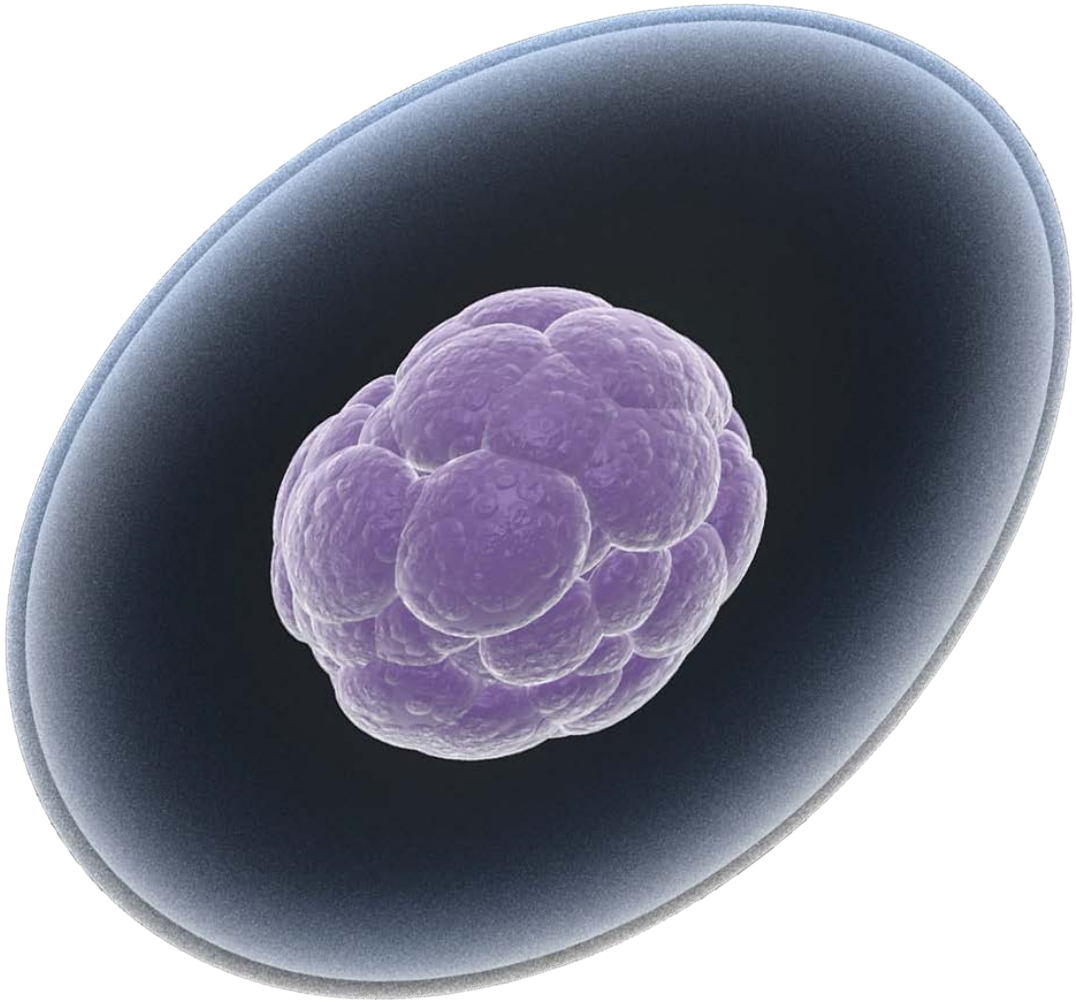
CS224w: Social and Information Network Analysis

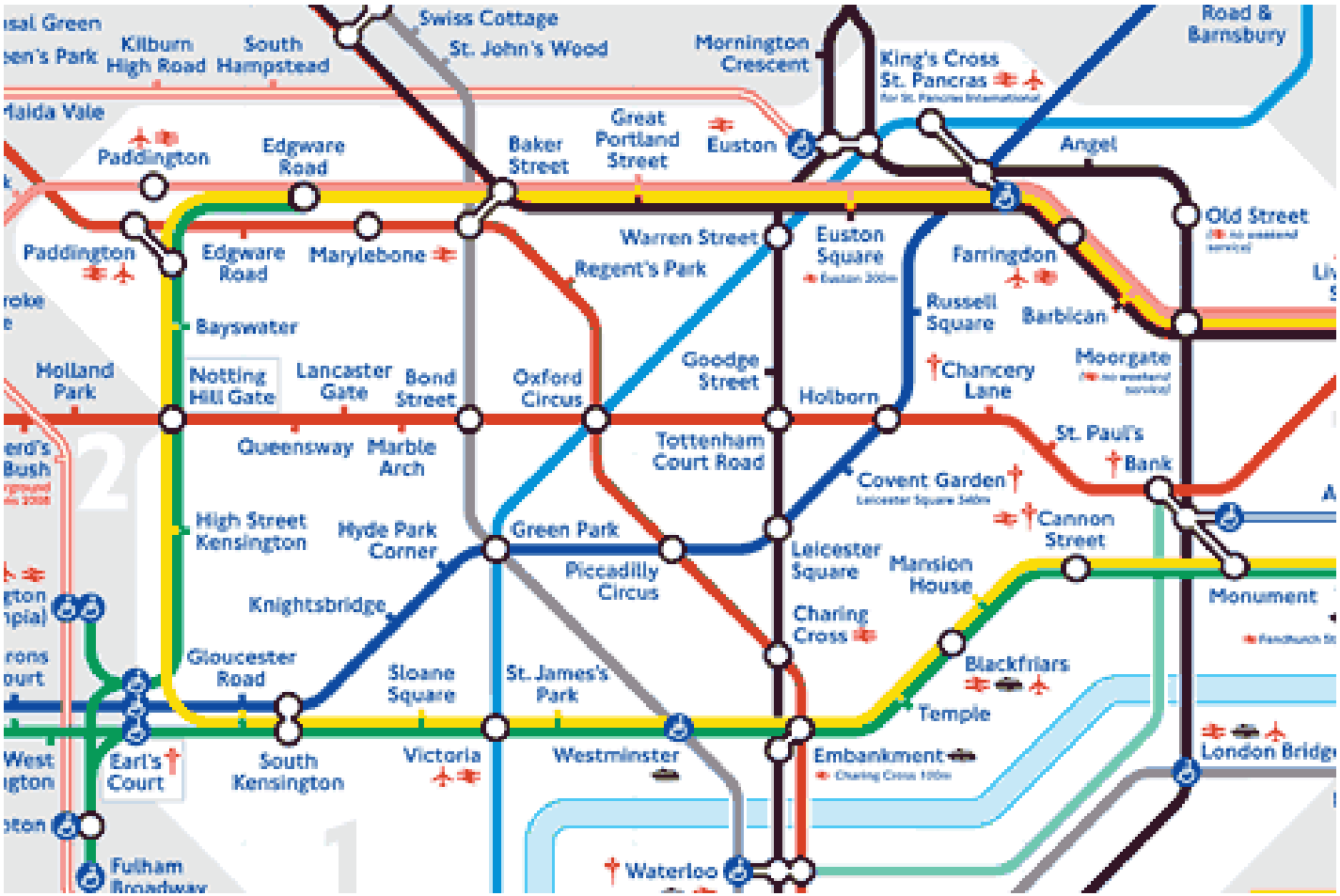
CS224w: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

What do the following things have in common?

World economy

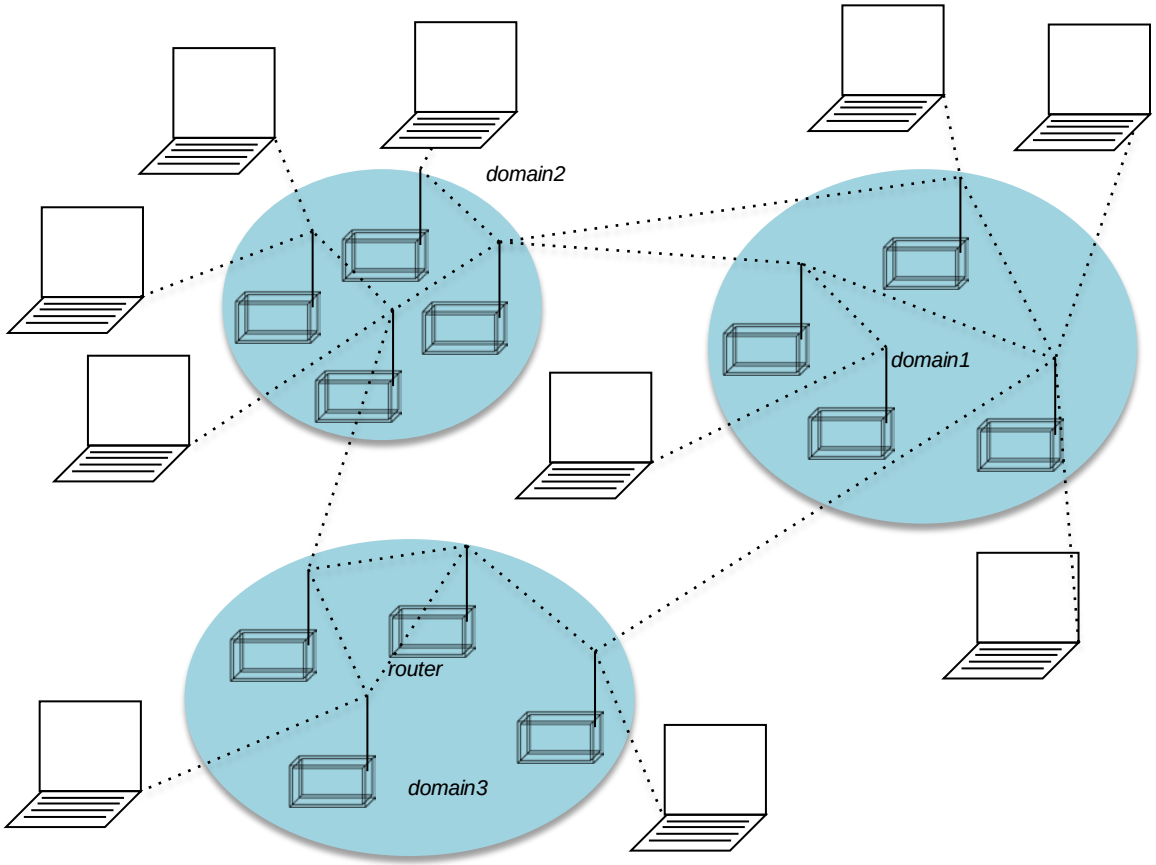


Human cell


Swiss Cottage
St. John's Wood
Kilburn High Road
South Hampstead
Mornington Crescent
King's Cross St. Pancras
Road & Barnsbury
Maida Vale
Paddington
Edgware Road
Baker Street
Great Portland Street
Euston
Angel
Old Street
Paddington
Edgware Road
Marylebone
Warren Street
Regent's Park
Euston Square
Farringdon
Russell Square
Barbican
Bayswater
Holland Park
Notting Hill Gate
Lancaster Gate
Bond Street
Oxford Circus
Goodge Street
Holborn
Chancery Lane
Moorgate
Queensway
Marble Arch
Tottenham Court Road
St. Paul's
Bank
Covent Garden
Cannon Street
High Street Kensington
Hyde Park Corner
Green Park
Piccadilly Circus
Leicester Square
Mansion House
Monument
Knightsbridge
Charing Cross
Gloucester Road
Sloane Square
St. James's Park
Blackfriars
Temple
Earl's Court
South Kensington
Victoria
Westminster
Embankment
London Bridge
Fulham Broadway
Waterloo

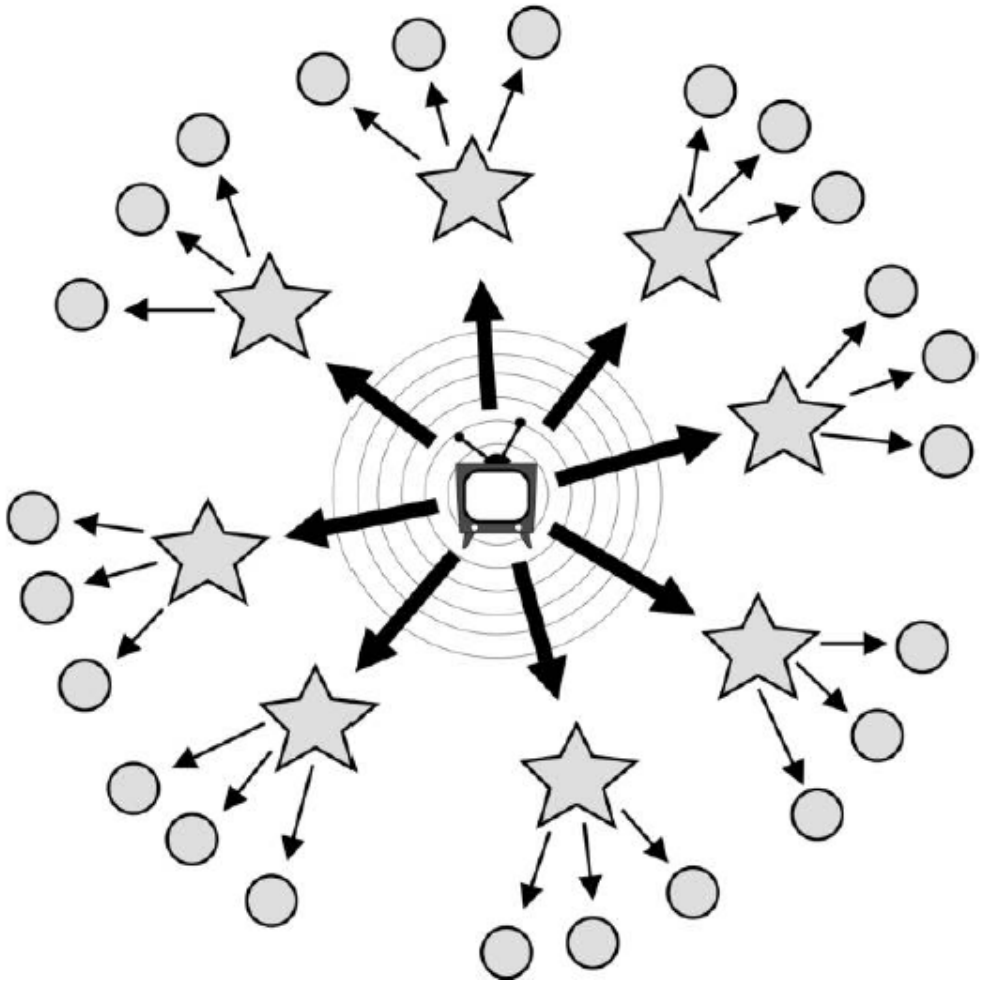
Railroads

Brain

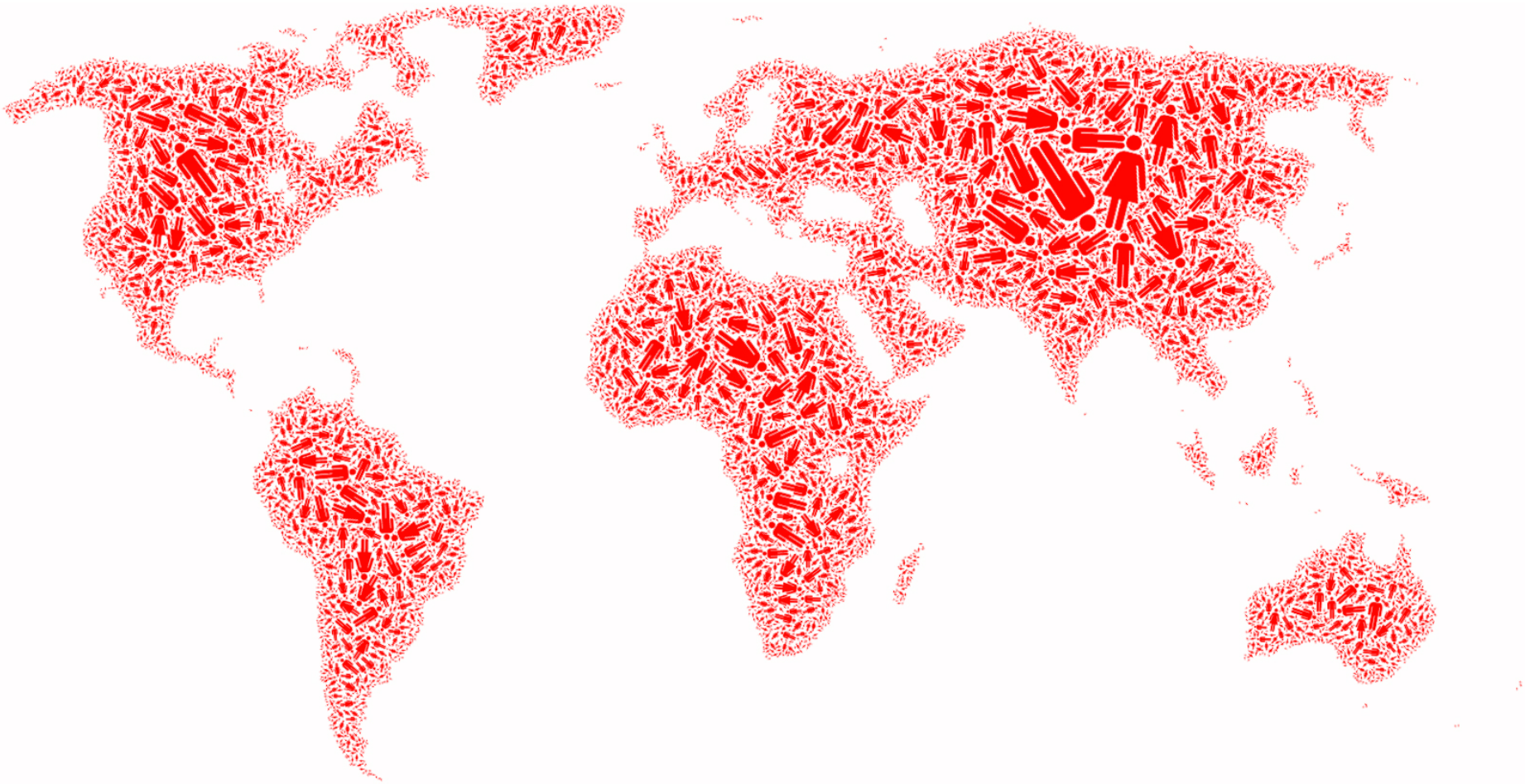


Internet

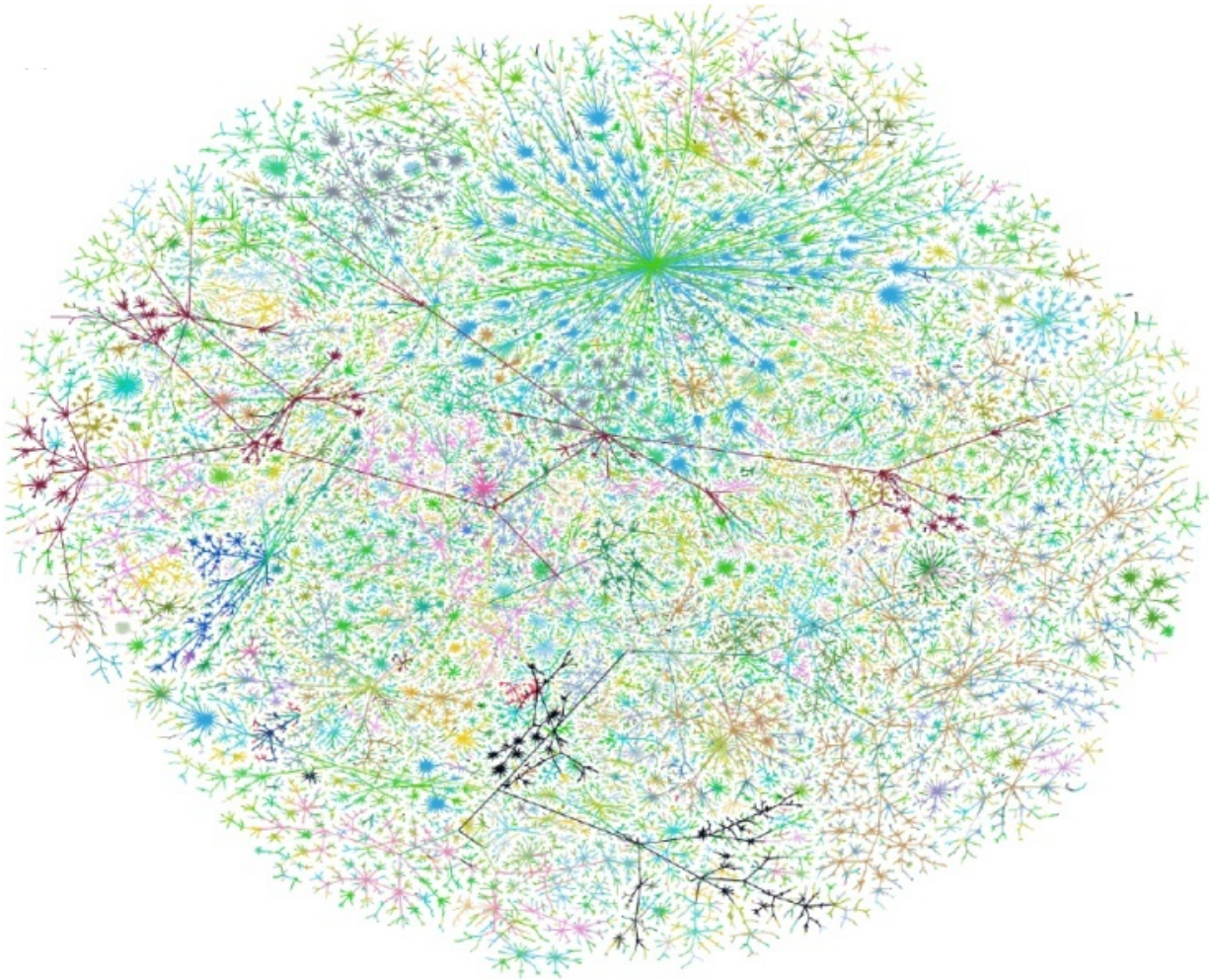
Friends & Family



Media & Information



Society



The Network!

Networks!!

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

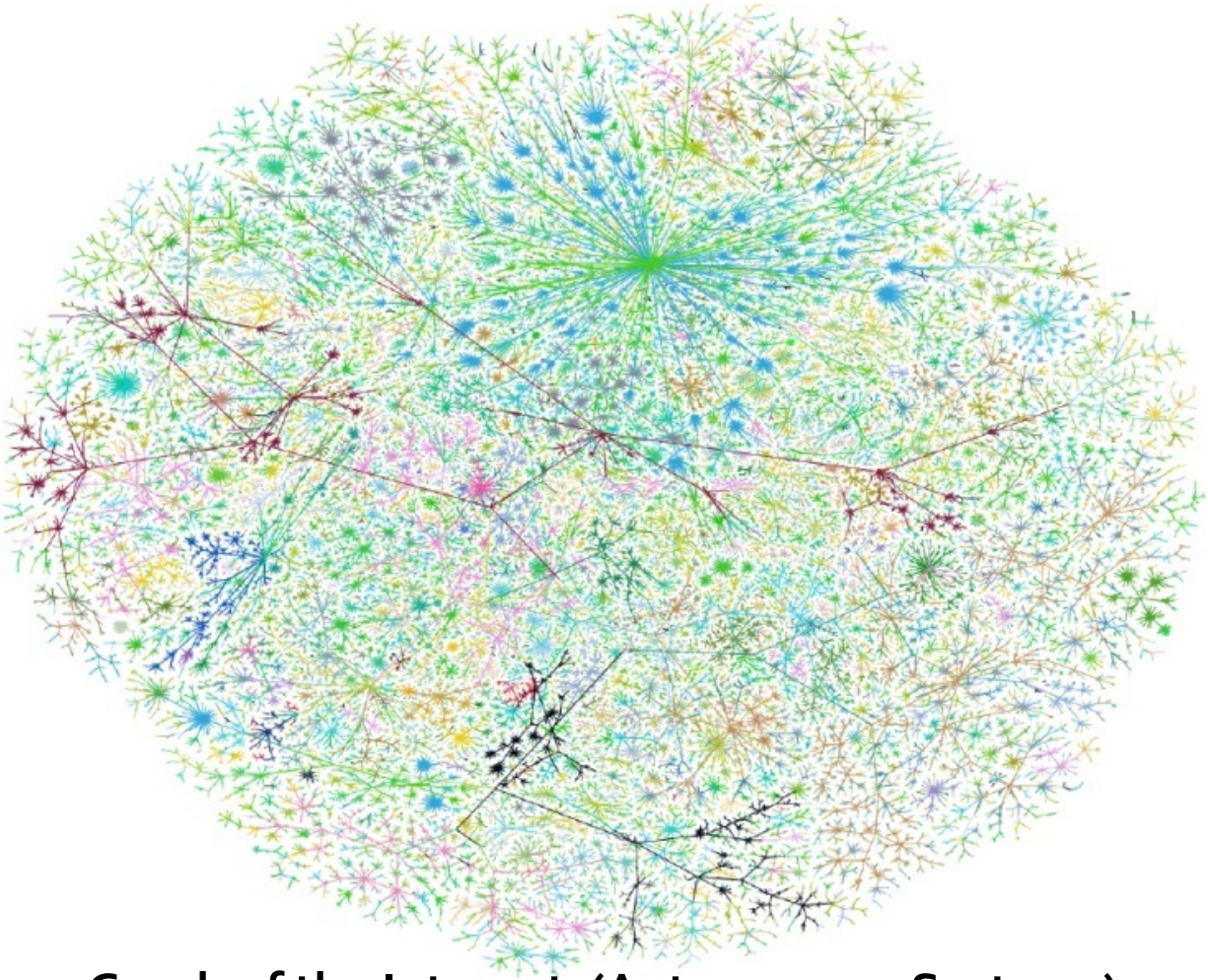
We will never understand these systems unless we understand the networks behind them!

Networks: Social

Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Networks: Communication

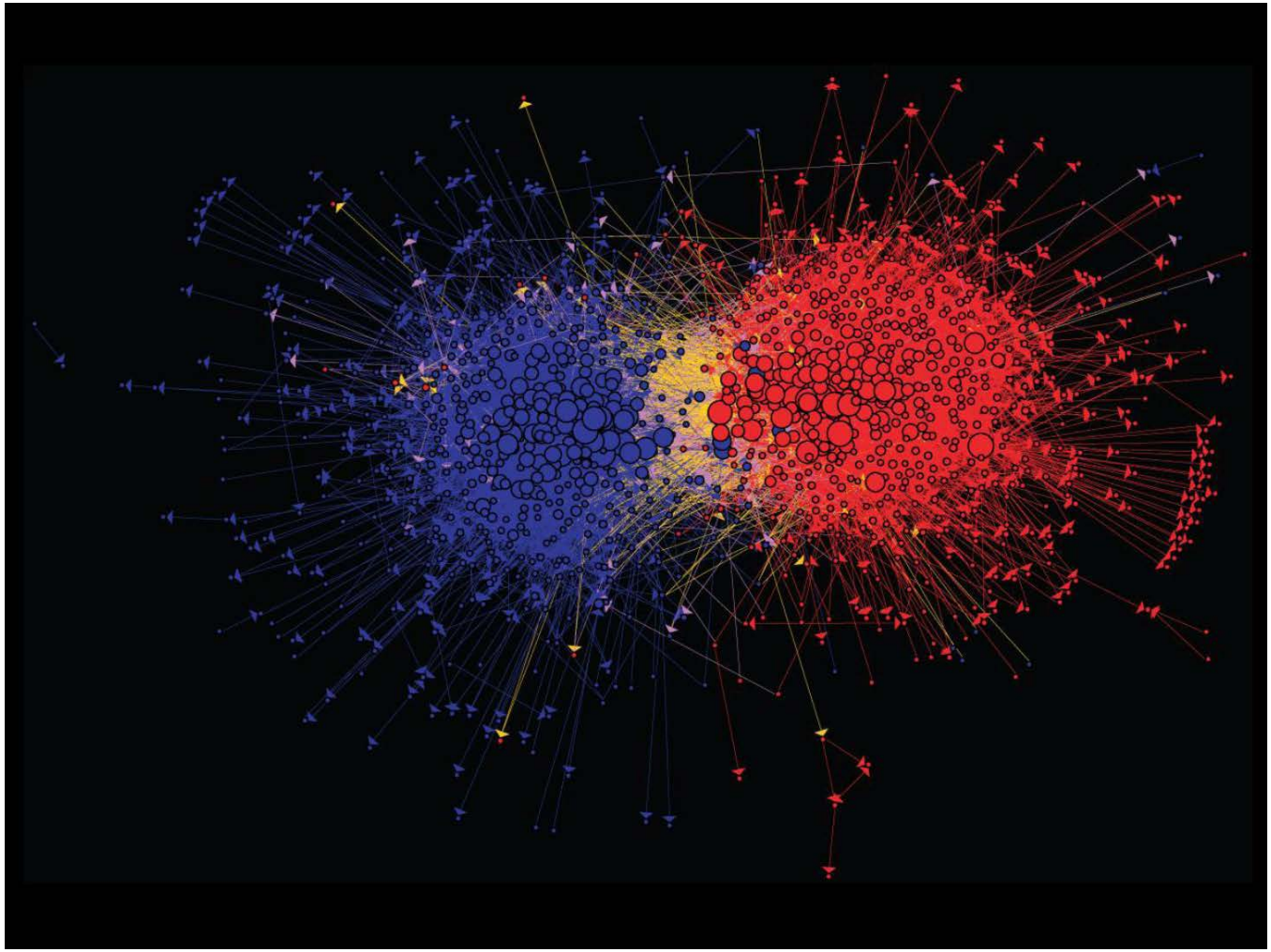


Graph of the Internet (Autonomous Systems)

Power-law degrees [Faloutsos-Faloutsos-Faloutsos, 1999]

Robustness [Doyle-Willinger, 2005]

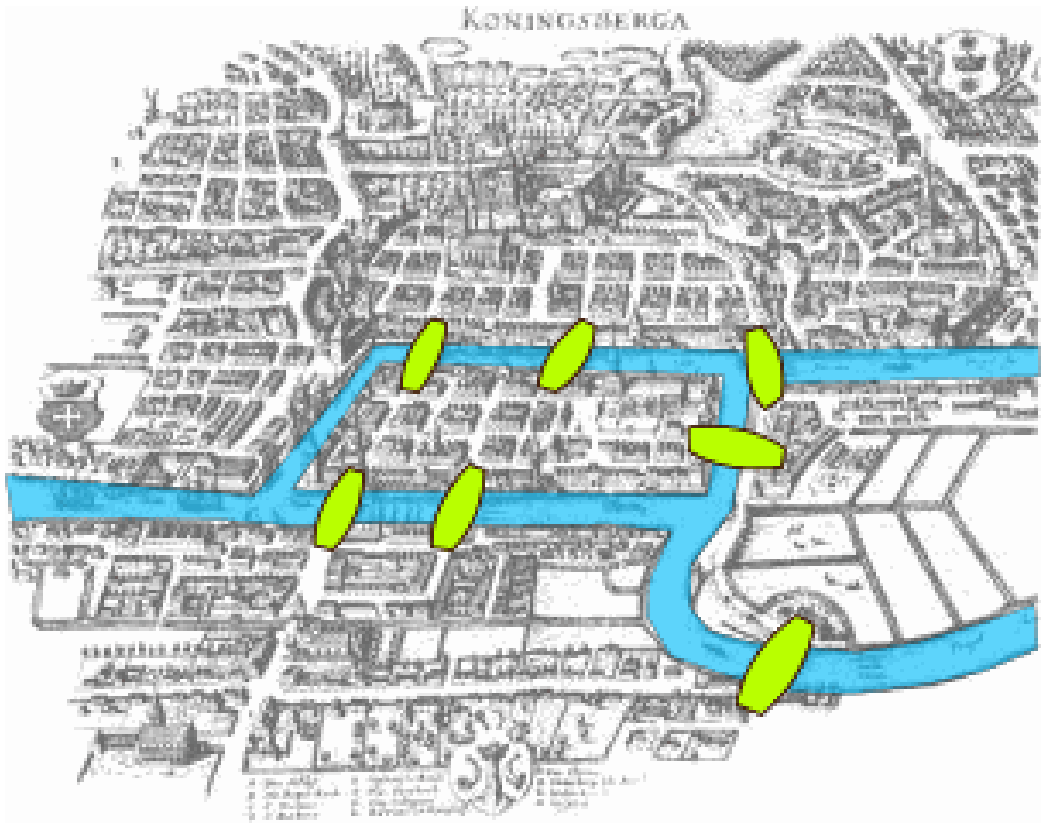
Networks: Media



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

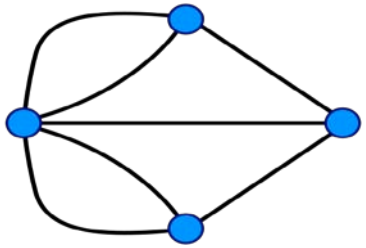
Networks: Technology



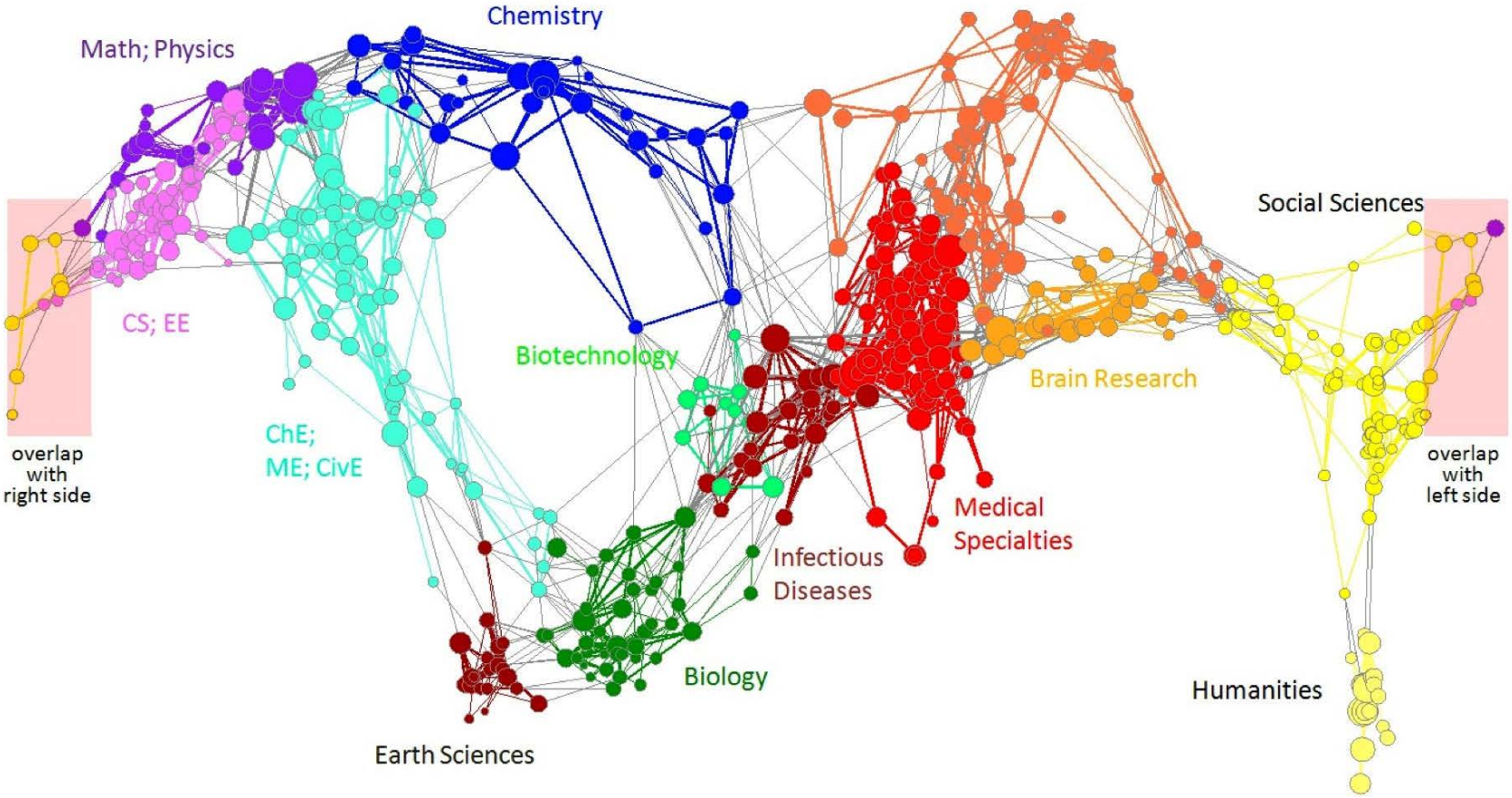
Seven Bridges of Königsberg

[Euler, 1735]

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



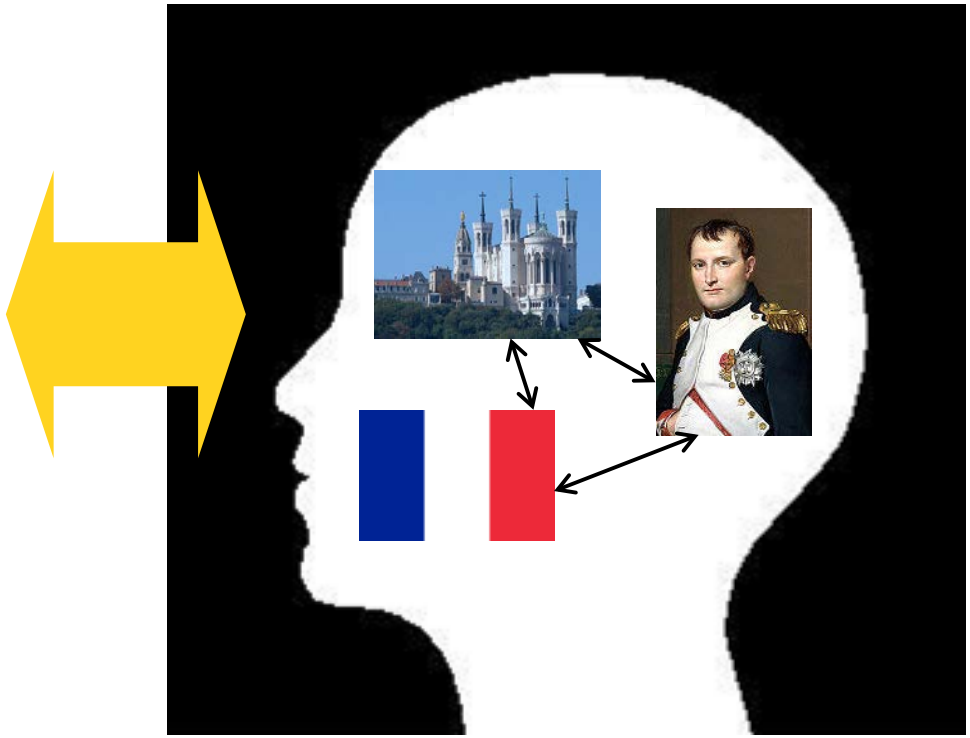
Networks: Information



Citation networks and Maps of science

[Börner et al., 2012]

Networks: Knowledge

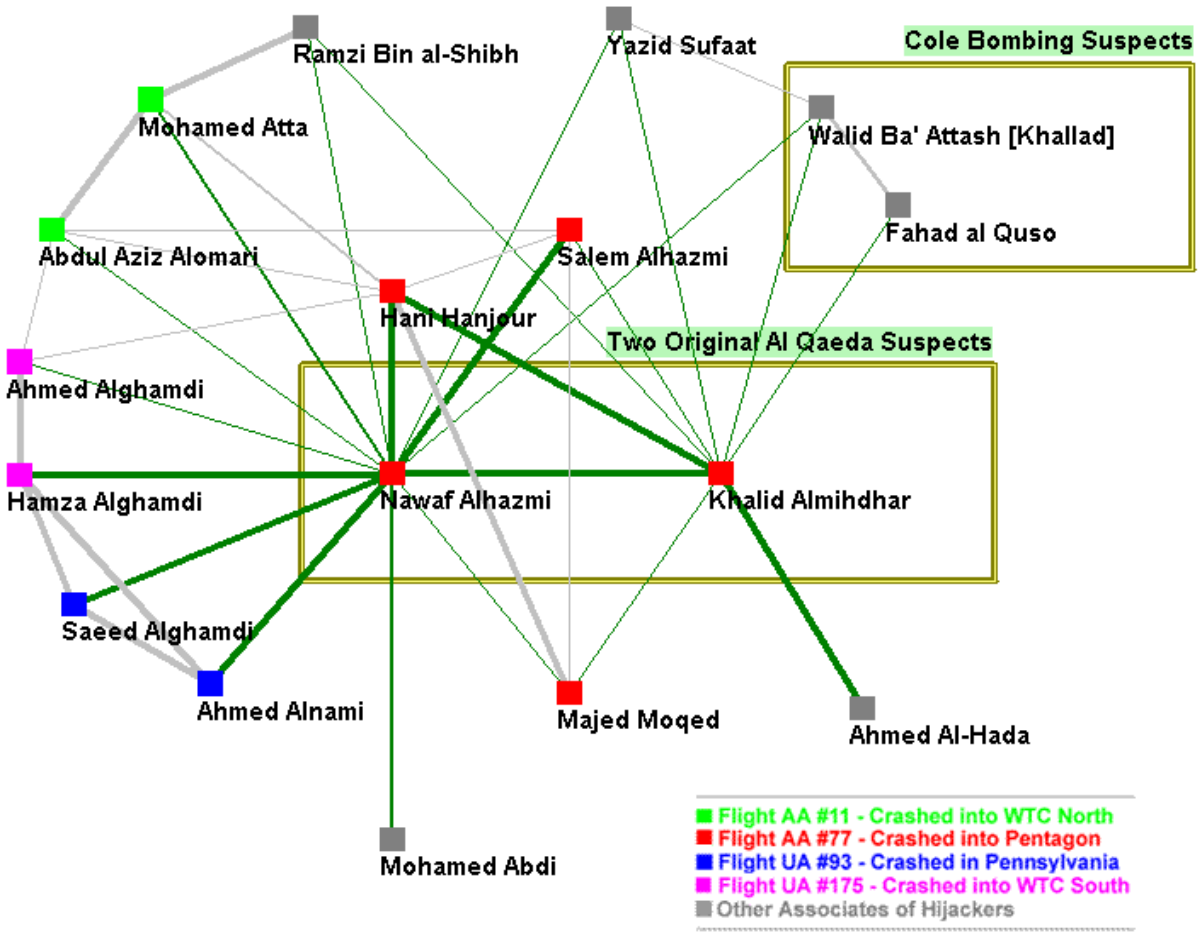


Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]

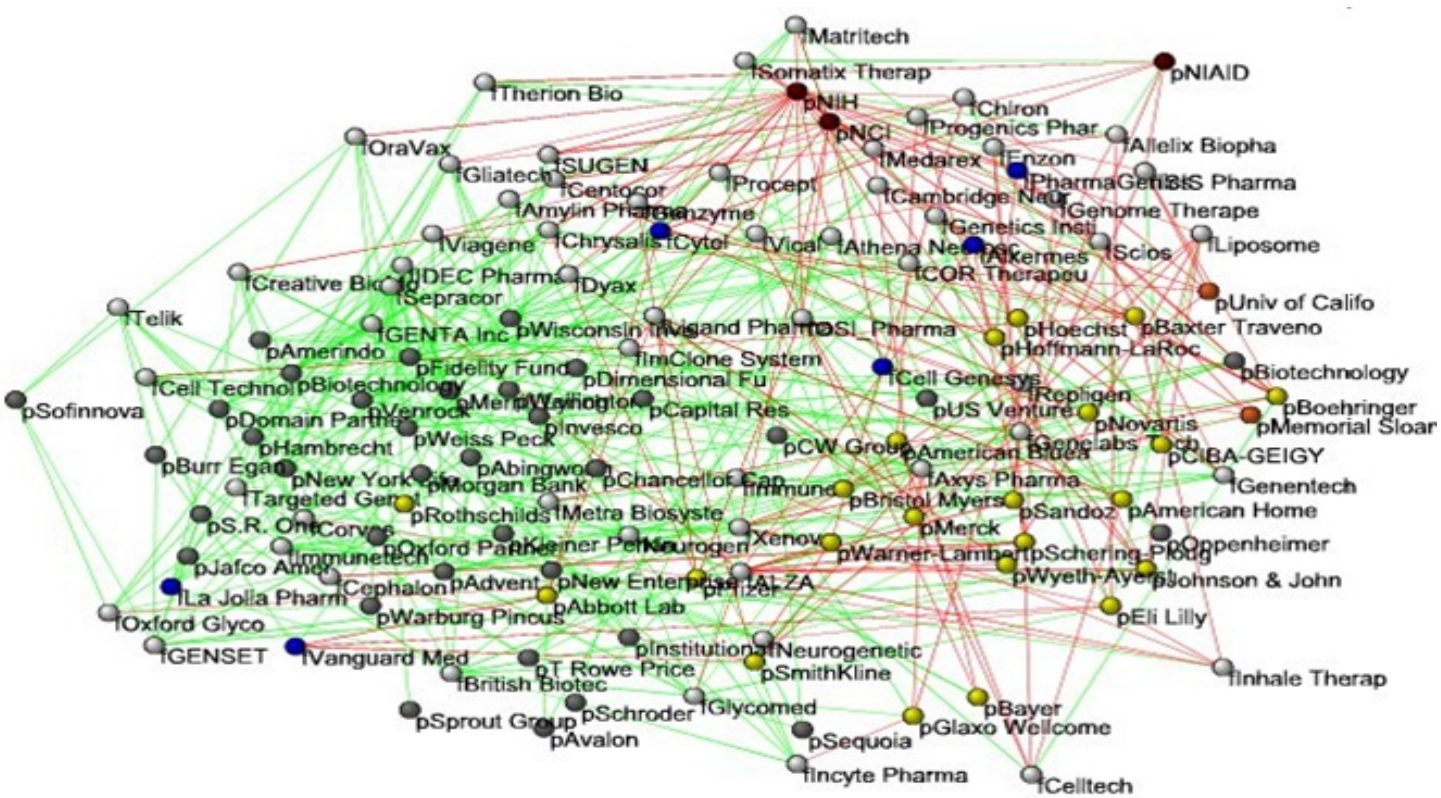
Networks: Organizations



9/11 terrorist network

[Krebs, 2002]

Networks: Economy



Nodes:

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

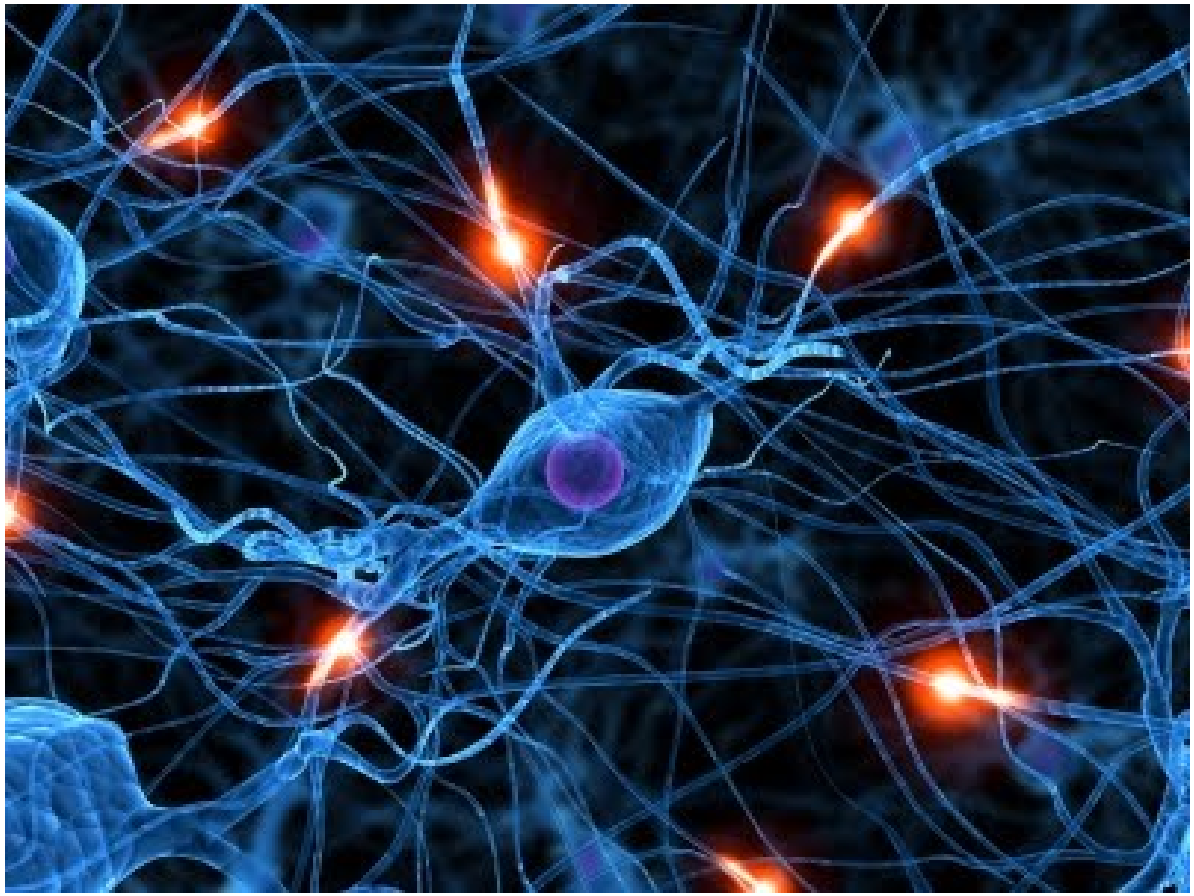
Links:

- Collaborations
- Financial
- R&D

Bio-tech companies

[Powell-White-Koput, 2002]

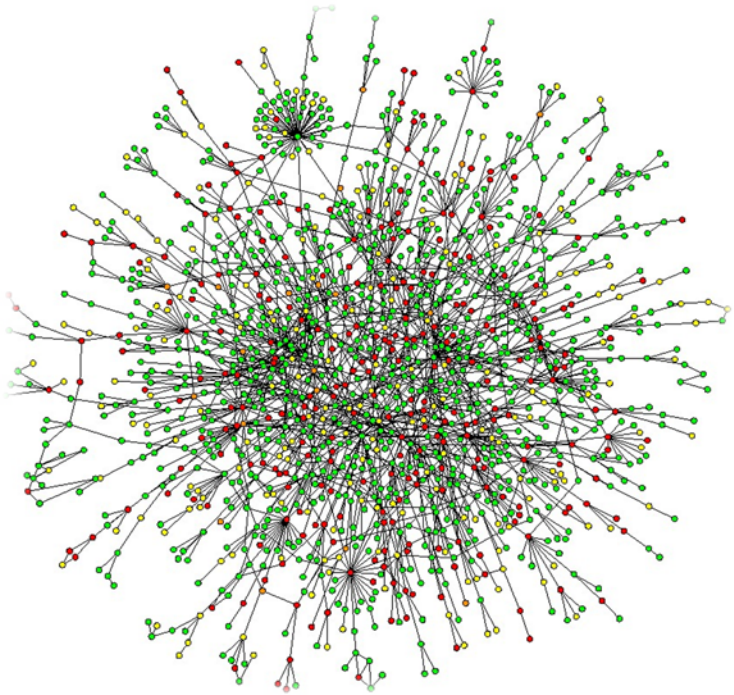
Networks: Brain

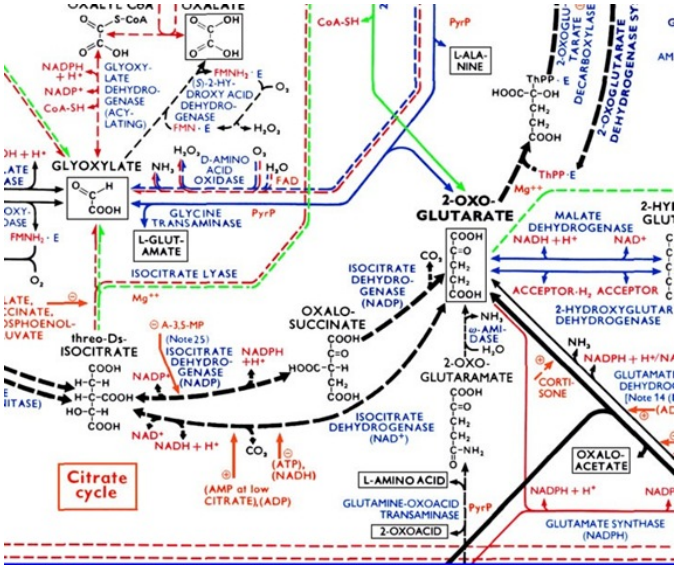


Human brain has between 10-100 billion neurons

[Sporns, 2011]

Networks: Biology





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

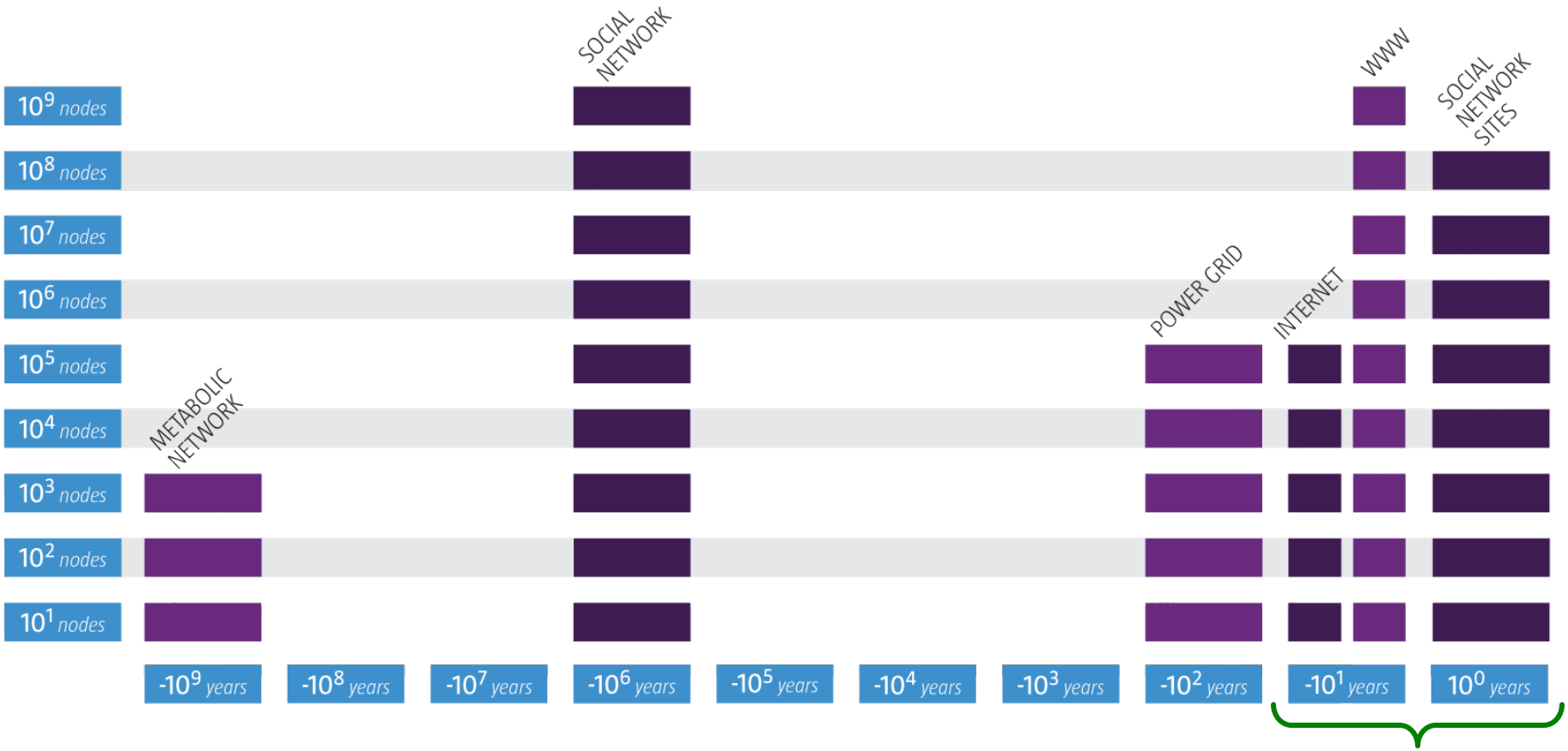
Metabolic networks:
Nodes: Metabolites and enzymes
Edges: Chemical reactions

But Jure, why should I care about networks?

Why Networks? Why Now?

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect
- **Shared vocabulary between fields**
  - Computer Science, Social science, Physics, Economics, Statistics, Biology
- **Data availability (/computational challenges)**
  - Web/mobile, bio, health, and medical
- **Impact!**
  - Social networking, Social media, Drug design

Networks: Why Now?



Age and size of networks

CS!!

Networks: Size Matters

- **Network data: Orders of magnitude**
  - 436-node network of email exchange at a corporate research lab [Adamic-Adar, SocNets '03]
  - 43,553-node network of email exchange at an 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]

Web – The Lab for Humanity

The Web is our "laboratory" for understanding the pulse of humanity.

Networks: Impact

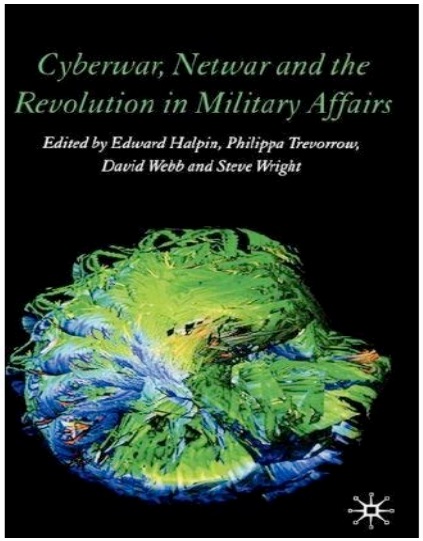
- **Google**
  Market cap: $300 billion
  (1y ago it was 250b)
- **Cisco**
  Market cap: $130 billion
  (1y ago it was 100b)
- **Facebook**
  Market cap: $114 billion
  (1y ago it was 50b)

Networks: Online

- **Communication networks:**
  - Intrusion detection, fraud
  - Churn prediction
- **Social networks:**
  - Link prediction, friend recommendation
  - Social circle detection, community detection
  - Social recommendations
  - Identifying influential nodes, Viral marketing
- **Information networks:**
  - Navigational aids

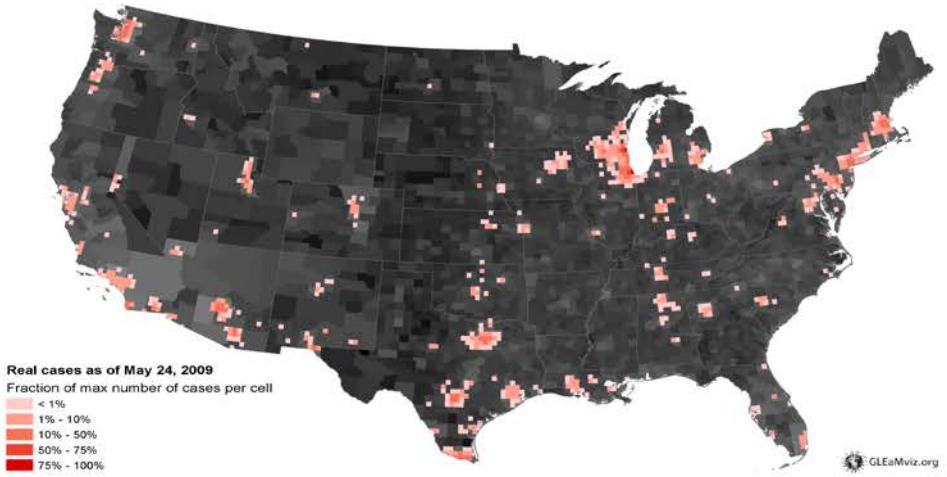
Networks: Impact

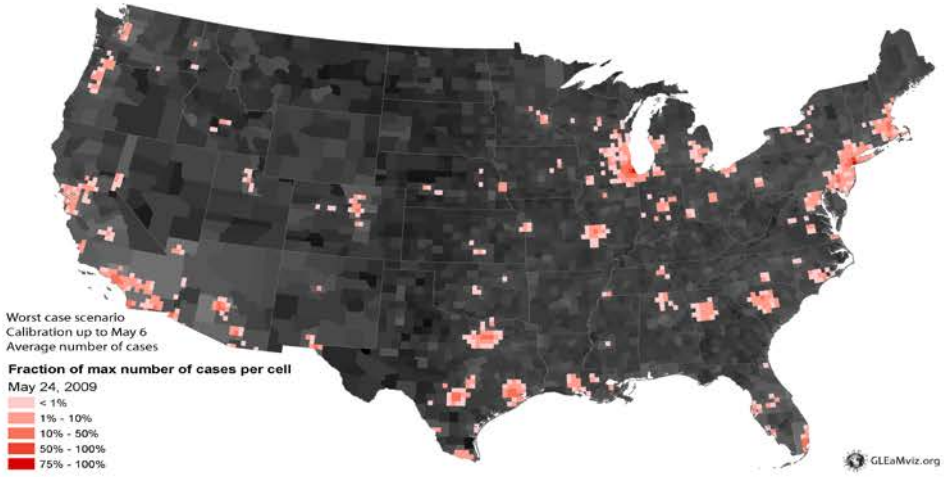
- **Intelligence and fighting (cyber) terrorism**



Networks: Impact

- **Predicting epidemics**





Real **Predicted**

Networks Really Matter

- If you were to understand the spread of diseases, **can you do it without social networks?**
- If you were to understand the structure of the Web, **it is hopeless without working with the Web's topology**
- If you want to understand dissemination of news or evolution of science, **it is hopeless without considering the information networks**

About the course

Reasoning about Networks

- **What do we hope to achieve from studying 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

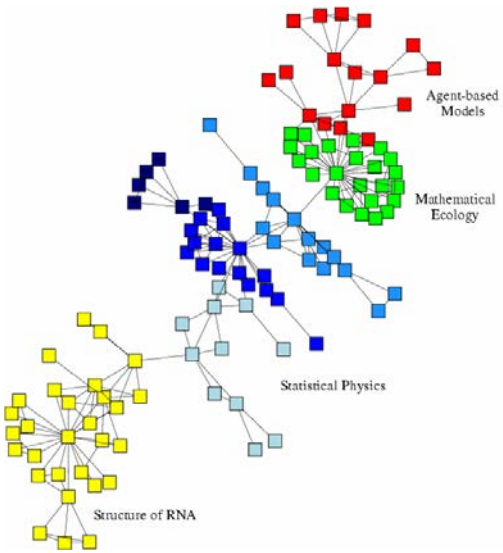
Reasoning about Networks

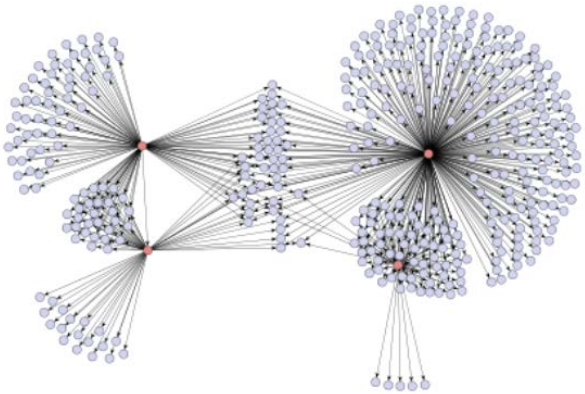
- **How do we reason about networks?**
  - **Empirical:** Study network data to find organizational principles
    - How do we measure and quantify networks?
  - **Mathematical models:** Graph theory, statistical models
    - Models allow us to understand behaviors and distinguish surprising from expected phenomena
  - **Algorithms** for analyzing graphs
    - Hard computational challenges

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
  - How do information and diseases spread?





How It All Fits Together

Observations	Models	Algorithms
Small diameter, Edge clustering	Small-world model, Erdös-Renyi model	Decentralized search
Scale-free	Preferential attachment, Copying model	PageRank, Hubs and authorities
Strength of weak ties, Core-periphery	Kronecker Graphs	Community detection: Girvan-Newman, Modularity
Densification power law, Shrinking diameters	Microscopic model of evolving networks	Link prediction, Supervised random walks
Patterns of signed edge creation	Structural balance, Theory of status	Models for predicting edge signs
Viral Marketing, Blogosphere, Memetracking	Independent cascade model, Game theoretic model	Influence maximization, Outbreak detection, LIM

Logistics: Course Assistants

Justin Cheng (head TA)

Christie Brandt

Peter Lofgren

Ashley Jin

Yoni Donner

Ashwin Apte

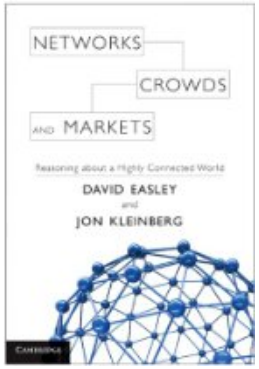
Zhemin Li

Bell Wang

See course website for office hour schedule!

Logistics: Website

- http://cs224w.stanford.edu
  - Slides posted just before the class
- Readings:
  - Mostly chapters from Easley&Kleinberg
  - Papers
- Optional readings:
  - Papers and pointers to additional literature
  - **This will be very useful for project proposals**



Logistics: Communication

- **Piazza Q&A website:**
  - http://piazza.com/stanford/fall2013/cs224w
  - **Please participate and help each other. Don't wait on TAs to wake up and reply!**
- **For e-mailing course staff, always use:**
  - cs224w-aut1314-staff@lists.stanford.edu
- **We will post course announcements to Piazza (make sure you check it regularly)**

Homeworks, Write-ups

- **Assignments are long and take time. Start early!**
  - A combination of: Data analysis, Algorithm design, and Math
- **How to submit?**
  - **Paper (Print code!):** In class and in cabinet in Gates
  - **SCPD:** Submit via SCPD
  - **Everyone: Code and write-ups** (proposal, milestone, final report) have to **also** be uploaded electronically at http://snap.stanford.edu/submit/
- **2 late periods for the quarter:**
  - 1 late period expires at the start of next class
  - Max 1 late period 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
  - Develop **scalable algorithms** for massive graphs
- **Performed in groups of 3 students**
- Project is the **main work** for the class
  - We will help with ideas, data and mentoring
  - Start thinking about this now
- Poster session with many external visitors

Course Schedule

Week	Assignment	Due on THU
2	**Homework 0**	October 3
3	**Homework 1**	October 10
4	**Project proposal**	October 17
5	**Homework 2**	October 24
	Work on the project	
7	**Homework 3**	November 7
8	**Project milestone**	November 14 (no late days!)
	Thanksgiving break	
9	**Homework 4**	November 21
	Project report	December 10, midnight (no late days!)
	Poster session	December 13 12-3pm

Work for the Course & Grading

- **Final grade will be composed of:**
  - **Homeworks: 48%**
    - Homeworks 1,2,3,4: 12% each
  - **Substantial class project: 50%**
    - Proposal: 20%
    - Project milestone: 20%
    - Final report: 50%
    - Poster presentation: 10%
  - **HW0, class+Piazza participation: 2%**

Prerequisites

- **No single topic in the course is too hard by itself**
- **But we will cover and touch upon many topics and this is what makes the course hard**
  - **Good background in:**
    - Algorithms and graph theory
    - Probability and Statistics
    - Linear algebra
  - **Programming:**
    - You should be able to write non-trivial programs
  - **4 recitation sessions:**
    - 2 to review programming tools (SNAP, SNAP.PY)
    - 2 to review basic mathematical concepts

Network Analysis Tools

- **We highly recommend SNAP:**
  - **SNAP C++:** more challenging but more scalable
  - **SNAP.PY:** Python ease of use, most of C++ scalability
    - We worked hard and just released SNAP.PY
    - **You (cs224w 2013 class) will help with the project!**
    - HW0 asks you to go over SNAP.PY tutorial and test functions, find bugs and help with documentation
      - This will benefit everyone in the class and beyond!
  - **We will give generous extra credit to students who will help us improve these tools!**
  - Other tools include NetworkX, JUNG, iGraph

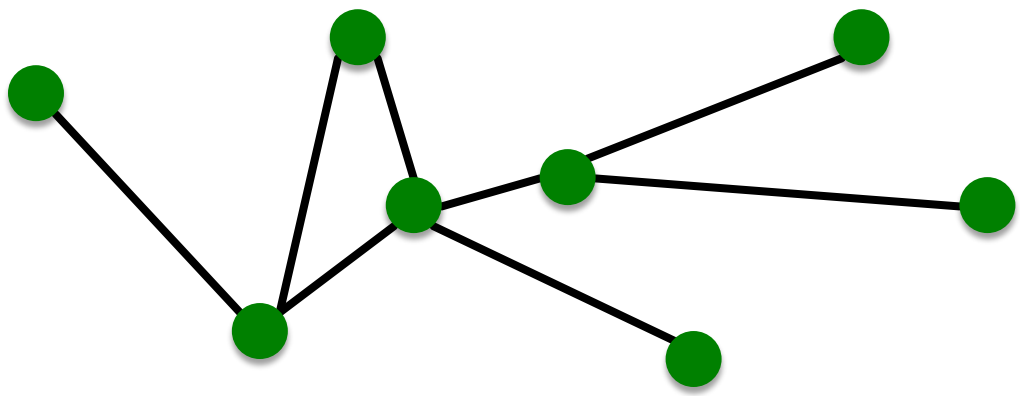
Starter Topic:
Structure of the Web Graph

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

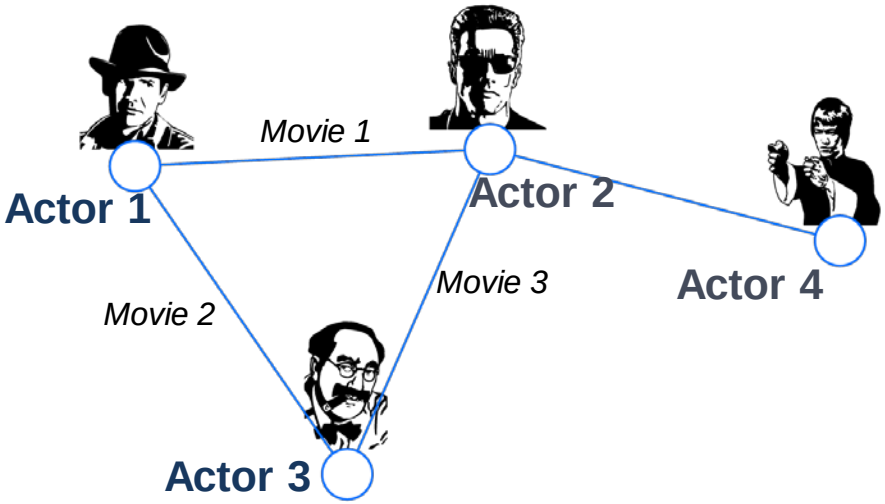
  Language: Network, node, link

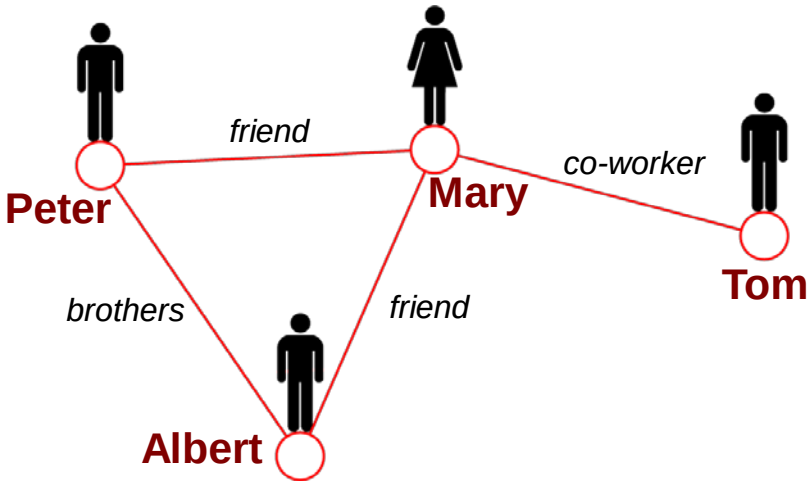
- **Graph** is mathematical representation of a network
  - Web graph, Social graph (a Facebook term)

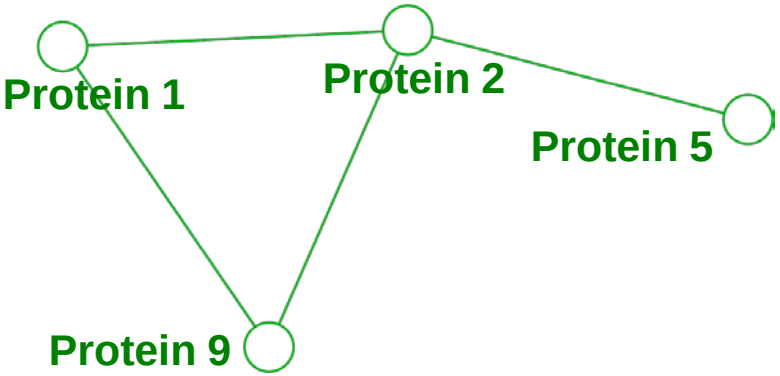
  Language: Graph, vertex, edge

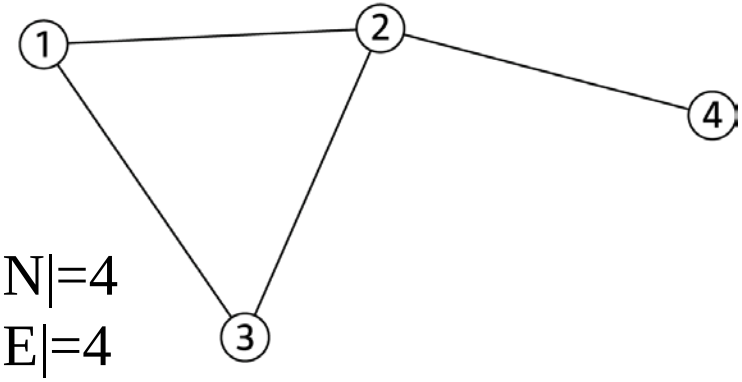
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably

Networks: Common Language









$|N|=4$

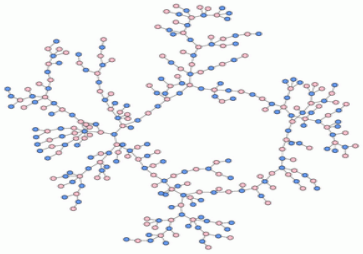
$|E|=4$

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**

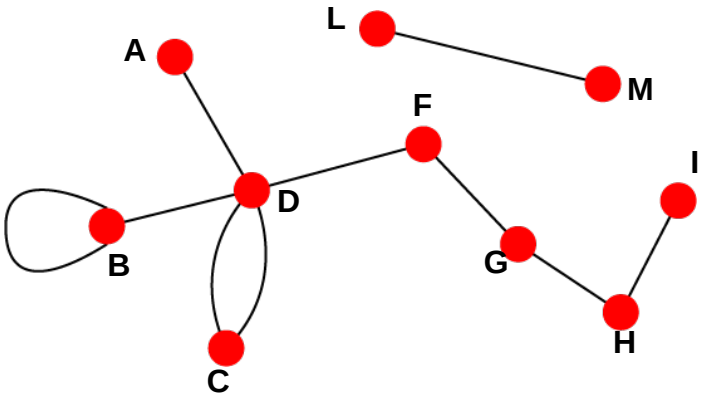


- **If you connect all papers with the same word in the title, you will be exploring what?** It is a network, nevertheless

Undirected vs. Directed Networks

Undirected

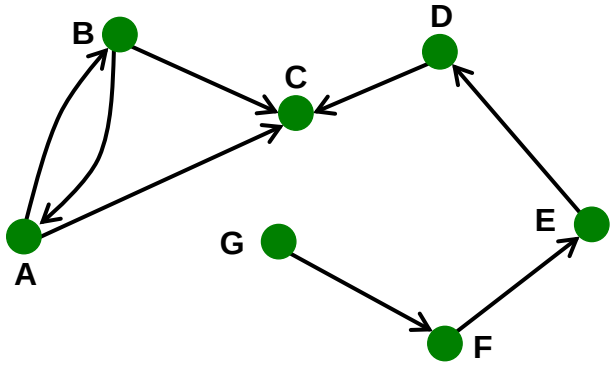
- **Links:** undirected (symmetrical, reciprocal)



- **Examples:**
  - Collaborations
  - Friendship on Facebook

Directed

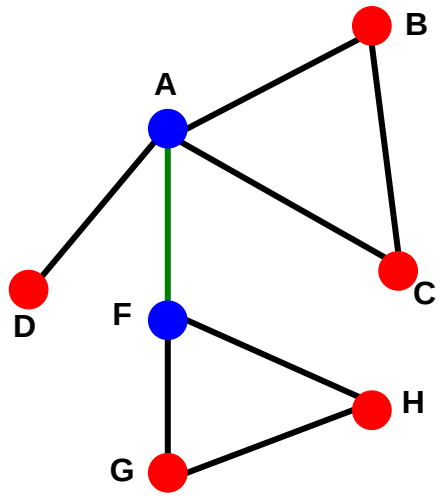
- **Links:** directed (arcs)

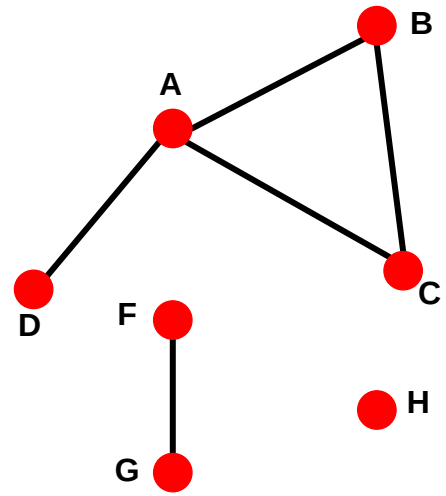


- **Examples:**
  - Phone calls
  - Following on Twitter

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





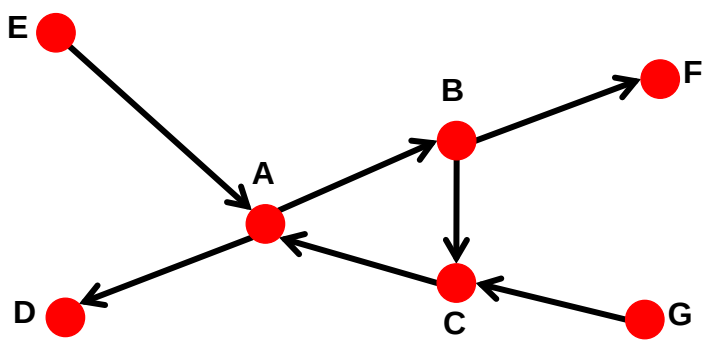
Largest Component:
Giant Component

Isolated node (node H)

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

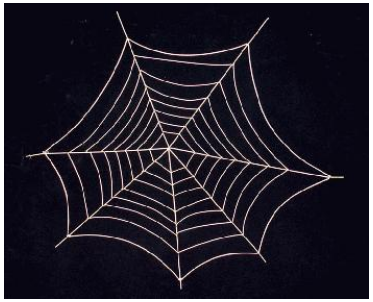
Connectivity of Directed Graphs

- **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 connected but not strongly connected (e.g., there is no way to get from F to G by following the edge directions).

Web as a Graph

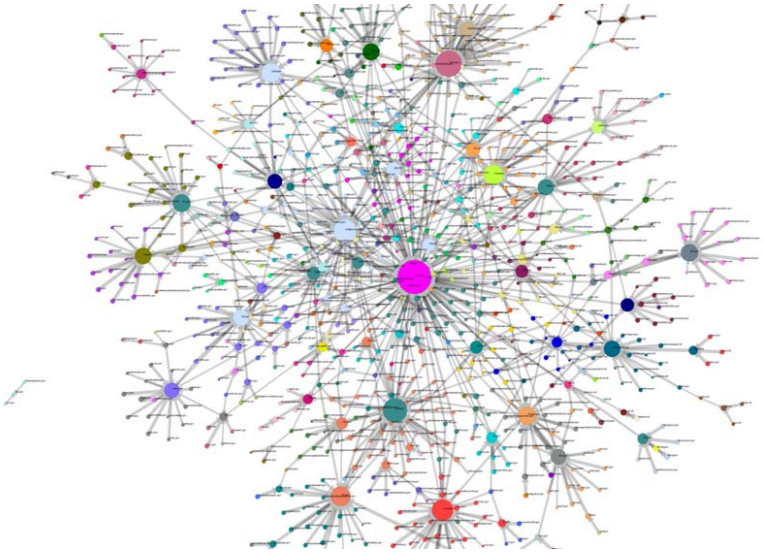


- **Q: What does the Web "look like"?**
- **Here is what we will do next:**
  - We will take a real system (i.e., the Web)
  - We will collect lots of Web data
  - We will represent the Web as a graph
  - We will use language of graph theory to reason about the structure of the graph
  - Do a computational experiment on the Web graph
  - **Learn something about the structure of the Web!**

Web as a Graph

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

- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks



  - **Side issue:** What is a node?
    - Dynamic pages created on the fly
    - "dark matter" – inaccessible database generated pages

The Web as a Graph

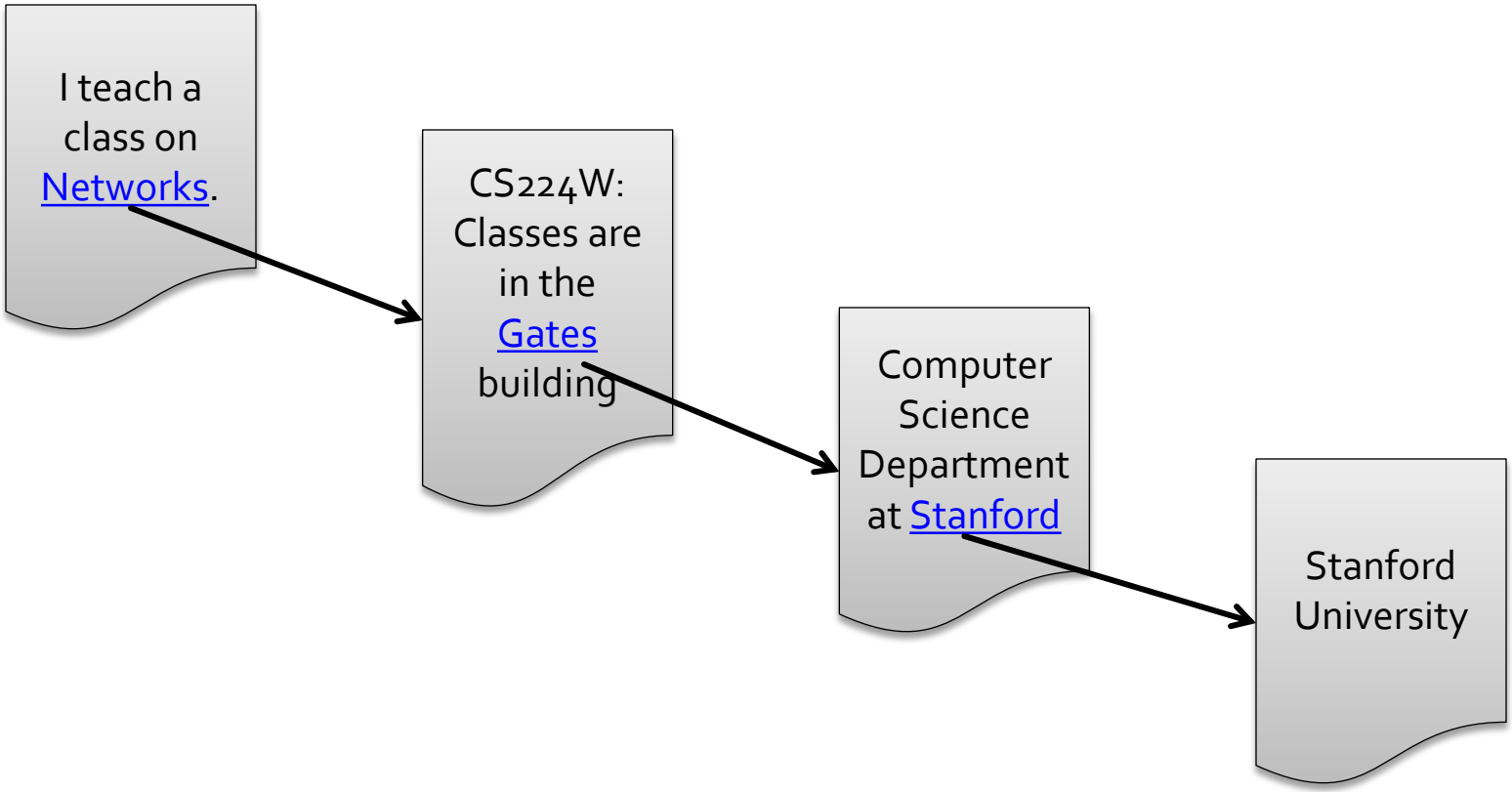
I teach a class on Networks.

CS224W: Classes are in the Gates building

Computer Science Department at Stanford

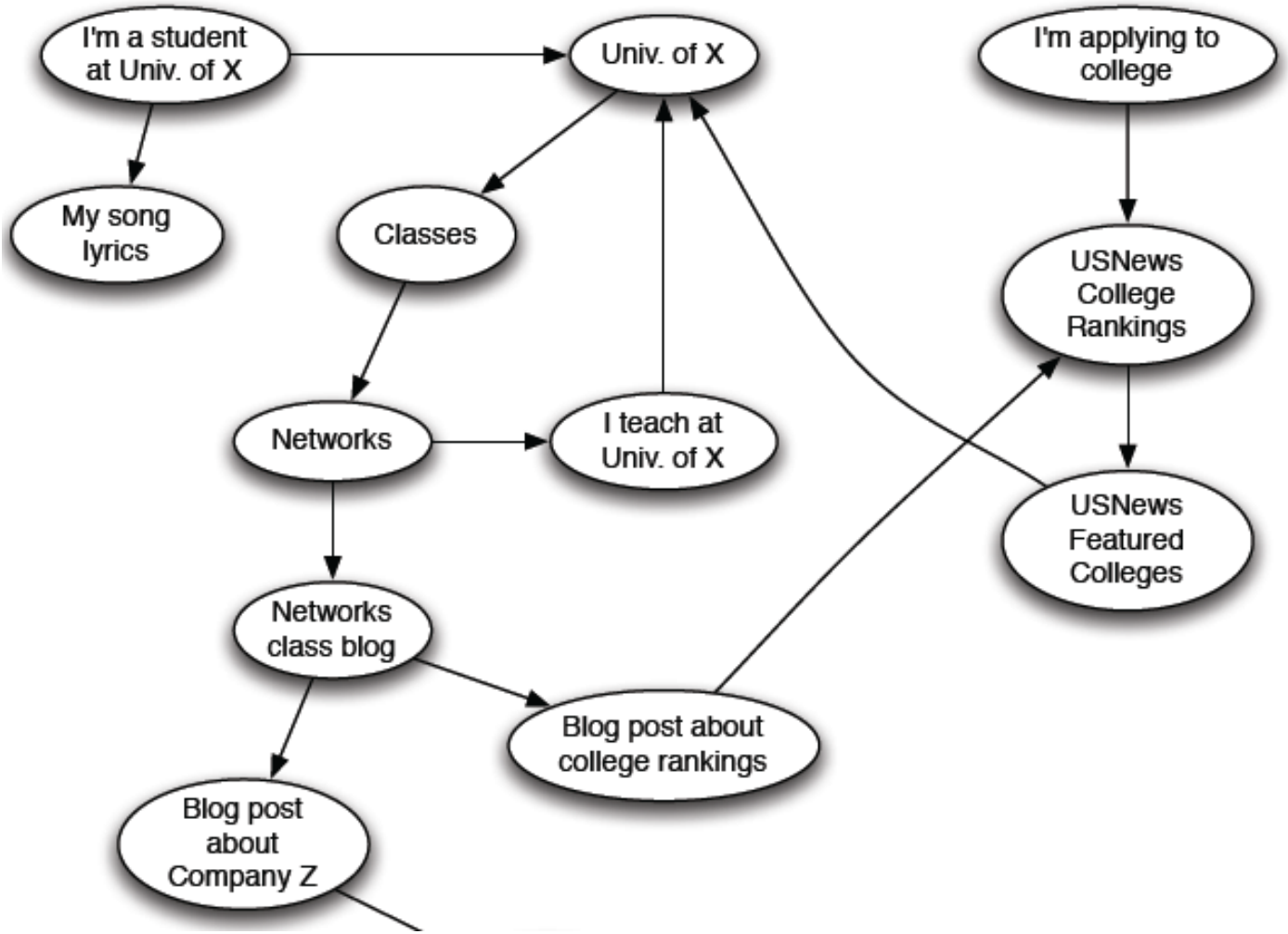
Stanford University

The Web as a Graph

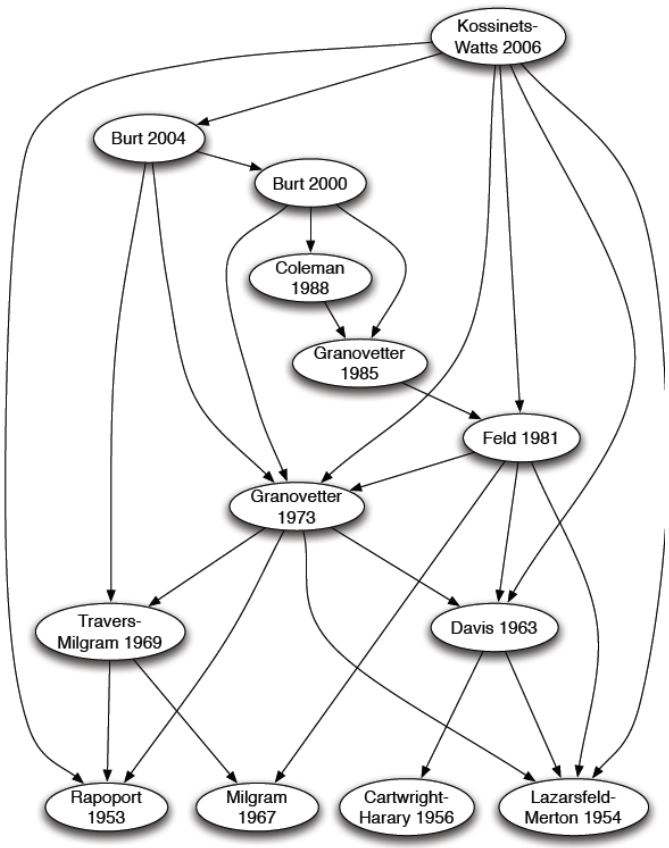


- In early days of the Web links were **navigational**
- Today many links are **transactional**

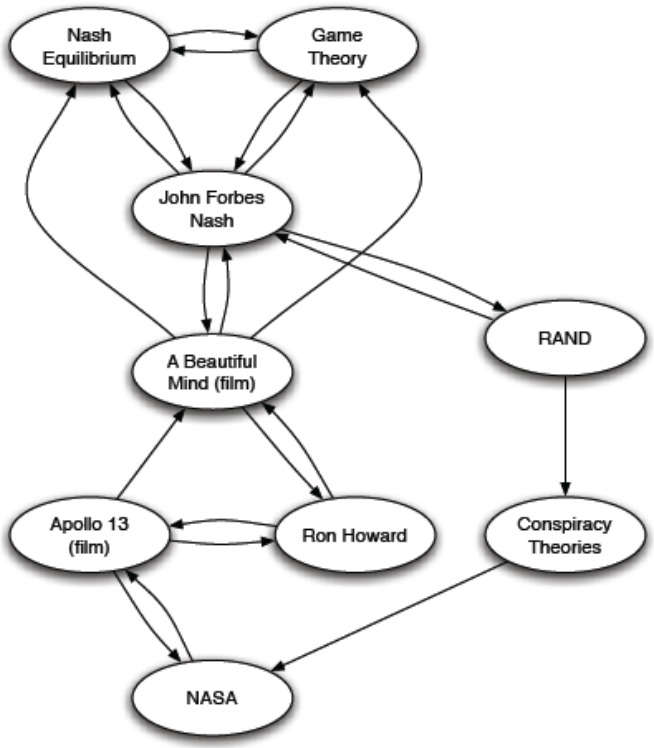
The Web as a Directed Graph



Other Information Networks



Citations



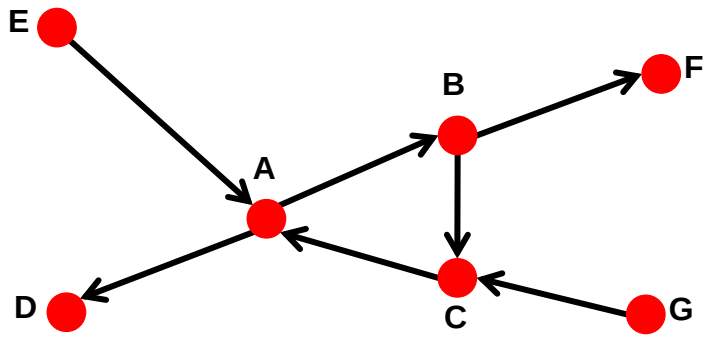
References in an Encyclopedia

What 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***?



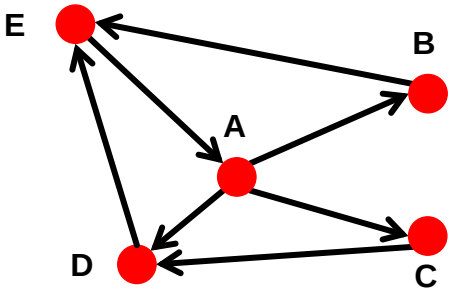
$$In(v) = \{w \mid w \text{ can reach } v\}$$
$$Out(v) = \{w \mid v \text{ can reach } w\}$$

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

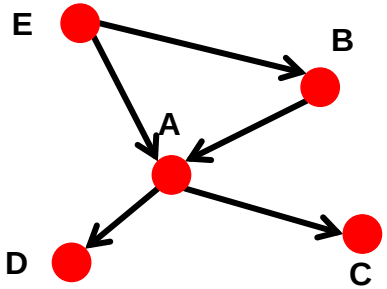
Directed Graphs

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

      $In(A)=Out(A)=\{A,B,C,D,E\}$



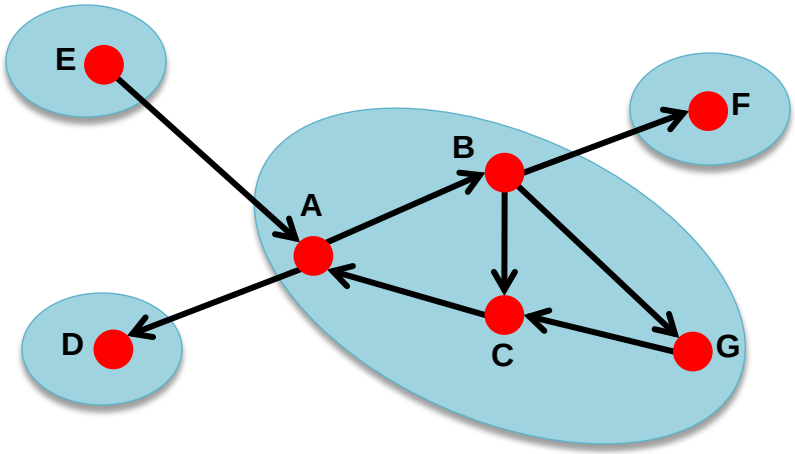
  - **DAG – Directed Acyclic Graph:**
    - Has no cycles: if $\boldsymbol{u}$ can reach $\boldsymbol{v}$, then $\boldsymbol{v}$ can not reach $\boldsymbol{u}$



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

Strongly Connected Component

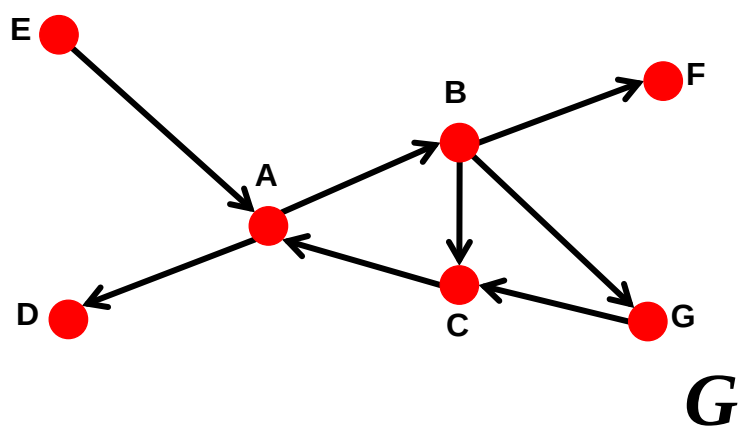
- **Strongly connected component (SCC)** is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



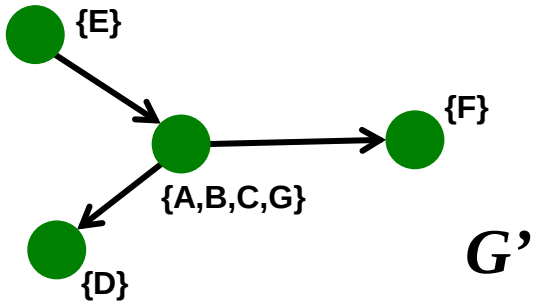
Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

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

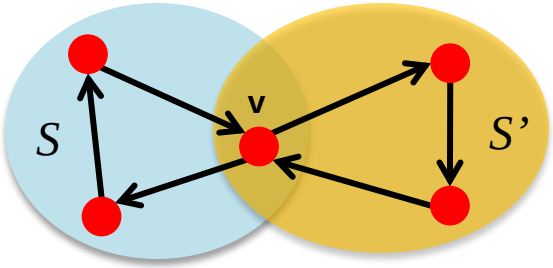


(1) Strongly connected components of graph G: {A,B,C,G}, {D}, {E}, {F}
(2) G' is a DAG:



Proof of (1)

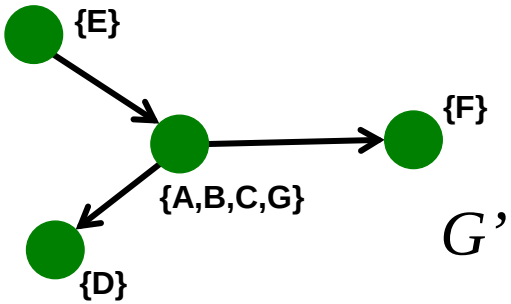
- **Claim: SCCs partitions nodes of G.**
  - This means: Each node is member of exactly 1 SCC
- Proof by contradiction:
  - Suppose there exists a node v which is a member of 2 SCCs S and S'

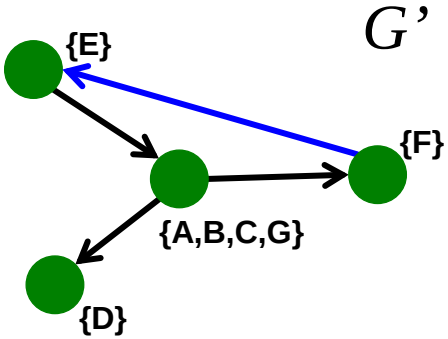


  - But then $S \cup S'$ is one large SCC!
    - Contradiction!

Proof of (2)

- **Claim: *G'* (graph of SCCs) is a DAG.**
  - This means: ***G'*** has no cycles
- Proof by contradiction:
  - Assume ***G'*** is not a DAG
  - Then ***G'*** has a directed cycle
  - Now all nodes on the cycle are mutually reachable, and all are part of the same SCC
  - But then ***G'*** is not a graph of connections between SCCs (SCCs are defined as maximal sets)
    - Contradiction!



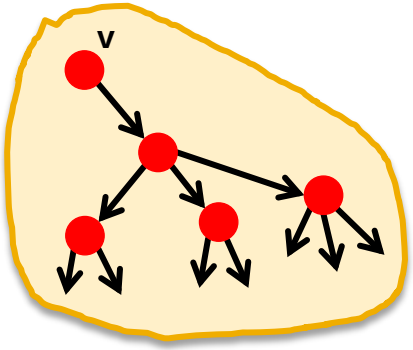


Now {A,B,C,G,E,F} is a SCC!

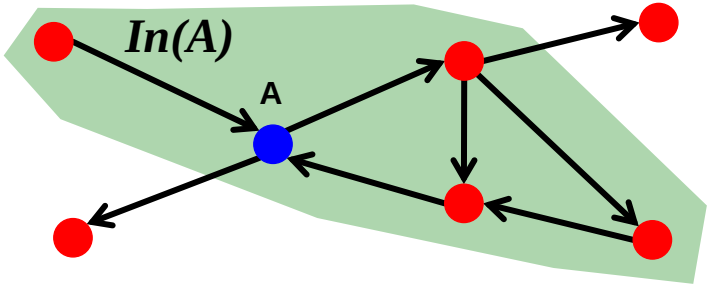
Graph Structure of the Web

- **Goal:** **Take a large snapshot of the Web and try to understand how its SCCs "fit together" as a DAG**
- **Computational issue:**
  - Want to find a 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

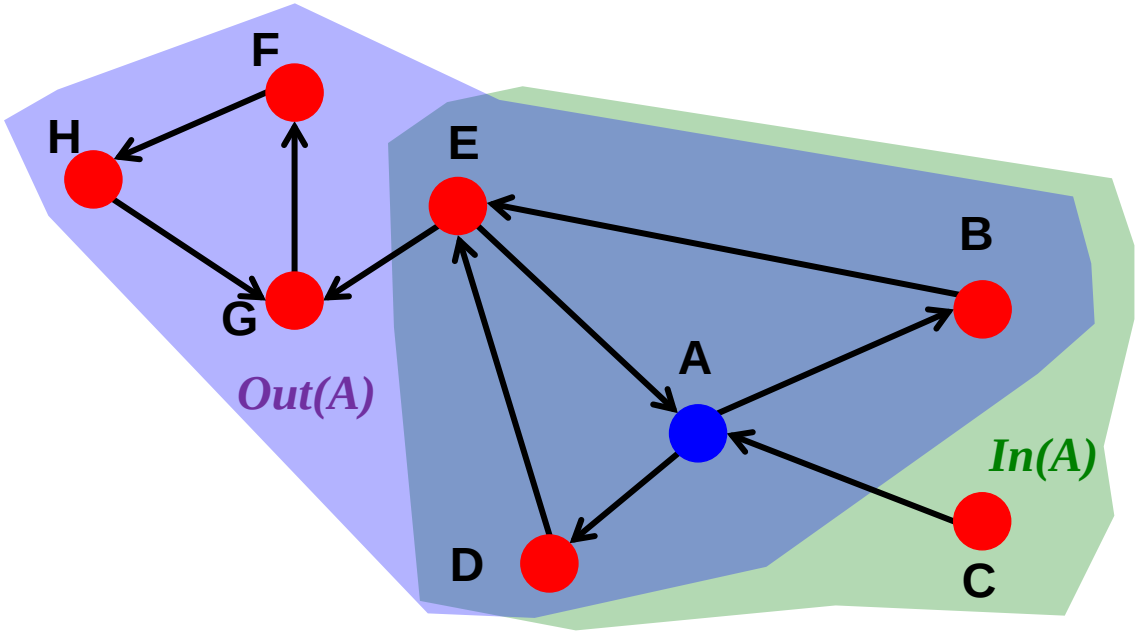


Out(v)



Out(A) ∩ In(A) = SCC

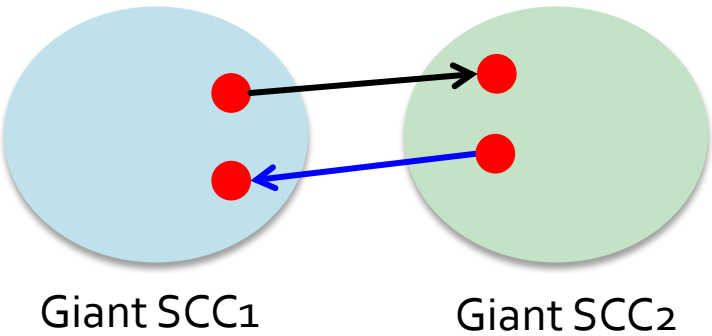
- **Example:**



- Out(A) = {A, B, D, E, F, G, H}
- In(A) = {A, B, C, D, E}
- So, SCC(A) = Out(A) ∩ In(A) = {A, B, D, E}

Graph Structure of the Web

- **There is a single giant SCC**
- **There won't be 2 giant SCCs**
- Heuristic argument:
  - It just takes 1 page from one SCC to link to the other SCC
  - If the 2 SCCs have millions of pages the likelihood of this not happening is very very small



Structure of the Web

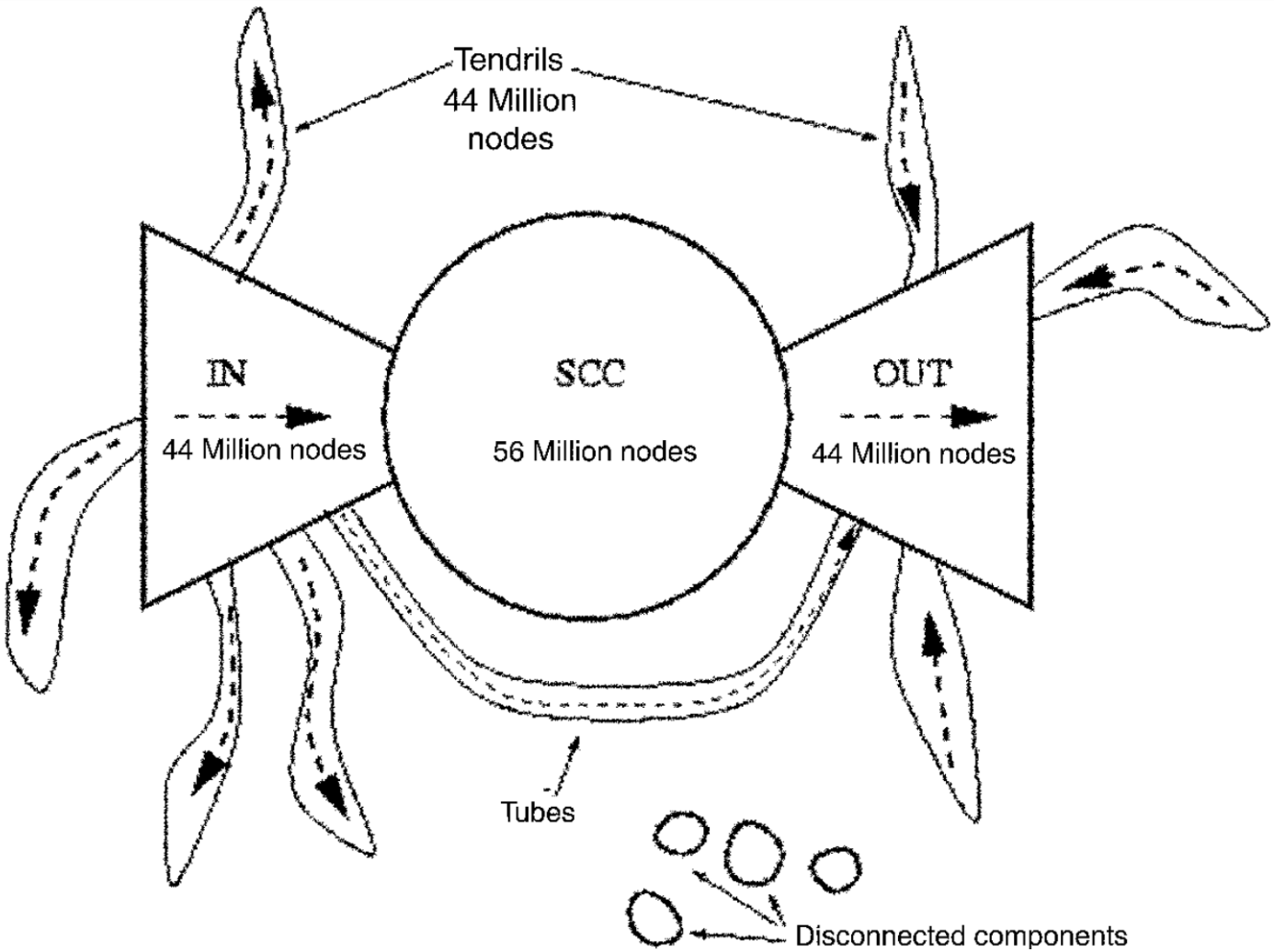
- **Broder et al., 2000:**
  - Altavista crawl from October 1999
    - 203 million URLS
    - 1.5 billion links
  - Computer: Server with 12GB of memory
- **Undirected version of the Web graph:**
  - 91% nodes in the largest weakly conn. component
  - Are hubs making the web graph connected?
    - Even if they deleted links to pages with in-degree >10 WCC was still ≈50% of the graph

Structure of the Web

- **Directed version of the Web graph:**
  - **Largest SCC:** 28% of the nodes (56 million)
  - Taking a random node v
    - **Out(v) ≈ 50%** (100 million)
    - **In(v) ≈ 50%** (100 million)

- **What does this tell us about the conceptual picture of the Web graph?**

Bow-tie Structure of the Web



203 million pages, 1.5 billion links [Broder et al. 2000]

What did We Learn/Not Learn ?

- **What did we learn:**
  - Some conceptual organization of the Web (i.e., the bowtie)
- **What did we 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: $\sim 1 / k^2$
    - 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.]