CS224W: Social and Information Network Analysis

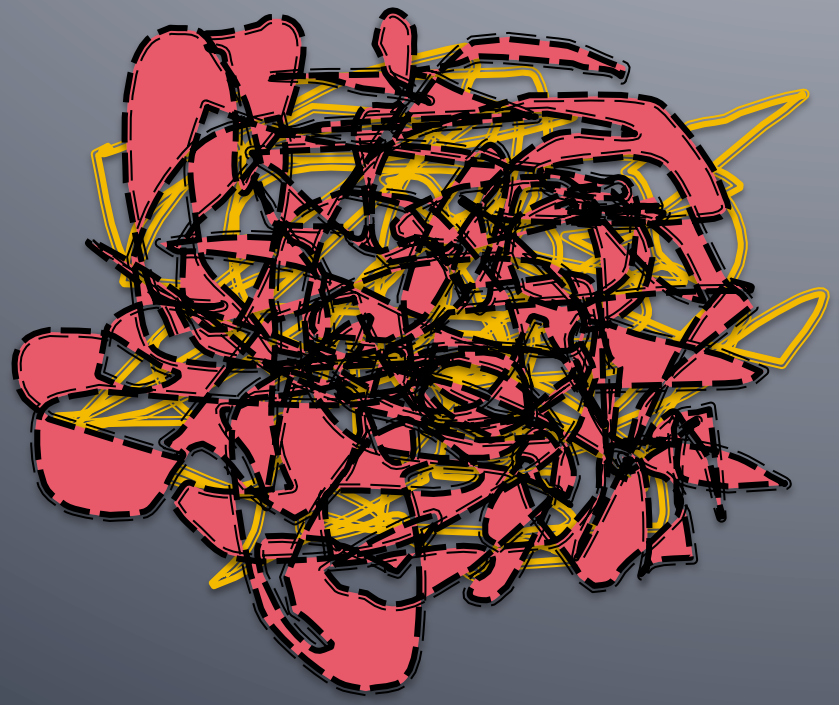
CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

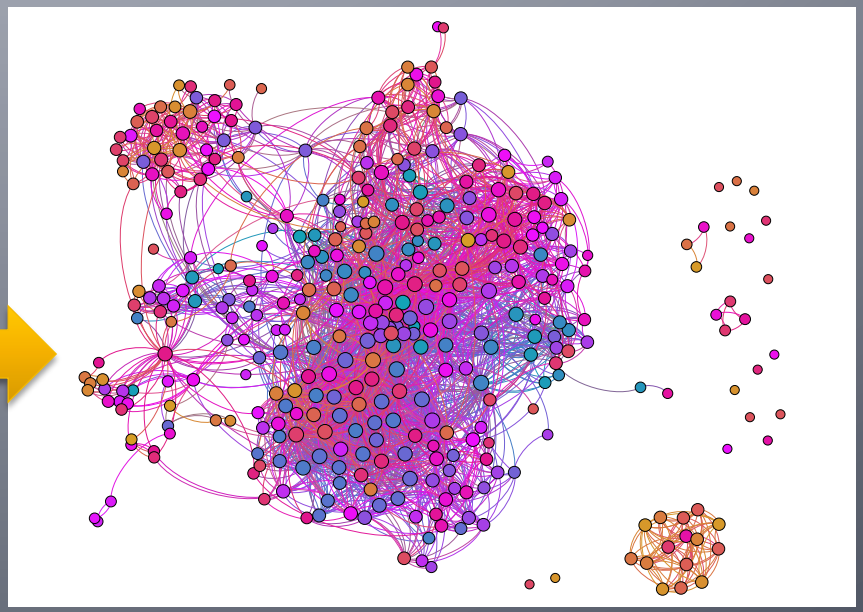
Networks & Complex Systems

- **Complex systems are around us:**
  - **Society** is a collection of six billion individuals
  - **Communication systems** link electronic devices
  - **Information** and **knowledge** is organized and linked
  - Interactions between thousands of **genes** regulate life
  - Our **thoughts** are hidden in the connections between billions of neurons in our brain

What do these systems have in common?
How can we represent them?

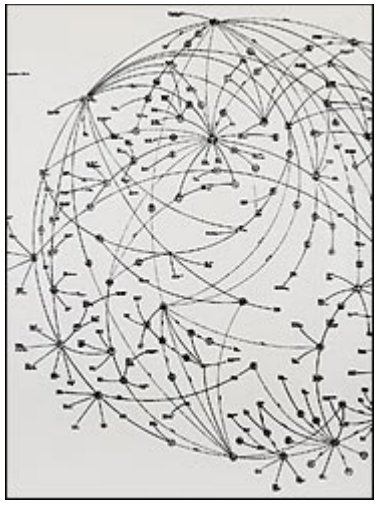
The network!

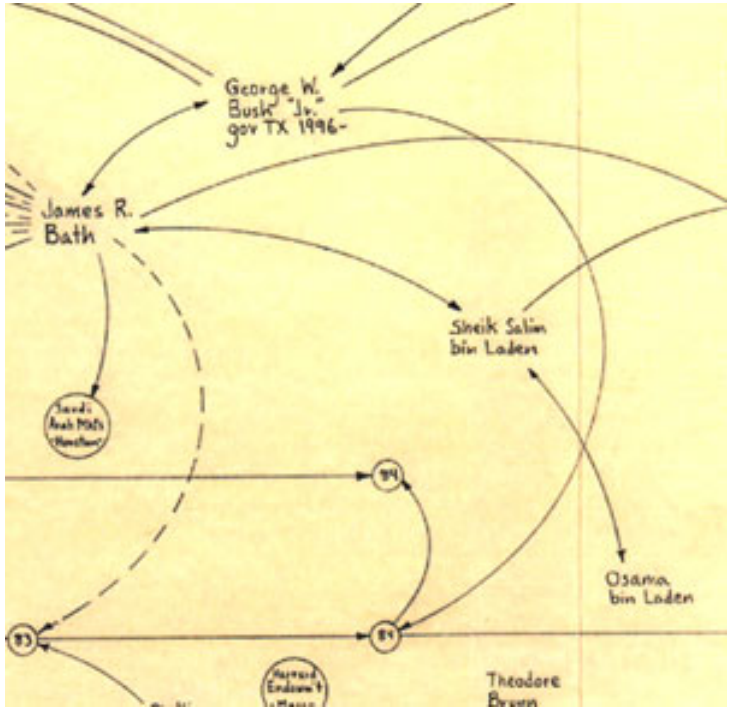




examples: Political/Financial Networks

- Mark Lombardi: tracked and mapped global financial fiascos in the 1980s and 1990s from public sources such as news articles





Understanding through visualization

- "I happened to be in the Drawing Center when the Lombardi show was being installed and several consultants to the Department of Homeland Security came in to take a look. They said they found the work revelatory, not because the financial and political connections he mapped were new to them, but because Lombardi showed them an elegant way to array disparate information and make sense of things, which they thought might be useful to their security efforts. I didn't know whether to find that response comforting or alarming, but I saw exactly what they meant."

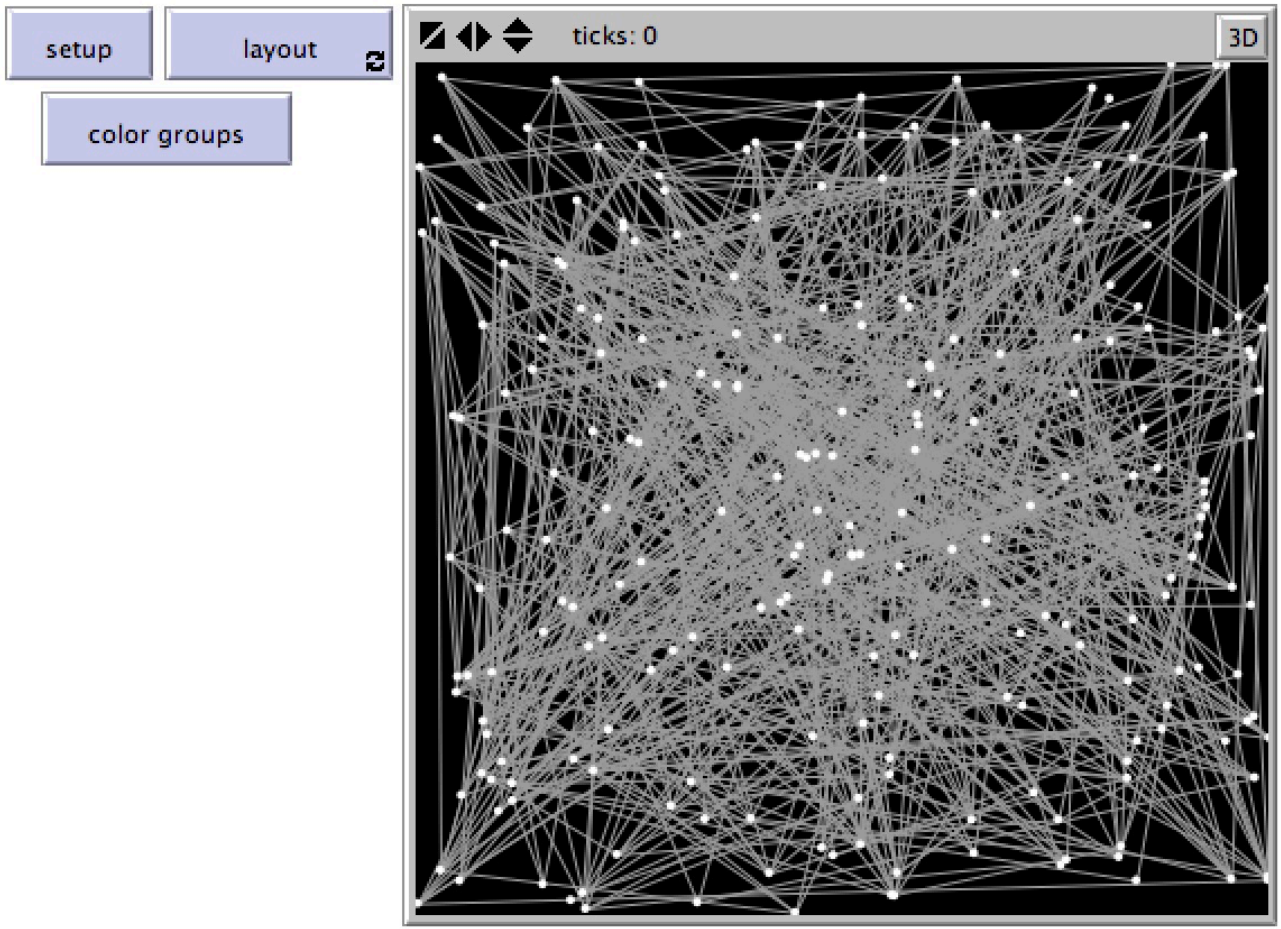
  Michael Kimmelman

  Webs Connecting the Power Brokers, the Money and the World

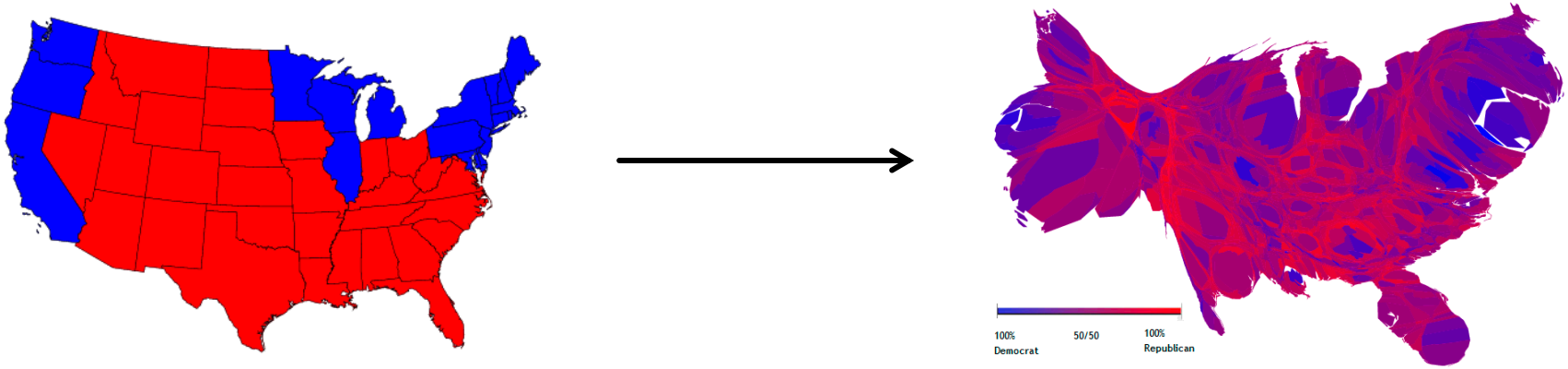
  NY Times November 14, 2003

When does SNA render things understandable?

- Is this just a random network?

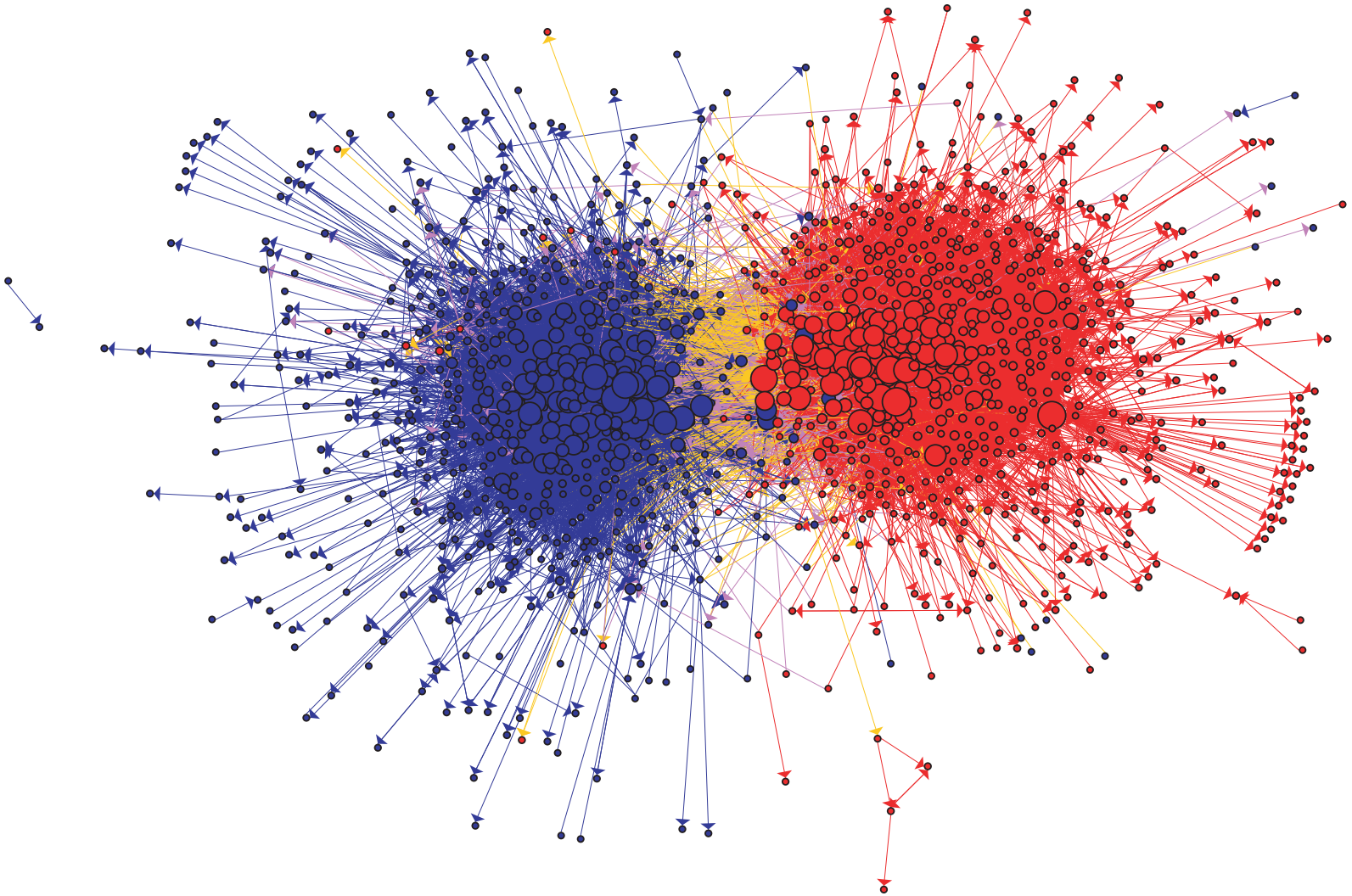


Mapping geography

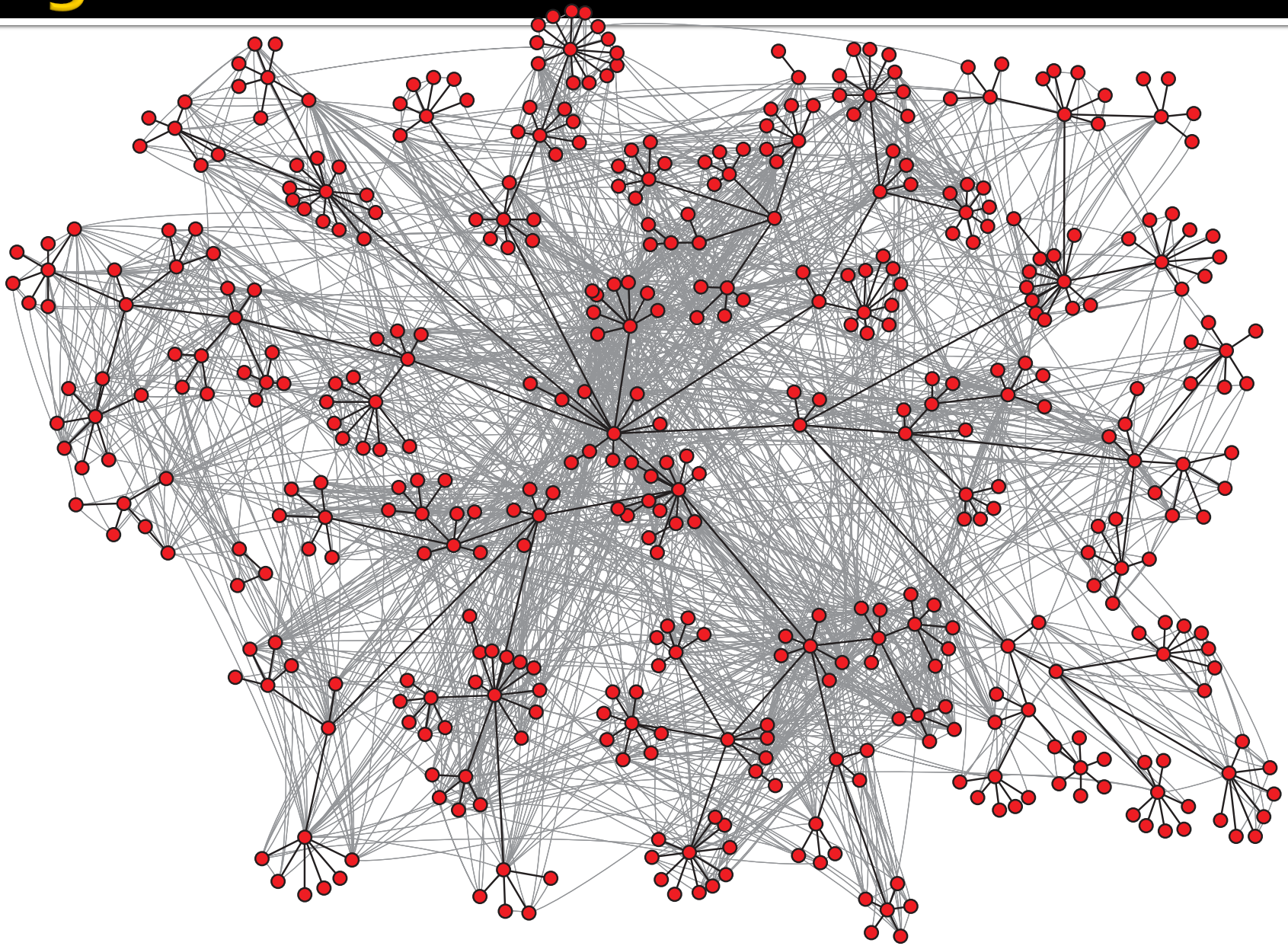


Michael Gastner, Mark Newman, PNAS 2005

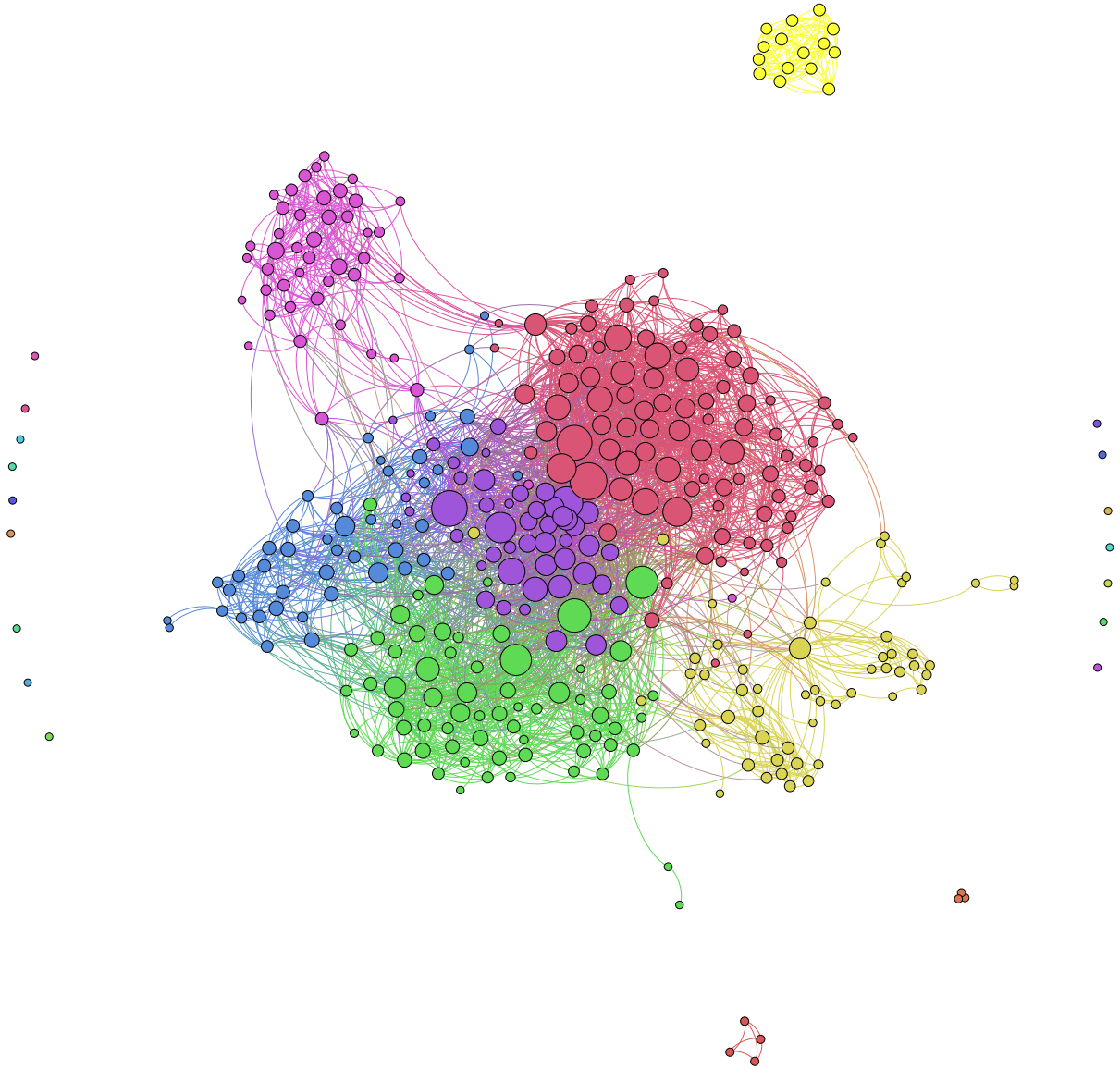
Mapping an online space



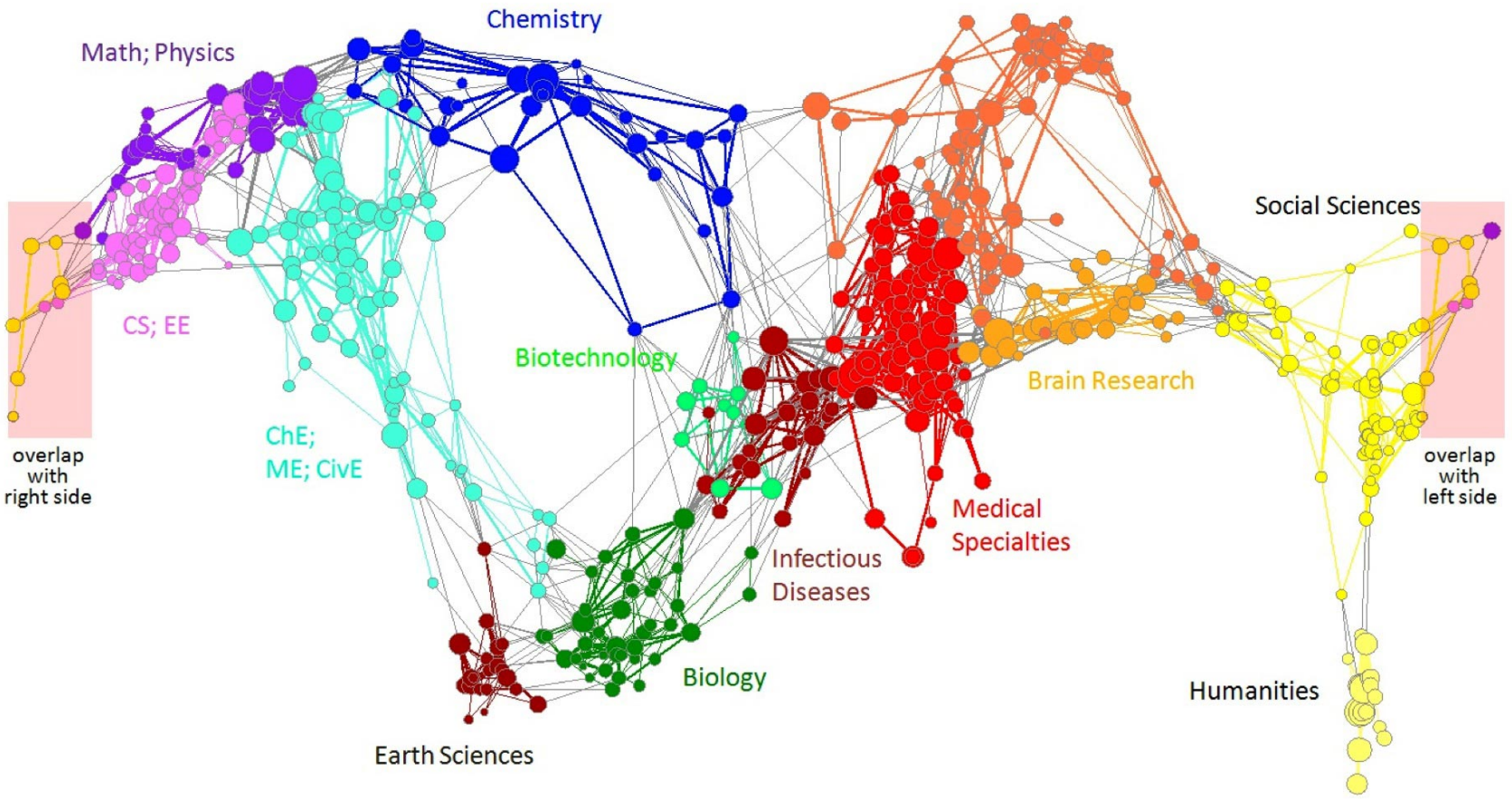
Organizations



Ego networks



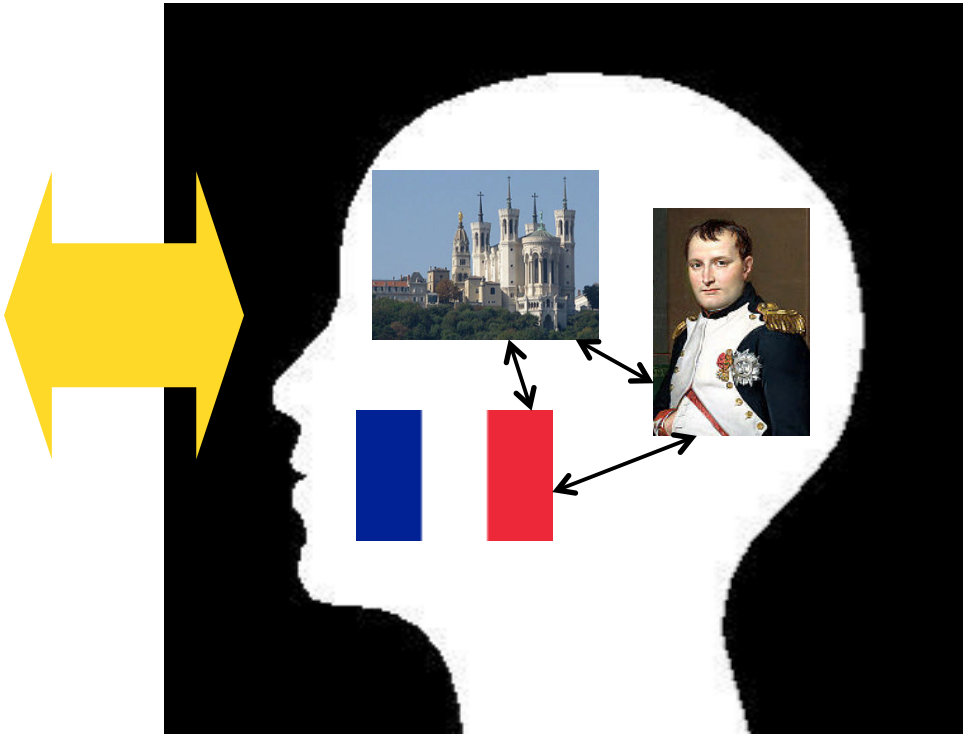
Networks: Information



Citation networks and Maps of science

[Börner et al., 2012]

Networks: Knowledge

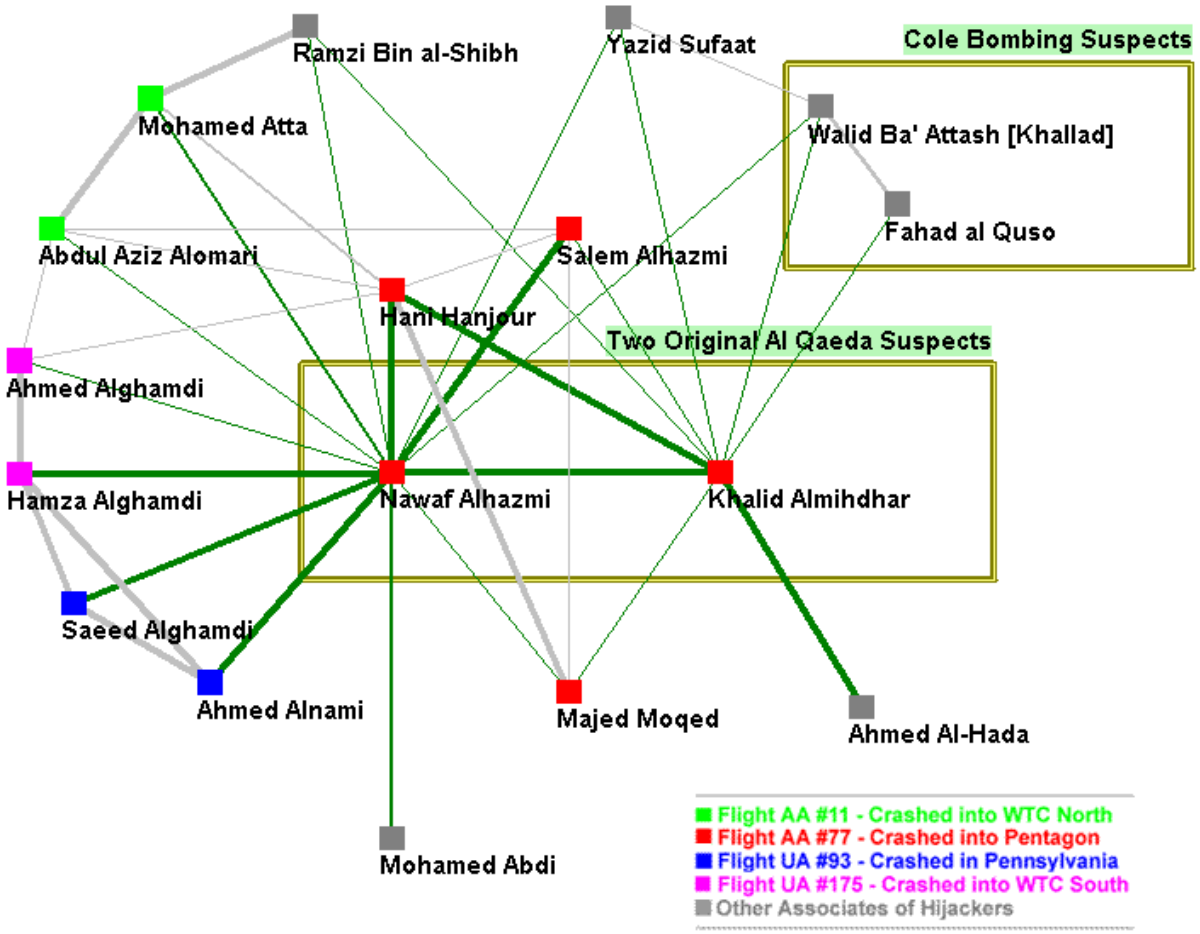


Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]

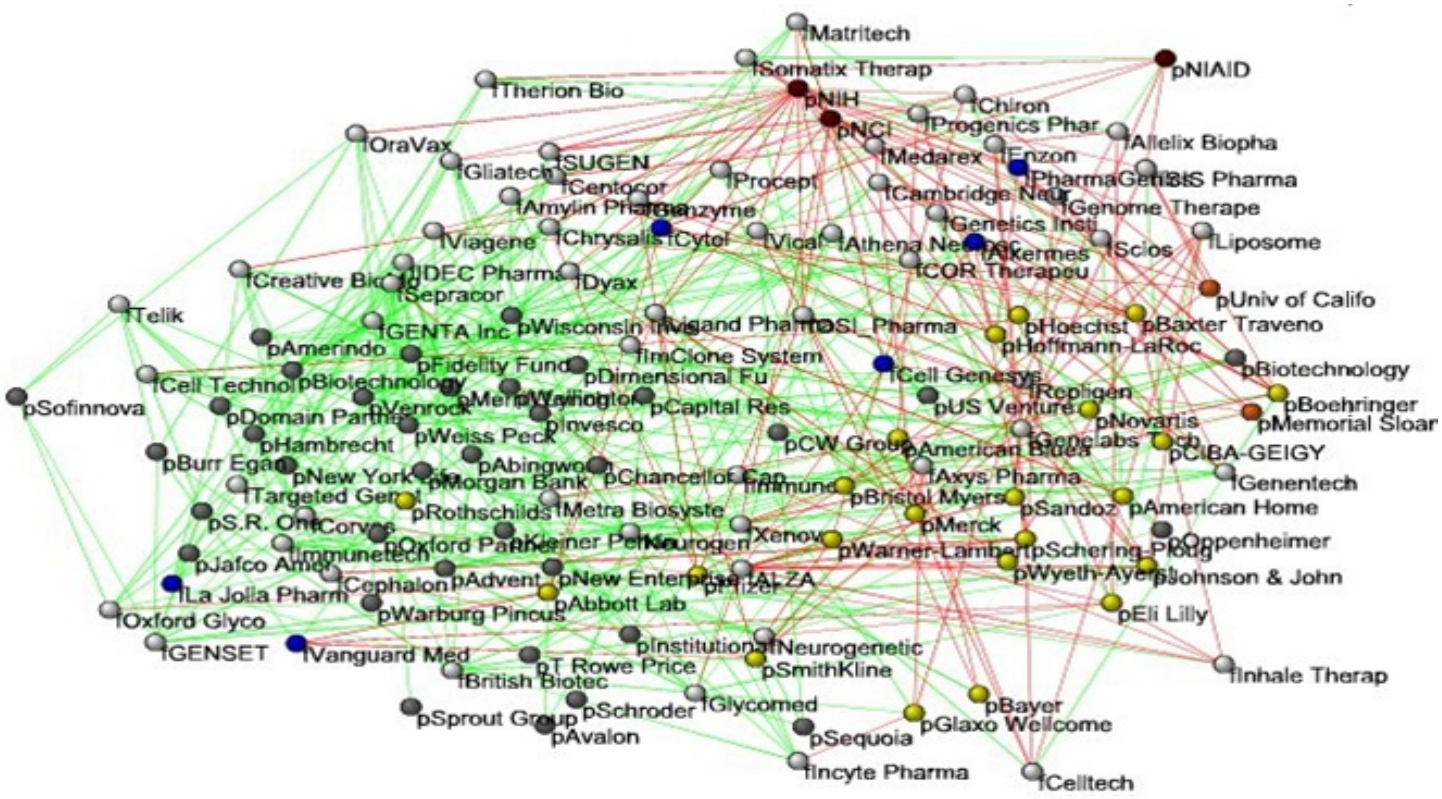
Networks: Organizations



9/11 terrorist network

[Krebs, 2002]

Networks: Economy



Nodes:

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

Links:

- Collaborations
- Financial
- R&D

Bio-tech companies

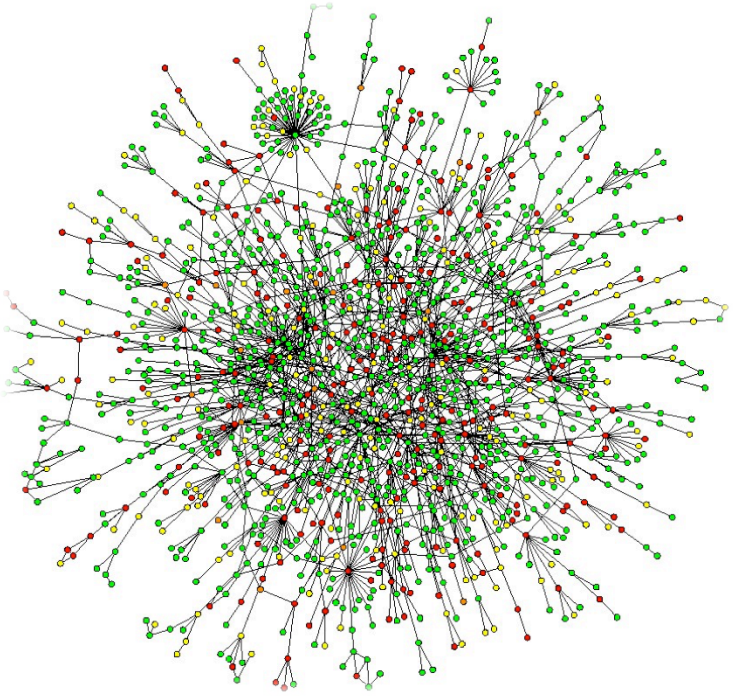
[Powell-White-Koput, 2002]

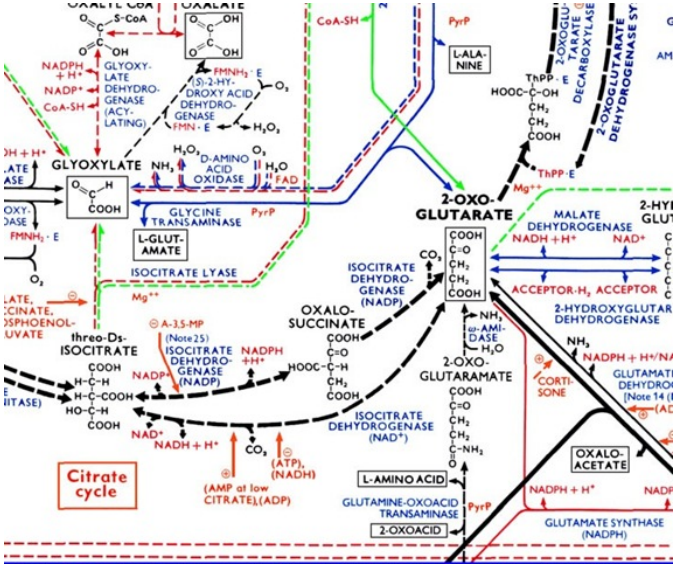
Networks: Brain

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

Networks!!

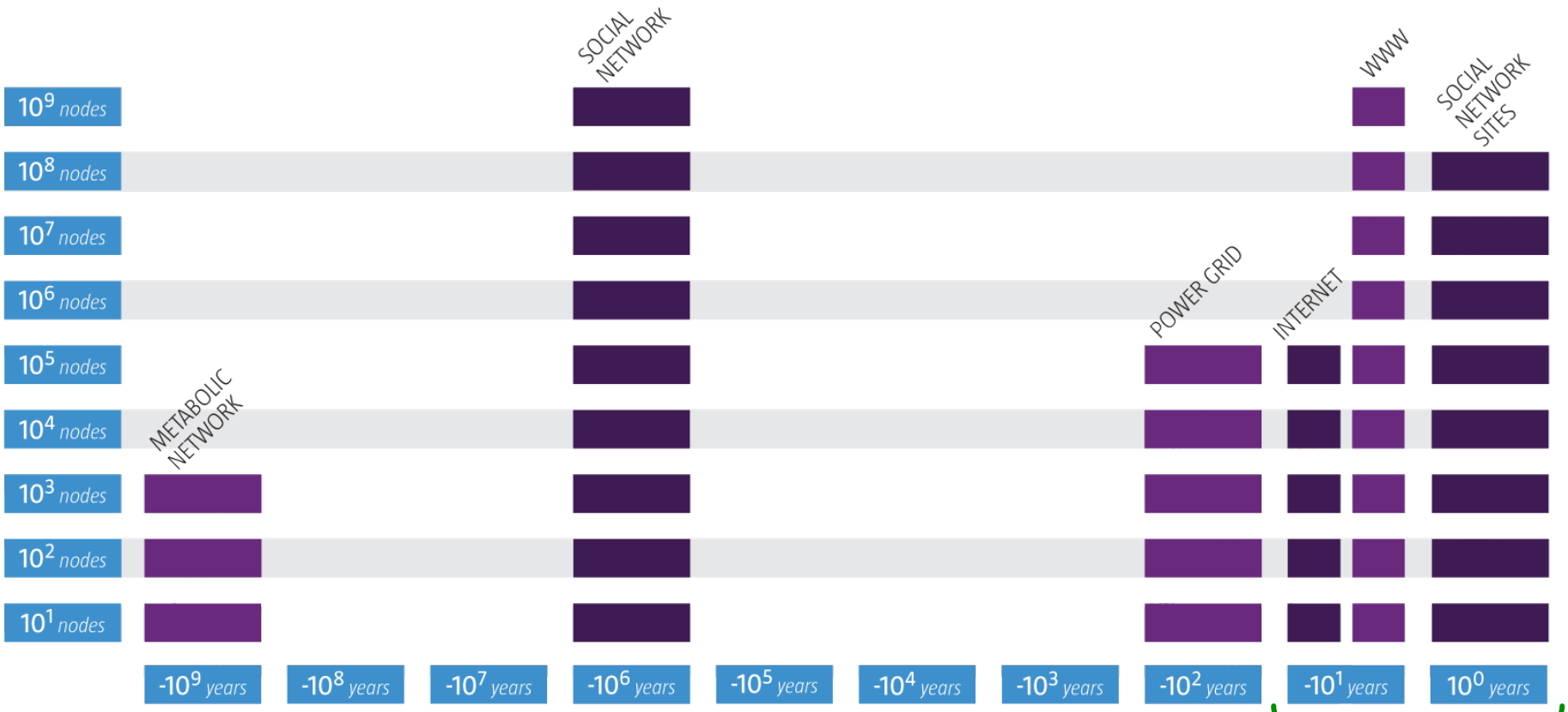
Behind many systems there is an intricate wiring diagram, **a network**, that defines the **interactions** between the components

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

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

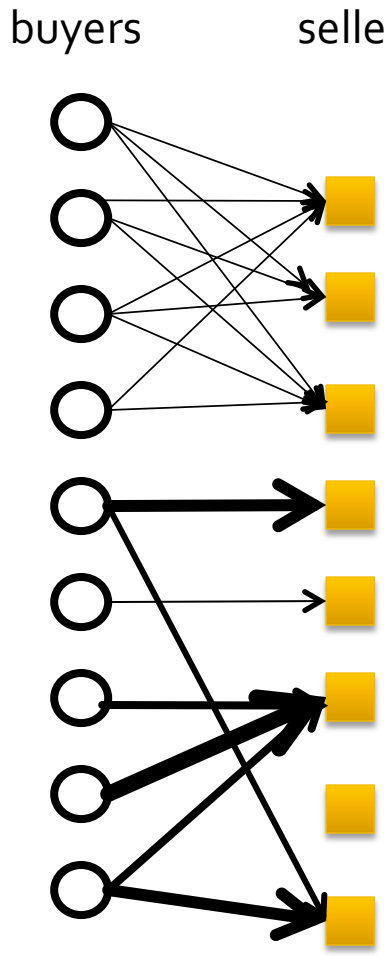
Networks: Size Matters

- **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]

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, Information virality
- **Information networks:**
  - Navigational aids

Detecting rating fraud



(thickness) purchase price

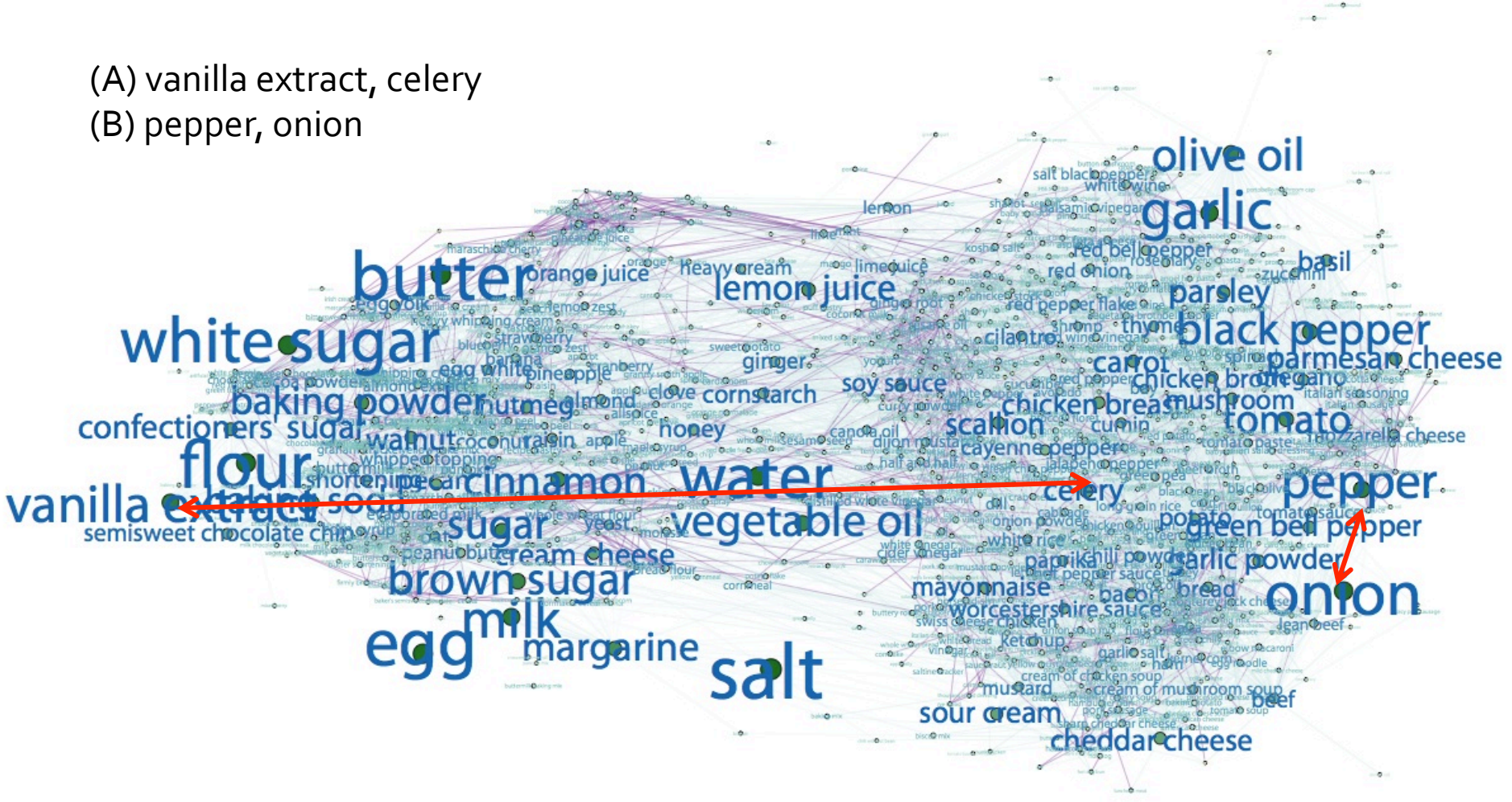
which ratings are likely to be fraudulent?

Where is a global pandemic most likely to hit?

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040961

Which link is most likely missing?

(A) vanilla extract, celery
(B) pepper, onion



Networks Really Matter

- If you want to understand the spread of diseases, **you need to figure out who will be in contact with whom**
- If you want to understand the structure of the Web, **you have to analyze the 'links'.**
- If you want to understand dissemination of news or evolution of science, **you have to follow the flow.**

About CS224W

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

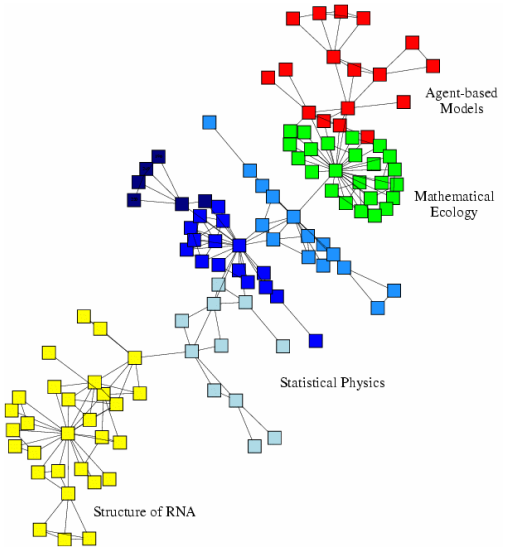
Reasoning about Networks

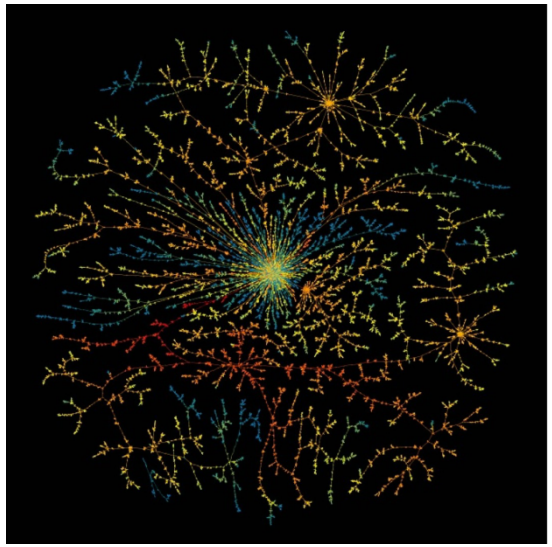
- **How do we reason about networks?**
  - **Empirical:** Study network data to find organizational principles
    - How do we measure and quantify networks?
  - **Mathematical models:** Graph theory and statistical models
    - Models allow us to understand behaviors and distinguish surprising from 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 come 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

Properties	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
Information virality, Memetracking	Independent cascade model, Game theoretic model	Influence maximization, Outbreak detection

Logistics: Course Assistants

Paris Syminelakis
(head TA)

Caroline Suen

Nihit Desai

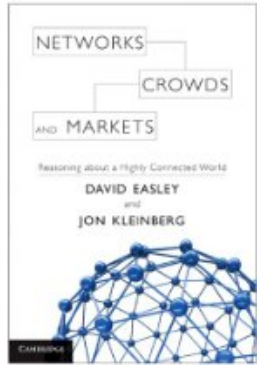
Rohit Mundra

Sameep Bagadia

Tim Althoff

Logistics: Website

- http://cs224w.stanford.edu
  - Slides posted just before the class
- **Readings:**
  - 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/fall2014/cs224w
    - Use access code "snap"
  - **Please participate and help each other! (2 % of grade)**
- **For e-mailing course staff, always use:**
  - cs224w-aut1516-staff@lists.stanford.edu
- We will post course announcements to Piazza (make sure you check it regularly)

Homework, Write-ups

- **Assignments are long and take time (10-20h) Start early!**
  - A combination of: Data analysis, Algorithm design, and Math
- **How to submit?**
  - **Upload via GradeScope (http://gradescope.com)**
    - **To register fill this form http://bit.ly/1DuETte**
    - **IMPORTANT: one answer per page!**
  - **Code and write-ups** (proposal, milestone, final report) have to **also** be uploaded at http://snap.stanford.edu/submit/
- **2 late periods for the quarter:**
  - 1 late period expires at the start of next class
  - You can use at most 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 up to 3 students**
  - **(all projects will be graded equally, regardless of group size)**
- 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
- **Read:** http://web.stanford.edu/class/cs224w/info.html#proj

Course Schedule

Week	Assignment	Due on THU
2	**Homework 0**	October 1
3	**Homework 1**	October 8
4	**Project proposal**	October 15
5	**Homework 2**	October 22
6	**Work on the project**	
7	**Homework 3**	November 5
8	**Project milestone**	November 12 (no late periods!)
9	**Homework 4**	November 19
	Thanksgiving break	
10	**Project report**	December 8, midnight (no late periods!)
	Poster session	December 9 8:30-11:30am

Work for the Course & Grading

- **Final grade will be composed of:**
  - **Homework: 48%**
    - Homework 1,2,3,4: 12% each
  - **Substantial class project: 50%**
    - Proposal: 20%
    - Project milestone: 20%
    - Final report: 50%
    - Poster presentation: 10%
  - **Piazza participation, snap code contribution: 2%**
    - Students between grades get extra credit for Piazza participation

Prerequisites

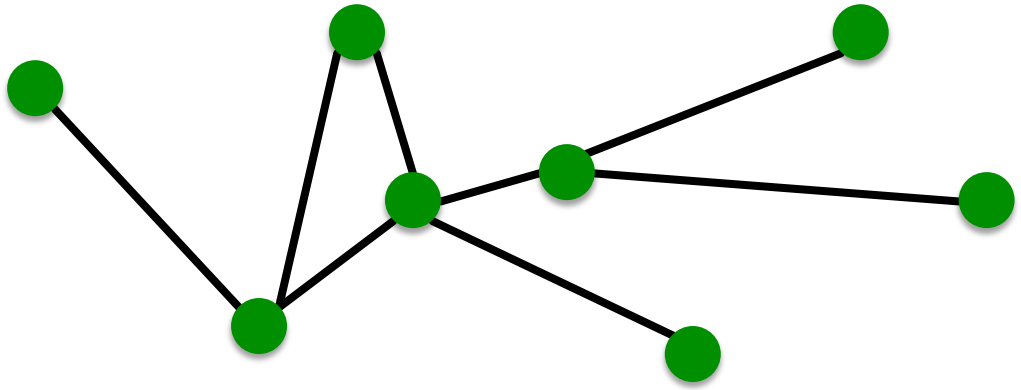
- **No single topic in the course is too hard by itself**
- **But we will cover and touch upon many topics and this is what makes the course hard**
  - **Good background in:**
    - Algorithms and graph theory
    - Probability and Statistics
    - Linear algebra
  - **Programming:**
    - You should be able to write non-trivial programs (in Python)
  - **2 recitation sessions (all in Nvidia auditorium):**
    - SNAP.PY: Friday, 9/25 (4:00-5:30)
    - Review of Probability and Linear Algebra: Friday, 10/2 (4:00-5:30)

Network Analysis Tools

- **We highly recommend SNAP:**
  - **SNAP C++:** more challenging but more scalable
  - **SNAP.PY:** Python ease of use, most of C++ scalability
    - HW0 asks you to do some very basic network analysis with snap.py
      - If you find HW0 difficult, this class is probably not for you
  - Other tools include NetworkX, JUNG, iGraph

Starter Topic:
Basic network properties

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

In most cases we will use the two terms interchangeably

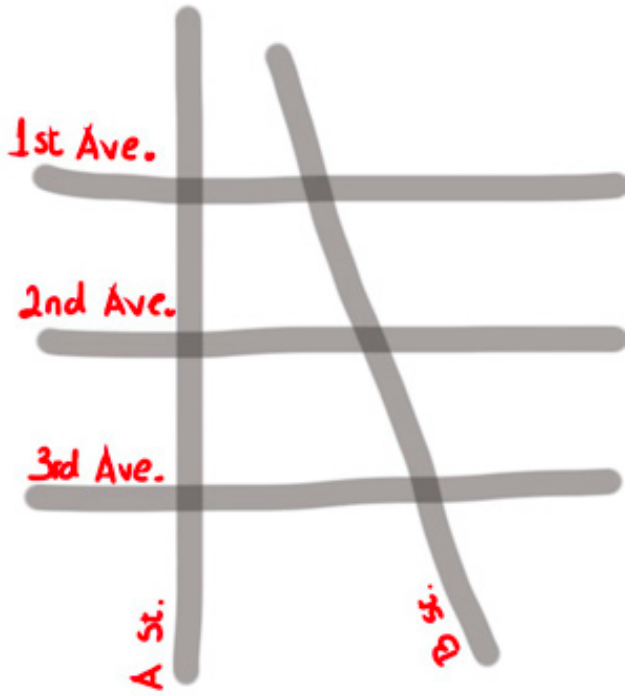
Network elements: edges

- Directed (also called arcs, links)
  - A -> B
    - A likes B, A gave a gift to B, A is B's child
- Undirected
  - A <-> B or A – B
    - A and B like each other
    - A and B are siblings
    - A and B are co-authors

How do you define a network?

The image shows 5 streets (A and B streets, and 1st, 2nd, and 3rd Avenue). How can a network be constructed from these streets?

- 1) Roads (A St., B St., 1st Ave, …) are nodes and an edge is drawn between every pair of roads that intersect.
- 2) Intersections are nodes (e.g. A St. and 1st Ave, B St. and 2nd Ave), and an edge is drawn between any two intersections that are directly connected by a segment of street with no intervening intersections.
- 3) Street blocks are nodes (e.g. the block between A and B, and 2nd and 3rd), and blocks that are adjacent (i.e. across the street from each other) have edges.

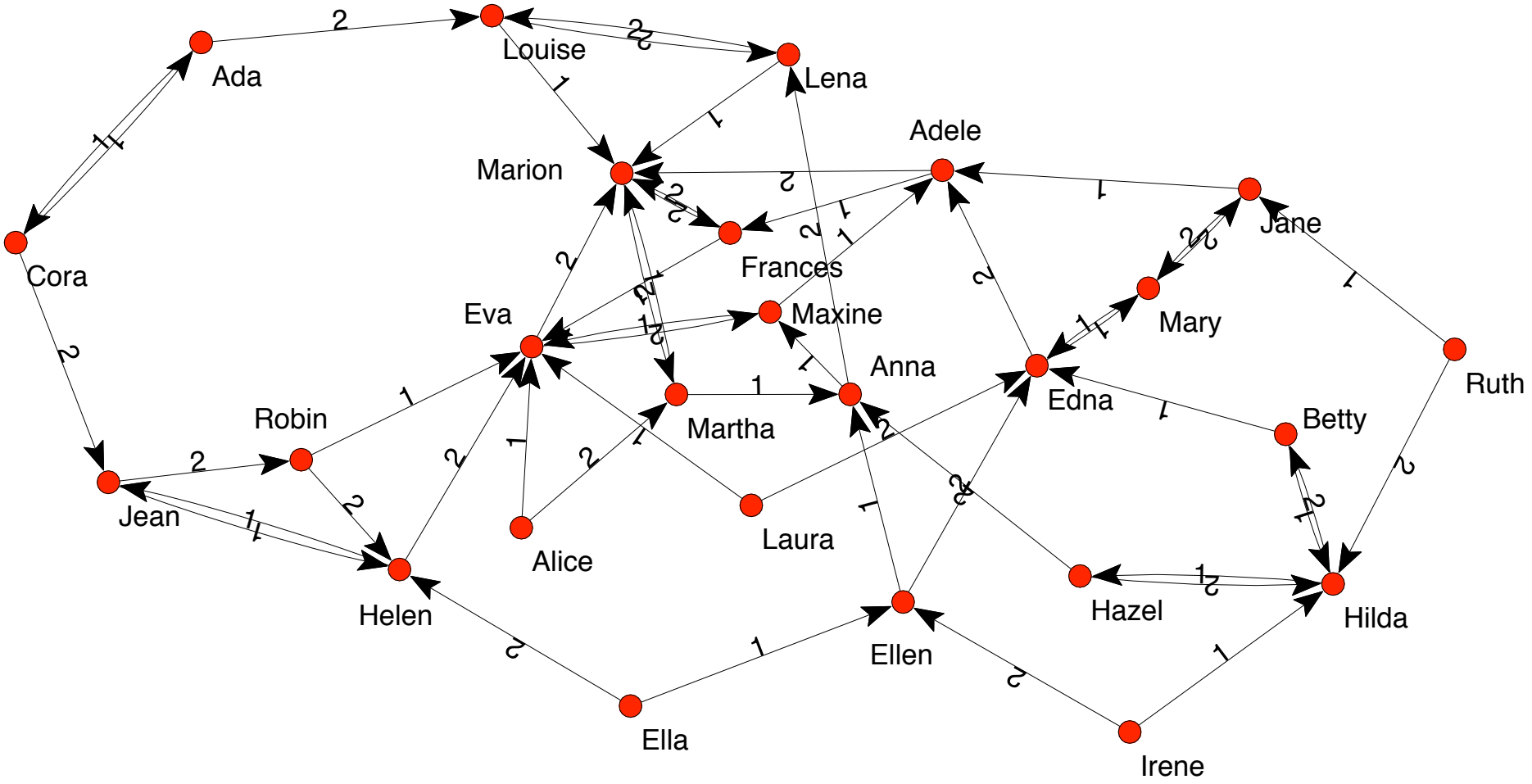


Edge attributes

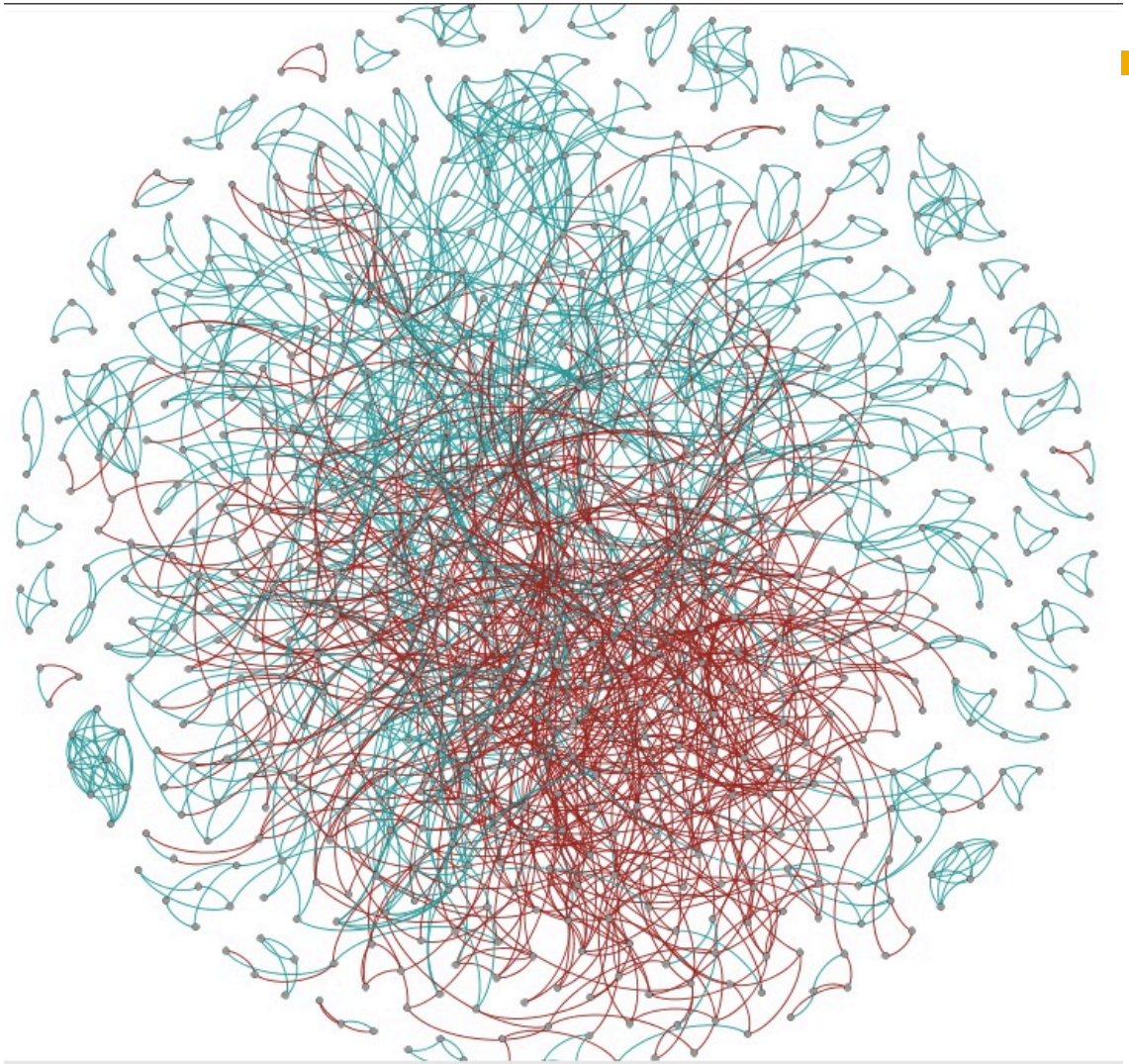
- Examples
  - weight (e.g. frequency of communication)
  - ranking (best friend, second best friend…)
  - type (friend, relative, co-worker)
  - properties depending on the structure of the rest of the graph: e.g. betweenness

Directed networks

- girls' school dormitory dining-table partners, 1st and 2nd choices (Moreno, *The sociometry reader*, 1960)



Positive and negative weights



- e.g. one person trusting/distrusting another
  - Research challenge: How does one 'propagate' negative feelings in a social network? Is my enemy's enemy my friend?

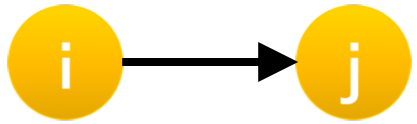
sample of positive & negative ratings from Epinions network

Data representation

- adjacency matrix
- edgelist
- adjacency list

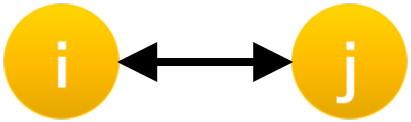
Adjacency matrices

- Representing edges (who is adjacent to whom) as a matrix
  - A_{ij} = 1 if node *i* has an edge to node *j*
    = 0 if node *i* does not have an edge to j

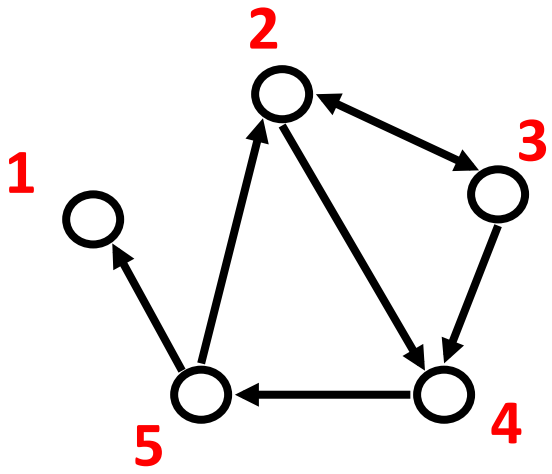


  - A_{ii} = 0 unless the network has self-loops

  - $A_{ij} = A_{ji}$ if the network is undirected, or if *i* and *j* share a reciprocated edge



Example adjacency matrix



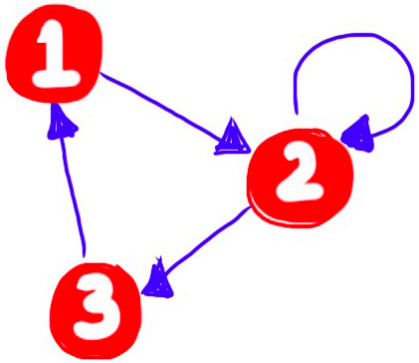
$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Which adjacency matrix represents the network?

A)
$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

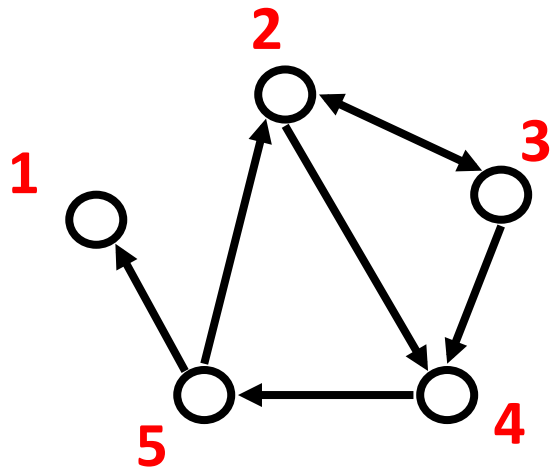
B)
$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

C)
$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$



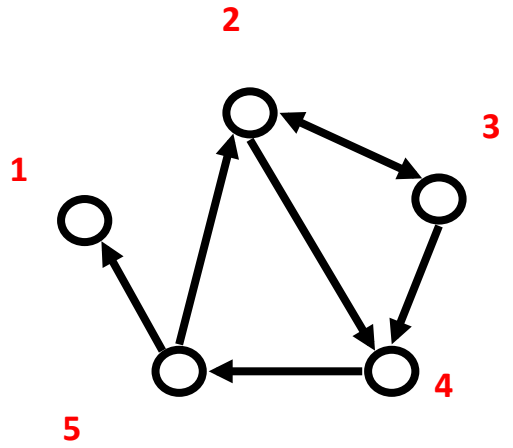
Edge list

- Edge list
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1



Adjacency lists

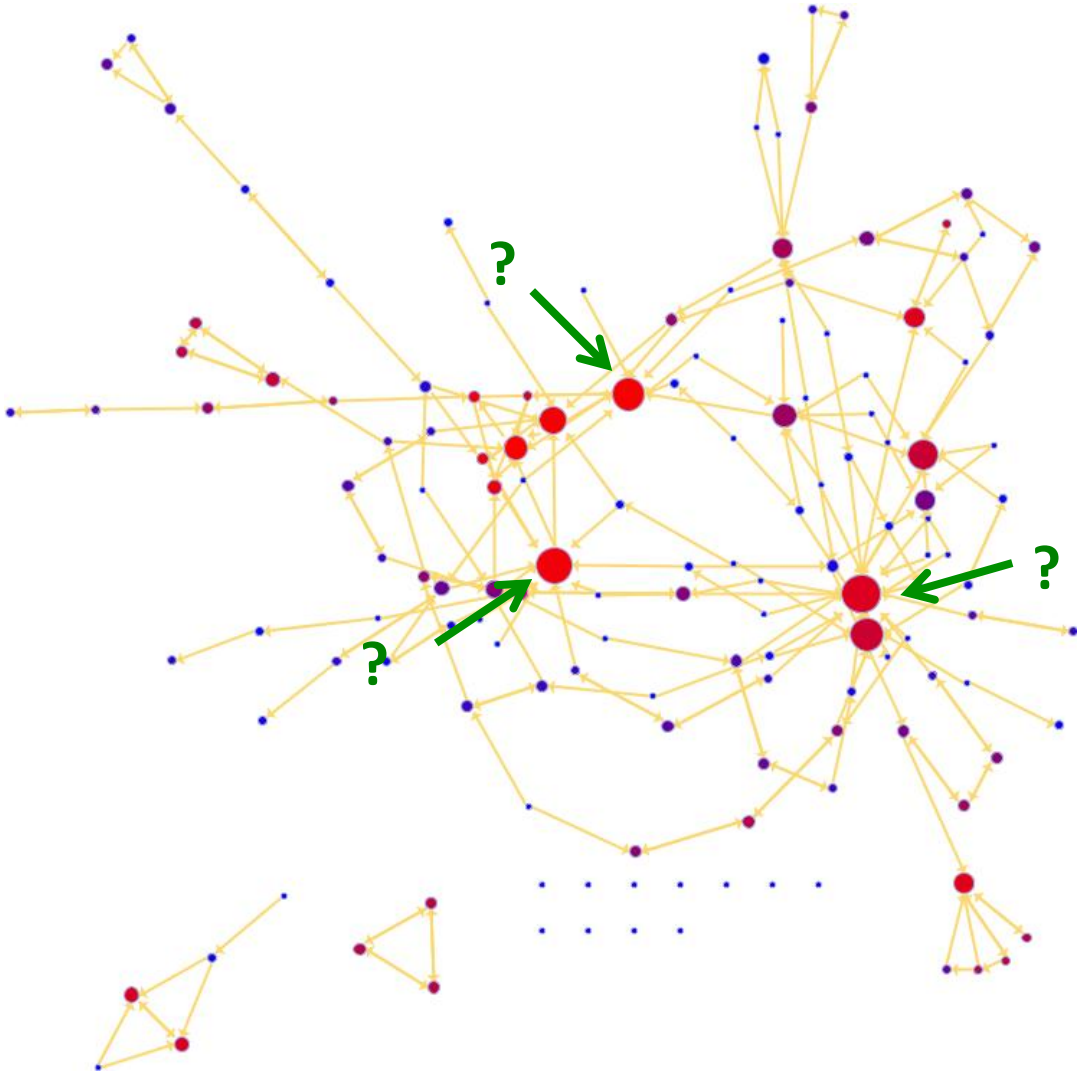
- Adjacency list
  - is easier to work with if network is
    - large
    - sparse
  - quickly retrieve all neighbors for a node
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2



Computing metrics

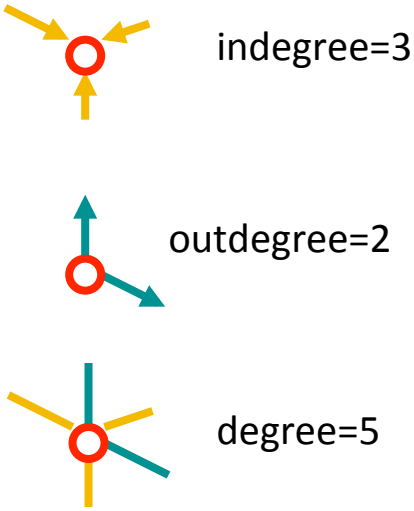
- degree & degree distribution
- connected components

Degree: which node is most connected?

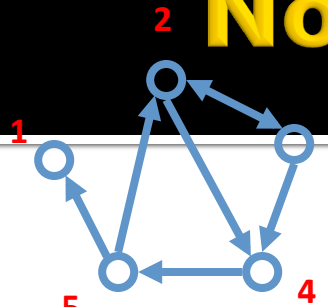


Nodes

- Node network properties
  - from immediate connections
    - indegree
      how many directed edges (arcs) are incident on a node
    - outdegree
      how many directed edges (arcs) originate at a node
    - degree (in or out)
      number of edges incident on a node
  - from the entire graph (next lecture)
    - centrality (betweenness, closeness)



Node degree from matrix values



- Outdegree(i) = $\sum_{j=1}^{n} A_{ij}$

A =

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

example: outdegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd *row*

$$\sum_{j=1}^{n} A_{3j}$$

- Indegree(j) = $\sum_{i=1}^{n} A_{ij}$

A =

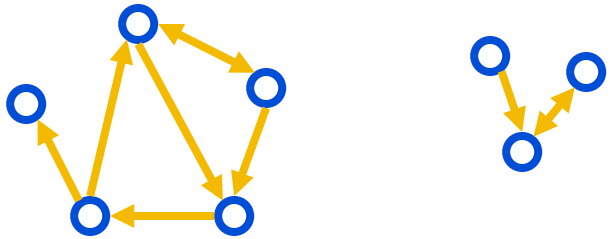
$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd *column*

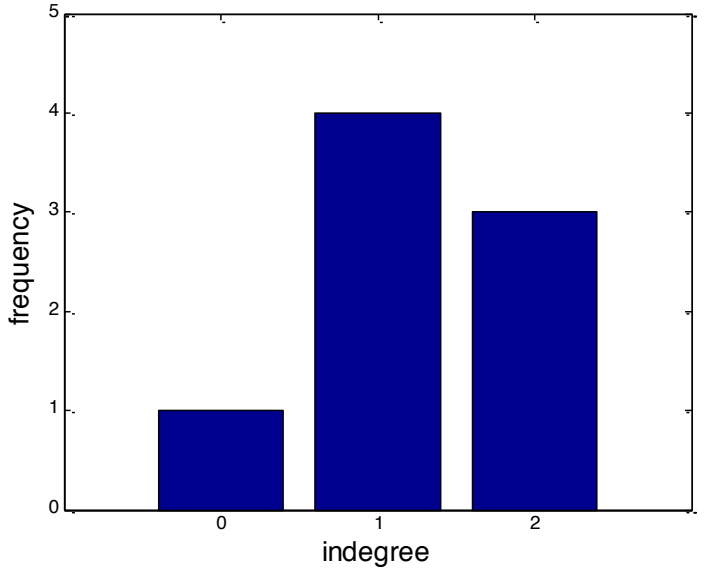
$$\sum_{i=1}^{n} A_{i3}$$

Network metrics: degree sequence and degree distribution

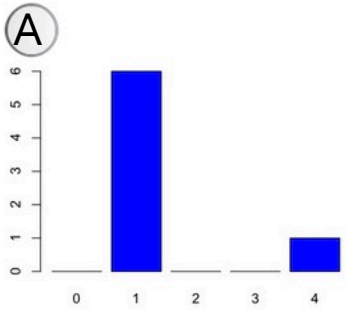
- Degree sequence: An ordered list of the (in,out) degree of each node
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]

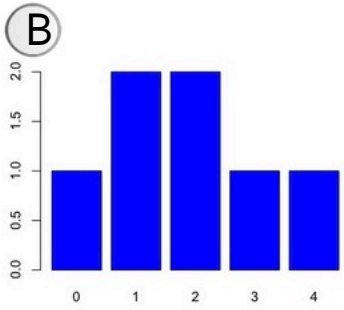


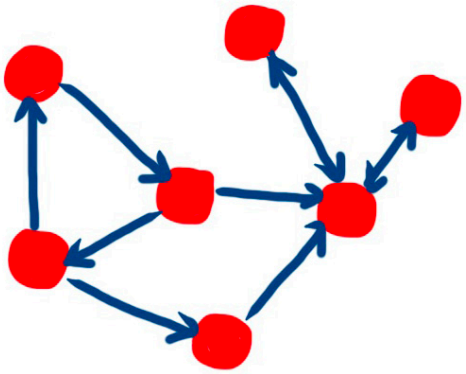
- Degree distribution: A frequency count of the occurrence of each degree
  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]

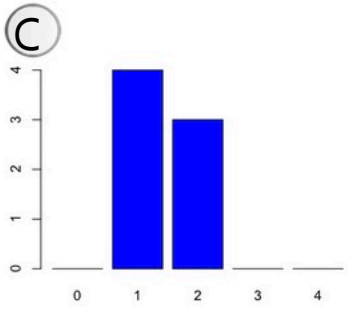


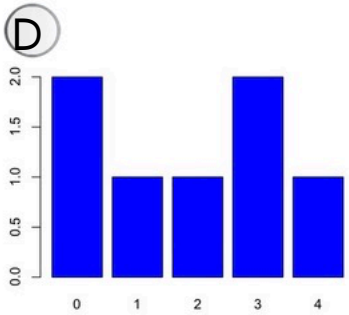
What is the degree distribution of this network



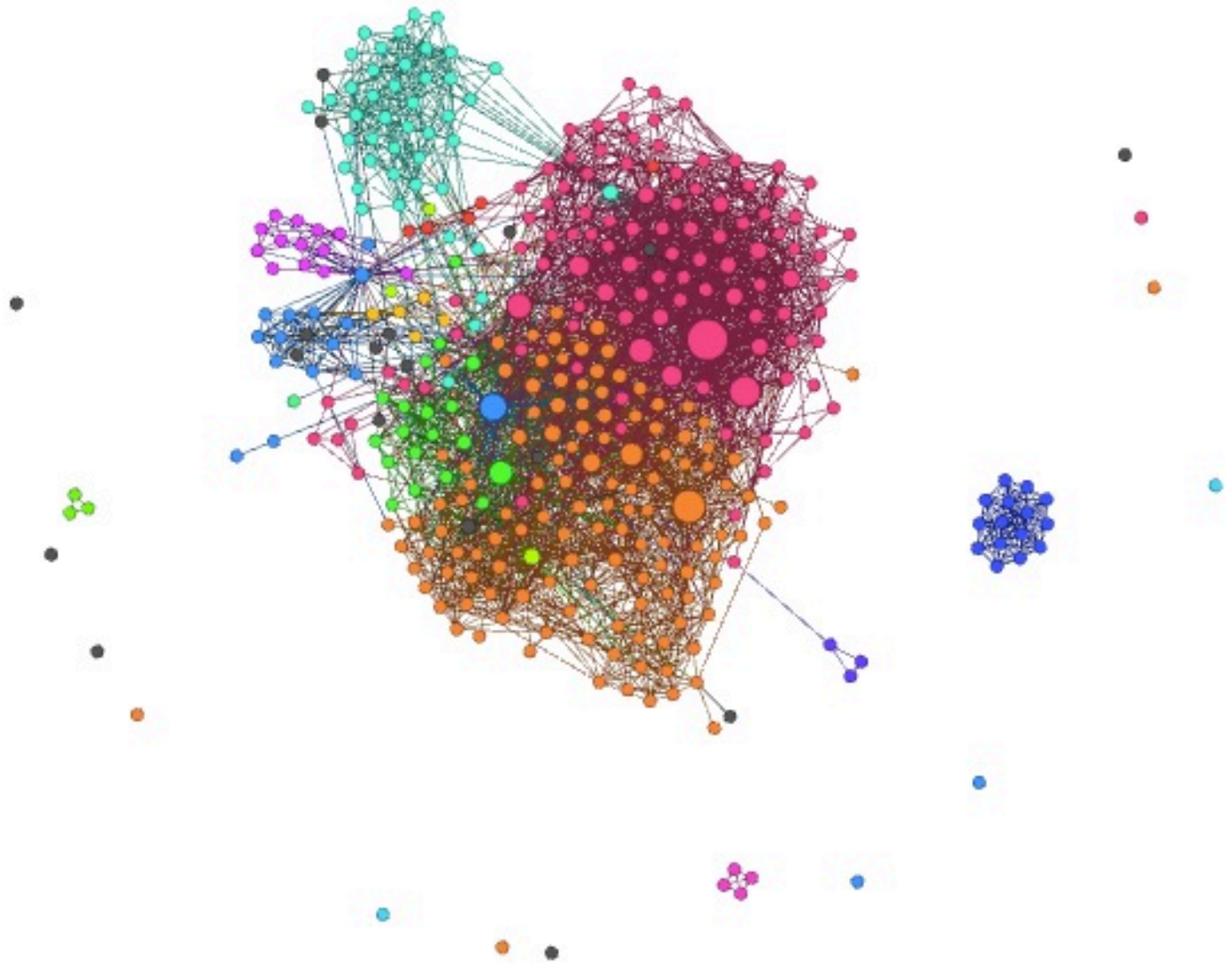






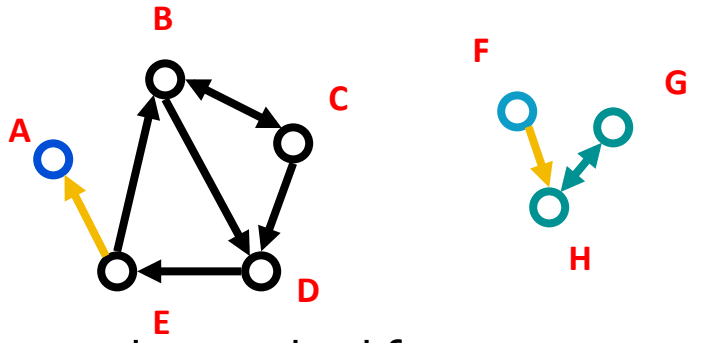


Is everything connected?

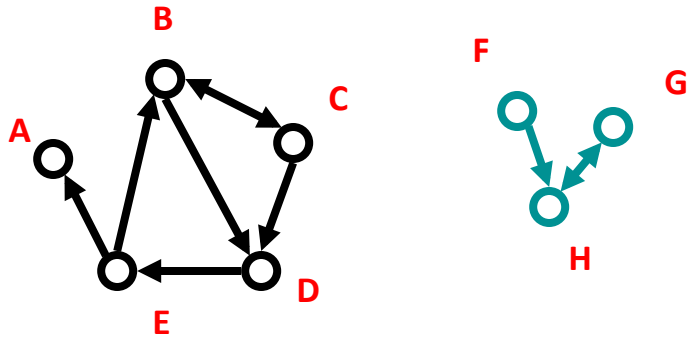


Connected components

- Strongly connected components
  - Each node within the component can be reached from every other node in the component by following directed links
    - Strongly connected components
      - B C D E
      - A
      - G H
      - F

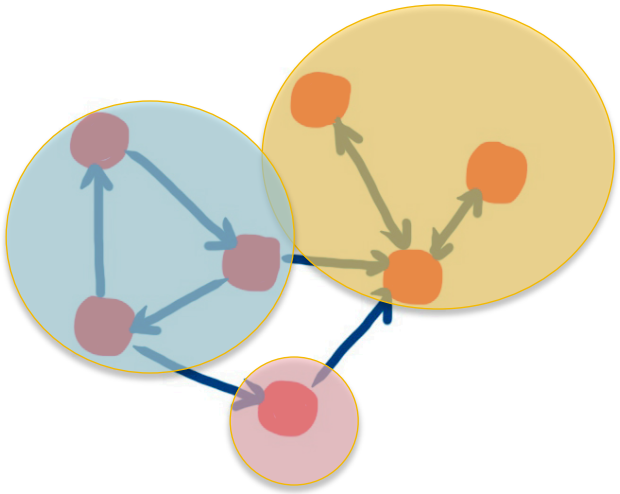


- Weakly connected components: every node can be reached from every other node by following links in either direction
  - Weakly connected components
    - A B C D E
    - G H F



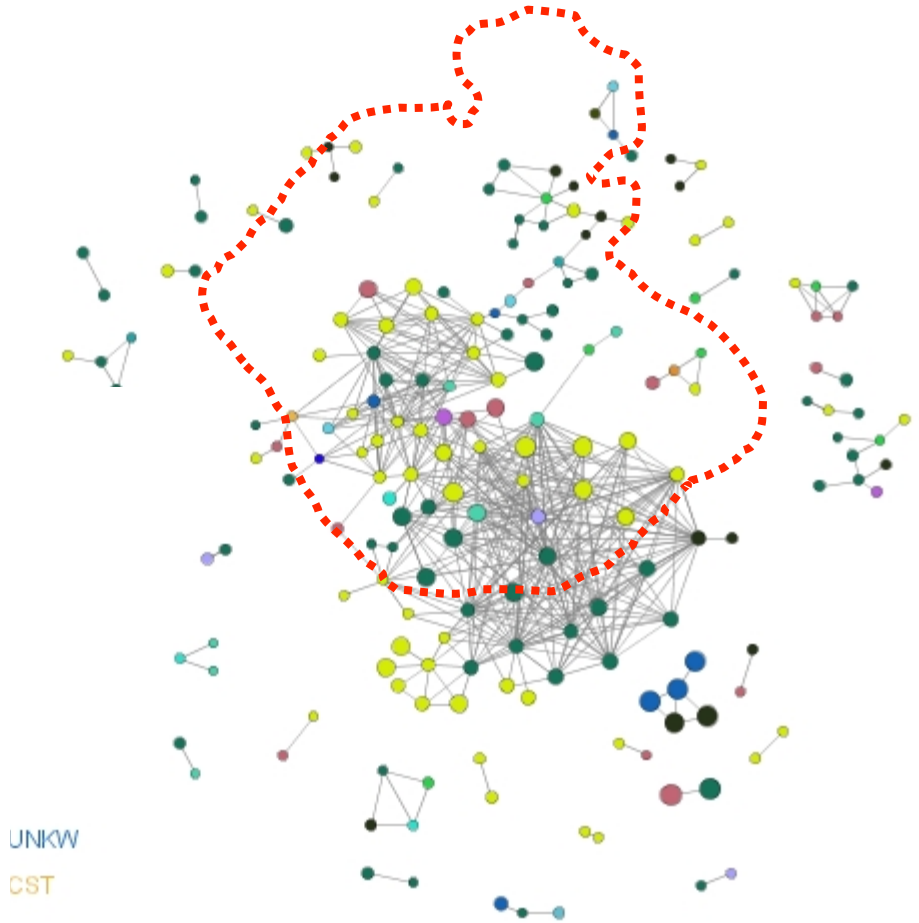
- In undirected networks one talks simply about ‘connected components’

How many strongly connected components are in this network?



Giant component

- if the largest component encompasses a significant fraction of the graph, it is called the **giant component**



Recap

- Complex systems can be analyzed and understood through a networks analysis lens
- Basic network properties:
  - degree and degree distribution
  - connected components
    - strong
    - weak
    - giant