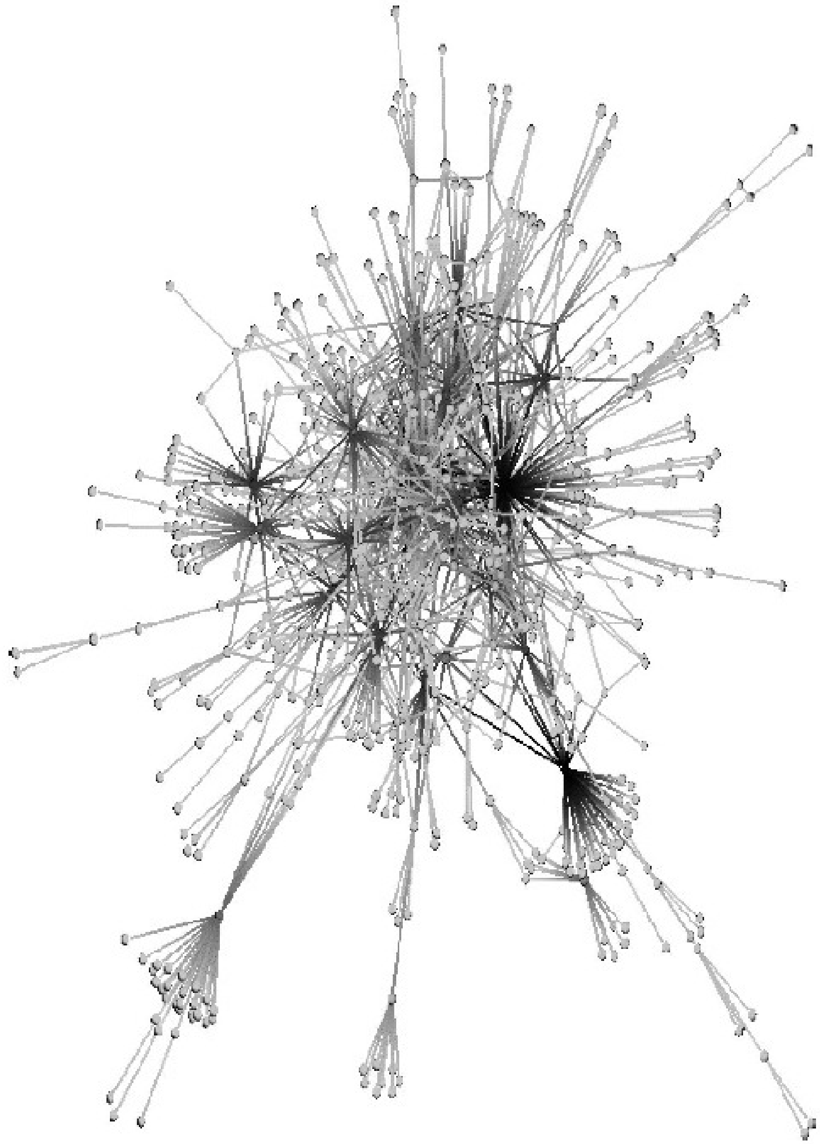
S. Masloc and K. Sneppen; M. E. J. Newman

The power grid
(Mitigating failure; Distributed sources)



M. E. J. Newman

Software call graph
(Uncovering design principles/robustness to mutation)



Chris Myers

Networks: basic properties

- Network made of nodes connected together by edges.

- Edges can be directed or undirected (i.e., one-way or two-way connections).

  Example one-way: Web pages.

  Example two-way: Family tree (relatives).

  Example hybrid: Road networks with some one-way and some two-way links (city of Boston prime example!).

- Geometric versus geometry free (e.g., Internet vs WWW)

- STRUCTURE (topology) and FUNCTION (information flow/ dynamics on the network)

Why do networks exist?
Physical, Biological, Social, Engineered

- More efficient control, esp through hierarchy?
- Robustness to noise and fluctuations?
- Can we learn function from structure?
- Can we apply these lessons to engineered systems?

  $\longrightarrow$ Would a modern power grid look like the one we have?

Lessons from existing networks?

All networks, all quantitatively different, each optimizes something different.

What are the key parameters that distinguish them?

Which kinds work best for which application?

General Considerations/Tradeoffs

- For what purpose are we building the network?

- CONNECTIVITY?
  - Preserve at all costs (Internet),
  - Or break at all costs (Immunology)?

- ROBUSTNESS to which failure modes?
  - Random failure (biology),
  - Or targeted attacks (technological).

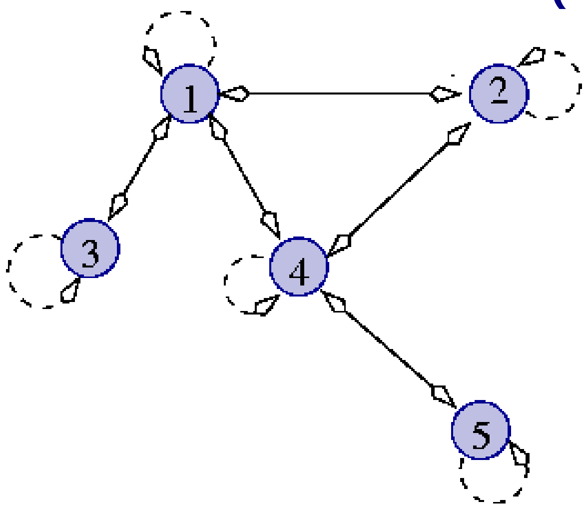
- Fully decentralized or some centralized control?

How do we represent a network as a mathematical object?

Matrix representation of a network: TOPOLOGY

Connectivity matrix, M:

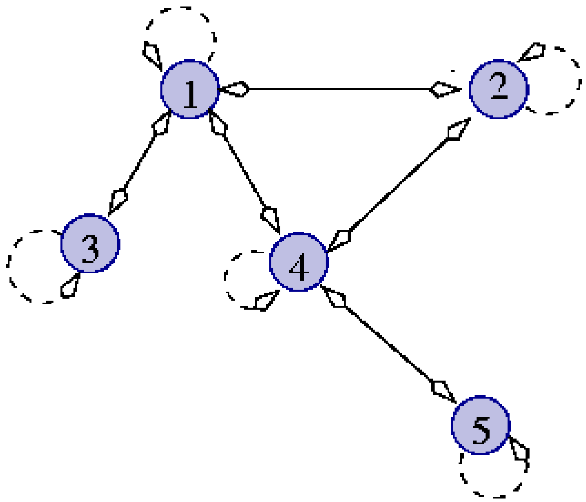
$$M_{ij} = \begin{cases} 1 \text{ if edge exists between } \; i \text{ and } j \\ 0 \text{ otherwise.} \end{cases}$$



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

The *degree* of a node, is how many links it has.

TOPOLOGY: Common measures of fine structure



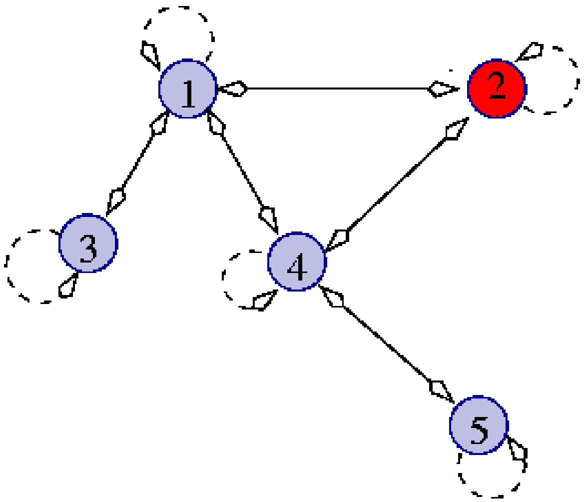
$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} = M$$

- Degree distribution (fraction of nodes with degree k)
- Clustering coefficient
- Diameter
- Betweenness centrality
- Assortative/dissortative mixing

Matrix representation of a network: ACTIVITY

(Spread of disease, routing of data, gossip spread/marketing)

Consider a random walk on the network. The state transition matrix, P:



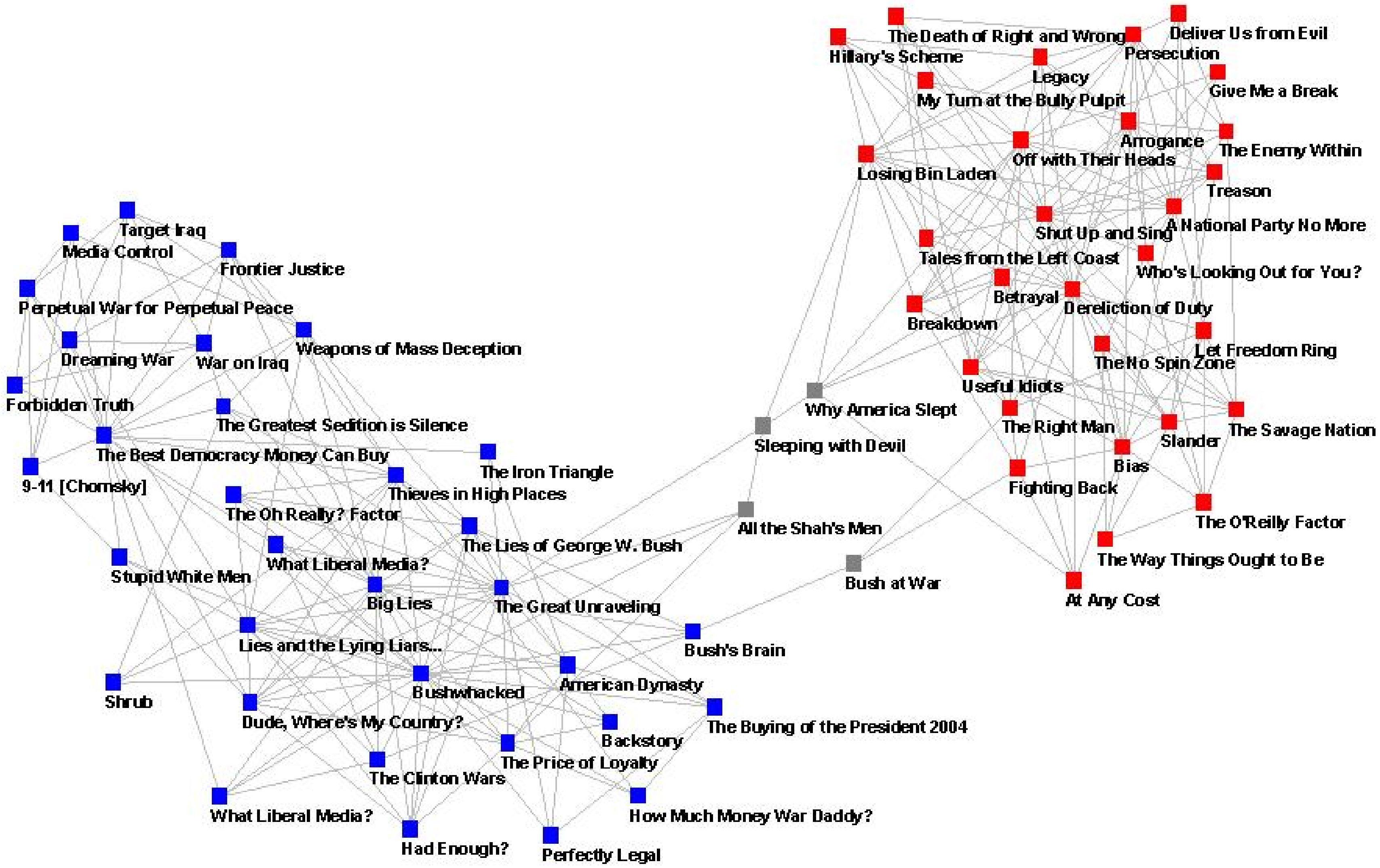
$$\begin{pmatrix} 1/4 & 1/3 & 1/2 & 1/4 & 0 \\ 1/4 & 1/3 & 0 & 1/4 & 0 \\ 1/4 & 0 & 1/2 & 0 & 0 \\ 1/4 & 1/3 & 0 & 1/4 & 1/2 \\ 0 & 0 & 0 & 1/4 & 1/2 \end{pmatrix} = P$$

The eigenvalues and eigenvectors convey much information.

Eigenvalues and eigenvectors of the state transition matrix

- The stationary distribution of a random walk on the graph:
  - “cover time” of a random walker
  - mixing time
  - occupancy probabilities
- Partitioning graphs into subgraphs/communities

Partitioning networks:
Community structure: Political Books USA, 2004



M. E. J. Newman

- Will focus on activity as characterized by state transition matrix (i.e., random walk on network) later.

- First, topology considerations...
  - Let's start with the classic ideas of Erdös and Rényi.

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).

Typical random graph: $\mathcal{G}(n, p)$

- $\mathcal{G}(n, p)$: 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 $[1 - p]$.)

$G(n,p)$

- We can 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).

Illustration of $G(n, p)$ generation process

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 $n \rightarrow \infty$. (A number of interesting properties of random graphs can be proven in this limit).

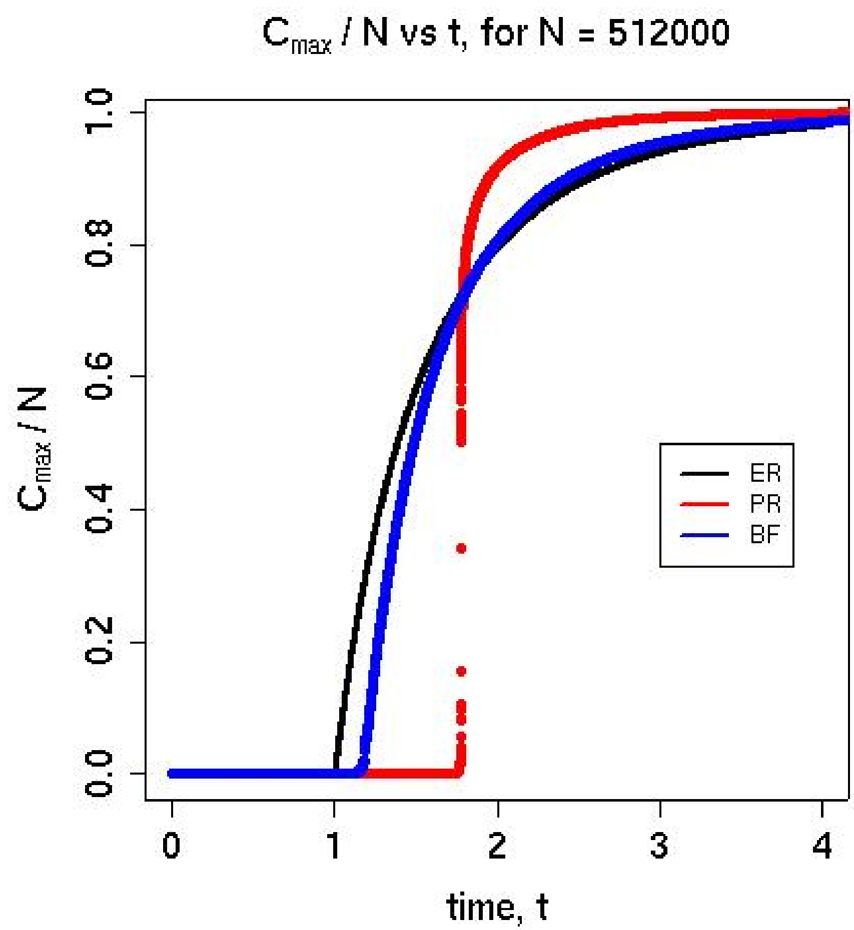
- Consider the size of the largest component of $G(n, p)$ as a function of p, $C_{max}(p)$.

- 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 the Giant Component

- For $p = 1/n$ (or equivalently $t = pE = E/n$), suddenly the largest component contains a finite fraction F of the total number of vertices, $C_{max} = Fn$, instead of a logarithmic fraction. All other components remain of size $O(\log n)$.



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

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:

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

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

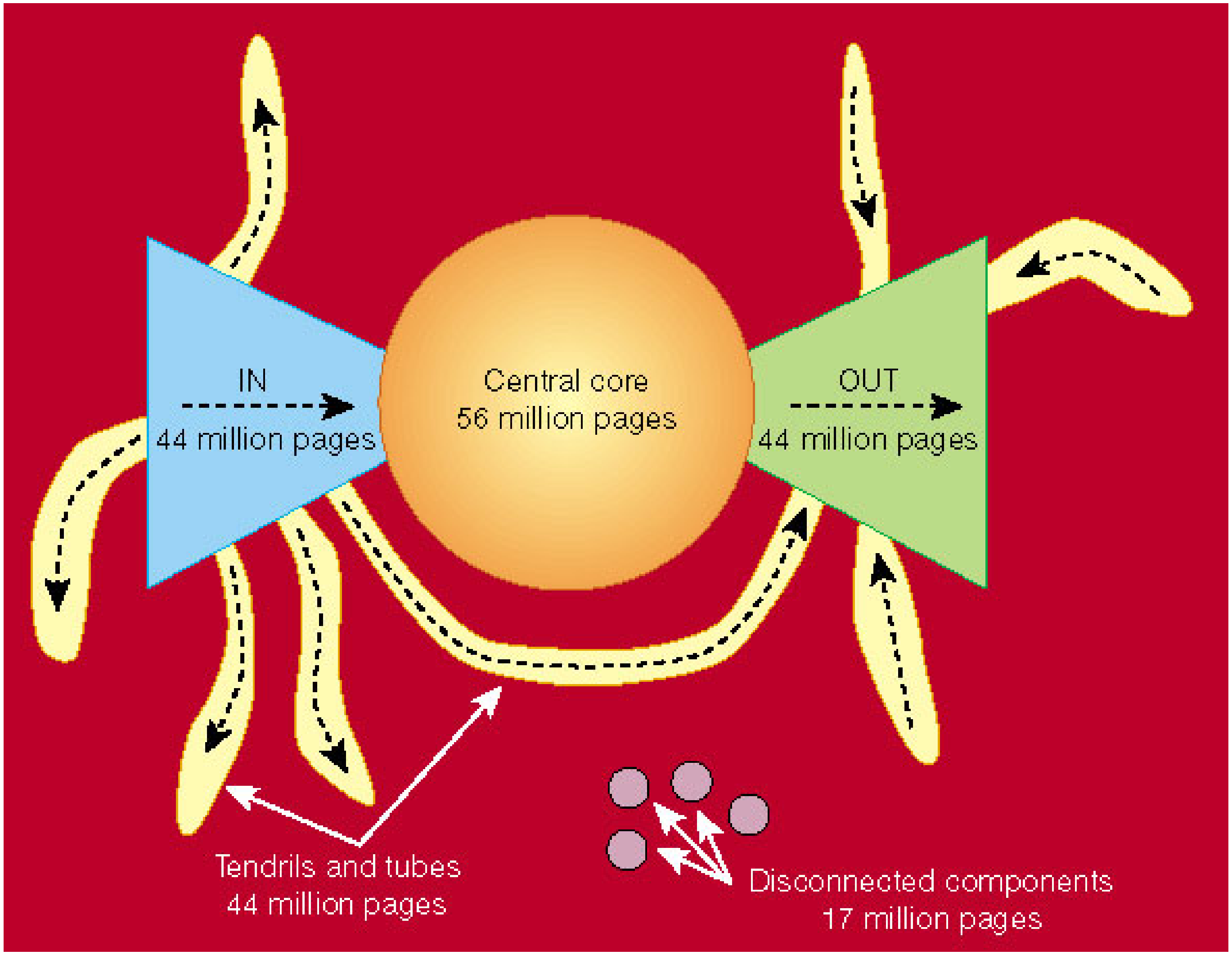
- At the phase transition, $z = np = 1$. The phase transition occurs when the average vertex degree is one!

Giant component observed in real-world networks

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

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

The giant component/Strongly Connected Component of the WWW



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

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 $n \to \infty$

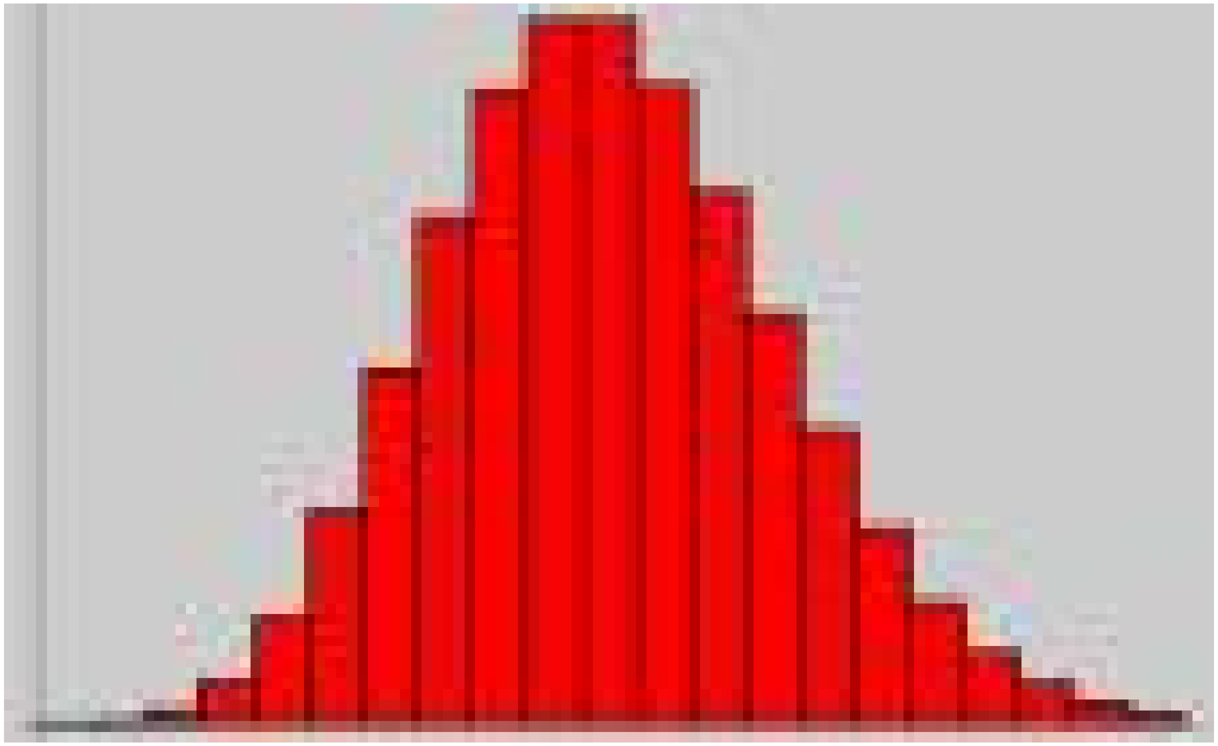
- Recall that $z = (n-1)p = np$ (for large n).

-

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

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

Much effort has gone into thinking about how to make a random graph have a degree distribution different from Poisson.

The configuration model (1970's)

- 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.

Summary: Terms introduced today

- Graph/network (also nodes and edges)
- Connectivity matrix (M)
- State transition matrix (random walk on M)
- Component
- Phase transition
- Degree distribution
- Graph diameter

Networks

- Network structures are pervasive – physical, biological, social, engineered

- Networks are made of discrete nodes – hard to envision continuum description

- Nodes can live in geometric, or geometry free space (Internet vs WWW)

- Need to consider:
  Topology of network
  Activity on network

- Can we learn lessons from existing networks?