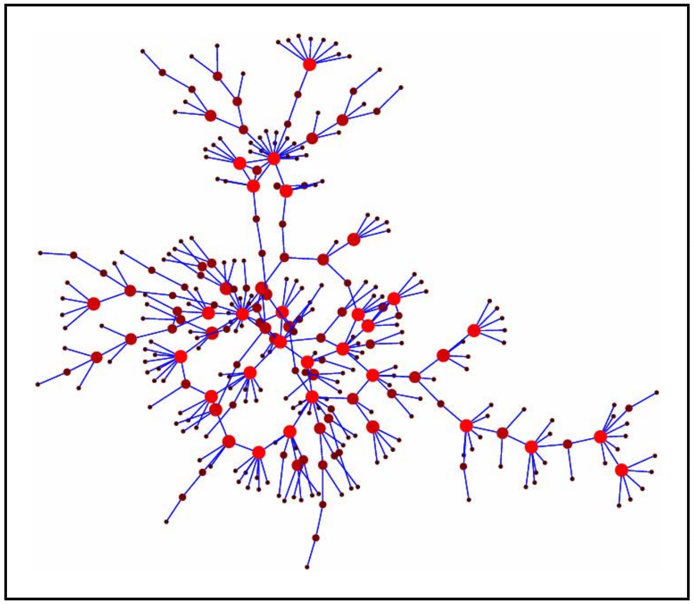
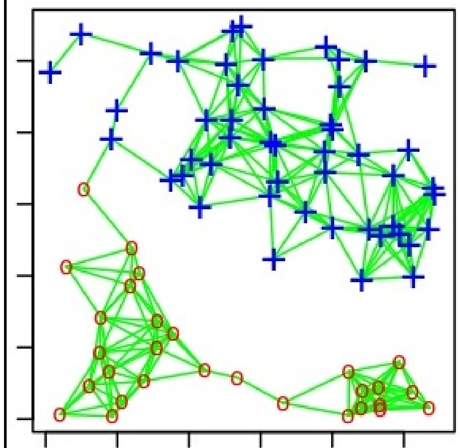
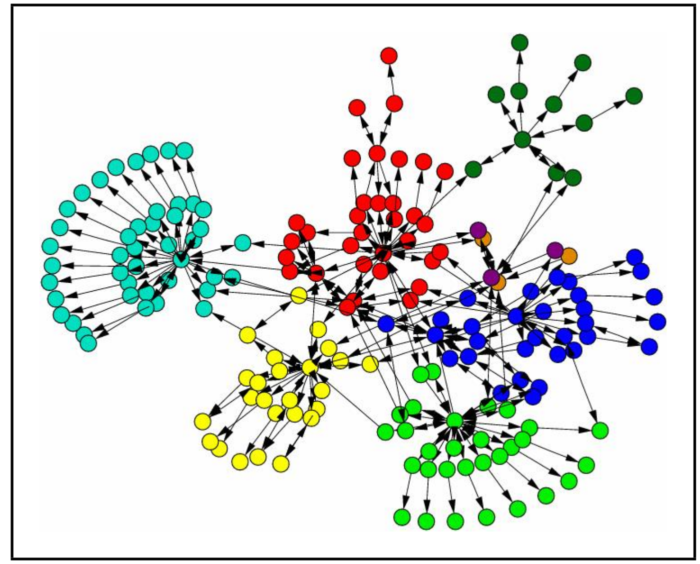
Statistical physics and network theory







Prof. Raissa D'Souza

University of California, Davis

Department of Computer Science

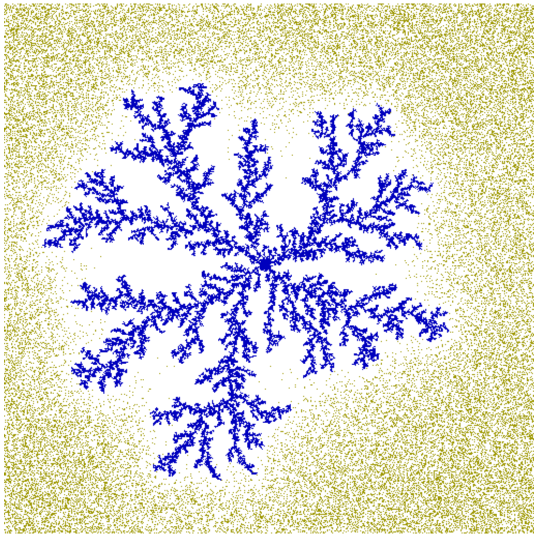
Mechanical and Aerospace Engineering

Graduate Group in Applied Math

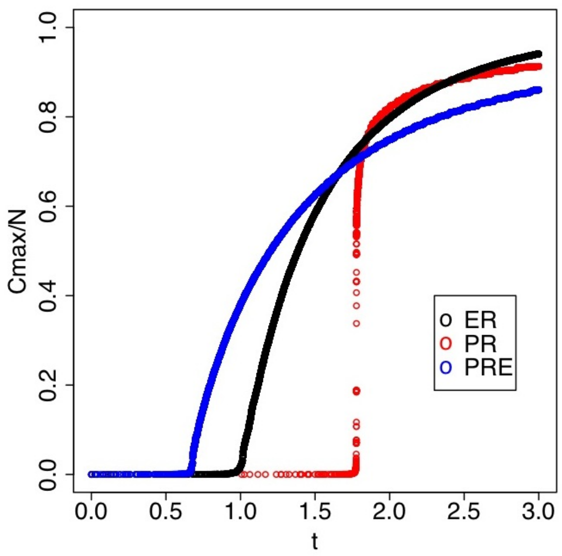
Raissa's Professional history: i.e., (How did I get here?)

- 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" (Physics and Theoretical Computer Science)
- Fall 2005-present, UC Davis:
  – Dept of Mechanical and Aeronautical Eng., Complexity Sciences Center, Grad Group Applied Math, Grad Group CS.
- 2007-present, External Faculty Member, Santa Fe Institute
- Fall 2009-present, UC Davis:
  – Dept of CS, Dept of Mech and Aero Eng., Complexity Sciences Center, Grad Group Applied Math.

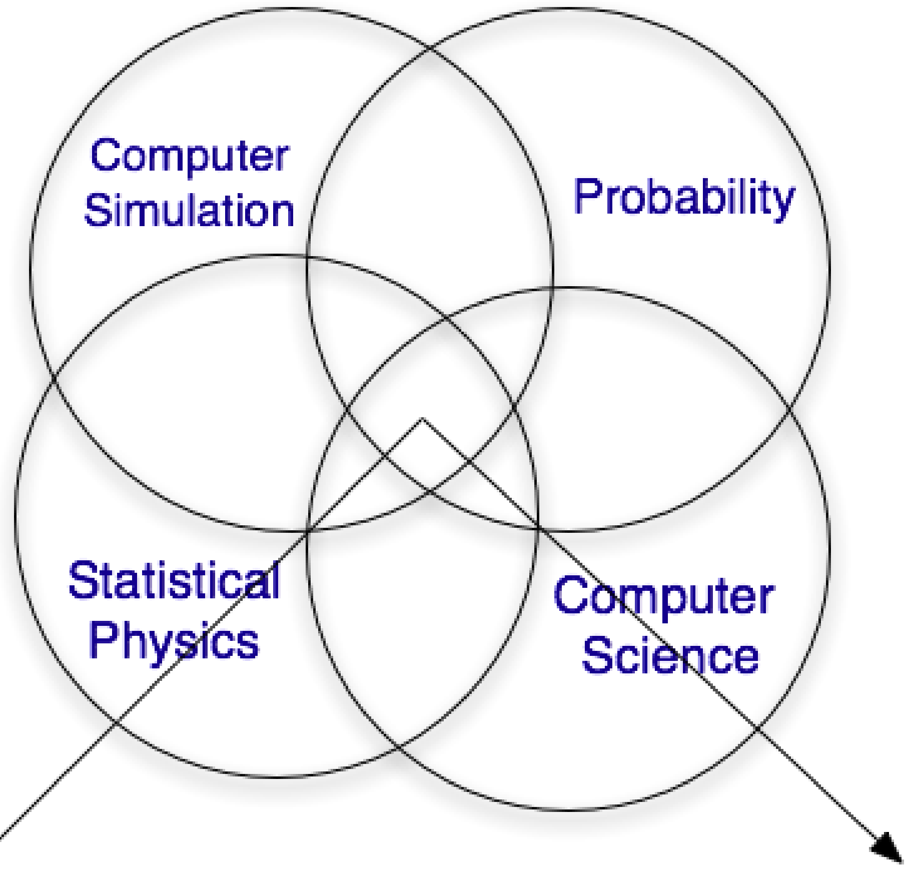
Complex systems



- Multiple length and time scales
- Irregular connectivity /
  Networked patterns of interaction
- Emergent behaviors
  (Collective behaviors not captured by equations of individual parts)
  - self organization
  - phase transtions



Fundamental connections



Lattice Models

- Ising
- Potts
- DLA
- BML

- Phase transitions
- Percolation
-Comp. Complexity

Network Models

- Boolean
- Circuit
- Random
- Bayesian

What is a Network?

- Topology (i.e., structure: nodes/vertices and edges/links)
  Measures of topology

- Activity (i.e., function, processes on networks, dynamics of nodes and edges)

Modeling networks

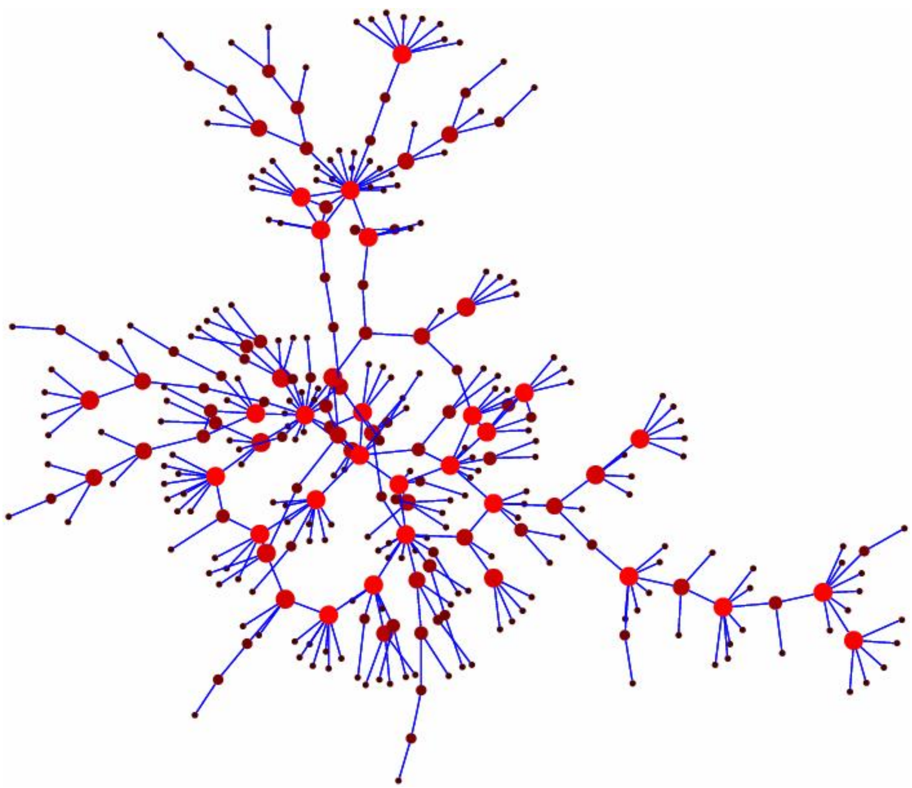
- Network growth

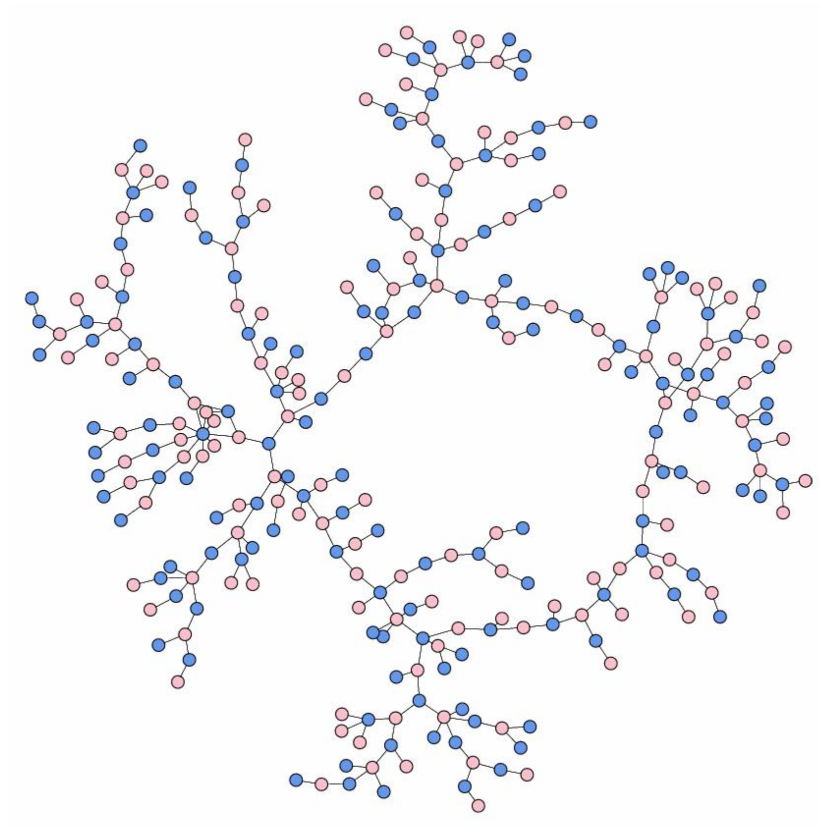
- Phase transitions

- Algorithms: analysis, growth/formation, searching and spreading

- Processes on networks

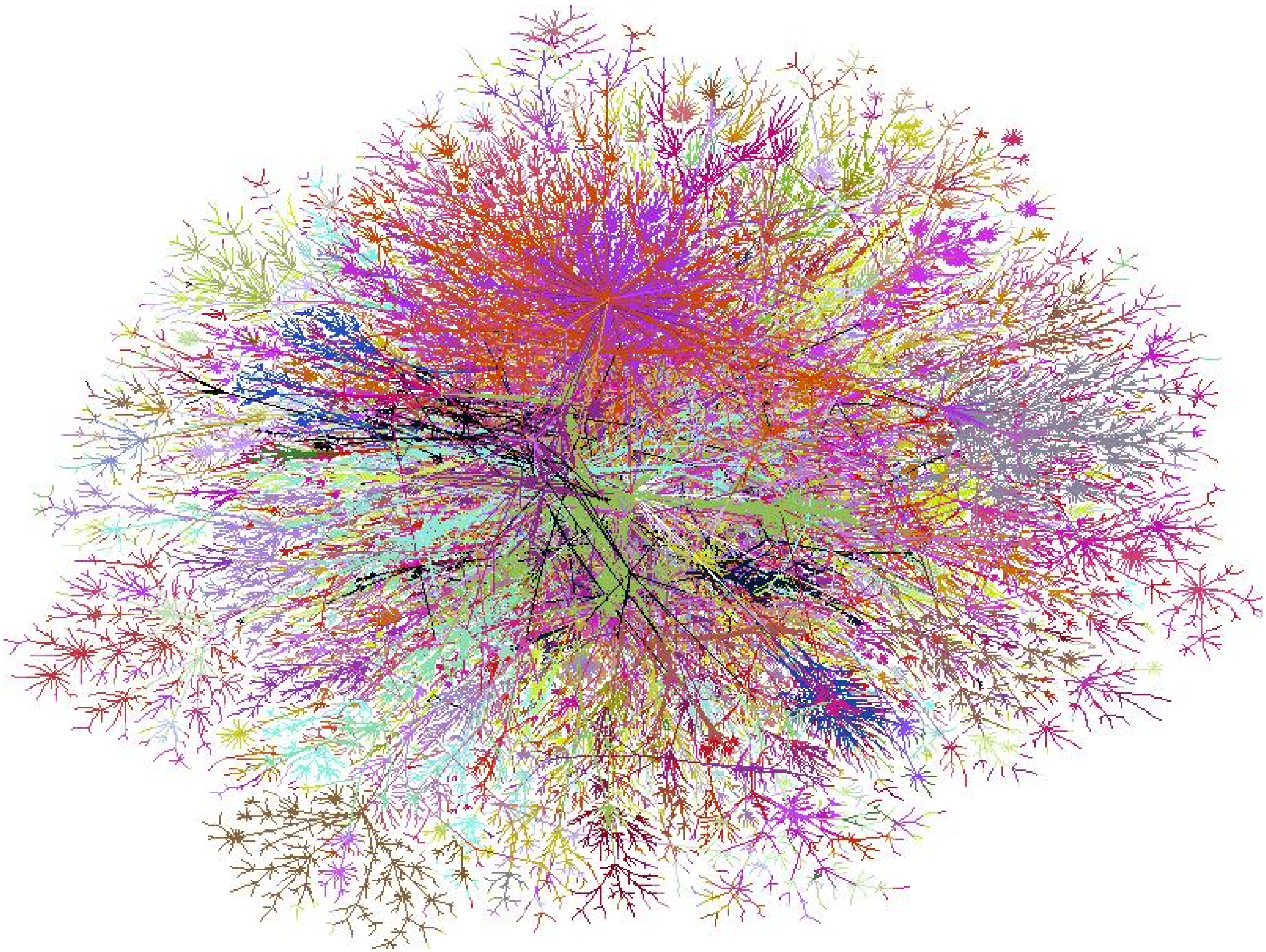
Example social networks
(Immunology; viral marketing; aliances/policy)





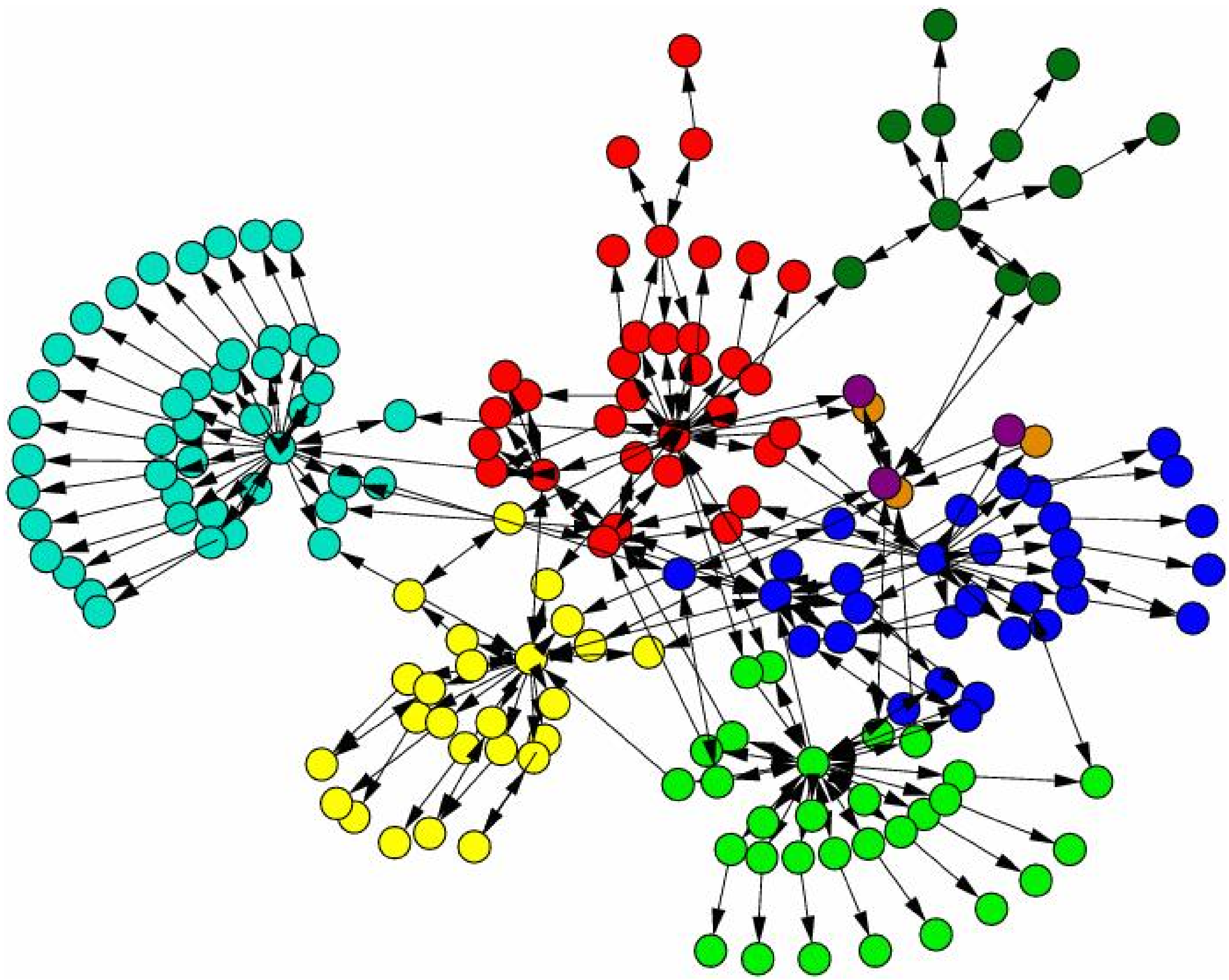
M. E. J. Newman

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



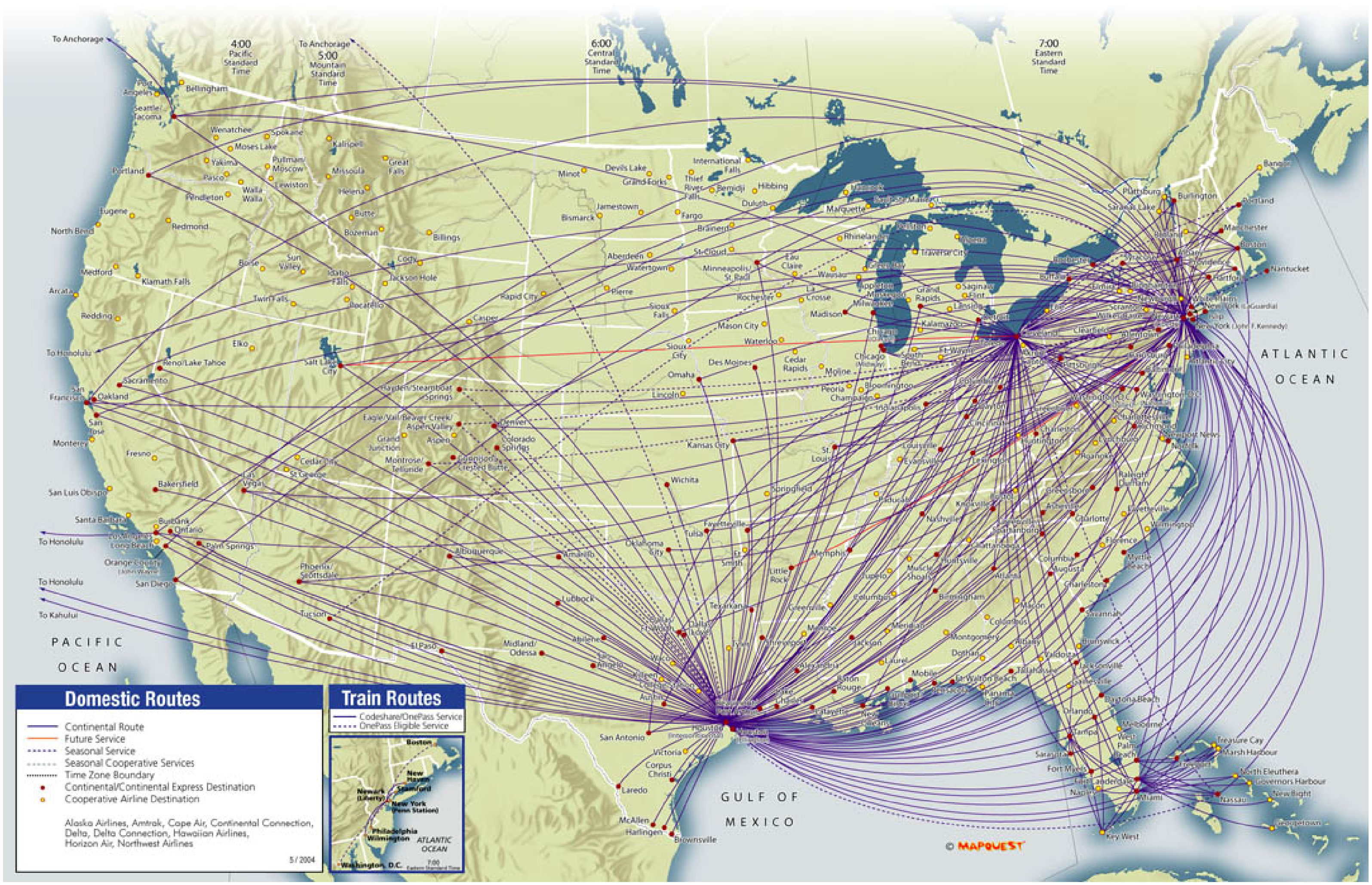
H. Burch and B. Cheswick

A typical web domain (Web search/organization and growth centralized vs. decentralized protocols)



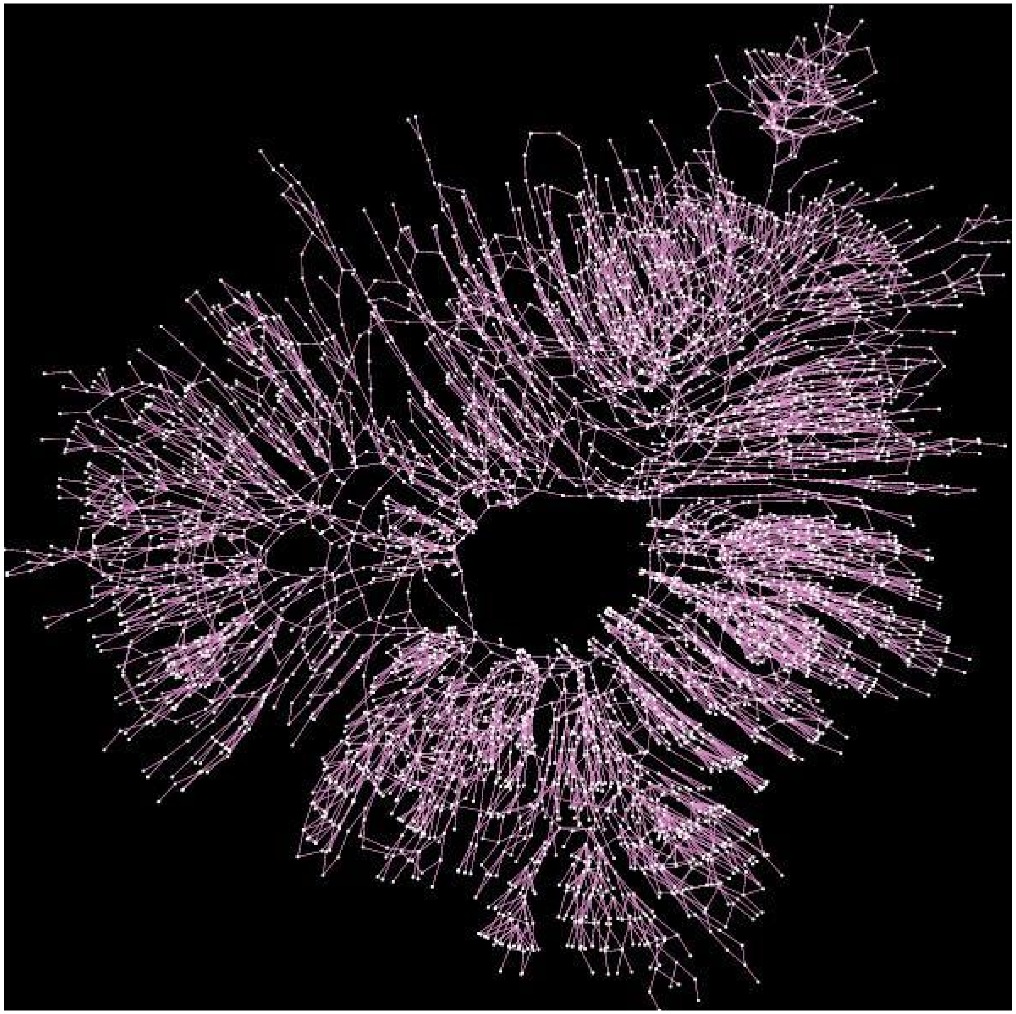
M. E. J. Newman

The airline network
(Optimization; dynamic external demands)



Continental Airlines

The power grid
(Mitigating failure; Distributed sources)

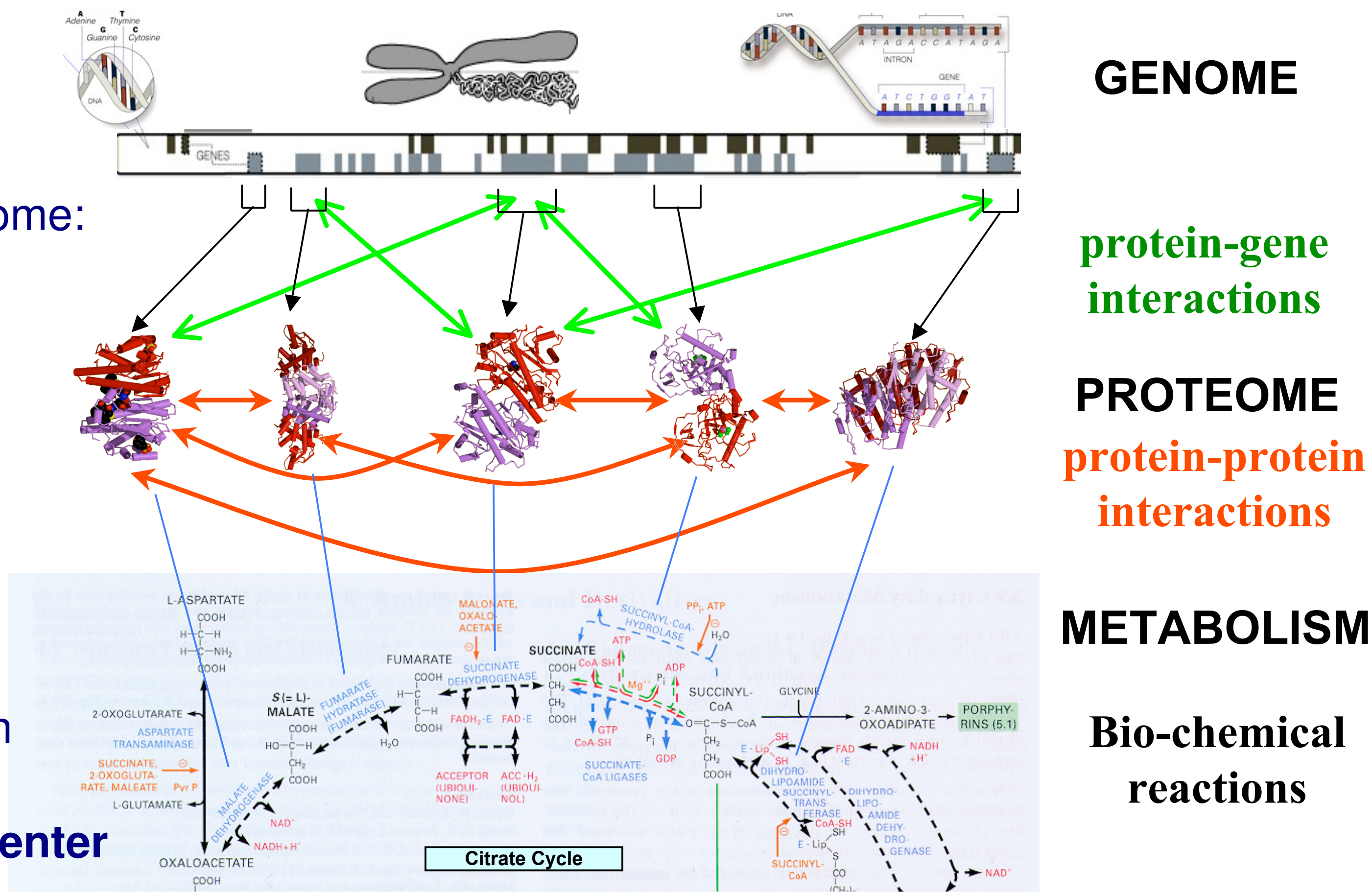


M. E. J. Newman

Biology: Networks at many levels

Control mechanisms / drug design/ gene therapy / biomarkers of disease

- Genome, Proteome:
  Dandekar Lab
- Metabolome:
  Fiehn Lab
- Data intergration
  BIOshare
  Lin, Genome Center



- Network structure / search for biomarkers:
  D’Souza

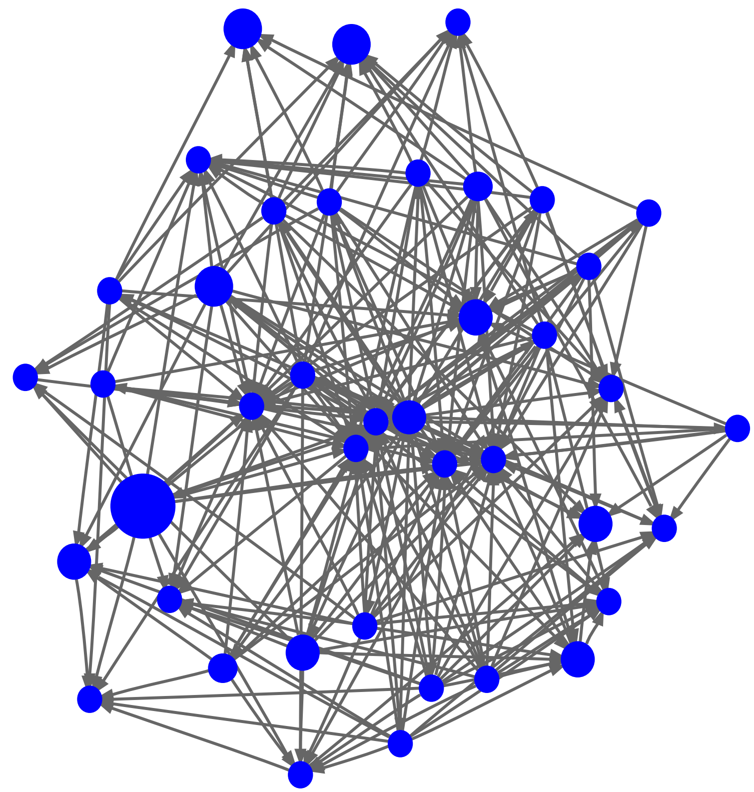
Software systems

(Highly evolveable, modular, robust to mutation, exhibit punctuated eqm)

Open-source software as a "systems" paradigm.

Networks:

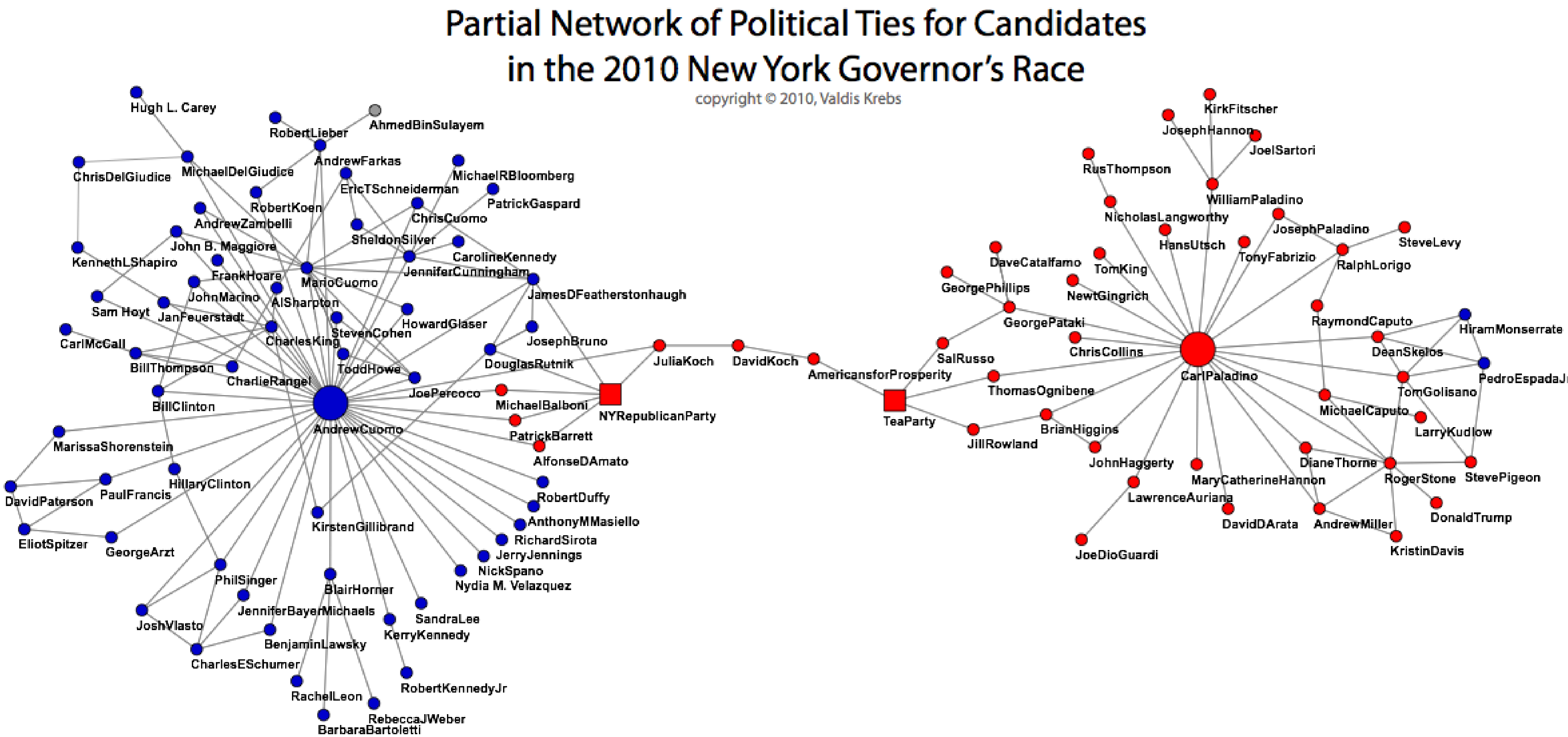
- Function calls
- Email communication
- Socio-Technical congruence



Bird, Devanbu, D'Souza, Filkov, Saul, Wen

A contemporary social network

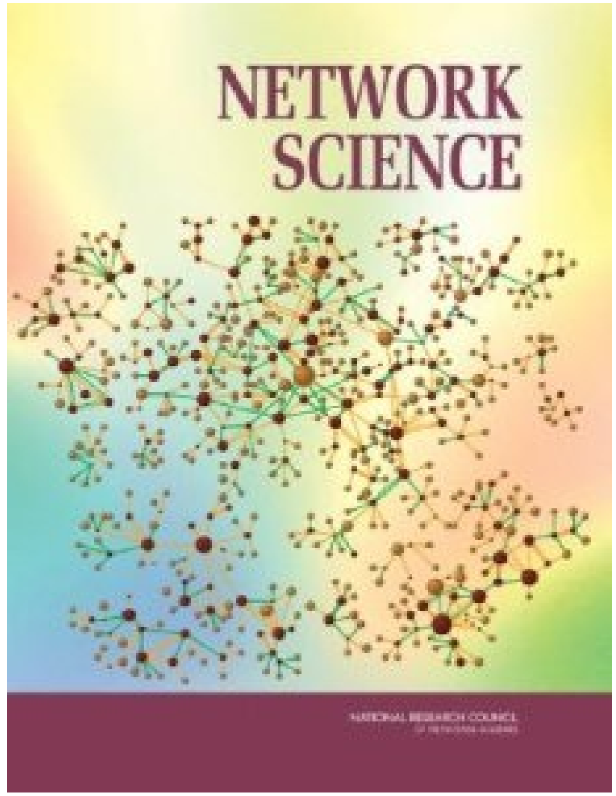
(Taken from http://www.thenetworkthinkers.com/)



Networks: Physical, Biological, Social

- **Geometric** versus **virtual** (Internet versus WWW).
- **Natural** /spontaneously arising versus **engineered** /built.
- **Directed** versus **undirected** edges.
- Each network **optimizes** something unique.
- Identifying **similarities** and fundamental **differences** can guide future design/understanding.
- Interplay of **topology** and **function** ?
- Unifying features:
  1) **Broad heterogeneity in node degree.**
  2) **Small Worlds** (Diameter $\sim \log(N)$)
  3) **Clustering** (Triangle motifs)

Natn Acam Sciences/Natn Research Council Study (2005)



“all our modern critical infrastructure relies on networks... too much emphasis on specific applications/jargon/disciplinary stovepipes... need a cross-cutting science of networks... Research for the 21st century”

Past decade, explosion of work and tools

SOFTWARE PACKAGES (excerpt Table 9.1 Newman's book)

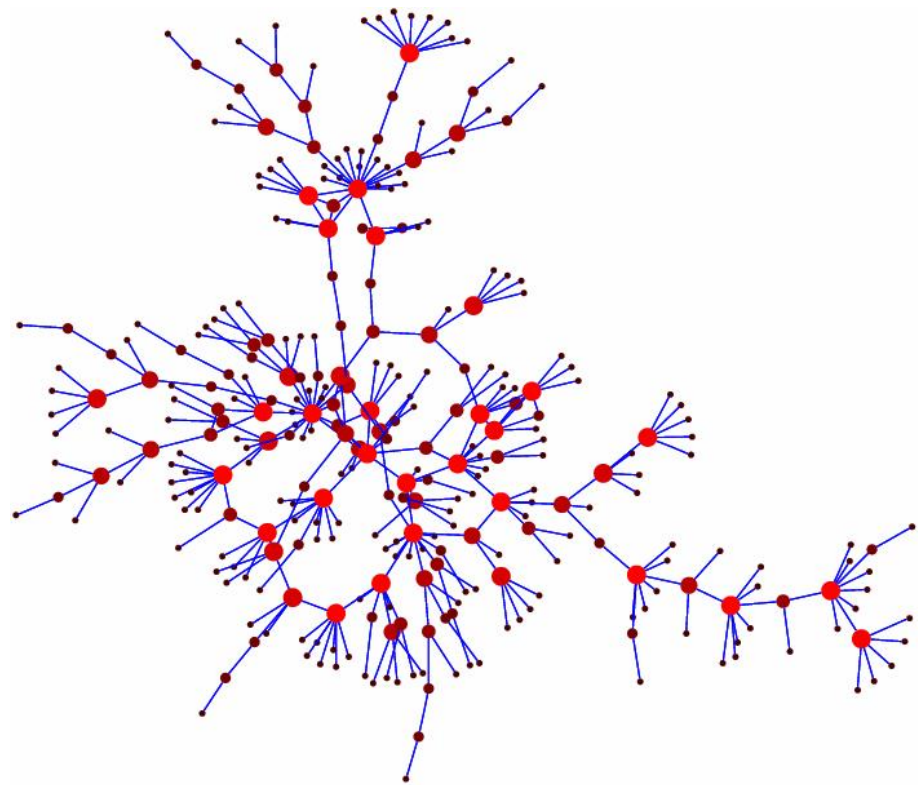
Name	Cost	OS	Description
NetworkX	Free	WML	Interactive analysis in Python
yEd	Free	WML	Interactive visualization
Graphviz	Free	WML	Visualization (Interactive gui M)
iGraph	Free	WML	C/R/Python libraries
Pajek	Free	W	Interactive SNA and visualization
Net Workbench	Free	WML	Interactive analysis and visualization

Past decade, explosion of work and tools

- **Models** of network topology:
  - random graphs
  - growth mechanisms
  - robustness and resilience
- **Measures** of network topology
- **Processes** on networks
  - Percolation
  - Epidemic spreading
  - Synchronization
  - Web search
- **Optimization**
  - User optimal versus system optimal
  - Braess's paradox
- Domain specifics and applications: CS, traffic, biology, social nets

Achievements of Single Network View

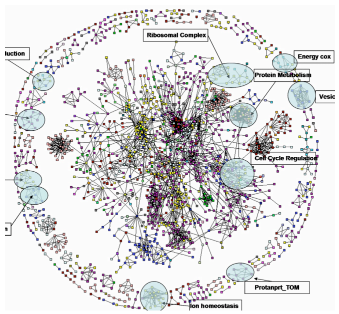
- **Power law** (broad scale) degree distributions ubiquitous.
- **Small world** effect (small diameter and local clusters).
- **Vulnerability** to "hub" removal / **resilience** to random removal.
- **Percolation**, spreading and epidemics (phase transitions)
- **Cascades**.
- **Synchronization**.
- Random walks / **Page rank**.
- **Communities** / subnetworks.
- **Structural roles** of nodes.



In reality a collection of interacting networks:

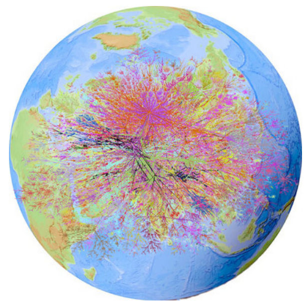
Networks:

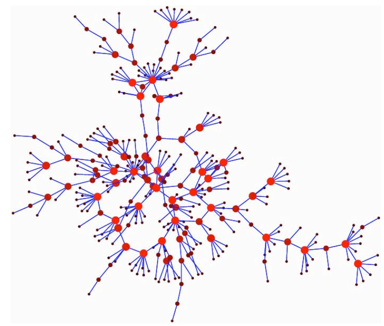
Transportation Networks/ Power grid
(distribution/ collection networks)



Biological networks
- protein interaction
- genetic regulation
- drug design

Computer networks





Social networks
- Immunology
- Information
- Commerce

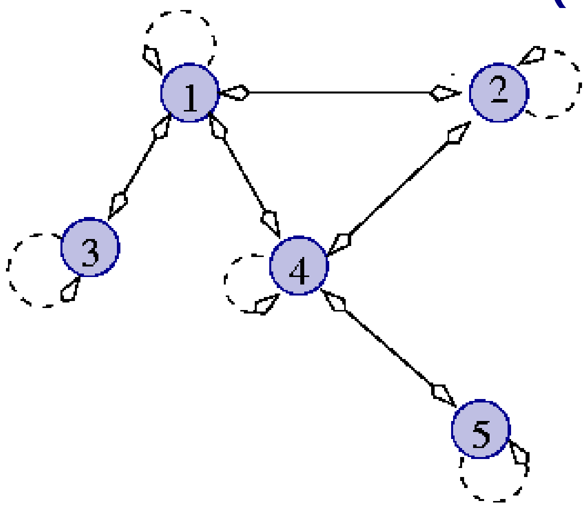
- E-commerce → WWW → Internet → Power grid → River networks.
- Biological virus → Social contact network → Transportation networks → Communication networks → Power grid → River networks.
  (Historical progression: Spatial waves (Black plague)   Regional outbreaks (ships)   Global pandemics (airplanes))

How do we represent a simple individual network as a mathematical object?

Representing a network: NETWORK TOPOLOGY

Connectivity matrix, M:

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



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

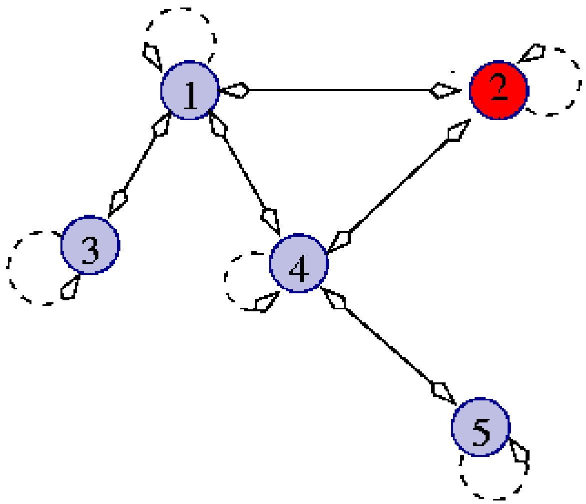
Node **degree** is number of links.

(Diameter, Clustering, Assortativity, Betweenness centrality)

Network Activity: FLOWS on NETWORKS

(Spread of disease, routing data, materials transport/flow)

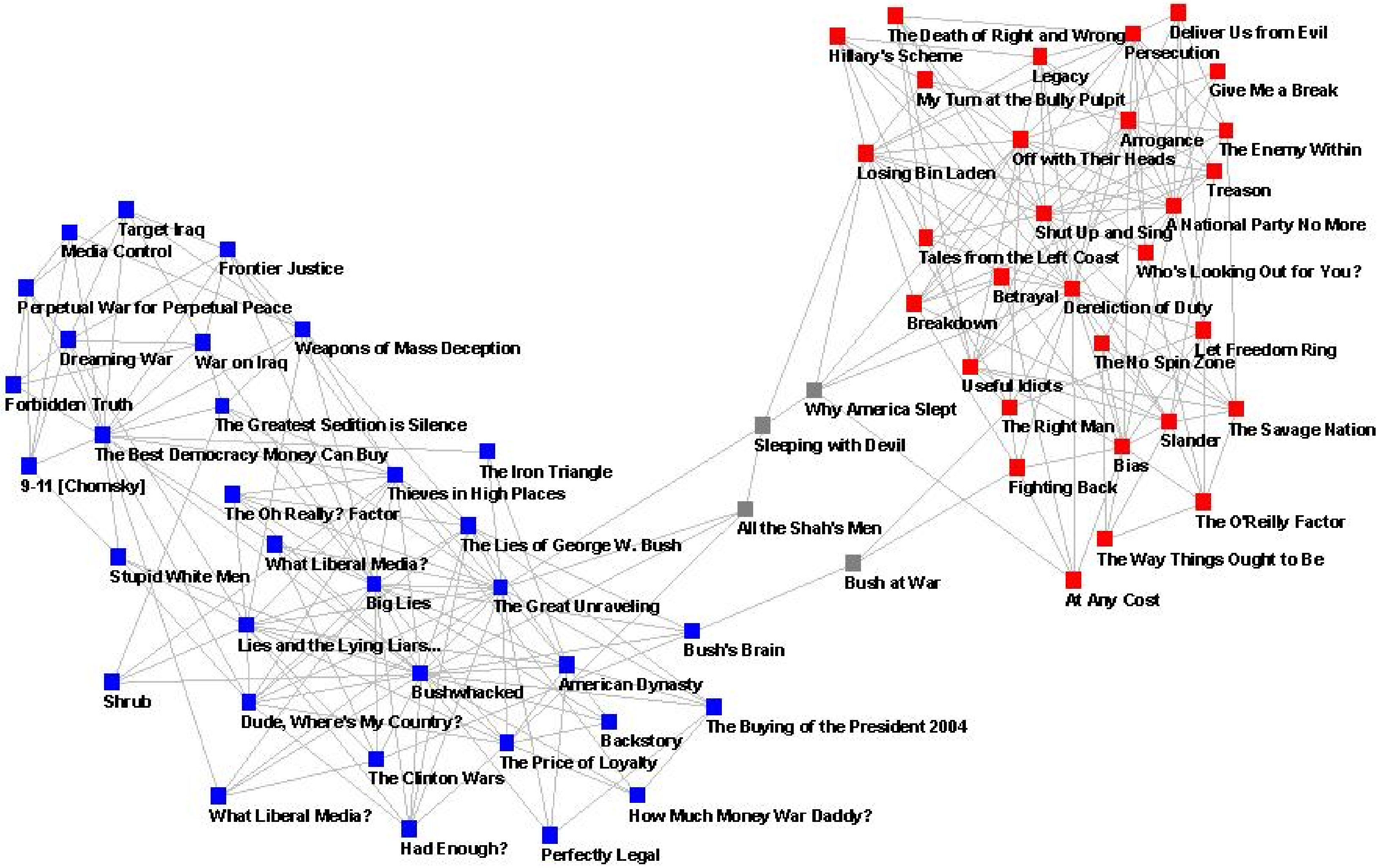
Random walk on the network has state transition matrix, P:



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

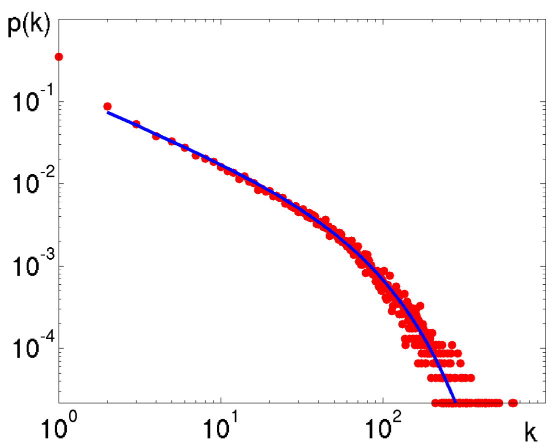
- Eigenvectors $\rightarrow$ **Page Rank**, Communities/modules.
- Eigenvalues $\rightarrow$ timescales.

Example Eigen-technique: Community structure (Political Books 2004)

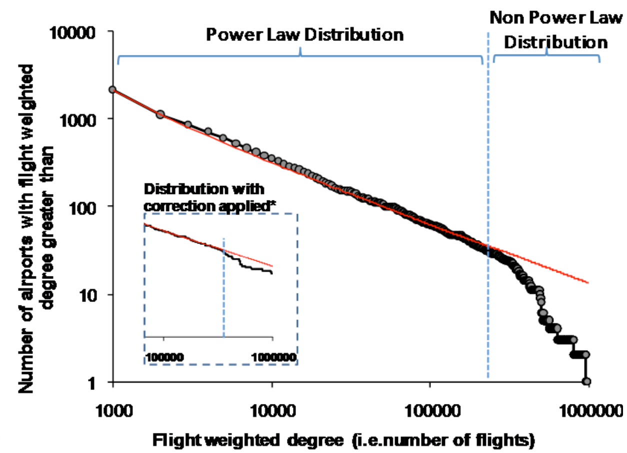


M. Girvan and M. E. J. Newman

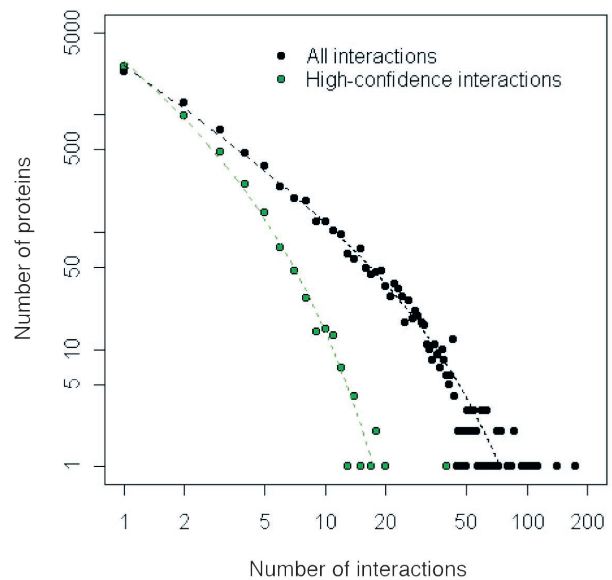
Commonly observed: broad scale degree distribution



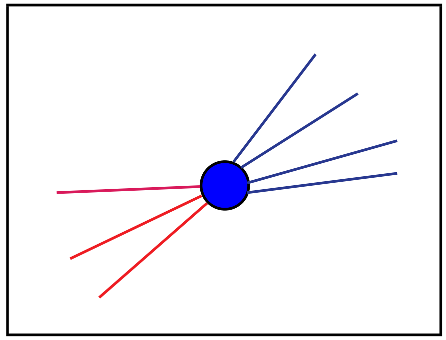
Social contacts
Szendröi and Csányi



Airport traffic
Bounova 2009



Protein interactions
Giot et al Science 2003



(node degree)

- Small data sets, power laws vs log normal, stretched-exponential, etc...
  - Exceptions: Power grids? Router-level Internet?

Power law with exponential tail

Ubiquitous empirical measurements:

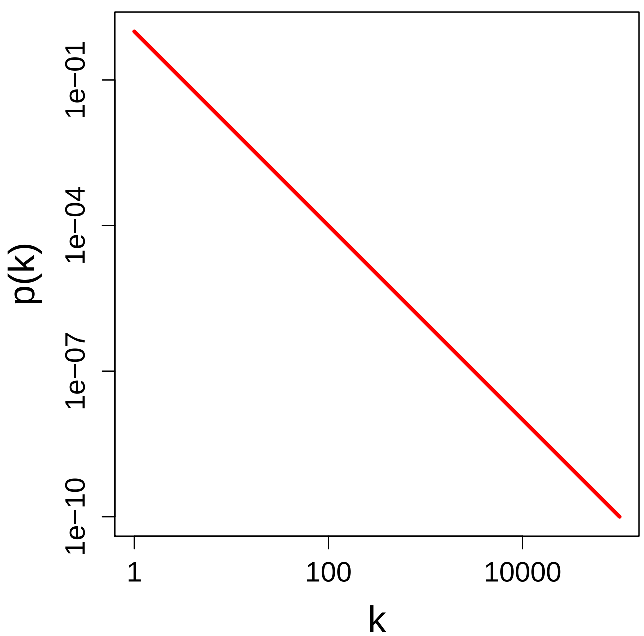
System with: $p(x) \sim x^{-B} \exp(-x/C)$	B	C
Full protein-interaction map of *Drosophila*	1.20	0.038
High-confidence protein-interaction map of *Drosophila*	1.26	0.27
Gene-flow/hydridization network of plants as function of spatial distance	0.75	10^5 m
Earthquake magnitude	1.35 - 1.7	$\sim 10^{21}$ Nm
Avalanche size of ferromagnetic materials	1.2 - 1.4	$L^{1.4}$
ArXiv co-author network	1.3	53
MEDLINE co-author network	2.1	~ 5800
PNAS paper citation network	0.49	4.21

What is a power law?

(Also called a "Pareto Distribution" in statistics).

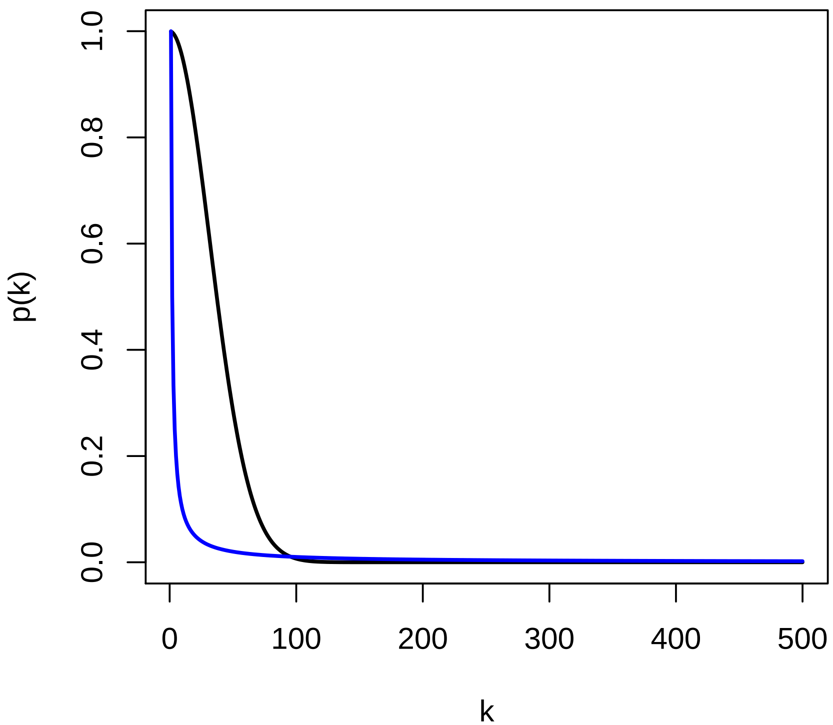
$$p_k \sim k^{-\gamma}$$

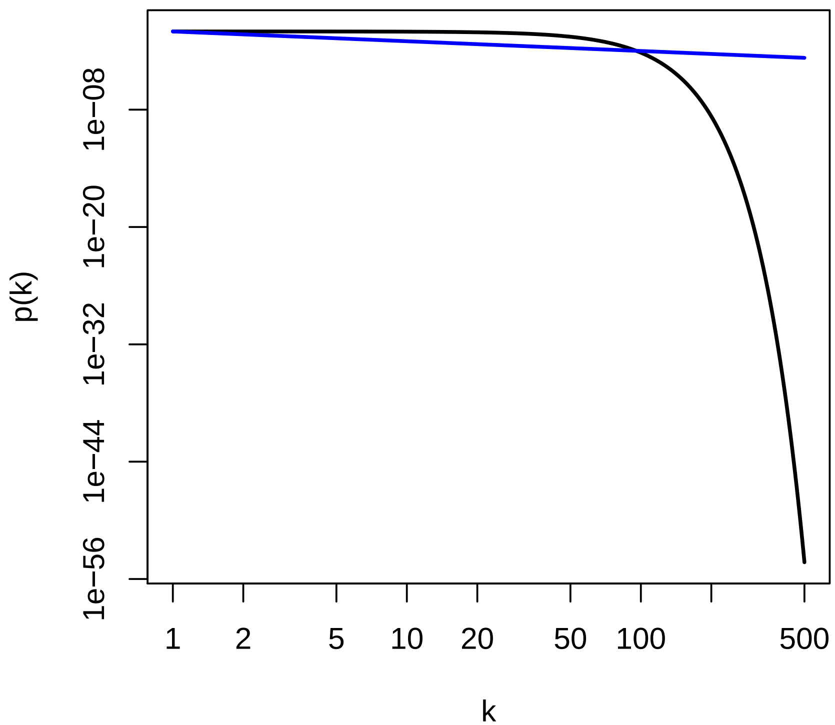
$$\ln p_k \sim -\gamma \ln k$$



Power Laws versus Bell Curves:
"Heavy tails"

- Power law distribution: $p_k \sim k^{-\gamma}$.
- Gaussian distribution: $p_k \sim \exp(-k^2/2\sigma^2)$.





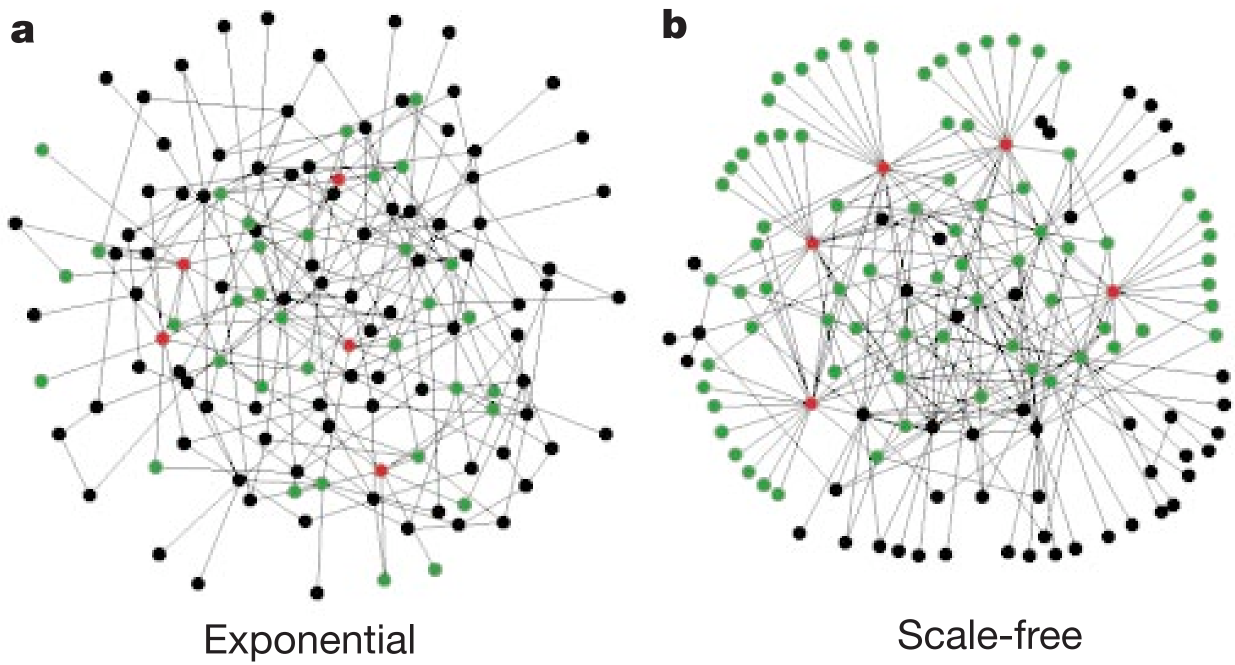
If $1 < \gamma < 2$, mean and variance $\rightarrow \infty$.

If $2 < \gamma < 3$ mean is finite, but variance $\rightarrow \infty$.

The flip-side of hubs

Random Graphs with Power Law degree distributions :

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



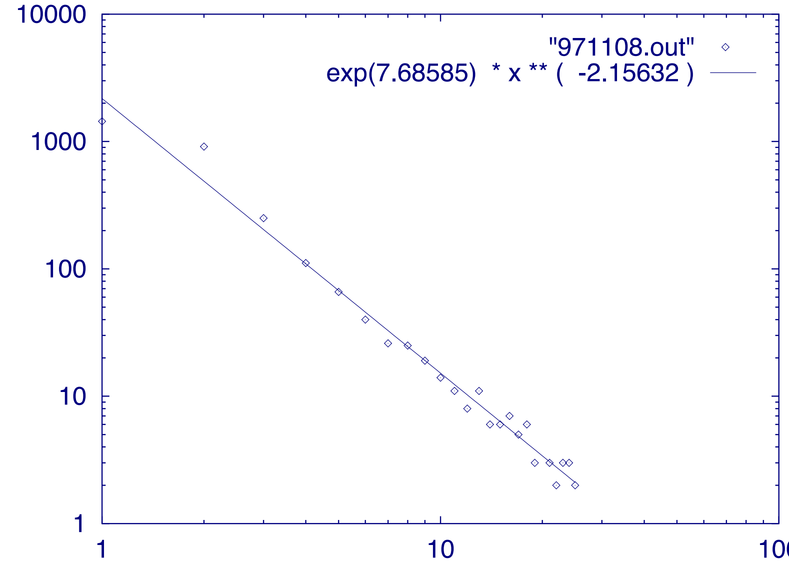
N=130, E=215, Red five highest degree nodes; Green their neighbors.

Exp has 27% of green nodes, SF has 60

- **Extremely "robust" to random failure (most nodes leaves).**
- **Extremely "fragile" to targeted attack (hubs are crucial).**

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

“The Achilles Heel of the Internet”



(a) Int-11-97

Faloutsos^3, *SIGCOMM* 1999

- “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 and scale-free networks (Extent of spreading)

More consequences of PLRGs with $2 < \gamma < 3$

- Callaway et. al. *Phys. Rev. Lett.* (2000).
- Cohen et. al. *Phys. Rev. Lett.* (2000) – Percolation threshold approaches zero for scale-free with exponent $2 < \gamma < 3$.
- Pastor-Satorras and Vespignani, *Phys. Rev. Lett.* (2001) – SIR and SIS models.

- **If $2 < \gamma < 3$ epidemic threshold approaches zero as system size goes to infinity.**

Modeling network growth

- **Random graphs** (Erdős-Rényi 1959, 1960; Gilbert 1959; Bollobás 1985)
- **Configuration models** (Bollobás 1980, Molloy and Reed 1995)
- **Randomly grown graphs** (CHKNS 2001)
- Growth by **Preferential Attachment** (Polya 1923; Yule 1925; Zipf 1949; Simon 1955; Price 1976; Barabási-Albert 1999)
- Growth by **copying** (WWW inspired) (Kumar et. al. FOCS 2000); **Copying with mutation** (bio inspired) (e.g., DMC – Vazquez et. al. 2003))
- **Non-linear PA** (Krapivsky-Redner 2000, 2001); **Geometric PA** (Flaxman-Frieze-Vera 2006); **PA from optimization** (DBCBK 2007)
- Growth with **feedback** (White-Kejzar-Tsallis-Farmer 2006; R.D.-Roy 2008)
- Growth with **choice** (Bohman-Frieze 2001, Achlioptas-D'Souza-Spencer 2009, ...)

Many network growth models produce power law degree distribution

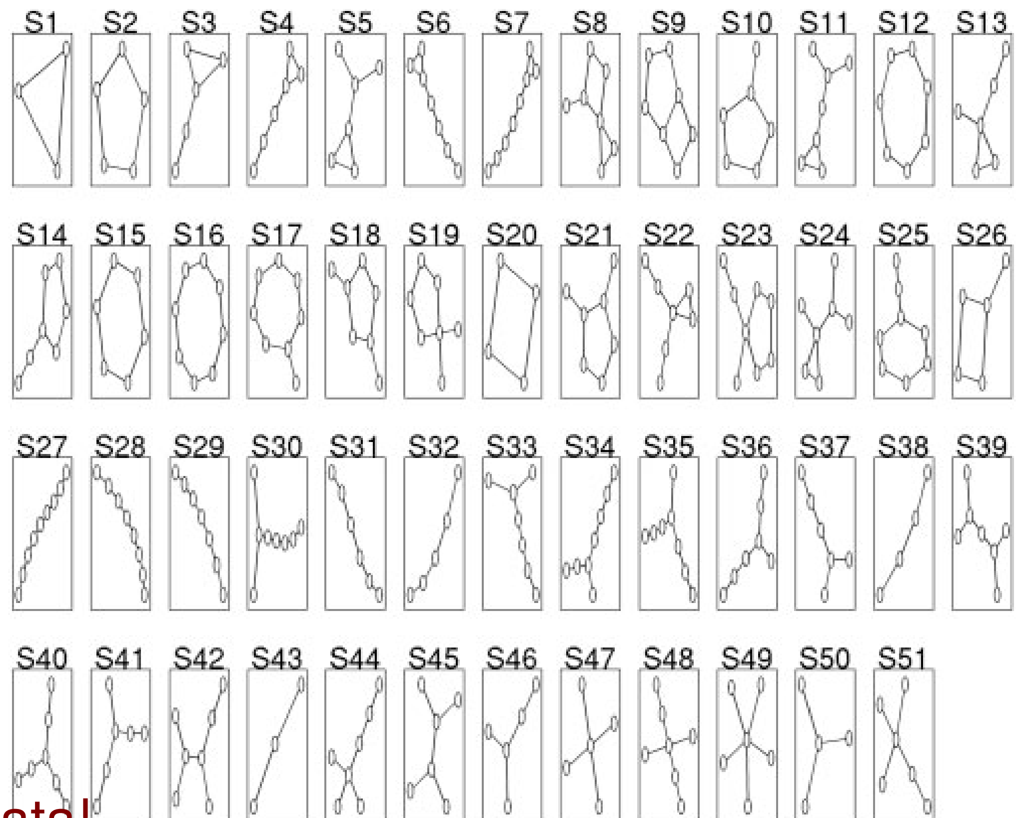
- Preferential attachment
- Copying models (WWW, biological networks, ...)
- Optimization models

Some outstanding challenges

- Incorporating additional attributes beyond degree
- Validation

Outstanding issue – Validation of growth model

- Currently more advances in **Model Selection** (selecting the best model from a set of choices)
- **Middendorf, Ziv, and Wiggins *PNAS* 102, 2005**
- ***Discriminative classification*** to compare seven proposed models. Determine which model best describes data of Drosophila protein interaction network.
- Build classifier:



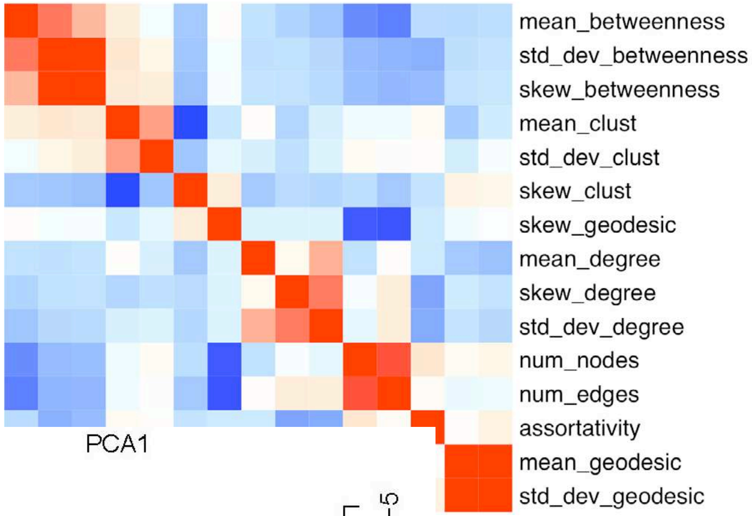
- DMC wins for Giot Drosophila data!

Outstanding issue – Fitting multiple network metrics

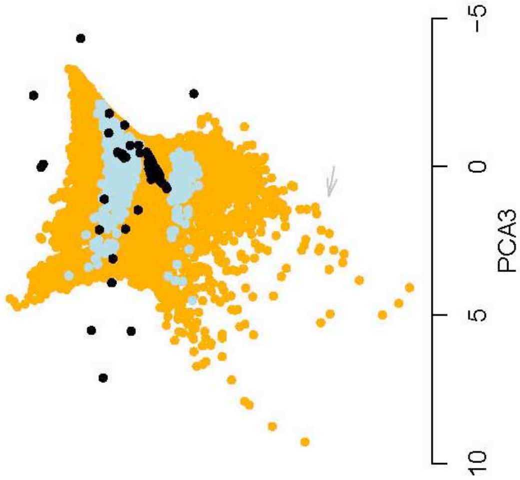
(e.g. Filkov et. al. *Europhysics Letters* 86 (2009))

- **Multiple network attributes**: degree distribution, clustering coefficient, betweenness distributions, higher-order loops, diameter, assortativity, degree-degree correlations, etc.

- **Heat map** of correlations:



- Again, **Model Selection** :
  Use PCA to compare real-world networks and competing models.



These results follow from the "statistical physics" treatment of networks

- Calculate properties of the ensemble of all random graphs with certain properties.

- Phase transitions:
  – Connectivity (emergence of the giant component)
  – Epidemic thresholds (truncated power law deg dist have $p_c > 0$)

- Master equations / typical evolution of growing network.

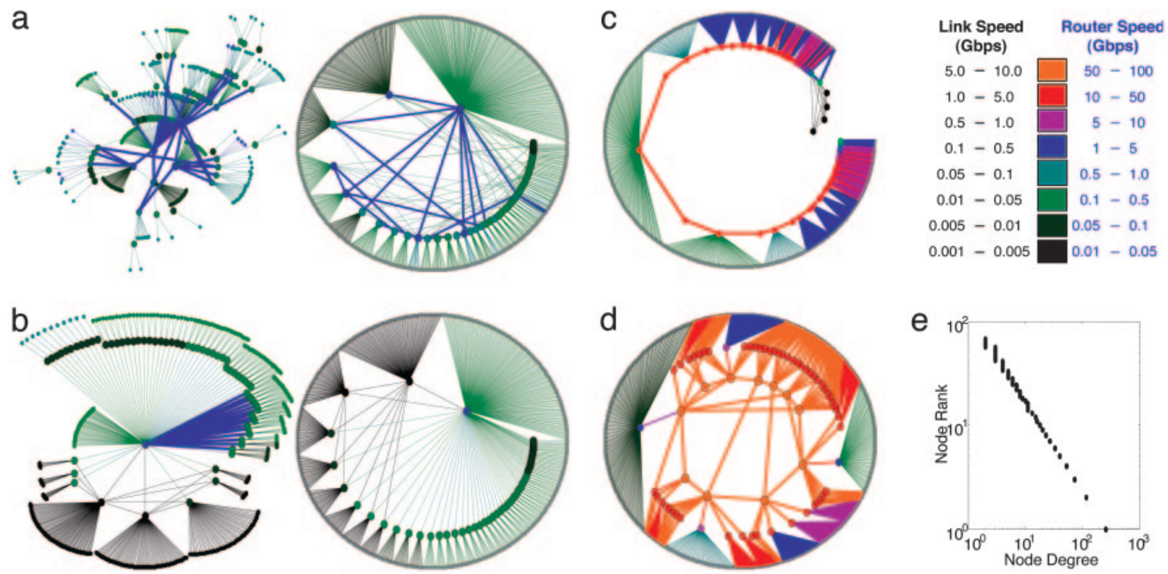
- Models of network growth

Ensemble versus an Individual engineered or biological system

- **More than just degree distribution: Fine structure**

Doyle, Alderson, Li, Low, Roughan, Shalunov, Tanaka, Willinger, PNAS **102** (4)2005.

All these have same deg dist:



- **"Scale-Free"** neglects correlations, additional structure.

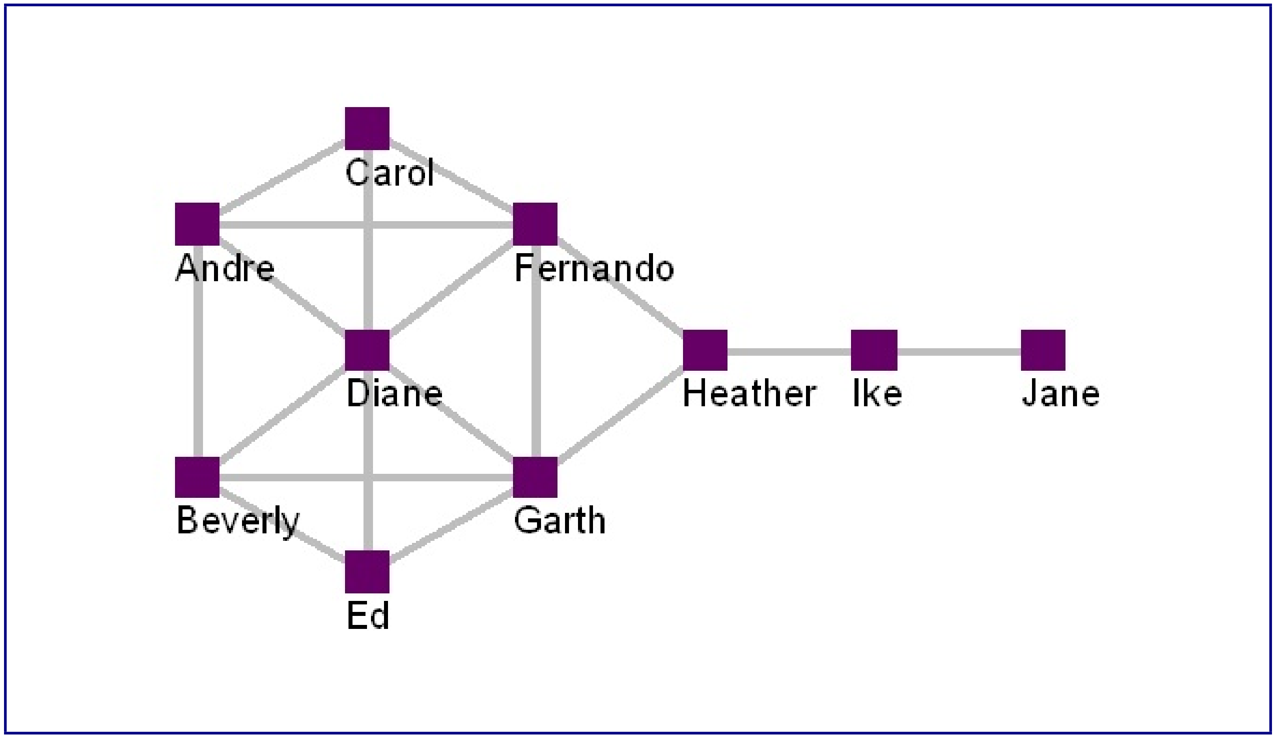
- **REDUNDANCY!** a key principle in engineering and bio. (Connect up the low degree nodes and scale-free resilient to targeted attack.)

Small scale interactions and individual differences can matter: (social systems and biology)

A classic example from Social Network Analysis (SNA)

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

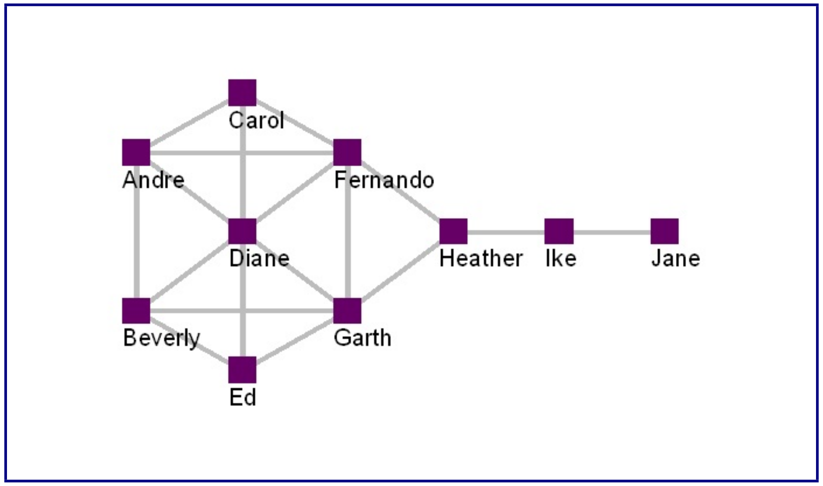
The "Kite Network"



Who is important and why?

See also Borgatti Science 323, 2009. Social networks vs random graphs

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.

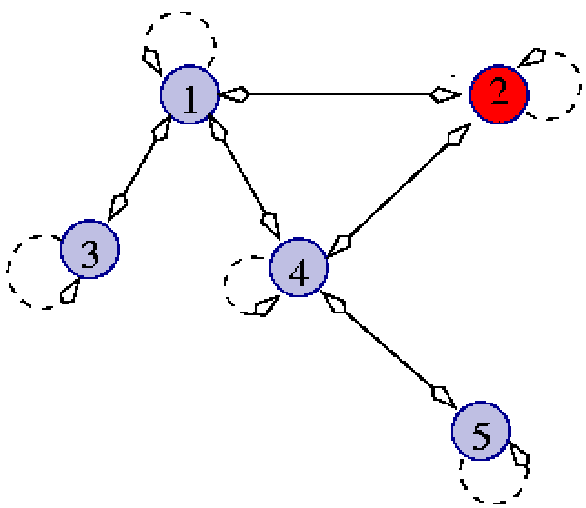
Beyond Topology

Network Activity: FLOWS on NETWORKS

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

FLOWS on NETWORKS : Random walks

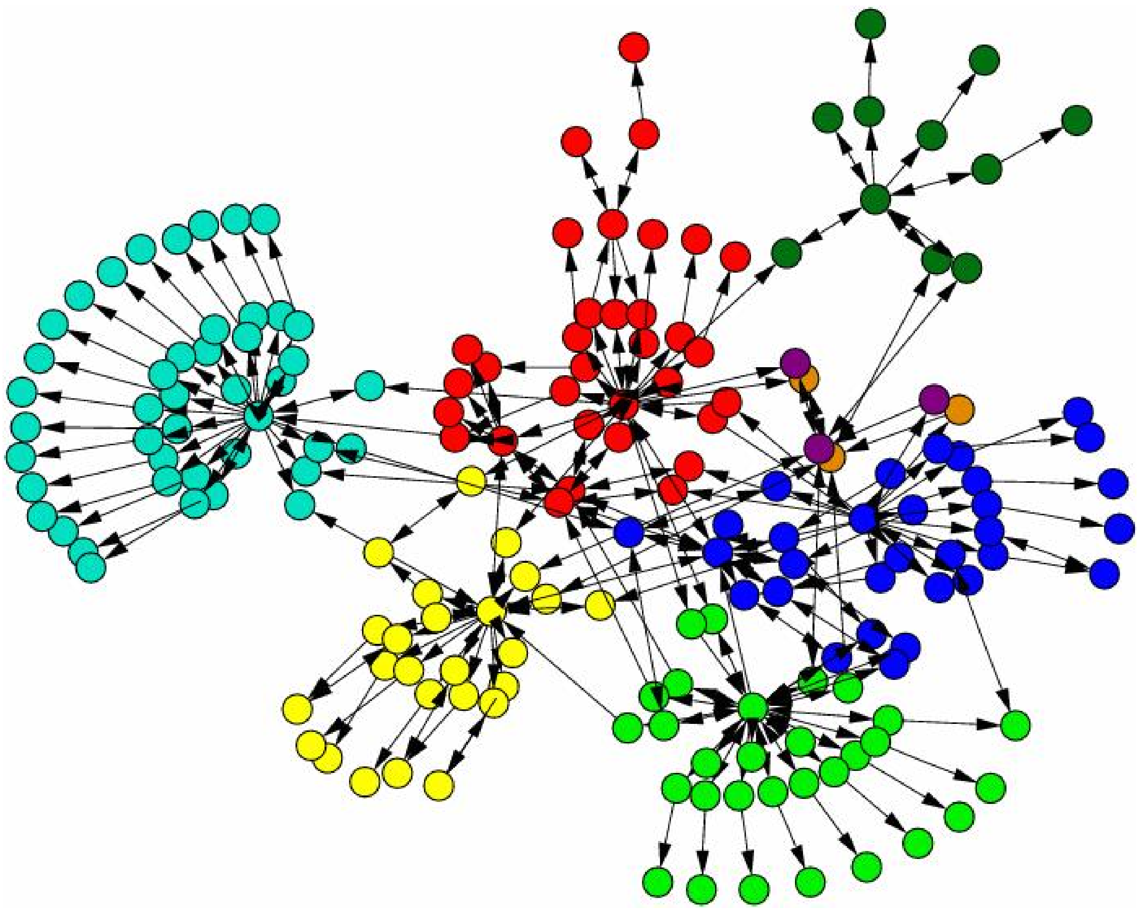
Random walk on the network has state transition matrix, P:



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

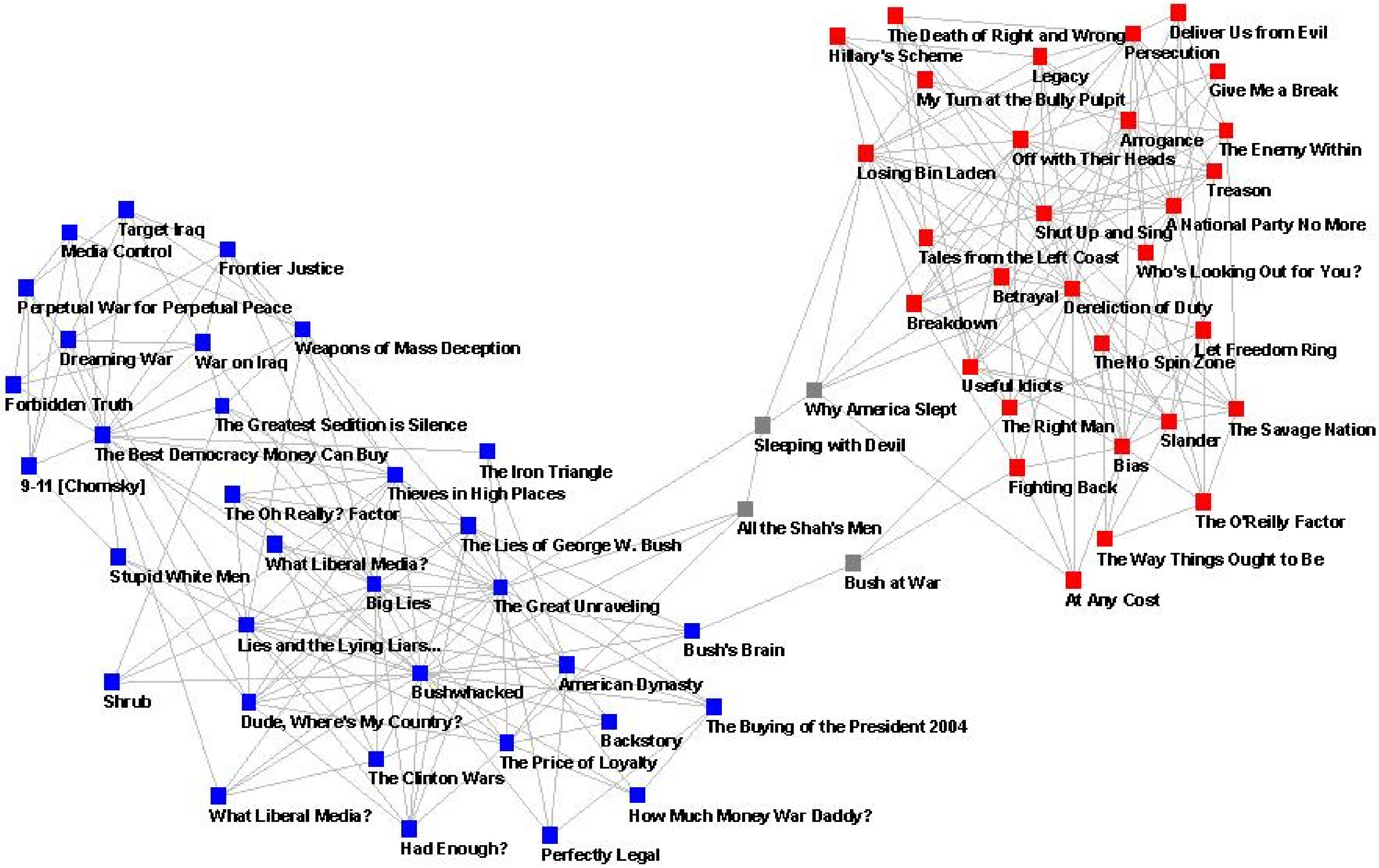
The eigenvalues and eigenvectors convey much information. Markov Chains, Spectral Gap.

Random walk on the WWW is the "Page Rank"



Page Rank of a node is the steady-state random walk occupancy probabilty.

Example Eigen-technique: Community structure (Political Books 2004)



M. Girvan and M. E. J. Newman

Beyond Topology and Flow

DYNAMICS OF NETWORKS

Key distinction:
Dynamics *of* networks vs dynamics *on* networks

- **Dynamical process taking place on a network substrate**

– Information flows
– Epidemic spreading
– Synchronization, etc.

- **Dynamics of the nodes and edges constituting the network**

– Network growth/decline; edge intermittency; etc.

Separation of timescales.

i.e., when is the network substrate essentially static over the timescale of the dynamical process? .... (e.g web crawling; routing tables in mobile ad hoc networks; epidemic spreading.)

Dynamical processes on (static) networks

(Durrett, *Random Graph Dynamics*, Cambridge Univ. Press (2006))

(Barrat, Barthélemy, Vespignani *Dyn. Proc. on Complex Networks*, Cambridge (2008))

- **Diffusion**
- **Spreading** / Percolation / Contact processes / Epidemics
- **Synchronization** – fireflies, Josephson junctions, ad hoc sensor networks ...
- **Searching** on networks (WWW, P2P, sensor nets, social nets)
- **Flows**: data, materials, transportation, biochemical (on many levels – genes, proteins, metabolites)
- **Traffic**, congestion, jamming
- **Cascades** / Avalanches / Sandpiles

Dynamics of the underlying network

- Network **growth** (long list here.... see previous slide)
- **Vulnerability** to node and edge deletion
- Network **shrinking** (nodes – banking, garment industry)
- **Online Social Networks**: shrinking diameter and densification
- Link **intermittency** (Internet)
- **Shifting centrality** (AS level internet)
- Ad hoc / MANETS (wireless radio communications)
  (self-org algorithms, capacity, scaling, energy conservation)
- **Coevolution**:
  (socio-technical – land use/population growth; task oriented social networks (OSS)); (biological – gene regulatory/ppi (conserved motifs) / ecosystems)

If there is time, we will discuss phase transitions here....

Concepts covered today

- Social, physical and biological networks
- Simple network metrics (recapped next page)
- Power law random graphs
- Model selection
- Random graphs
- Phase transitions in connectivity
- Random walks on networks

Outstanding challenges

- How do we connect network **structure** to **function?**
  - Degree
  - Clustering Coefficient
  - Motifs
  - Betweeness Centrality
  - Assortativity
  - Flow and transport
  - Growth/evolution mechanisms.
- **Interacting networks**
- **Strategic interactions** / Game theory on networks