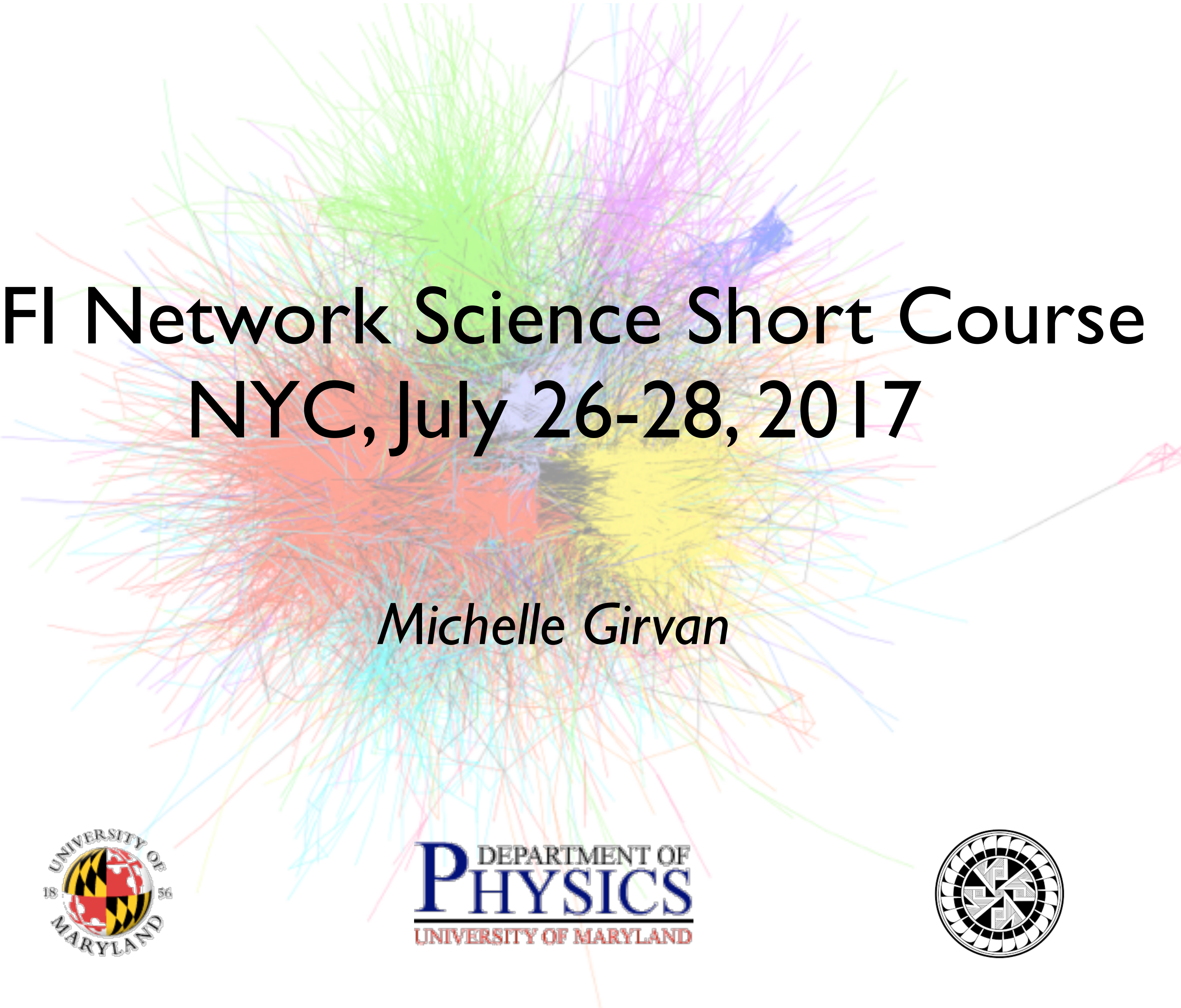
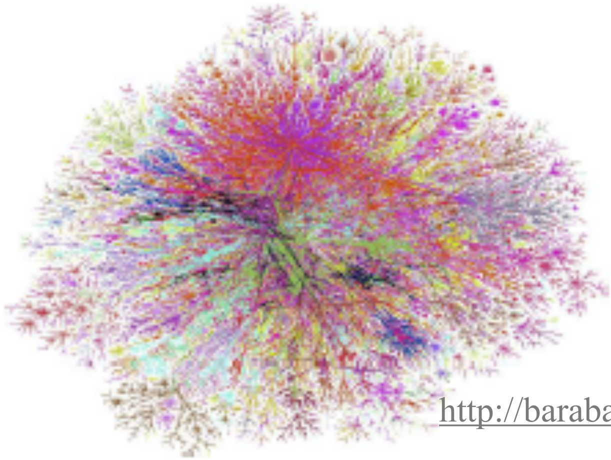


SFI Network Science Short Course
NYC, July 26-28, 2017

Michelle Girvan

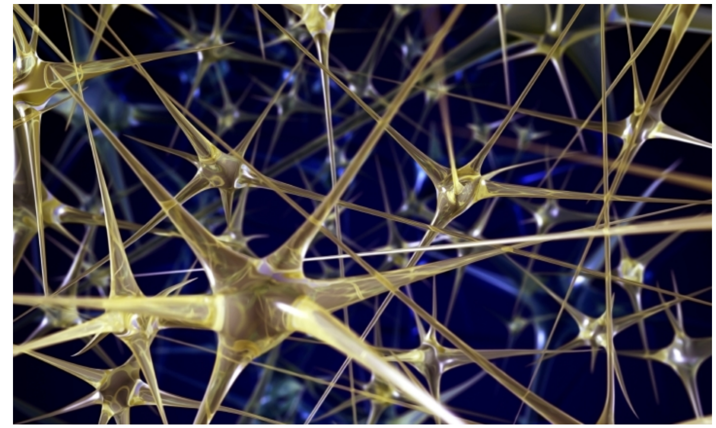
DEPARTMENT OF PHYSICS
UNIVERSITY OF MARYLAND

Examples of Complex Networks



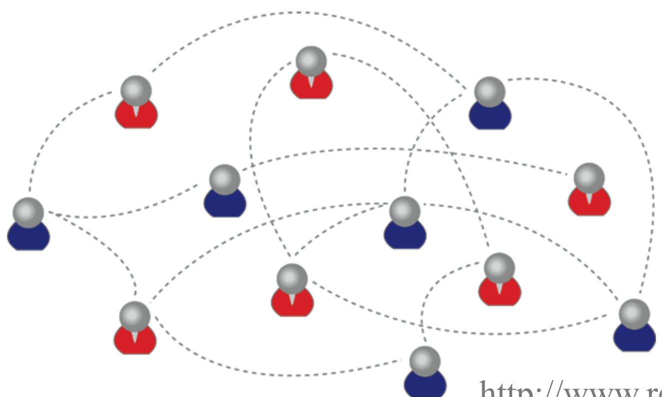
http://barabasilab.com/gallery

The Internet



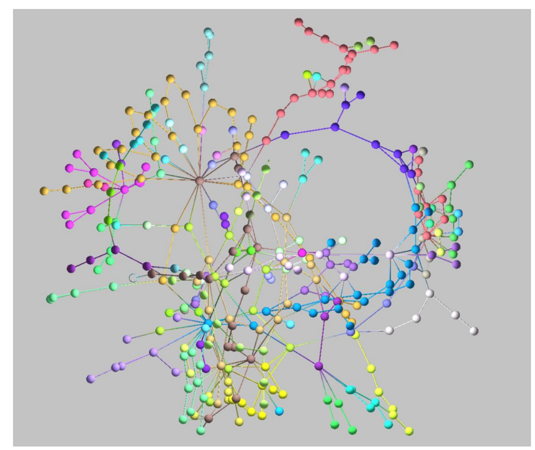
http://www.zmescience.com

A Neural Network



http://www.relenet.com/

A Social Network



http://www4.toulouse.inra.fr/toxalim_eng

A Metabolic Network

A network approach is appropriate when the role of a node or link in a system depends critically on the system's connections

Overview of the Short Course

- **Day 1: Network Basics**
  - Intro to Network Science (Michelle Girvan)
  - Applications of Network Science (Hernan Makse)
  - Hands-on activity: Network Data Analysis (David Darmon)
- **Day 2: Network Dynamics**
  - Diffusion Dynamics (Damon Centola)
  - Ecological Networks (Neo Martinez)
  - Agent-based Network Models (Bill Rand)
  - Hands-on activity: Building a Network Diffusion Model (David Darmon)
- **Day 3: Next Generation Network Science for Complex, Big Data**
  - Cooperation Networks (Ole Peters)
  - Data-driven Explorations of the Ecology of Human Stories (Peter Dodds)
  - Networks in the Wild: Generative Models and Multilayer Networks (Daniel Larremore)
  - Hands-on activity: Complex Network Visualization (David Darmon)

Intro to Network Science

- **Goal:** To provide an introduction to the field of network science, starting with foundational work on key structural features of networks and begin to talk about how structure connects to function/dynamics

- **Outline:**

  » Systems of study and questions of interest

  » Key structural properties and network models that help us understand them

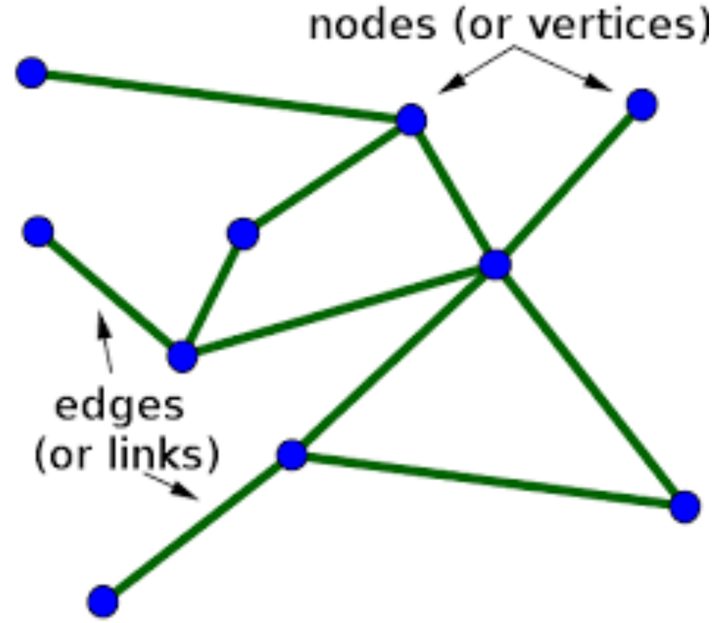
  » Connecting structure to function, first steps (more to come tomorrow)

Selected References

- Mark Newman, *Networks: An Introduction*, Oxford University Press, 2010.
- David Easley and Jon Kleinberg, *Networks, Crowds and Markets*, Cambridge University Press, 2010.
- A. Barabasi, *Network Science*, available online, print version to be published by Cambridge University Press in 2015.

What is a network?

A collection of nodes (vertices) and links (edges)



System	Nodes	Links
Wolrd-Wide Web	webpages	hyperlinks
Internet	computers	IP links
Citation Network	papers	citations
Social Network	individuals	friendships
Subway	stations	tracks
Neuronal network	neurons	synapses

The field of *Network Science* exists at the intersection of….

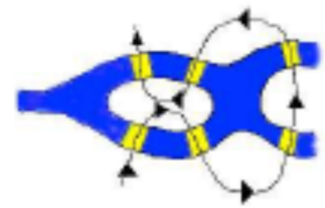
- Graph theory
- Statistical physics
- Computer science: algorithms and networks
- Dynamical systems
- Social network theory
- Complex systems
- and more….

The *Network Science* Approach

Traditional Questions:

Social Networks:
Who is the most important person in the network?

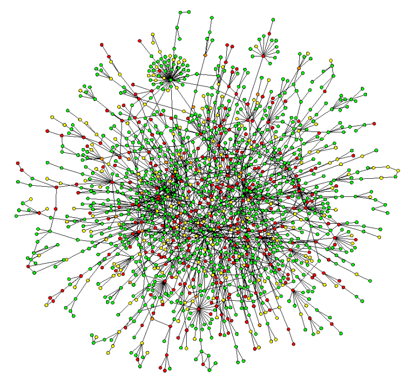
Graph Theory:
Does there exist a cycle through the network that uses each edge exactly once?



Network Science Questions:

How do different network structures influence the system's robustness to node or edge removal?

How does the network structure influence the system's dynamics or function?



Areas of Network Science Research

Structural Complexity

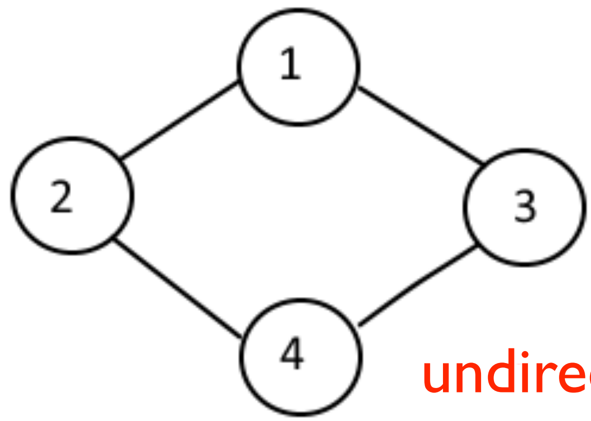
- The connection pattern between elements could have complex structural features, far from perfectly regular or perfectly random.
- The network could include different classes of nodes
- The edges could be heterogeneous with different weights, directions, signs. The network could also include different classes of edges.

Dynamical Complexity

- Dynamics on the network: processes could be taking place on a fixed network.
- Dynamics of the network: the network itself could be evolving in time.
- Dynamics of and on the network: processes could be taking place on a network that is evolving in time.

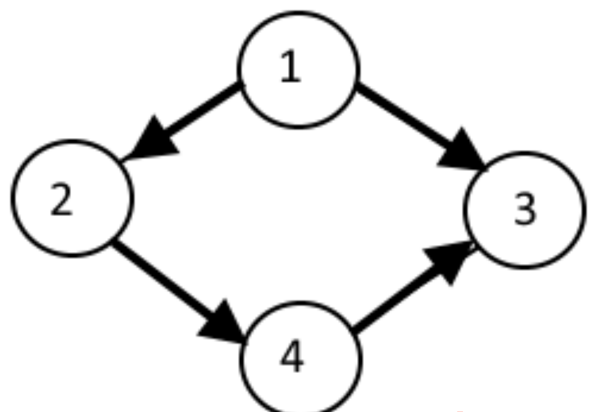
Representing a Network: The Adjacency Matrix

$$A_{ij} = \begin{cases} 1 & \text{if there is a link from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$



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

undirected network



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

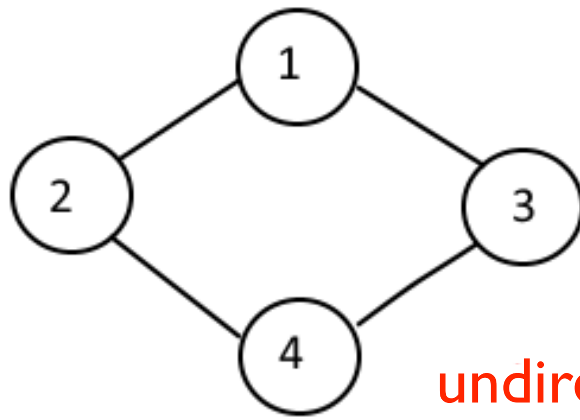
directed network

We will be working with undirected, unweighted networks unless otherwise specified.

Other Representations

Edge List | **Adjacency List**

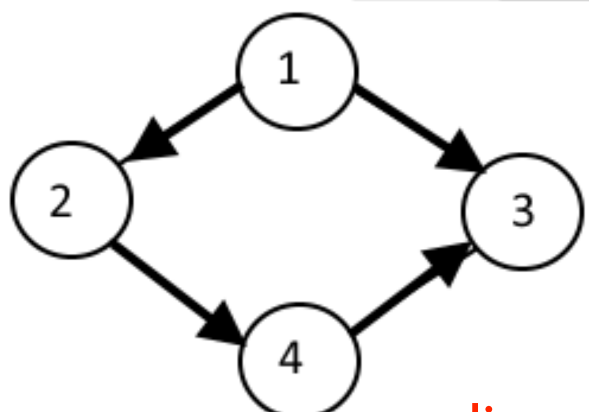
```
1 2
1 3
2 4
3 4
```



undirected network

```
1: 2 3
2: 1 4
3: 1 4
4: 2 3
```

```
1 2
1 3
2 4
4 3
```



directed network

```
1: 2 3
2: 4
3:
4: 3
```

Erdös-Renyi (ER) random graphs

- Consider a system of N nodes
- Connect each pair of nodes with probability p
- If we desire M edges, on average, then $p = \frac{2M}{N(N-1)}$
- Mean degree z (number of edges connected to a node) $z = \frac{2M}{N}$

Three different network realizations N=10, p =1/6

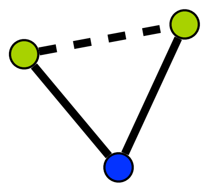
Some Network Measures

- Average path length

Let d_{ij} be the number of edges in the shortest path between nodes i and j. Then the average path length ℓ is

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}$$

- Clustering



C = Probability that two of a node's neighbors are themselves connected

In an ER random graph: $C_{\text{rand}} \sim 1/N$ (if the average degree is held constant)

$$C_{\text{rand}} = p = \frac{2M}{N(N-1)} = \frac{zN}{N(N-1)}$$

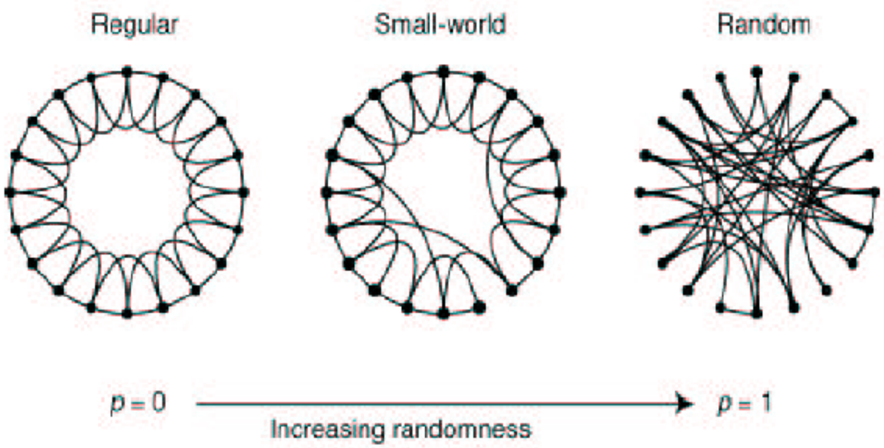
$$C_{\text{rand}} \sim \frac{1}{N}$$

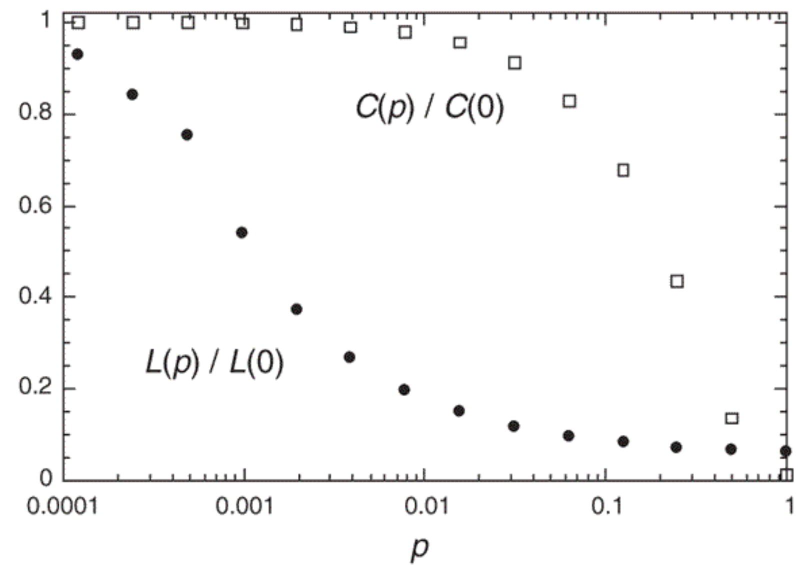
Network	N	ℓ	C	C_{rand}
movie actors	225266	3.65	0.79	0.00027
neural network	282	2.65	0.28	0.05
power grid	4941	18.7	0.08	0.0005

Table from Watts & Strogatz, *Nature* (1998)

Watts-Strogatz 'Small World' Model

Watts and Strogatz introduced this simple model to show how networks can have both short path lengths and high clustering.

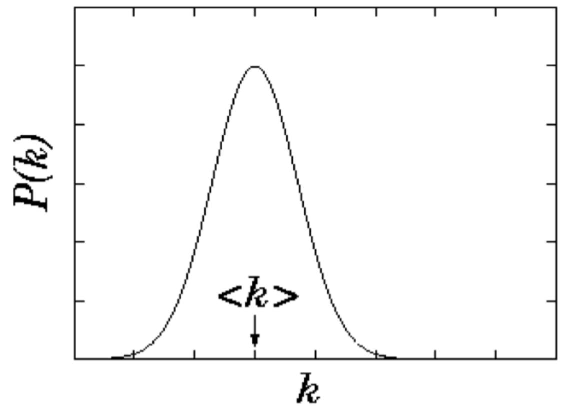


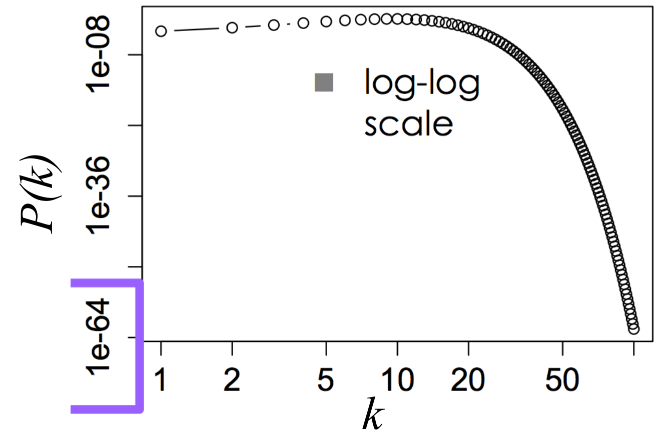


D. J. Watts and S. H. Strogatz, *Collective dynamics of "small-world" networks*, Nature 1998

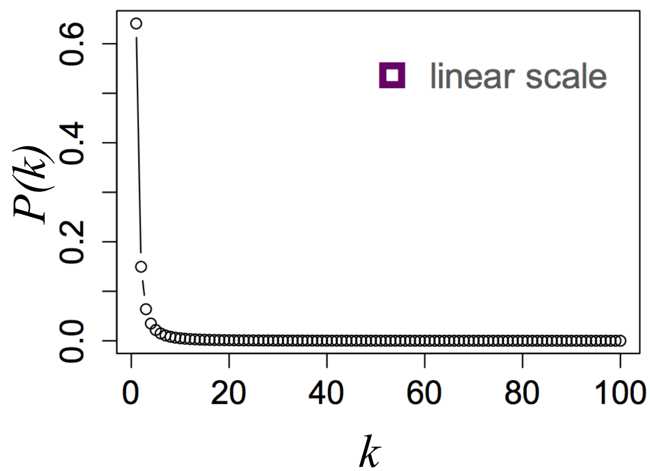
Degree Distributions

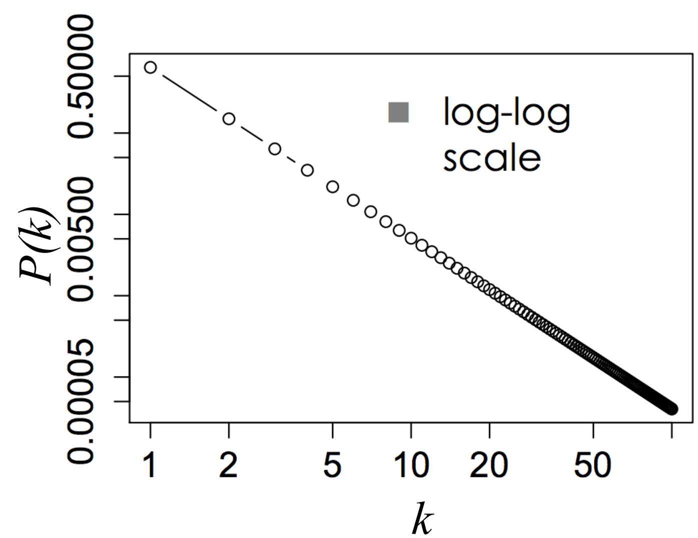
Sharply-peaked degree distributions:





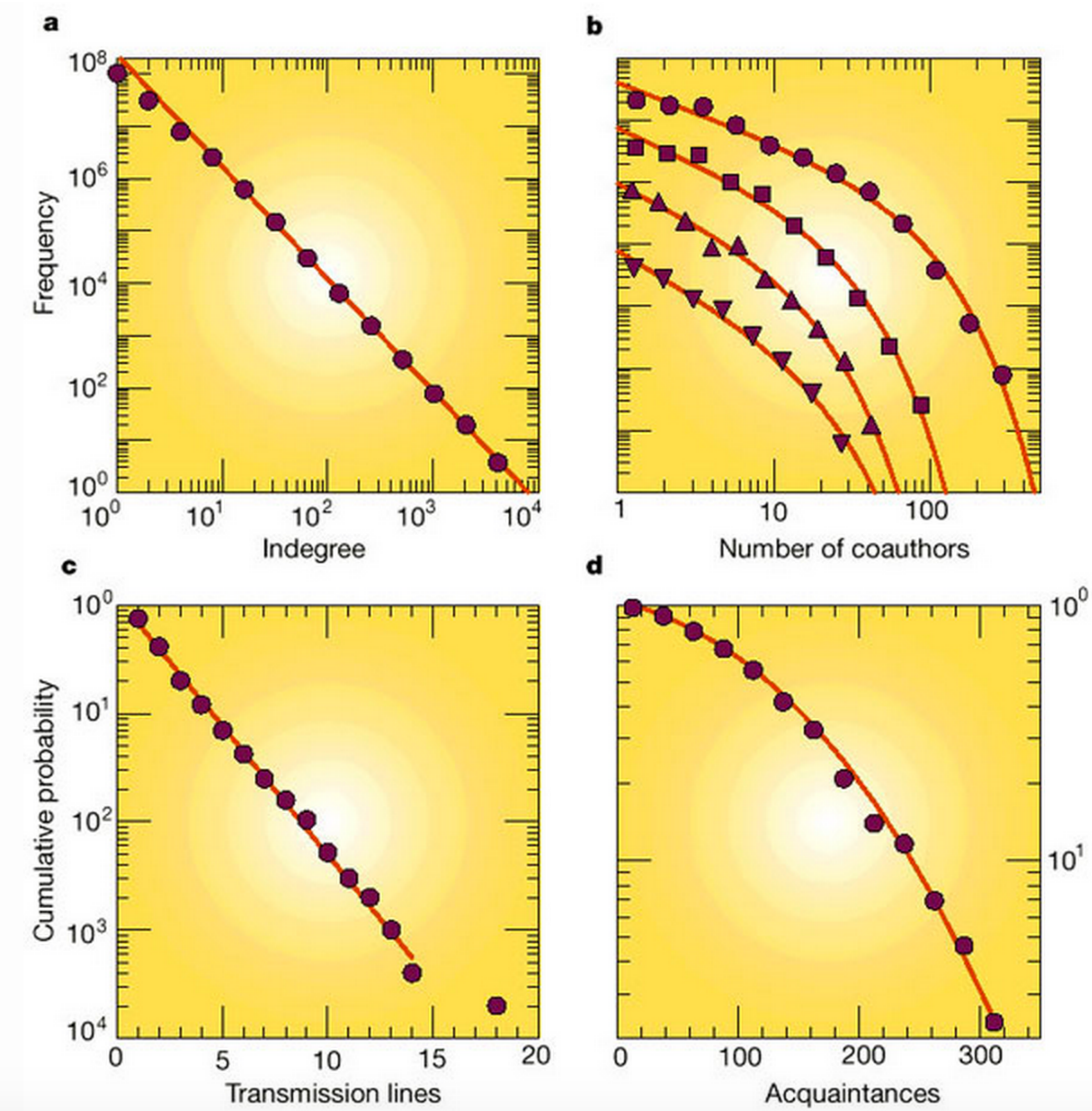
Scale-free degree distributions:





Images from the Barabasi Lab

Degree distributions for various networks



(a) World-Wide Web

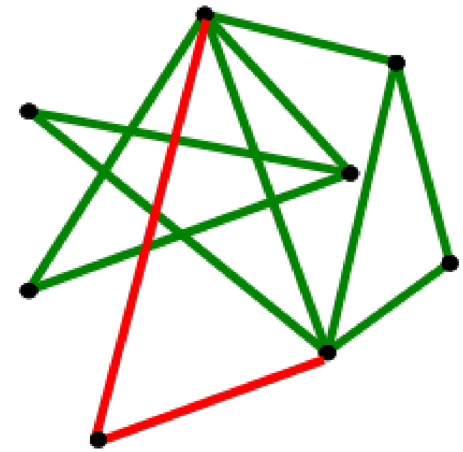
(b) Coauthorship networks: computer science, high energy physics, condensed matter physics, astrophysics

(c) Power grid of the western United States and Canada

(d) Social network of 43 Mormons in Utah

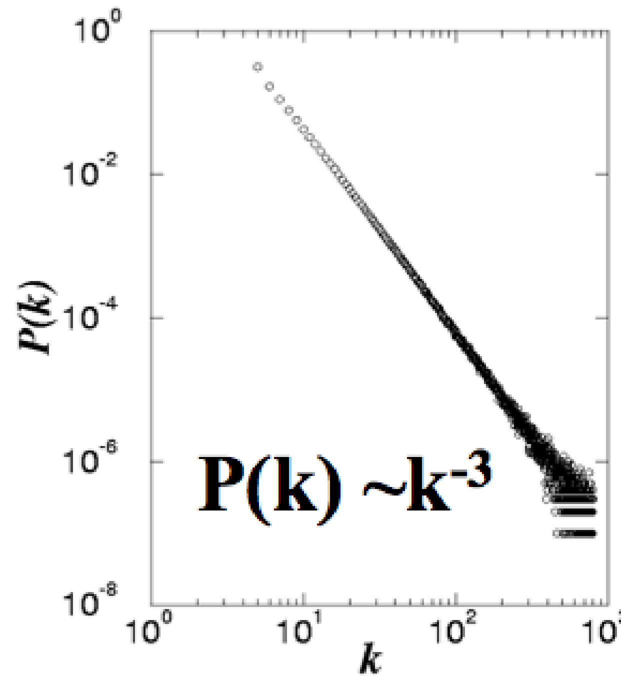
from Strogatz, "Exploring Complex Networks," *Nature* 2001

How might power law degree distributions arise?



One possible answer: Barabasi-Albert model of preferential attachment

- Growth - At each time step, we add a node with *m* new edges (connecting to nodes already existent in the system)
- Preferential attachment - The probability that a new node connects to an existing node *i* depends on the connectivity, k_i of that node.



A.L.Barabasi, R. Albert, *Science* (1999)

How does degree distribution matter?

A simple example

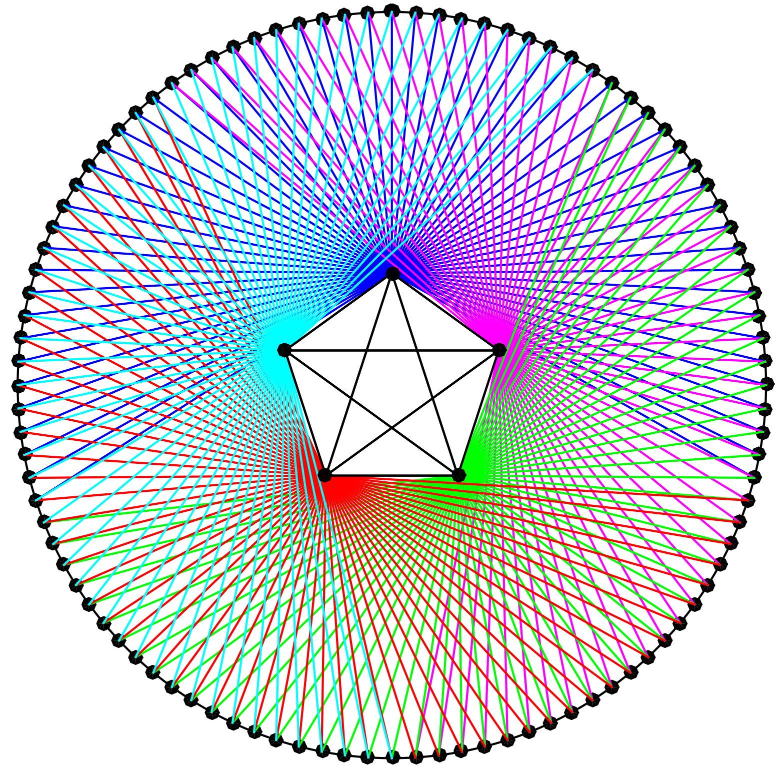
Let's consider 2 networks…

Network 1: 105 people

- Everyone has exactly 8 friends

Network 2: 105 people

- 100 "regular folk" each have 5 friends
- 5 "big wigs"
  - each have 60 regular folk friends
  - all big wigs are friends
- Average number of friends = 7.8095



Network 2

Q: How many friends do your friends have, on average?

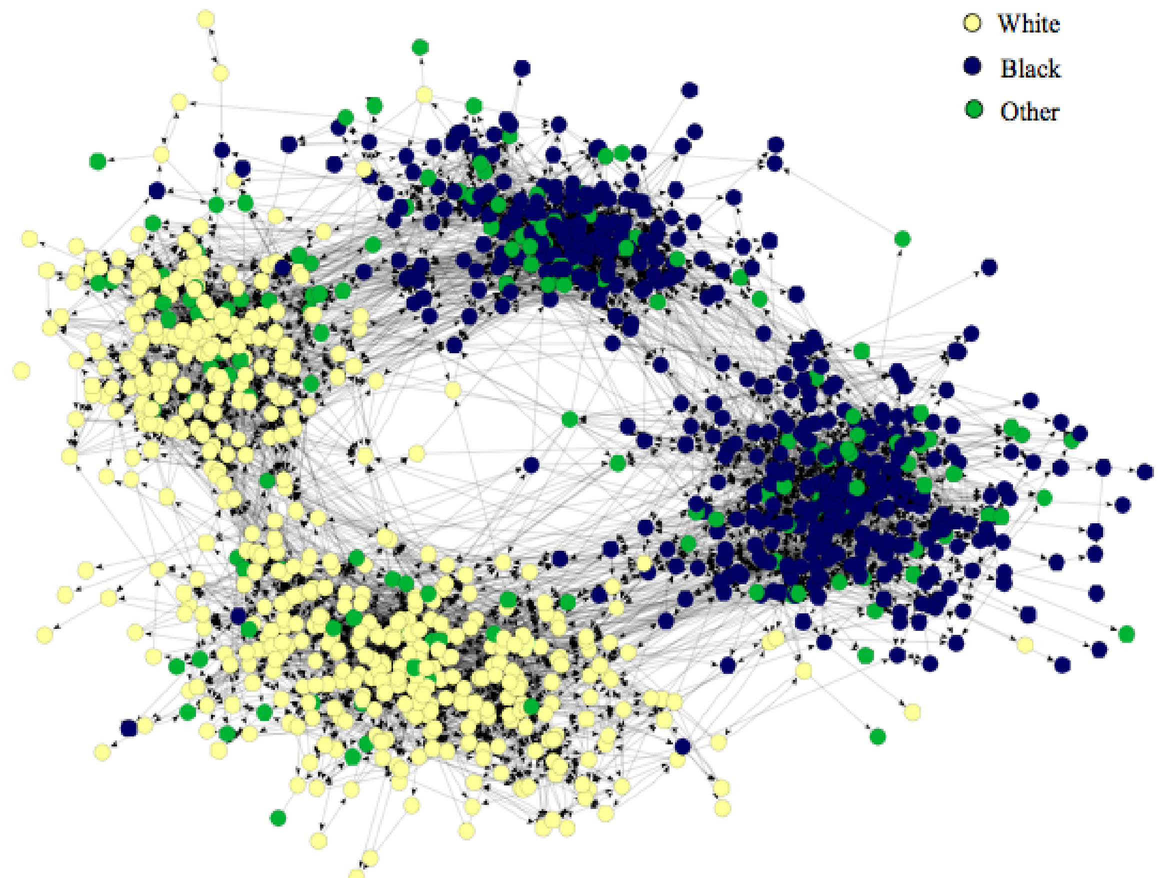
Network 1: 8   Network 2: ≈40

Assortative Mixing

In assortatively mixed networks, like vertices tend to connect preferentially to one another.

Assortative mixing by traits:

Friendship network of students in a U.S. school. Vertices are color coded according to race, as marked, and the split from left to right in the figure is clearly primarily along lines of race. The split from top to bottom reflects a division between middle school and high school students.

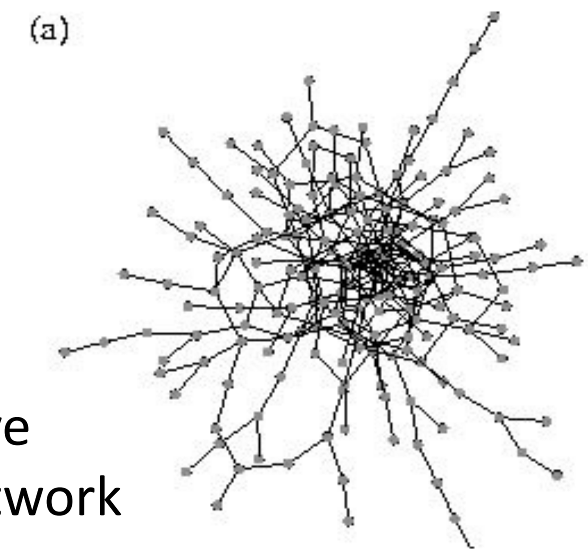


Assortative Mixing by Degree

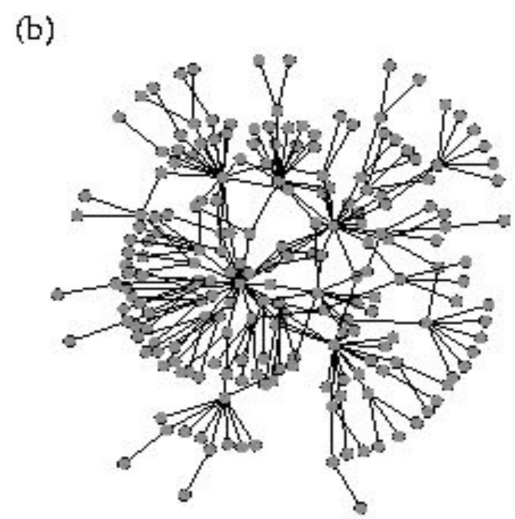
- A network is said to be assortatively mixed by degree if high degree vertices tend to connect preferentially to other high degree vertices
- A network is disassortatively mixed by degree if high degree vertices tend to connect preferentially to low degree vertices.
- Assortativity by degree is measured by the assortativity coefficient, r, which is the Pearson correlation coefficient of the excess degrees of the two nodes at the end of an edge:

$$r = \frac{\langle ij \rangle - \langle i \rangle \langle j \rangle}{\sigma_i^2}$$

where $\langle \cdot \rangle$ denotes the average over edges



Assortative
Scale-free network



Disassortative
Scale-free network

Measured assortativity for various networks

	network	type	size n	assortativity r
social	physics coauthorship	undirected	52 909	0.363
	biology coauthorship	undirected	1 520 251	0.127
	mathematics coauthorship	undirected	253 339	0.120
	film actor collaborations	undirected	449 913	0.208
	company directors	undirected	7 673	0.276
	email address books	directed	16 881	0.092
technol.	Internet	undirected	10 697	−0.189
	World-Wide Web	directed	269 504	−0.067
	software dependencies	directed	3 162	−0.016
biological	protein interactions	undirected	2 115	−0.156
	metabolic network	undirected	765	−0.240
	neural network	directed	307	−0.226
	marine food web	directed	134	−0.263
	freshwater food web	directed	92	−0.326

M.E.J Newman and M. Girvan, *Mixing Patterns and Community Structure in Networks* (2002).

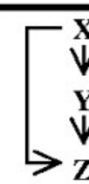
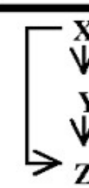
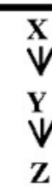
Network Motifs

Motifs:

Subgraphs that have a significantly higher density in the observed network than in the randomizations of the same.

Randomized networks:

Ensemble of maximally random networks preserving the degree distribution (or some other feature(s)) of the original network.

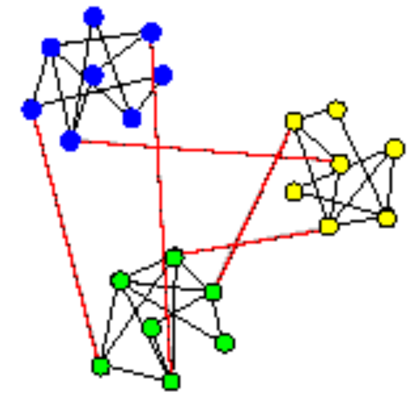
Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$
Gene regulation (transcription)				
E. coli	424	519	40	7 ± 3
*S. cerevisiae**	685	1,052	70	11 ± 4
Neurons				
C. elegans†	252	509	125	90 ± 10
Food webs				
Little Rock	92	984	3219	3120 ± 50
Ythan	83	391	1182	1020 ± 20
St. Martin	42	205	469	450 ± 10
Chesapeake	31	67	80	82 ± 4
Coachella	29	243	279	235 ± 12
Skipwith	25	189	184	150 ± 7
B. Brook	25	104	181	130 ± 7

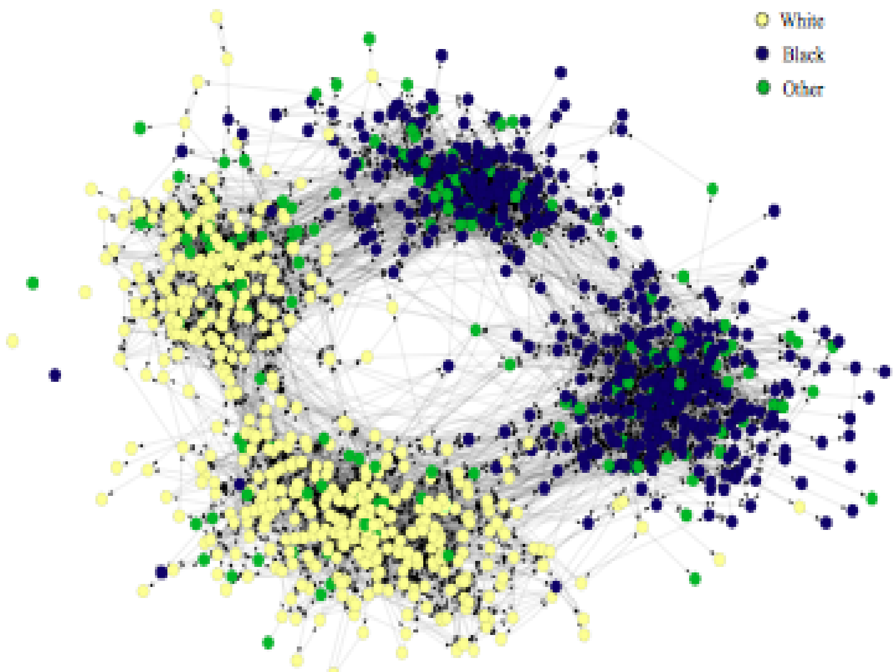
R. Milo et al., Science **298**, 824 (2002)

Configuration Model Random Networks

- Suppose we want to consider networks that have a specified degree distribution, but are otherwise maximally random
- This model of networks is called the configuration model
- To generate configuration model networks:
  - » Start with a degree sequence that lists the degrees, k_i, for each node i.
  - » Put k_i copies of each node i into a 'bin'
  - » Randomly draw pairs of nodes from the bin and connect them by edges
  - » Caveat: this process can generate double edges and self loops

Community Structure in Networks





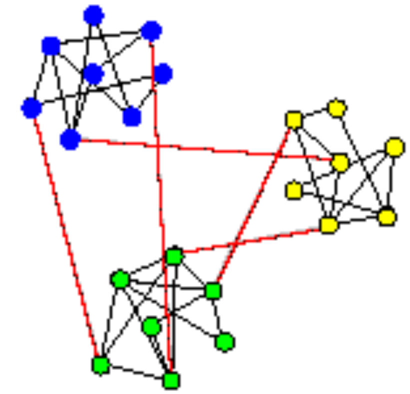
Community structure can arise from assortative mixing by traits

The challenge of community detection:

- The goal: Given an arbitrary network, develop a method to divide the network into groups, or communities, such that within-group edges are relatively dense.
- Important caveat: We do not want to specify the number of groups a priori.
- We can quantify the strength of a given partition of a network into k communities using the modularity function.

References:
Girvan and Newman, PNAS 2002
Newman and Girvan, PRE 2004
Review: Porter et al., AMS, 2009

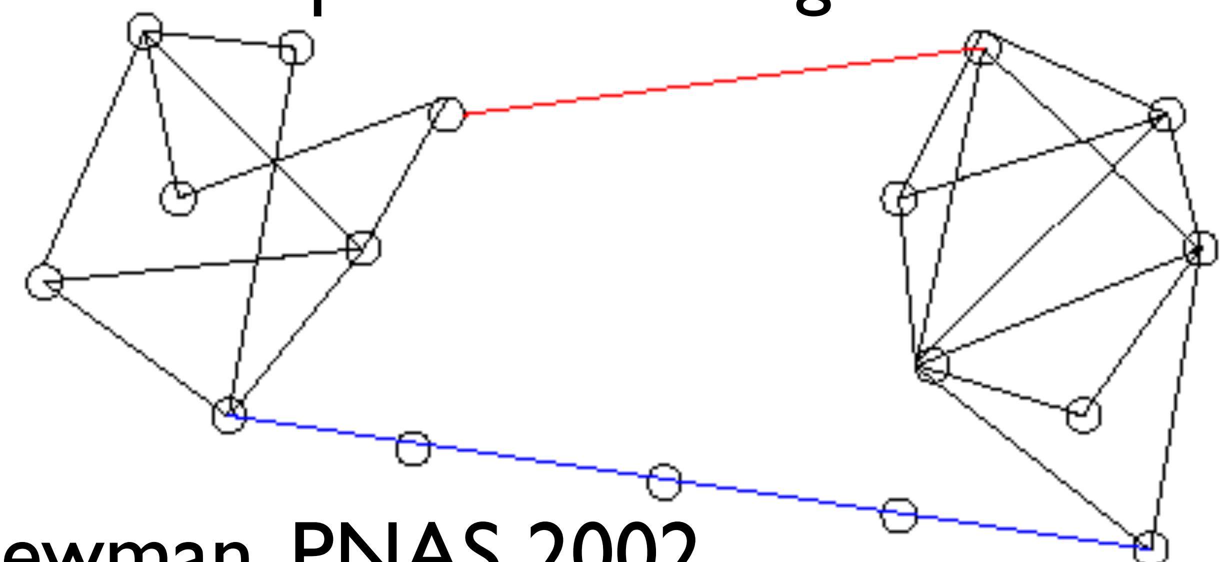
Betweenness-based Method for Detecting Communities



- We can try to detect community structure by identifying boundaries.
- Boundary detection based on centrality indices:
  - Node betweenness: The betweenness centrality of a vertex i is the number of shortest paths between pairs of other vertices which run through i.
  - Edge betweenness: Similarly, the betweenness of an edge j is the number of shortest paths between pairs of nodes which run along j.

Algorithm for Detecting Communities

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweennesses for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.



Girvan and Newman, PNAS 2002

Quantifying community structure

- **Technical details:** The strength of a given partition of a network into *k* communities can be quantified by the modularity function:

$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

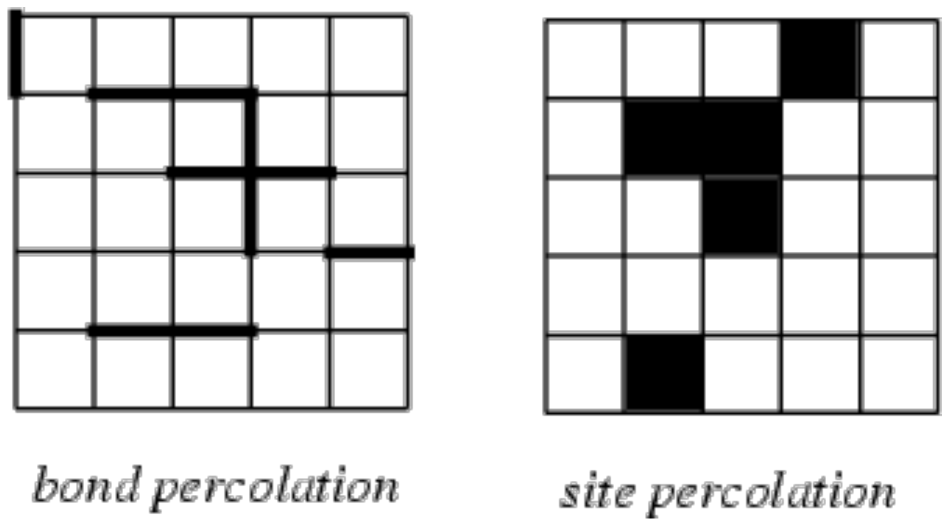
where e_i is the number of edges that connect vertices in community *i*, d_i is the number of edge ends that connect to vertices in community *i*, and *m* is the total number of edges.

- **Layperson description:** The modularity measures observed within-community density vs. expected within community density.

- Many community identification algorithms work by finding the partition of the network that maximizes the modularity.

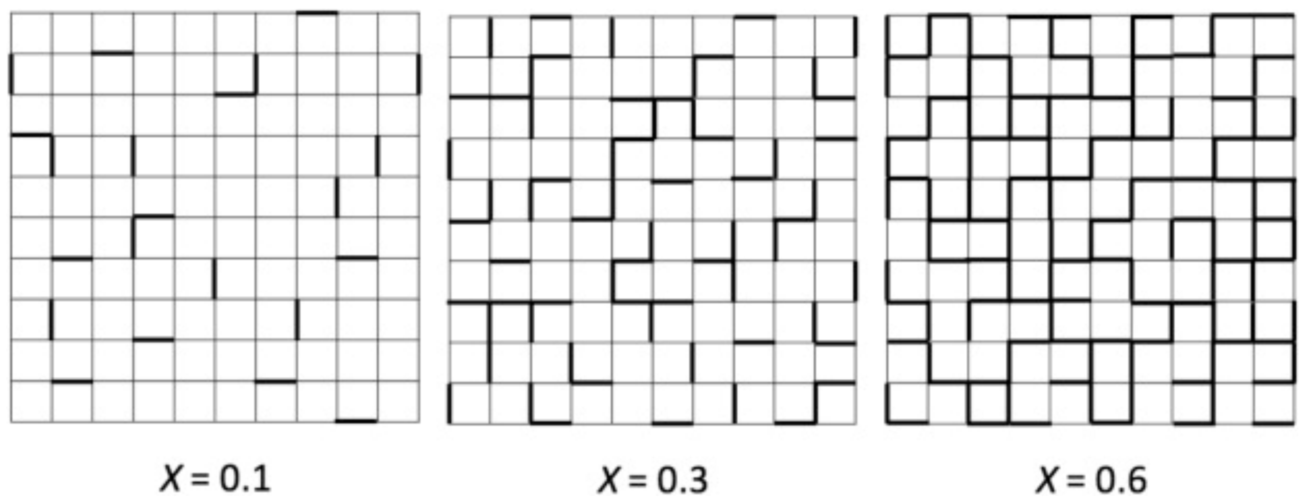
Newman and Girvan, PRE 2004

Connecting Structure to Dynamics: Percolation as a simple example

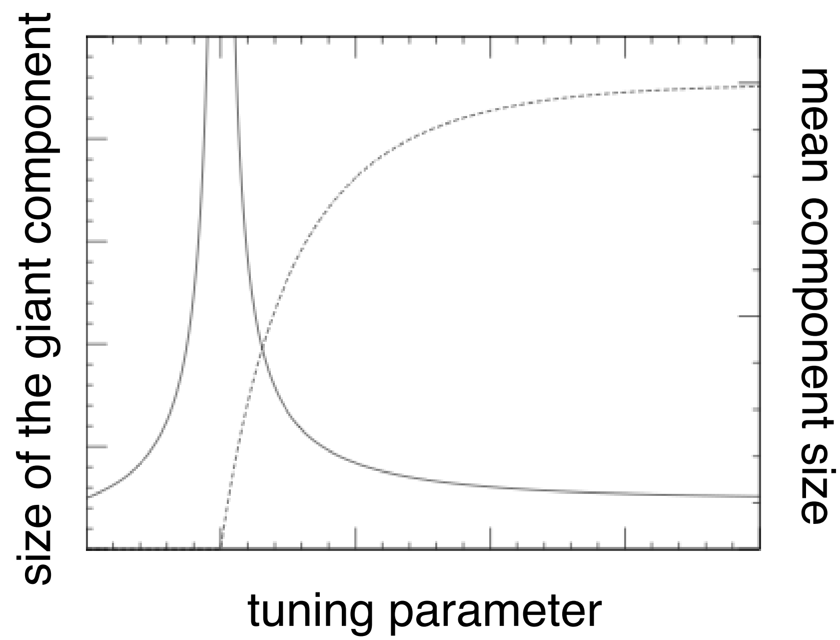


Ordinary Percolation on Lattices: Fill in each link (bond percolation) or site (site percolation) with probability p and ask questions about the sizes of connected components.

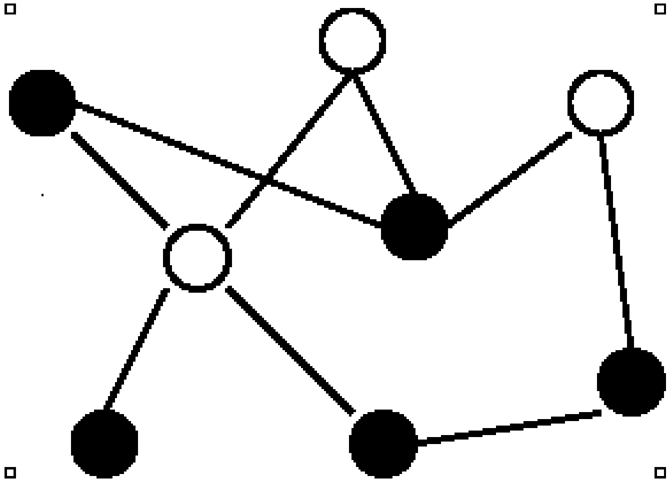
Q: What happens as we increase the occupation probability?



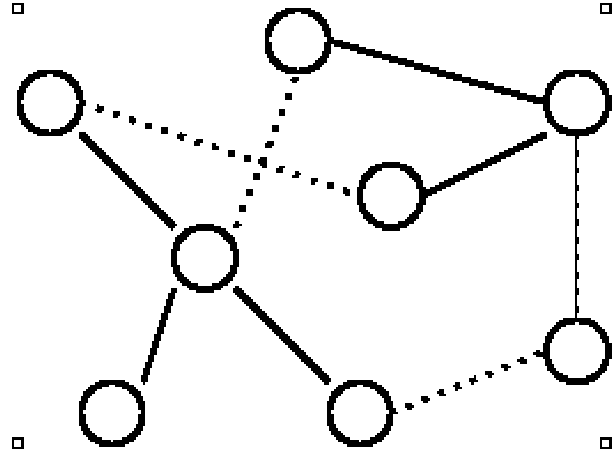
- For low values of X, we see small islands of connected components.
- At a critical value of p, a giant component forms. A giant component is a connected component that occupies a finite fraction of the system, in the limit of infinite system size. At the critical point, there is a power law distribution of the size of connected components.
- Above the critical value, the giant component occupies an increasingly large fraction of the system. If we look at the mean component size excluding the giant component, we observe a characteristic component size.



Percolation on Complex Networks



Site Percolation



Bond Percolation

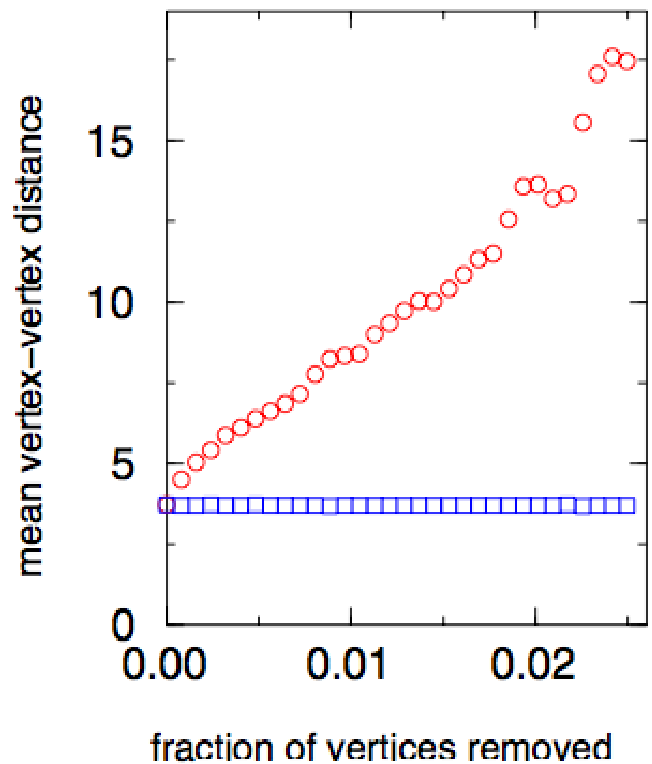
- Percolation can be extended to networks of arbitrary topology.
- We say the network percolates when a giant component forms.
- Simple models of disease spread can be mapped to bond percolation.

How does percolation relate to network resilience?

- We consider the resilience of the network to the removal of its vertices (site percolation) or edges (bond percolation).
- As vertices (or edges) are removed from the network, the average path length will increase.
- Ultimately, the giant component will disintegrate.
- Networks vary according to their level of resilience to vertex (or edge) removal.

Robustness and fragility of scale free networks

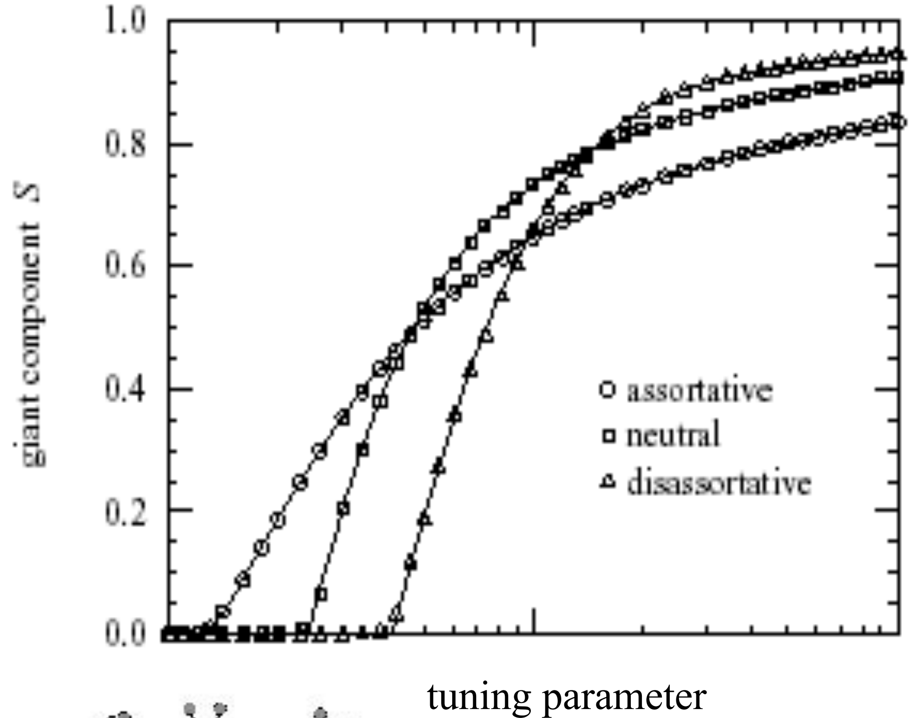
Mean vertex–vertex distance on a graph representation of the Internet at the autonomous system level, as vertices are removed one by one. If vertices are removed in random order (squares), distance increases only very slightly, but if they are removed in order of their degrees, starting with the highest degree vertices(circles), then distance increases sharply. We say the network is resilient to random removal of vertices, but sensitive to targeted removal.

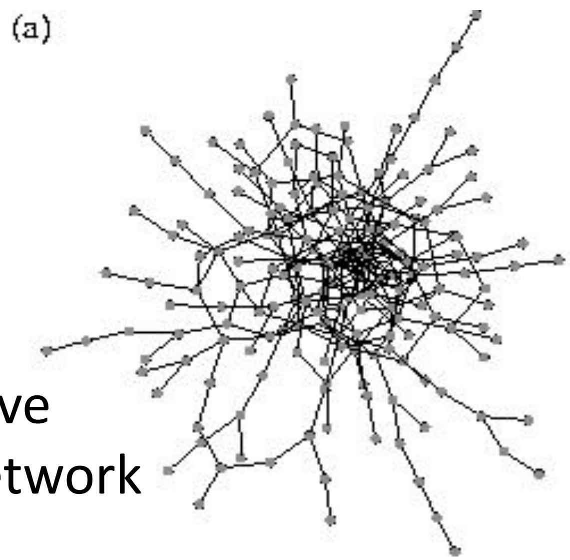


R. Albert, H. Jeong, and A.-L. Barabasi, *Attack and error tolerance of complex networks*, Nature (2000).

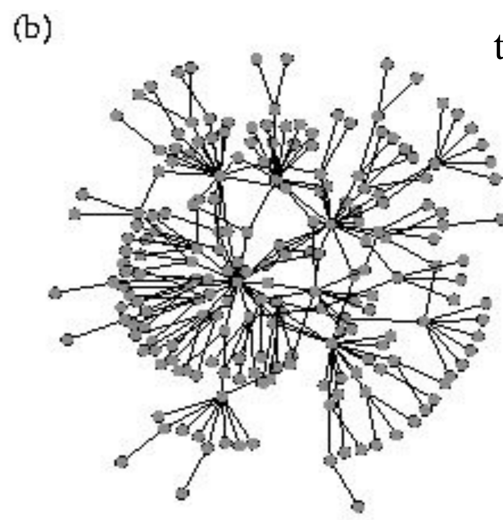
How does assortative mixing affect the percolation transition?

- Assortatively mixed networks exhibit an earlier transition to percolation as the tuning parameter is increased
- Well into the percolating regime, the size of the giant component is larger for disassortative compared to assortative networks





Assortative
Scale-free network



Disassortative
Scale-free network

M. E. J. Newman, *Mixing patterns in networks,* PRE (2003)

Recap of Key Concepts

- **Average path length:** avg. number of links in the shortest path (geodesic) between pairs of nodes
- **Clustering:** Probability that two of your friends of themselves connected; triangle closure
- **Degree of node:** number of links connected to the node
- **Degree distribution:** distribution of the node degrees across the system
- **Assortativity (disassortativity):** Propensity for nodes with similar (different) degrees to connect to one another
- **Betweenness of node X:** Number of shortest paths (geodesics) between pairs of nodes that go through node X
- **Communities/modules:** Groups of nodes that are densely connected

Concluding Remarks

- The emerging field of network science demonstrates how characterizing complex connectivity patterns can be key to understanding many systems.
- Foundational work in this area gives us insight into the role of network topology in numerous applications.
- Many open questions remain. Areas of active research include:
    - Uncertainty in networks
    - Control in networks
    - Temporal networks
    - Multiplex networks