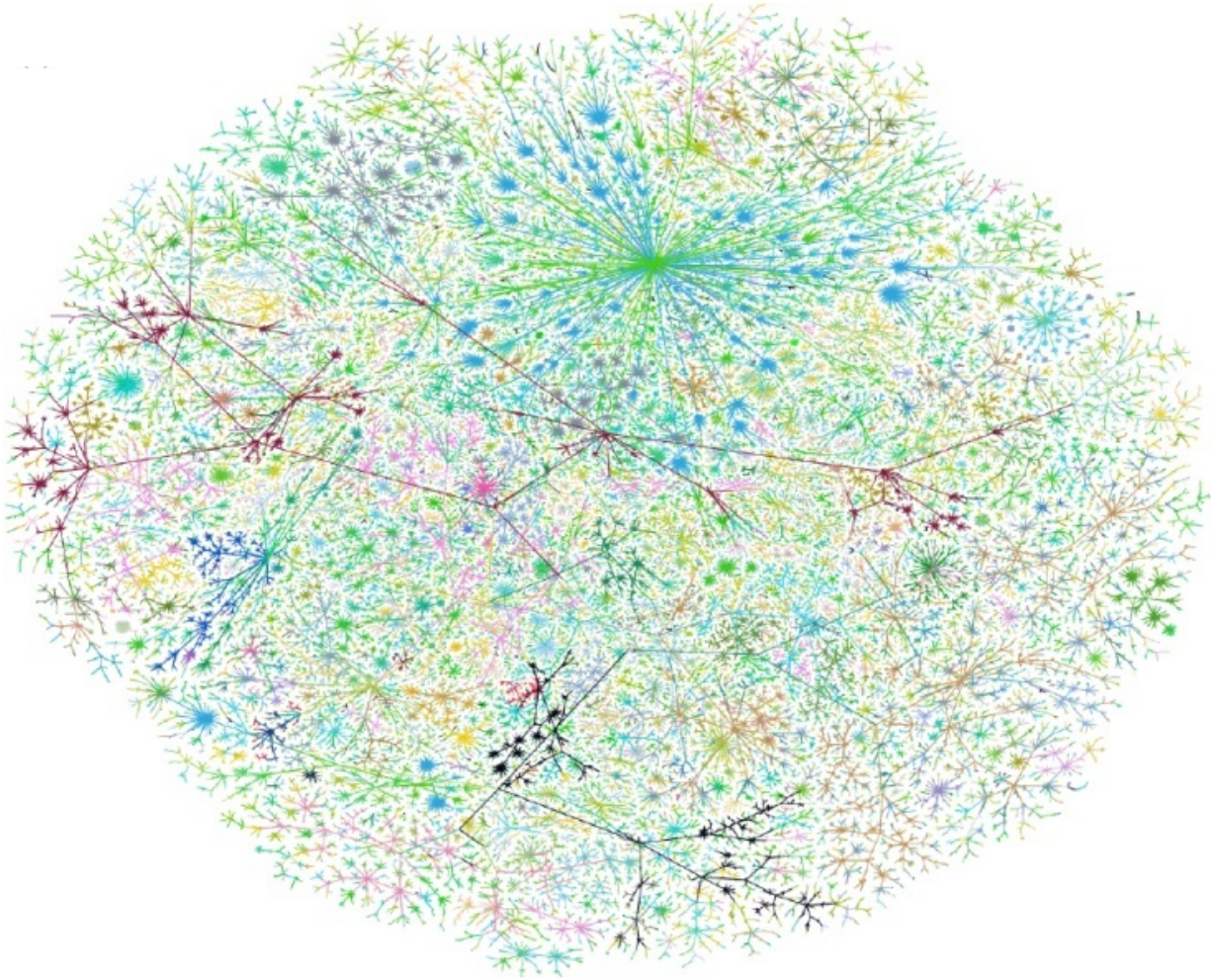
CS224W: Social and Information Network Analysis

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

Networks & Complex Systems

- **Complex systems are 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** 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?



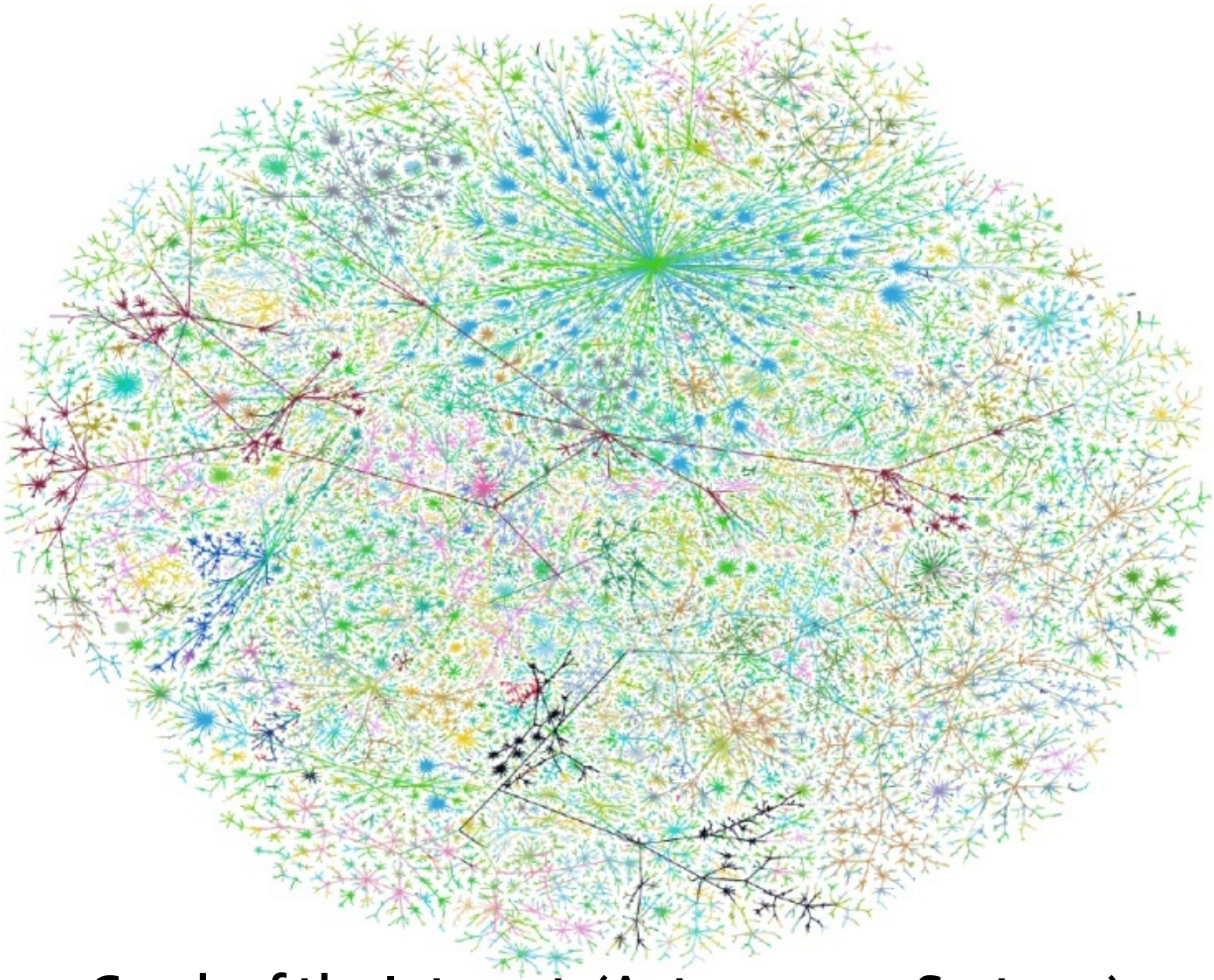
The Network!

Networks: Social

Facebook social graph

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

Networks: Communication

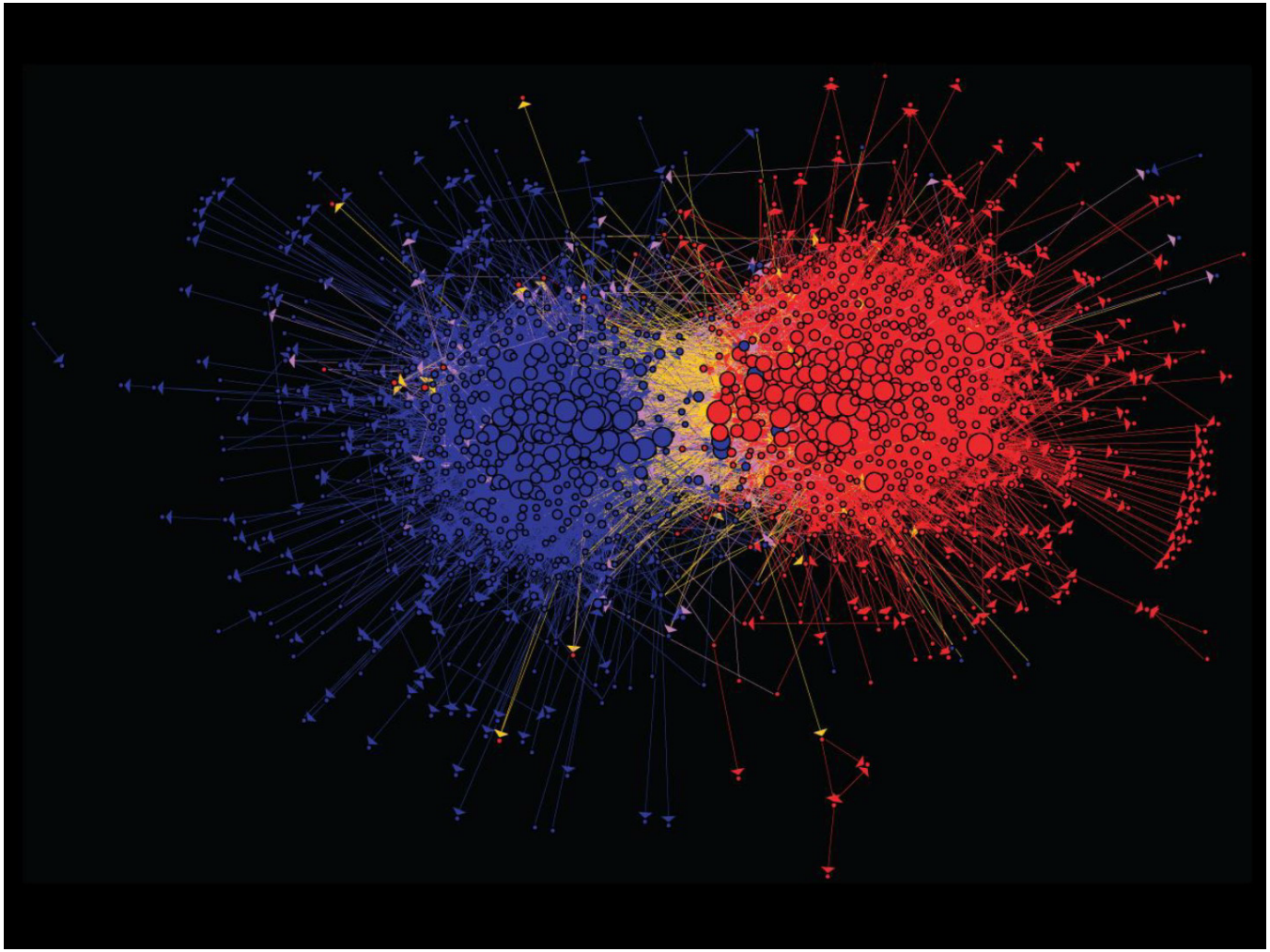


Graph of the Internet (Autonomous Systems)

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

Robustness [Doyle-Willinger, 2005]

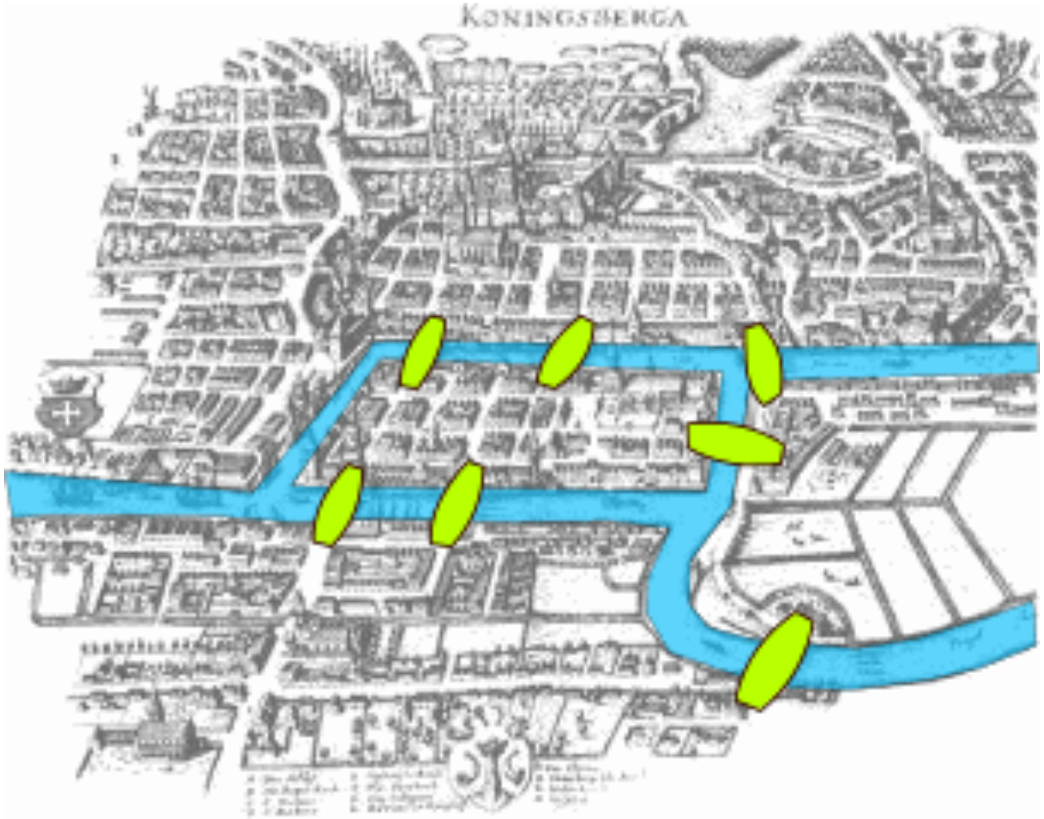
Networks: Media



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

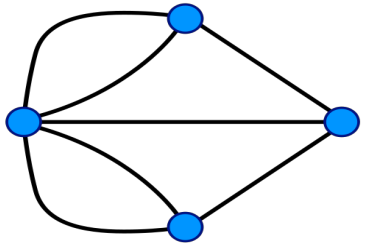
Networks: Technology



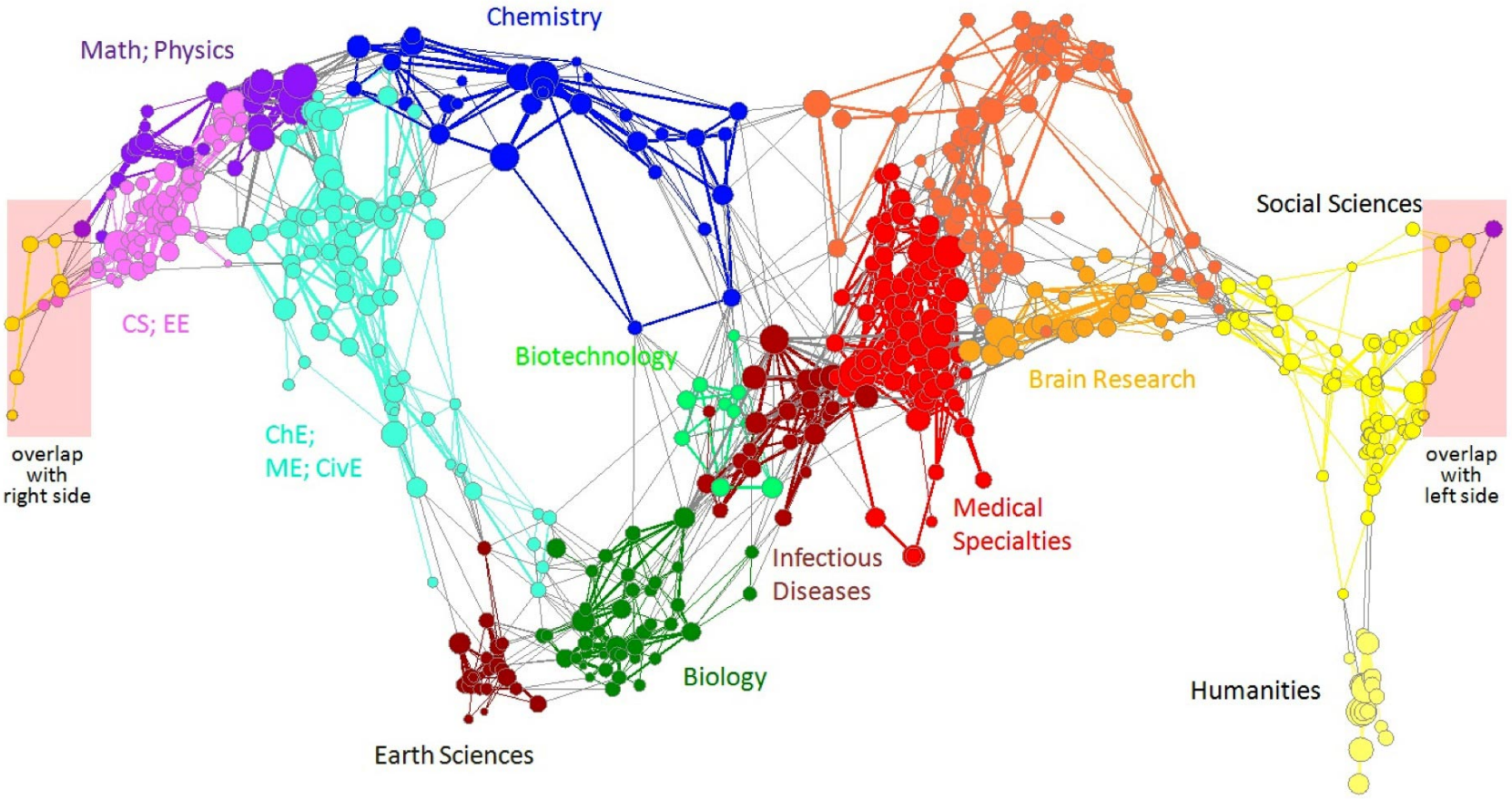
Seven Bridges of Königsberg

[Euler, 1735]

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



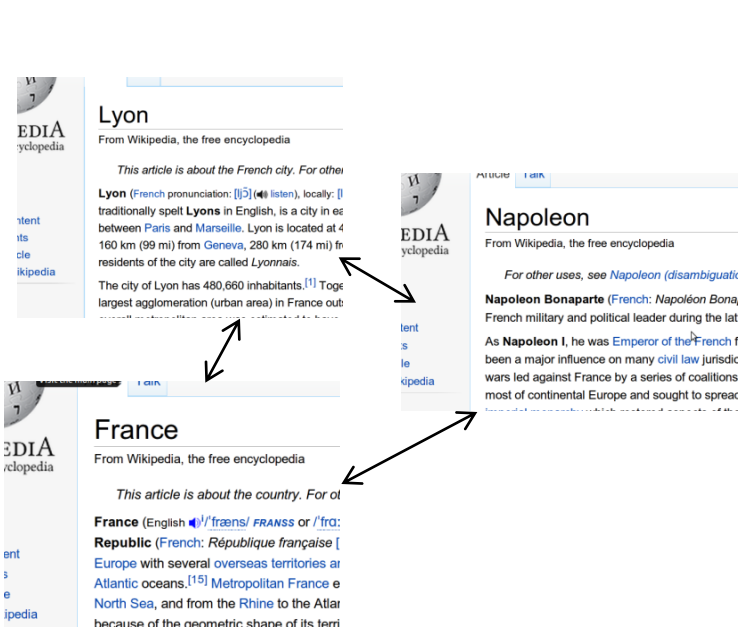
Networks: Information

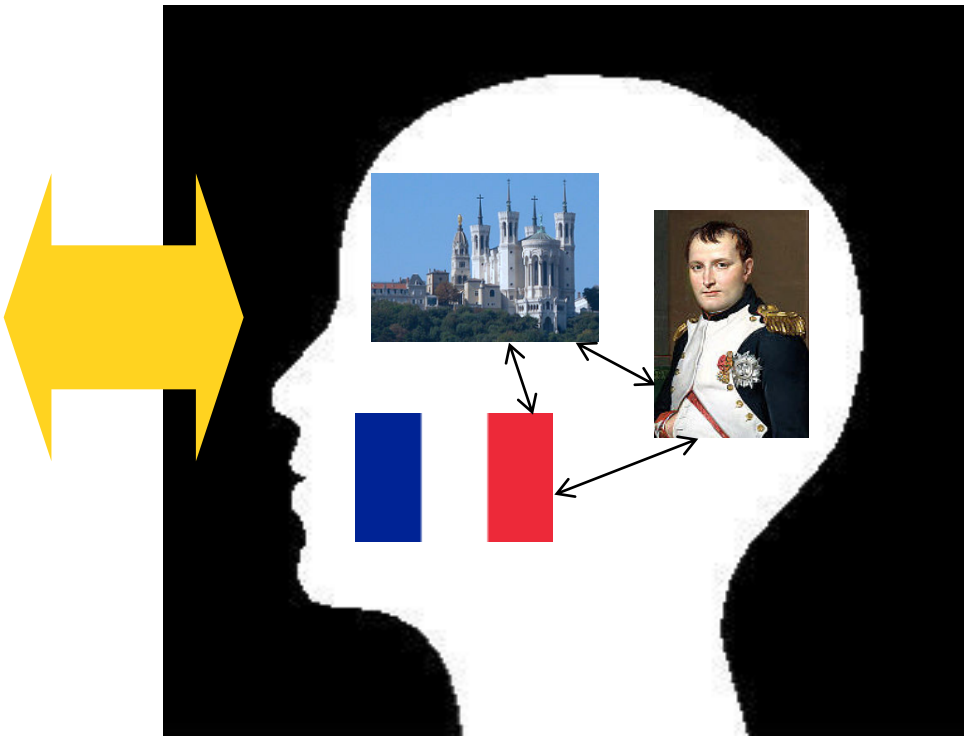


Citation networks and maps of science

[Börner et al., 2012]

Networks: Knowledge



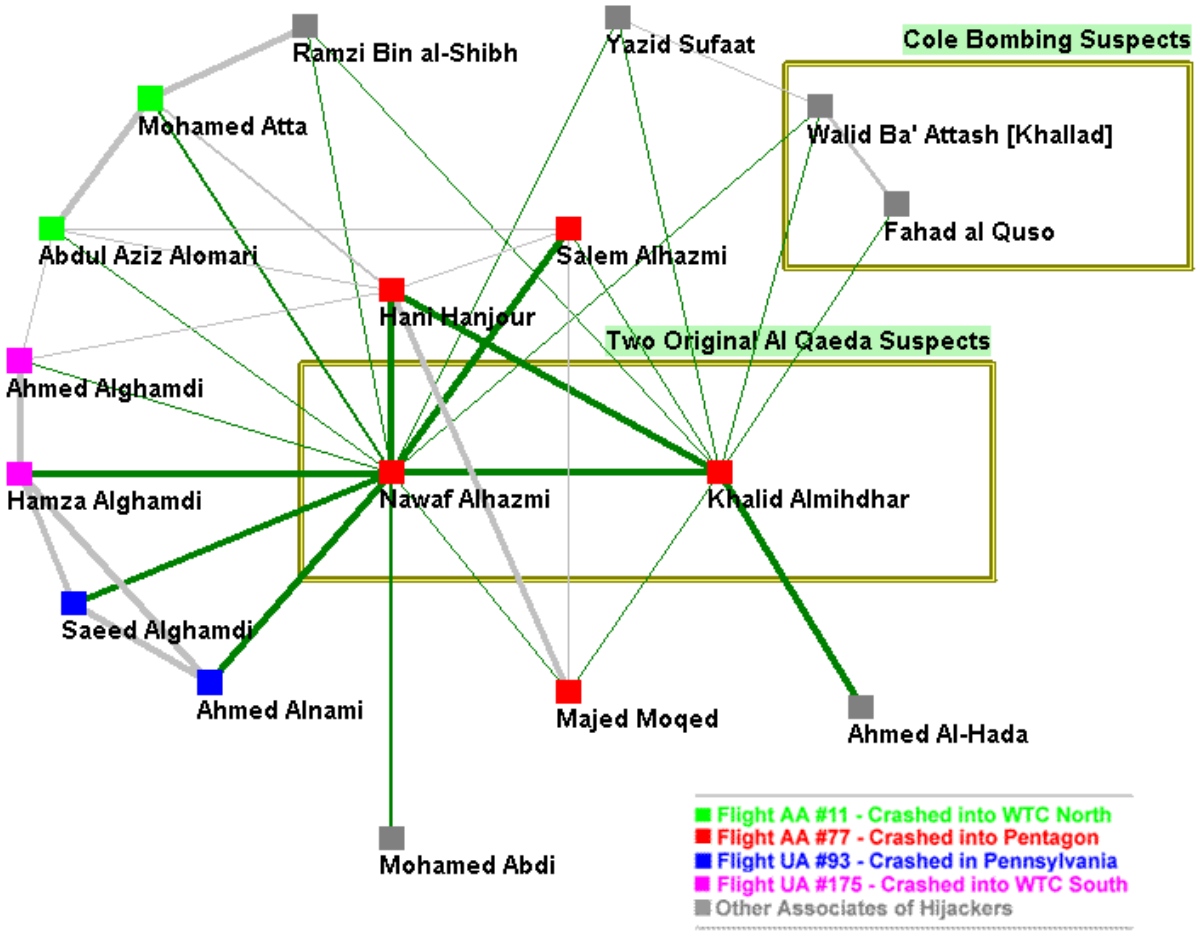


Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]

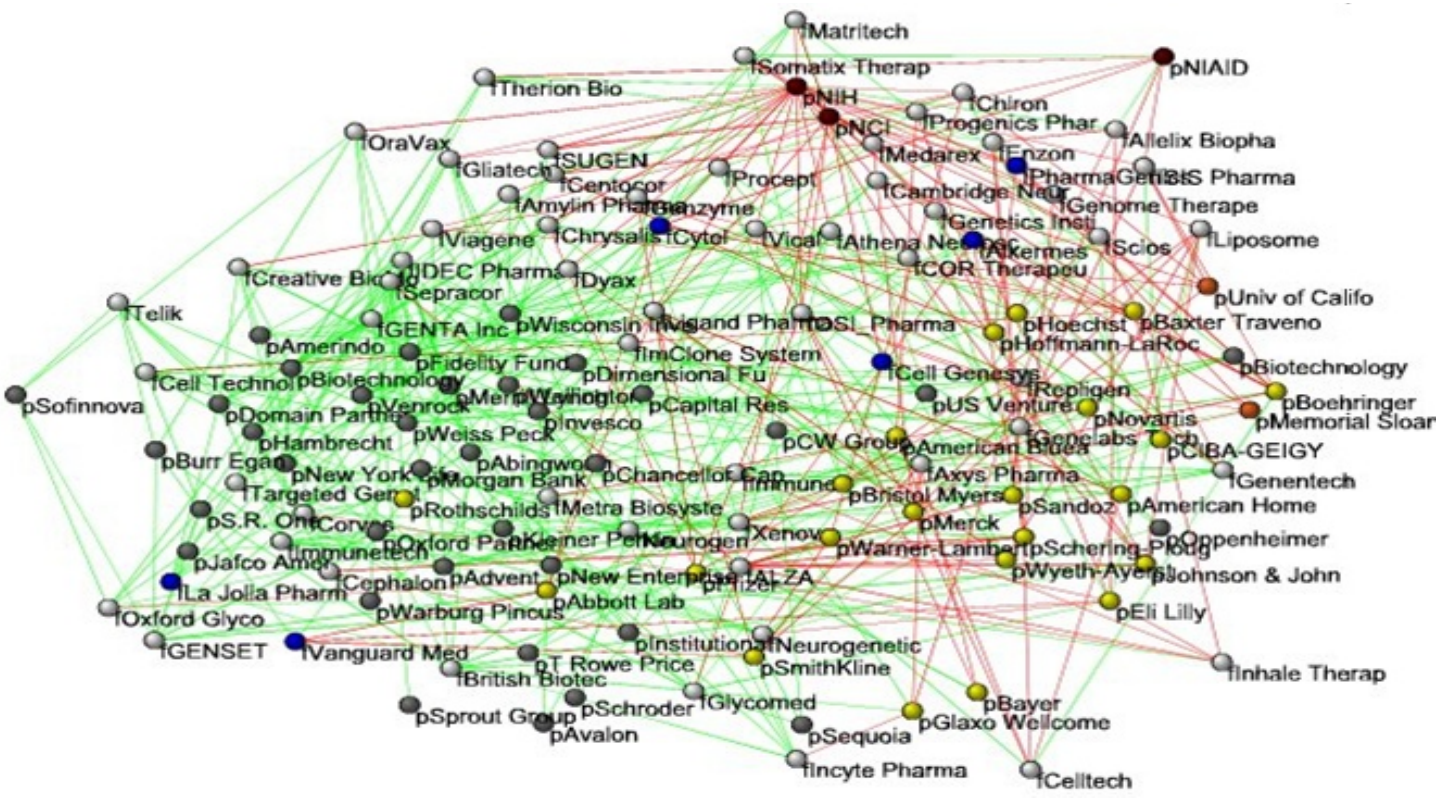
Networks: Organizations



9/11 terrorist network

[Krebs, 2002]

Networks: Economy



Nodes:

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

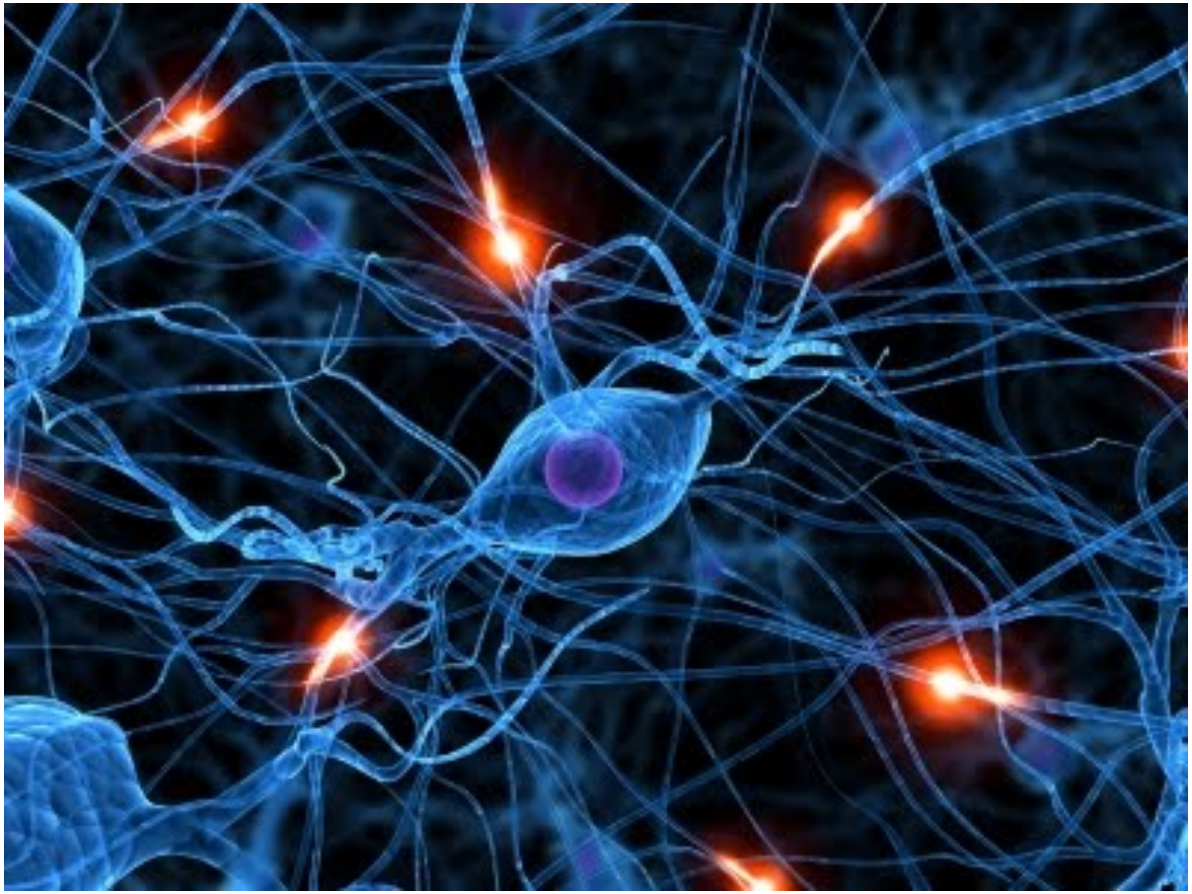
Links:

- Collaborations
- Financial
- R&D

Bio-tech companies

[Powell-White-Koput, 2002]

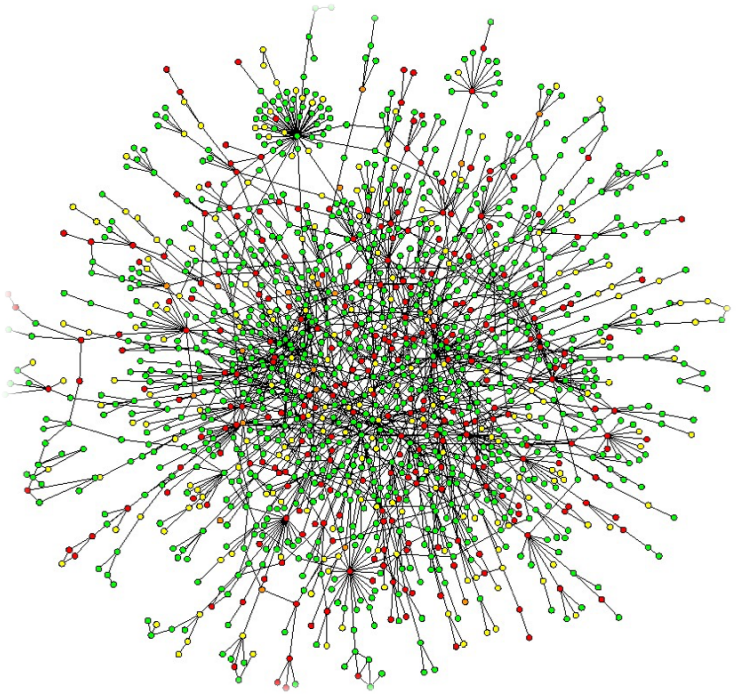
Networks: Brain

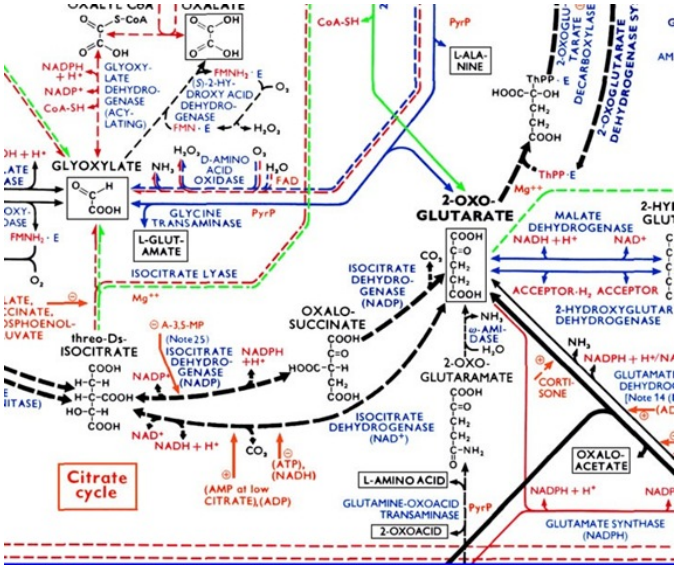


Human brain has between 10-100 billion neurons

[Sporns, 2011]

Networks: Biology





Protein-protein interaction (PPI) networks:

Nodes: Proteins

Edges: 'Physical' interactions

Metabolic networks:

Nodes: Metabolites and enzymes

Edges: Chemical reactions

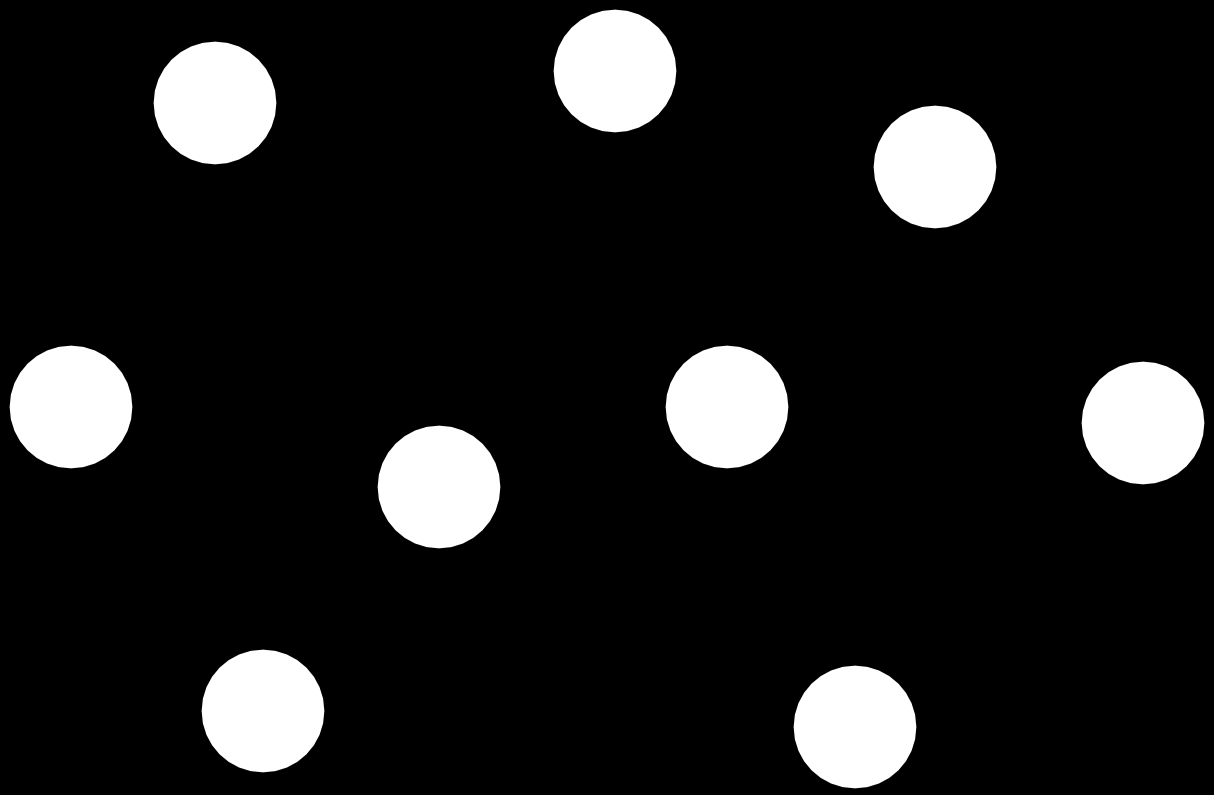
Networks!!

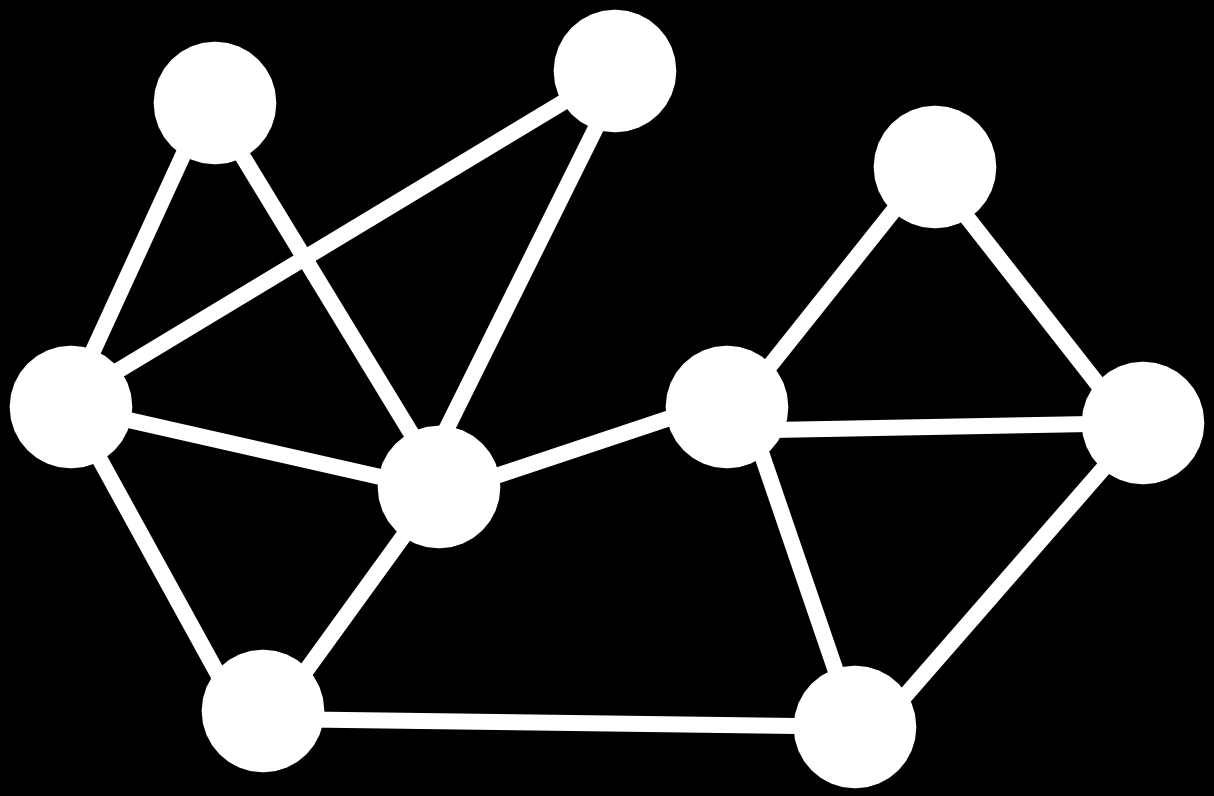
Behind many systems there is an intricate wiring diagram, **a network**, that defines the **interactions** between the components

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

Why Networks?

Networks are a general language for describing complex system





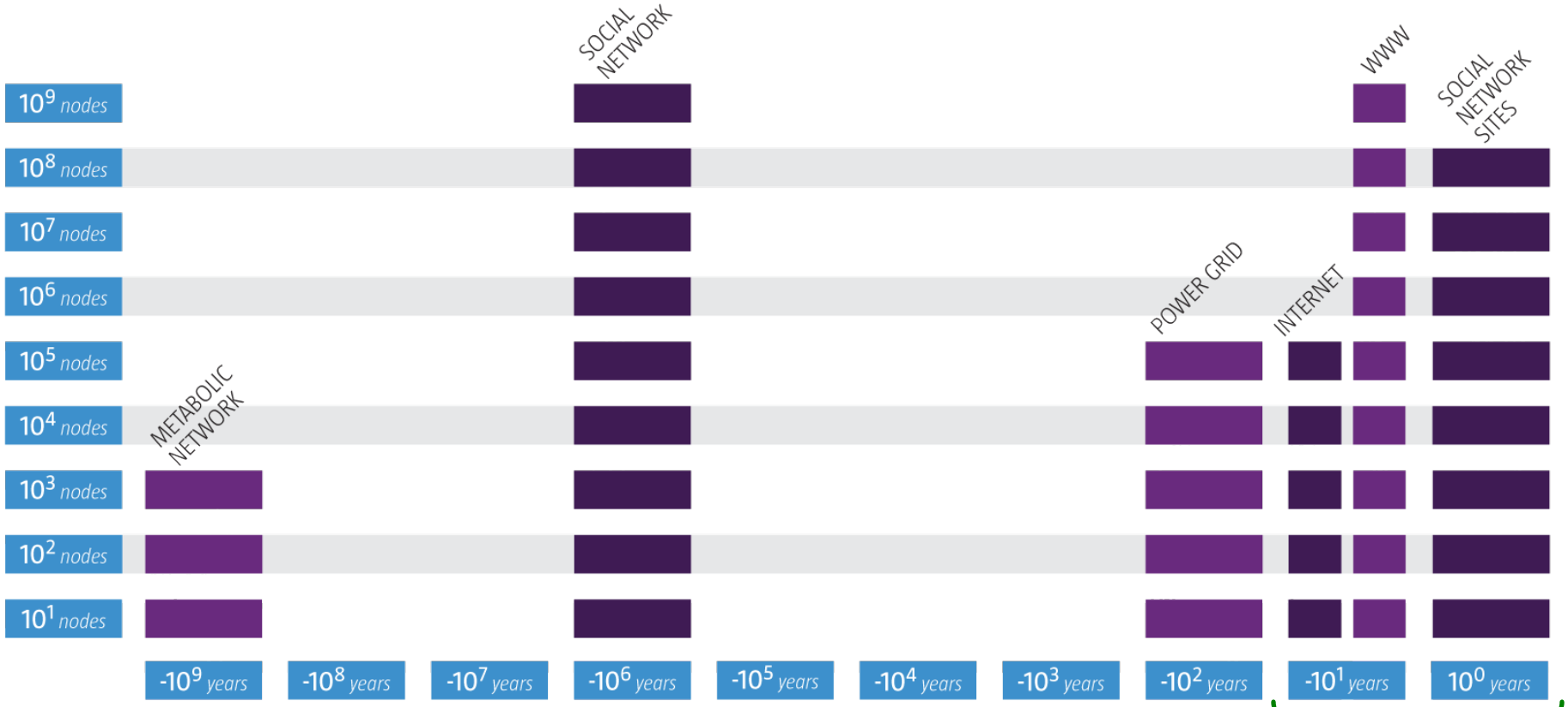
Network!

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

Networks: Size Matters

- **Network data: Orders of magnitude**
  - 436-node network of email exchanges 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 a "laboratory" for understanding the pulse of humanity.

Networks: Impact

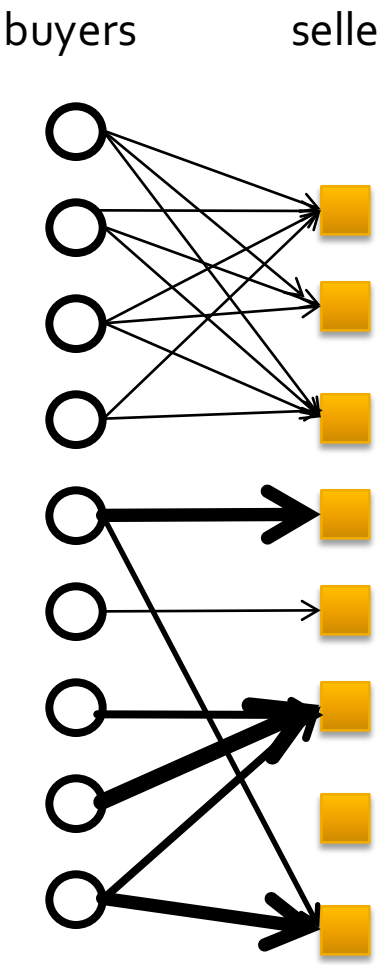
licensed under Attribution-NonCommercial-ShareAlike 2.0 Germany | Ludwig Gatzke | http://flickr.com/photos/stabilo-boss/

- **Google**
  Market cap:
  $541 billion
  (1y ago it was 394b)
- **Cisco**
  Market cap:
  $154 billion
  (1y ago it was 130b)
- **Facebook**
  Market cap:
  $359 billion
  (1y ago it was 201b)

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 Fraud



(thickness) purchase price

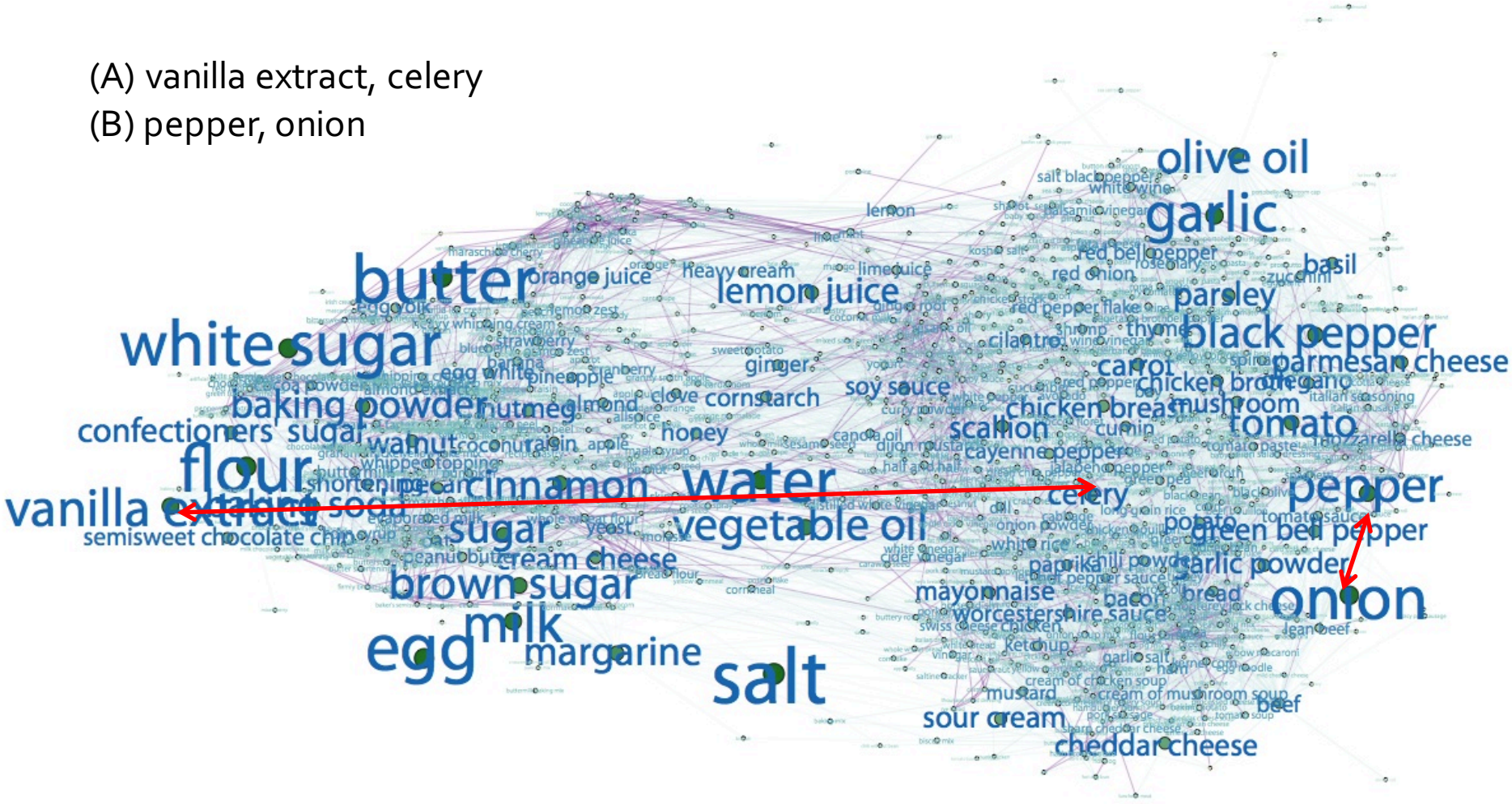
Which transactions are likely to be fraudulent?

Modeling Epidemics

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

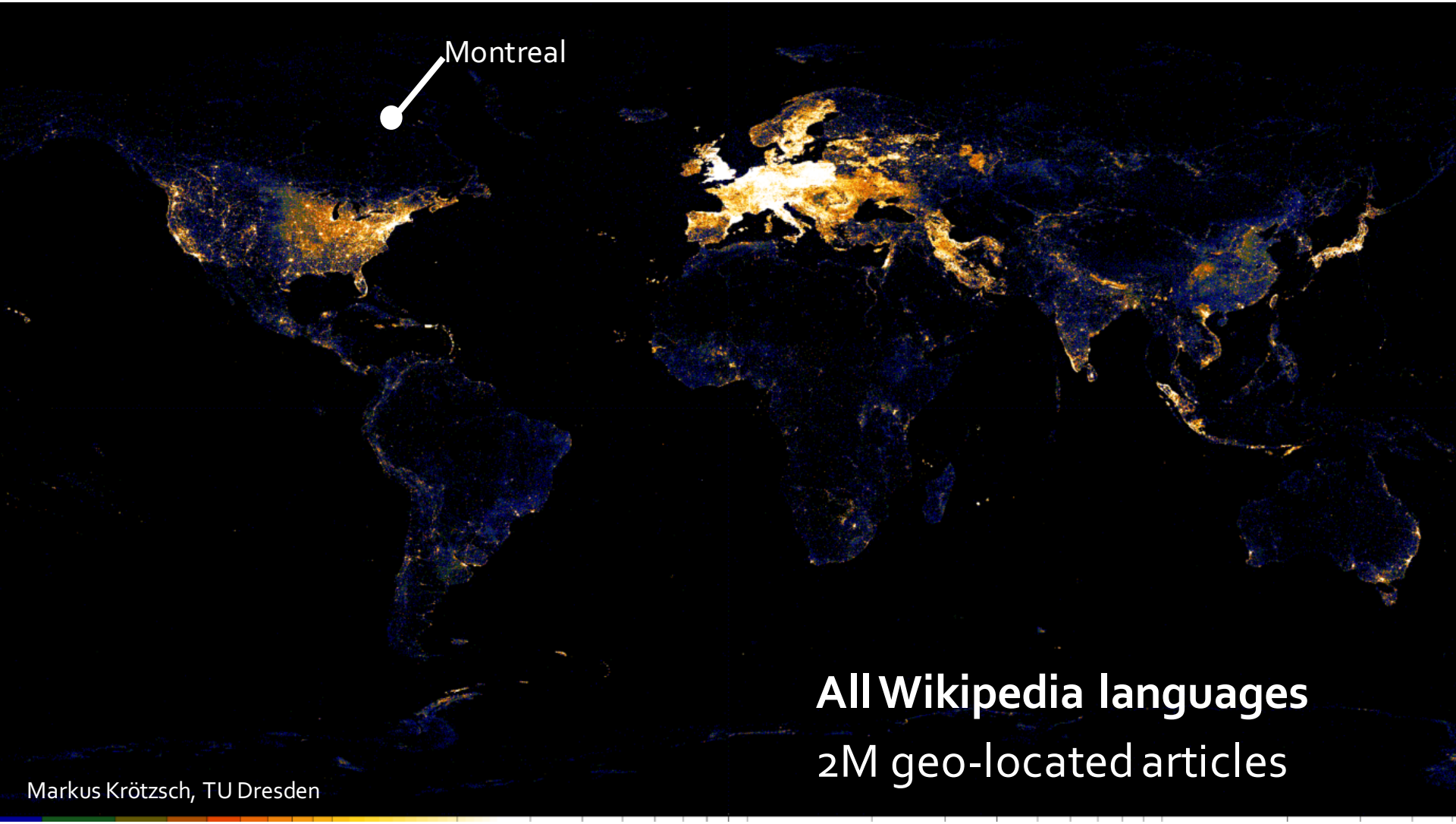
Which link is missing?

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

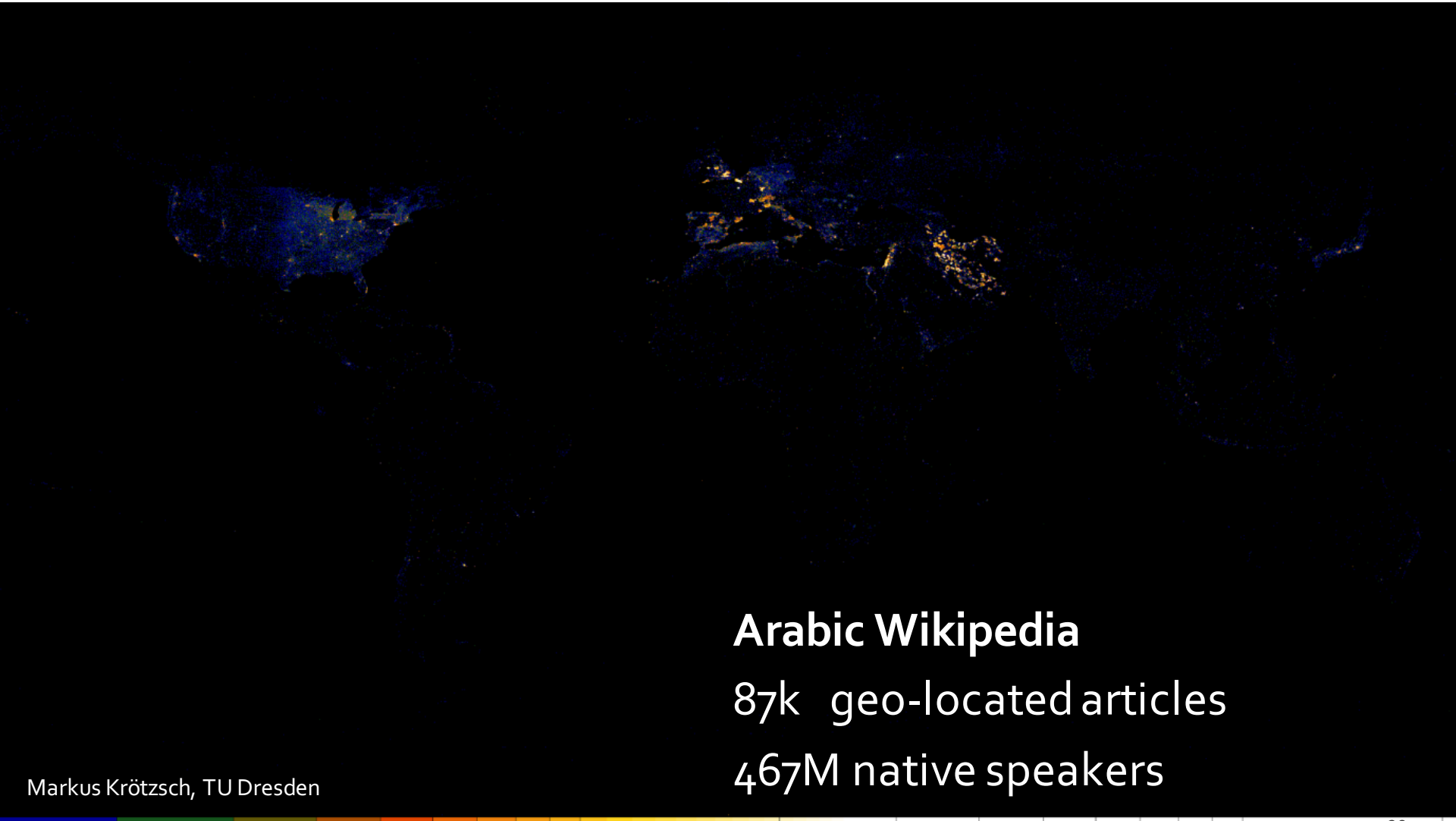


Recipe recommendation using ingredient networks. Teng et al., 2012.

Knowledge on Wikipedia



Knowledge on Wikipedia



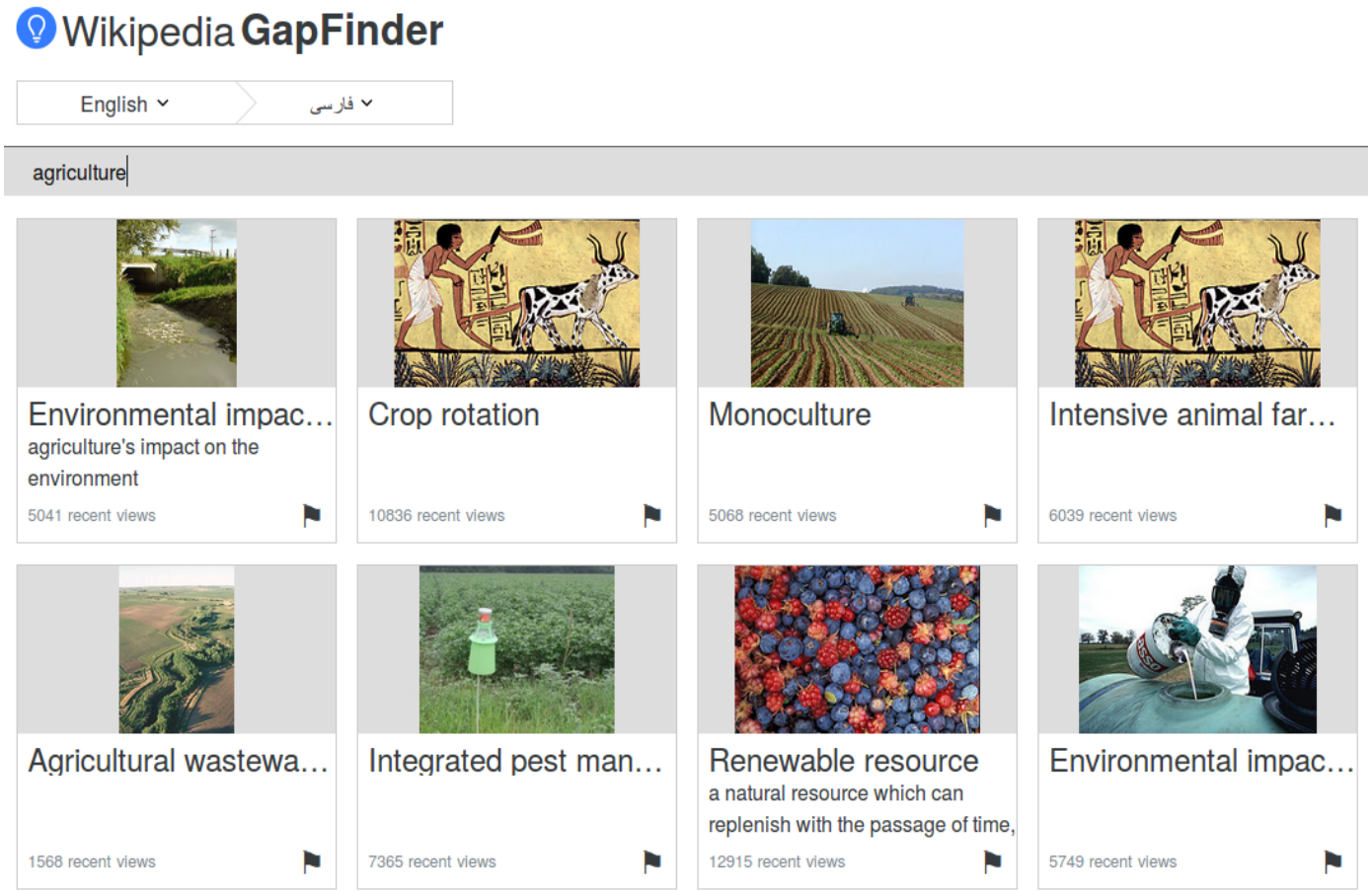
Arabic Wikipedia

87k geo-located articles

467M native speakers

Markus Krötzsch, TU Dresden

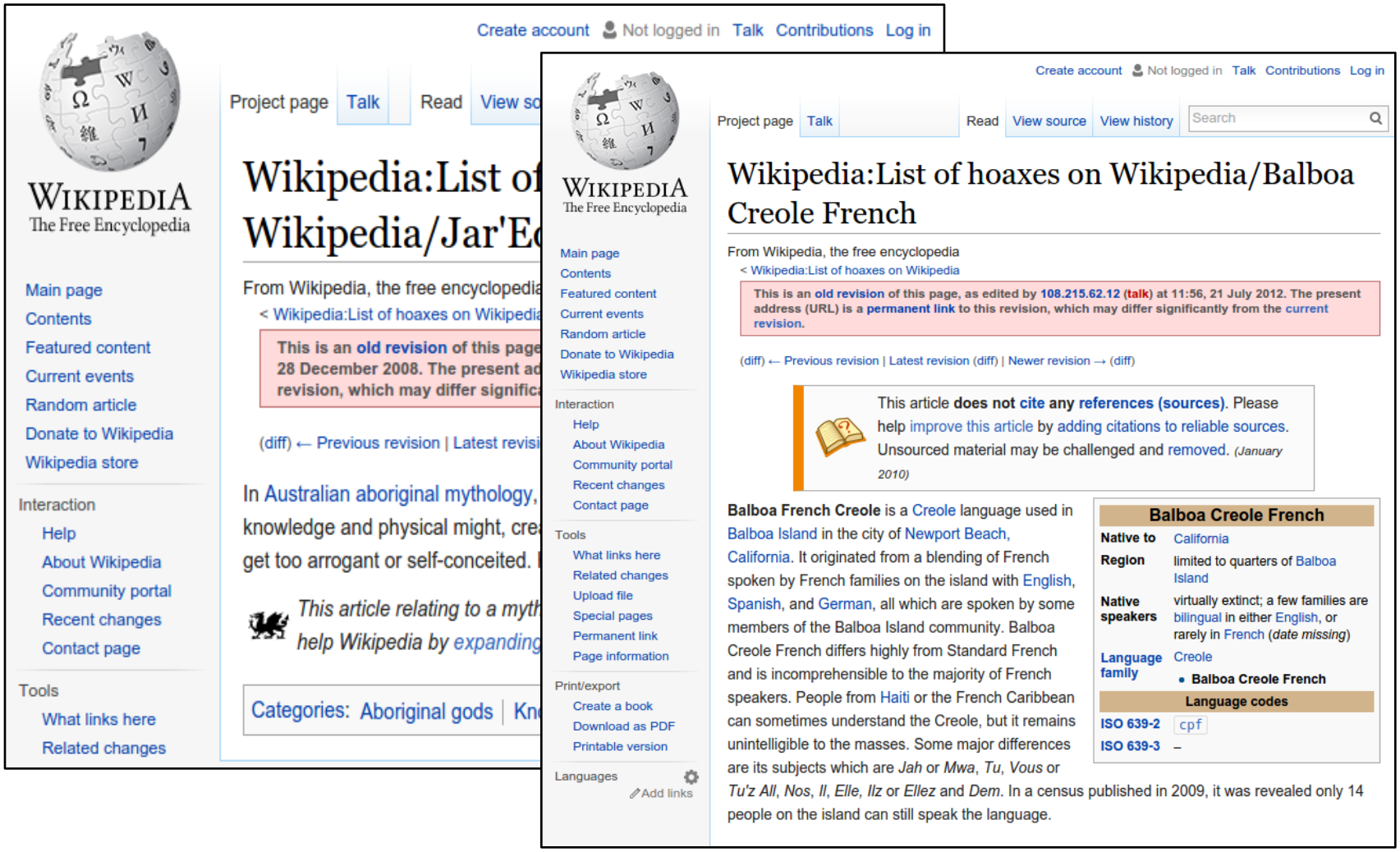
Recommending Missing Articles



Missing articles on agriculture in Farsi.

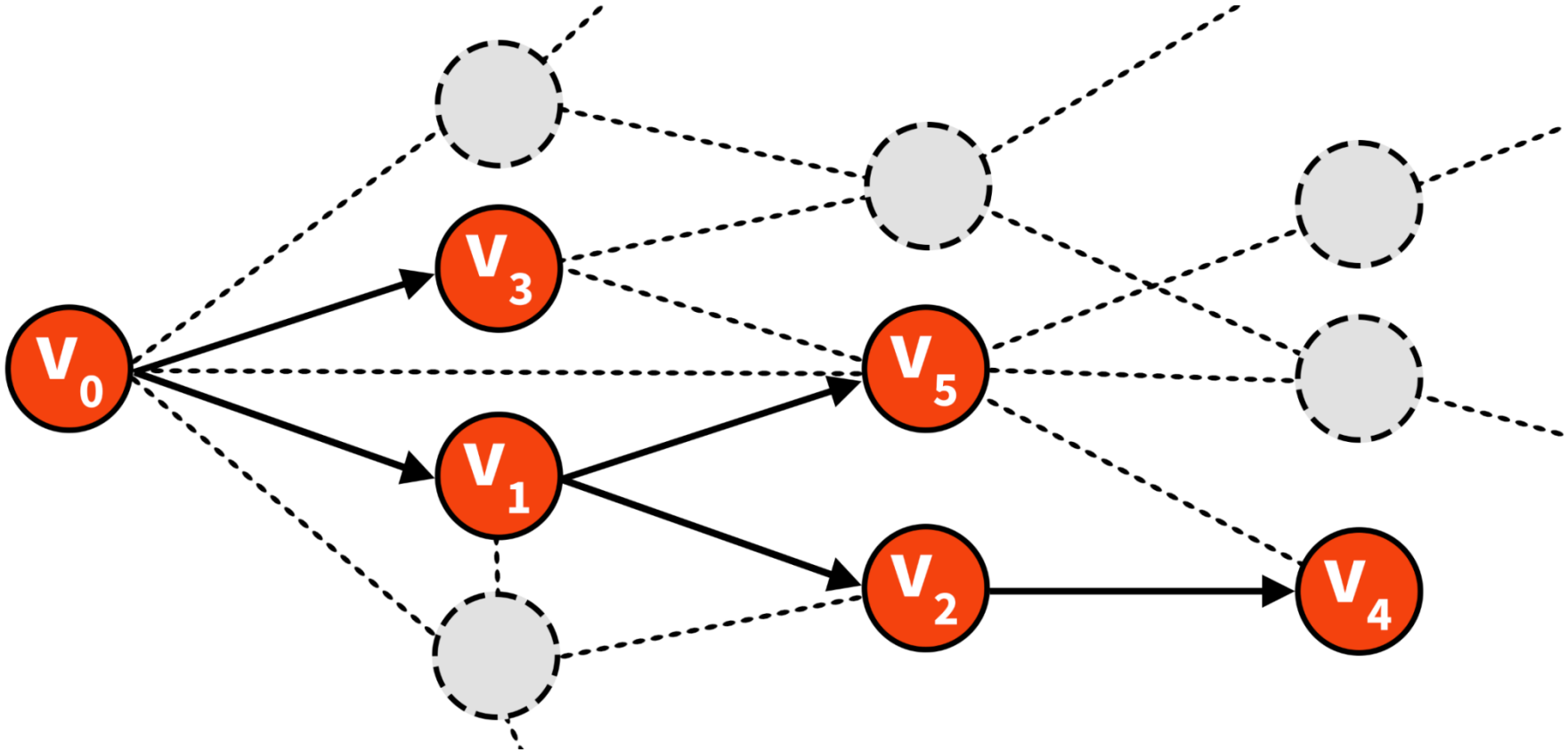
Growing Wikipedia Across Languages via Recommendation. Wulczyn et al. WWW '16

Can you detect a hoax?



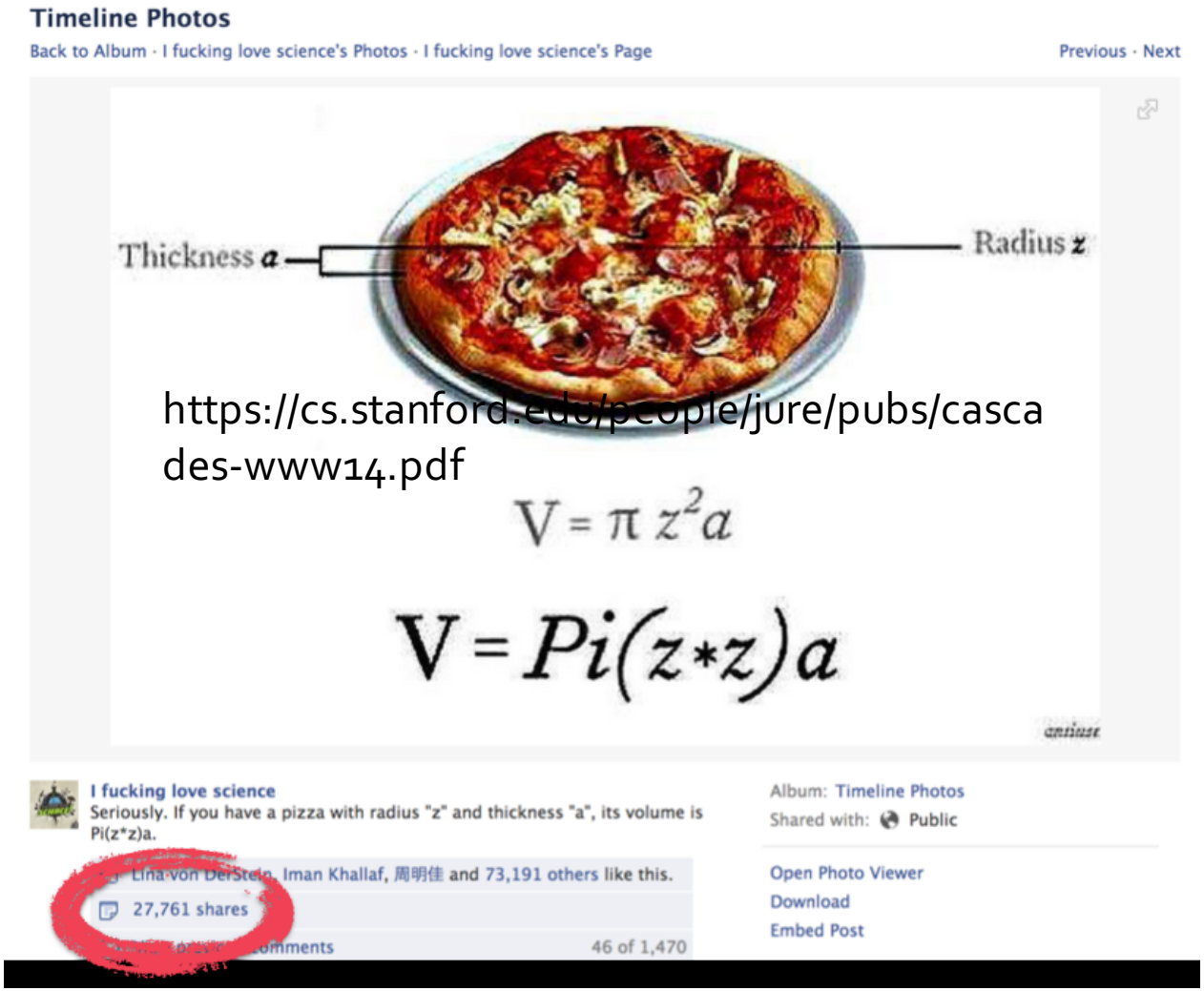
Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. Kumar et al. WWW '16.

How is information going to spread?



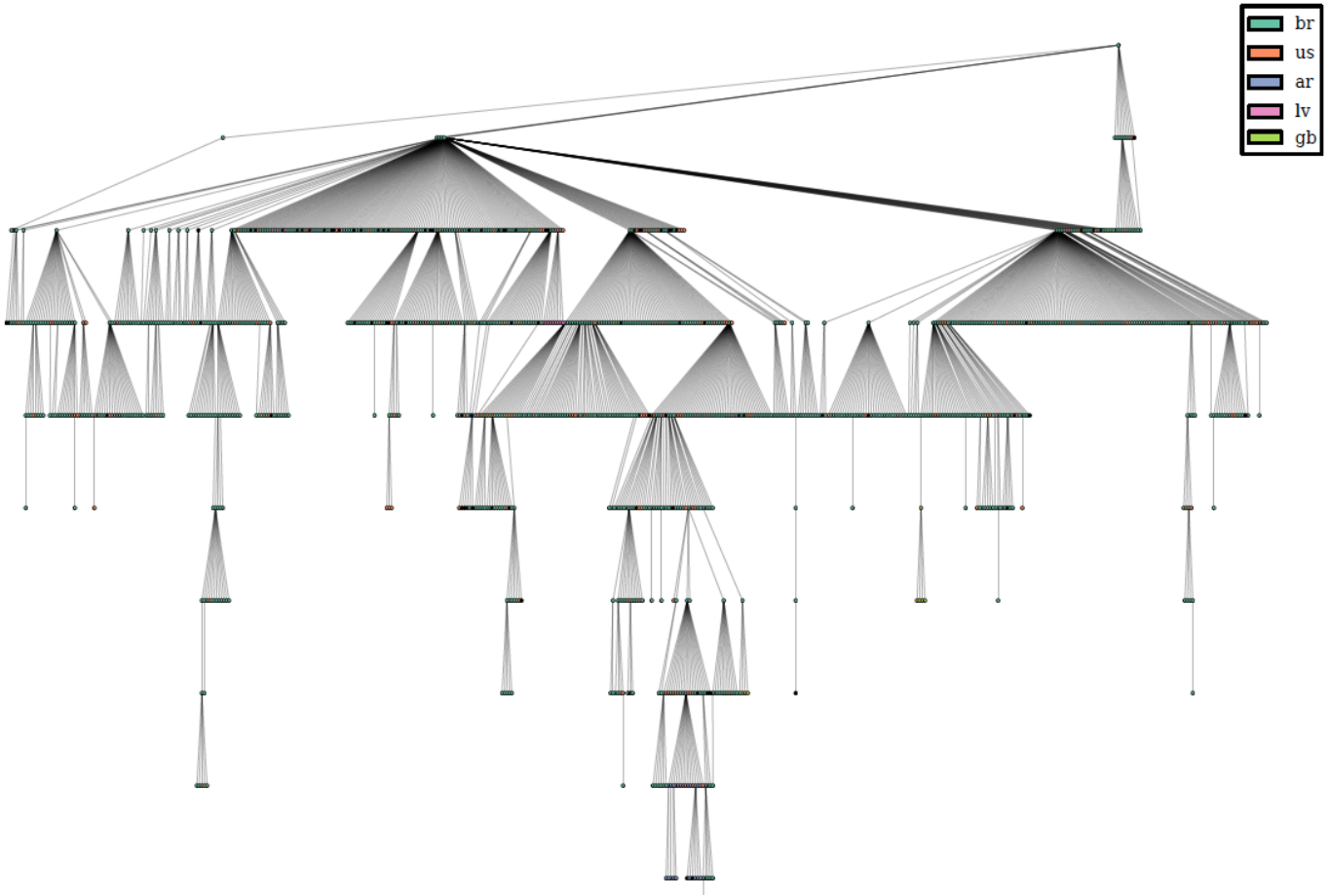
Information cascade in a social network

Facebook Information Cascades



Can cascades be predicted? Cheng et al., WWW '14.

LinkedIn Adoption Cascade



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.

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

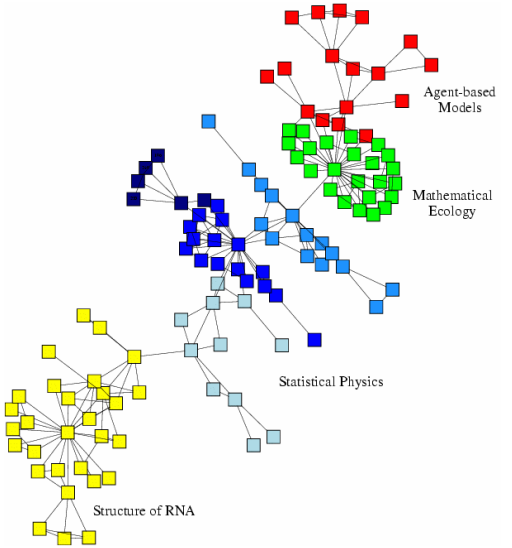
Reasoning about Networks

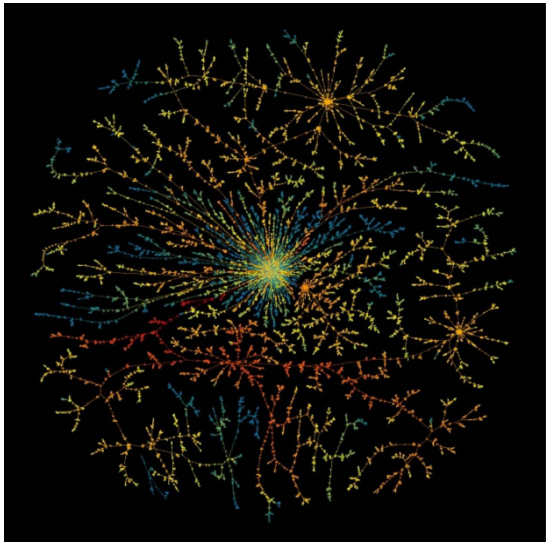
- **How do we reason about networks?**
  - **Empirical:** Study network data to find organizational principles
    - How do we measure and quantify networks?
  - **Mathematical models:** Graph theory and statistical models
    - Models allow us to understand behaviors and distinguish surprising from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

- Small diameter, Edge clustering
- Scale-free
- Strength of weak ties, Core-periphery
- Densification power law, Shrinking diameters
- Patterns of signed edge creation
- Information virality, Memetracking

Models

- Small-world model, Erdös-Renyi model
- Preferential attachment, Copying model
- Kronecker Graphs
- Microscopic model of evolving networks
- Structural balance, Theory of status
- Independent cascade model, Game theoretic model

Algorithms

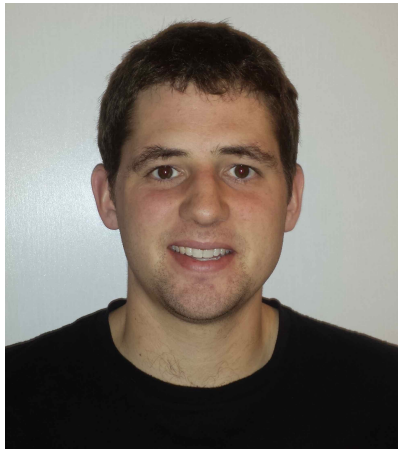
- Decentralized search
- PageRank, Hubs and authorities
- Community detection: Girvan-Newman, Modularity
- Link prediction, Supervised random walks
- Models for predicting edge signs
- Influence maximization, Outbreak detection, LIM

Logistics: Course Assistants

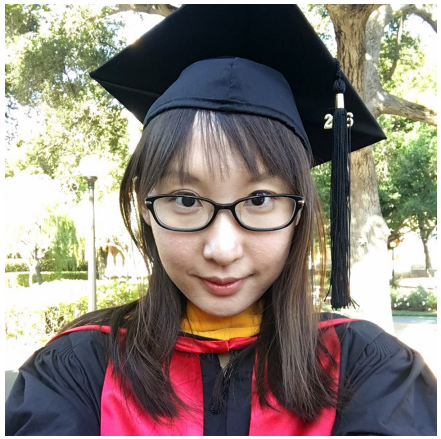
Will Hamilton
(head TA)

Jessica Su

Ben Ulmer



Austin Benson

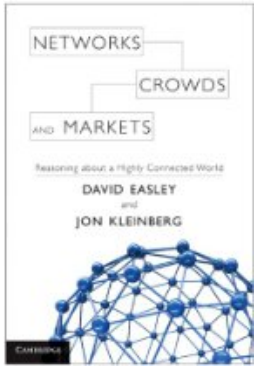


Zhedi Liu

Leon Yao

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/fall2016/cs224w
    - Use access code "snap"
  - **Please participate and help each other! (2 % of grade)**
- **For e-mailing course staff, always use:**
  - cs224w-aut1617-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 MJ2ZDM**
      - **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/
- **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**
  - **(all projects will be graded equally, regardless of 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 23:59 PST
2	**Homework 0**	October 6
3	**Homework 1**	October 13
4	**Project proposal**	October 20 (no late periods!)
5	**Homework 2**	October 27
6	**Work on the project**	
7	**Homework 3**	November 10
8	**Project milestone**	November 17 (no late periods!)
9	**Thanksgiving break**	
10	**Homework 4**	December 1
11	**Project report**	TBD (no late periods!)
	Poster session	**December 13 3:30-6:30pm**

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

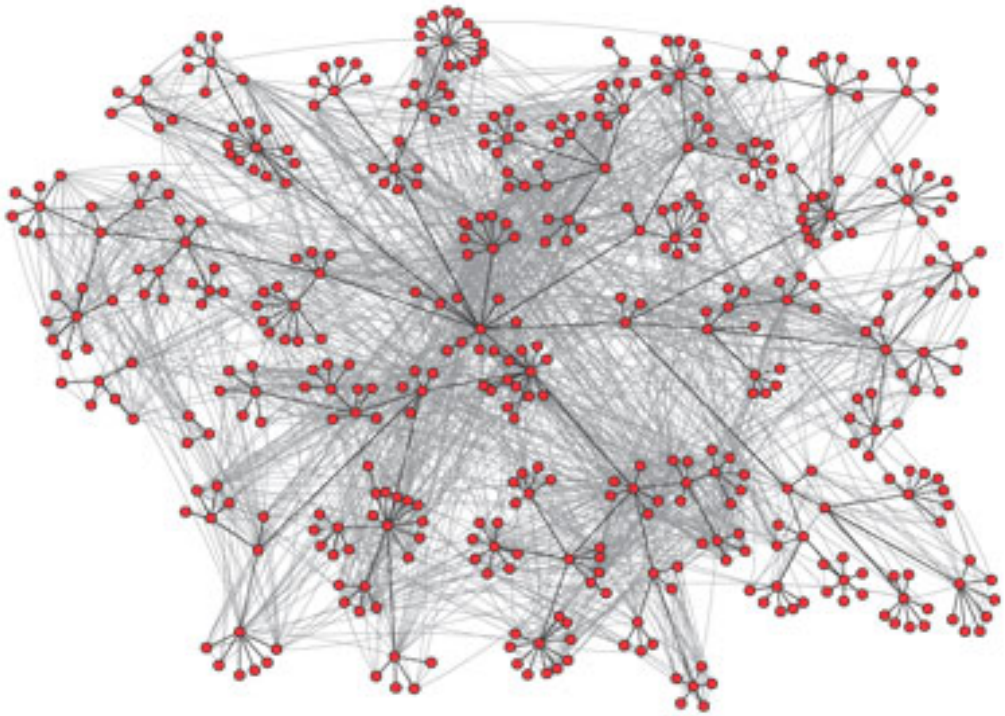
- **No single topic in the course is too hard by itself**
- **But we will cover and touch upon many topics and this is what makes the course hard**
  - **Good background in:**
    - Algorithms and graph theory
    - Probability and statistics
    - Linear algebra
  - **Programming:**
    - You should be able to write non-trivial programs (in Python)
  - **2 recitation sessions:**
    - SNAP.PY: Friday, 9/30 (12:30-1:20pm, Huang 18)
    - Review of Probability, Linear Algebra, and Proof Techniques: Mon, 10/3 (3:00-3:50pm, Gates B03)

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:
Structure of the Web Graph

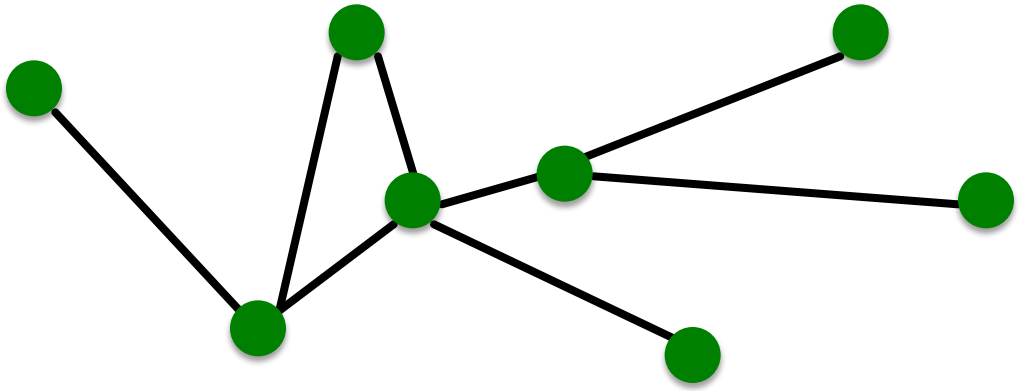
Structure of Networks?



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

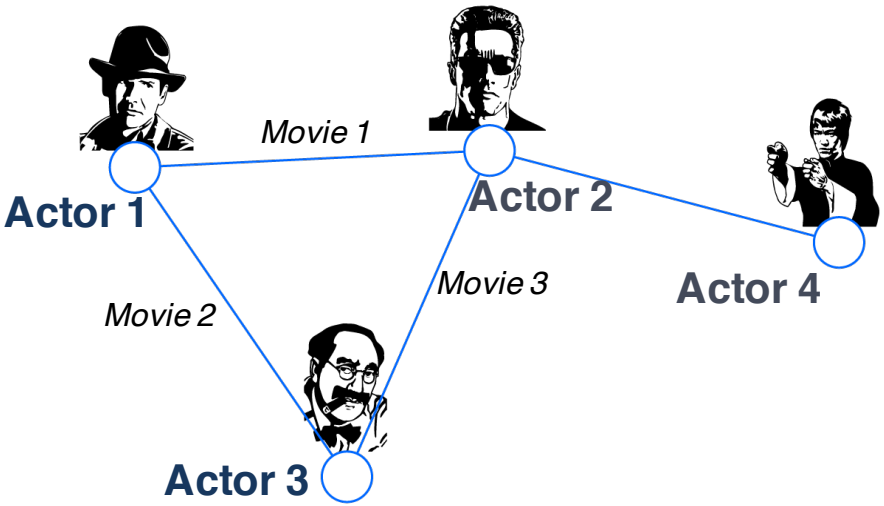
  Language: Network, node, link

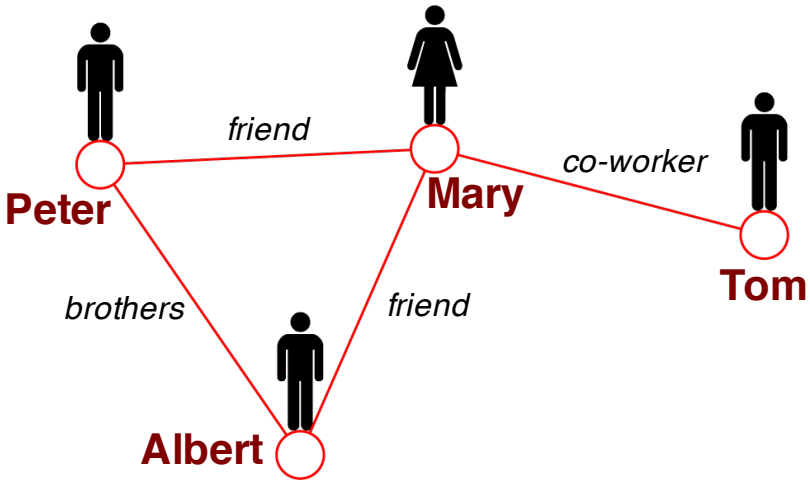
- **Graph** is a mathematical representation of a network
  - Web graph, Social graph (a Facebook term)

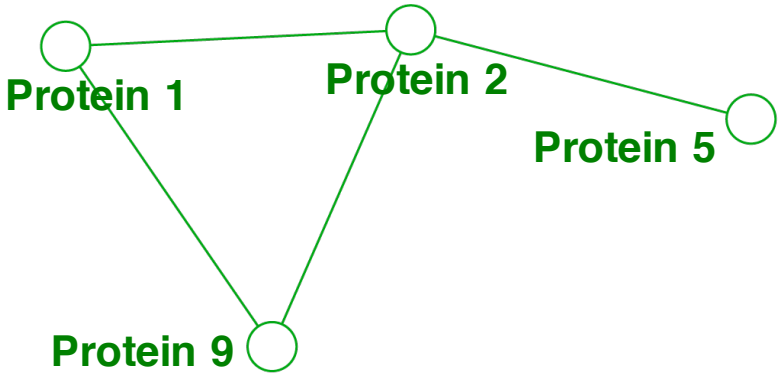
  Language: Graph, vertex, edge

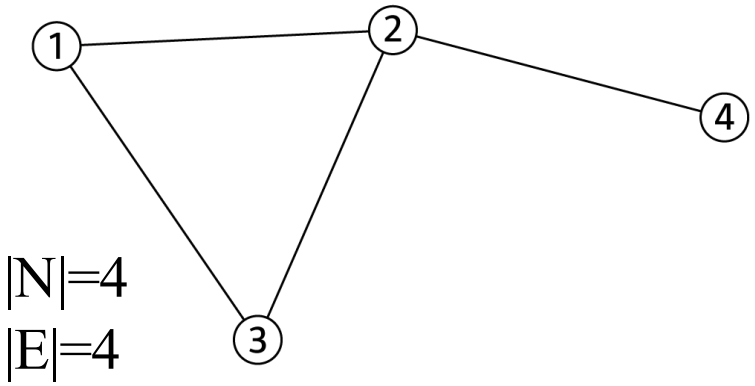
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably

Networks: Common Language



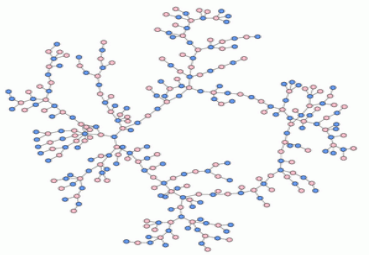


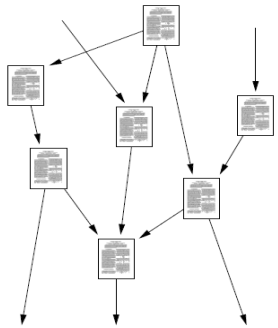




Choosing Proper Representations

- 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