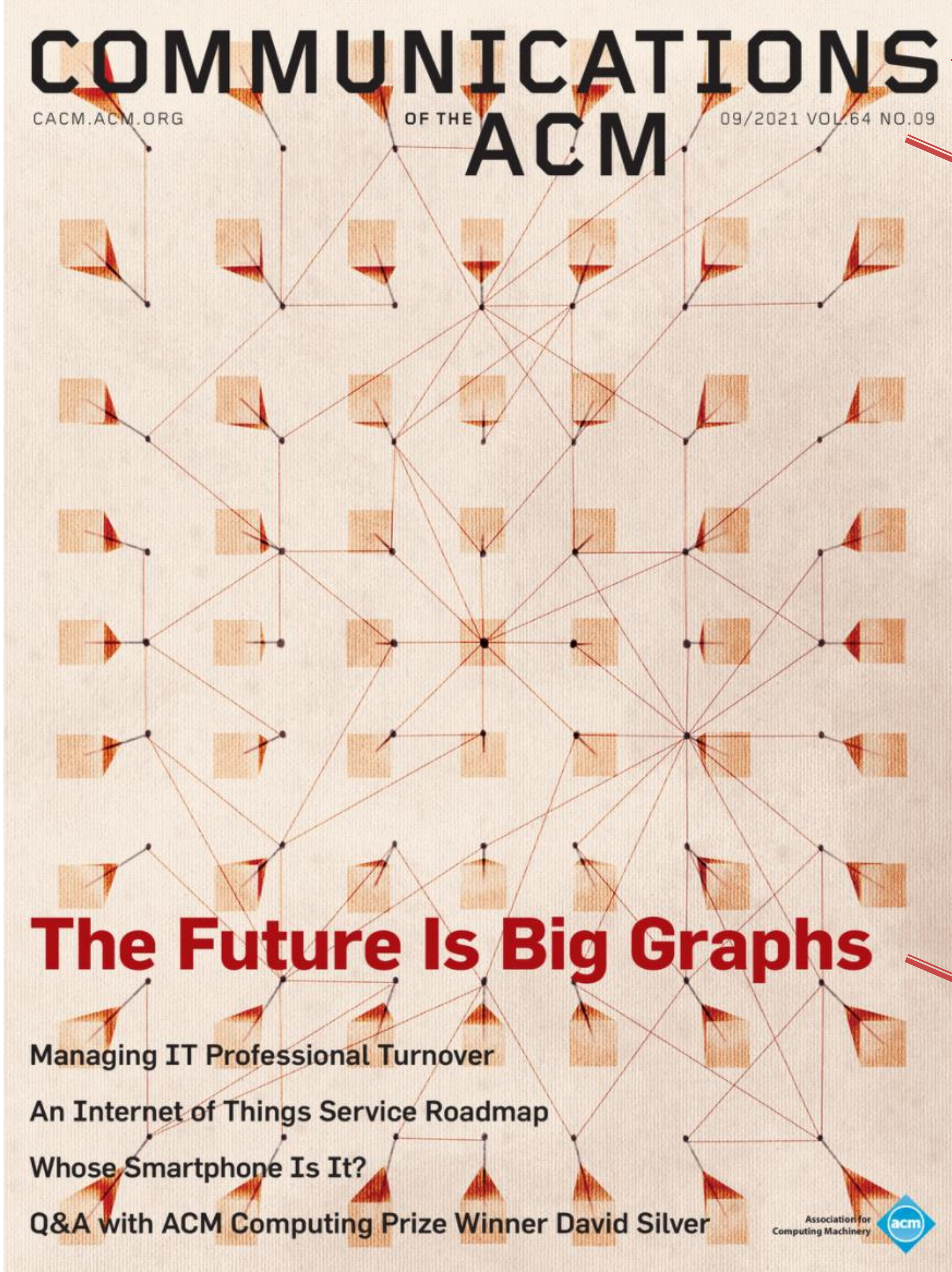
EECS4414/5414: Information Networks

Thanks to Jure Leskovec, Stanford and Panayiotis Tsaparas, Univ. of Ioannina for material in the slides


COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

09/2021 VOL.64 NO.09

The Future Is Big Graphs

Managing IT Professional Turnover

An Internet of Things Service Roadmap

Whose Smartphone Is It?

Q&A with ACM Computing Prize Winner David Silver

Association for Computing Machinery

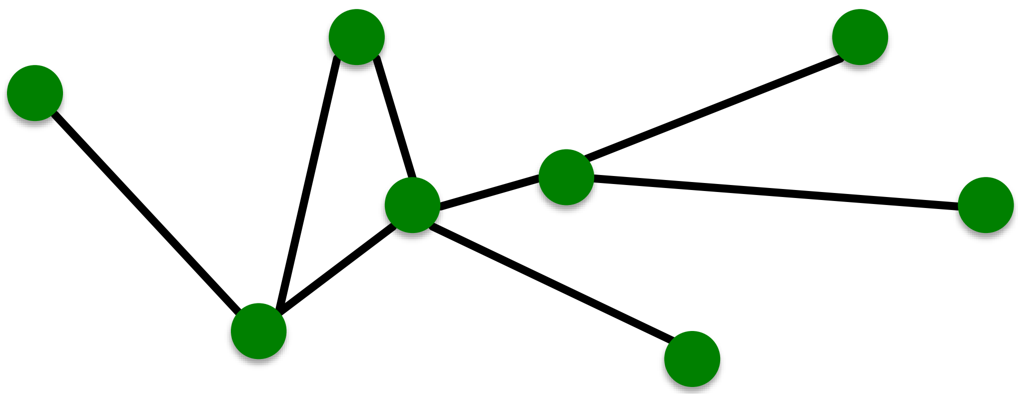
?

Issue of Sep 2021

Graphs

what is a network or a graph?

Network Components

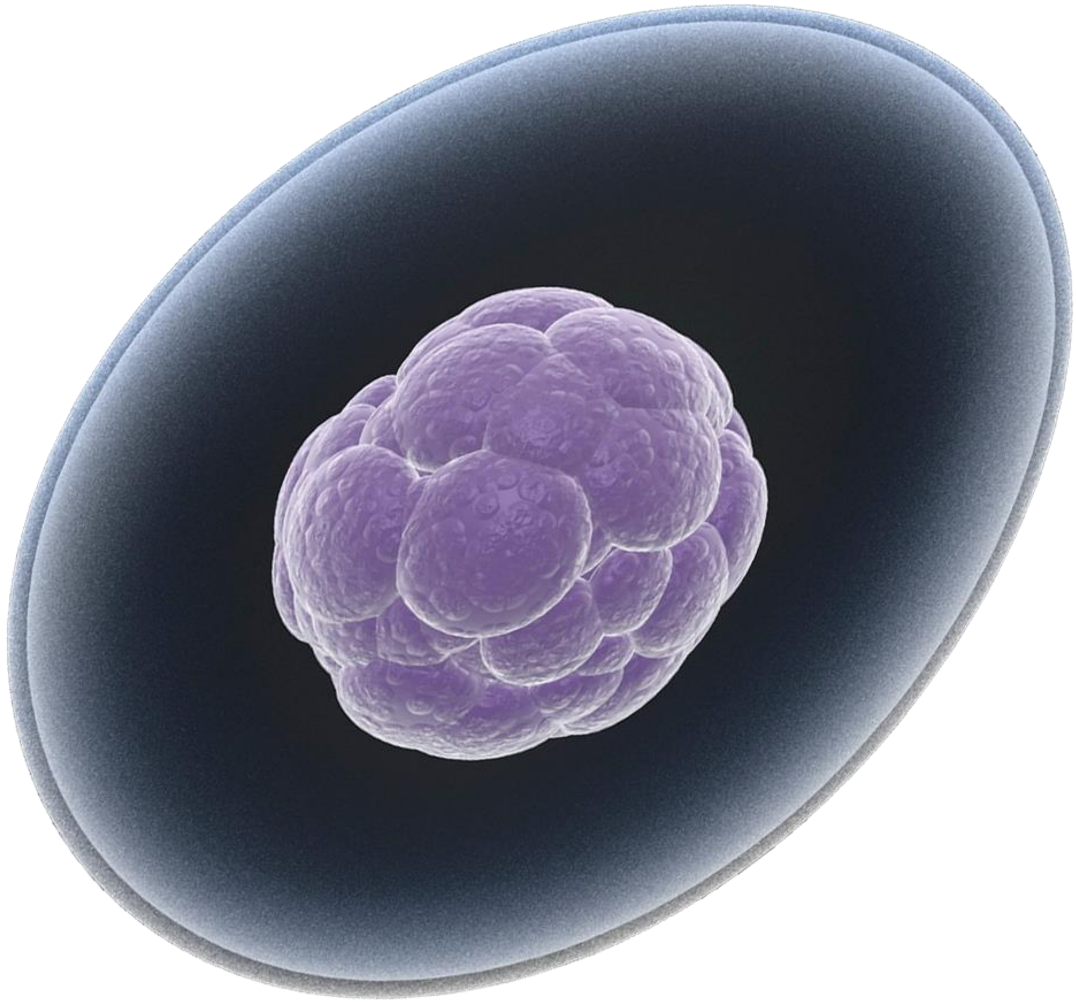


- **Network** (or **Graph**) $G(N,E)$
  - **Objects:** nodes (vertices) N
  - **Relationships:** links (edges) E

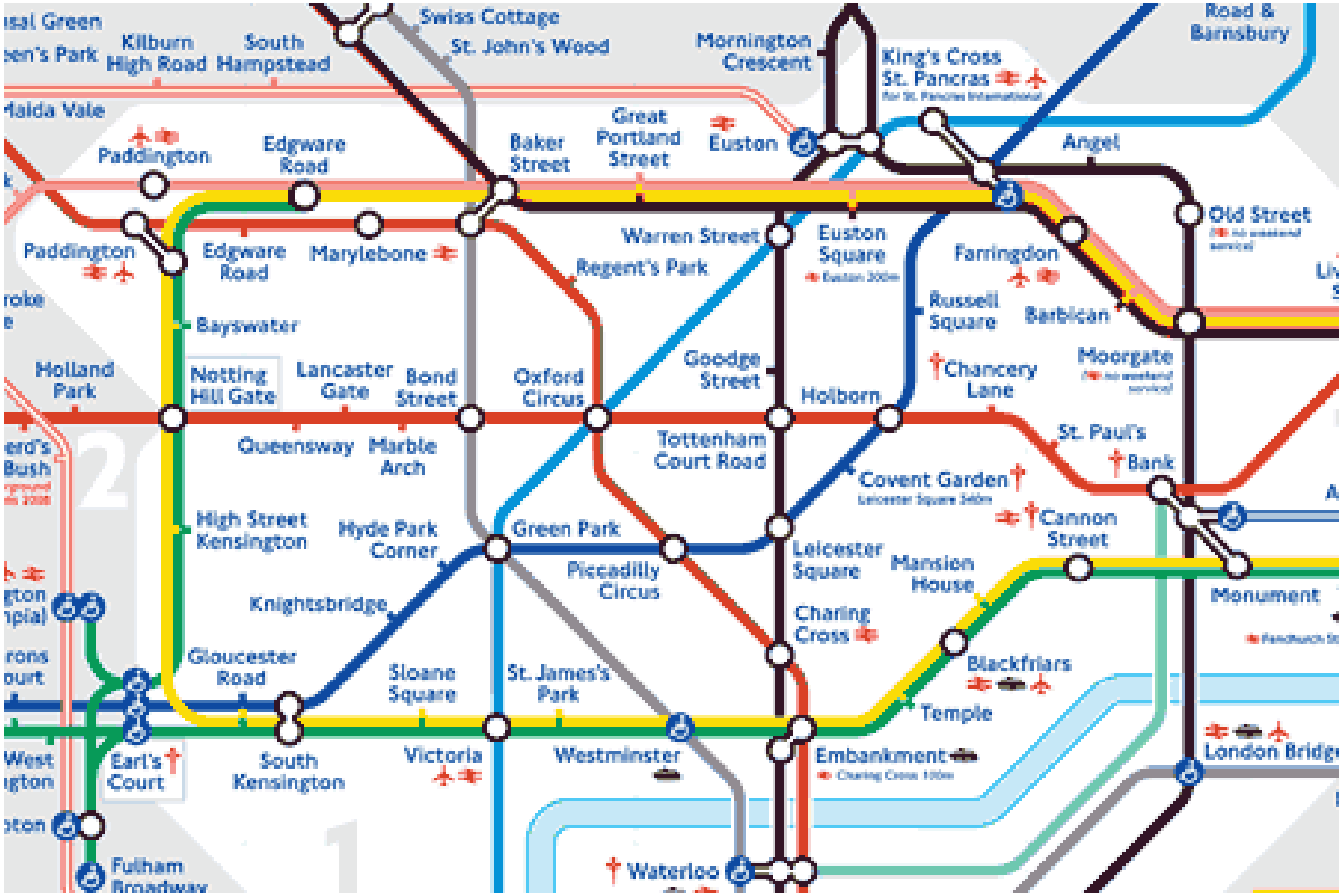
Built on the mathematics of **graph theory**

networks are ubiquitous

World economy



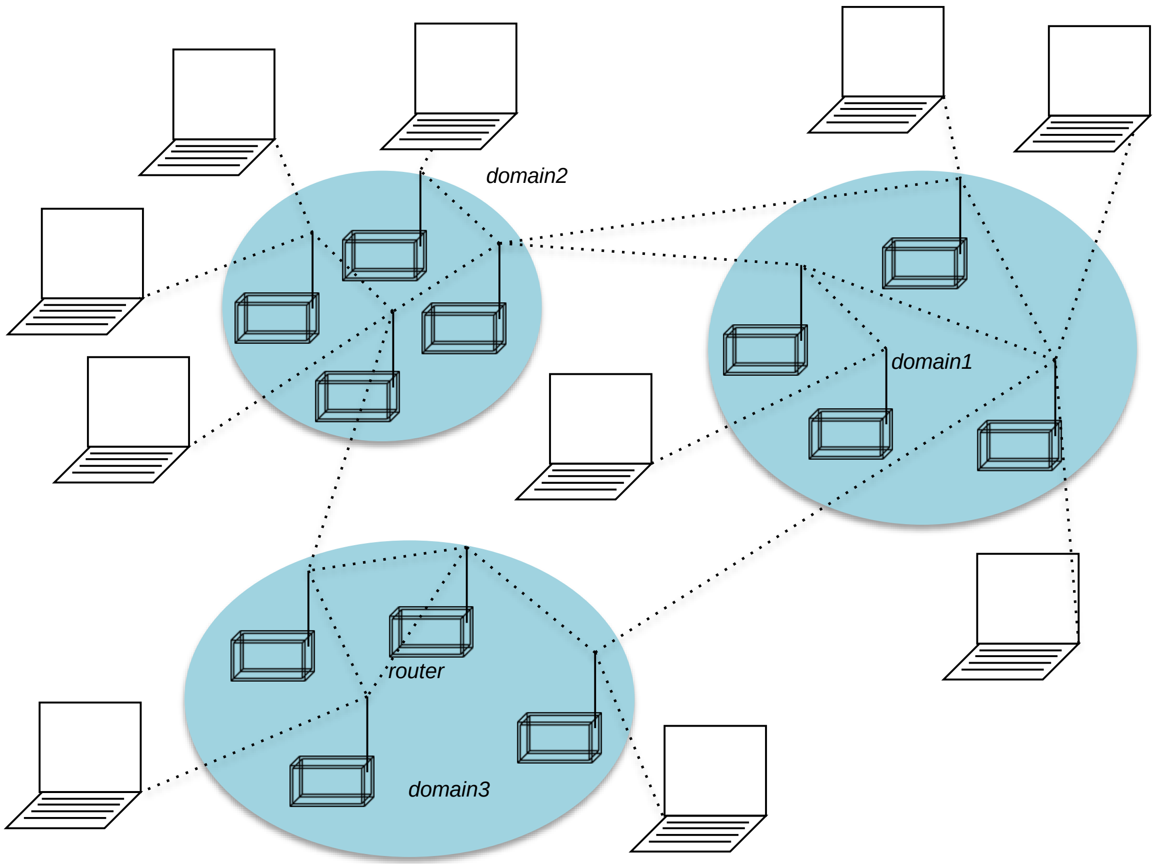
Human cell



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
Paddington
Edgware Road
Marylebone
Warren Street
Euston Square
Farringdon
Old Street
Regent's Park
Bayswater
Russell Square
Barbican
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
High Street Kensington
Hyde Park Corner
Green Park
Leicester Square
Mansion House
Cannon Street
Piccadilly Circus
Knightsbridge
Charing Cross
Monument
Gloucester Road
Sloane Square
St. James's Park
Blackfriars
Temple
Earl's Court
South Kensington
Victoria
Westminster
Embankment
London Bridge
Waterloo
Fulham Broadway

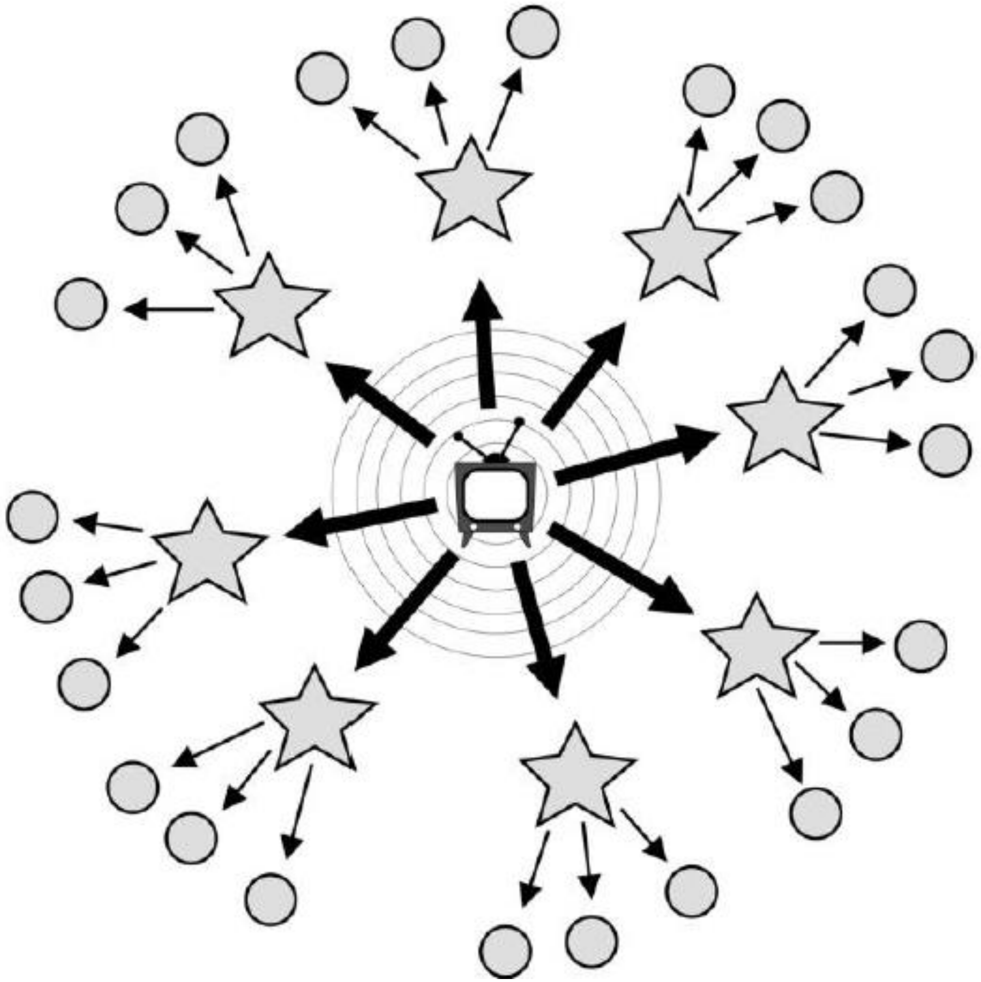
Railroads

Brain

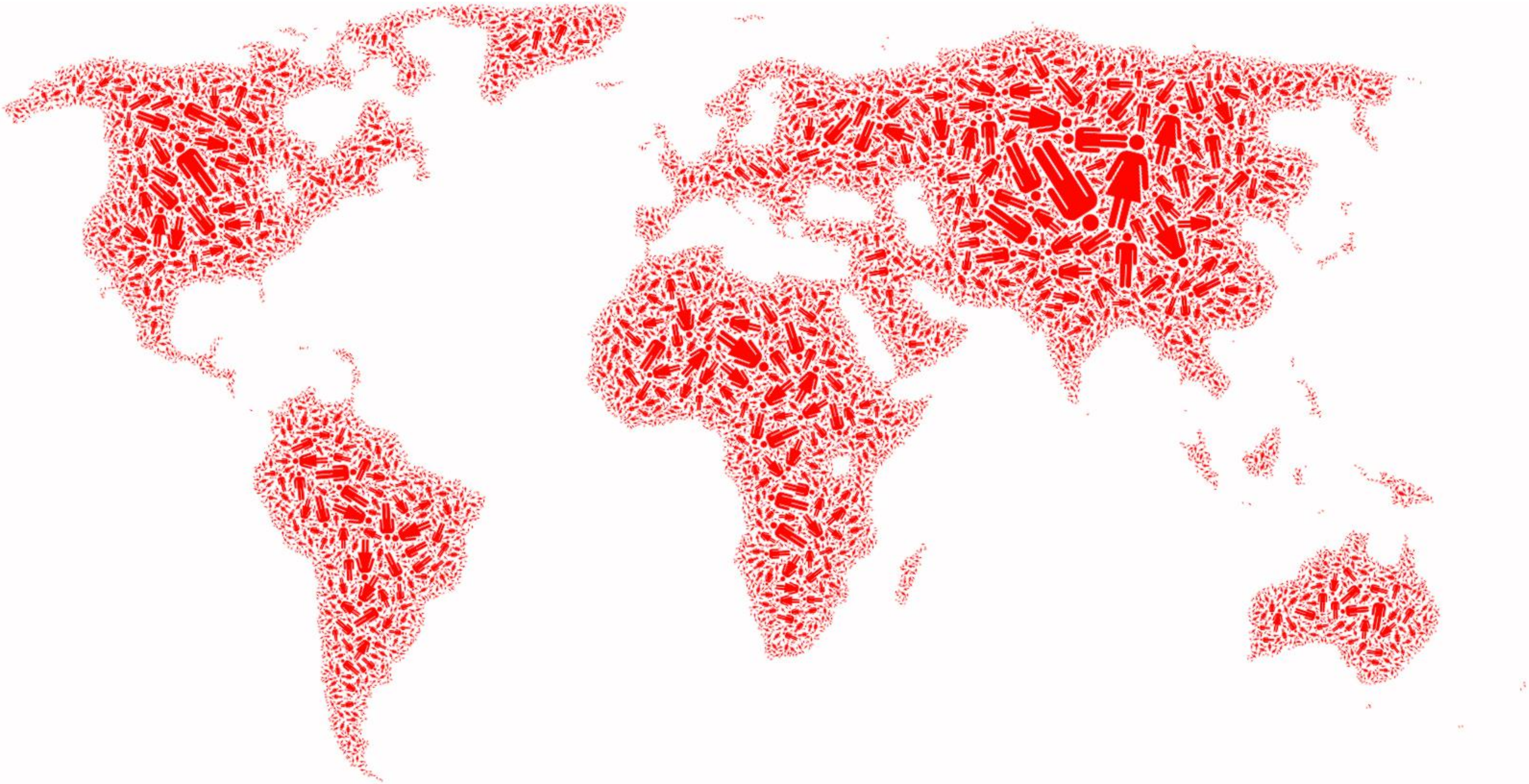


Internet

Friends & Family

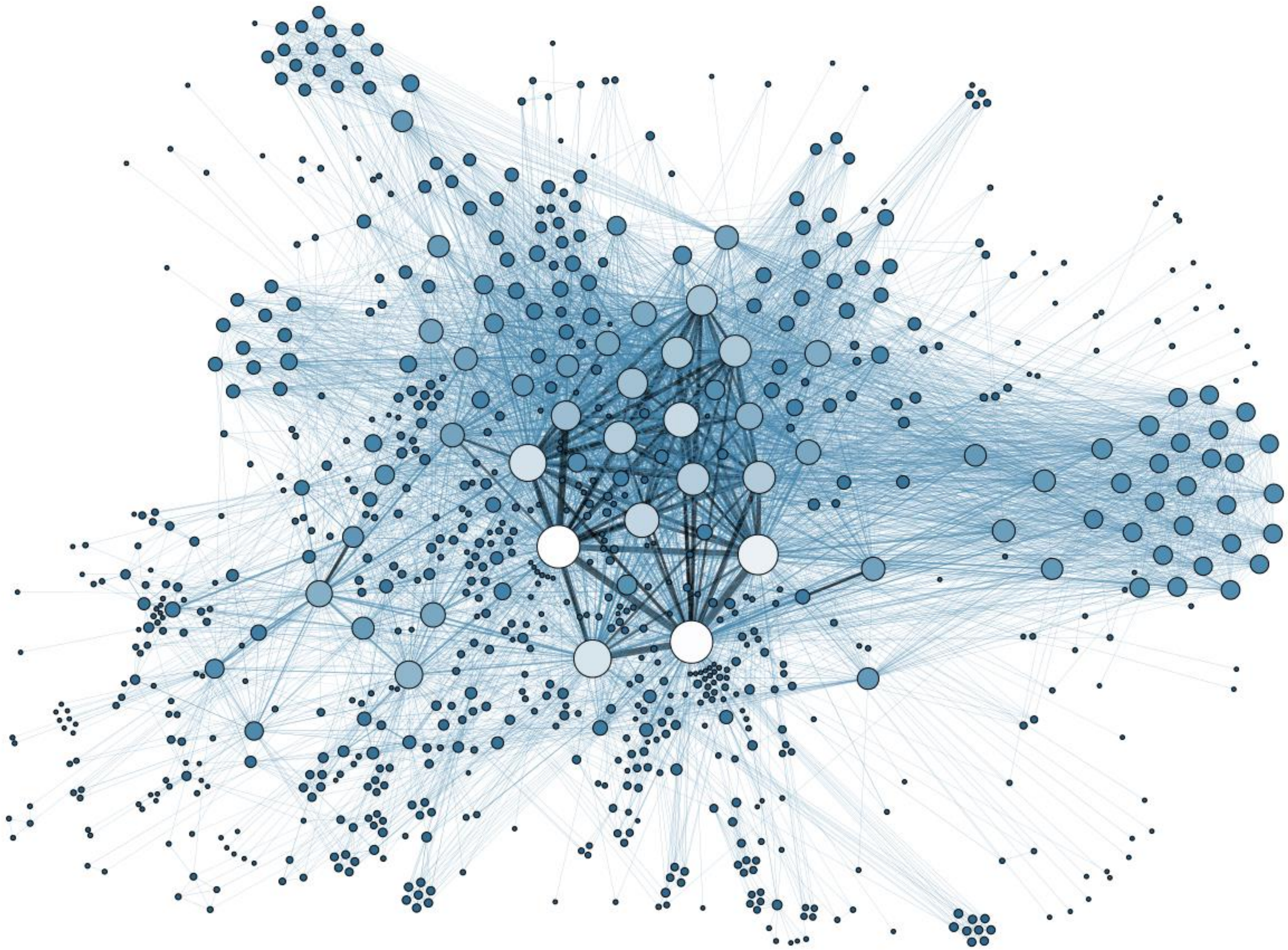


Media & Information



Society

What do the following things have in common?



Complex systems that can be modeled as Networks!

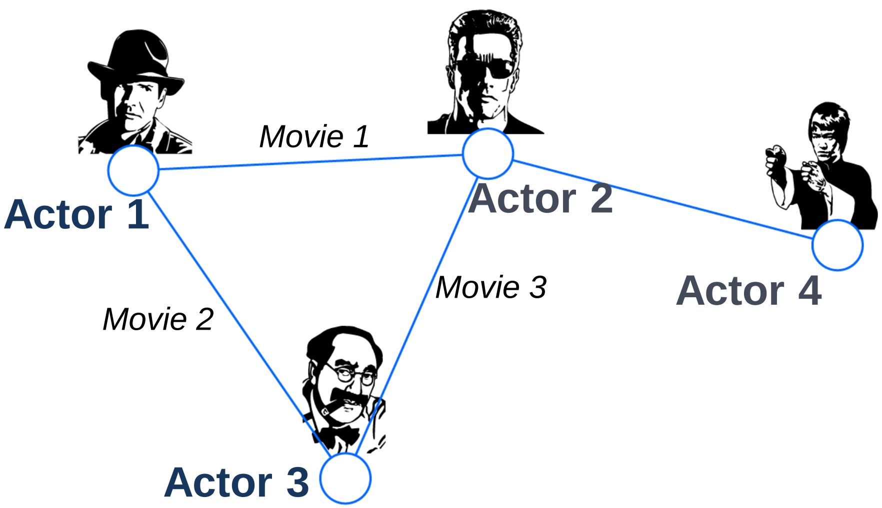
Networks!

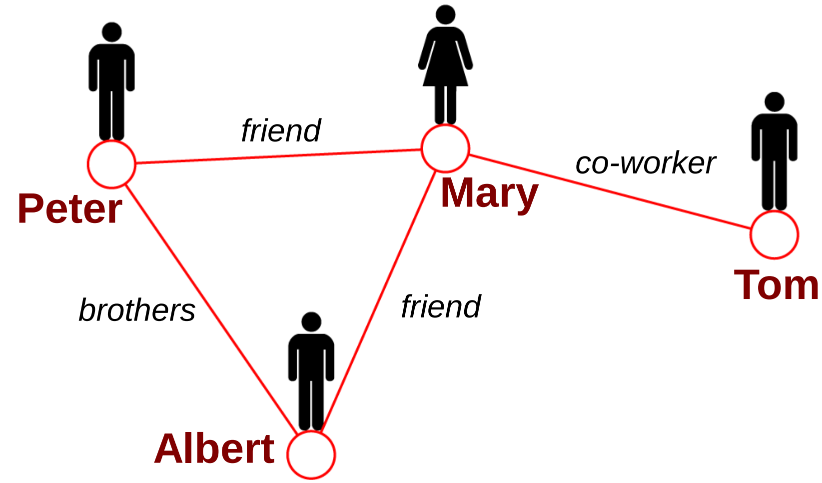
Behind many systems there is an intricate wiring diagram, **a network**, that defines the **interactions** between the components

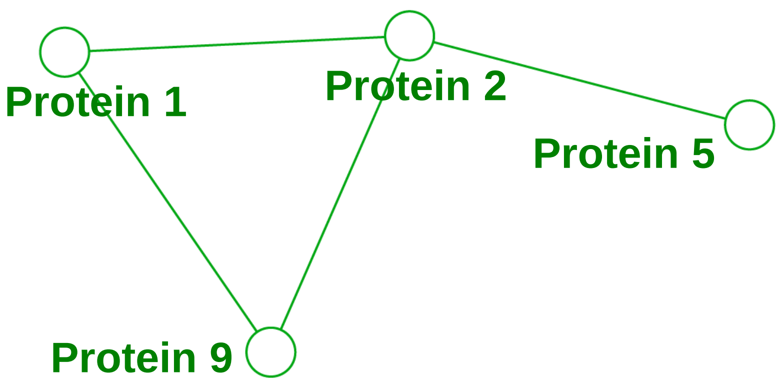
We will never understand these systems unless we understand the networks behind them!

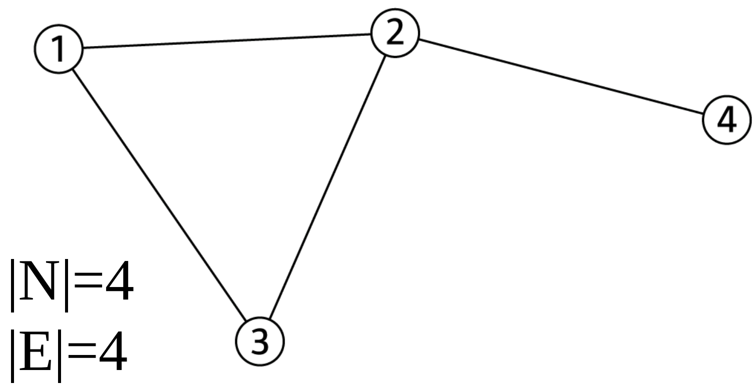
But, why should <u>we</u> care about networks? Why now?

Why Networks?



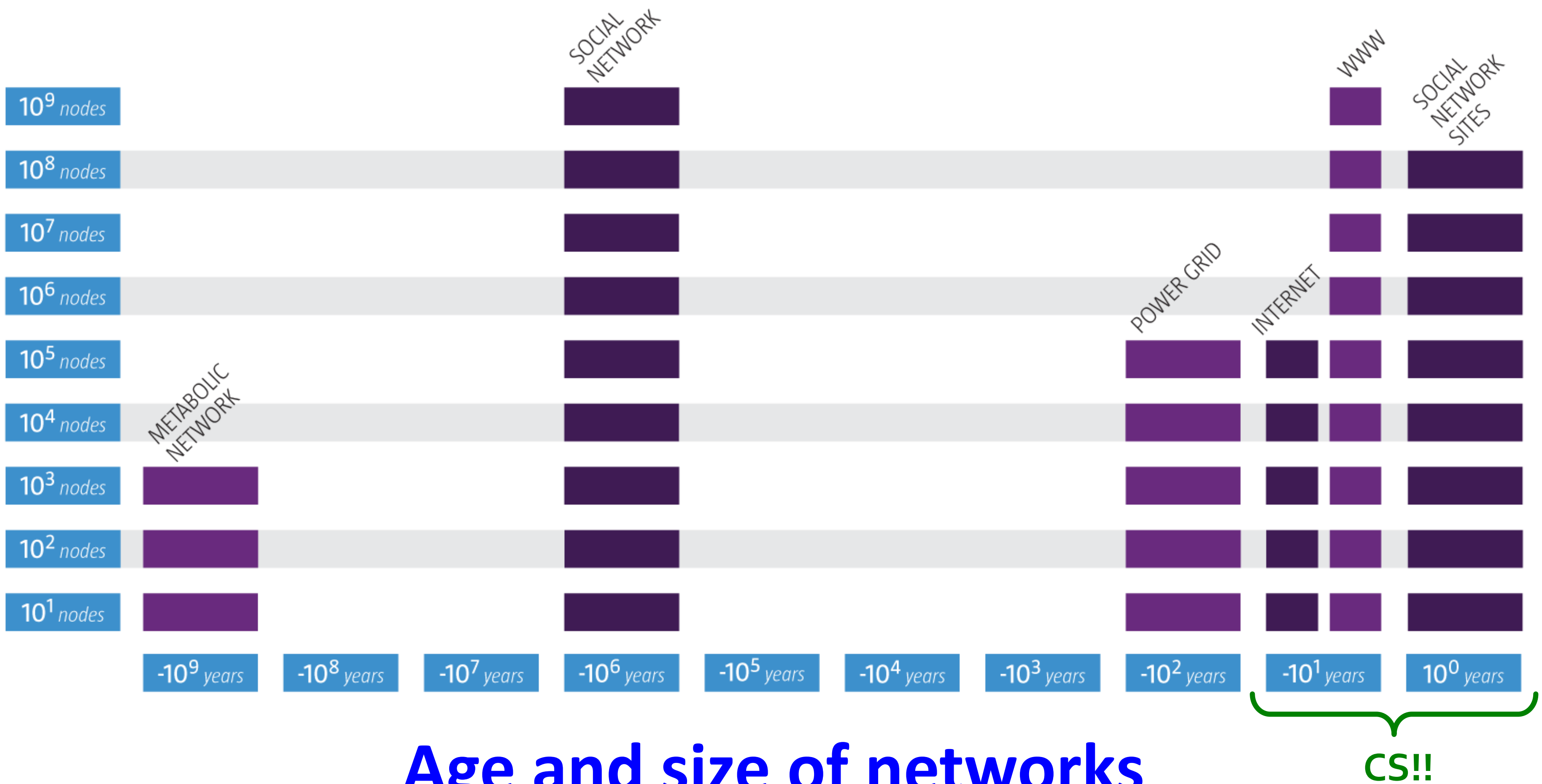






Universal language for describing complex data

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]

How can we *study* networks?

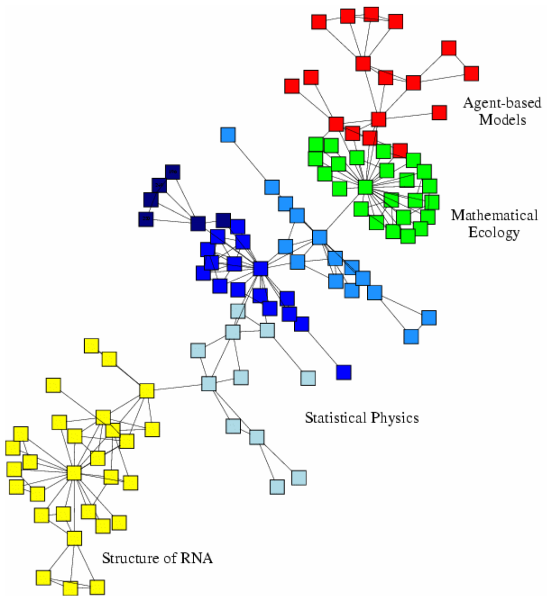
Network Analysis

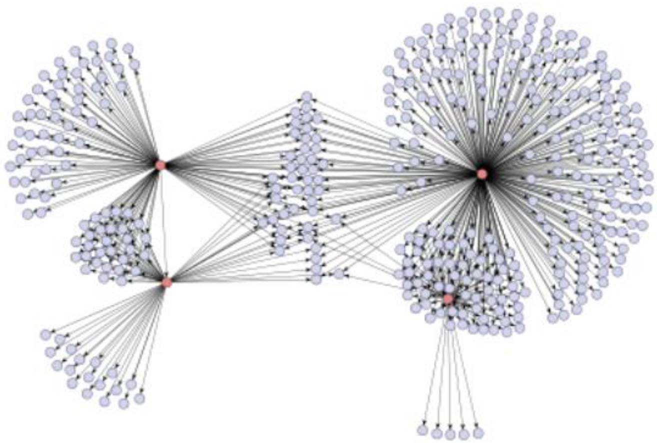
network analysis helps to reveal the *underlying dynamics* of these systems, not easily observable before

what do we study in networks?

Networks: Structure & Process

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





how do we reason about networks?

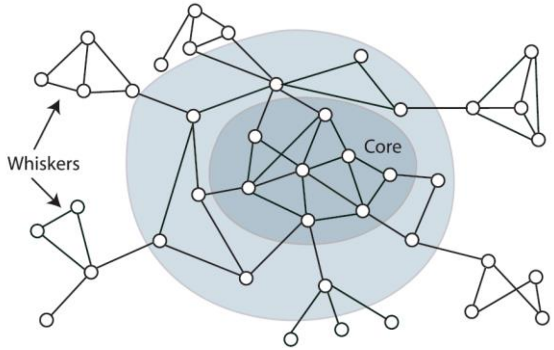
Reasoning About Networks

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

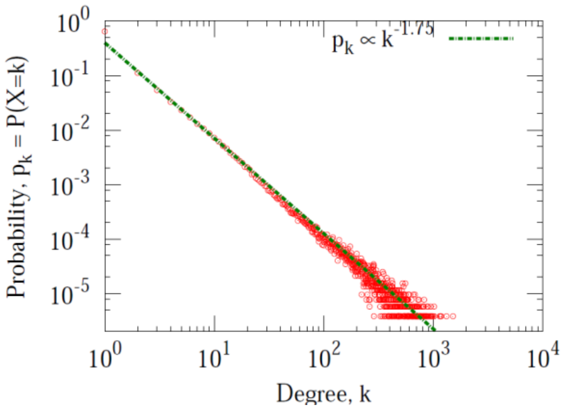
Properties

- **Six degrees of separ.**

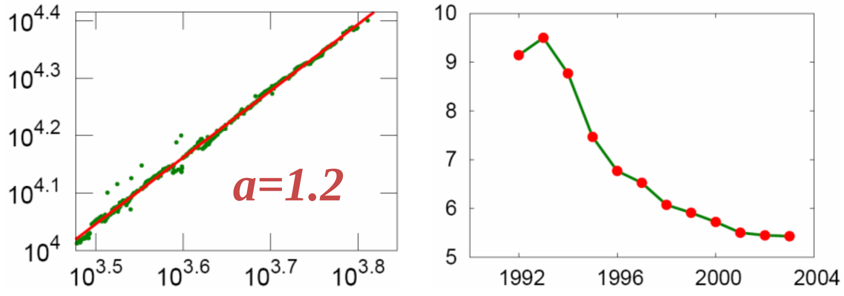
- **Strength of weak ties**



- **Power-law degrees**

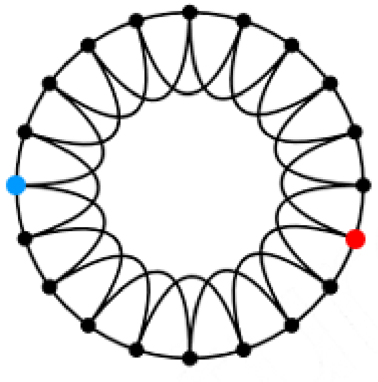


- **Densif. power law, Shrinking diameter**

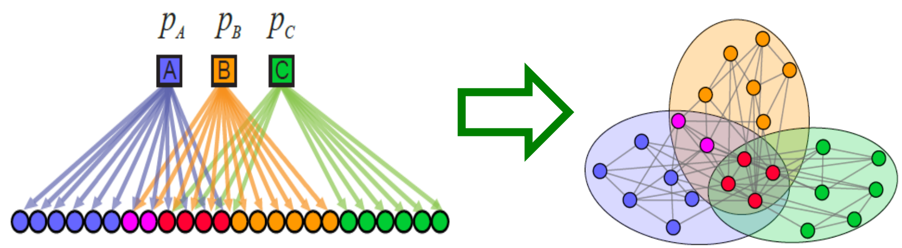


Models

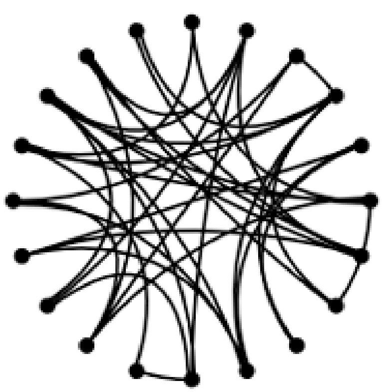
- Erdös-Renyi model



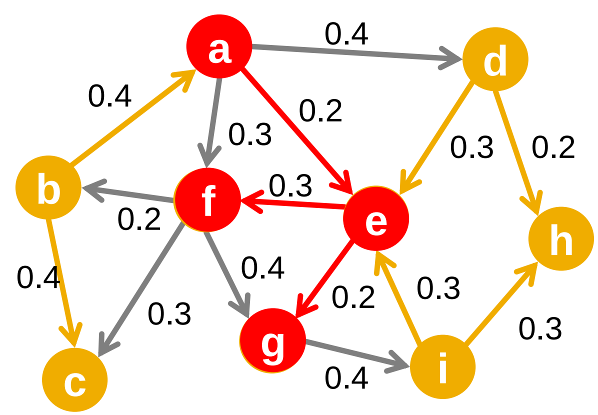
- Community model



- Small-world model

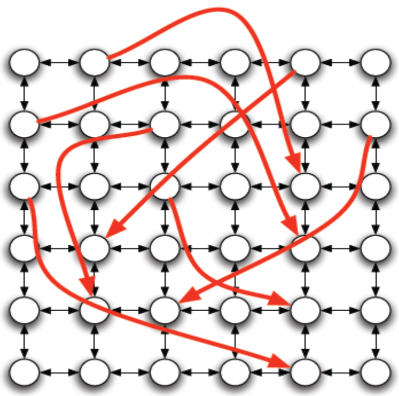


- Cascade model

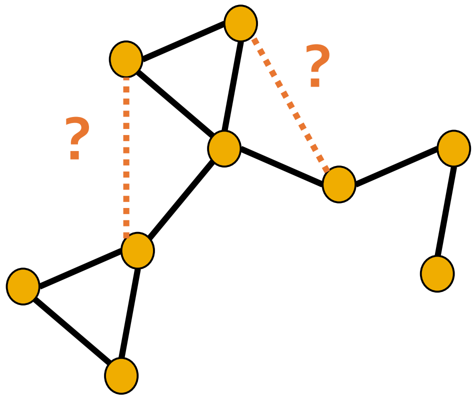


Algorithms

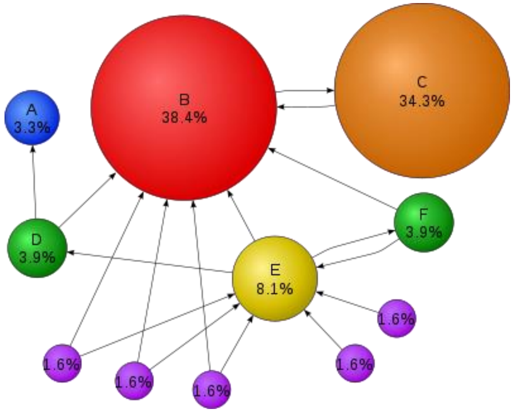
- **Decentralized search**



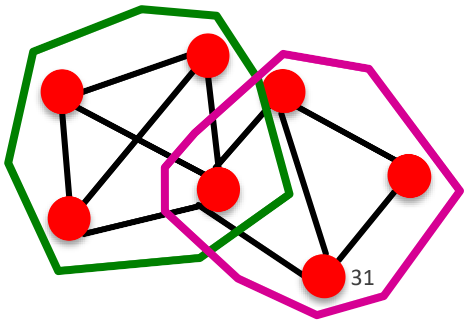
- **Link prediction**



- **Link analysis**

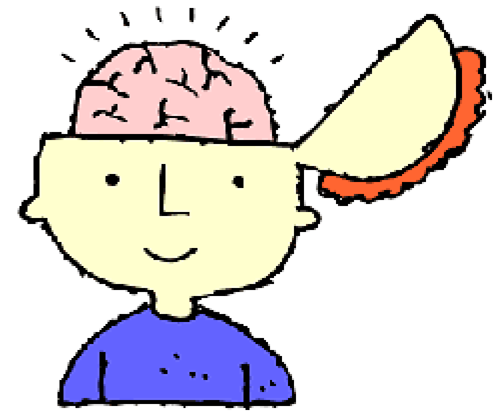


- **Community detection**



Map of Superpowers

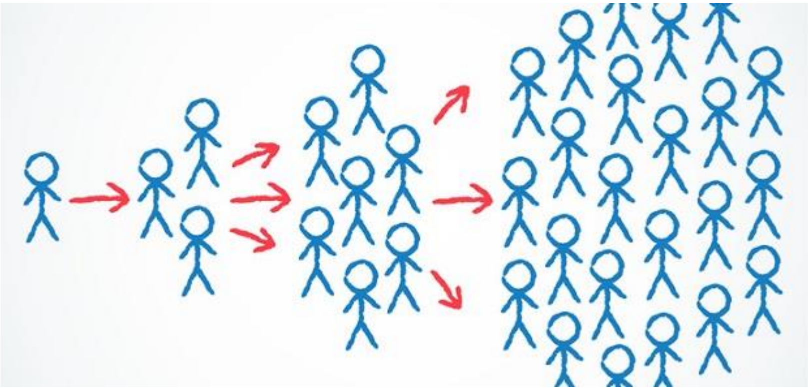
Properties	Models	Algorithms
Small diameter, Edge clustering	Small-world model, Erdös-Renyi model	Decentralized search
Scale-free	Preferential attachment, Copying model	PageRank, Hubs and authorities
Strength of weak ties, Core-periphery	Community-affiliation Graph Model	Community detection: Girvan-Newman, Modularity
Densification power law, Shrinking diameters	Microscopic model of evolving networks	Link prediction, Supervised random walks
Information virality, reproductive number	Independent cascade model, Game theoretic model, SIR	Influence maximization, Outbreak detection, LIM



Applying Our Superpowers

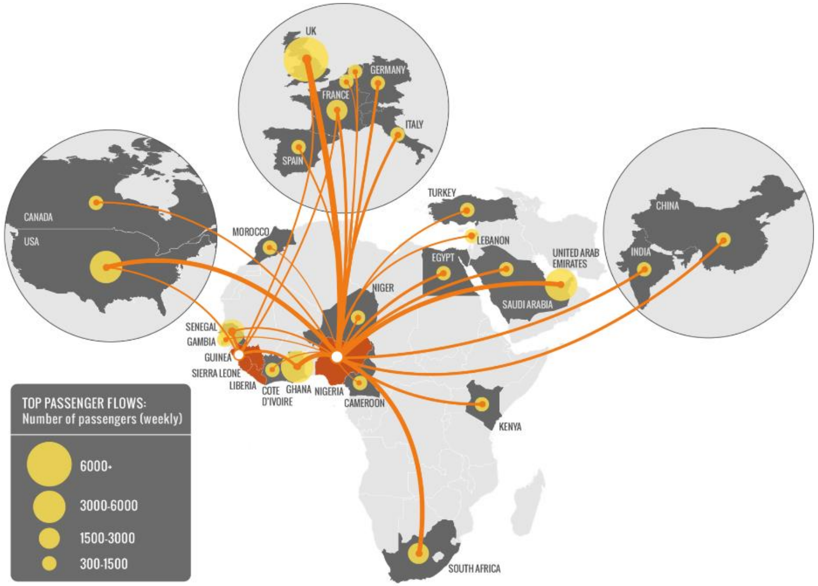
- Social media analytics

- Viral marketing

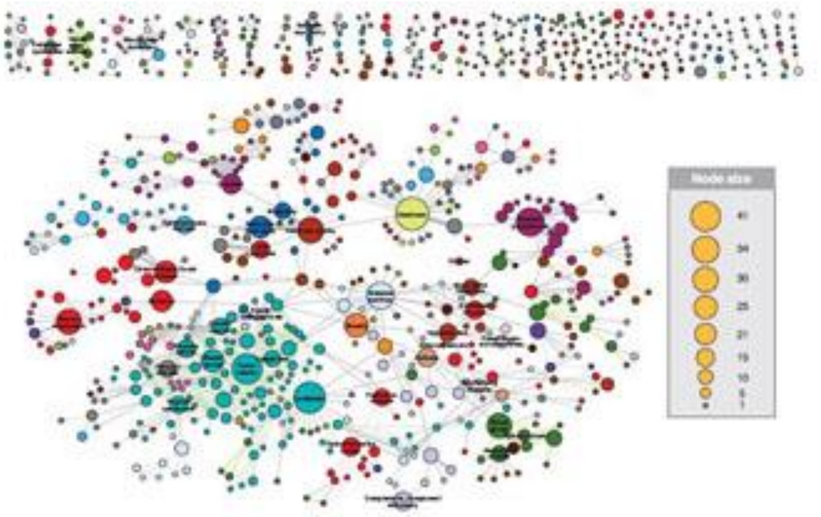


Applying Our Superpowers

- **Predicting epidemics: Ebola**



- **Drug design**



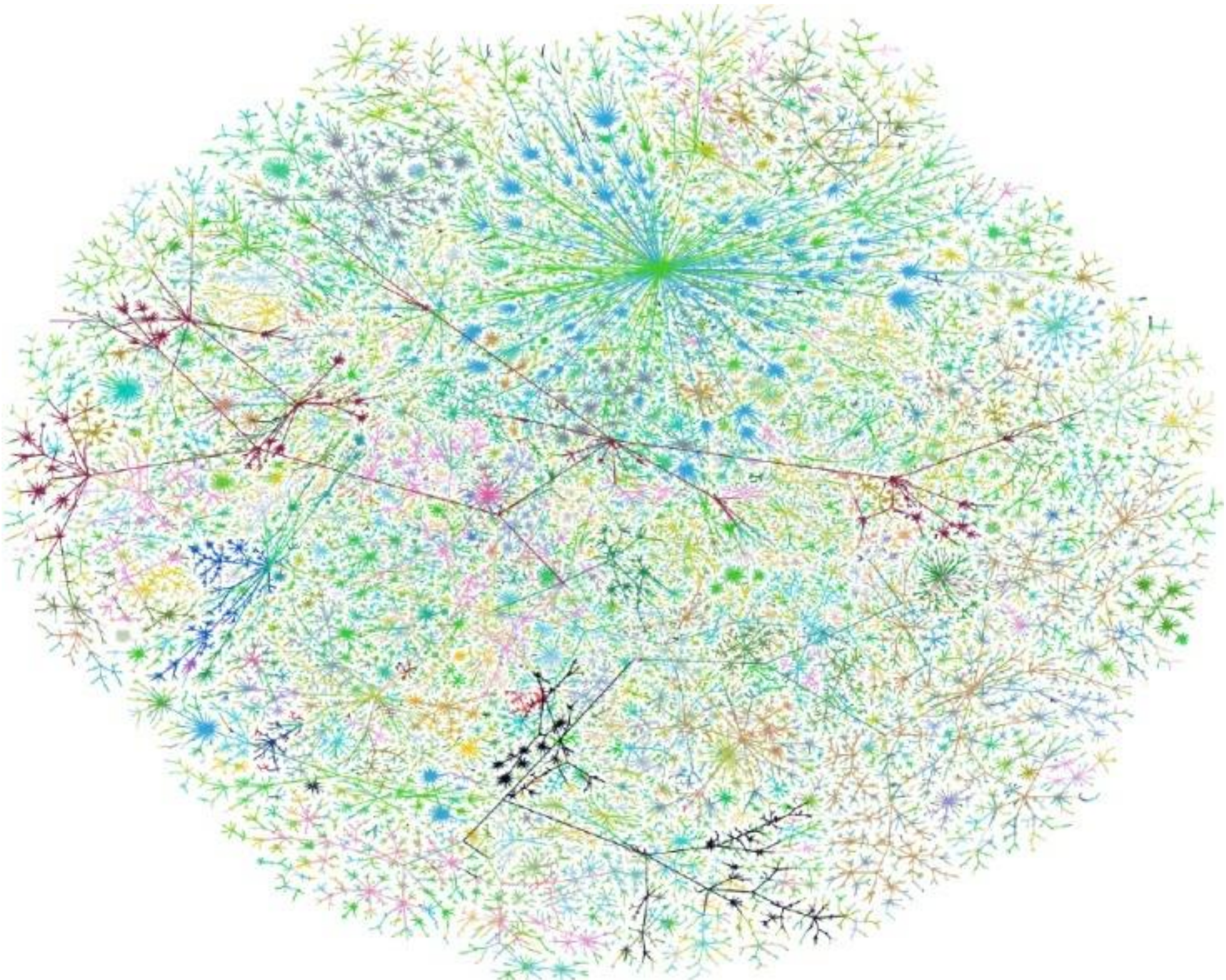
examples of network studies

Networks: Social

Facebook social graph

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

Networks: Communication

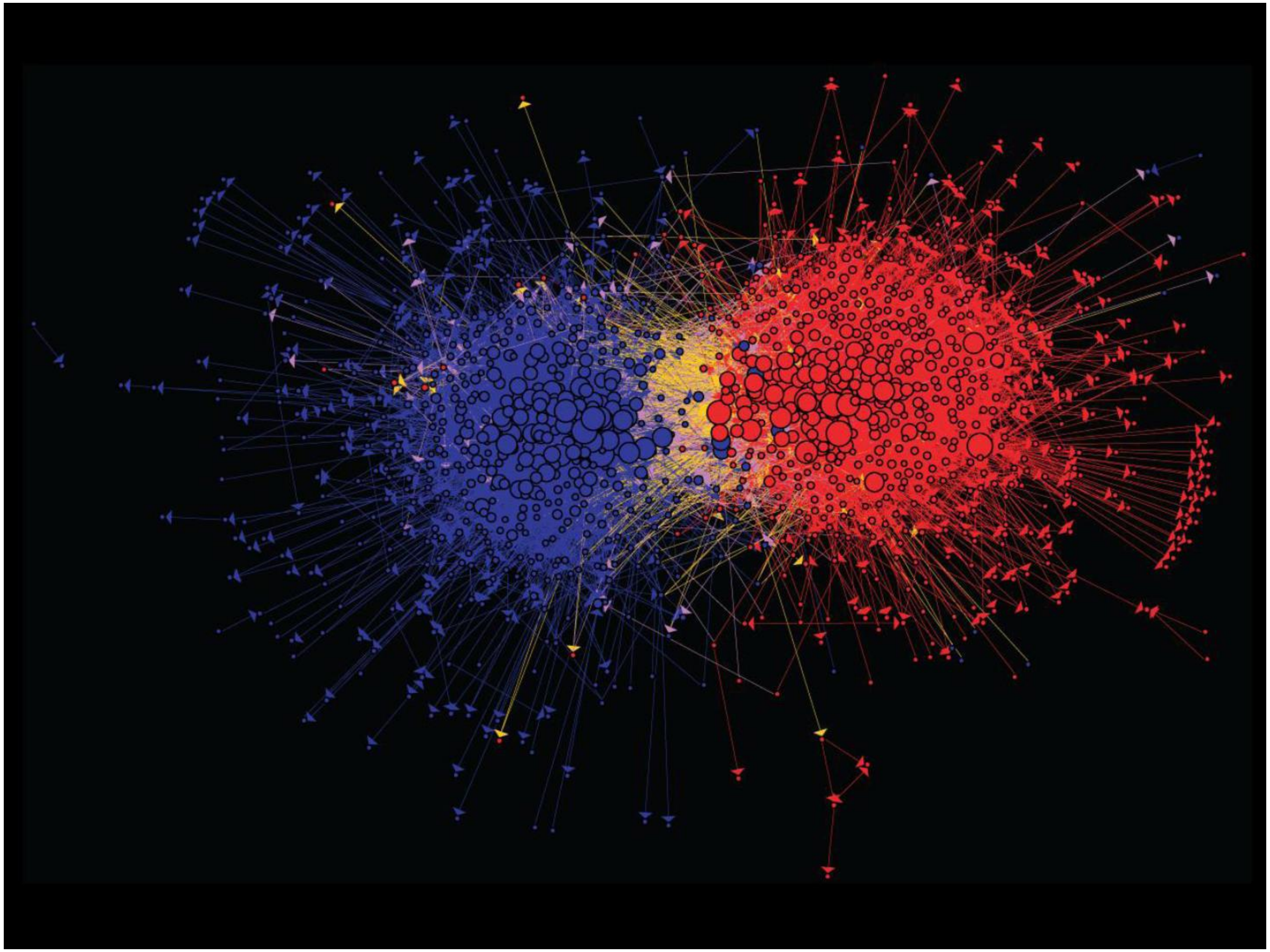


Graph of the Internet (Autonomous Systems)

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

Robustness [Doyle-Willinger, 2005]

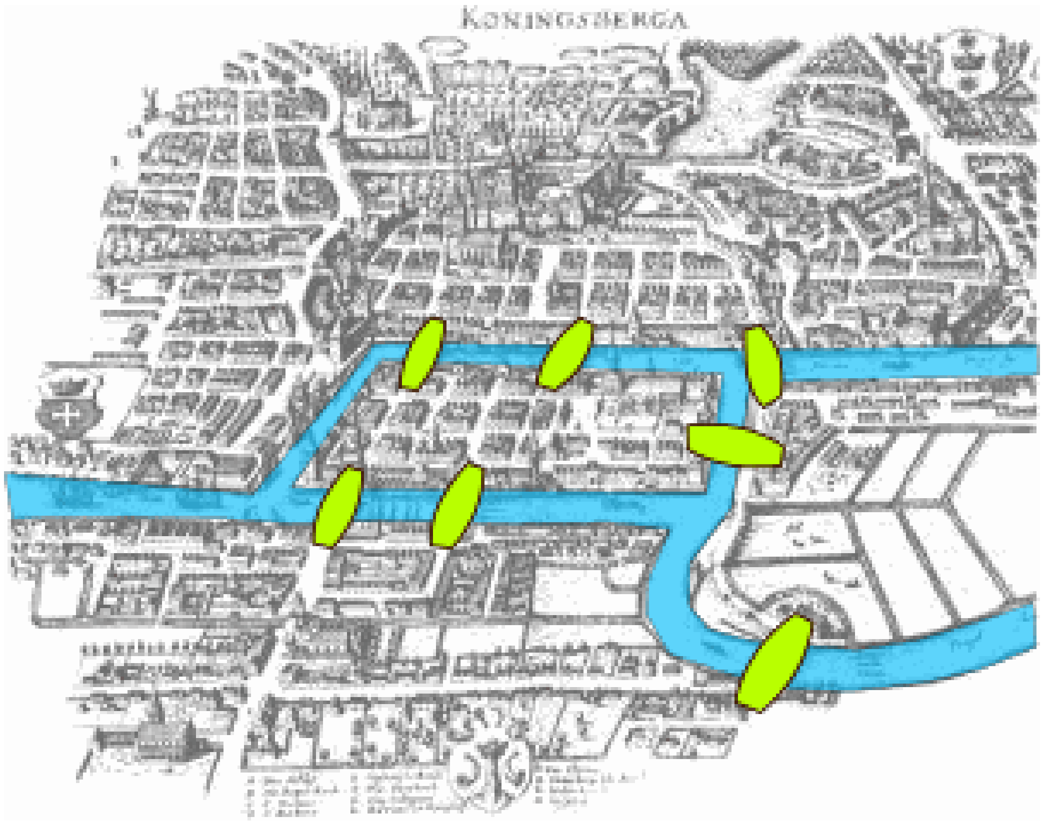
Networks: Media



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

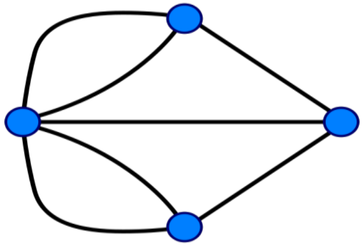
Networks: Infrastructure



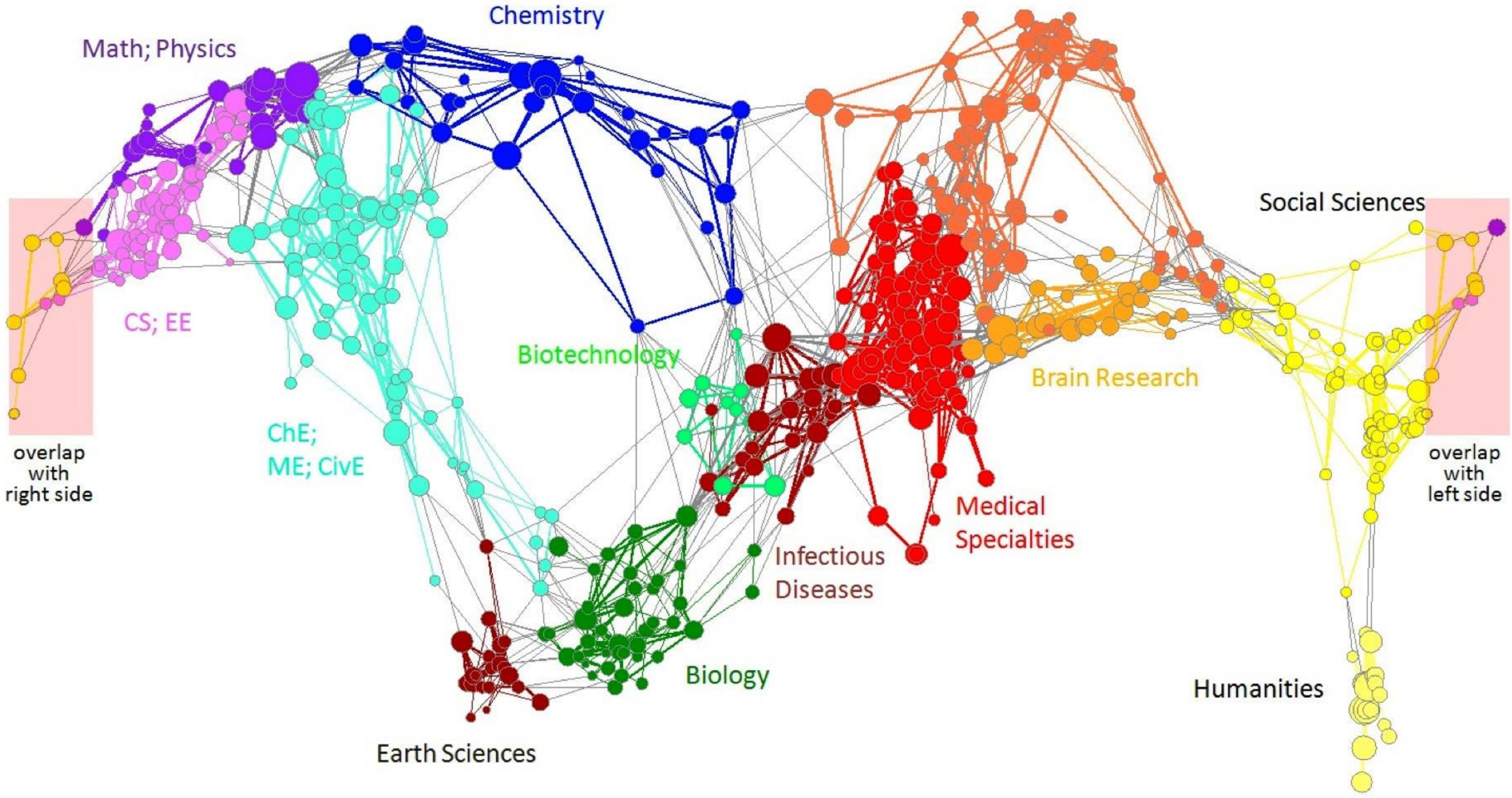
Seven Bridges of Königsberg

[Euler, 1735]

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



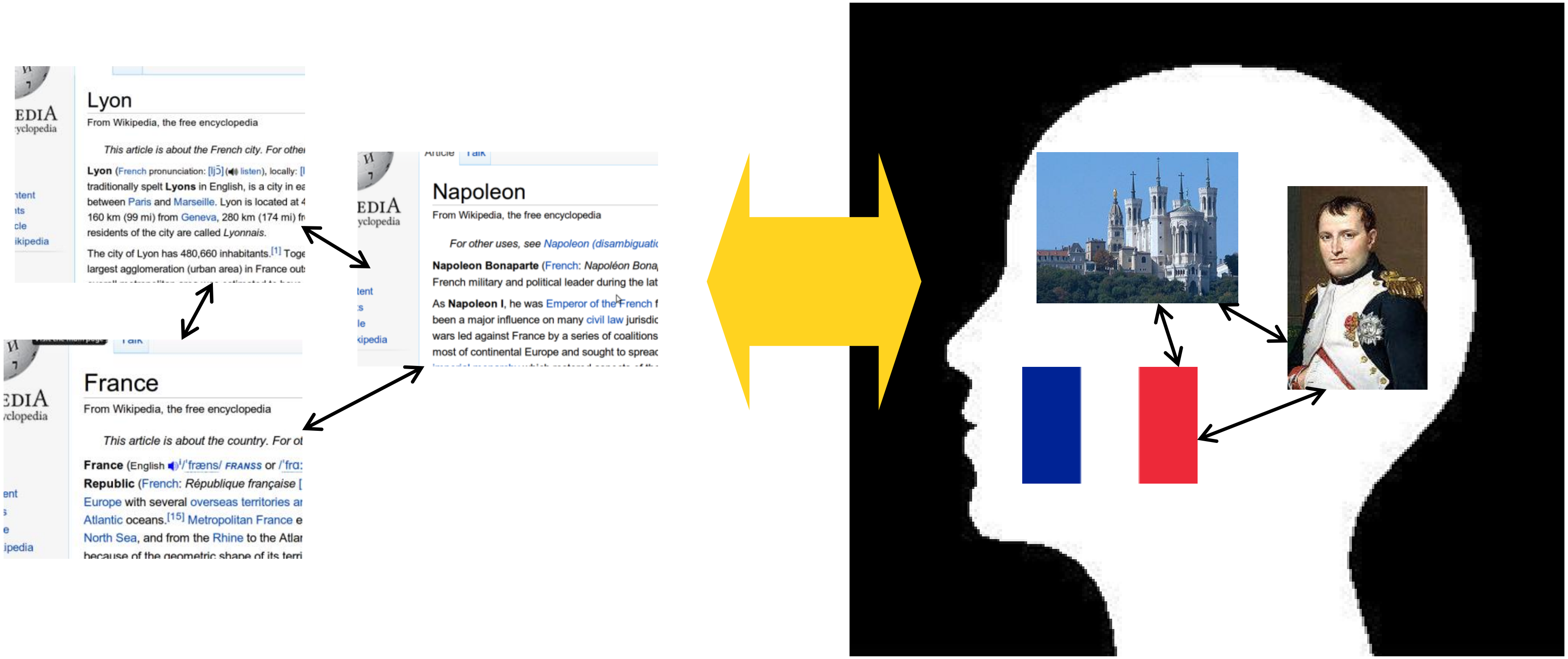
Networks: Citation



Citation networks and Maps of science

[Börner et al., 2012]

Networks: Knowledge

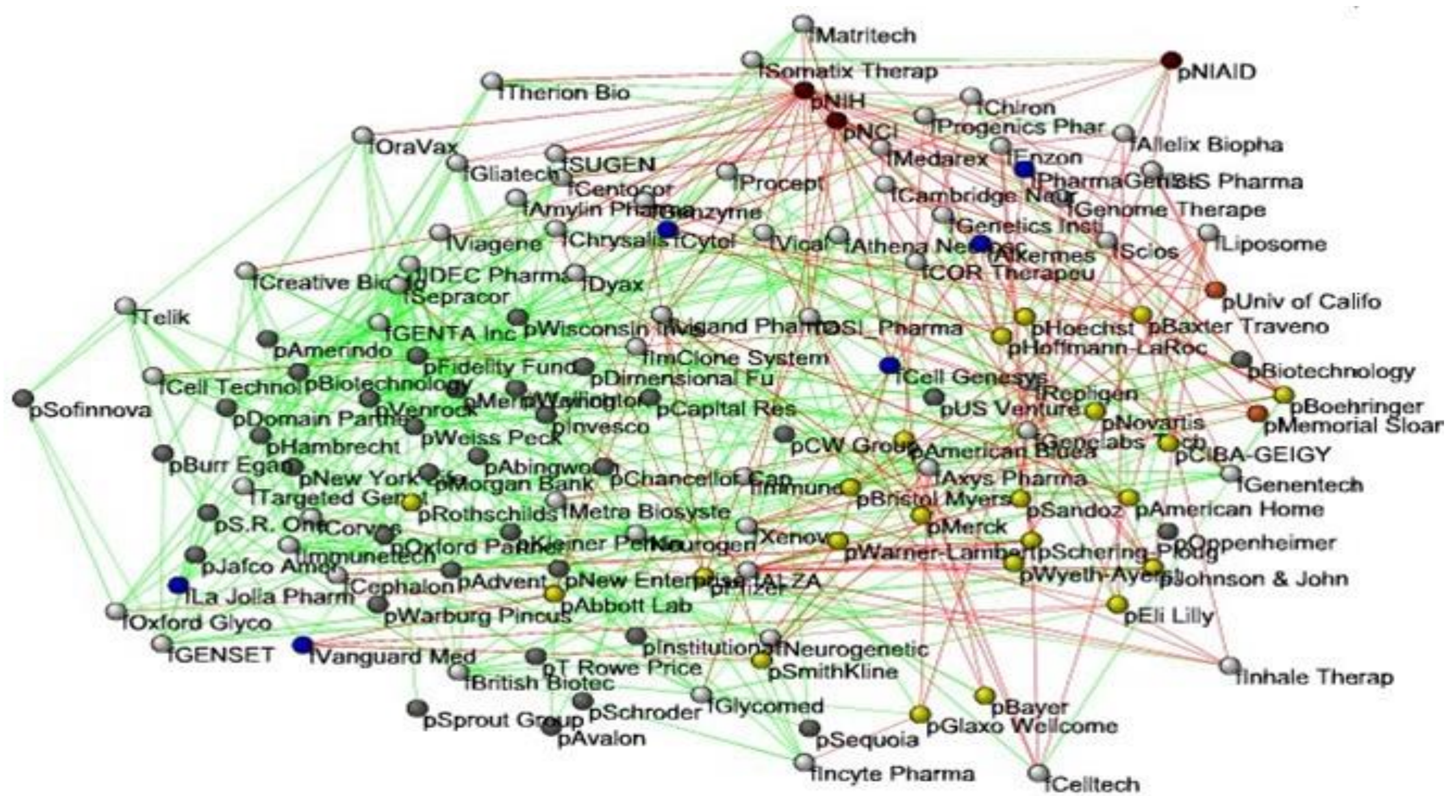


Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]

Networks: Economy



Nodes:

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

Links:

- Collaborations
- Financial
- R&D

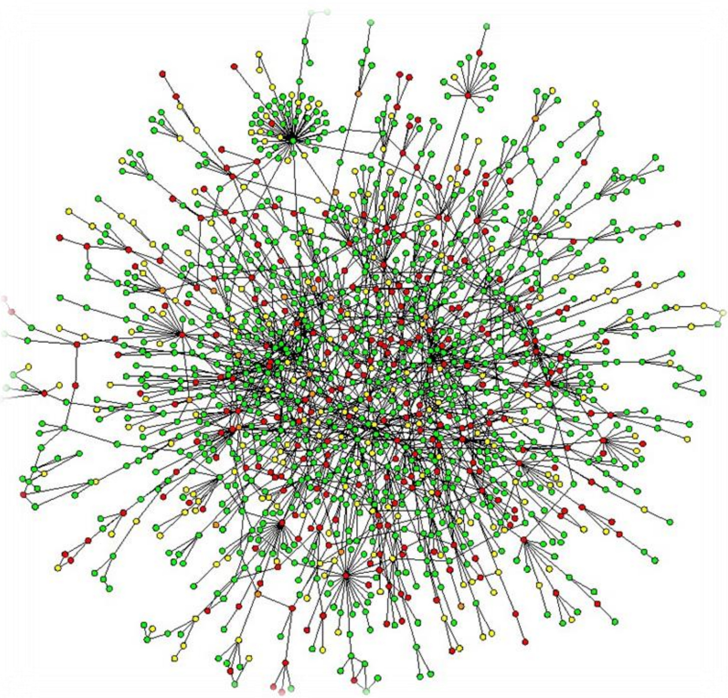
Bio-tech companies

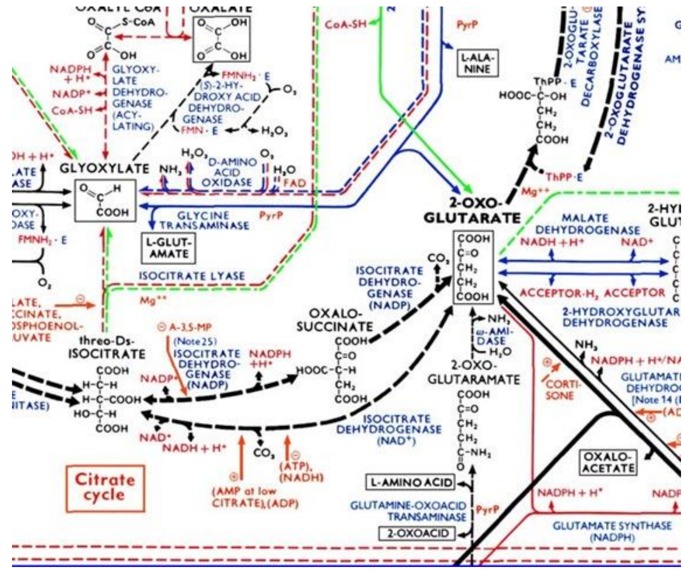
[Powell-White-Koput, 2002]

Networks: Brain

Human brain has between
~100 billion neurons, ~1,000 trillion synapses
[Sporns, 2011]

Networks: Biology





Protein-Protein Interaction Networks:

Nodes: Proteins

Edges: ‘physical’ interactions

Metabolic networks:

Nodes: Metabolites and enzymes

Edges: Chemical reactions

Web – The Lab for Humanity

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

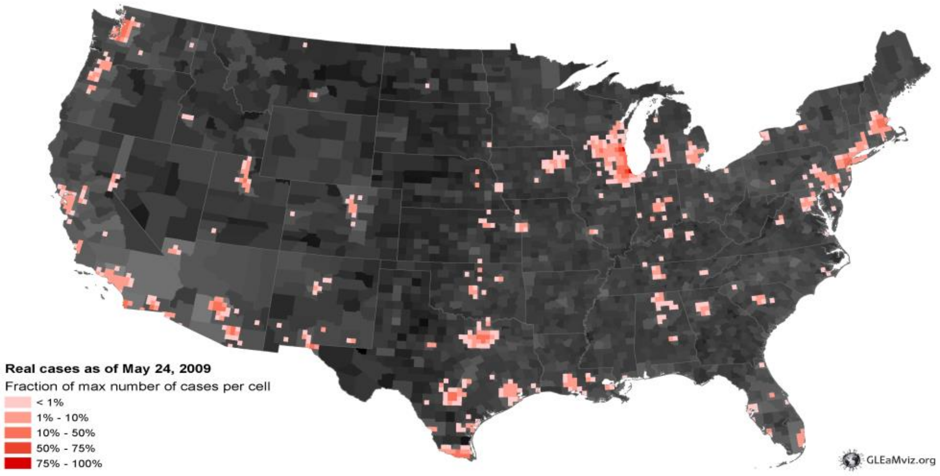
examples of network analysis impact

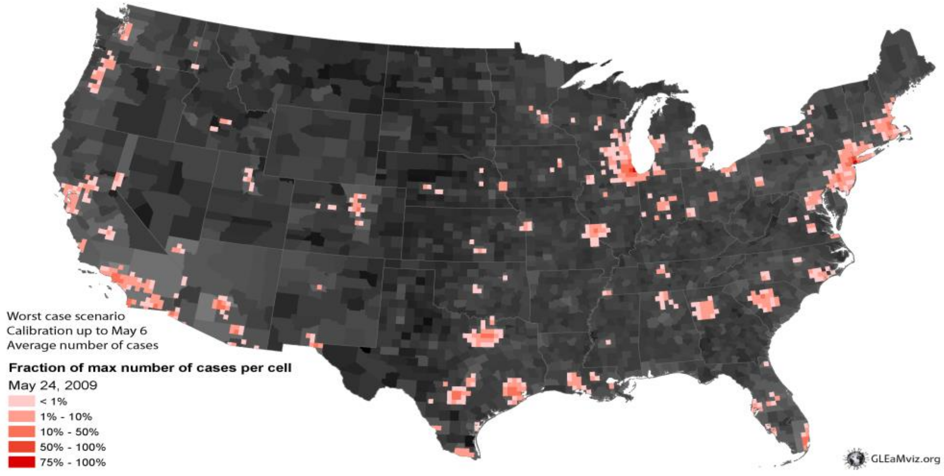
Networks: Impact

- **Google (~Australia?)**
  Market cap: $1700 billion
- **Cisco (~Hungary?)**
  Market cap: $200 billion
- **Meta (~Switzerland?)**
  Market cap: $940 billion

Networks: Impact

- **Predicting epidemics**





Real

Predicted

Networks Really Matter

- If you want to understand the spread of diseases, **can you do it without social networks?**
- If you want 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**

EECS4414/5414 Administrivia

Logistics: Communication

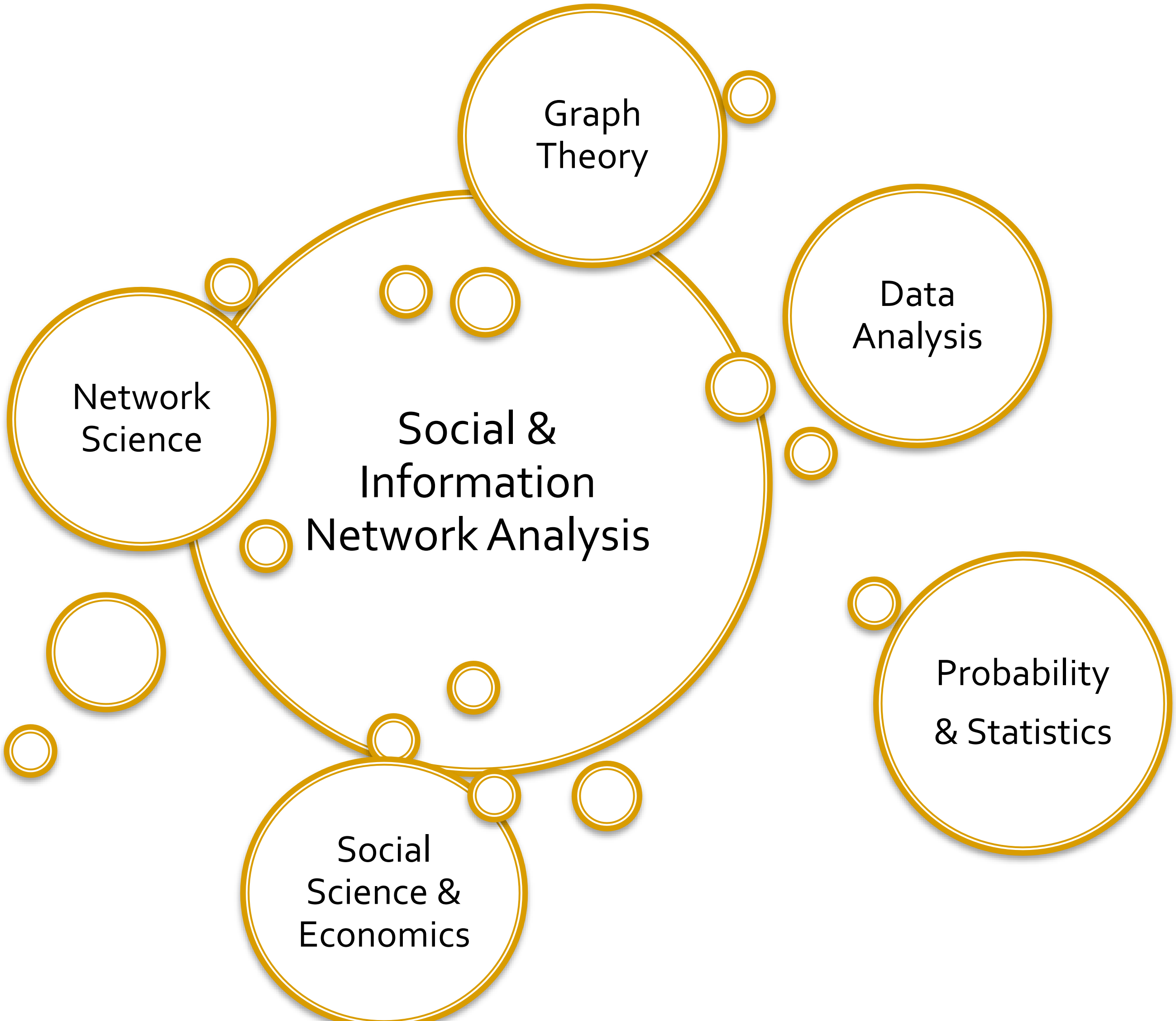
- **Website**
  - http://www.eecs.yorku.ca/~papaggel/courses/eecs4414/
- **eClass/Moodle**
  - https://eclass.yorku.ca/course/view.php?id=100607
- **Piazza Q&A website**
  - https://piazza.com/yorku.ca/summer2024/eecs44145414
  - You need to register with your ***yorku.ca*** email

    Please participate and help each other!
- **e-mail for personal issues**
  - papaggel@eecs.yorku.ca

Prerequisites

- **Course Prerequisites**
  - **EECS-3421: Introduction to Database Systems**
  - **EECS-3101: Design and Analysis of Algorithms**
  - **MATH-2030: Elementary Probability**
  - **General 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 or C)

Course Intellectual Content



Topics Covered

Component I

basic graph theory, network measurements, network models

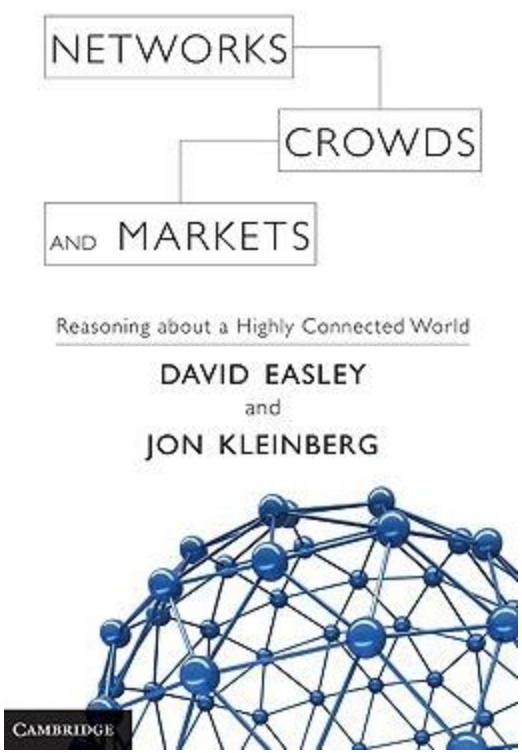
Component II

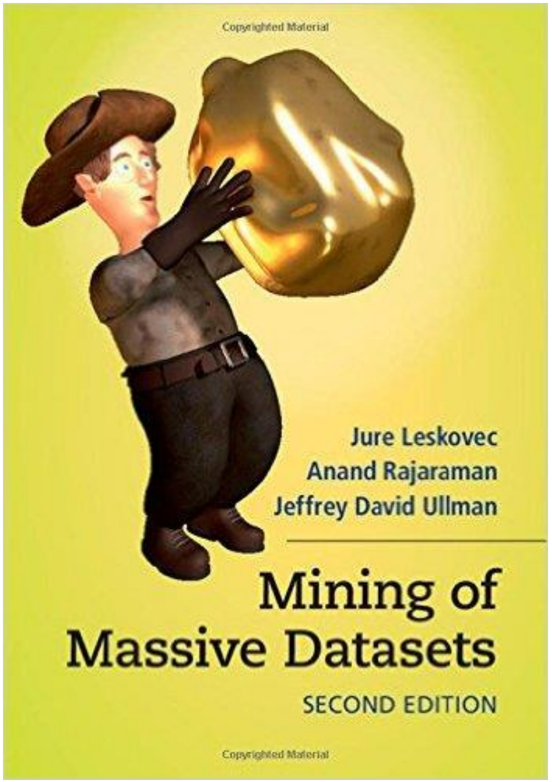
link analysis, link prediction, network ties, community detection, graph partitioning

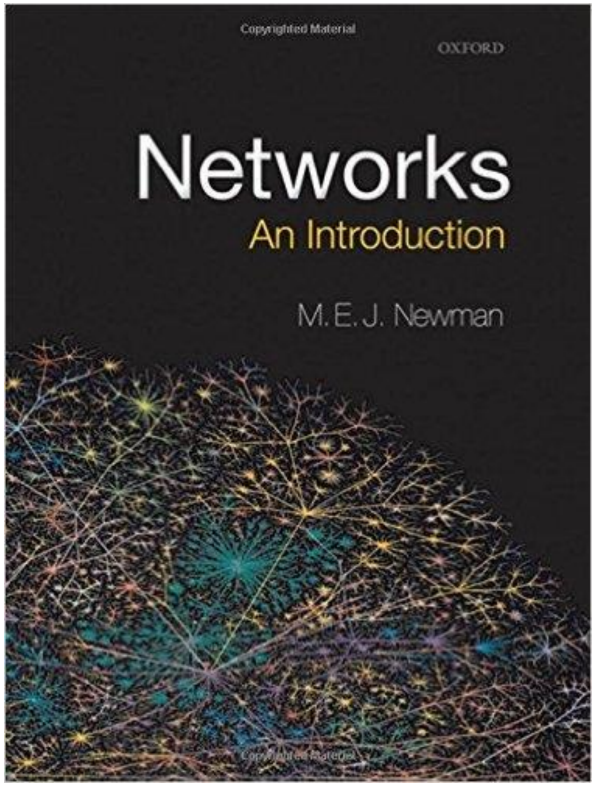
Component III

information cascades & epidemics, graph mining, machine learning with graphs, influence maximization, connections to problems in the social sciences and economics

“Suggested” Textbooks







+ a few more reference books

+ recent research papers on topics covered

Coursework

Work	Weight	Comment
2 Assignments	20%	A1: 10% A2: 10%
Research Project (team large project + report in research paper format)	40%	Proposal: 10% Project milestone: 20% In-class presentation: 20% Final report & code: 50%
Final Exam*	40%	Final exam grade must be > 40%

* **There will be no final exam for grad students (marking scheme to be determined):**

- **either:** Project: 80%, but **more substantial** and in **teams of two**
- **or:** Project: 60% + A lecture/presentation on a relevant research topic: 20%

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**
  - Graduate students: team of up to 2
- Project is the **main work** for the class
  - We will help with ideas, and mentoring
  - Need to start thinking **about this now**
- Class presentation

Network Analysis Tools

- **Highly recommend SNAP:**
  - **SNAP C++:** more challenging but more scalable
  - **SNAP.PY:** Python ease of use, most of C++ scalability
- Other tools include:
  - **NetworkX**
  - **JUNG**
  - **iGraph**
  - **…**

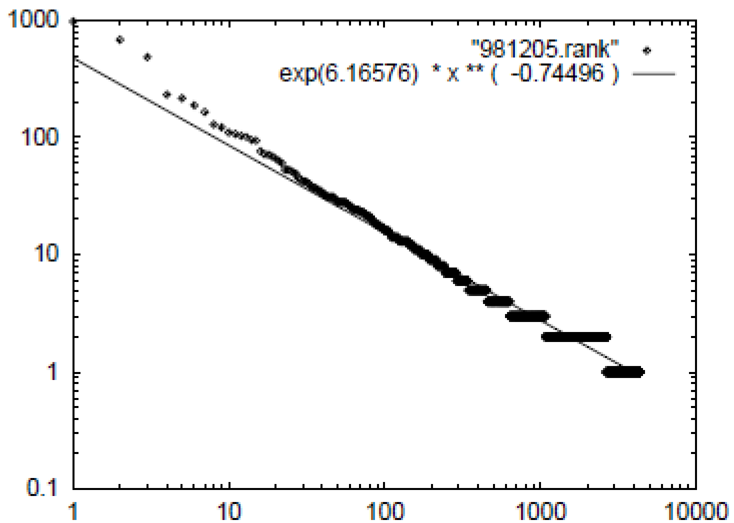
Example Research Questions/ Topics

Topics

- Measuring real networks
- Modeling the evolution of networks
- Identifying important nodes in the graph
- Finding communities in graphs
- Link prediction and recommendation
- Modeling information cascades in networks
- …

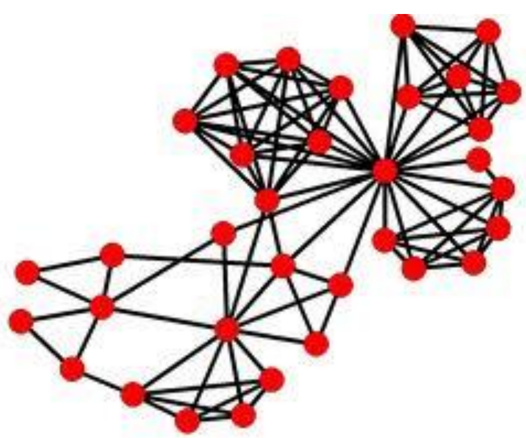
Understanding Large Graphs

- What does a network look like?
  - Measure different properties to understand the structure



(a) Int-12-98

degree of nodes



Triangles in the graph

Modeling Real Networks

- Real life networks are not "random"
- Can we define a model that generates graphs with statistical properties similar to those in real life?
- The rich-get-richer model

We need to accurately model the mechanisms that govern the evolution of networks (for prediction, simulations, understanding)

Ranking Nodes on the Web

- Is my home page as important as the facebook page?
- We need algorithms to compute the importance of nodes in a graph
- The PageRank Algorithm
  - A success story of network use

It is impossible to create a web search engine without understanding the web graph

Link Prediction

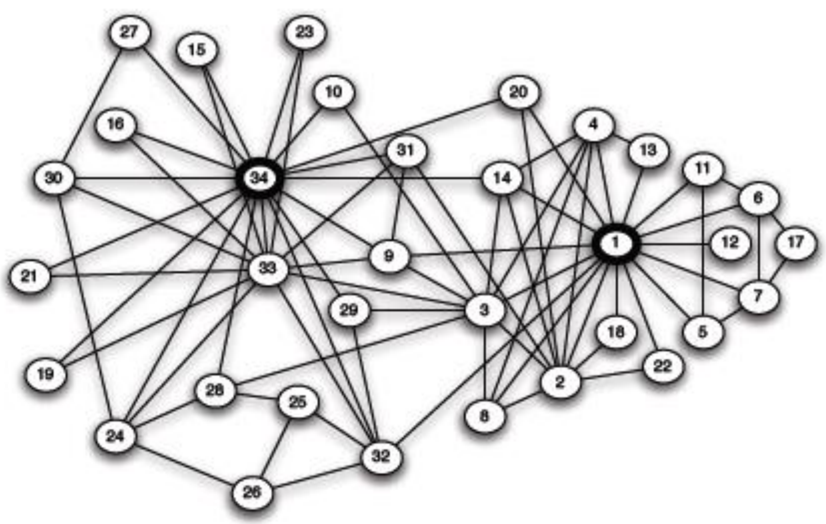
- Given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'.
- Applications
  - Accelerate the growth of a social network (e.g., Facebook, LinkedIn, Twitter)
  - Maximize information cascades

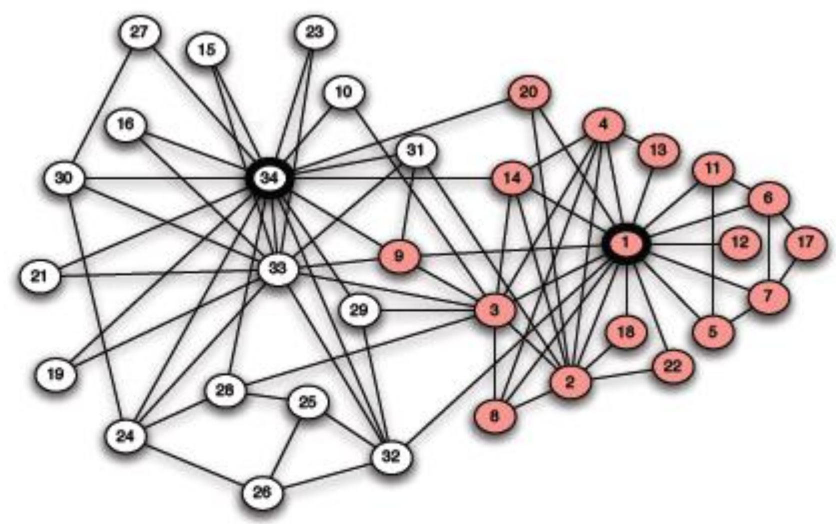
How do we predict future links?

Clustering and Communities

- What is community?
  - “Cohesive subgroups are subsets of actors among whom there are relatively strong, direct, intense, frequent, or positive ties.” [Wasserman & Faust ‘97]

Karate club example [W. Zachary, 1970]





Information/Virus Cascade

- How do viruses spread between individuals? How can we stop them?
- How does information propagates in social and information networks? What items become viral? Who are the influencers and trend-setters?
- We need models and algorithms to answer these questions

Online advertising relies heavily on online social networks and word-of-mouth marketing. There is currently need for models for understanding the spread of Covid-19 virus.

Mining Social Media

- Social Media (Twitter, Facebook, Instagram) have supplanted the traditional media sources
  - Information is generated and disseminated by users
- Interesting problems:
  - Automatically detect events using Twitter
    - Earthquake response
    - Crisis detection and management
  - Sentiment mining
  - Track the evolution of events: socially, geographically, over time
  - …

Research in Graph Mining

- Current hot research topics:
  - Graph representation learning
  - Graph neural networks
  - Graph attention mechanisms
  - Graph generative models
  - Graph classification, clustering, anomaly detection
  - Dynamic graph analysis and mining
- Relevant research conferences
  - Data Mining: KDD, ICDM, WSDM, WWW, …
  - ML: ICML, NeurIPS, ECML/PKDD, …

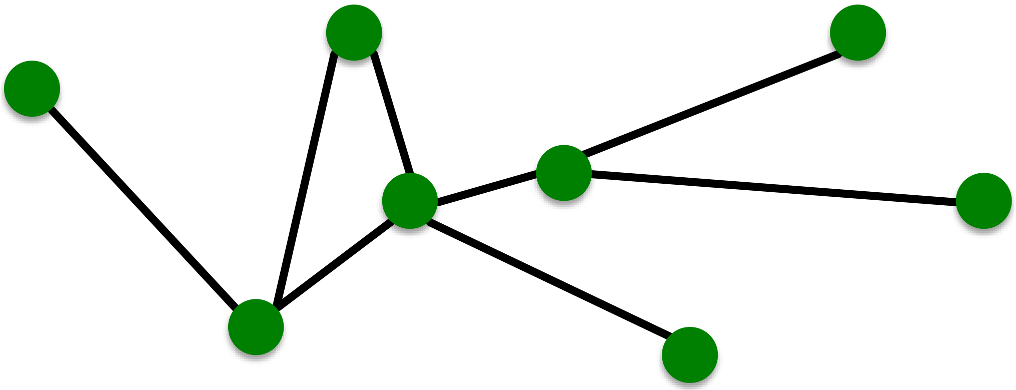
Starter Topic:
Structure of the Web Graph

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

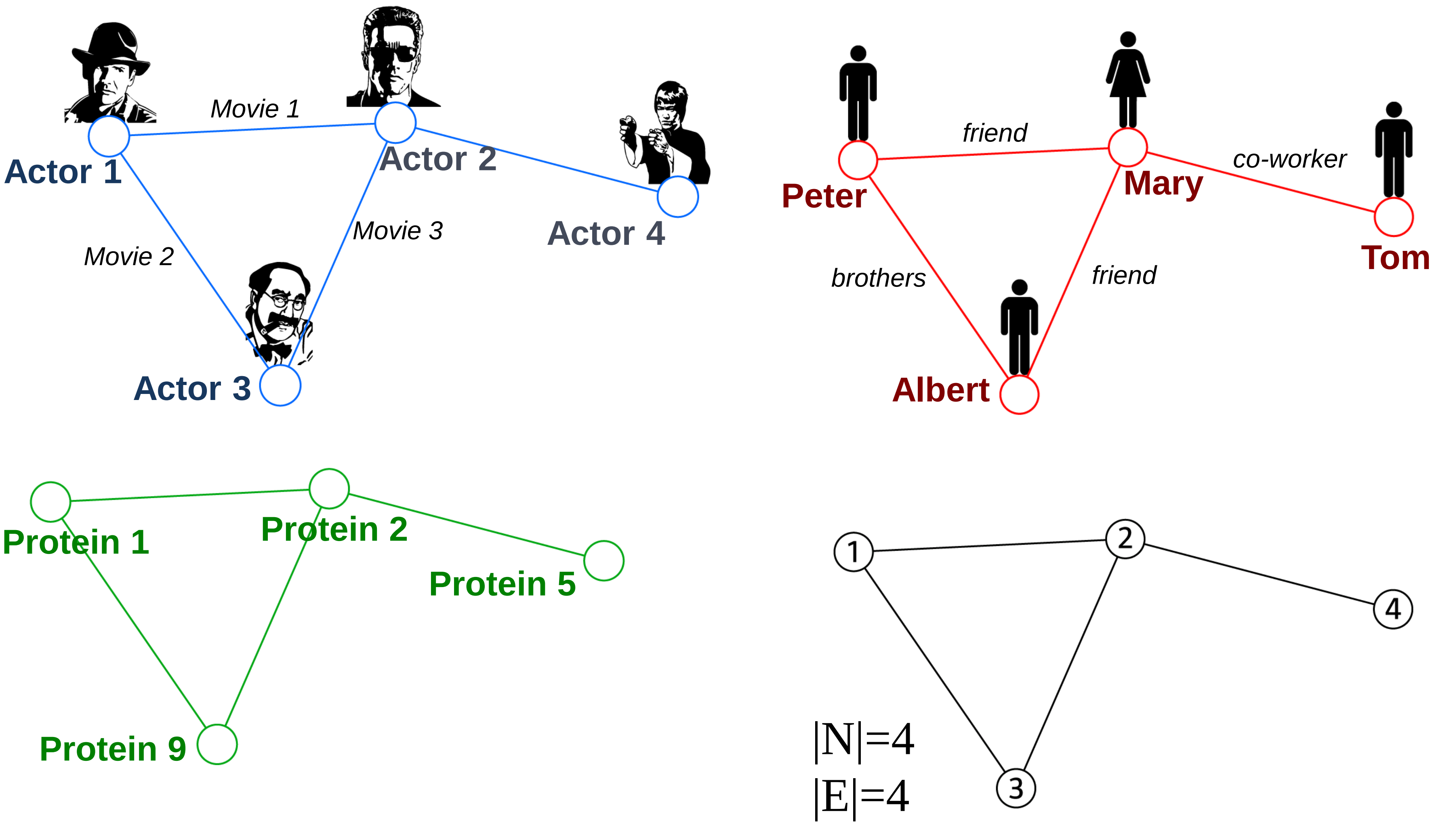
  Language: Network, node, link

- **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

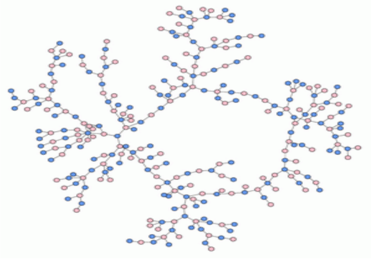


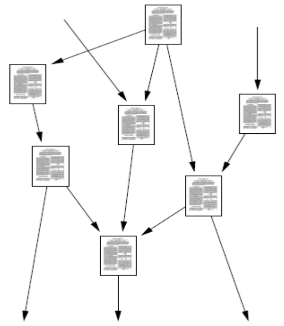
Choosing Proper Representation

- **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

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**



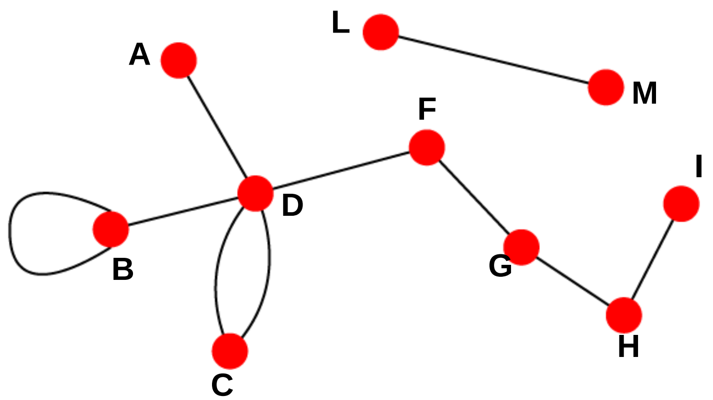


- **If you connect all papers with the same word in the title, you will be exploring what?** It is a network, nevertheless

Undirected vs. Directed Networks

Undirected

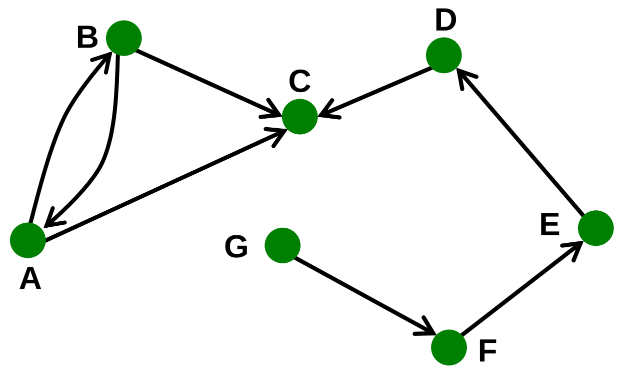
- **Links:** undirected (symmetrical, reciprocal)



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

Directed

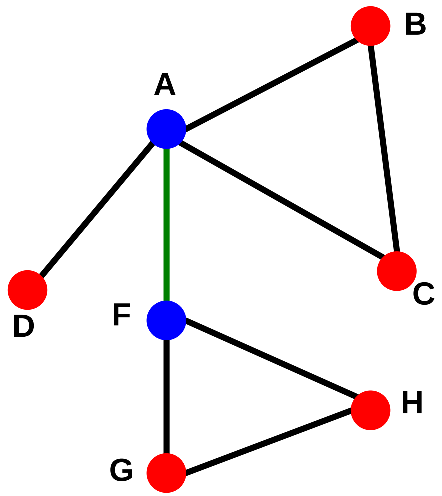
- **Links:** directed (arcs)

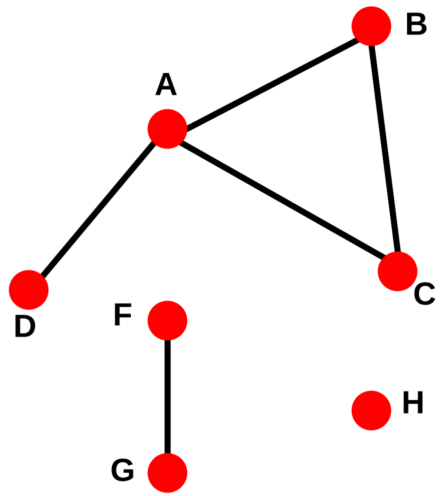


- **Examples:**
  - Phone calls
  - Following on Twitter

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





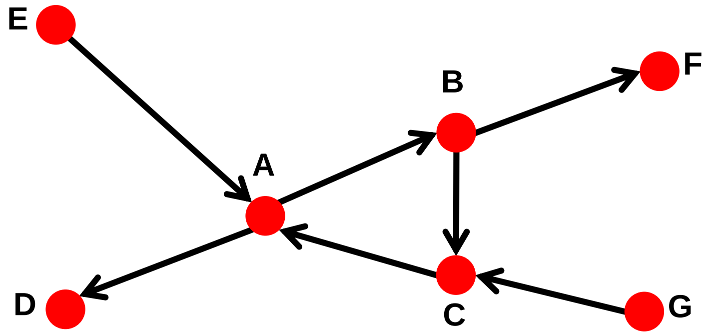
Largest Component:
Giant Component

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

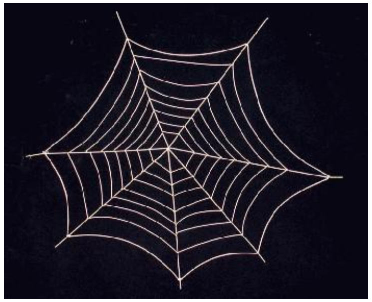
- **Strongly connected directed graph**
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)
- **Weakly connected directed graph**
  - is connected if we disregard the edge directions



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"?**



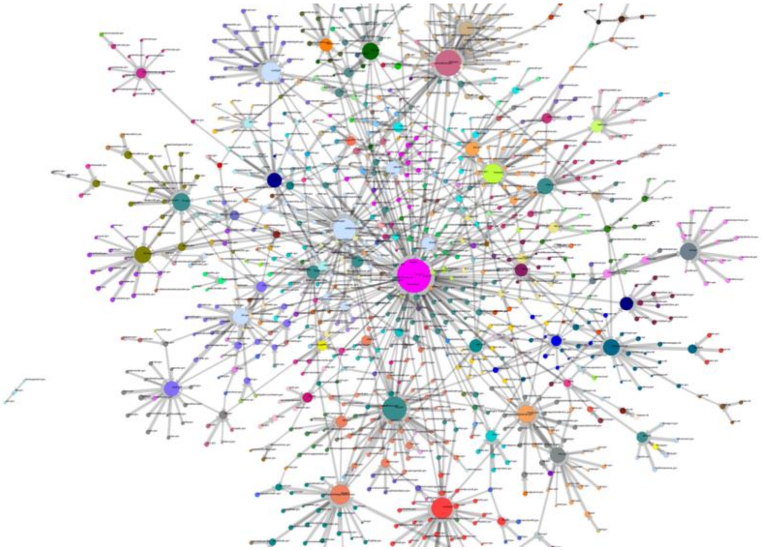
- **Here is what we will do next:**
  - We will take a real system (i.e., the Web)
  - We will represent the Web as a graph
  - We will use language of graph theory to reason about the structure of the graph
  - Do a computational experiment on the Web graph
  - **Learn something about the structure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

Web as a Graph

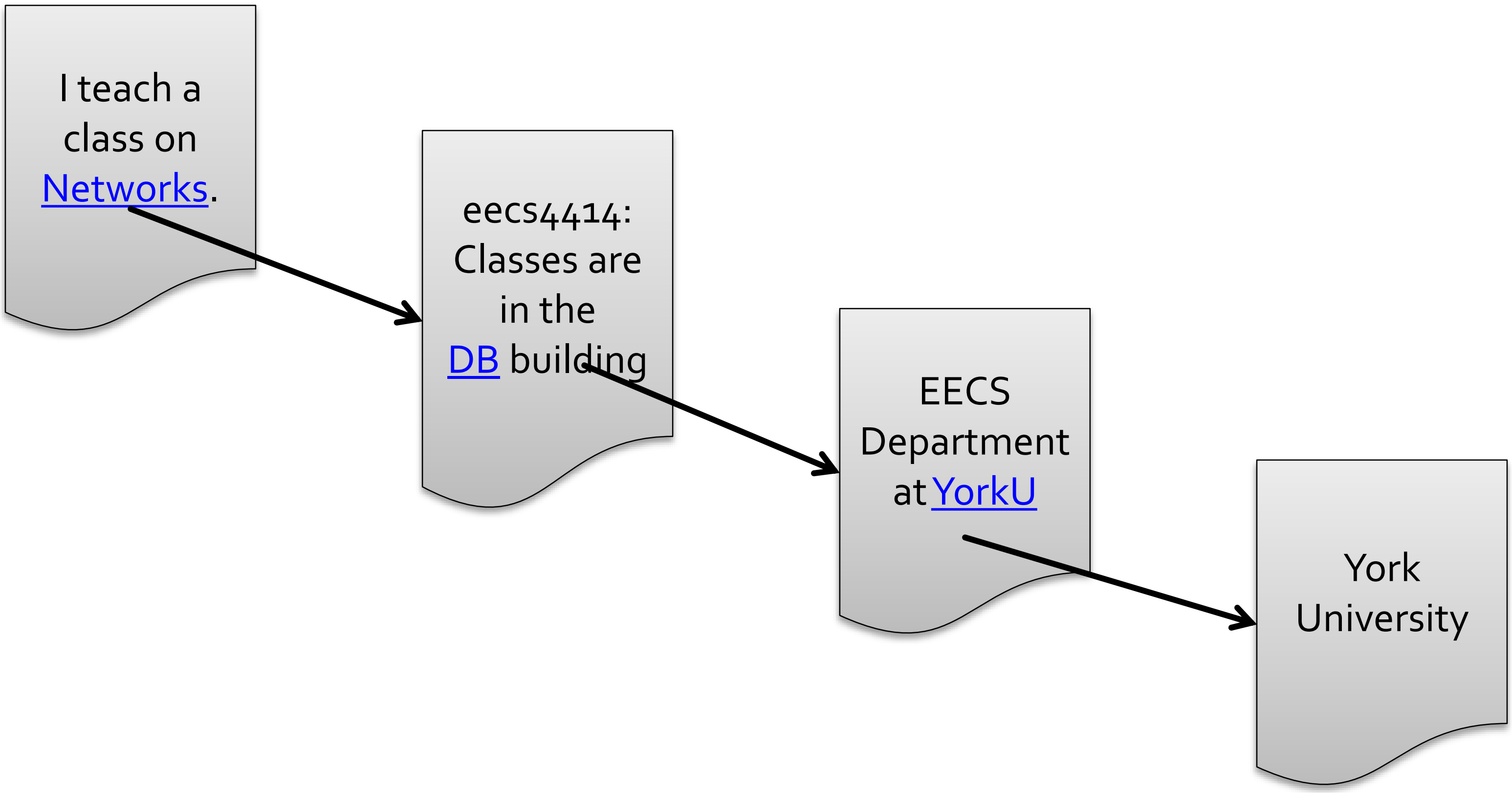
Q: What does the Web "look like" at a global level?

- **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

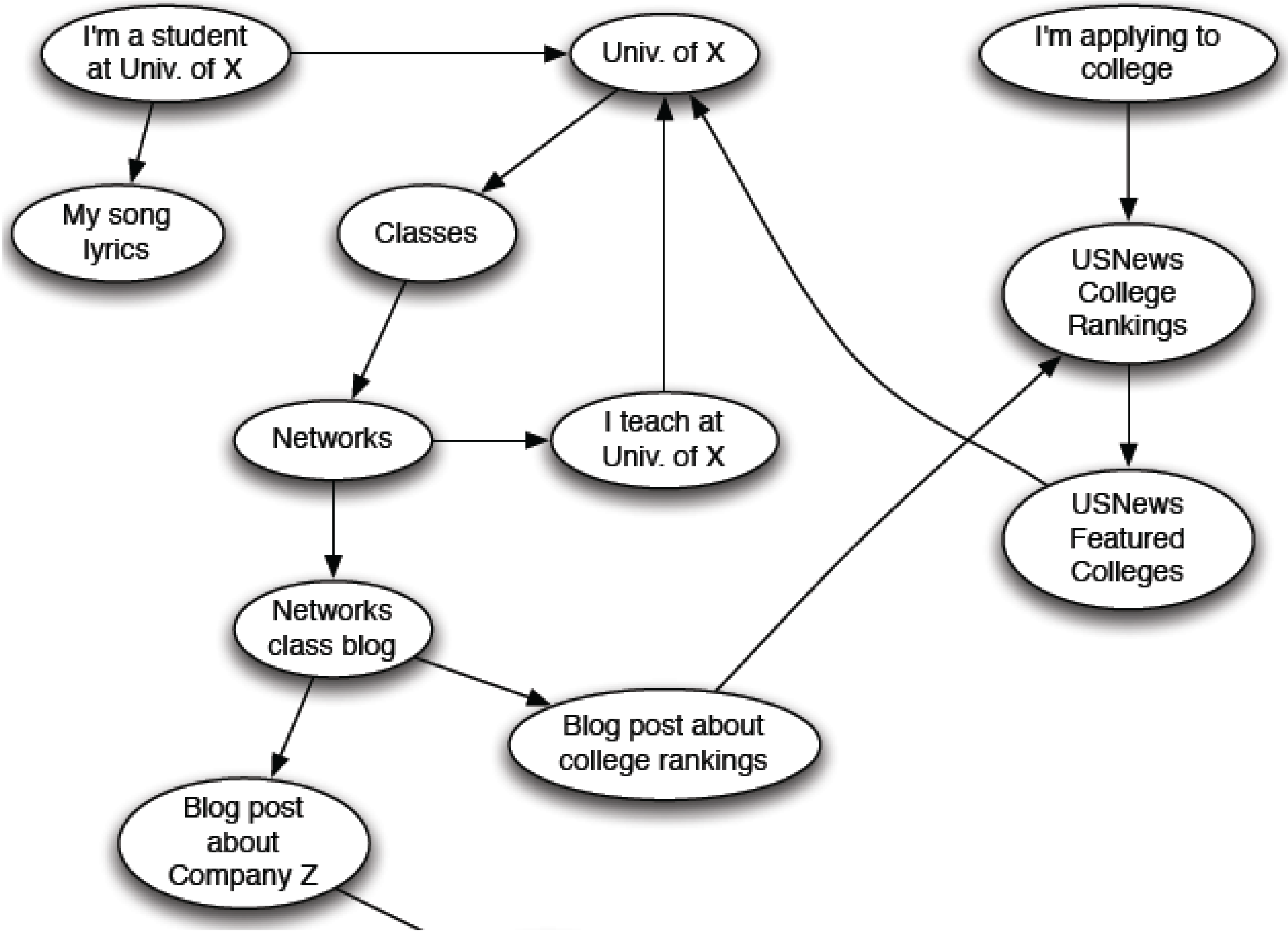


The Web as a Graph



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

The Web as a Directed Graph

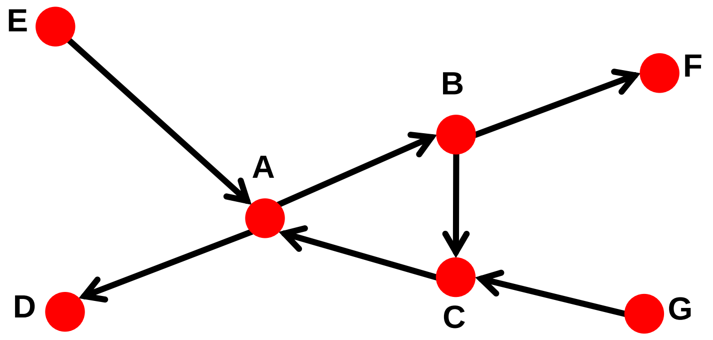


What Does the Web Look Like?

- **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?

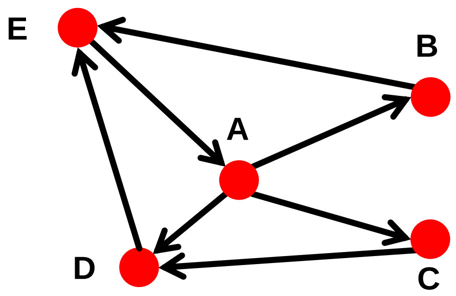


$$In(v) = \{w \mid w \text{ can reach } v\}$$
$$Out(v) = \{w \mid v \text{ can reach } w\}$$

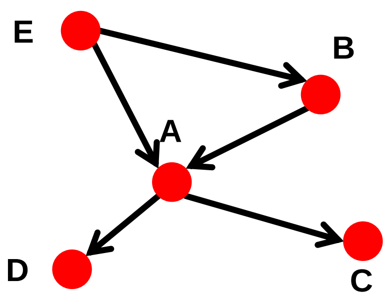
For example:
In(A) = {A,B,C,E,G}
Out(A)={A,B,C,D,F}

Directed Graphs

- **Two types of directed graphs:**
  - **Strongly connected:**
    - Any node can reach any node via a directed path
      $In(A)=Out(A)=\{A,B,C,D,E\}$



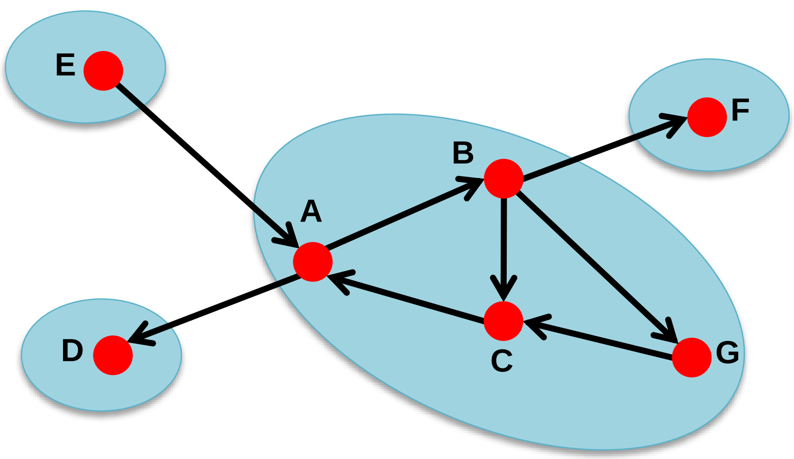
  - **DAG – Directed Acyclic Graph:**
    - Has no cycles: if $\boldsymbol{u}$ can reach $\boldsymbol{v}$, then $\boldsymbol{v}$ can not reach $\boldsymbol{u}$



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

Strongly Connected Component

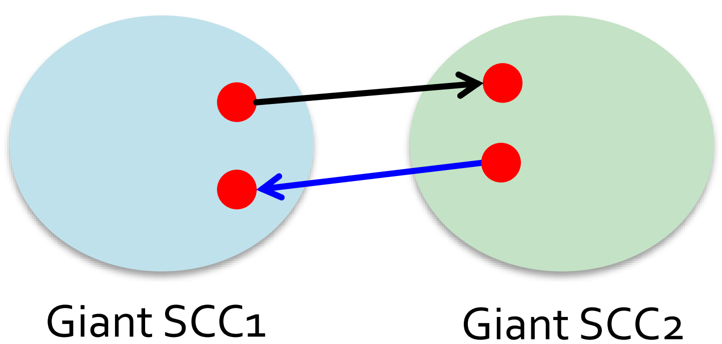
- **Strongly connected component (SCC)** is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



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

Graph Structure of the Web

- **There is a single giant SCC**
  - That is, there won't be two 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



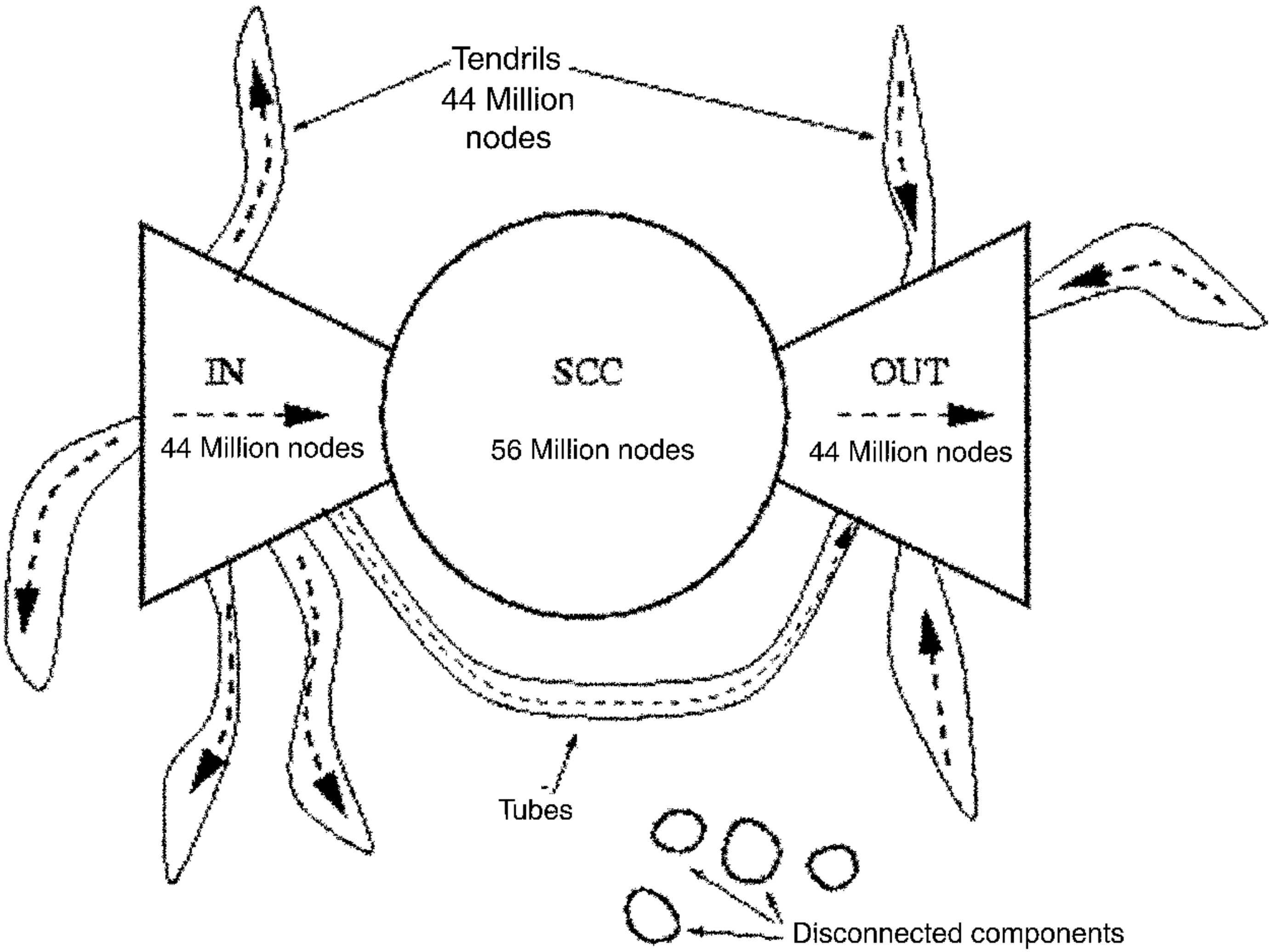
Structure of the Web

- **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: $\sim k^{-2}$
    - 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.]