Network Analysis

Rayid Ghani

What we're going to cover

- What is network analysis useful for?
- How to create networks?
- How to analyze networks?

Graphs & Networks

- Graphs have been studied by lots of people for a long time
- Networks are typically represented as graphs
- Networks are the underlying structure of many natural phenomena

Karate Club

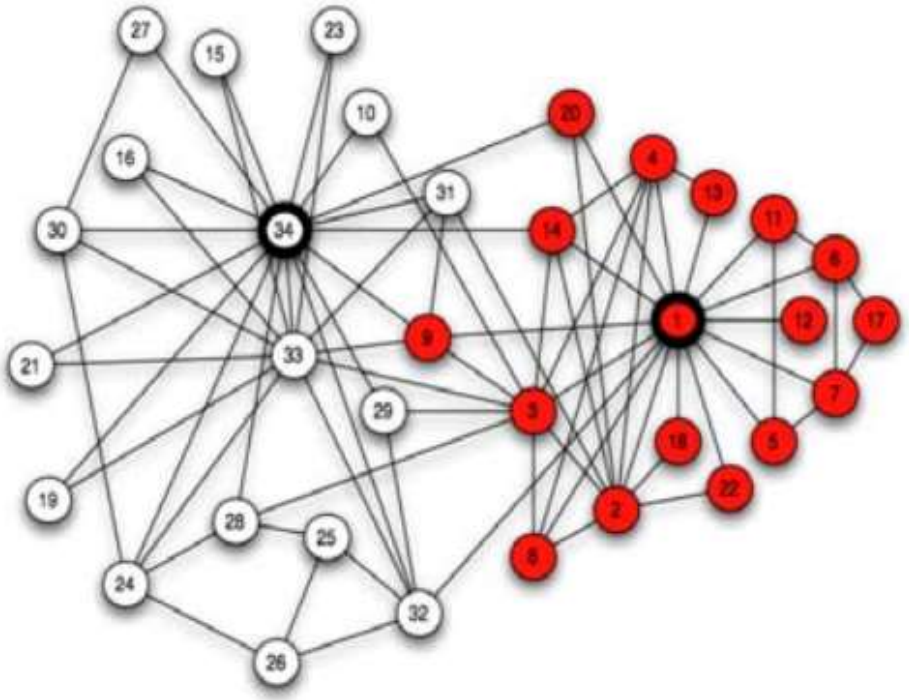
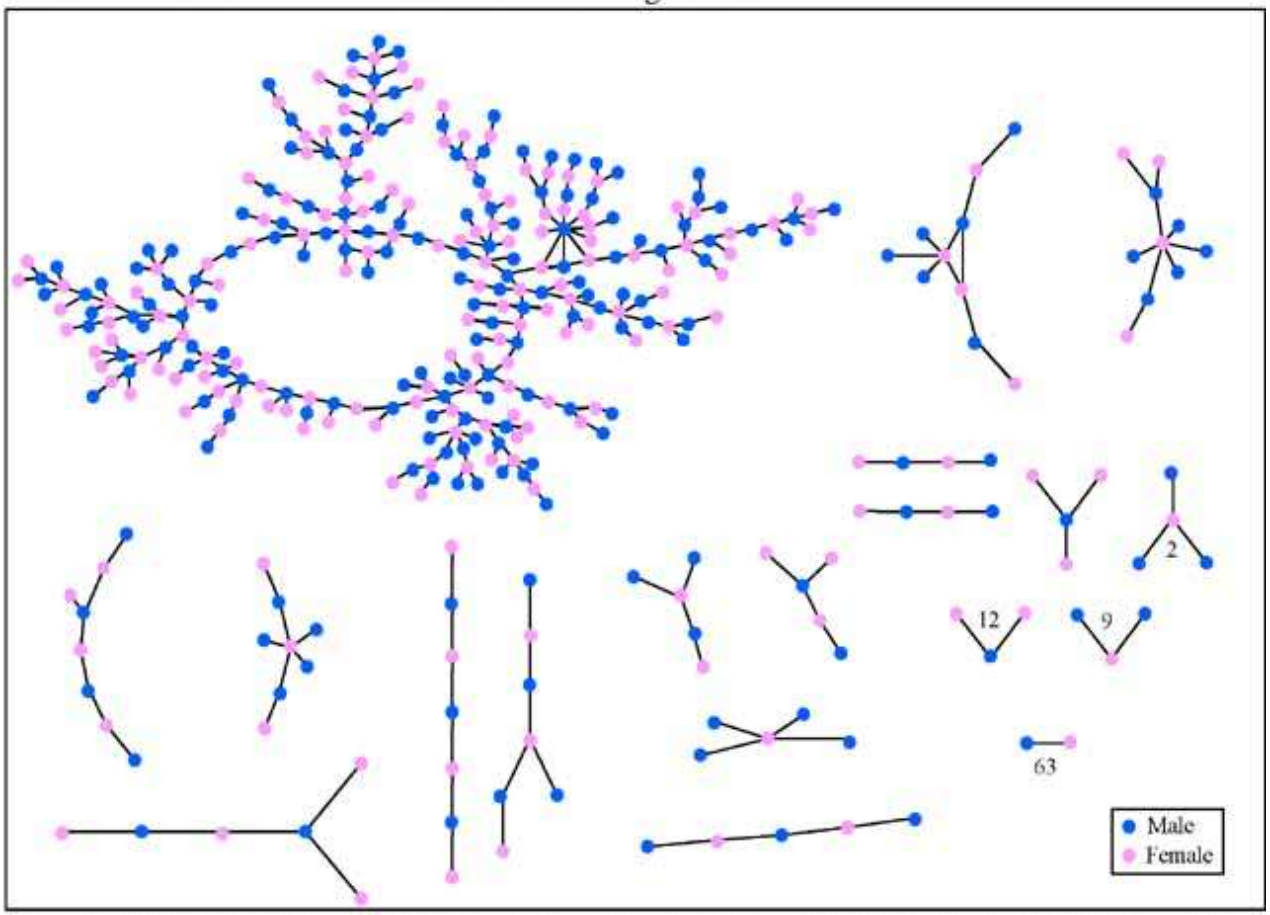


Figure: The social network of friendships within a 34-person karate club provides clues to the fault lines that eventually split the club apart (Zachary, 1977)

Adapted from Figure 1 (p. 456) in Zachary, Wayne W. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 33, no. 4 (1977): 452-473.

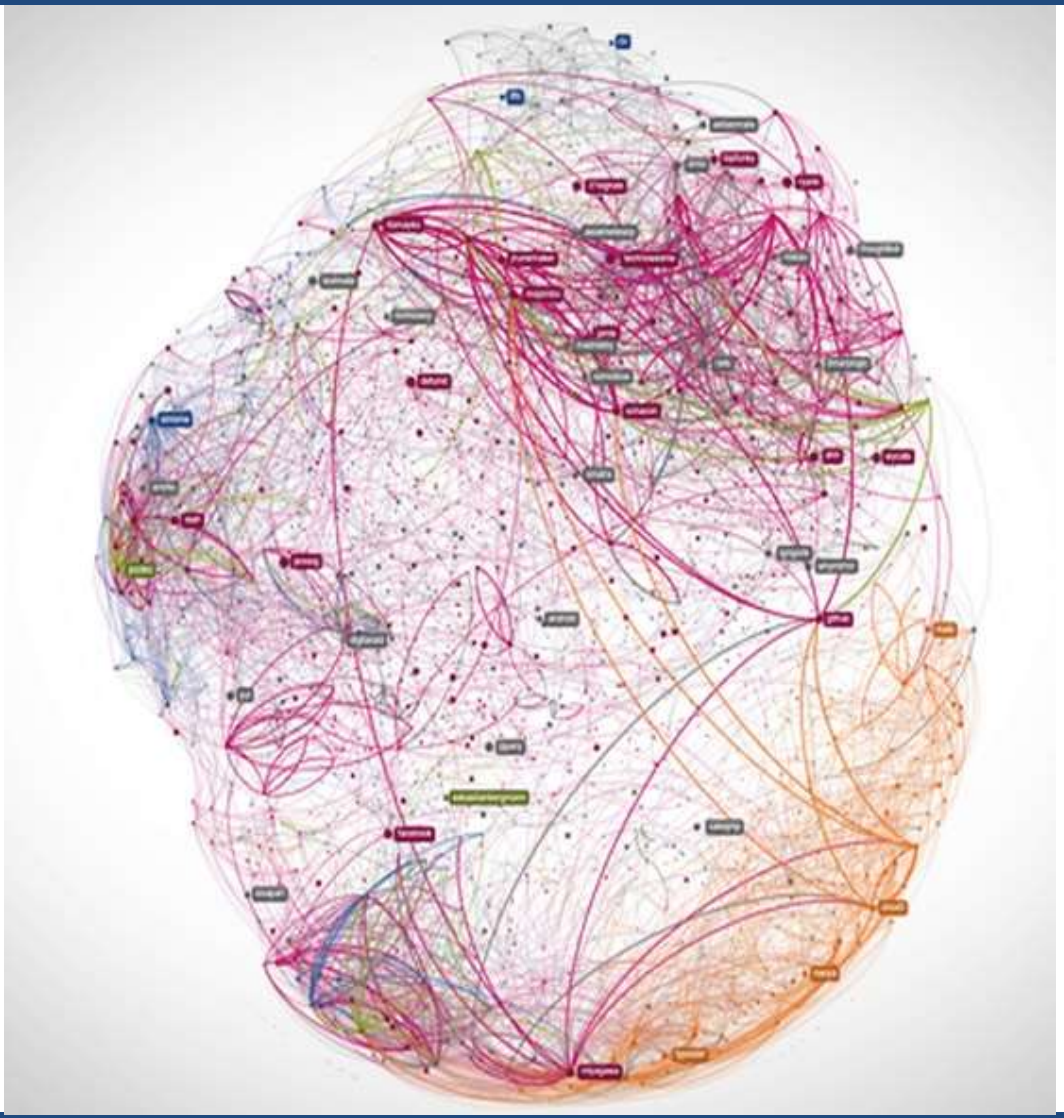
Dating Network

The Structure of Romantic Relations at "Jefferson High School"

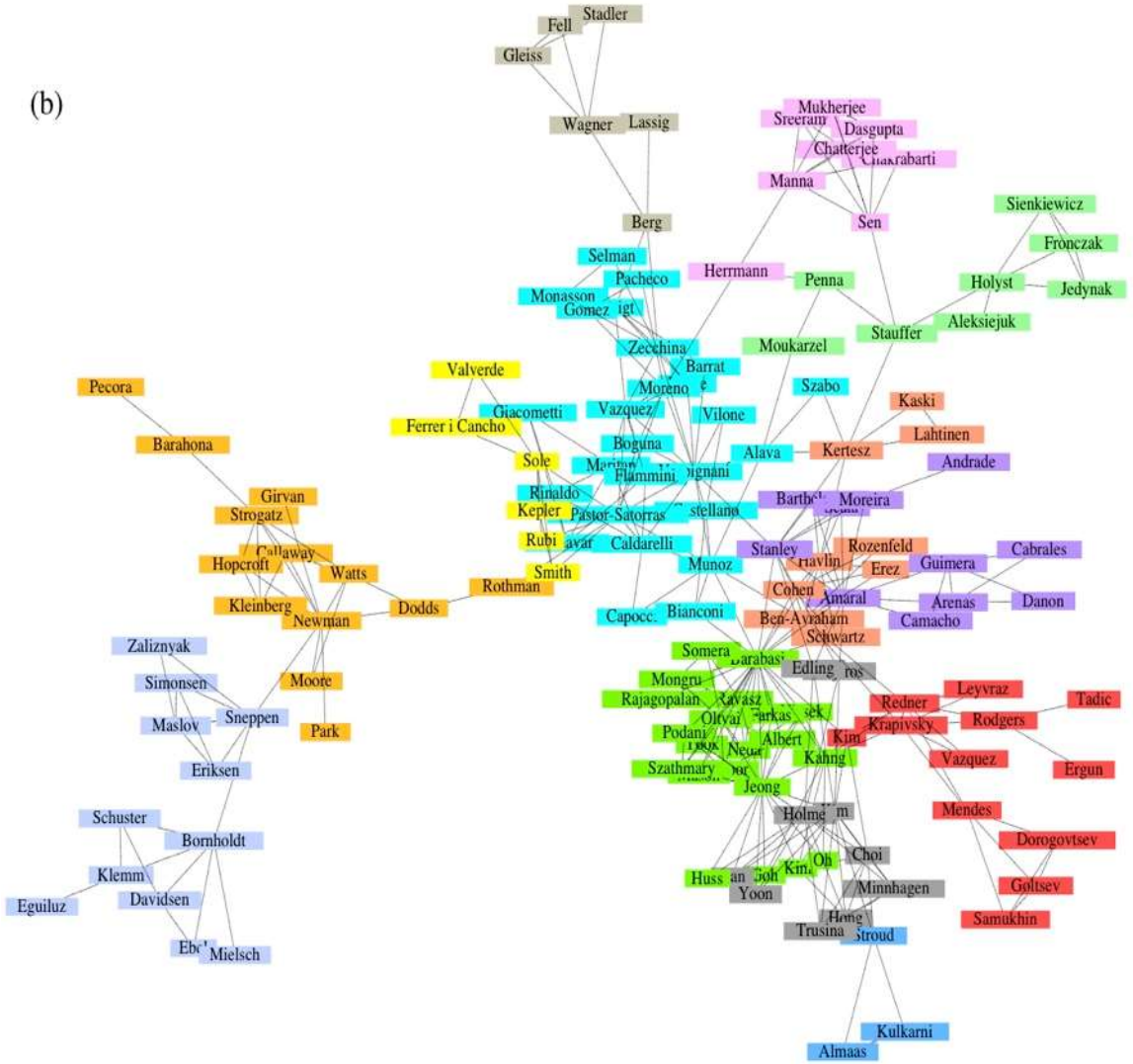


Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

Github collaboration networks



Collaboration Networks



Collaboration Network:
Nodes: Scientists
Links: Joint publications

Physical Review:
1893 – 2009.

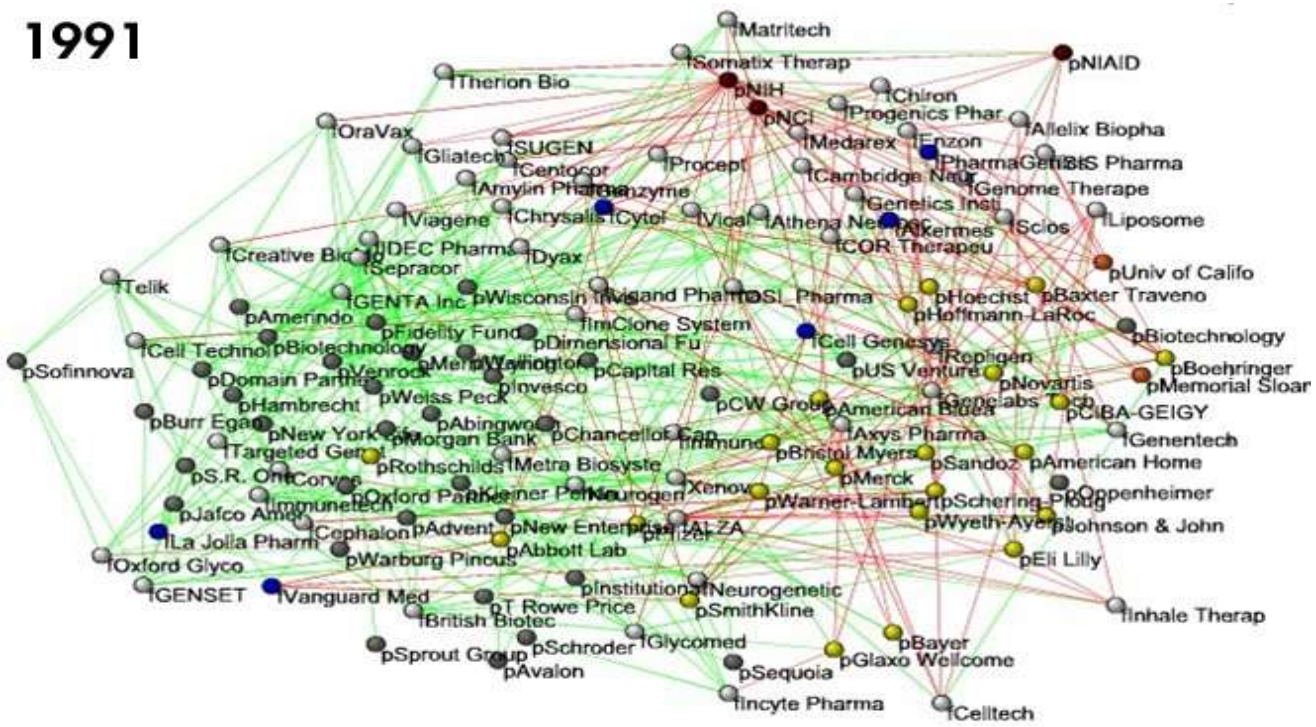
N=449,673
L=4,707,958.

Nodes:

- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

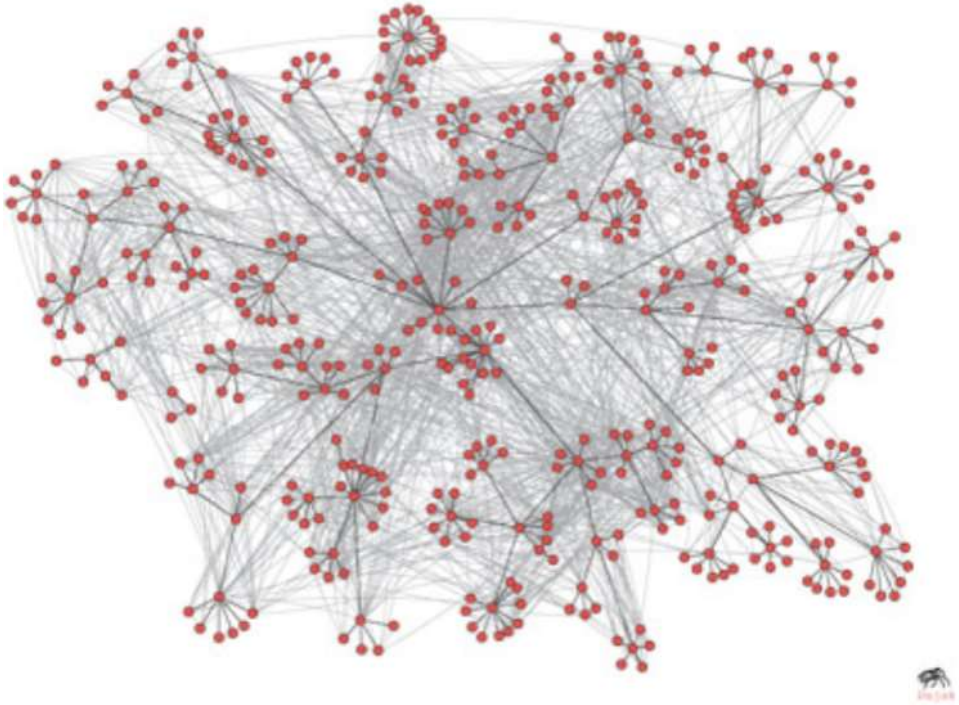
Links:

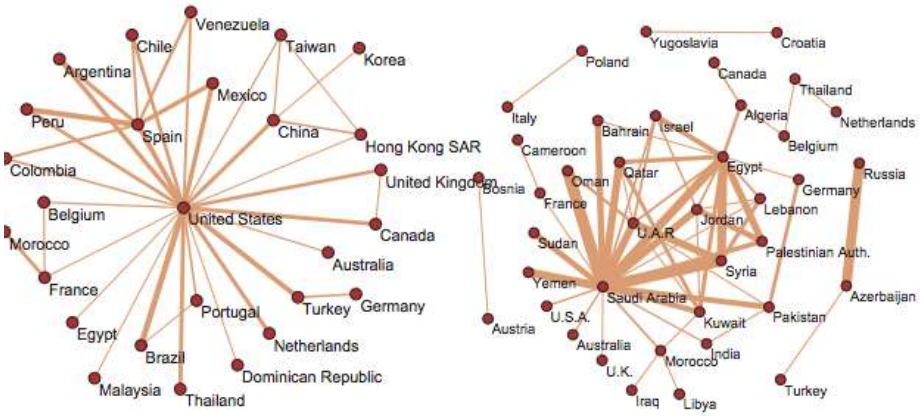
- Collaborations
- Financial
- R&D



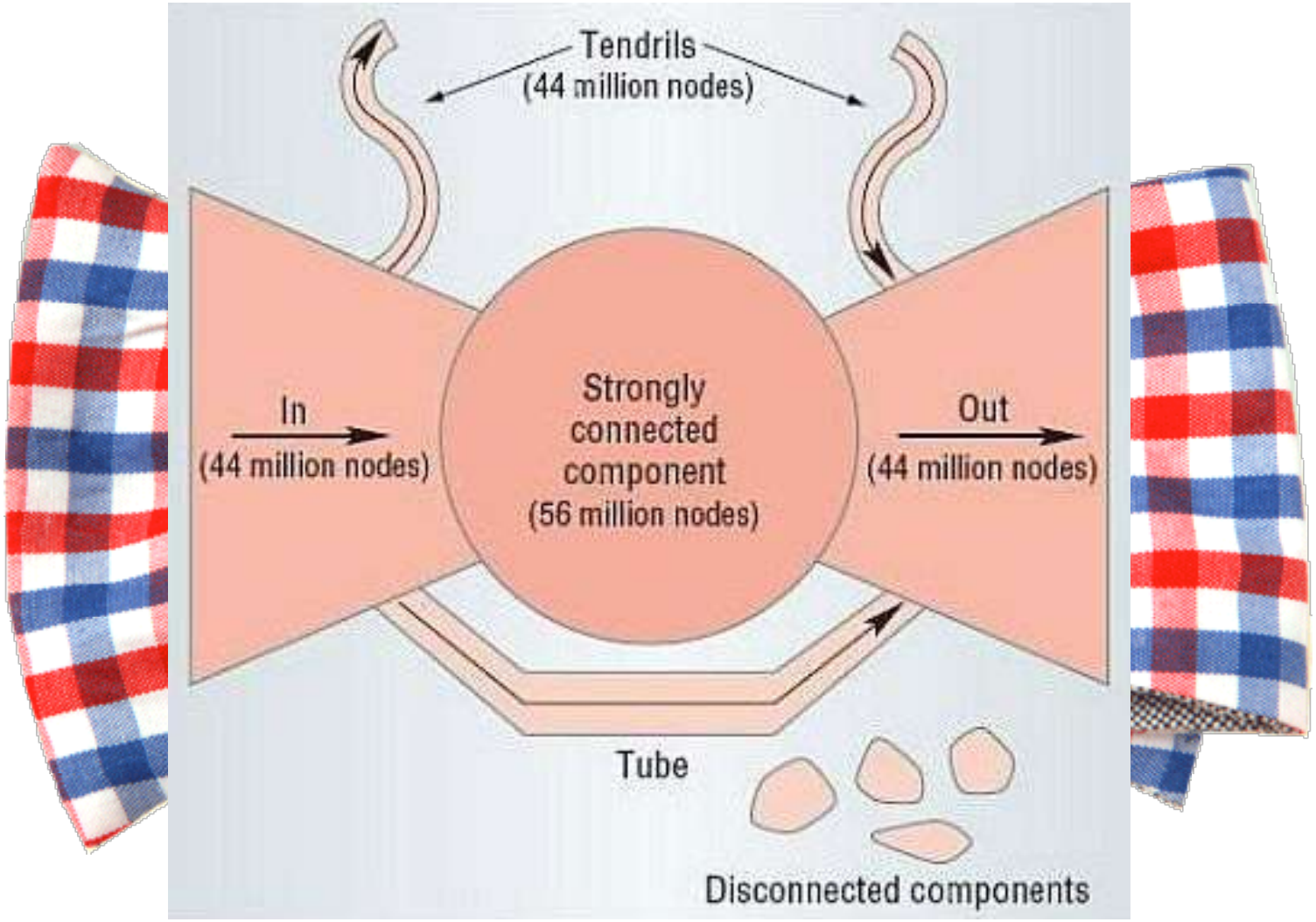
http://ecclectic.ss.uci.edu/~drwhite/Movie

Communication Networks





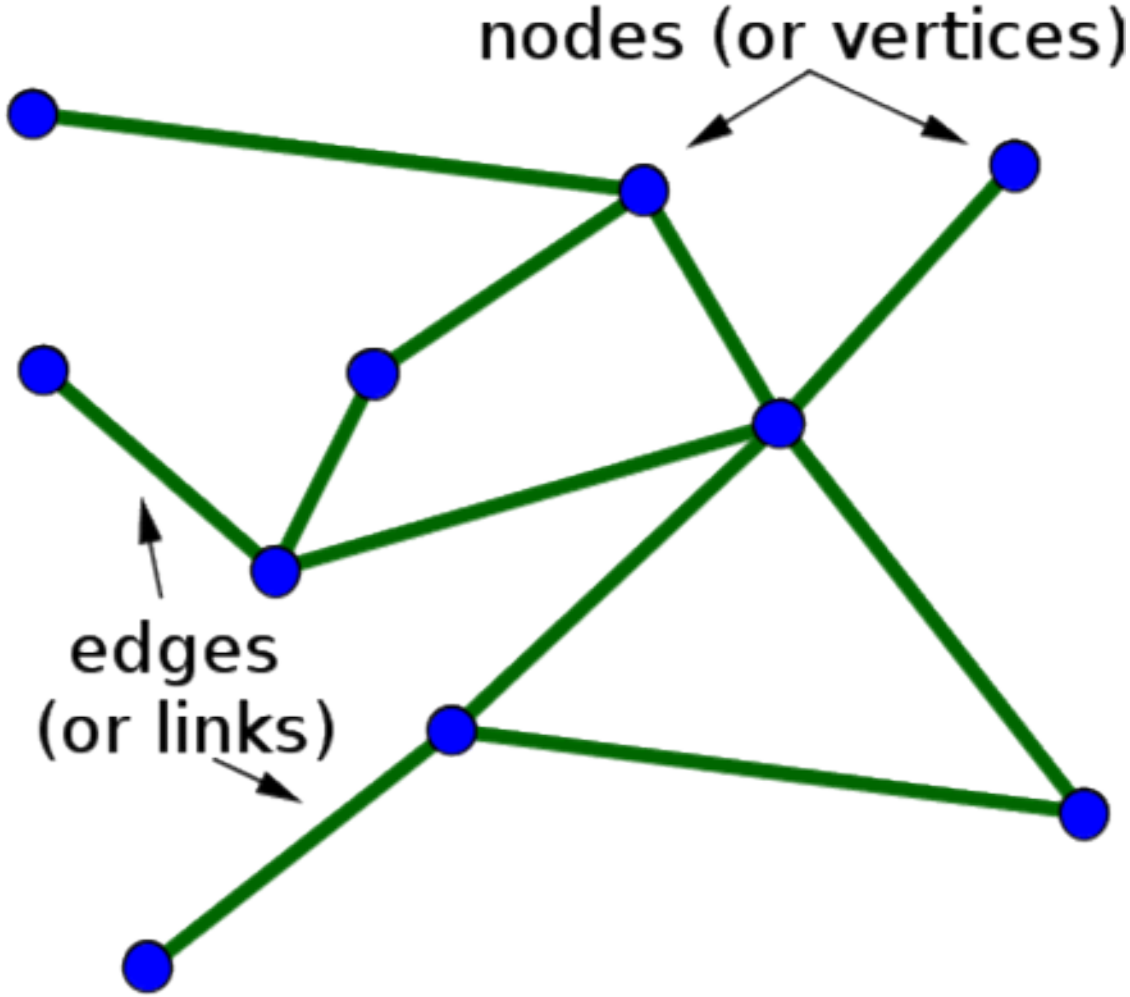
Information Linking Networks



Some questions we can answer

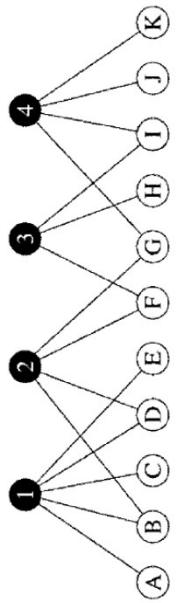
- Does connectedness affect outcomes? For individuals? For groups?
- How do groups evolve, naturally or as a result of a policy change?
- How do you propagate information through a community efficiently?
- How do you prevent propagation through a community?

Networks (graphs)



Types of graphs/networks

- Directed/undirected
- Multi graphs (multiple edges between nodes)
- Hyper graphs (edges connecting multiple nodes)
- Bipartite graphs (e.g., papers to authors)
- Weighted networks
- Different type nodes and edges
- Evolving networks:
  - Nodes and edges only added
  - Nodes, edges added and removed



Graph Vocabulary

- Nodes
- Edges
- Paths: sequence of nodes with each consecutive pair in the sequence is connected by an edge
- Cycles: > 2 edges, in which the first and last nodes are the same, but otherwise all nodes are distinct.
- Directed or undirected
- Adjacency matrix
- Distance: length of the shortest path between 2 nodes
- Weighted/unweighted
- Connected: if for every pair of nodes, there is a path between them
- Connectivity:
- Components
  - Giant Components
- Bipartite

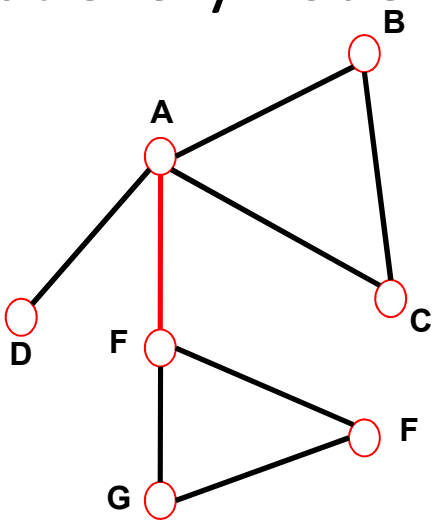
Graph Theory

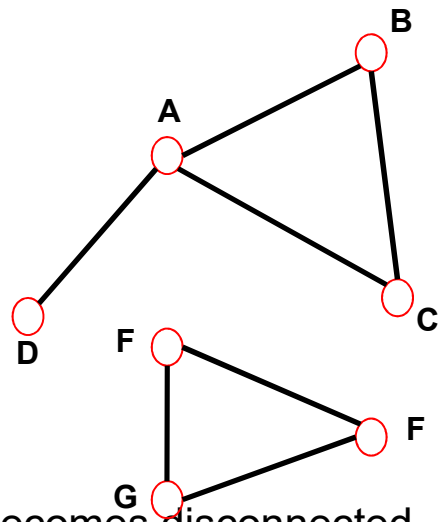
"Terminological jungle, in which any newcomer may plant a tree" (John Barnes)

Connectivity

A Connected Component is a subset of the nodes such that:

- every node in the subset has a path to every other; and
- the subset is not part of some larger set with the property that every node can reach every other.





Largest Component:
Giant Component

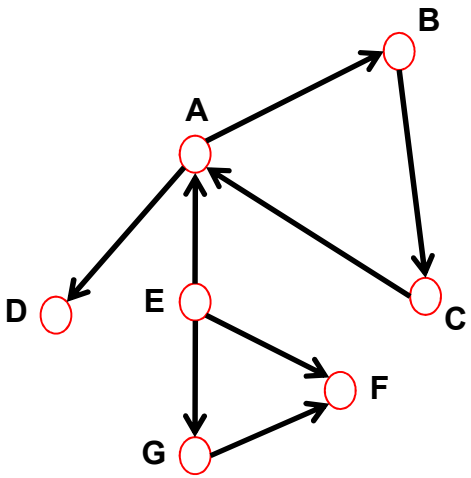
The rest: **Isolates**

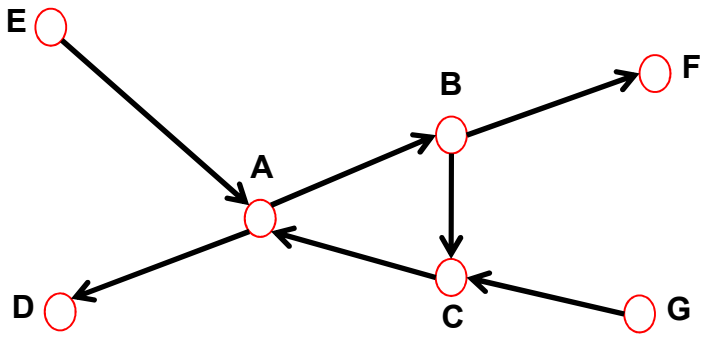
Bridge: if we erase it, the graph becomes disconnected.

Connectivity

Strongly connected directed graph: has a path from each node to every other node and vice versa (e.g. AB path and BA path).

Weakly connected directed graph: it is connected if we disregard the edge directions.

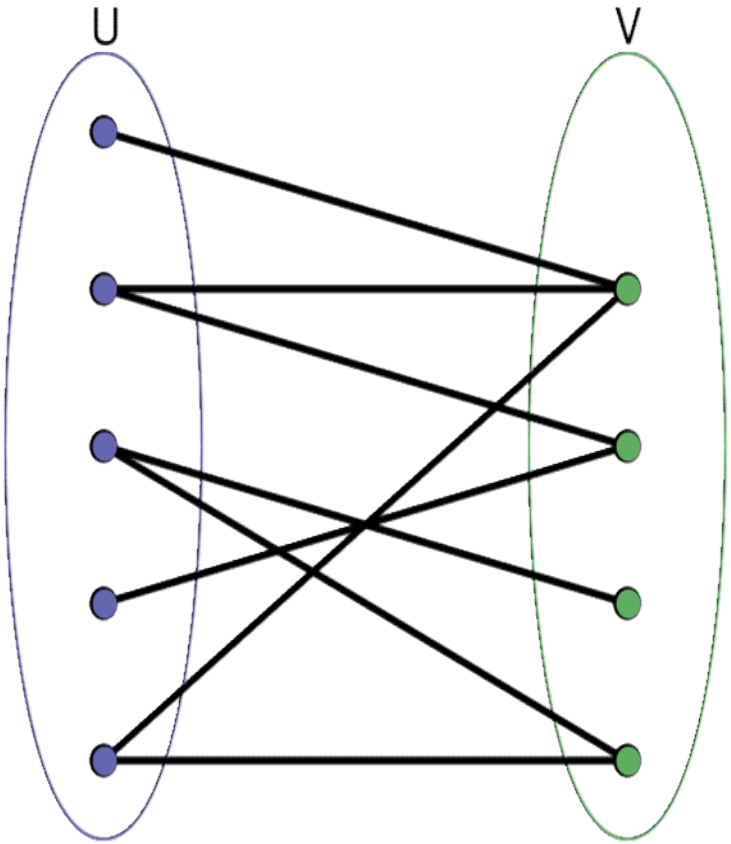




In-component: nodes that can reach the scc,
Out-component: nodes that can be reached from the scc.

Bipartite Graphs

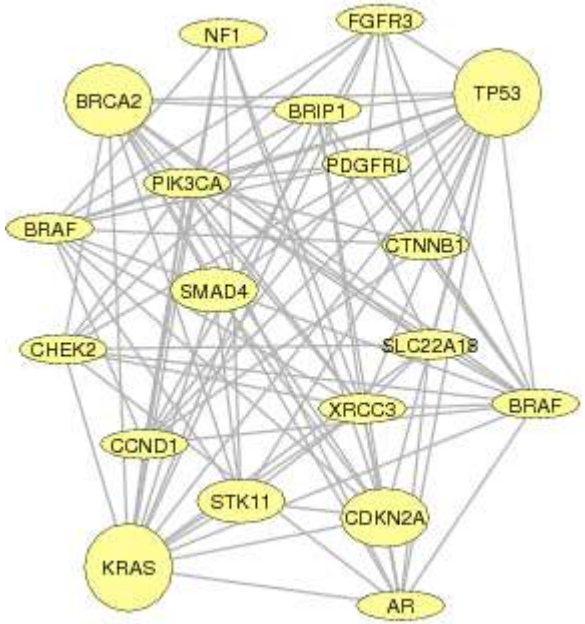
Nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



Examples:

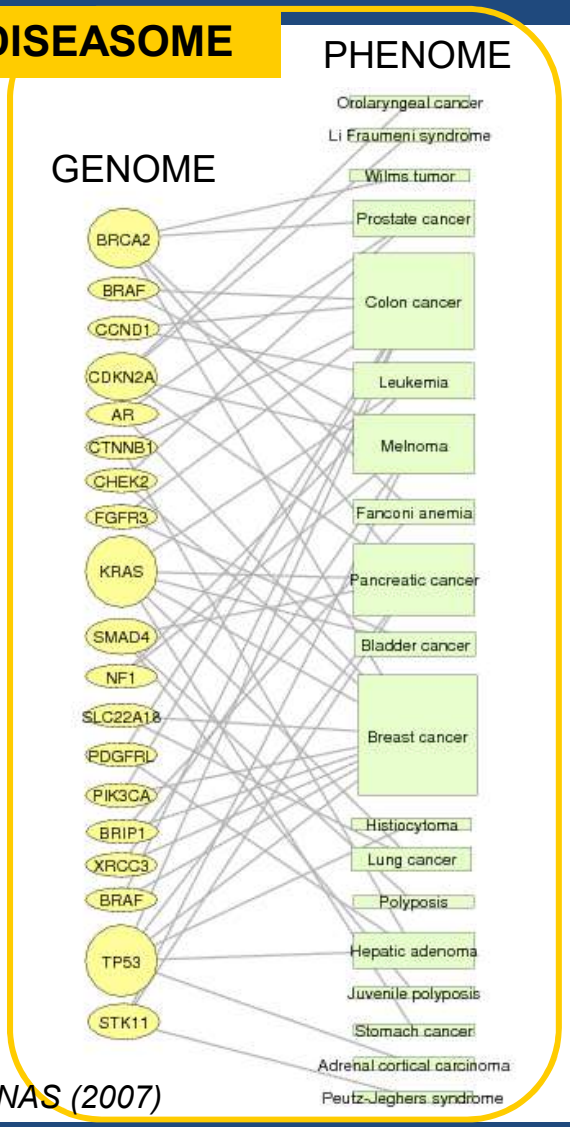
Collaboration networks
Disease network (diseasome)

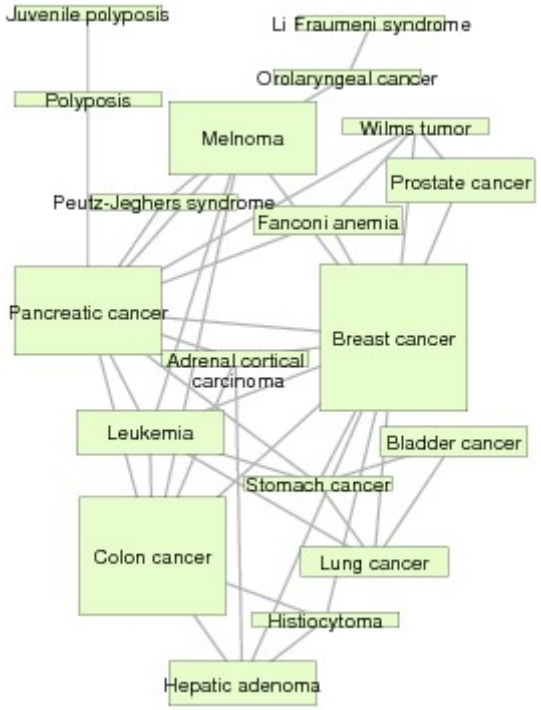
Disease Network



Gene network

DISEASOME





Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Representation: How to turn something into a graph?

- How you turn relationships into a graph defines what you can study, analyze, and infer.

- This is often the most important decision to make

Study of Networks

- Sociologists were first to study networks:
  - Study of patterns of connections between people to understand society
  - Typical questions: Centrality and connectivity
  - Limited to small graphs (~10 nodes) and properties of individual nodes and edges

- **Large** networks (e.g., web, internet, on-line social networks) with millions of nodes

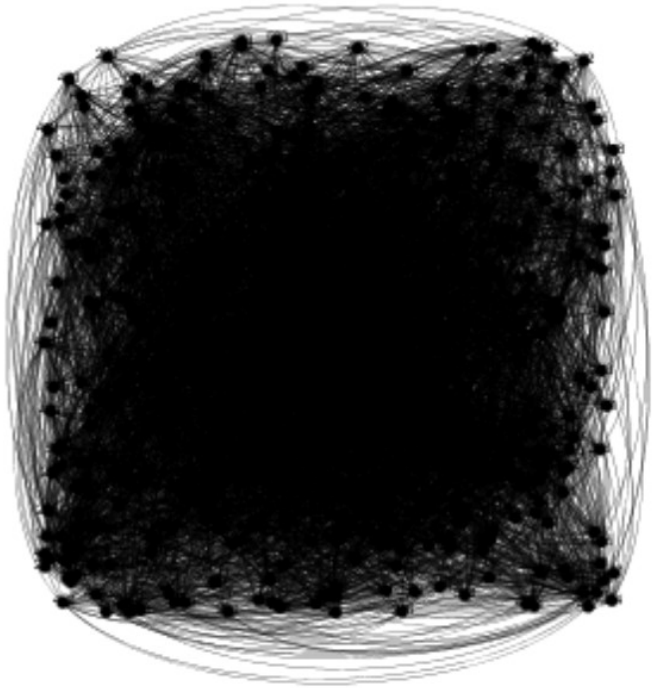
- Many traditional questions not useful anymore:
  - Traditional: What happens if a node *U* is removed?
  - Now: What percentage of nodes needs to be removed to affect network connectivity?

- Focus moves from a single node to study of **statistical** properties of the network as a whole

- Can not draw (plot) the network and examine it

Study of Networks

- What does the network "looks like" even if I can't look at it?
- Statistical methods and tools to quantify large networks

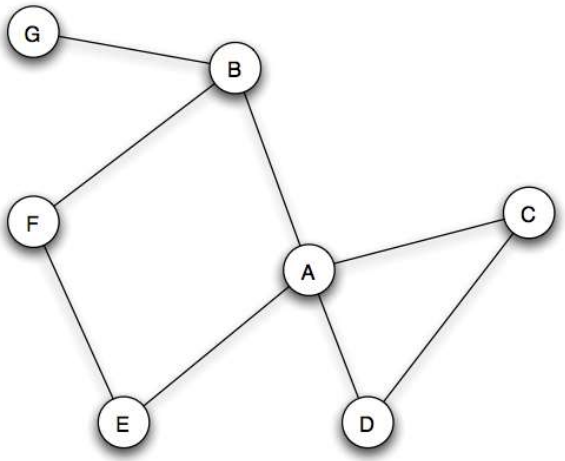


Network Measures

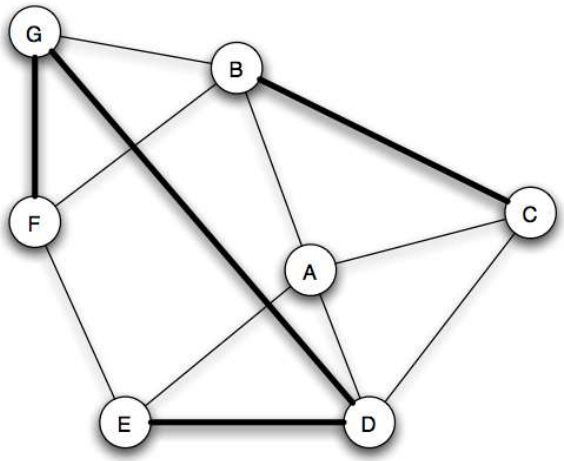
- Degree measures
  - In-degree
  - Out-degree
  - Average
  - Distribution
- Diameter: maximum (shortest) distance between any pair of nodes in the graph
- Average Path length: Avg # steps along the shortest paths of all possible pairs of nodes (efficiency)
- Clustering coefficient (average)

Triads

If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future



(a) *Before new edges form.*

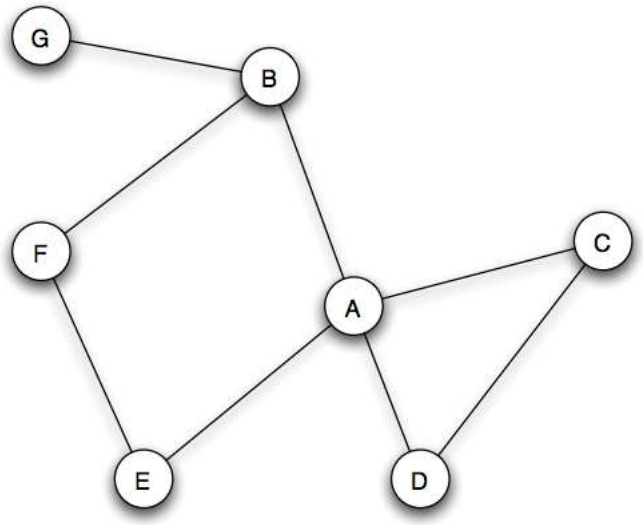


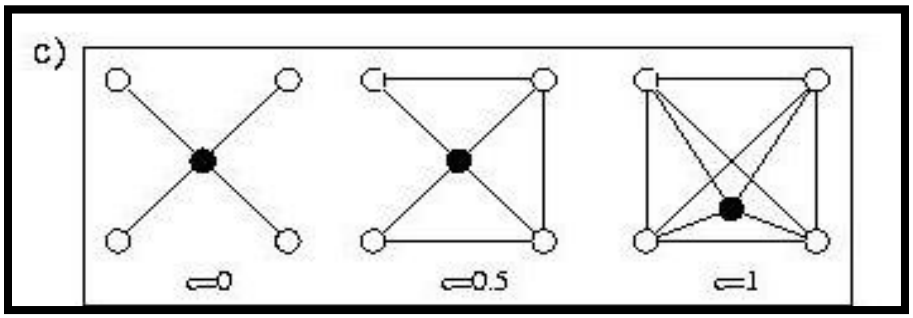
(b) *After new edges form.*

Clustering Coefficient

- **Clustering coefficient: probability that two randomly selected friends of A are friends with each other**

What % of your neighbors are connected to each other?

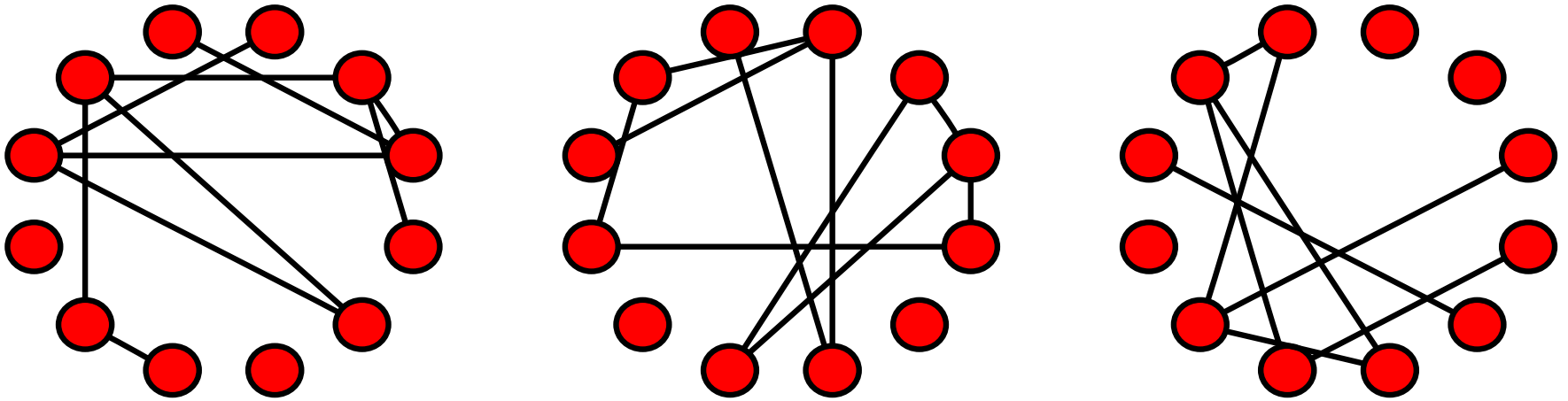




Graph Models

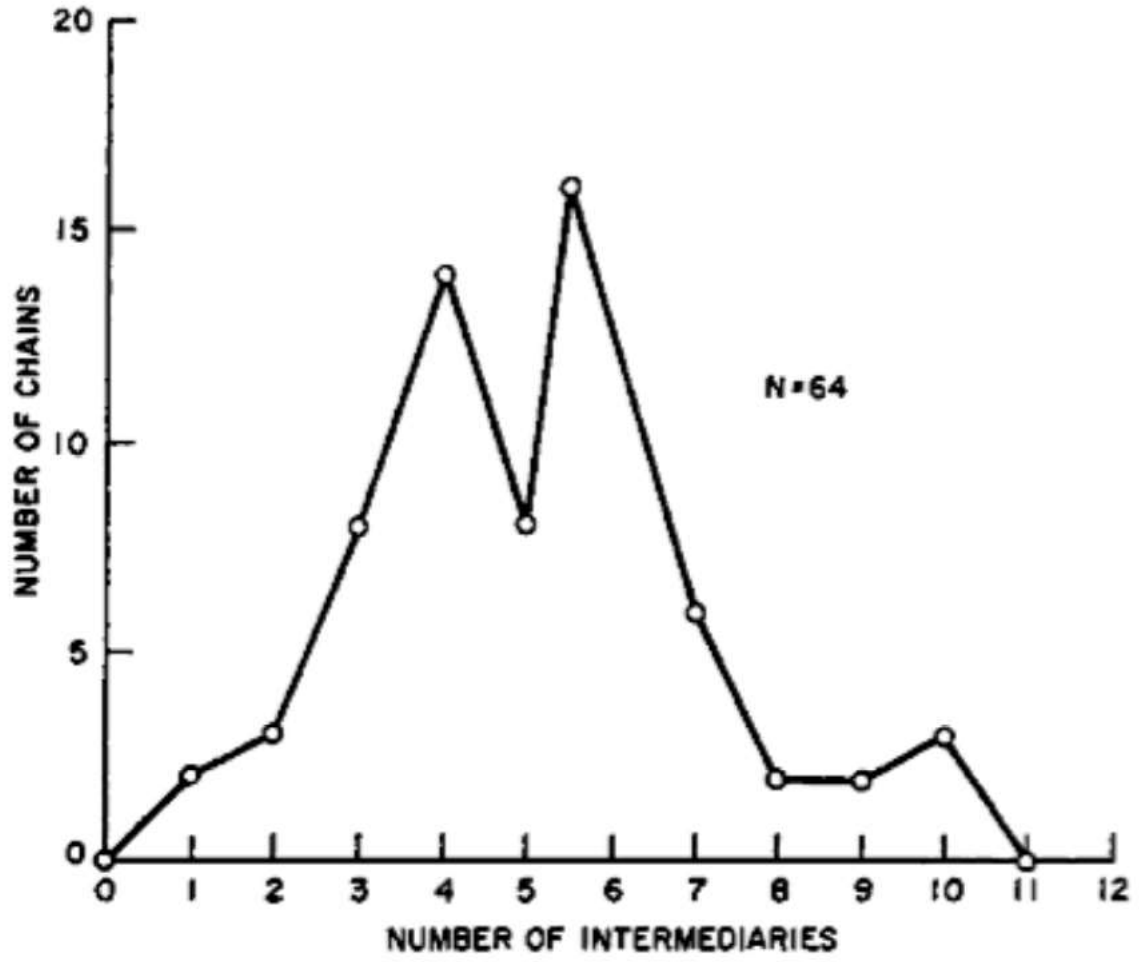
- Random Graph: A **random graph** is a graph of N labeled nodes where each pair of nodes is connected by a preset probability **p**.

p=1/6
N=12



Small World: Six Degrees

- Early 20 ame up with the hat any two people a n were linked th
- The soci Small World P d.
- He aske d a folder to a targ uld then do likewise Milgram to measure k the original
- 42 of th an number of interr



Milgram Experiment

HOW TO TAKE PART IN THIS STUDY

1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

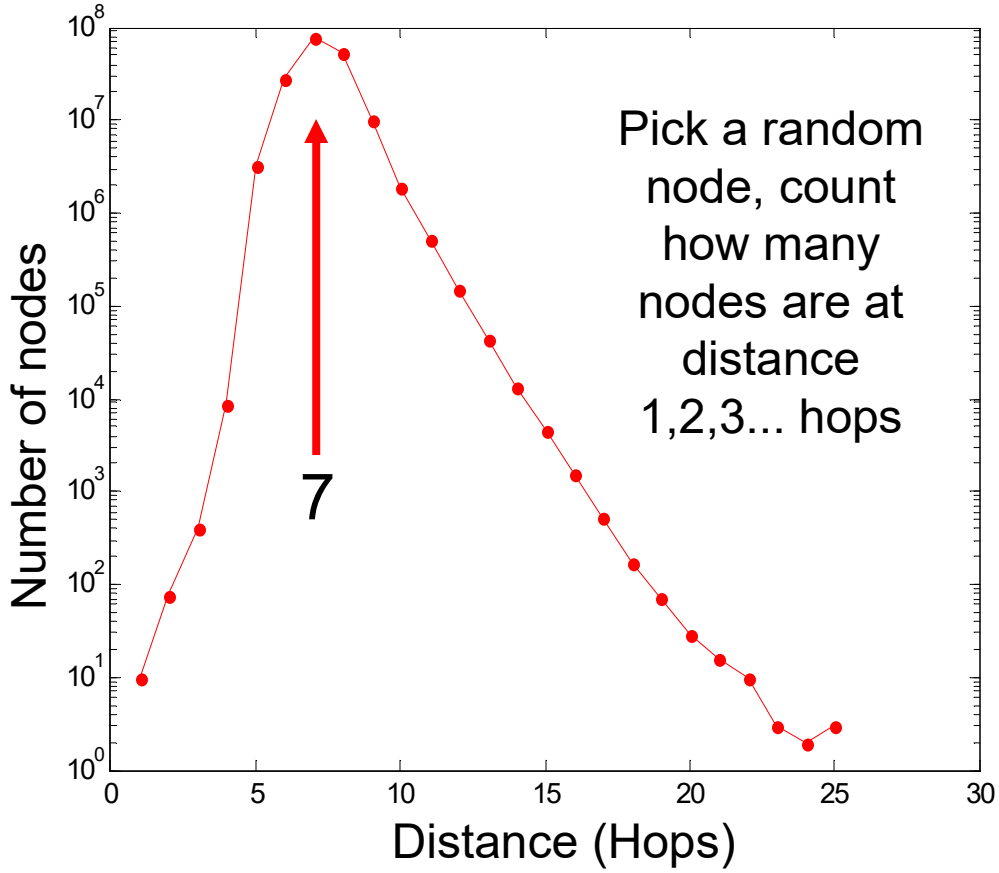
2. DETACH ONE POSTCARD. FILL IT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.

3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

4. IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder to a friend, relative or acquaintance, but it must be someone you know on a first name basis.

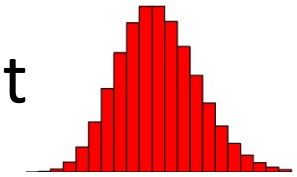
Small-world effect (2)

- Distribution of shortest path lengths
- Microsoft Messenger network
  - 180 million people
  - 1.3 billion edges
  - Edge if two people exchanged at least one message in one month period

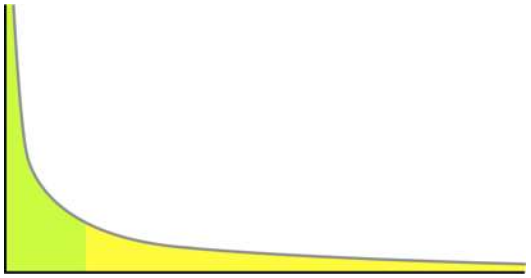


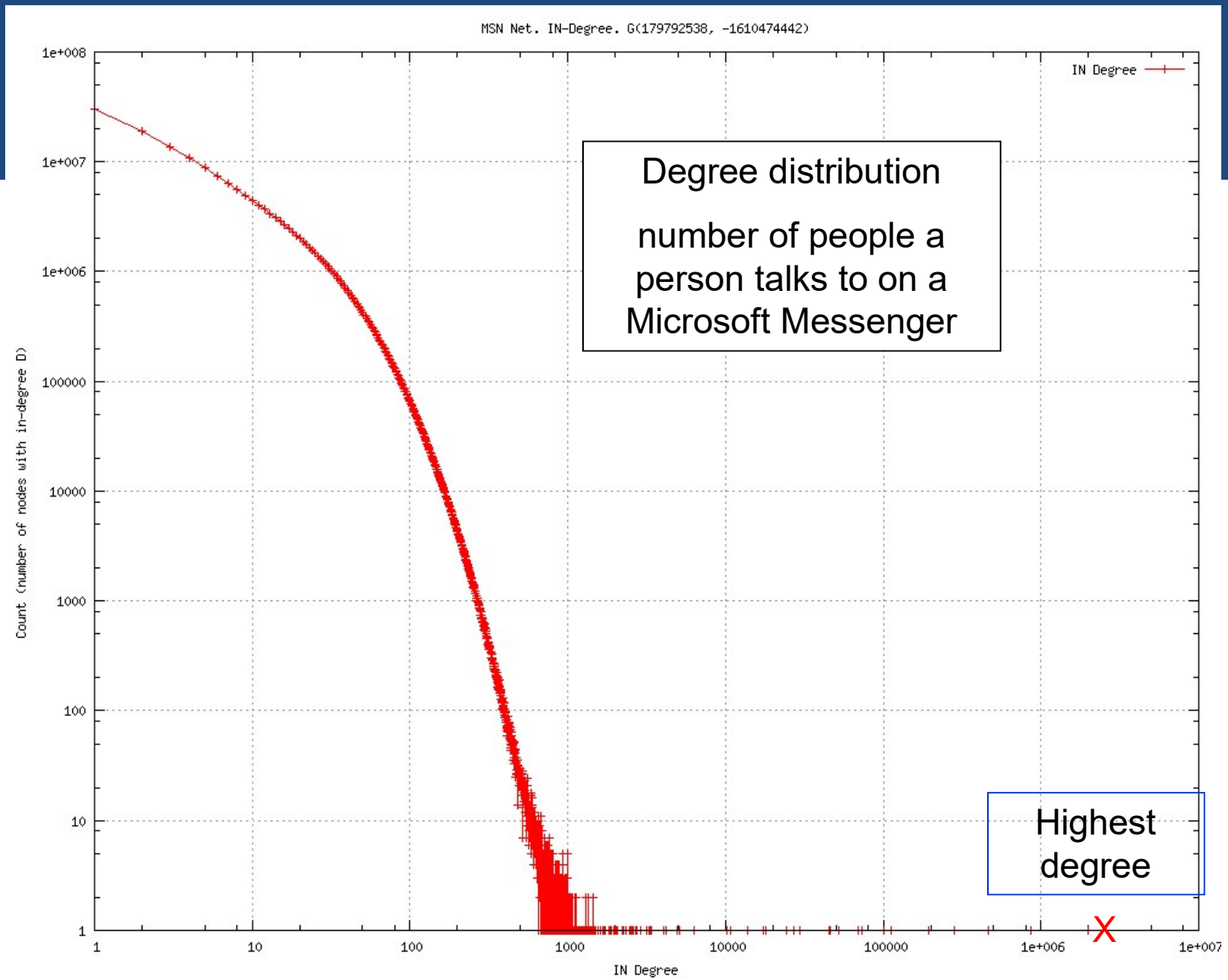
Degree distributions (1)

- Let p_k denote a fraction of nodes with degree k
- We can plot a histogram of p_k vs. k
- In a Erdos-Renyi random graph degree distribut follows Poisson distribution



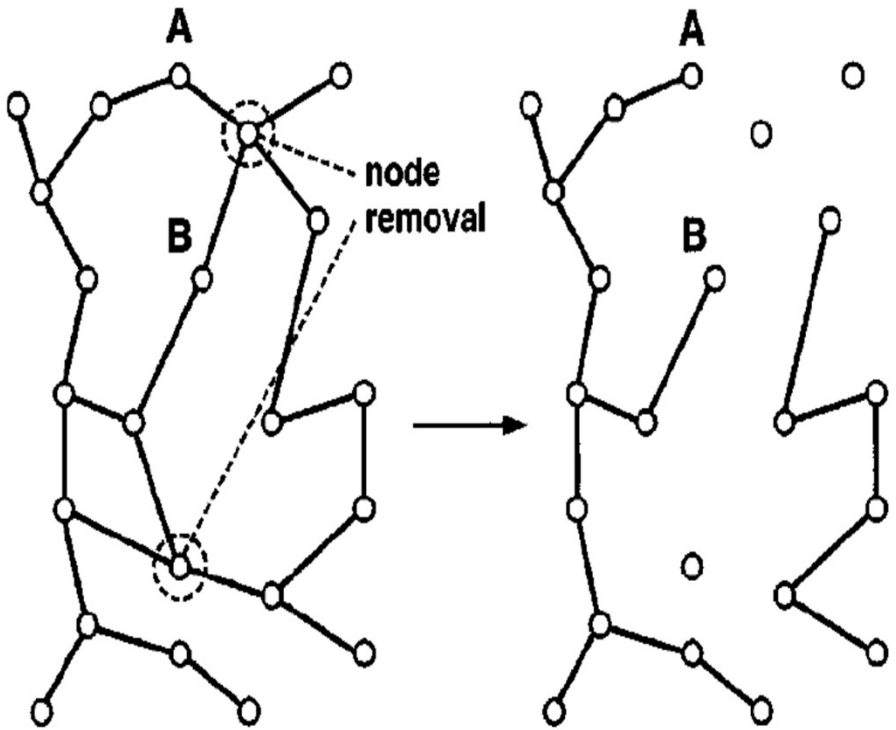
- Degrees in real networks are heavily skewed to the right
- Distribution has a long tail of values that are far above the mean
- Heavy (long) tail:
  - Amazon sales
  - word length distribution, …




MSN Net. IN-Degree. G(179792538, -1610474442)
IN Degree
Count (number of nodes with in-degree D)
1e+008
1e+007
1e+006
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
1e+006
1e+007
IN Degree
Degree distribution
number of people a person talks to on a Microsoft Messenger
Highest degree
Count
Node degree

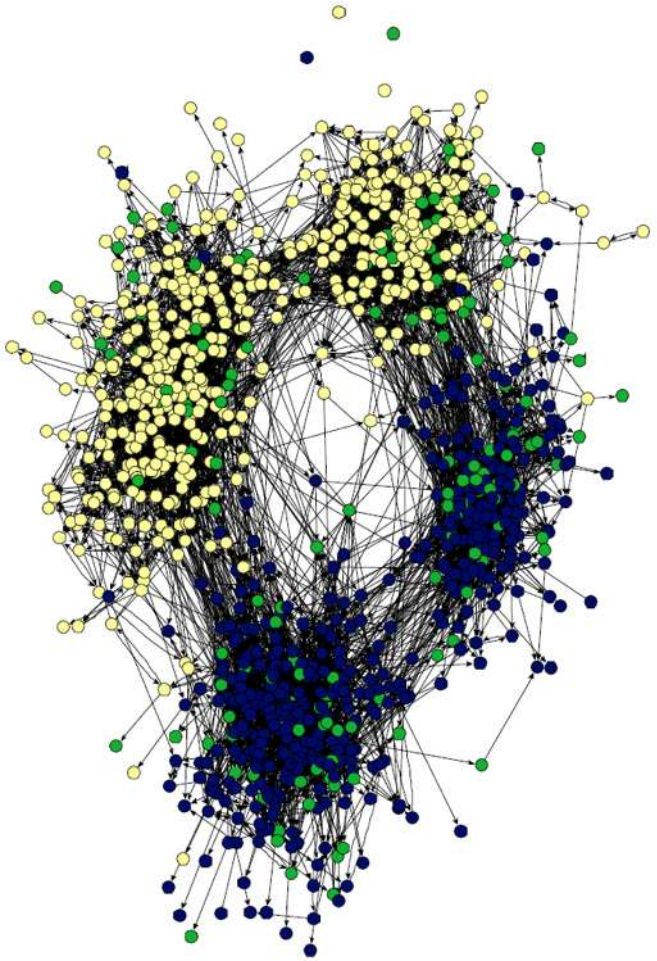
Network resilience (1)

- We observe how the connectivity (length of the paths) of the network changes as the vertices get removed
- Vertices can be removed:
  - Uniformly at random
  - In order of decreasing degree
- It is important for epidemiology
  - Removal of vertices corresponds to vaccination



Community structure

- Most social networks show community structure
  - groups have higher density of edges within than accross groups
  - People naturally divide into groups based on interests, age, occupation, …
- How to find communities:
  - Spectral clustering (embedding into a low-dim space)
  - Hierarchical clustering based on connection strength
  - Combinatorial algorithms
  - Block models
  - Diffusion methods



Friendship network of children in a school

Examples of Social Network Analysis

- 436-node network of email exchange at a corporate research lab [Adamic-Adar, SocNets '03]
- 43,553-node network of email exchange at a university [Kossinets-Watts, Science '06]
- 4.4-million-node network of declared friendships on a blogging community [Liben-Nowell et al., PNAS '05]
- 240-million-node network of communication on Microsoft Messenger [Leskovec-Horvitz, WWW '08]
- 800-million-node Facebook network [Backstrom et al. '11]

Network Analysis Tools

- SNAP
- JUNG
- NetworkX
- Pajek
- NodeXL
- igraph

- Visualization
  - Gephi
  - GraphViz
  - Tulip

Playing around at home

- Exercise from http://snap.stanford.edu/proj/snap-icwsm/

Best online resource

- https://github.com/briatte/awesome-network-analysis

Useful Online Sources

- Tina Eliassi-Rad: Information in Networks Spring 2013 course at Rutgers
- Barabasi: Network Science Book Project
- Jure Leskovec: Social and Information Network Analysis Fall 2013 course at Stanford

- Book: http://www.cs.cornell.edu/home/kleinber/networks-book/

(lots of slides in this presentation borrowed from these sources)