CS224W: Social and Information Network Analysis

CS224W: Social and Information Network Analysis

Jure Leskovec, Stanford University

http://cs224w.stanford.edu

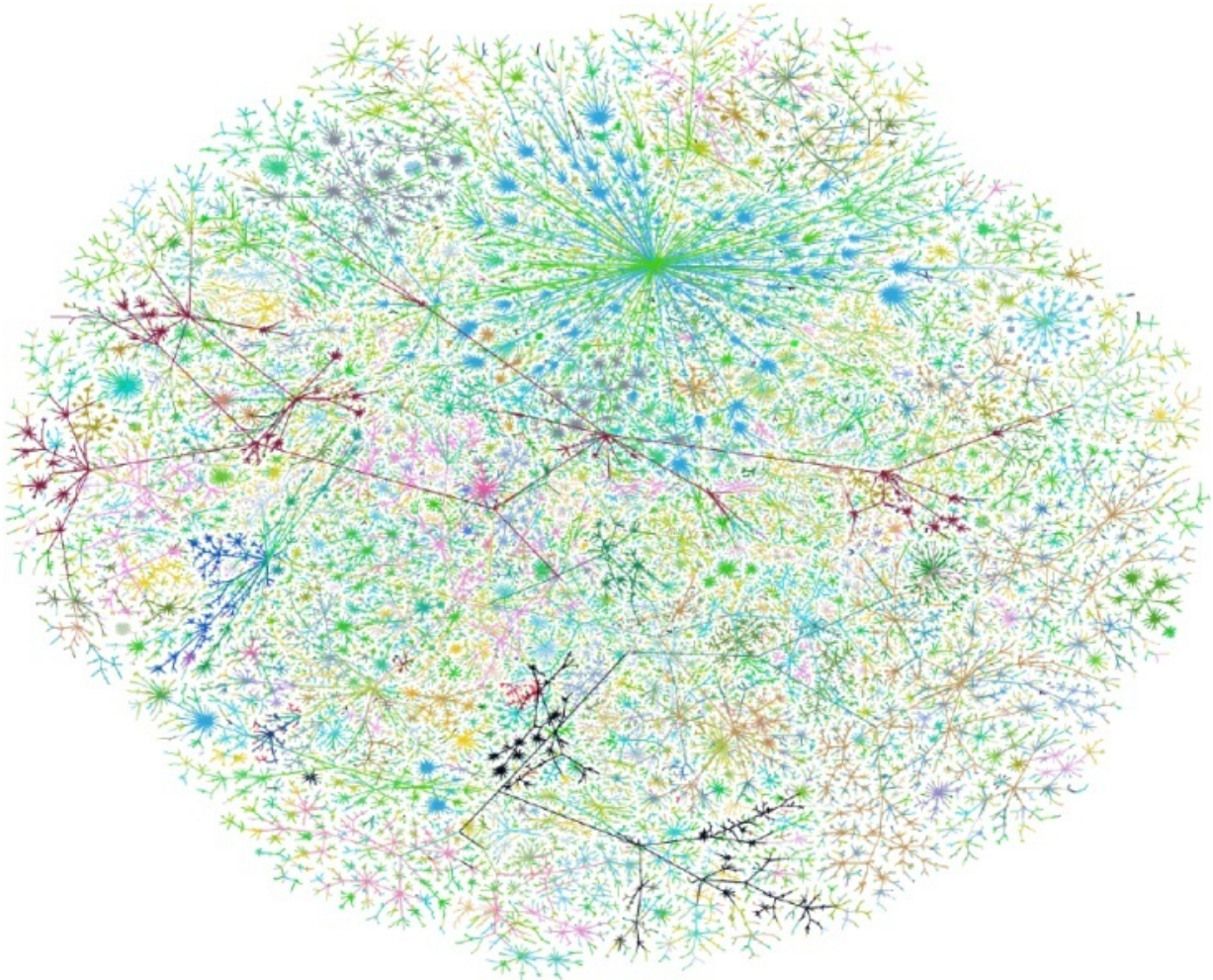
Networks & Complex Systems

- **We are surrounded by hopelessly complex systems:**
  - Society is a collection of six billion individuals
  - Communication systems link electronic devices
  - Information and knowledge is organized and linked
  - Thousands of genes in our cells work together in a seamless fashion
  - Our thoughts are hidden in the connections between billions of neurons in our brain
- These systems, random looking at first, display signatures of order and self-organization

Networks

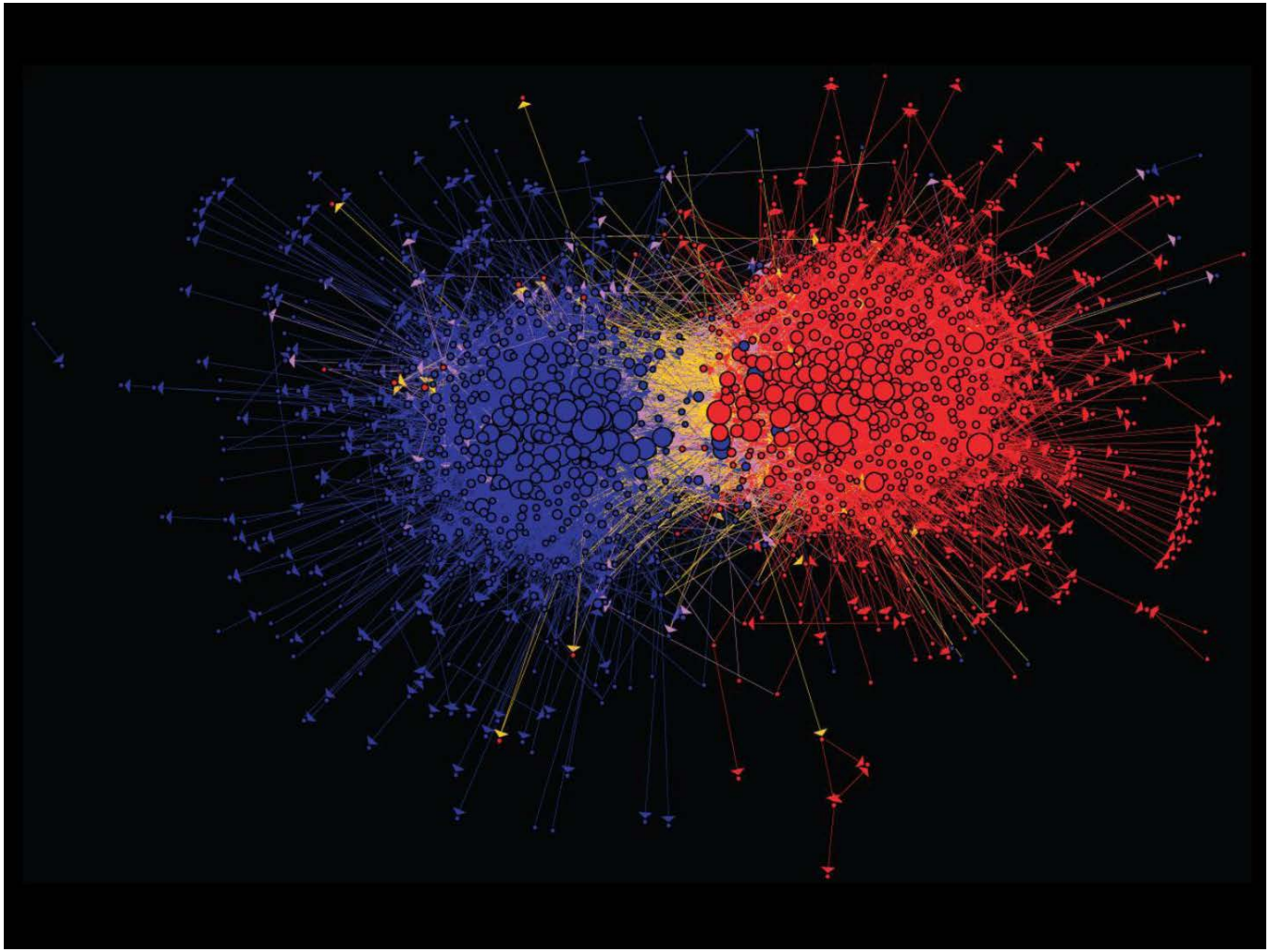
- Each such system can be represented as a **network**, that defines the **interactions** between the components

Networks: Communication



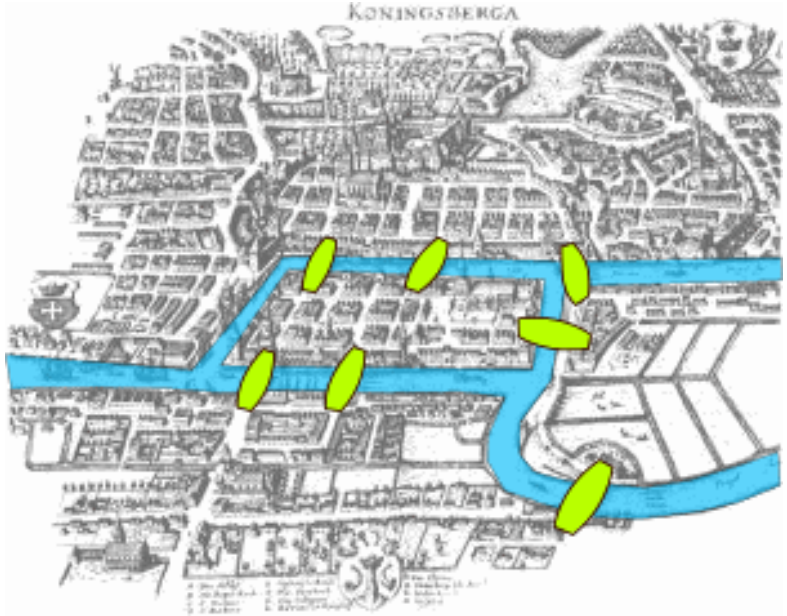
Graph of the Internet
(Autonomous Systems)

Networks: Information



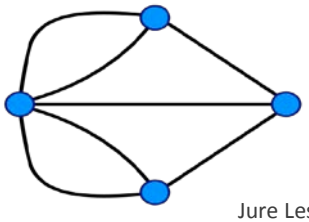
Connections between political blogs

Networks: Technology



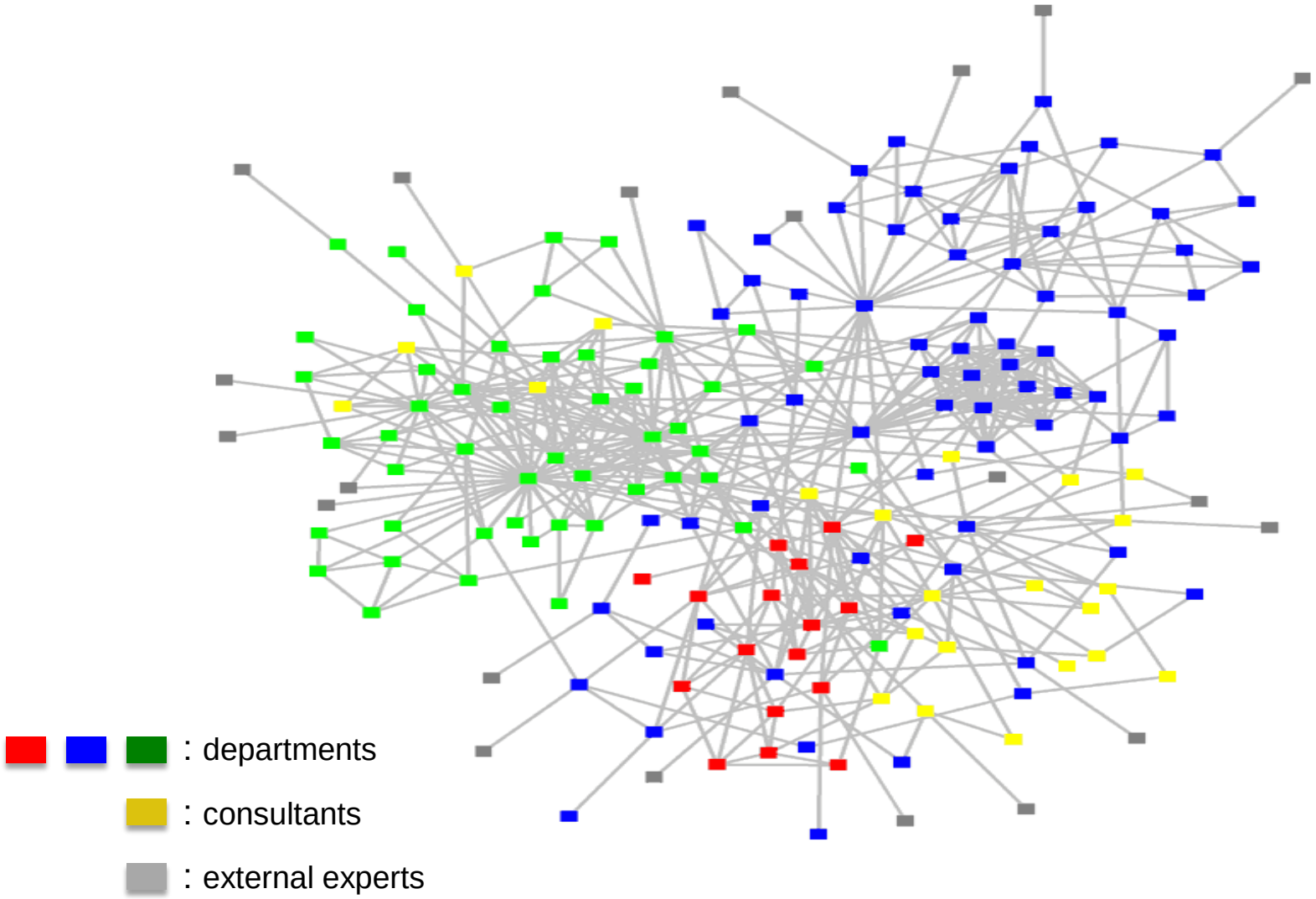
Seven Bridges of Königsberg (Euler 1735)

Return to the starting point by traveling each link of the graph once and only once.

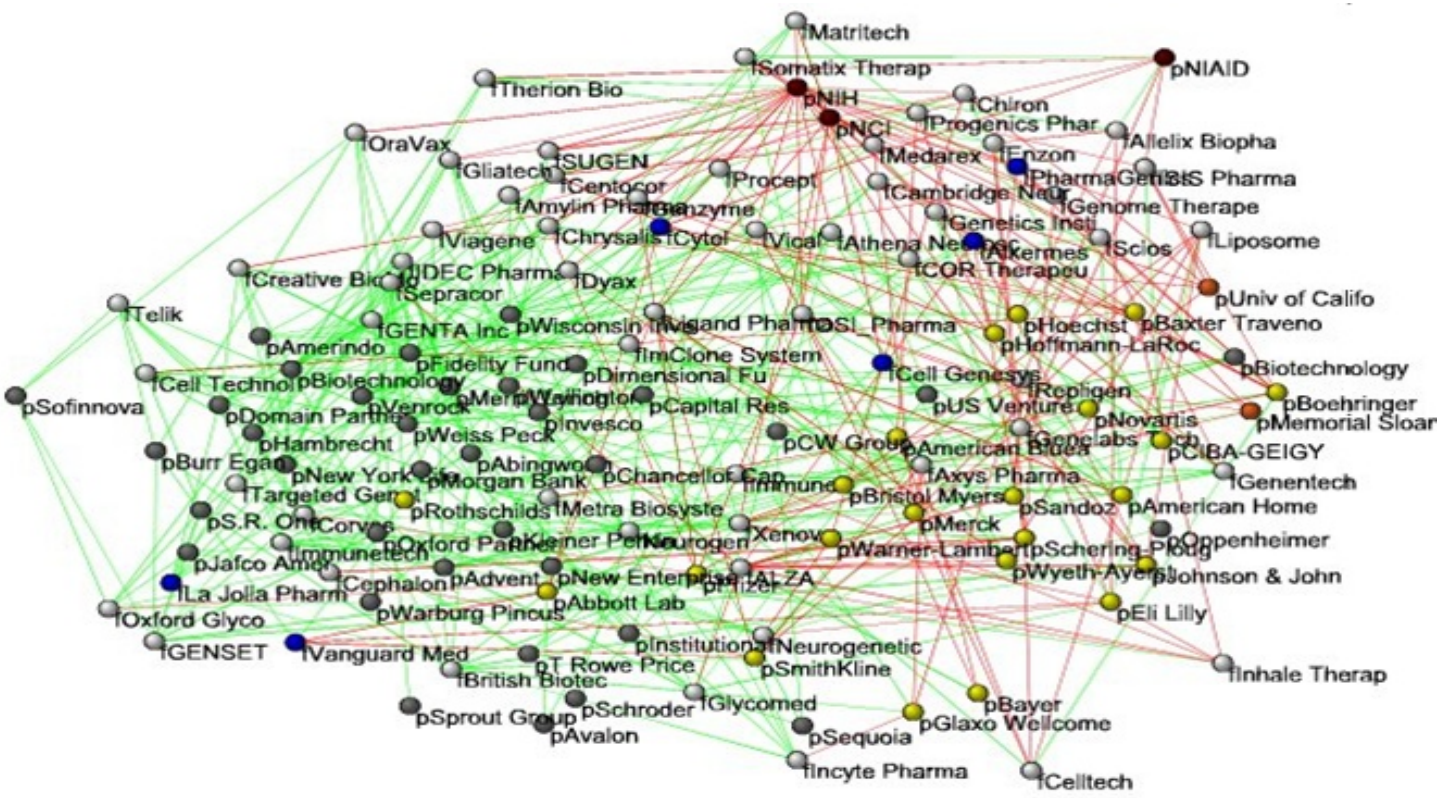


London Underground

Networks: Organizations



Networks: Economy



Nodes:

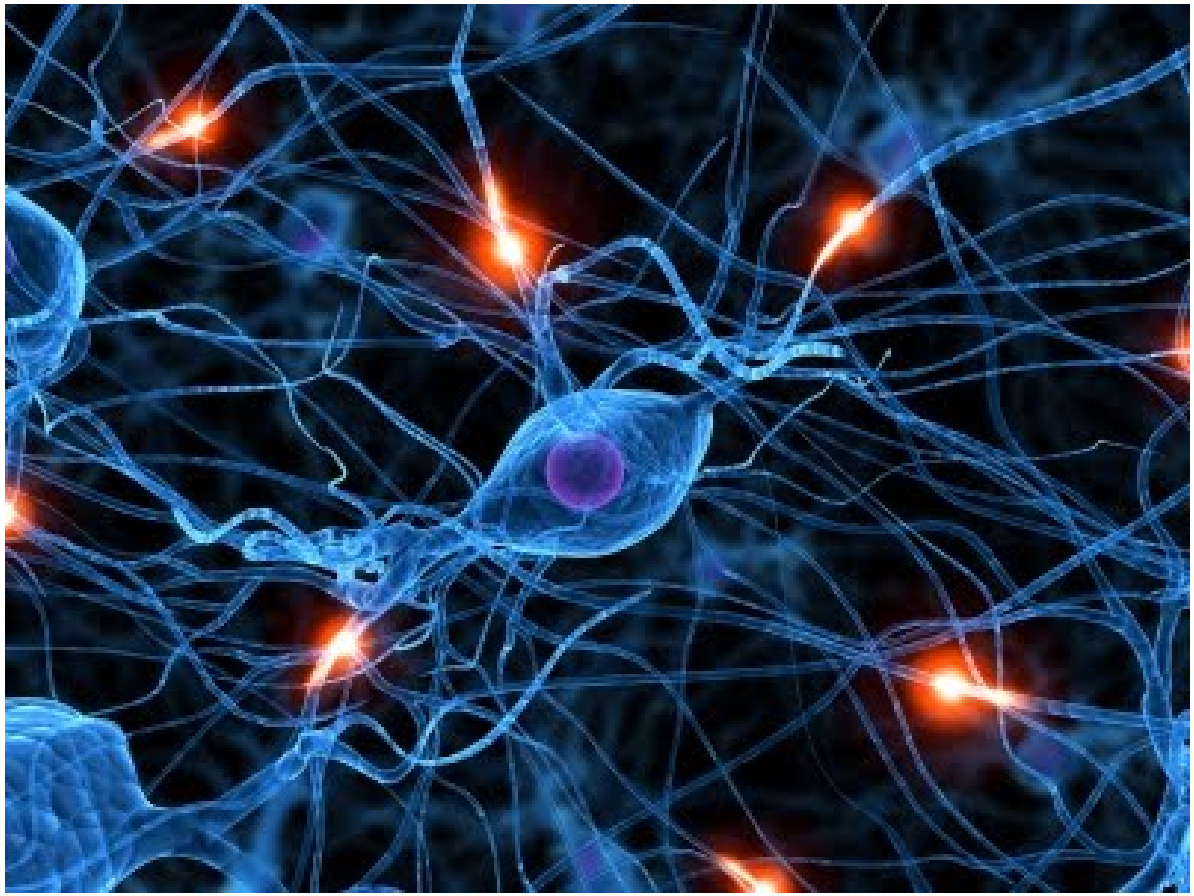
- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

Links:

- Collaborations
- Financial
- R&D

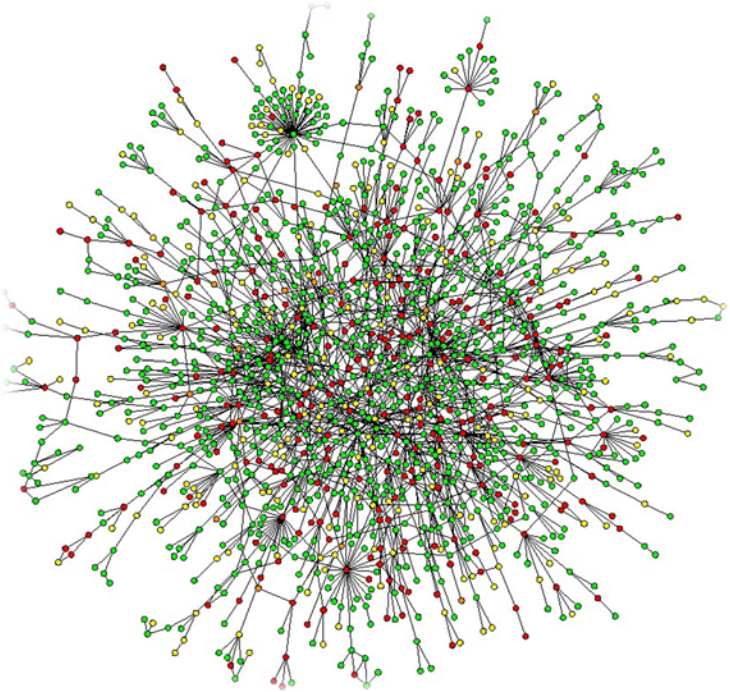
Bio-tech companies, 1991

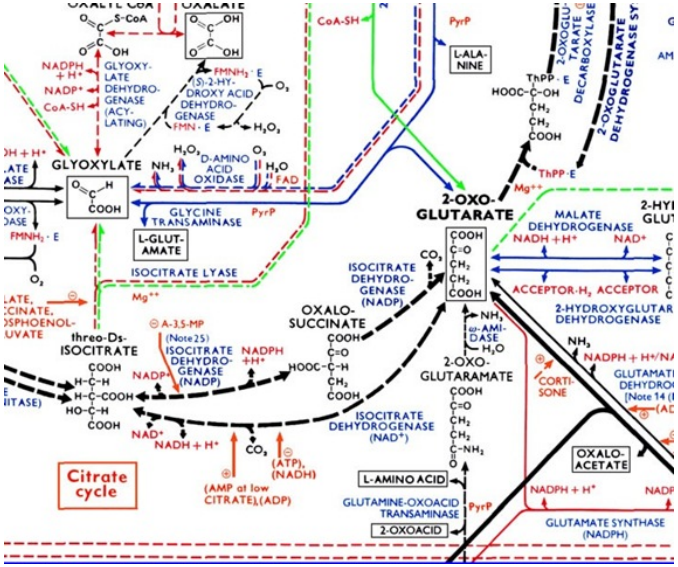
Networks: Brain



Human brain has between 10-100 billion neurons

Networks: Cells





Protein-Protein Interaction Networks:
Nodes: Proteins
Edges: 'physical' interactoins

Metabolic networks:
Nodes: Metabolites and enzymes
Edges: Chemical reactions

Why Networks?

Behind each such system there is an intricate wiring diagram, **a network**, that defines the **interactions** between the components

We will never understand a complex system unless we understand the networks behind it

Reasoning about Networks

How do we reason about networks

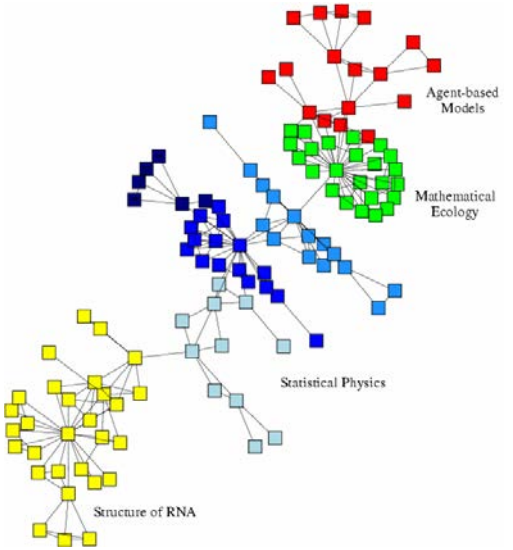
- Empirical: Study network data to find organizational principles
- Mathematical models: Probabilistic, graph theory
- Algorithms for analyzing graphs

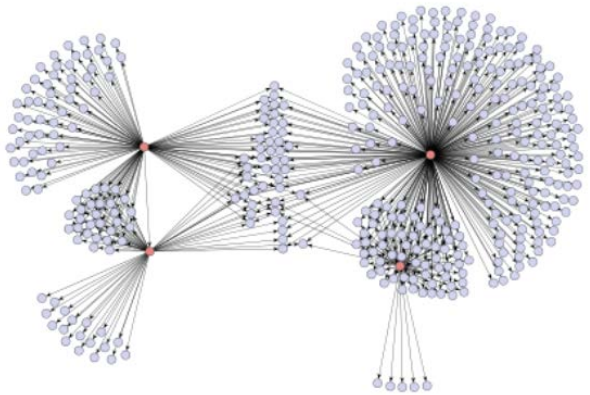
- What do we hope to achieve from models of networks?
  - Patterns and statistical properties of network data
  - Design principles and models
  - Understand why networks are organized the way they are (Predict behavior of networked systems)

Networks: Structure & Process

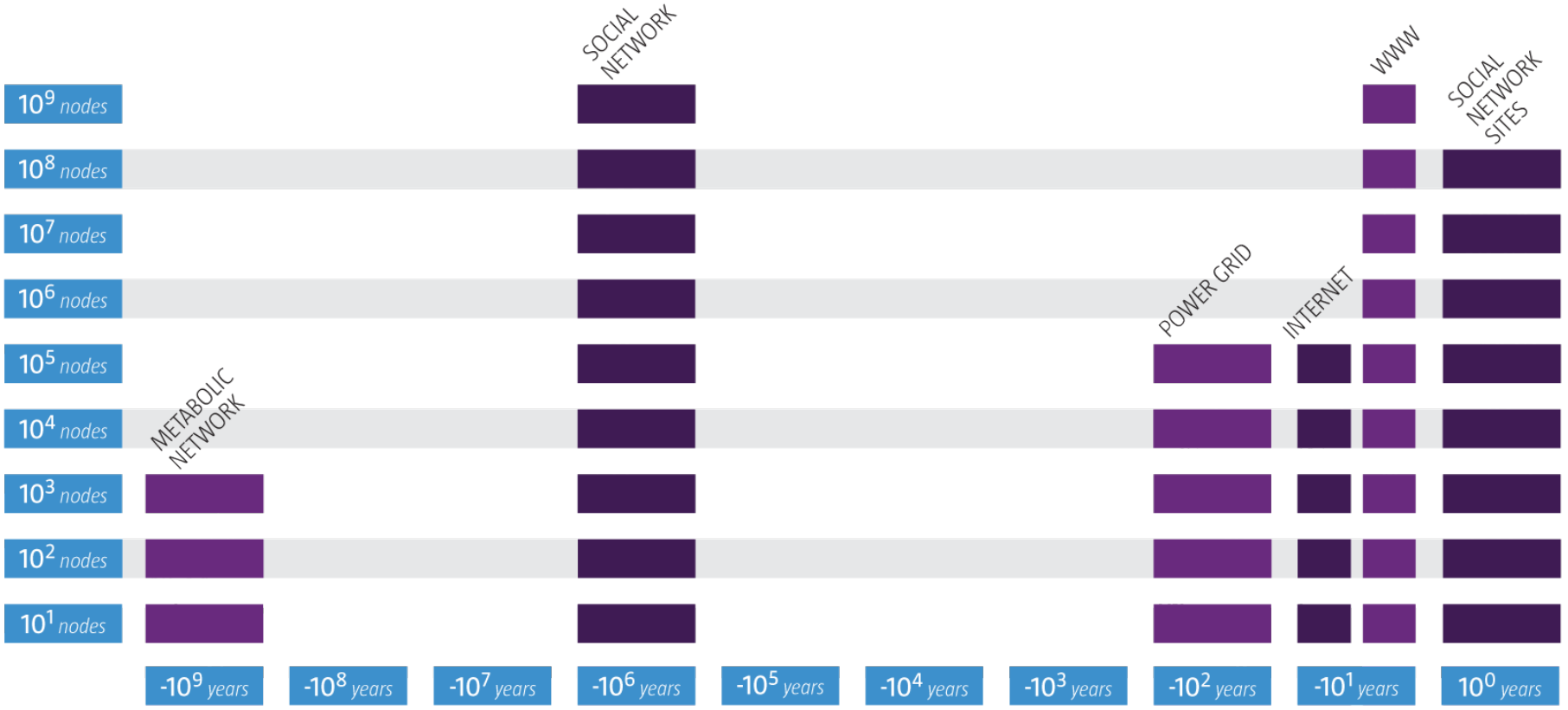
What do we study in networks?

- Structure and evolution:
  - What is the structure of a network?
  - Why and how did it became to have such structure?
- Processes and dynamics:
  - Networks provide "skeleton" for spreading of information, behavior, diseases



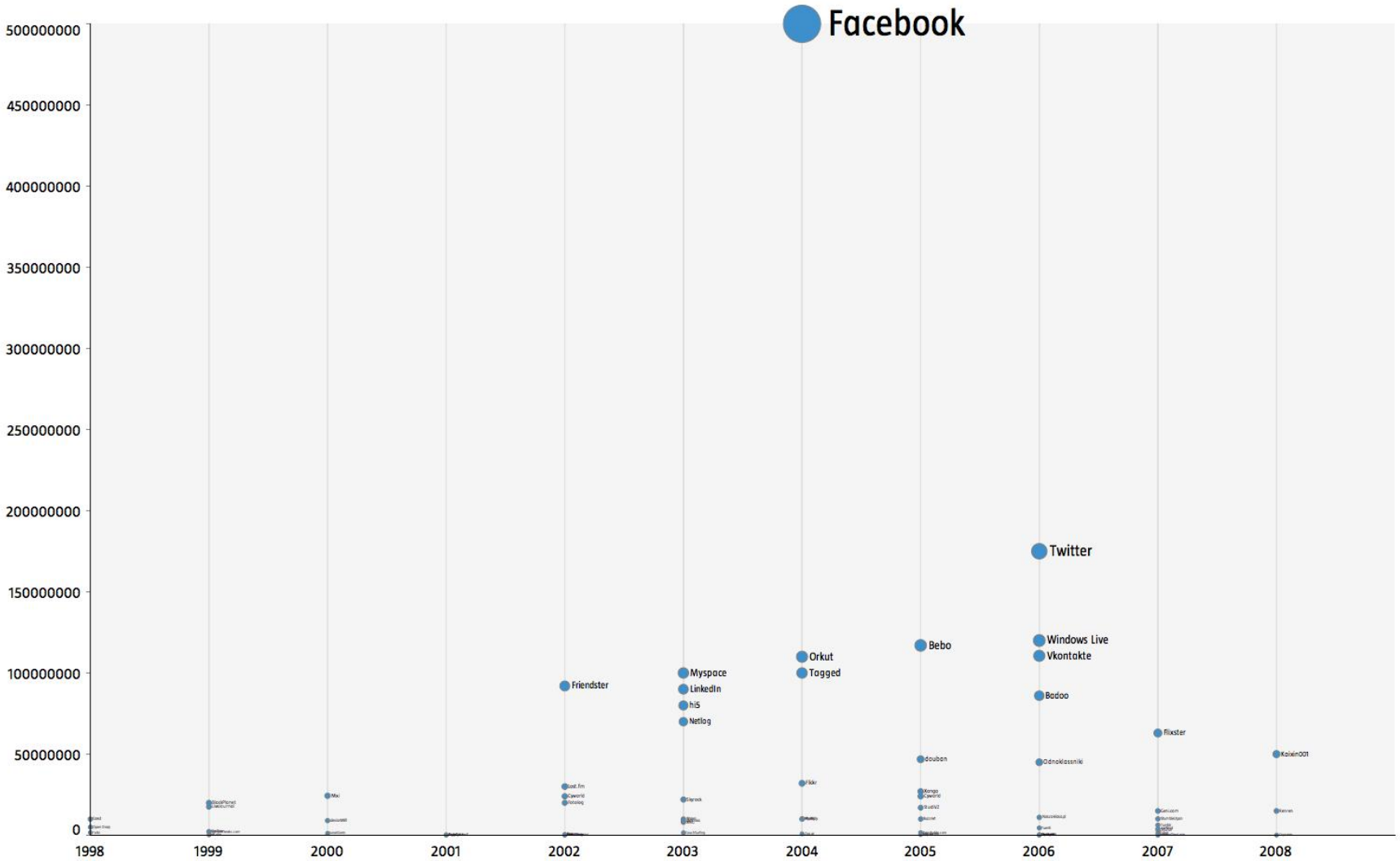


Networks: Why Now?

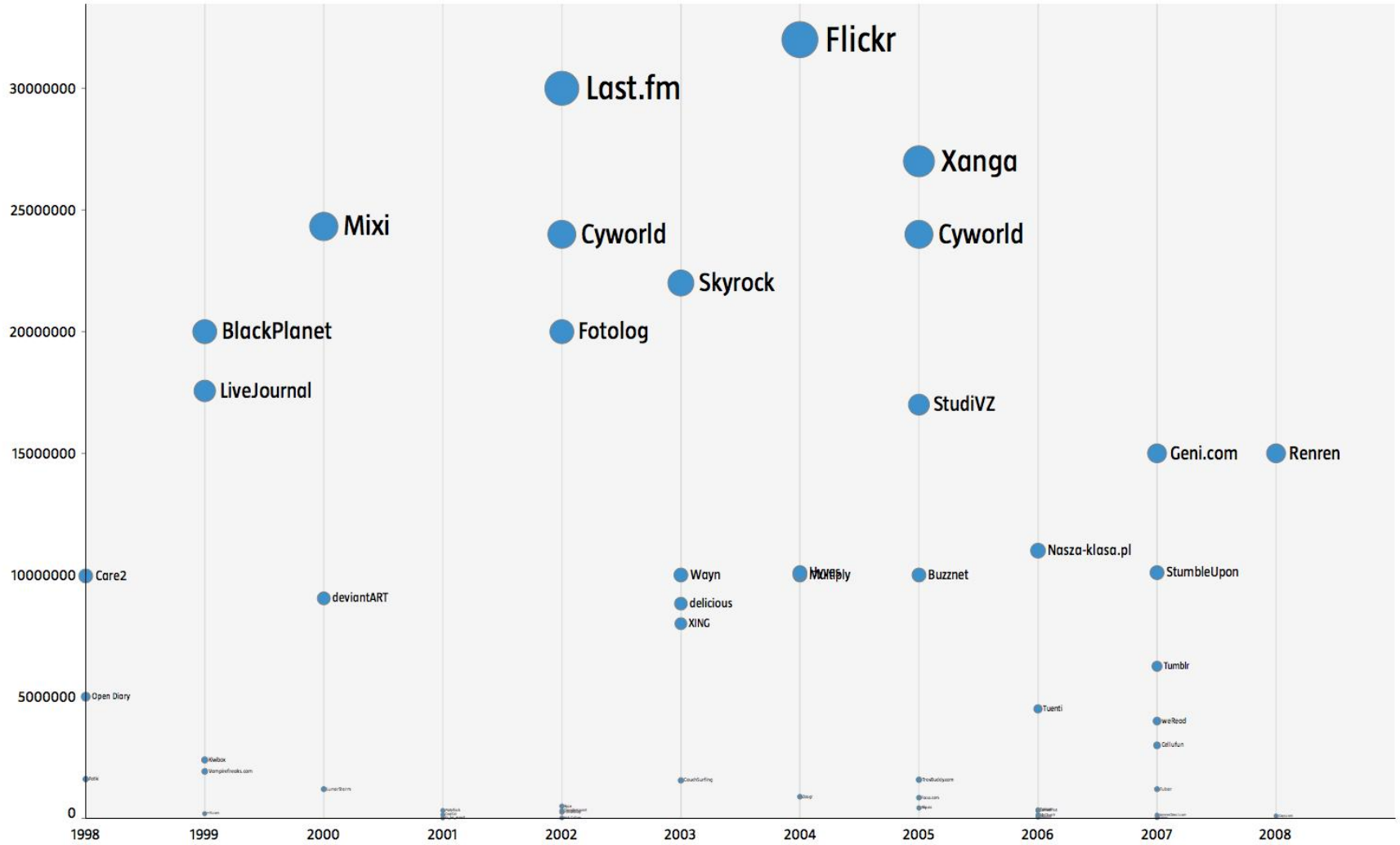


Age and size of networks

Networks: Why Now?



Networks: Why Now?



Why Networks? Why Now?

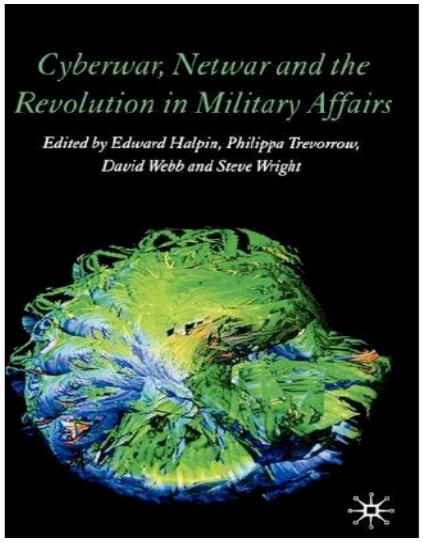
Why is the role of networks expanding?

- Data availability
  - Rise of the Web 2.0 and Social media
- Universality
  - Networks from various domains of science, nature, and technology are more similar than one would expect
- Shared vocabulary between fields
  - Computer Science, Social science, Physics, Economics, Statistics, Biology

Networks: Impact

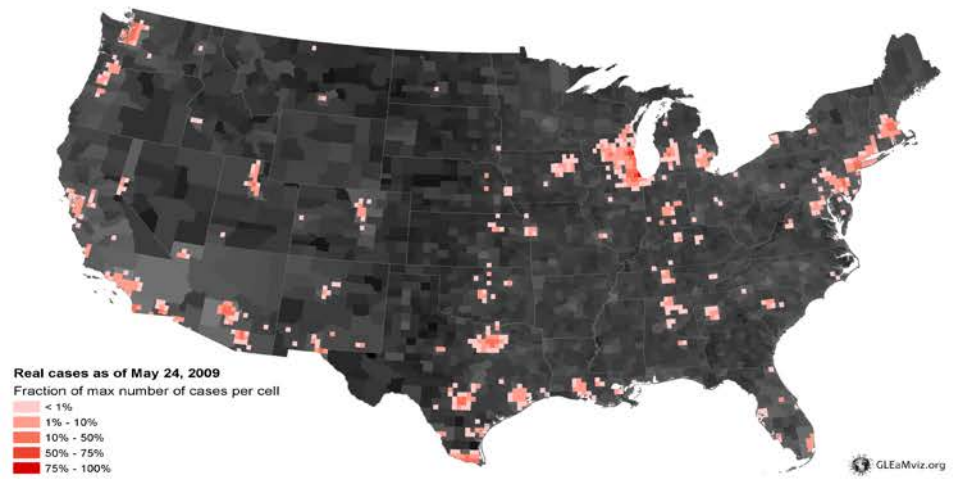
Networks: Impact

- Intelligence and fighting (cyber) terrorism

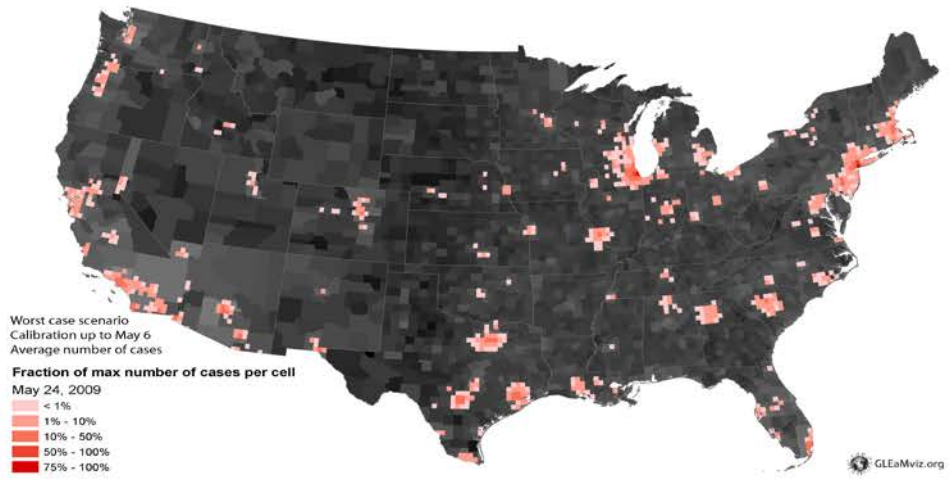


Networks: Impact

- **Predicting epidemics**



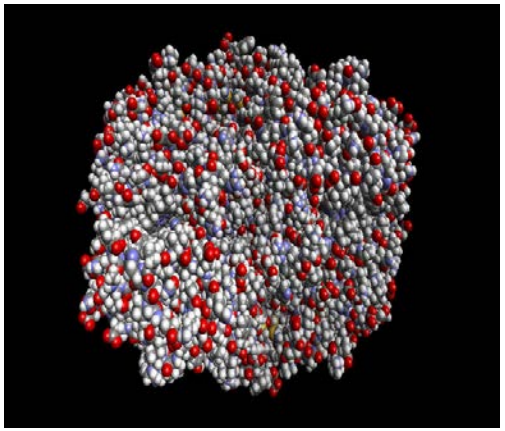
Real

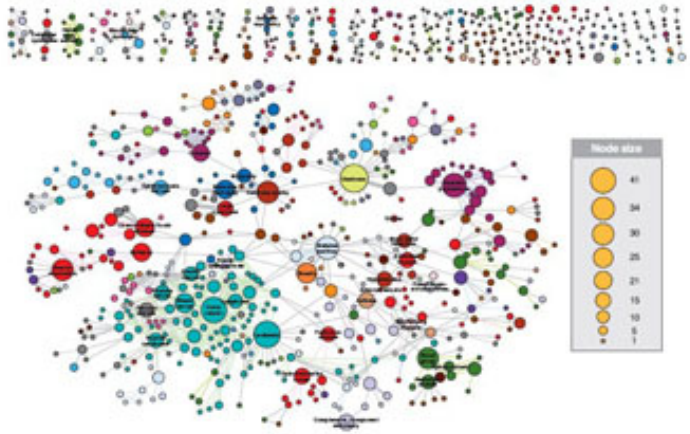


Predicted

Networks: Impact

- Interactions of human disease
- Drug design





Networks Really Matter

- If you were to understand the spread of diseases, **can you do it without networks?**
- If you were to understand the WWW structure and information, **hopeless without invoking the Web's topology.**
- If you want to understand human diseases, **it is hopeless without considering the wiring diagram of the cell**.

Course Logistics

Course Syllabus

- Covers a wide range of **network analysis techniques** – from basic to state-of-the-art
- You will learn about things you heard about:

  Six degrees of separation, small-world, page rank, network effects, P2P networks, network evolution, spectral graph theory, virus propagation, link prediction, power-laws, scale free networks, core-periphery, network communities, hubs and authorities, bipartite cores, information cascades, influence maximization, …

- Covers algorithms, theory and applications
- It's going to be fun ☺

Prerequisites

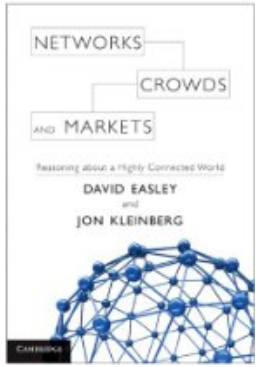
- **Good background in:**
  - Algorithms
  - Graph theory
  - Probability and Statistics
  - Linear algebra
- **Programming:**
  - You should be able to write non-trivial programs
- **4 recitation sessions:**
  - 2 to review basic mathematical concepts
  - 2 to review programming tools (SNAP, NetworkX)

Logistics

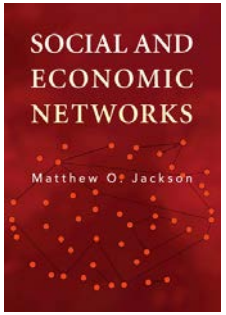
- **Course website:**
  http://cs224w.stanford.edu
- Slides posted at least 30 min before the class
- Required readings:
  - Mostly chapters from Easley&Kleinberg book
  - Papers
- Optional readings:
  - Papers and pointers to additional literature
  - This will be very useful for reaction paper and project proposal

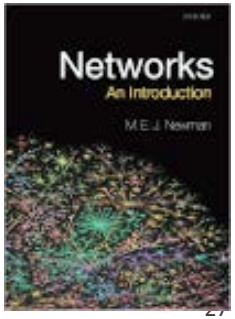
Text Books

- Recommended textbook:
  - D. Easley, J. Kleinberg: *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*
  - Freely available at: http://www.cs.cornell.edu/home/kleinber/networks-book/



- Optional books:
  - Matthew Jackson: *Social and Economic Networks*
  - Mark Newman: *Networks: An introduction*





Work for the Course

Assignment	Due on
Homework 1	October 13
Reaction paper	October 20
Project proposal	October 27
Homework 2	November 3
Competition	November 10
Project milestone	November 17
Project write-up	December 11 (no late days!)
Project poster presentation	December 16 12:15-3;15pm

Grading

- Final grade will be composed of:
  - 2 homeworks: 15% each
  - Reaction paper: 10%
  - **Substantial class project: 60%**
    - Proposal: 15%
    - Project milestone: 15%
    - Final report: 60%
    - Poster session: 10%

Homeworks, Write-ups, Reports

- **Assignments** (homeworks, write-ups, reports) **take time. Start early!**
- How to submit?
  - **Paper:** Box outside the class and in Gates basement
    - We will grade on paper!
  - **You should also submit electronic copy:**
    - 1 PDF/ZIP file (writeups, experimental results, code)
    - Submission website: http://www.stanford.edu/class/cs224w/submit/
  - **SCPD:** Only submit electronic copy & send us email
- 7 late days for the quarter:
  - Max 4 late days per assignment

Course Projects

- **Substantial course project:**
  - Experimental evaluation of algorithms and models on an interesting network dataset
  - A theoretical project that considers a model, an algorithm and derives a rigorous result about it
  - An in-depth critical survey of one of the course topics and offering a novel perspective on the area
- Performed in groups of (exactly) 3 students
- Project is the **main work** for the class

Course Assistants

Borja Peleato (head TA)

Chenguang Zhu

Evan Rosen

Dakan Wang

Communication

- **Piazza Q&A website:**
  - http://piazza.com/stanford/fall2011/cs224w
    - If you don't have @stanford.edu email address, send us email and we will register you to Piazza
- For e-mailing course staff, always use:
  - cs224w-aut1112-staff@lists.stanford.edu
- For course announcements subscribe to:
  - cs224w-aut1112-all@lists.stanford.edu

Sitting-in & Auditing

- **You are welcome to sit-in and audit the class**
  - Please send us email saying that you will be auditing the class
- To receive announcements, subscribe to the mailing list:
  - cs224w-aut1112-all@lists.stanford.edu

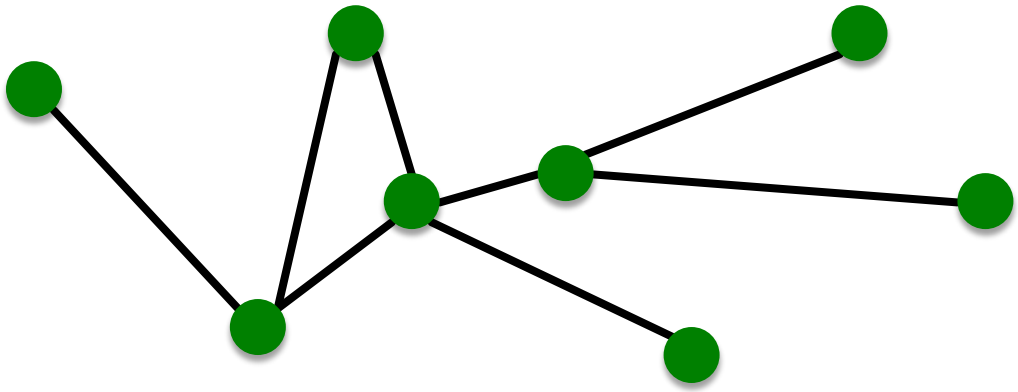
Starter Topic: Structure of Networks

Structure of Networks?

Network is a collection of objects where some pairs of objects are connected by links

What is the structure of the network?

Components of a Network



- Objects: nodes, vertices — N
- Interactions: links, edges — E
- System: network, graph — $G(N,E)$

Networks or Graphs?

- **Network** often refers to real systems
  - Web, Social network, Metabolic network

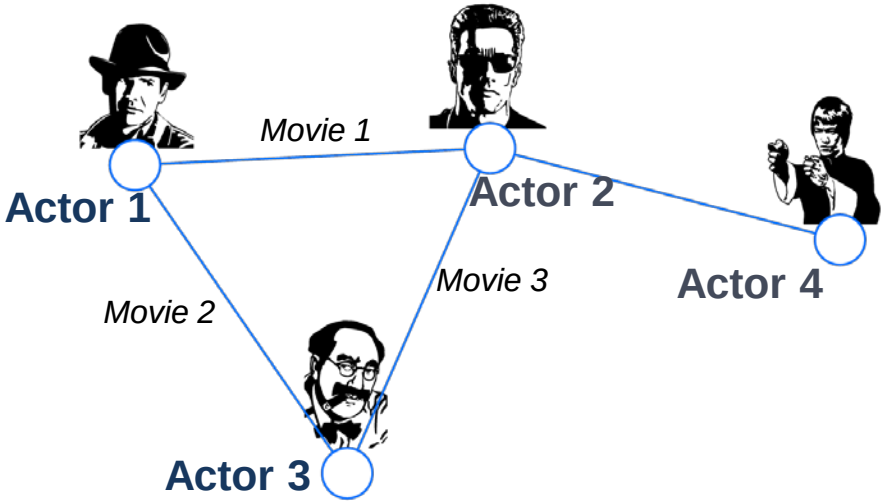
  Language: Network, node, link

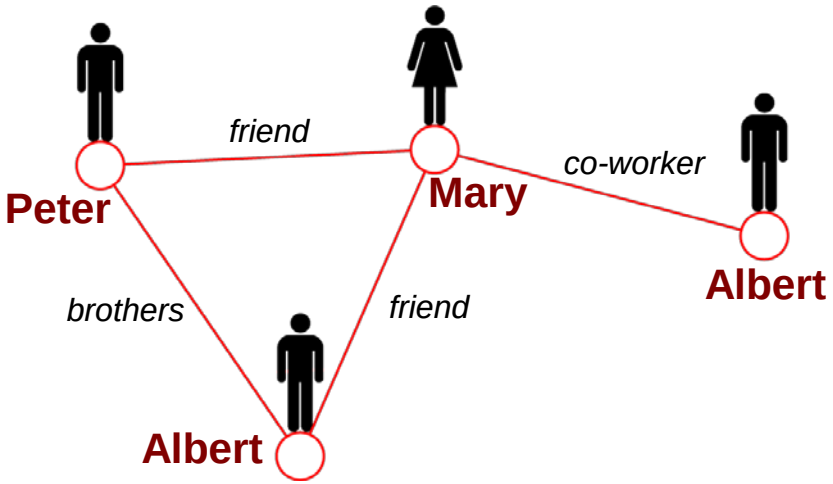
- **Graph:** mathematical representation of a network
  - Web graph, Social graph (a Facebook term)

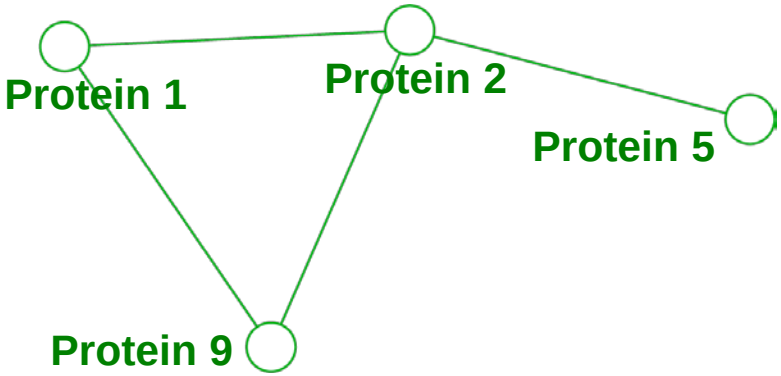
  Language: Graph, vertex, edge

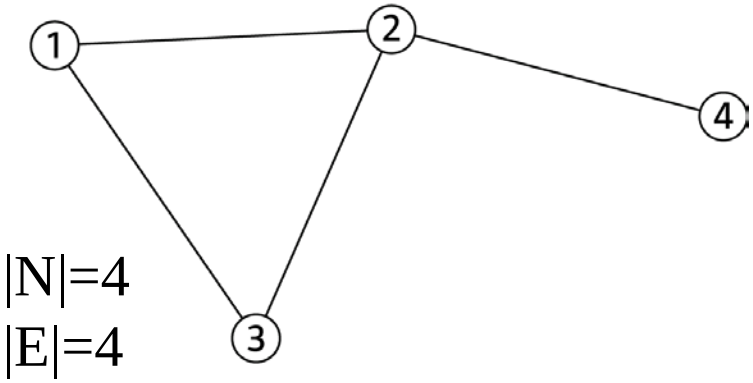
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably

Networks: Common Language







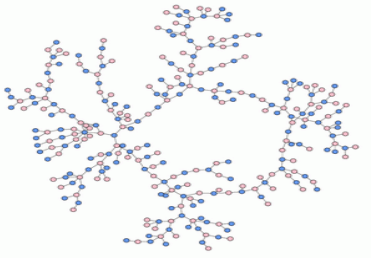


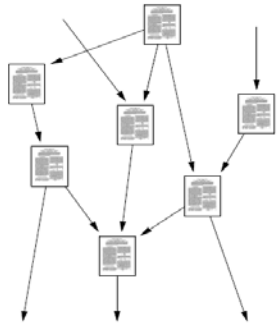
Choosing Proper Representation

- **Choice of the proper network representation determines our ability to use networks successfully:**
  - In some cases there is a unique, unambiguous representation
  - In other cases, the representation is by no means unique
  - The way you assign links will determine the nature of the question you can study

Choosing Proper Representation

- If you connect individuals that work with each other, you will explore a **professional network**
- If you connect those that have a sexual relationship, you will be exploring **sexual networks**
- If you connect scientific papers that cite each other, you will be studying the **citation network**



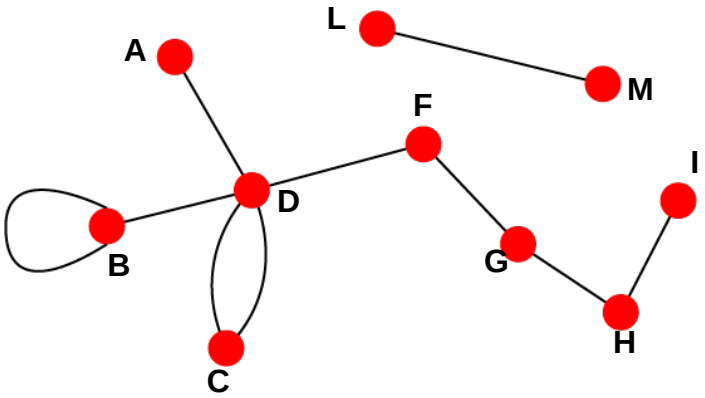


- If you connect all papers with the same word in the title, you will be exploring what? It is a network, nevertheless

Undirected vs. Directed Networks

Undirected

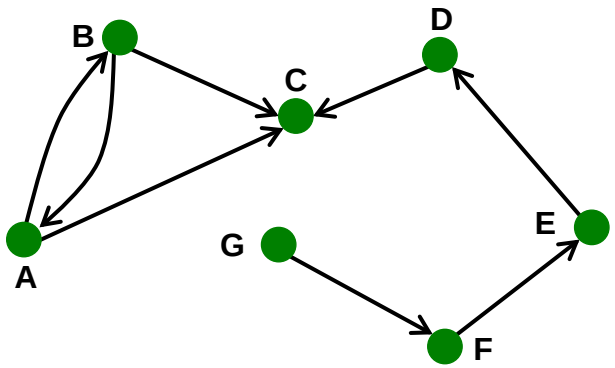
- Links: undirected (symmetrical)



- Undirected links:
  - Collaborations
  - Friendship on Facebook

Directed

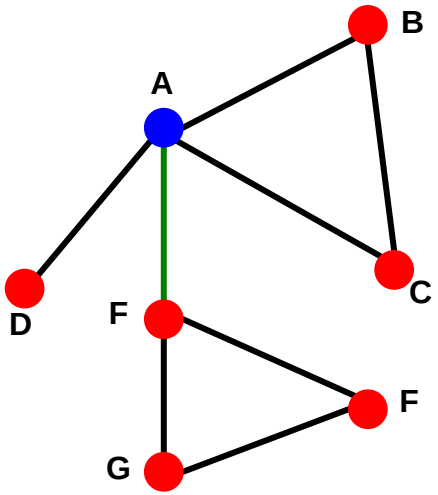
- Links: directed (arcs)

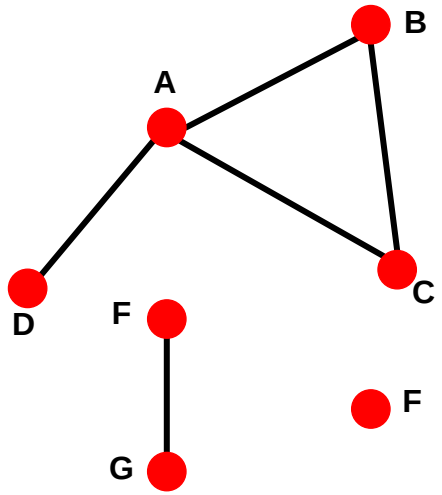


- Directed links:
  - Phone calls
  - Following on Twitter

Connectivity of Graphs

- Connected (undirected) graph:
  - Any two vertices can be joined by a path.
- A disconnected graph is made up by two or more connected components





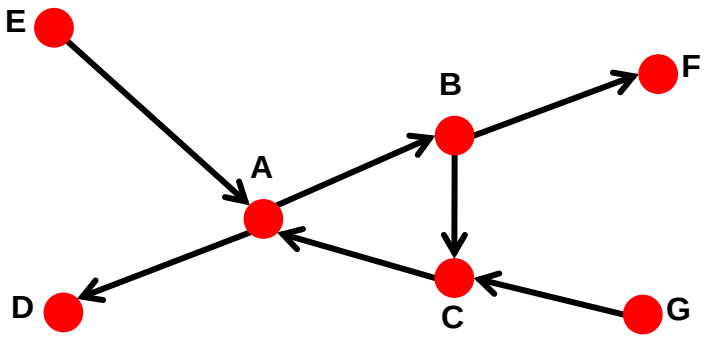
Largest Component:
Giant Component

Isolated node (F)

Bridge edge: If we erase it, the graph becomes disconnected.
Articulation point: If we erase it, the graph becomes disconnected.

Connectivity of Directed Graphs

- Strongly connected directed graph
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)
- Weakly connected directed graph
  - is connected if we disregard the edge directions



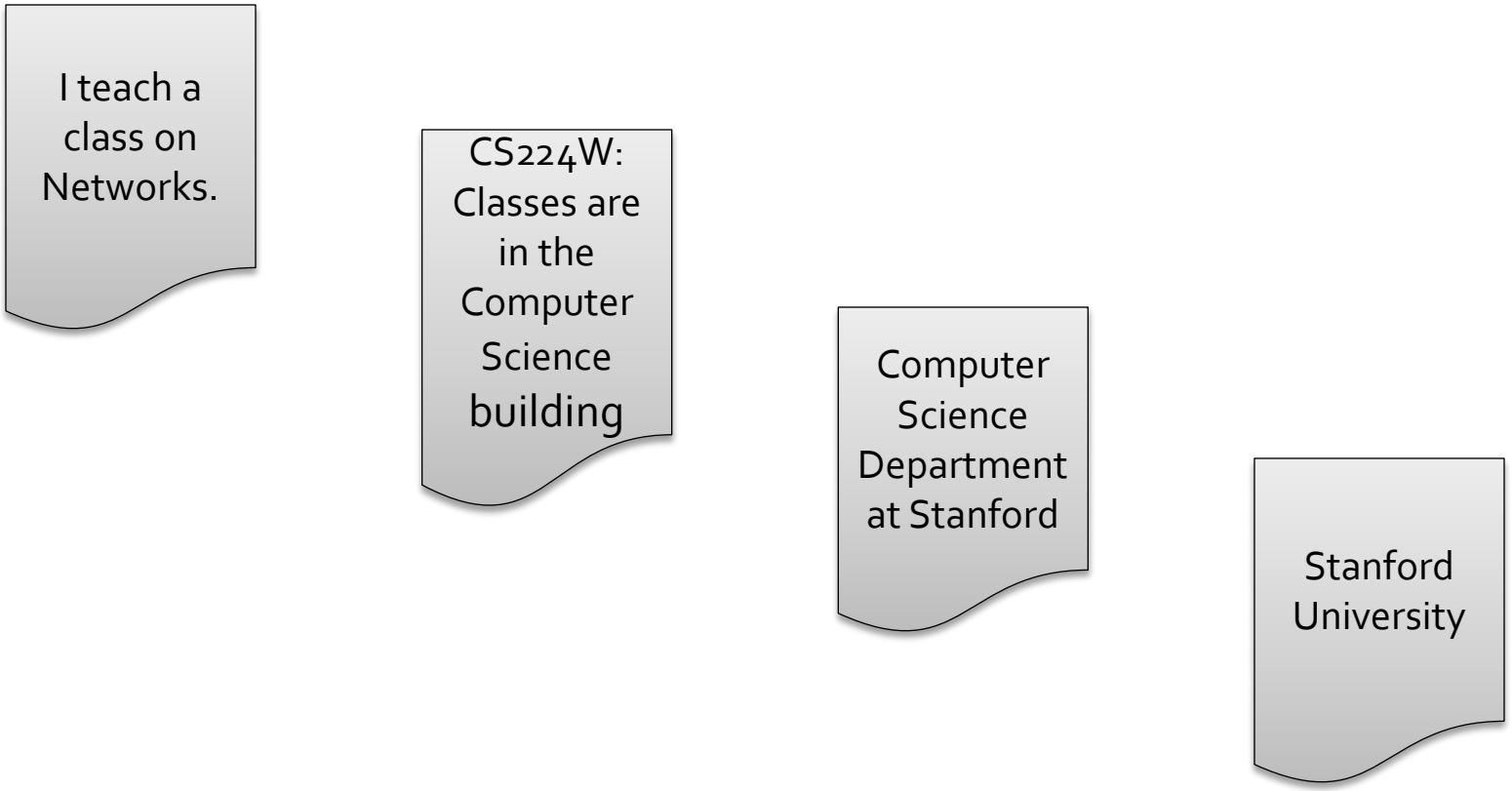
Graph on the left is not strongly connected.

Web as a Graph

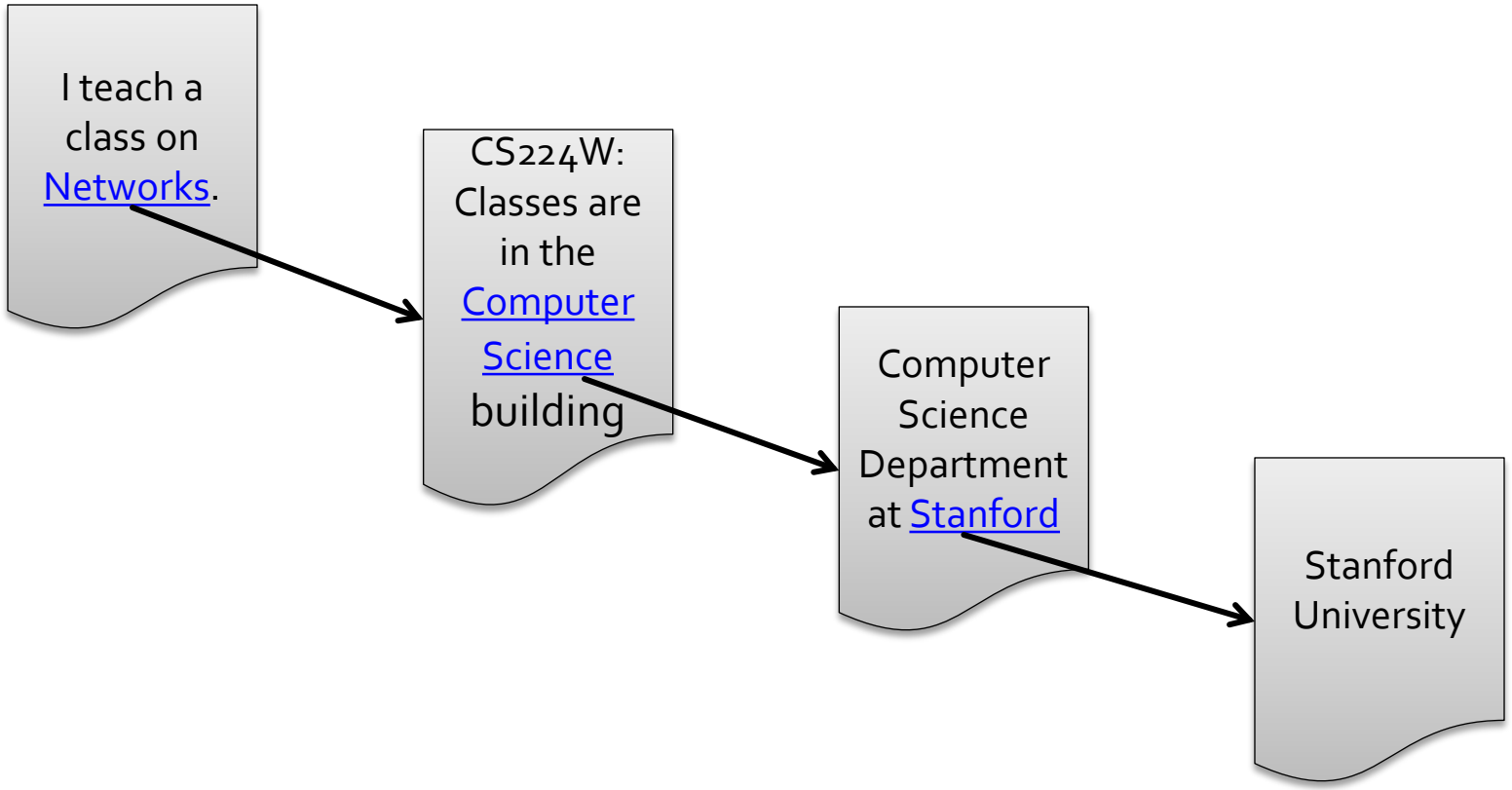
Q: What does Web "look like" at a global level?

- **Web as a graph:**
  - Nodes = pages
  - Edges = hyperlinks
- What is a node?
  - Problems:
    - Dynamic pages created on the fly
    - "dark matter" – inaccessible database generated pages

The Web as a Graph

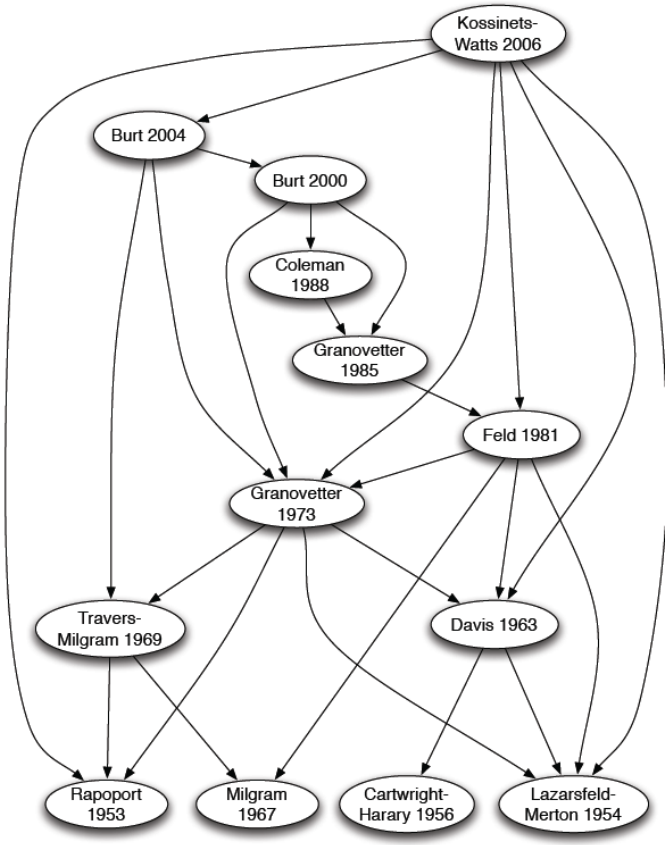


The Web as a Graph

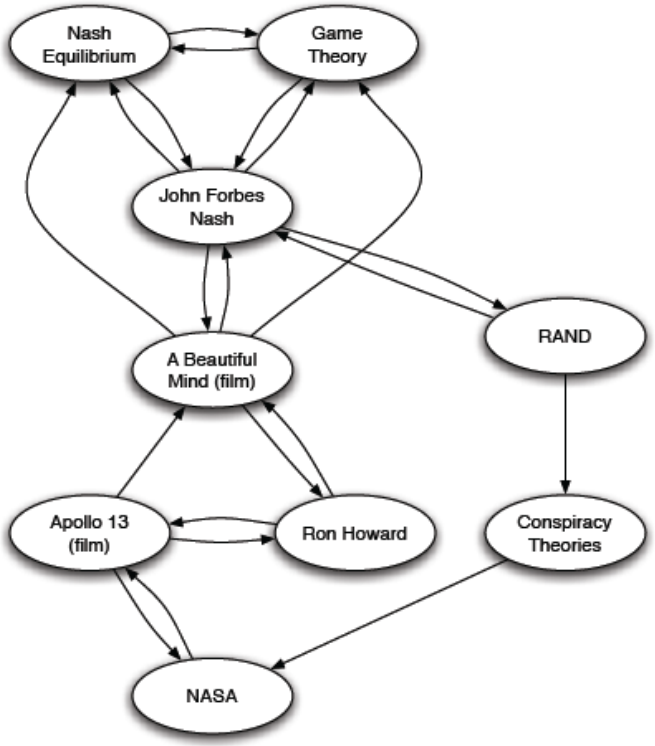


- In early days of the Web links were navigational
- Today many links are transactional

Other Information Networks

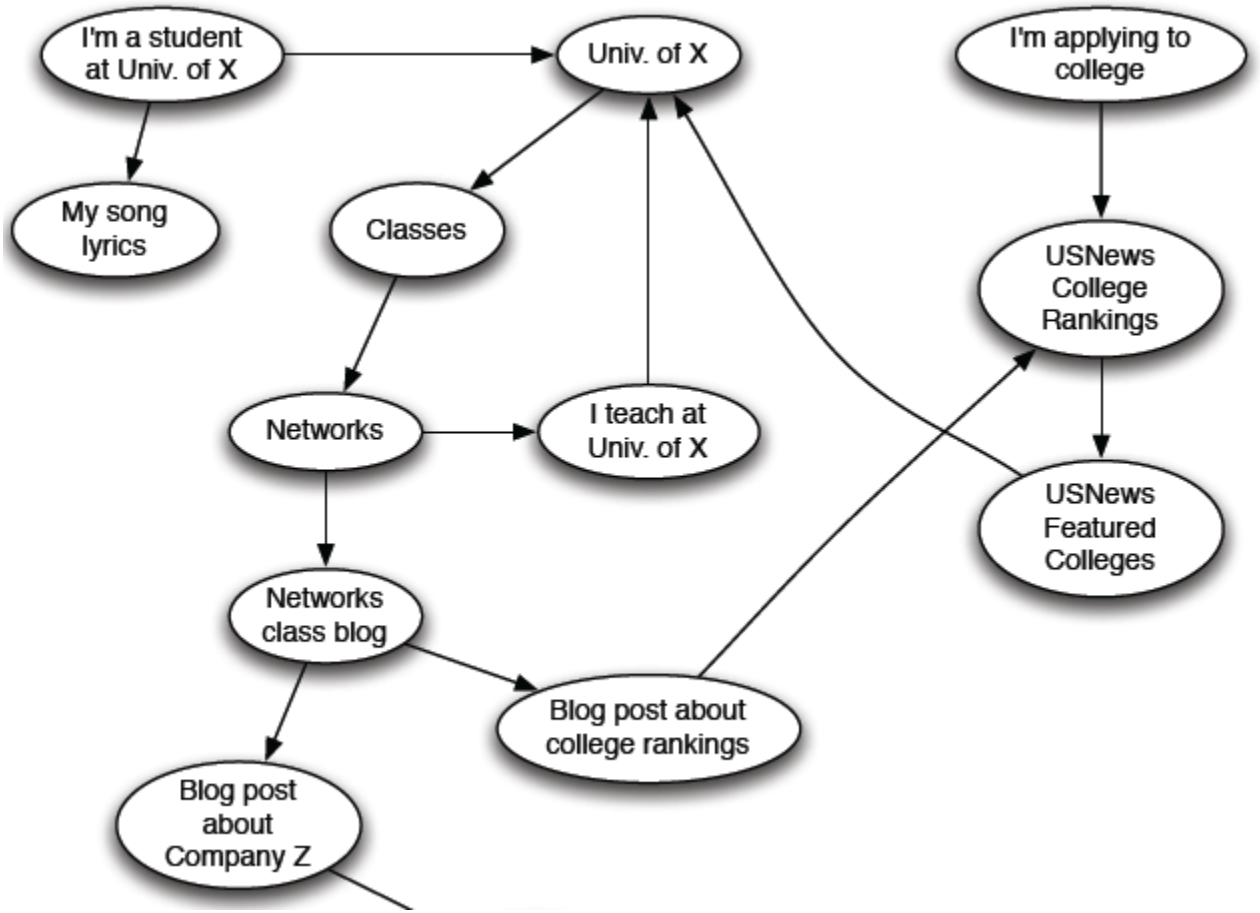


Citations



References in an Encyclopedia

Web as a Directed Graph

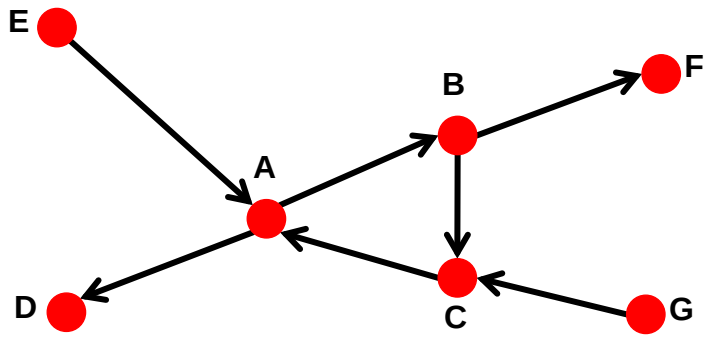


How does the Web look like?

- How is the Web linked?
- What is the "map" of the Web?

Web as a directed graph [Broder et al. 2000]:

- Given node v, what can v reach?
- What other nodes can reach v?



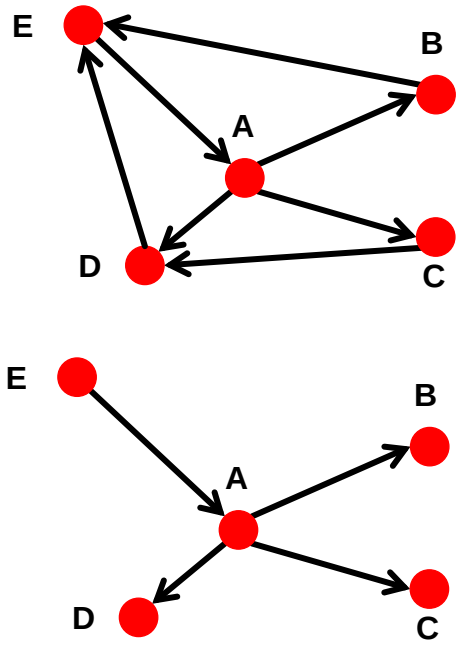
$$In(v) = \{w \mid w \text{ can reach } v\}$$
$$Out(v) = \{w \mid v \text{ can reach } w\}$$

In(A) = {B,C,E,G}
Out(A)={B,C,D,F}

Directed Graphs

- Two types of directed graphs:
  - Strongly connected:
    - Any node can reach any node via a directed path

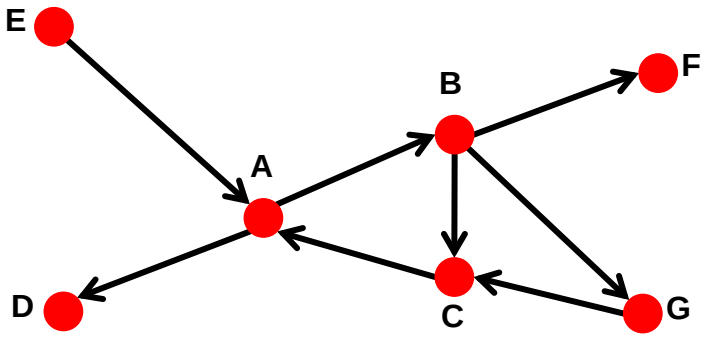
      $In(A)=Out(A)=\{A,B,C,D,E\}$
  - DAG – Directed Acyclic Graph:
    - Has no cycles: if u can reach v, then v can not reach u



- Any directed graph can be expressed in terms of these two types

Strongly Connected Component

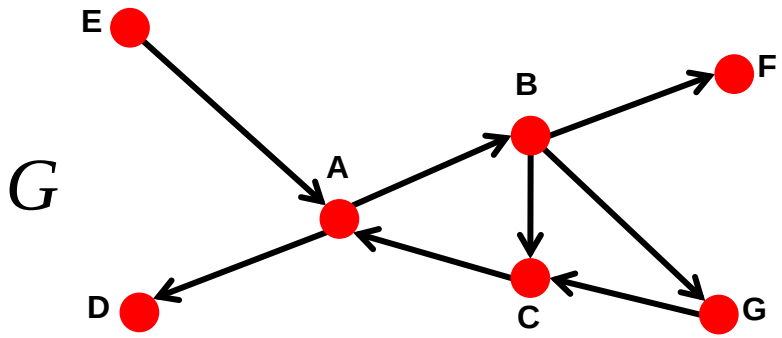
- Strongly connected component (SCC) is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



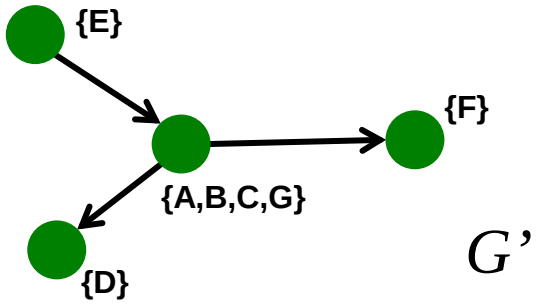
Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

Strongly Connected Component

- Fact: Every directed graph is a DAG on its SCCs
  - (1) SCCs partitions the nodes of G
    - Each node is in exactly one SCC
  - (2) If we build a graph G' whose nodes are SCCs, and with an edge between nodes of G' if there is an edge between corresponding SCCs in G, then G' is a DAG

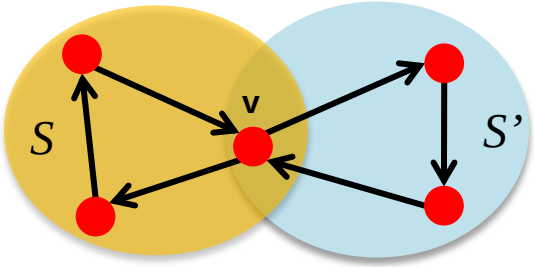


(1) Strongly connected components of graph G: {A,B,C,G}, {D}, {E}, {F}
(2) G' is a DAG:

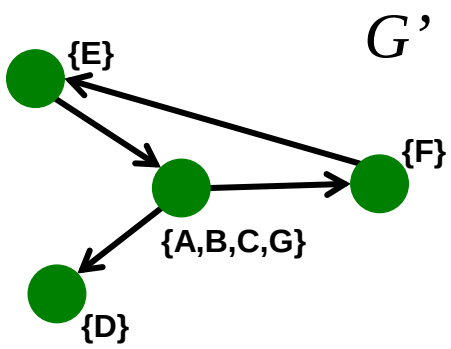


Proof

- **Why is (1) true?** SCCs partitions the nodes of G.
  - Suppose node v is a member of 2 SCCs S and S'.
  - Then $S \cup S'$ is one large SCC:



- **Why is (2) true?** G' (graph of SCCs) is a DAG
  - If G' is not a DAG, then we have a directed cycle.
  - Now all nodes on the cycle are mutually reachable, and all are part of the same SCC.

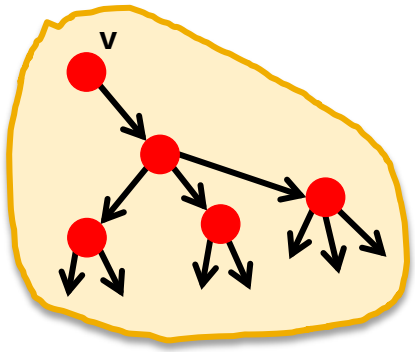


Now {A,B,C,G,E,F} is a SCC

Graph Structure of the Web

- Take a large snapshot of the Web and try to understand how it's SCCs "fit together" as a DAG
- Computational issue:
  - Want to find SCC containing node v?
  - Observation:
    - $Out(v)$ … nodes that can be reached from v
    - SCC containing v is: $Out(v) \cap In(v)$

      $= Out(v,G) \cap Out(v,\overline{G})$, where $\overline{G}$ is G with all edge directions flipped

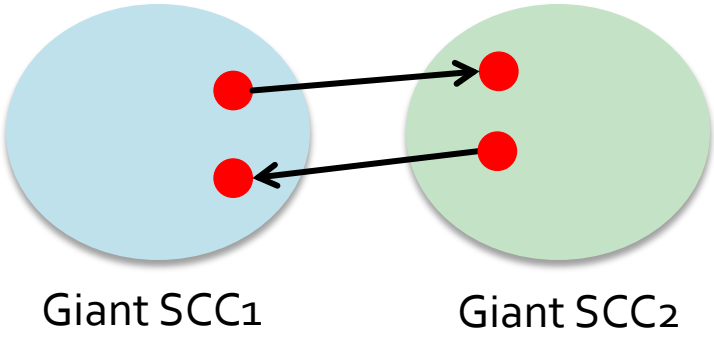


$Out(v)$

[Broder et al., 'oo]

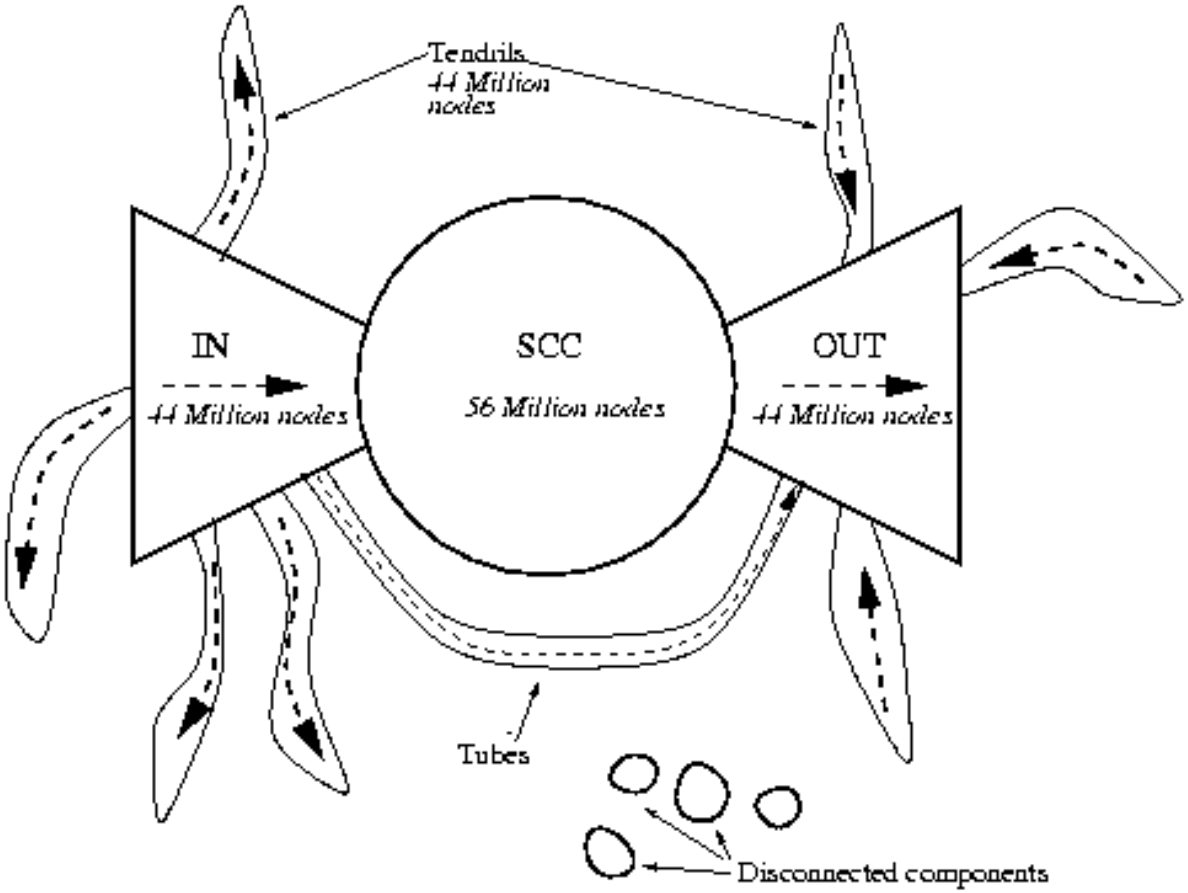
Graph structure of the Web

- There is a giant SCC
- There won't be 2 giant SCCs:
  - Just takes 1 page from one SCC to link to the other SCC
  - If the components have millions of pages the likelihood of this is very large



[Broder et al., 'oo]

Bow-tie structure of the Web



- 250 million pages, 1.5 billion links

What did We Learn/Not Learn ?

- Learn:
  - Some conceptual organization of the Web (i.e., bowtie)
- Not learn:
  - Treats all pages as equal
    - Google's homepage == my homepage
  - What are the most important pages
    - How many pages have k in-links as a function of k?
      The degree distribution: ~$1/k^{2+\varepsilon}$
    - Link analysis ranking -- as done by search engines (PageRank)
  - Internal structure inside giant SCC
    - Clusters, implicit communities?
  - How far apart are nodes in the giant SCC:
    - Distance = # of edges in shortest path
    - Avg = 16 [Broder et al.]