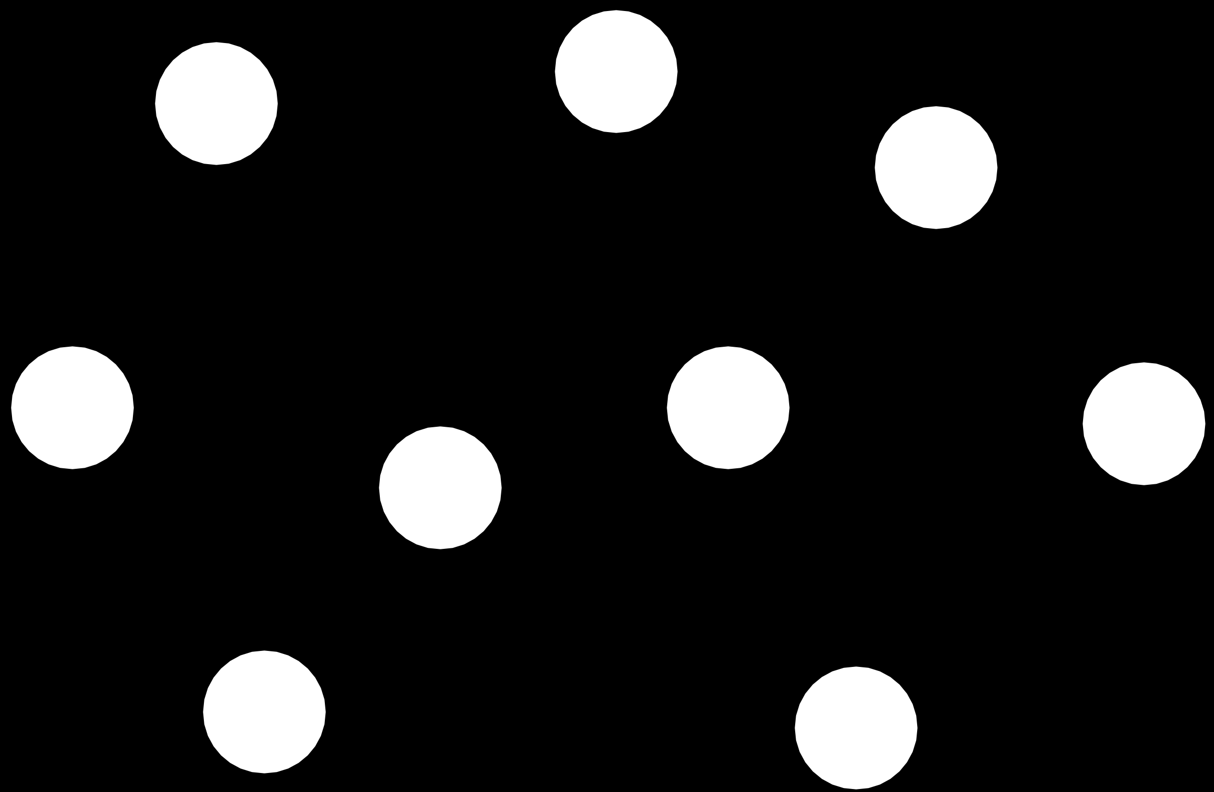
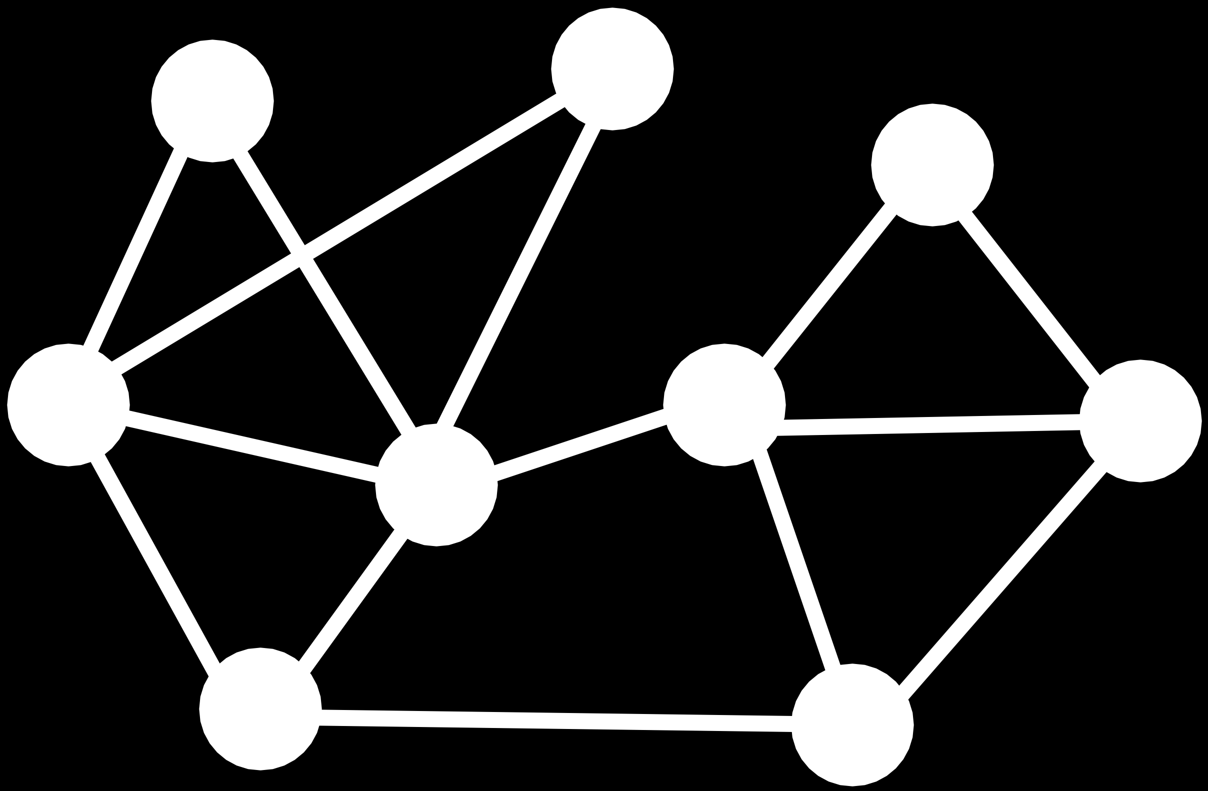
CS224W: Analysis of Networks

CS224W: Analysis of Network
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

Why Networks?

Networks are a general language for describing complex systems



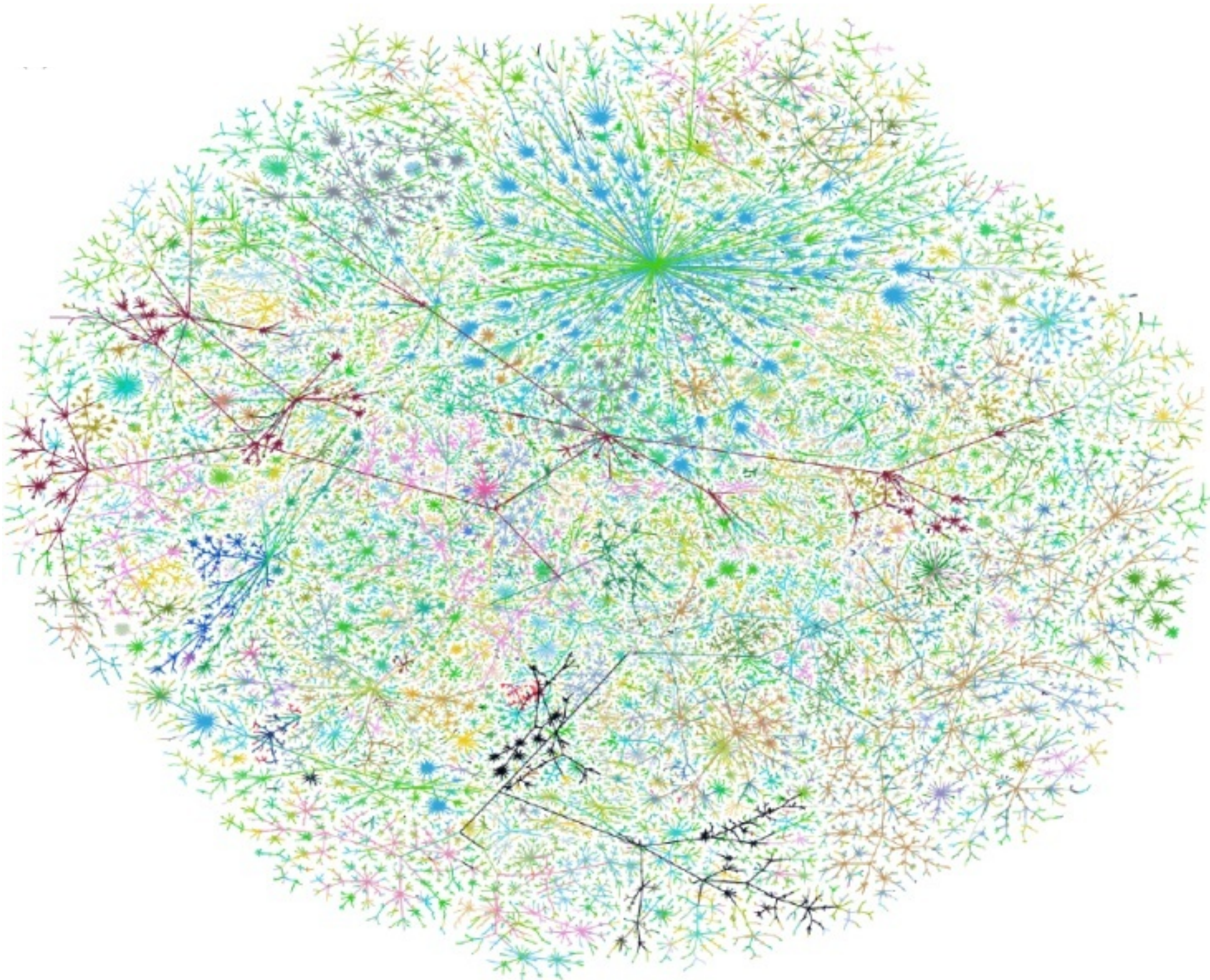


Interactions!

Networks & Complex Systems

- **Complex systems are all 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/proteins** 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!

Networks!!

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

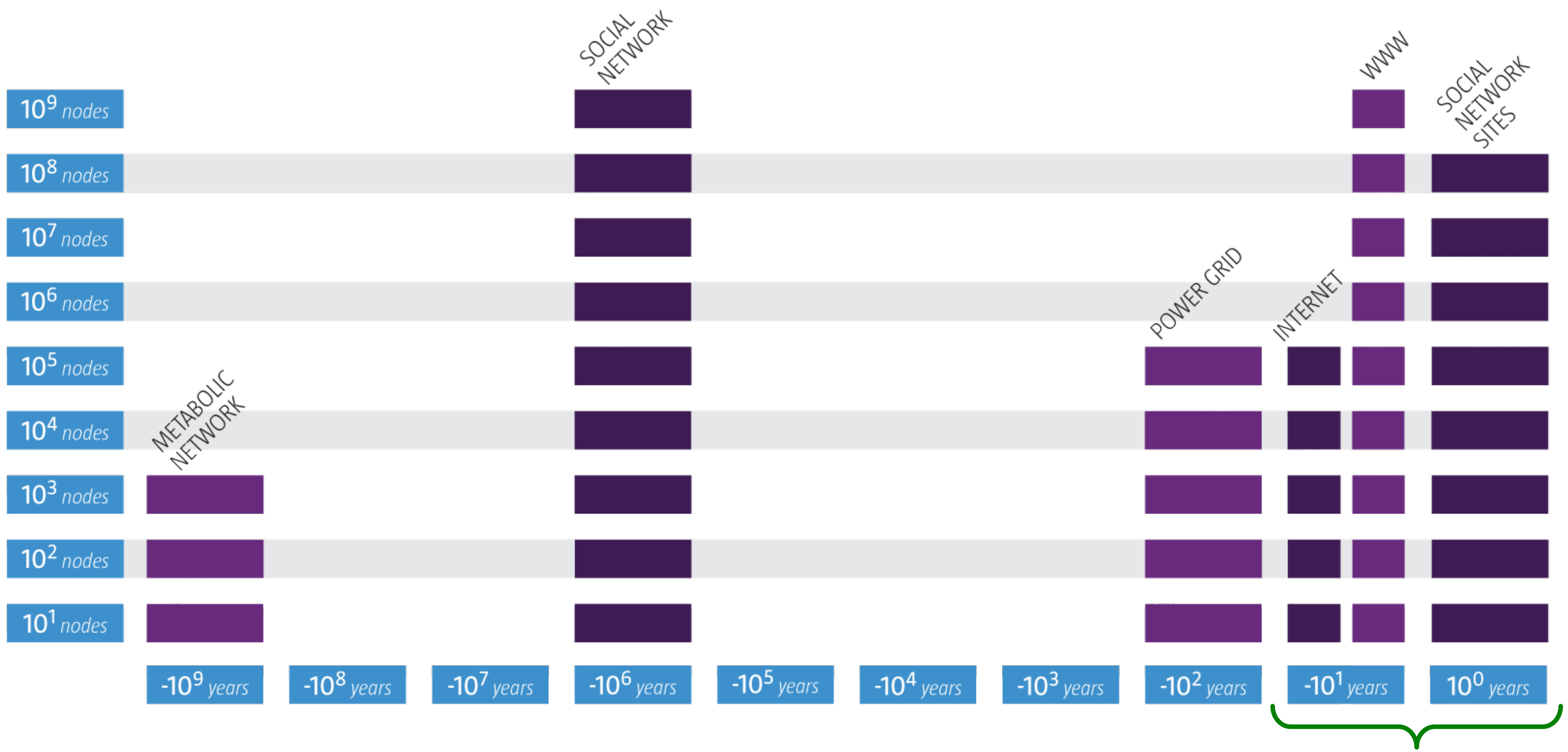
We will never understand these systems unless we understand the networks behind them!

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

Web – The Lab for Humanity

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

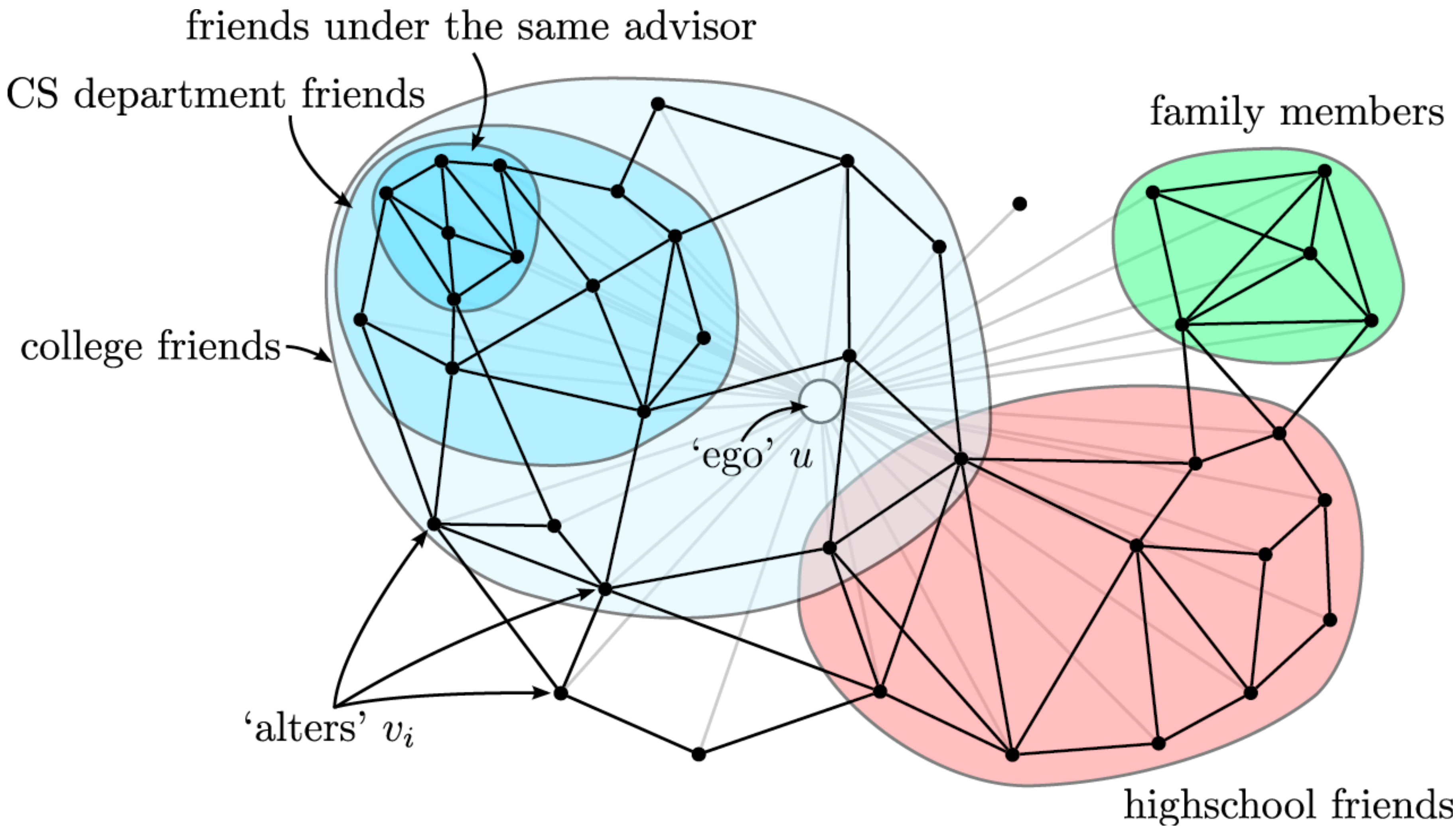
Networks and Applications

(1) Networks: Social

Facebook social graph

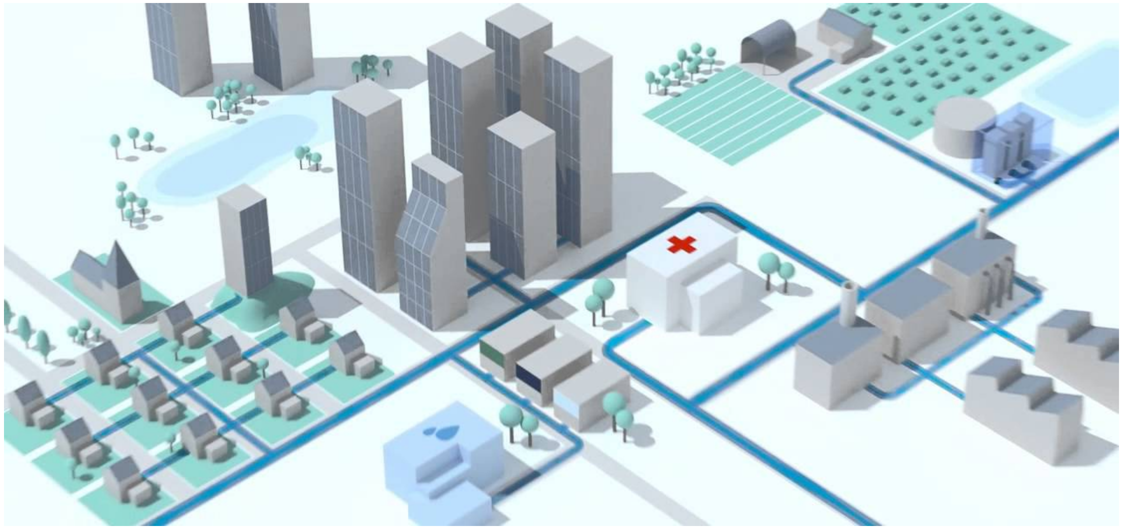
4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Application: Social Circle Detection



Discover circles and why they exist

(2) Networks: Infrastructure



Water supply distribution networks

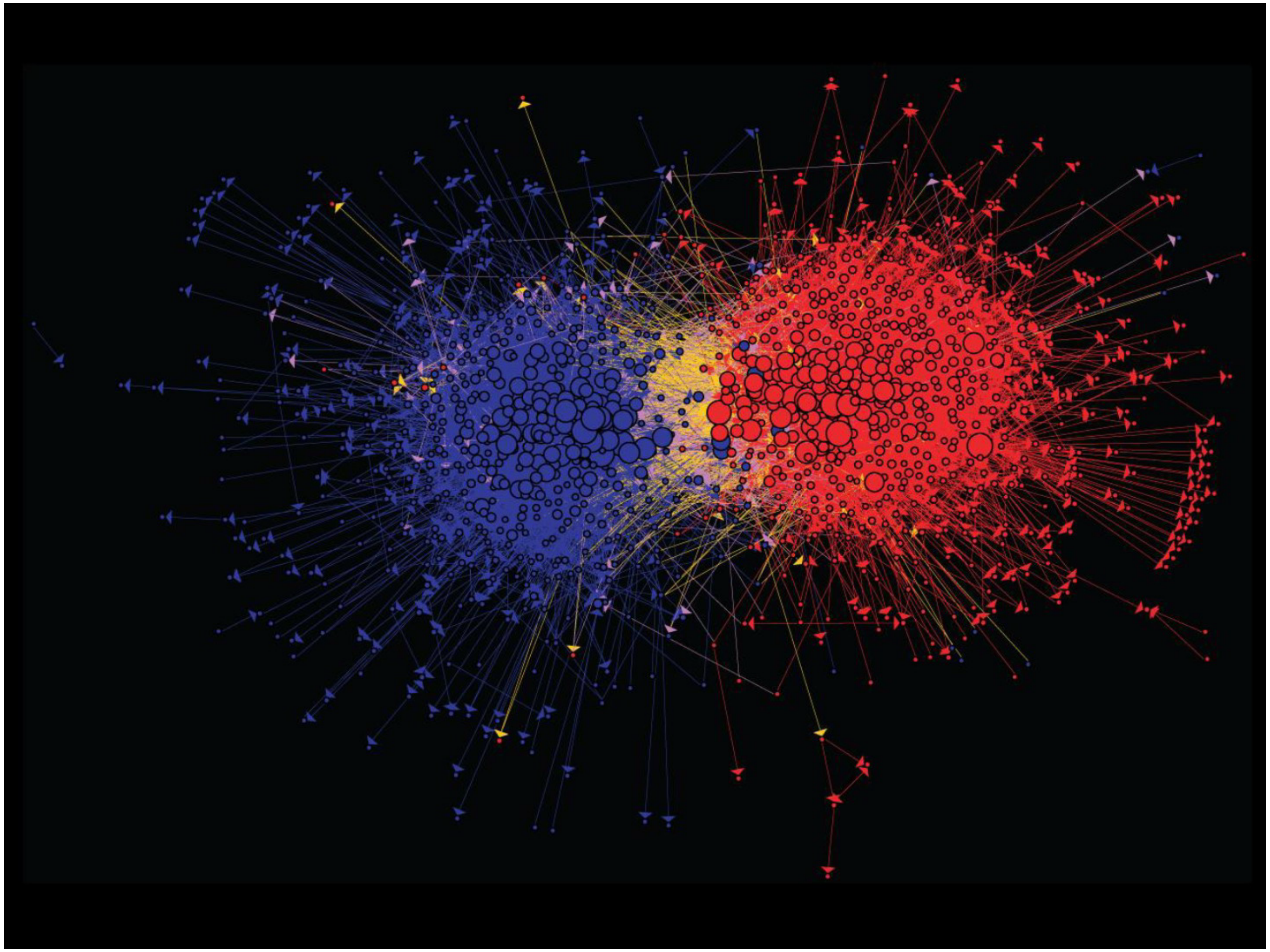
Airline networks

Application: Modeling Epidemics

- **Infrastructure networks are crucial for modeling epidemics**

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

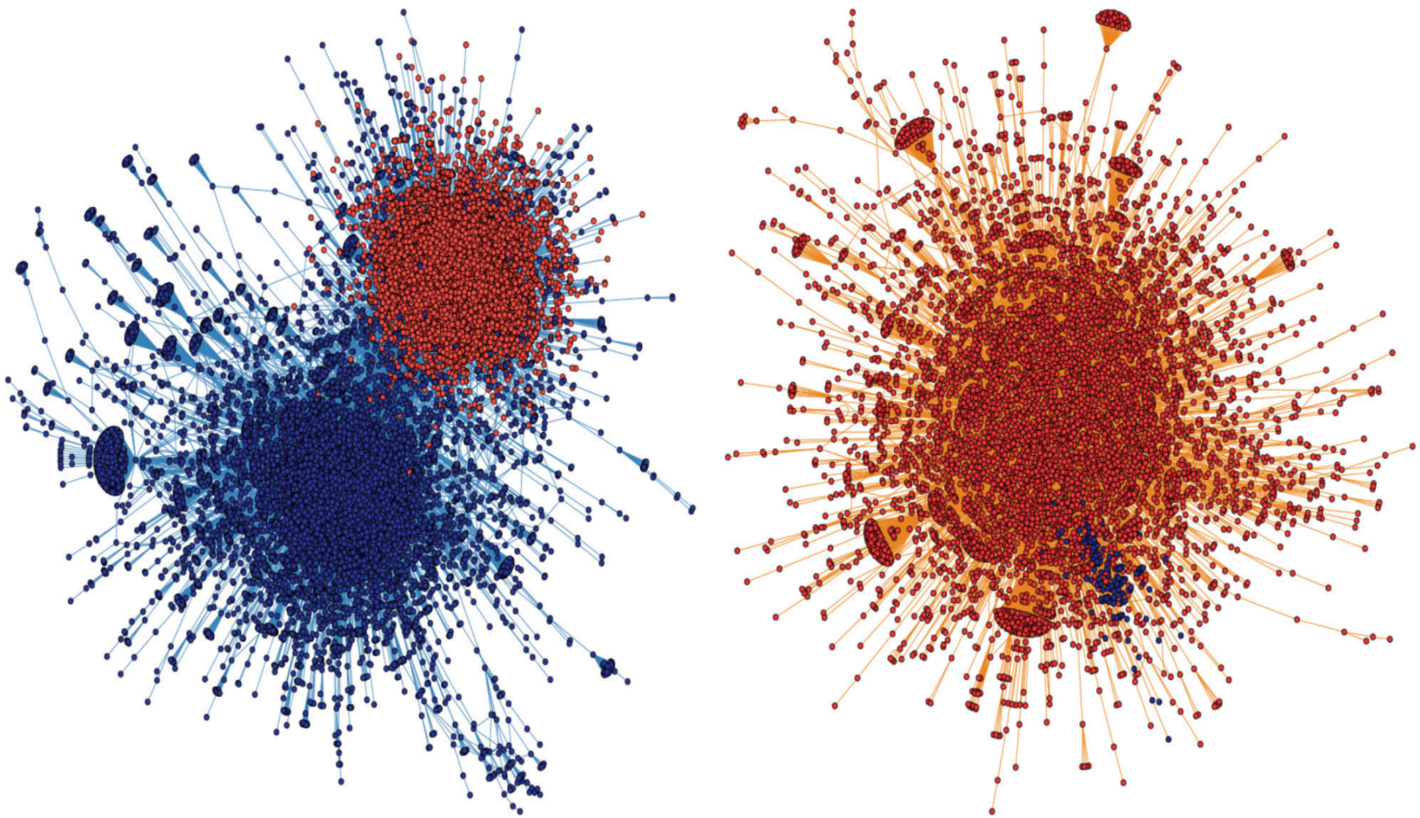
(3) Networks: Online Media



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

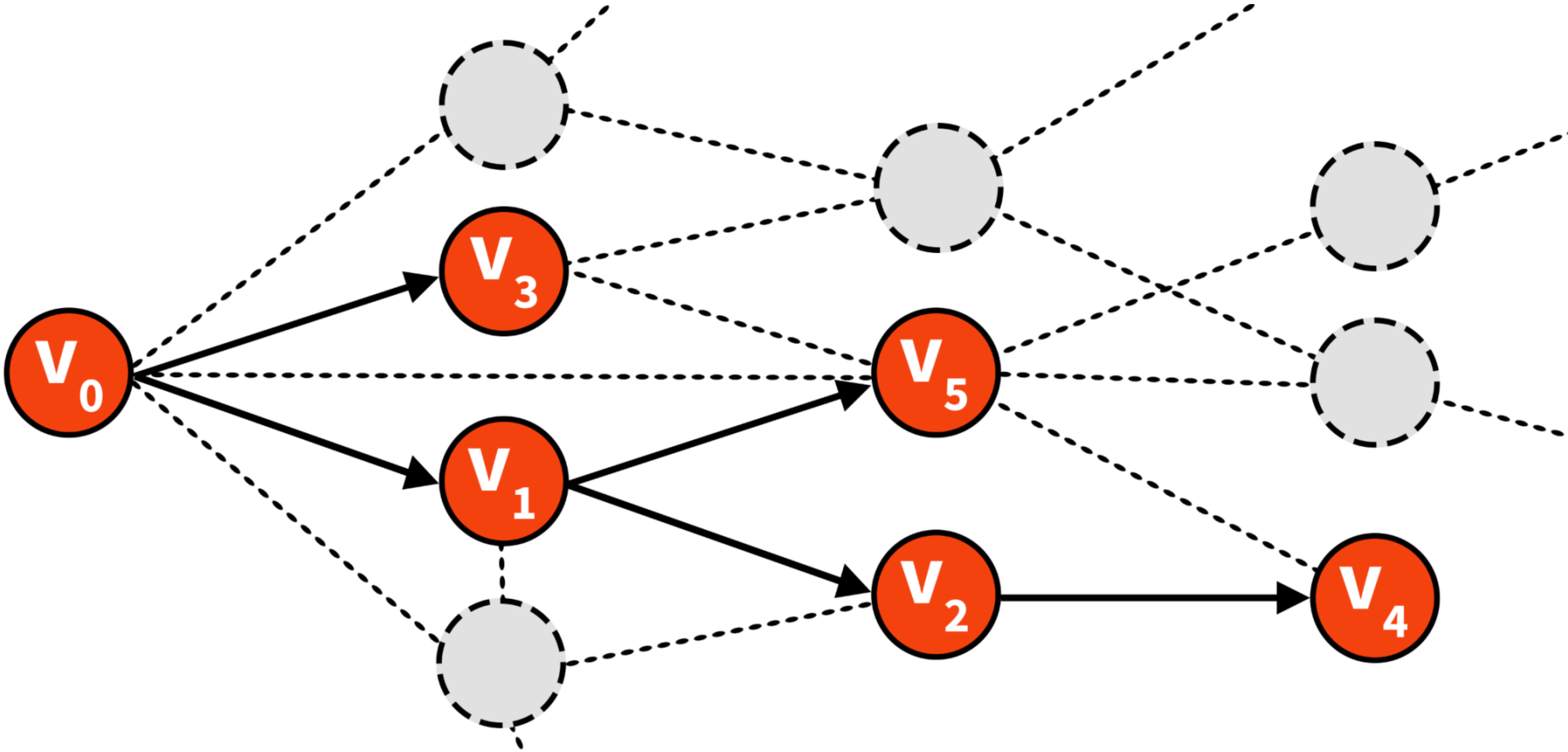
Application: Polarization on Twitter



- **Retweet networks:**
  Polarized (left), Unpolarized (rigth)

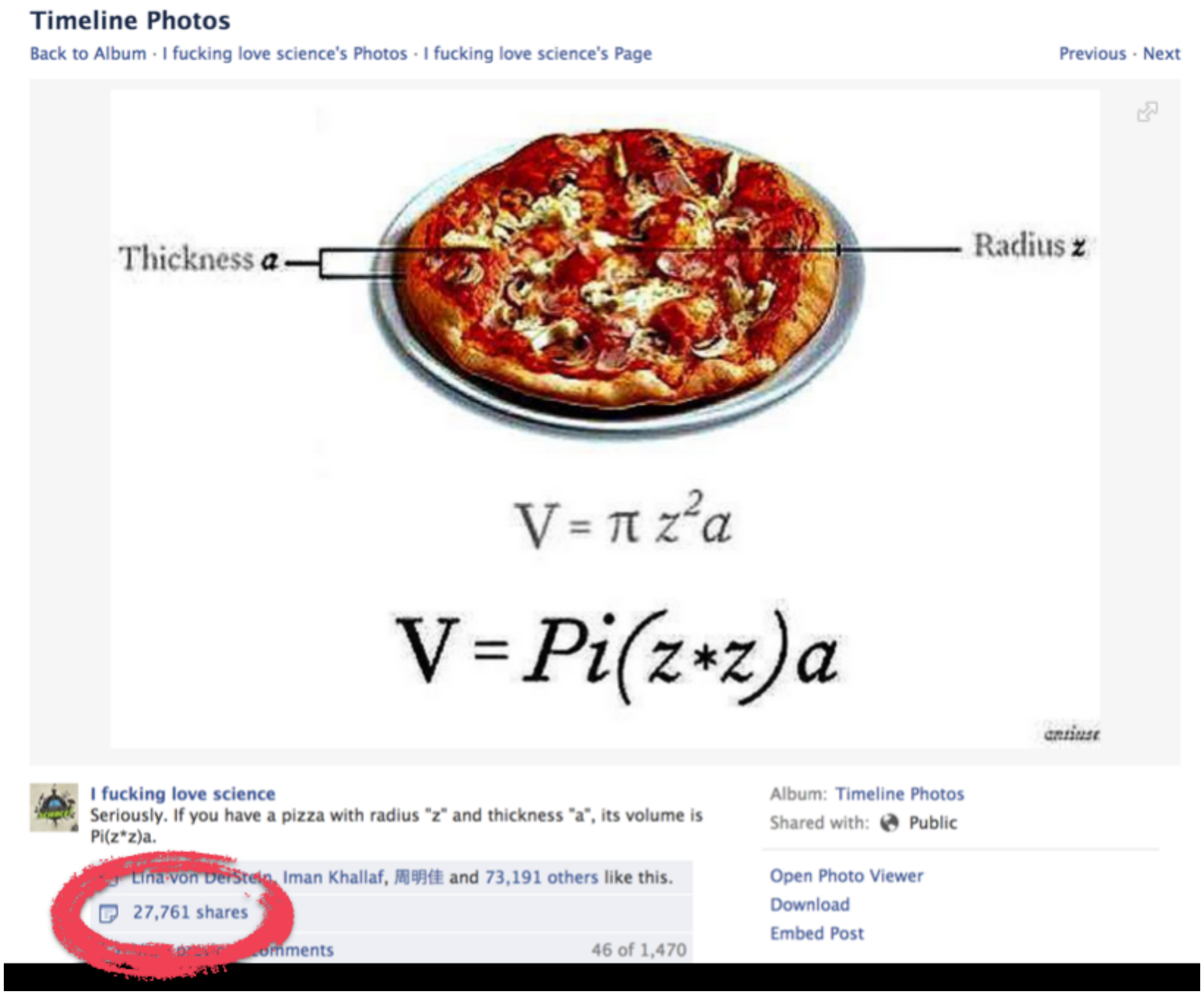
Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. "Political Polarization on Twitter." (2011)

Application: Information Diffusion



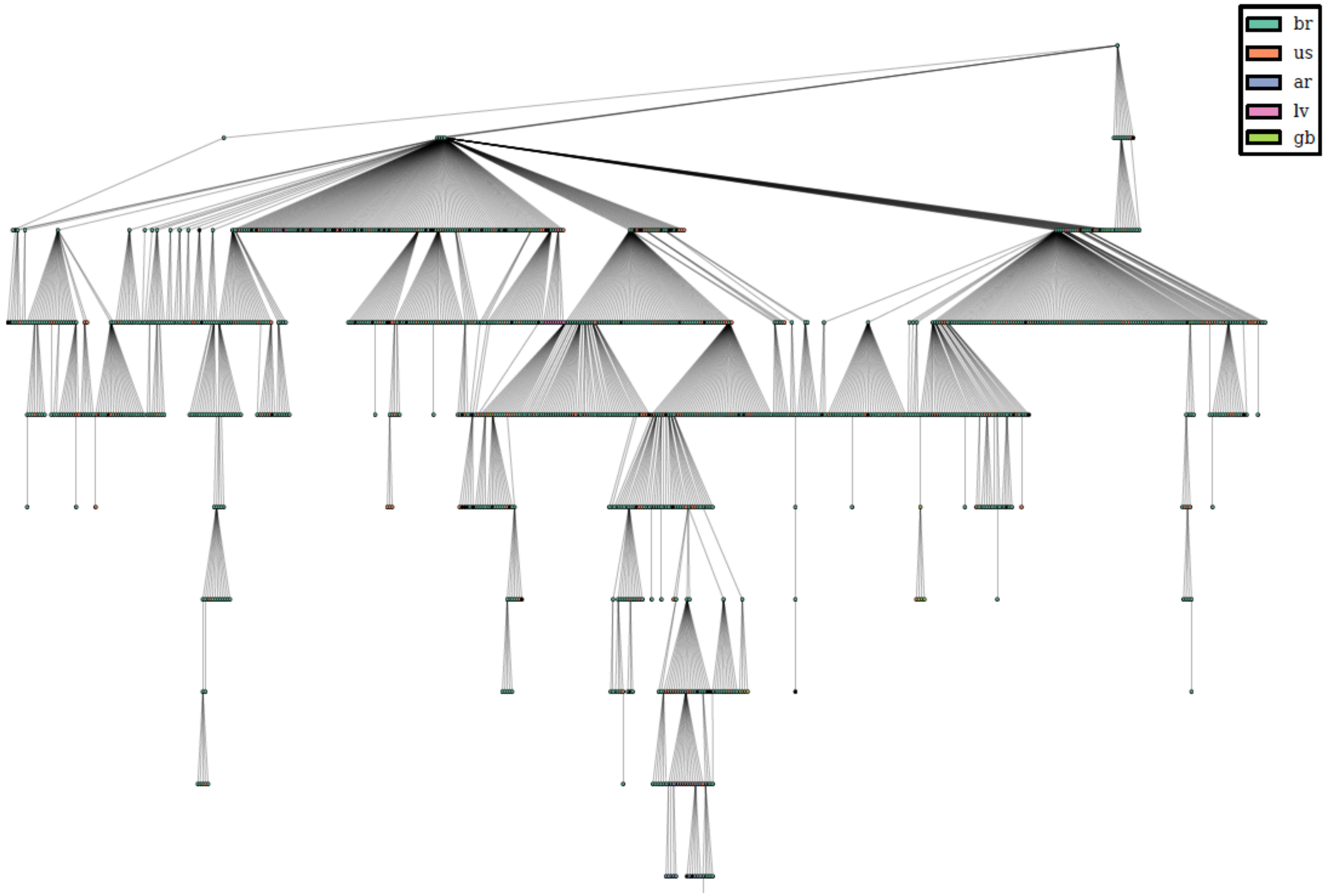
Information cascade in a social network

Facebook Information Cascades



[Can cascades be predicted?]() Cheng et al., WWW '14.

Application: Product Adoption

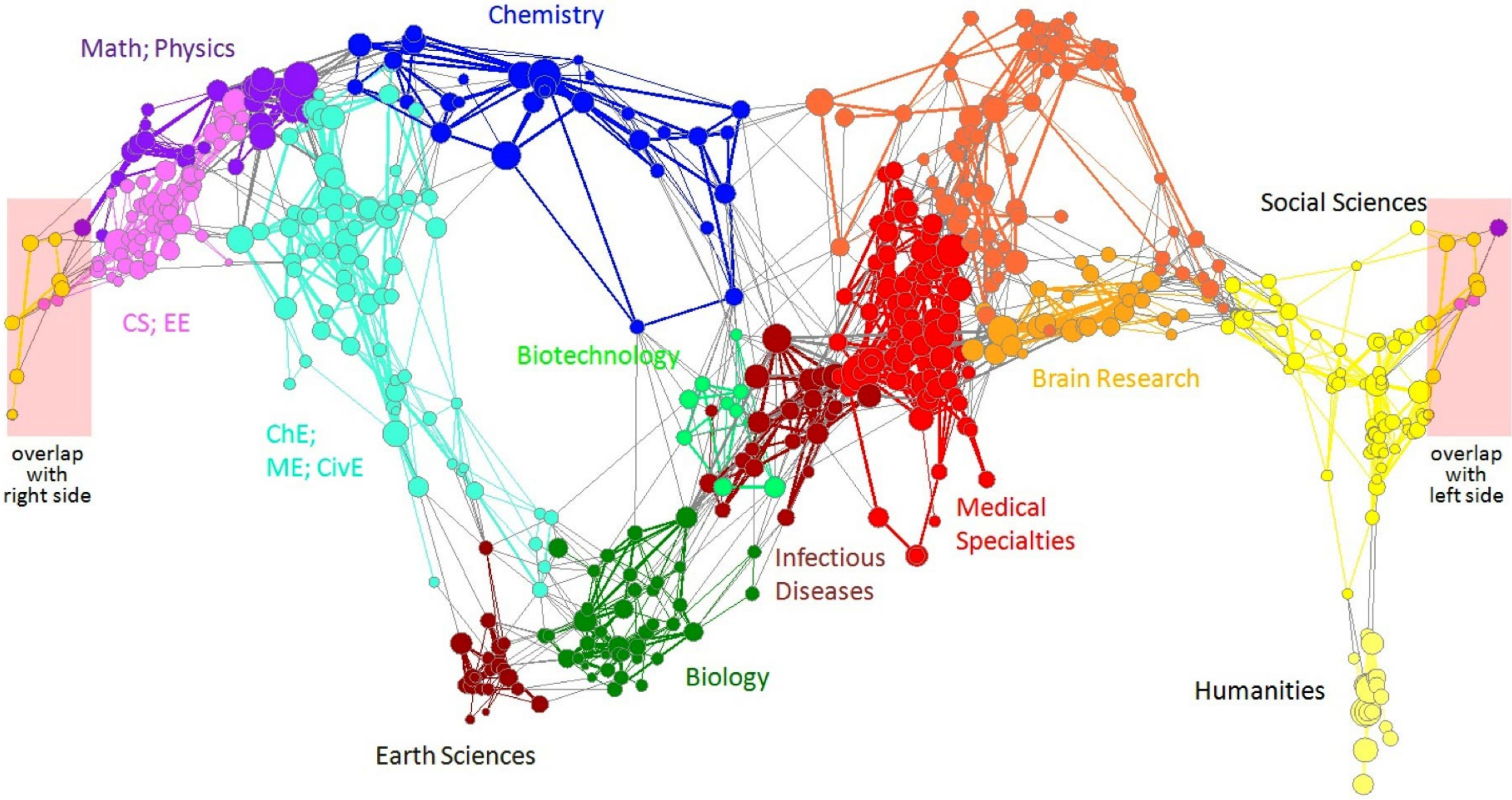


60-90% of LinkedIn users signed up due to an invitation from another user.

Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily.

Anderson et al., WWW '15.

(4) Networks: Information, Knowledge

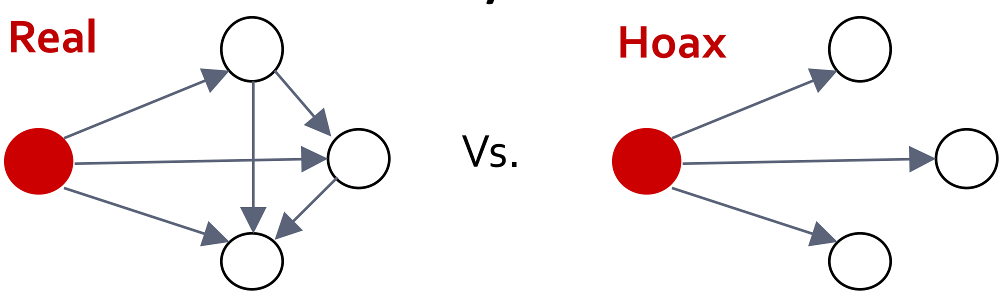


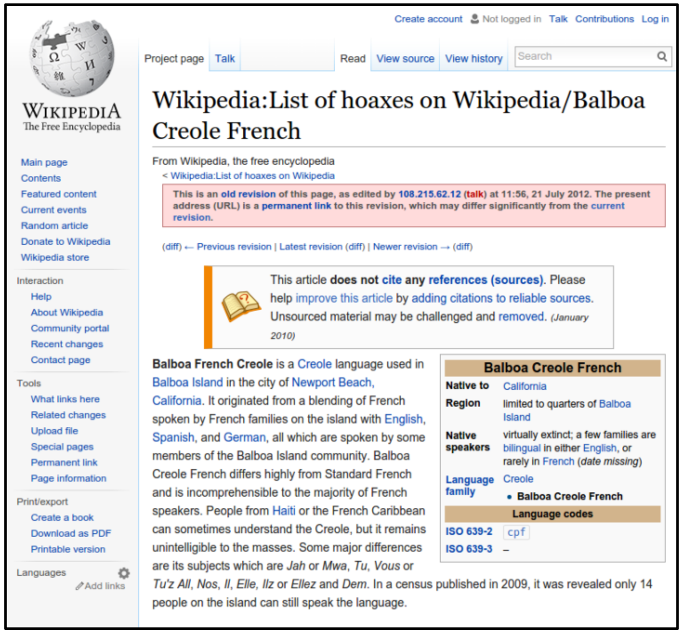
Citation networks and maps of science

[Börner et al., 2012]

Application: Misinformation

- **Q: Is a given Wikipedia article a hoax?**
  - Real articles link more coherently:





Create account Not logged in Talk Contributions Log in

Project page Talk Read View source View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information

Print/export
Create a book
Download as PDF
Printable version

Languages
Add links

Wikipedia:List of hoaxes on Wikipedia/Balboa Creole French

From Wikipedia, the free encyclopedia
< Wikipedia:List of hoaxes on Wikipedia

This is an old revision of this page, as edited by 108.215.62.12 (talk) at 11:56, 21 July 2012. The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

This article **does not cite any references (sources)**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2010)*

Balboa French Creole is a Creole language used in Balboa Island in the city of Newport Beach, California. It originated from a blending of French spoken by French families on the island with English, Spanish, and German, all which are spoken by some members of the Balboa Island community. Balboa Creole French differs highly from Standard French and is incomprehensible to the majority of French speakers. People from Haiti or the French Caribbean can sometimes understand the Creole, but it remains unintelligible to the masses. Some major differences are its subjects which are *Jah* or *Mwa*, *Tu*, *Vous* or *Tu'z All*, *Nos*, *Il*, *Elle*, *Ilz* or *Ellez* and *Dem*. In a census published in 2009, it was revealed only 14 people on the island can still speak the language.

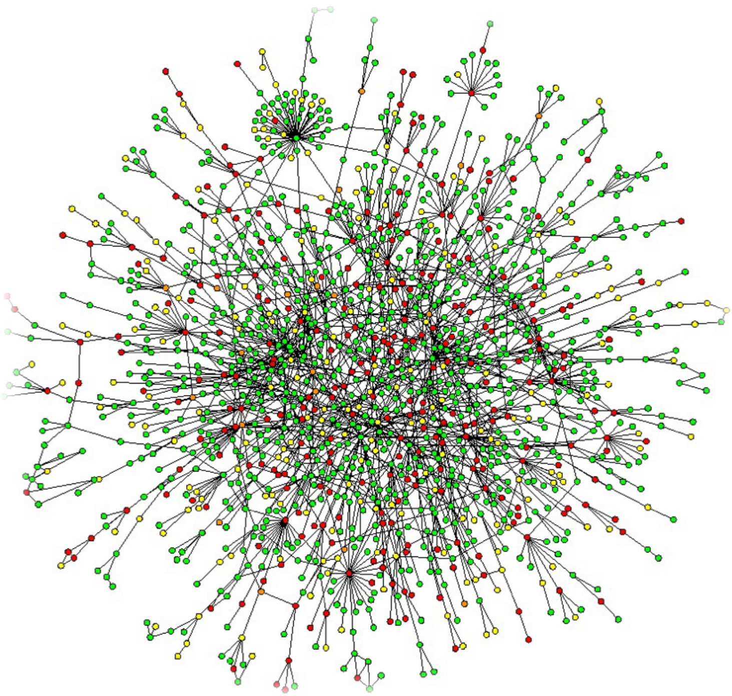
Balboa Creole French	
Native to	California
Region	limited to quarters of Balboa Island
Native speakers	virtually extinct; a few families are bilingual in either English, or rarely in French (*date missing*)
Language family	Creole • **Balboa Creole French**
Language codes	
ISO 639-2	cpf
ISO 639-3	–

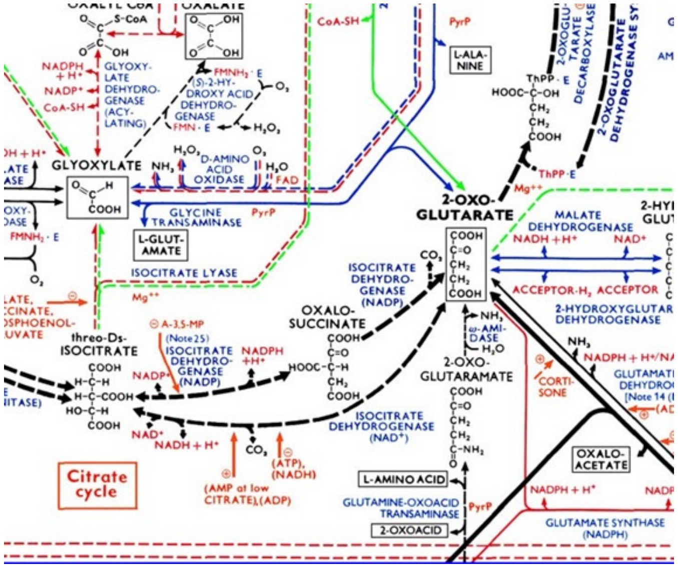
Hoax article detection performance:

50%	66%	86%
Random	Human	WWW '16

[Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes](). Kumar et al. WWW '16.

(5) Networks: Biology





Protein-protein interaction (PPI) networks:

Nodes: Proteins

Edges: ‘Physical’ interactions

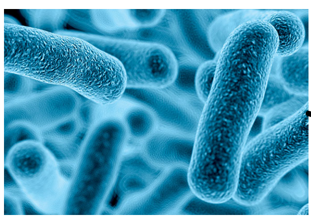
Metabolic networks:

Nodes: Metabolites and enzymes

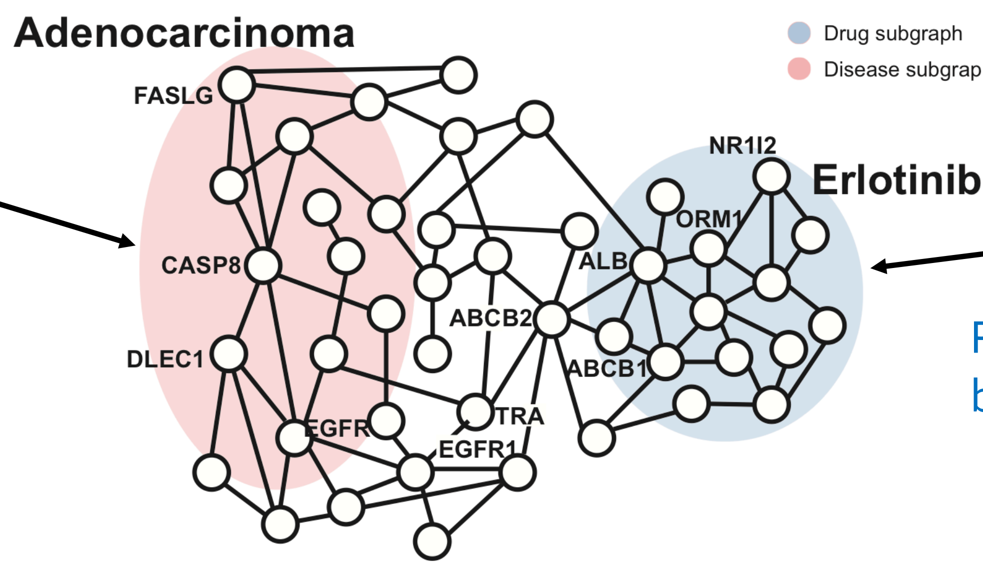
Edges: Chemical reactions

Application: Drug Repurposing

- **Q:** Can we predict therapeutic uses of a drug?
- **Insight:** Proteins are worker molecules in a cell. Protein interaction networks capture how the cell works.

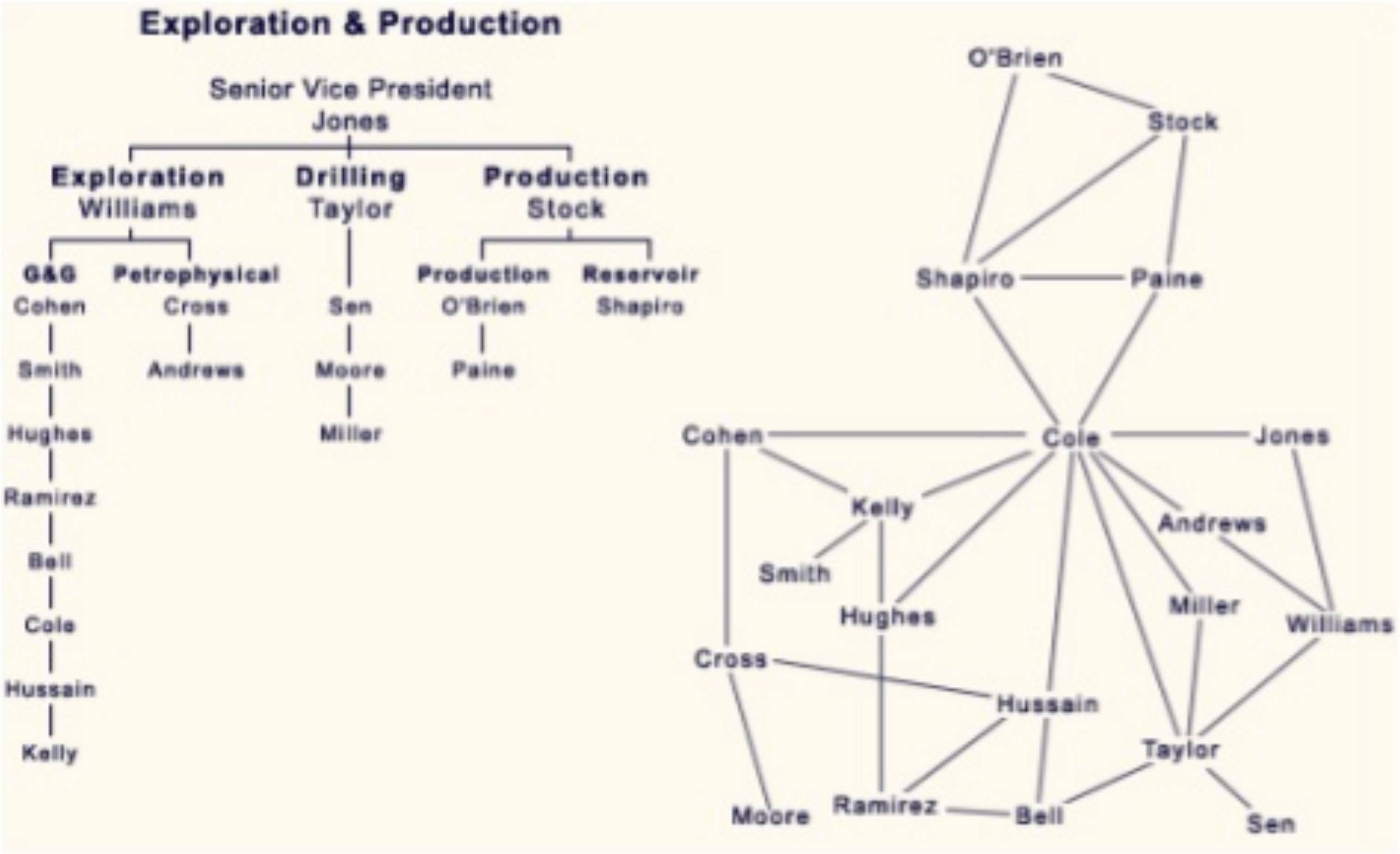


Proteins targeted by a disease



Proteins targeted by a drug

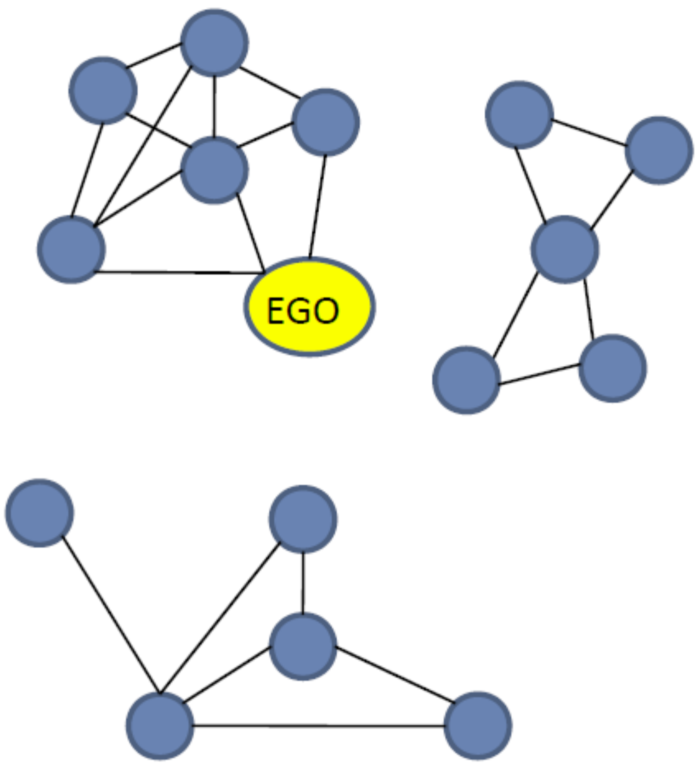
(6) Networks: Organizations



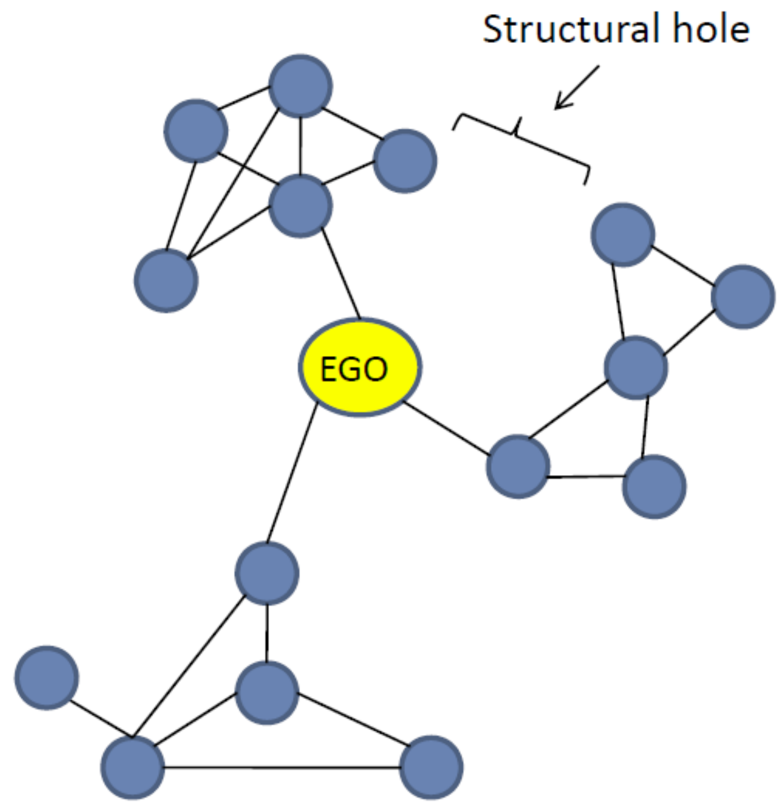
Who are the central nodes in the organization?

[Krebs, 2002]

Application: Employee Success



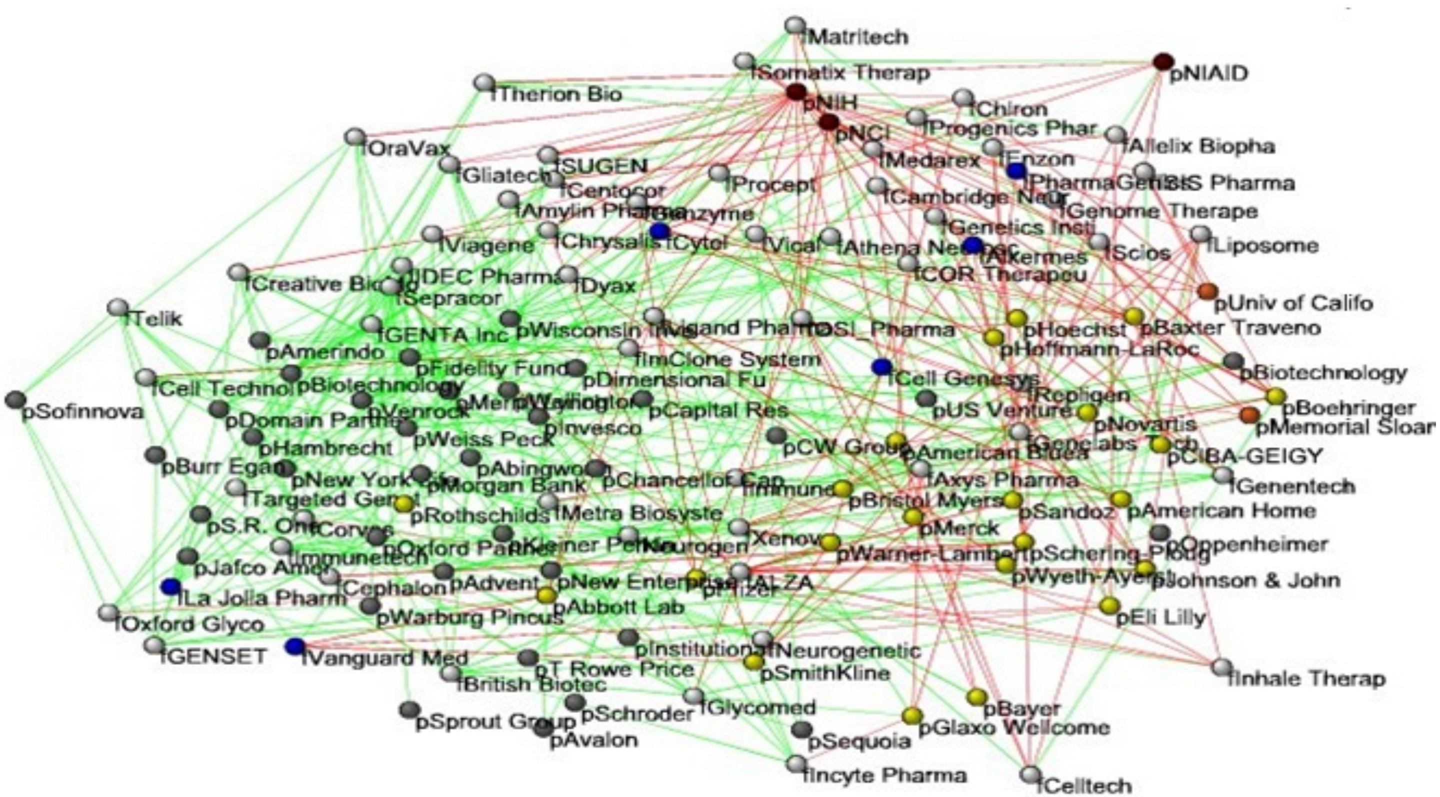
Few structural holes



Many structural holes

Structural Holes provide ego with access to novel information, power, freedom

Networks: Economic Networks



Nodes:

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

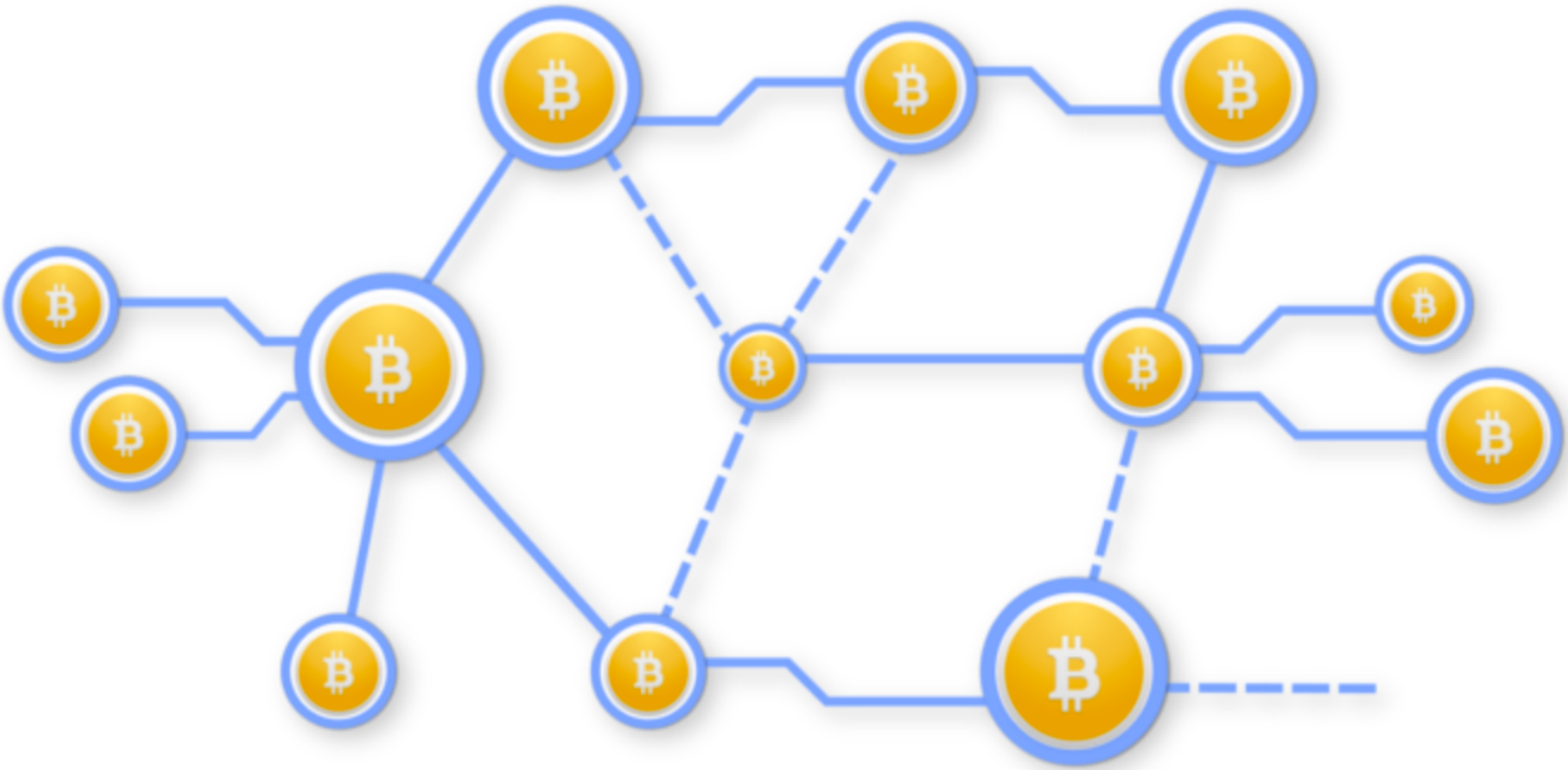
Links:

- Collaborations
- Financial
- R&D

Bio-tech companies: Why companies succeed?

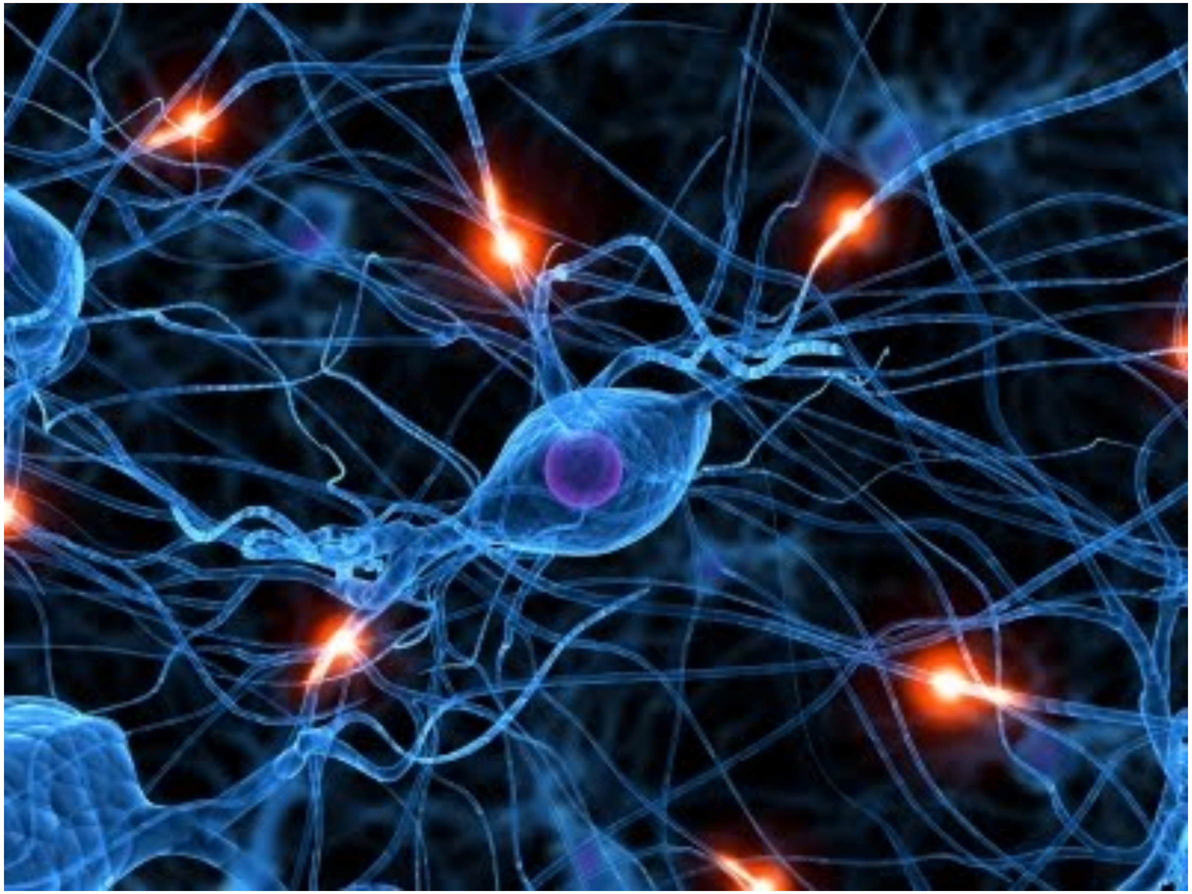
[Powell-White-Koput, 2002]

Networks: Transactions



- Detecting fraud and money laundering

Networks: Brain



Human brain has between 10-100 billion neurons

[Sporns, 2011]

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

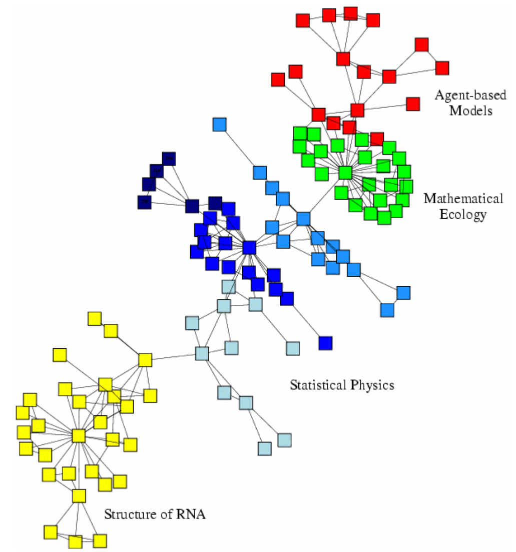
Reasoning about Networks

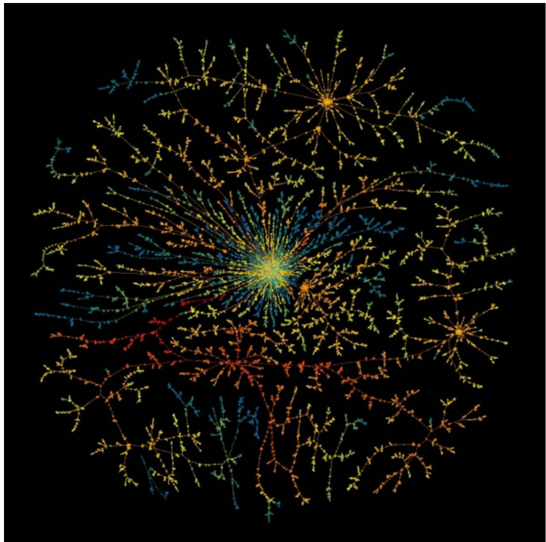
- **How do we reason about networks?**
  - **Empirical:** Study network data to find organizational principles
    - How do we measure and quantify networks?
  - **Mathematical models:** Graph theory and statistical models
    - Models allow us to understand behaviors and distinguish surprising from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

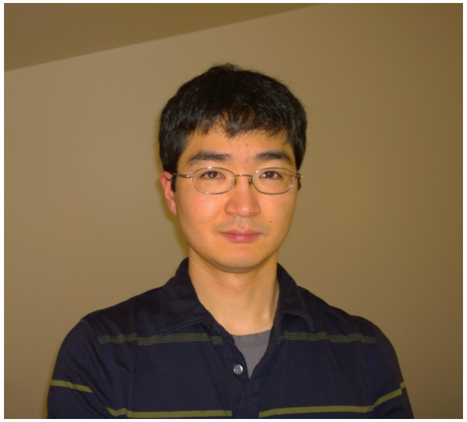




How It All Fits Together

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, LIM

Logistics: Course Assistants



Anthony Kim (Head TA)

Ziyi Yang

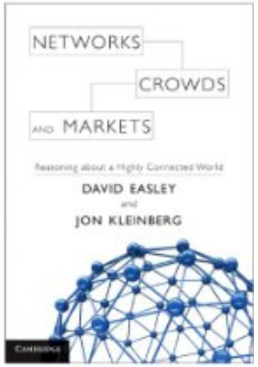
Anunay Kulshrestha

Silviana Ilcus

Praty Sharma

Logistics: Website

- http://cs224w.stanford.edu
  - Slides posted the night 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/fall2017/cs224w
    - Use access code "snap"
  - **Please participate and help each other! (2% of the grade)**
- **For e-mailing course staff, always use:**
  - cs224w-aut1718-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 use the code MRW7PY**
      - **Use your Stanford email (if non-SCPD) and include your Stanford ID # (everyone)**
    - **IMPORTANT: One answer per page!**
  - **Code and project write-ups** (proposal, milestone, final report) have to **also** be uploaded at http://snap.stanford.edu/submit/
- **Total of 2 late periods for the quarter:**
  - Late period expires on Monday at 23:59 Pacific Time
  - You can use at most 1 late period per assignment
  - No late periods for submissions related to final project

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**
  - Fine to have groups of 1 or 2. The team size will be taken under consideration when evaluating the scope of the project in breadth and depth. But 3 person teams can be more efficient.
  - Project is the **main work** for the class
  - We will help with ideas, data and mentoring
  - Start thinking about this now!
  - Ok to combine projects. Clearly indicate, which part of the project is done for CS224W and which part is done for the other class.
- Poster session with many external visitors
- **Read:** http://cs224w.stanford.edu/info.html#proj

Course Schedule

Week	Assignment	Due on (23:59 PST)
2	Homework 0	October 5
3	Homework 1	October 12
4	Project proposal	October 19 (no late periods!)
5	Homework 2	October 26
6	Work on the project	
7	Homework 3	November 9
8	Project milestone	November 16 (no late periods!)
9	Thanksgiving break	
10	Homework 4	November 30
11	Project report	**Sun December 10** (no late periods!)
	Poster session	**Mon, December 11** **12:15-3:15pm**

Work for the Course & Grading

- **Final grade will be composed of:**
  - **Homework: 48%**
    - Homework 1,2,3,4: 11.75% each, HW0: 1%
  - **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:**
    - Review of Probability, Linear Algebra, and Proof Techniques: **Thursday, 9/28 (4:30-5:20pm, Gates B03)**
    - SNAP.PY review and installation party: **Friday, 9/29 (4:30-5:20pm, Gates B03)**

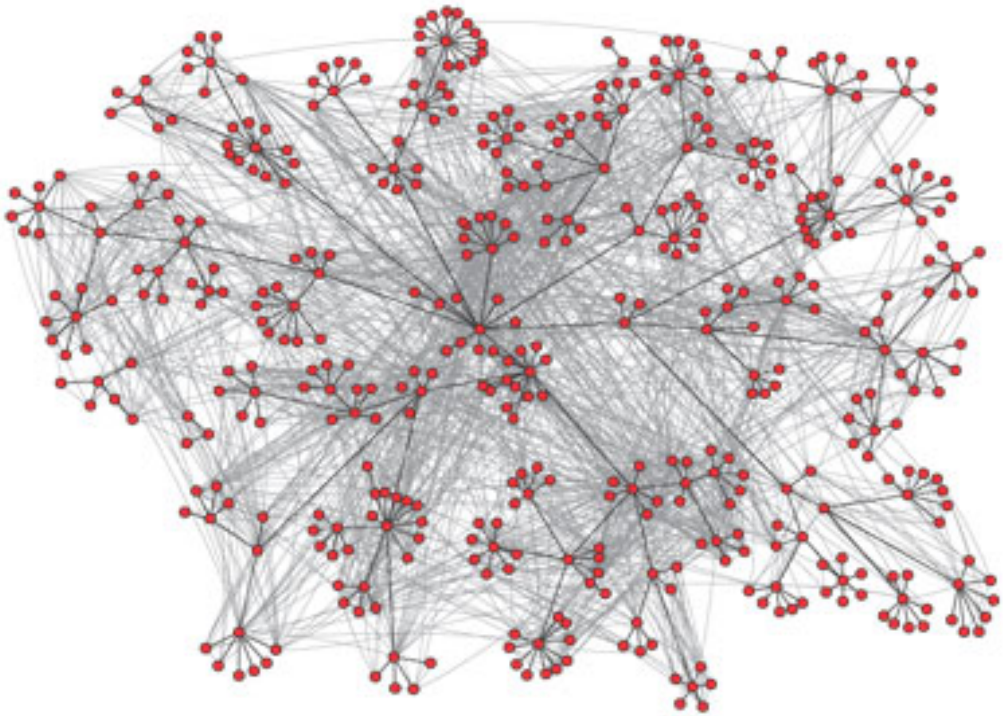
Network Analysis Tools

- **We highly recommend SNAP:**
  - **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
  - **SNAP C++:** more challenging but more scalable
  - Other tools include NetworkX, iGraph

SNAP.PY review and installation party:
Friday, 9/29 (4:30-5:20pm, Gates B03)

Starter Topic:
Structure of Graphs

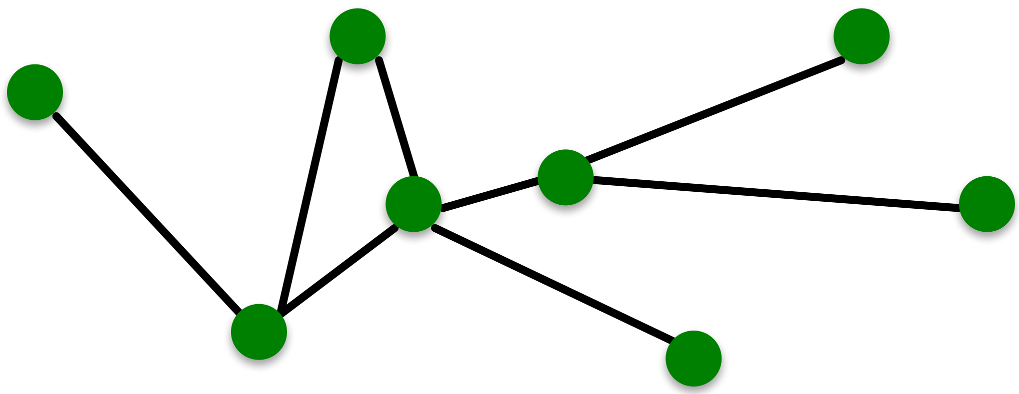
Structure of Networks?



A 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

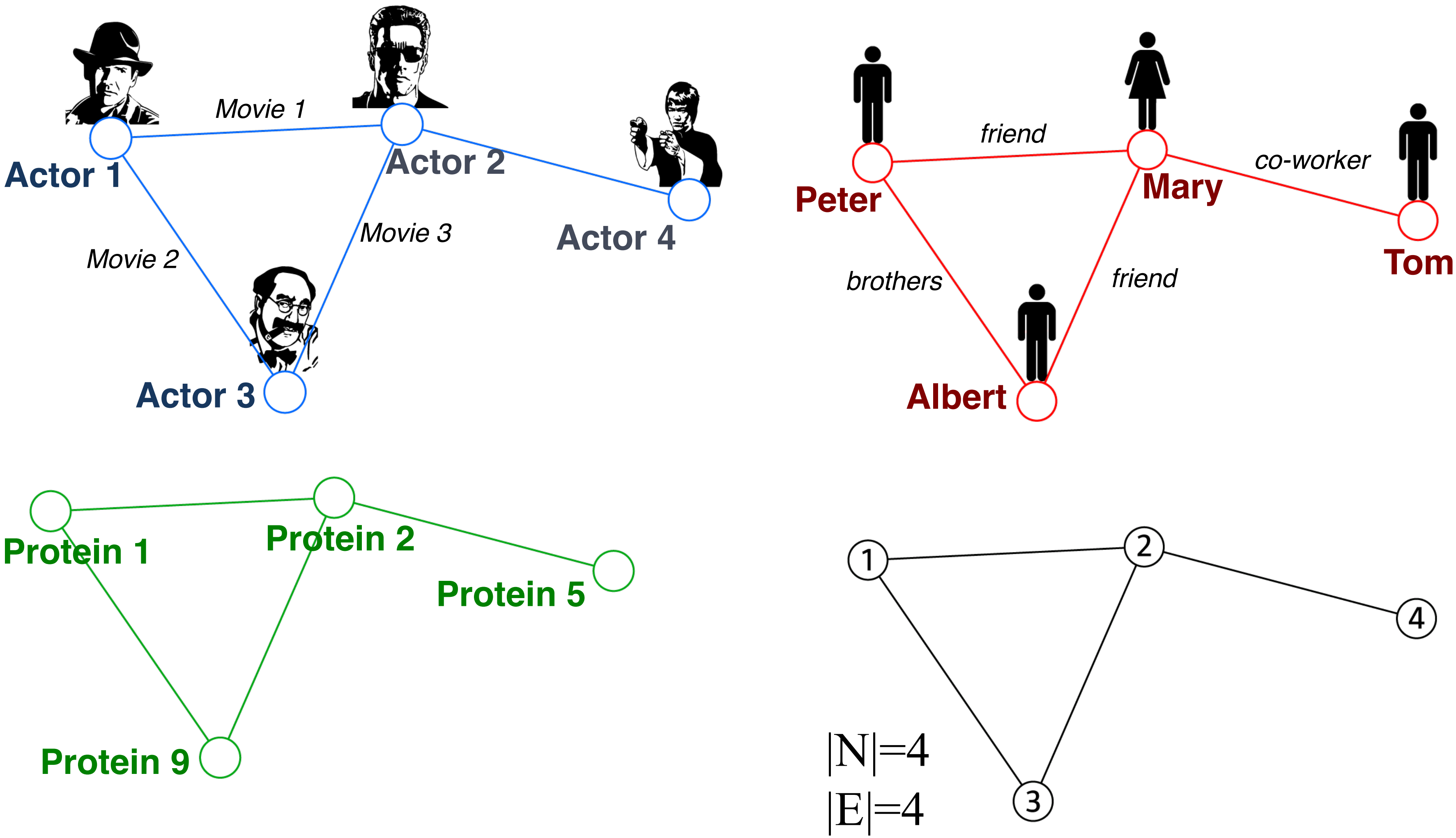
  Language: Network, node, link

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

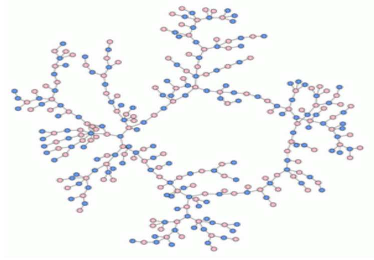
  Language: Graph, vertex, edge

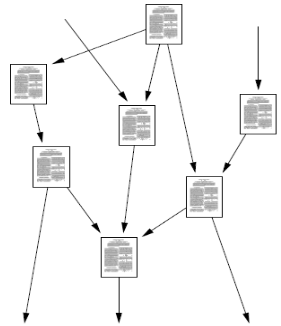
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably

Networks: Common Language



Choosing Proper Representations





- 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

How do you define a network?

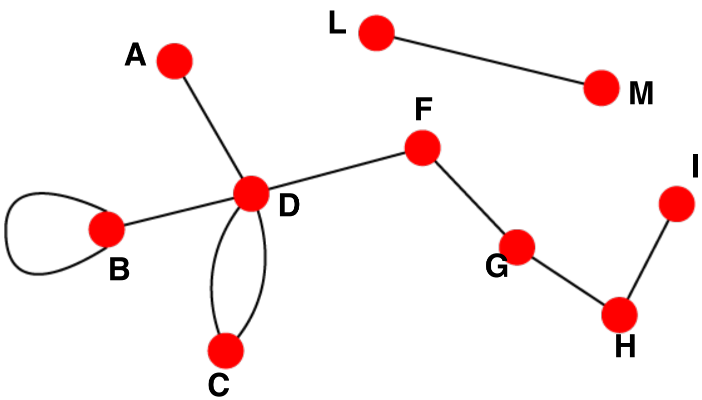
- **How to build a graph:**
  - **What are nodes?**
  - **What are edges?**
- **Choice of the proper network representation of a given domain/problem determines our ability to use networks successfully:**
  - In some cases there is a unique, unambiguous representation
  - In other cases, the representation is by no means unique
  - The way you assign links will determine the nature of the question you can study

Choice of Network Representation

Directed vs. Undirected Graphs

Undirected

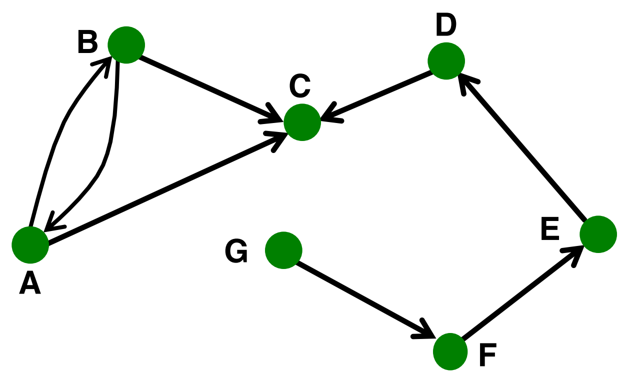
- **Links:** undirected (symmetrical, reciprocal)



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

Directed

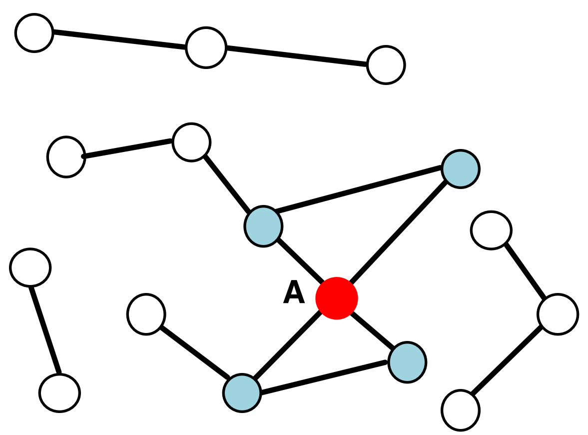
- **Links:** directed (arcs)



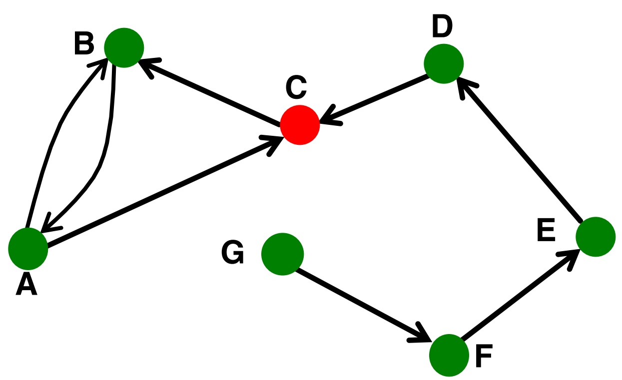
- **Examples:**
  - Phone calls
  - Following on Twitter

Node Degrees

Undirected



Directed



Source: Node with $k^{in} = 0$

Sink: Node with $k^{out} = 0$

Node degree, k_i: the number of edges adjacent to node i

$$k_A = 4$$

Avg. degree: $\bar{k} = \langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2E}{N}$

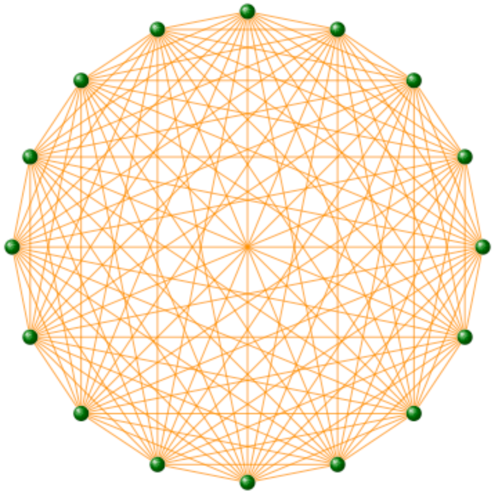
In directed networks we define an **in-degree** and **out-degree.**

The (total) degree of a node is the sum of in- and out-degrees.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

$$\bar{k} = \frac{E}{N} \qquad \overline{k^{in}} = \overline{k^{out}}$$

Complete Graph



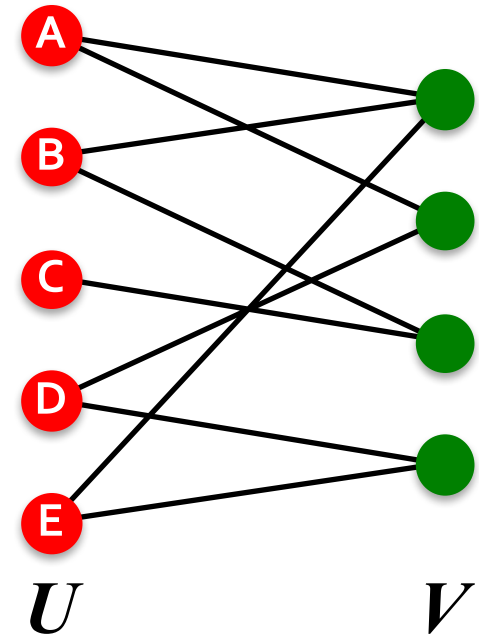
The **maximum number of edges** in an undirected graph on N nodes is

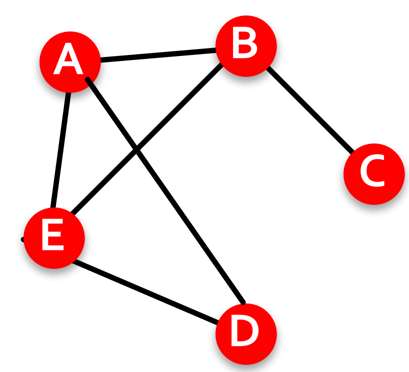
$$E_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

An undirected graph with the number of edges $E = E_{max}$ is called a **complete graph**, and its average degree is N-1

Bipartite Graph

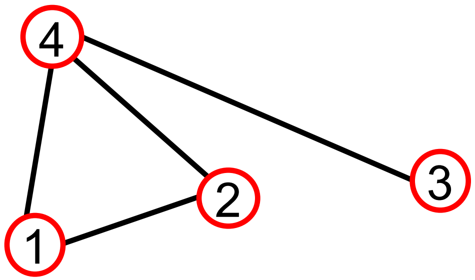
- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are **independent sets**
- **Examples:**
  - Authors-to-papers (they authored)
  - Actors-to-Movies (they appeared in)
  - Users-to-Movies (they rated)
- **"Folded" networks:**
  - Author collaboration networks
  - Movie co-rating networks

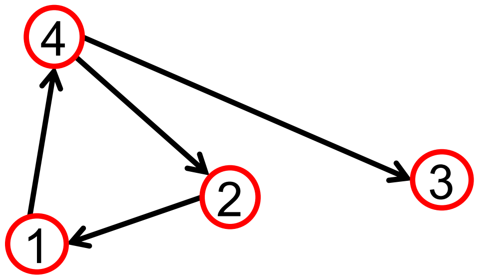




Folded version of the graph above

Representing Graphs: Adjacency Matrix





$A_{ij} = 1$ if there is a link from node ***i*** to node ***j***

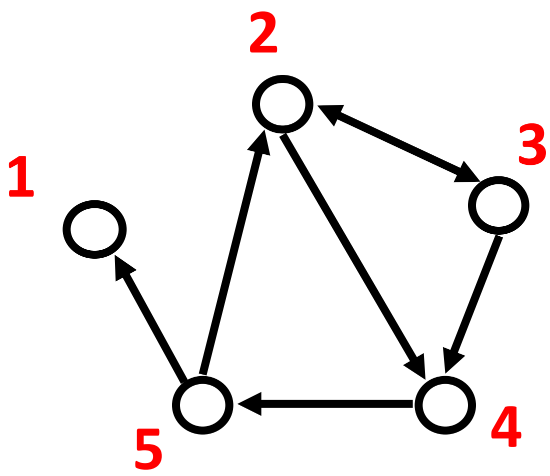
$A_{ij} = 0$ otherwise

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

Note that for a directed graph (right) the matrix is not symmetric.

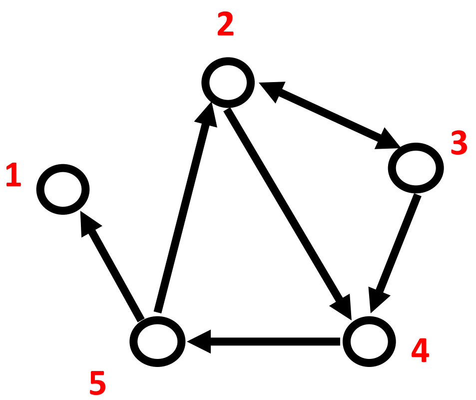
Representing Graphs: Edge list

- **Represent graph as a set of edges:**
  - (2, 3)
  - (2, 4)
  - (3, 2)
  - (3, 4)
  - (4, 5)
  - (5, 2)
  - (5, 1)



Representing Graphs: Adjacency list

- **Adjacency list:**
  - Easier to work with if network is
    - Large
    - Sparse
  - Allows us to quickly retrieve all neighbors of a given node
    - 1:
    - 2: 3, 4
    - 3: 2, 4
    - 4: 5
    - 5: 1, 2



Networks are Sparse Graphs

Most real-world networks are sparse

$$E \ll E_{max} \quad (\text{or } \bar{k} \ll N-1)$$

WWW (Stanford-Berkeley):	N=319,717	$\langle k \rangle$=9.65
Social networks (LinkedIn):	N=6,946,668	$\langle k \rangle$=8.87
Communication (MSN IM):	N=242,720,596	$\langle k \rangle$=11.1
Coauthorships (DBLP):	N=317,080	$\langle k \rangle$=6.62
Internet (AS-Skitter):	N=1,719,037	$\langle k \rangle$=14.91
Roads (California):	N=1,957,027	$\langle k \rangle$=2.82
Proteins (S. Cerevisiae):	N=1,870	$\langle k \rangle$=2.39

(Source: *Leskovec et al., Internet Mathematics, 2009*)

Consequence: Adjacency matrix is filled with zeros!

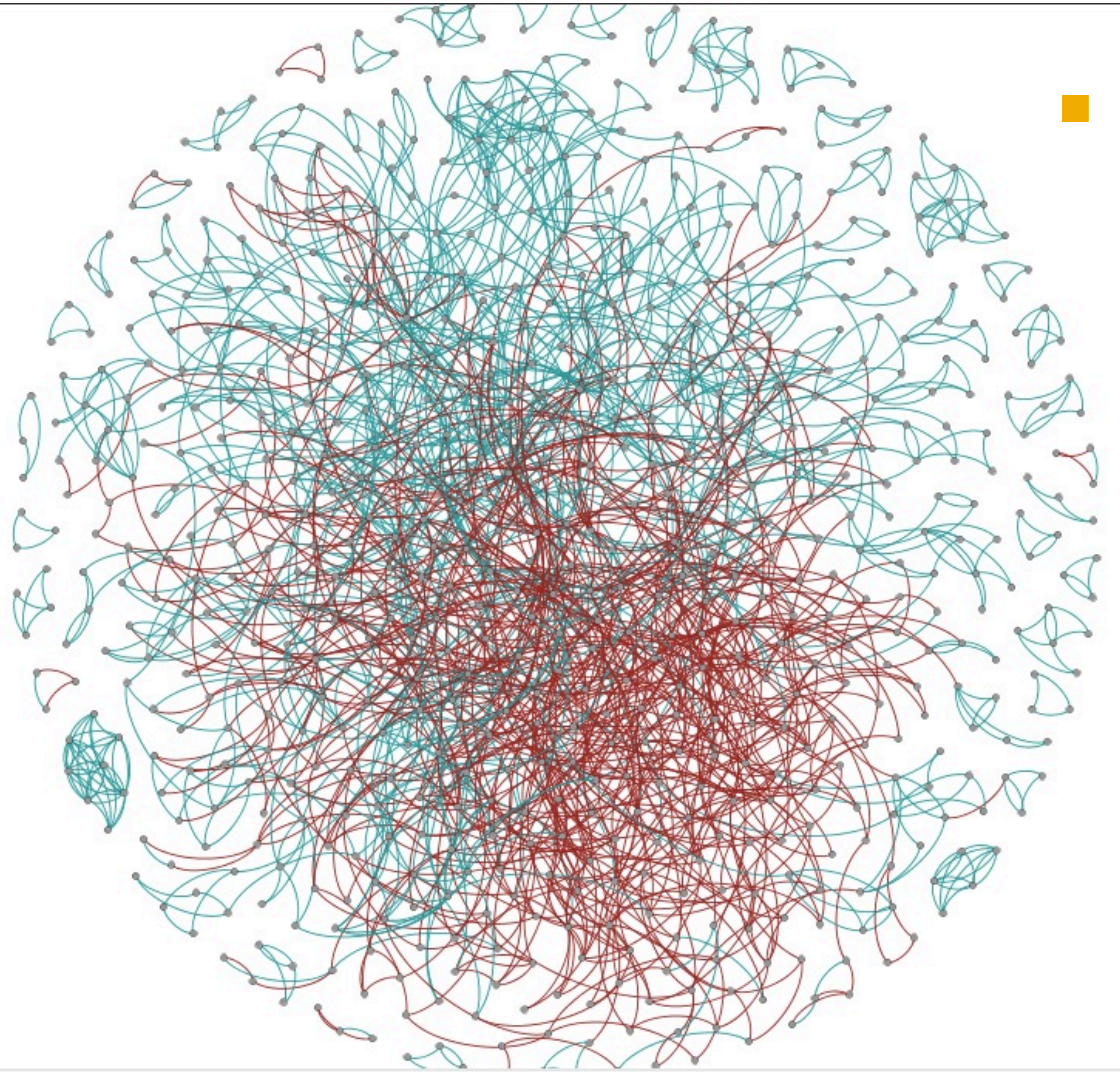
(**Density of the matrix (E/N^2):** WWW=1.51×10^{-5}, MSN IM = 2.27×10^{-8})

Edge Attributes

Possible options:

- Weight (e.g. frequency of communication)
- Ranking (best friend, second best friend…)
- Type (friend, relative, co-worker)
- Sign: Friend vs. Foe, Trust vs. Distrust
- Properties depending on the structure of the rest of the graph: number of common friends

Positive and Negative Weights

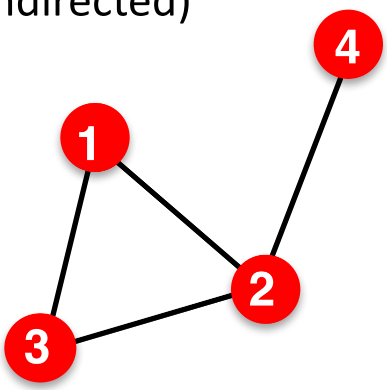


- **One person trusting/distrusting another**
  - Research challenge: How does one 'propagate' negative feelings in a social network? Is my enemy's enemy my friend?

sample of positive & negative ratings from Epinions network

More Types of Graphs

- **Unweighted**
  (undirected)



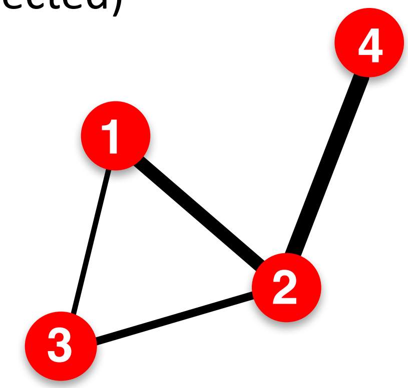
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \bar{k} = \frac{2E}{N}$$

Examples: Friendship, Hyperlink

- **Weighted**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

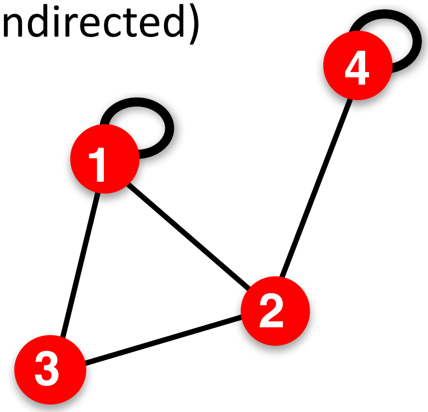
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \bar{k} = \frac{2E}{N}$$

Examples: Collaboration, Internet, Roads

More Types of Graphs

- **Self-edges (self-loops)**
  (undirected)



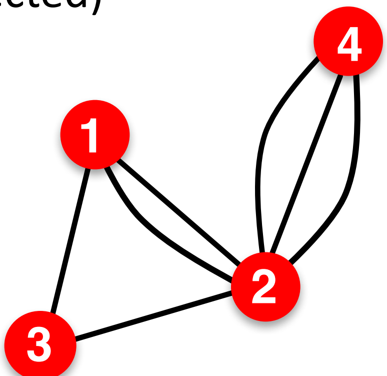
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2} \sum_{i,j=1, i \neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}$$

Examples: Proteins, Hyperlinks

- **Multigraph**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

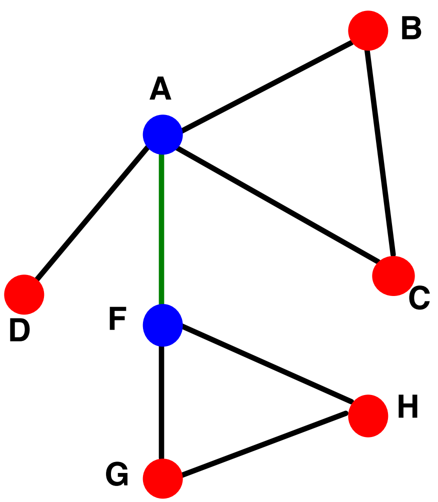
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

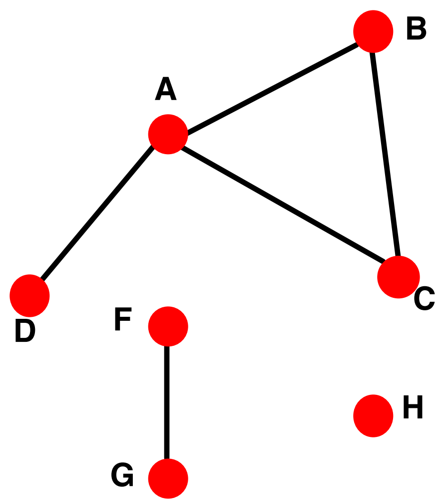
$$E = \frac{1}{2} \sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \bar{k} = \frac{2E}{N}$$

Examples: Communication, Collaboration

Connectivity of Undirected 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





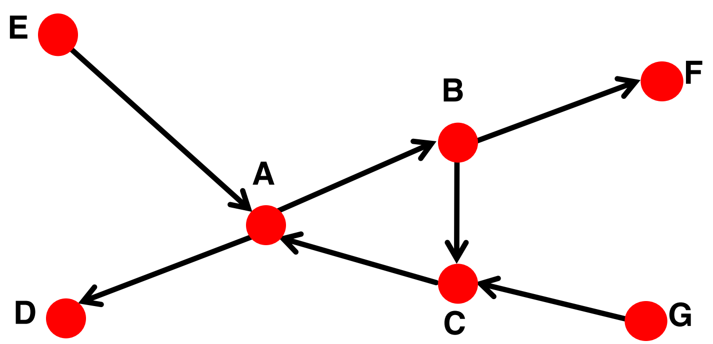
Largest Component:
Giant Component

Isolated node (node H)

Bridge edge: If we erase the **edge**, the graph becomes disconnected.
Articulation node: If we erase the **node**, 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).

Network Representations

Email network >> directed multigraph with self-edges

Facebook friendships >> undirected, unweighted

Citation networks >> unweighted, directed, acyclic

Collaboration networks >> undirected multigraph or weighted graph

Mobile phone calls >> directed, (weighted?) multigraph

Protein Interactions >> undirected, unweighted with self-interactions