CentraleSupélec

Learning in Networks

CentraleSupélec - UESTC

Fragkiskos Malliaros

Thursday, July 26, 2018

About Me

- Undergrad at the University of Patras, Greece
- Ph.D. in CS at Ecole Polytechnique, Paris
- Postdoc researcher at UC San Diego
- Assistant Professor at CentraleSupélec (since October 2017)

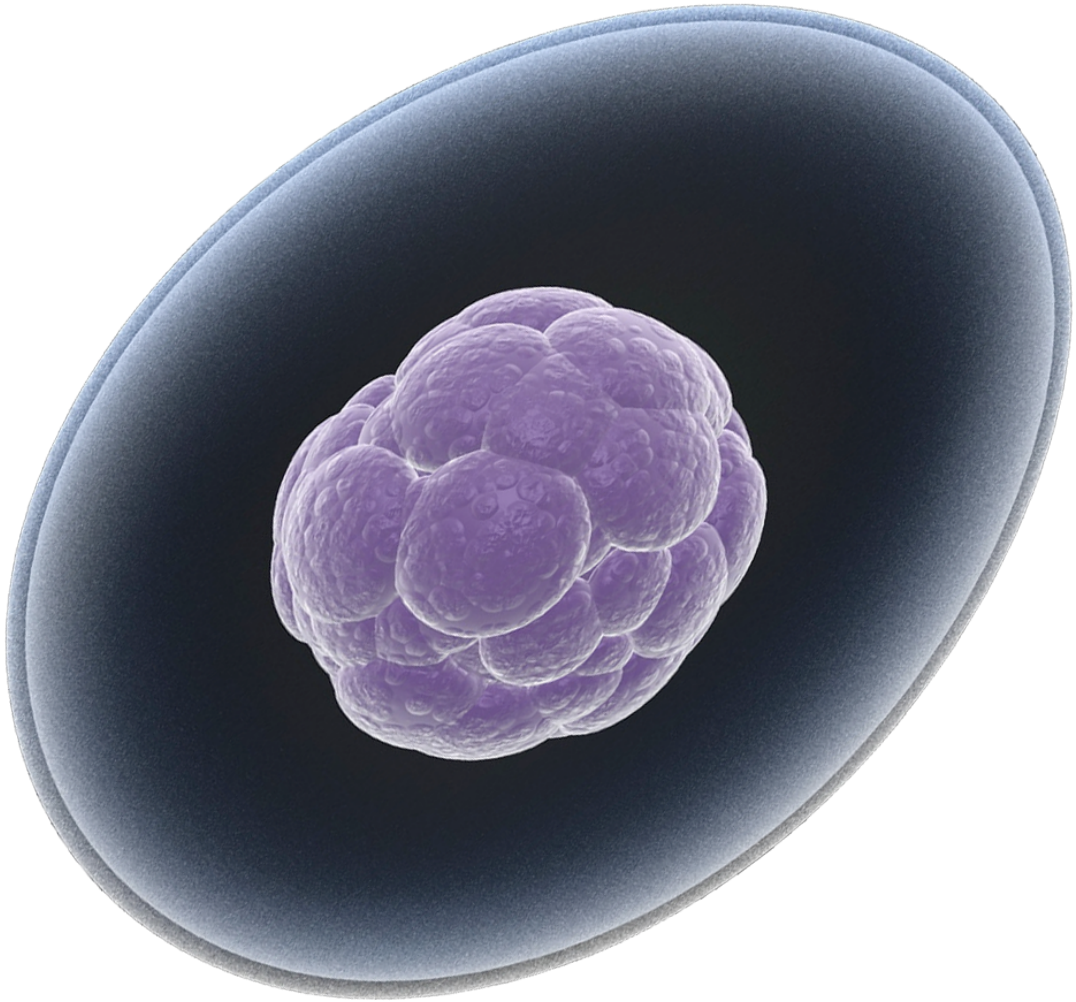
Research interests: Data science, ML, graph mining, text mining and NLP

Material of Today's Lecture

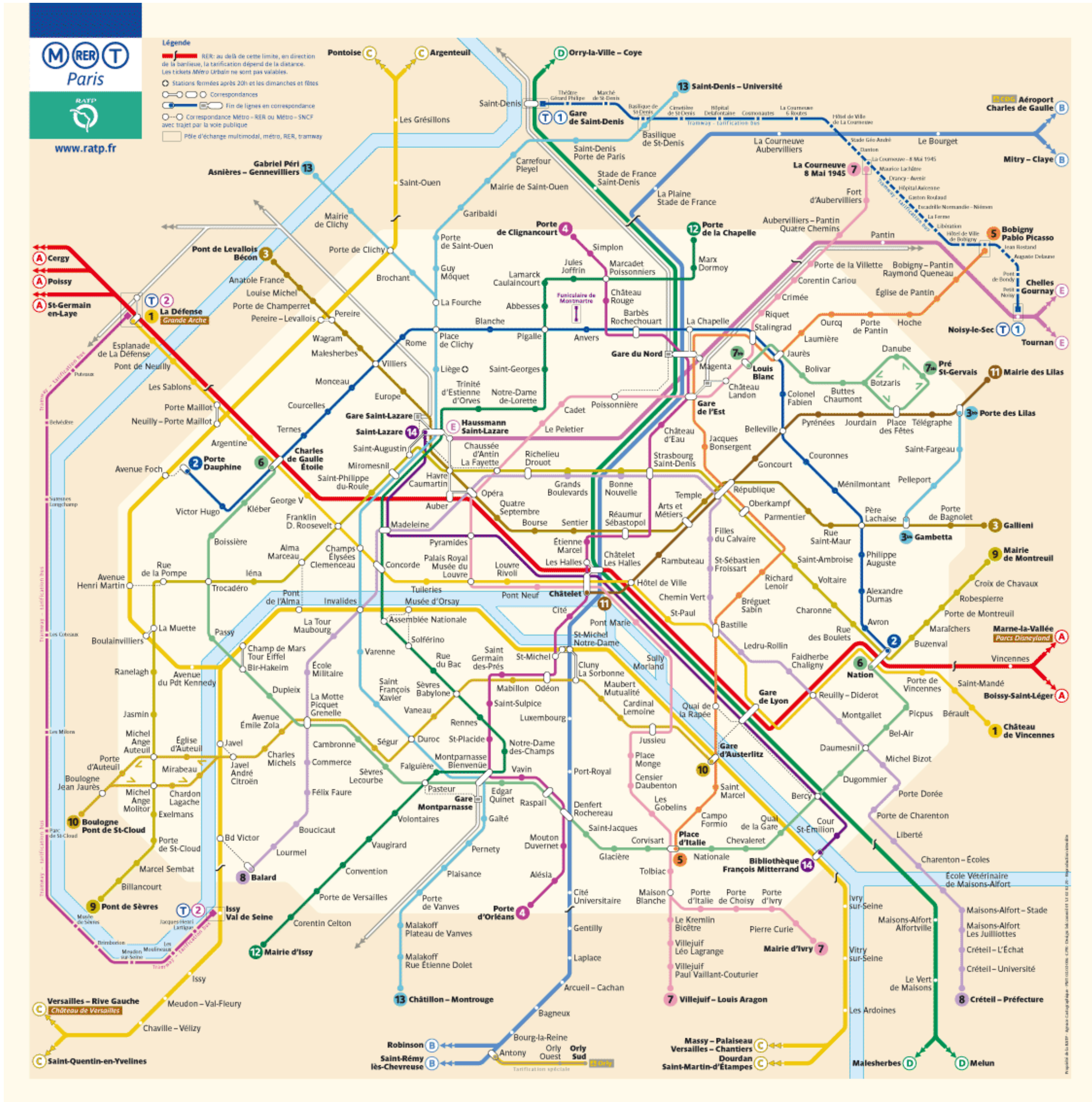
http://fragkiskos.me/networks_UESTC.pdf

What do the following things have in common?

World economy



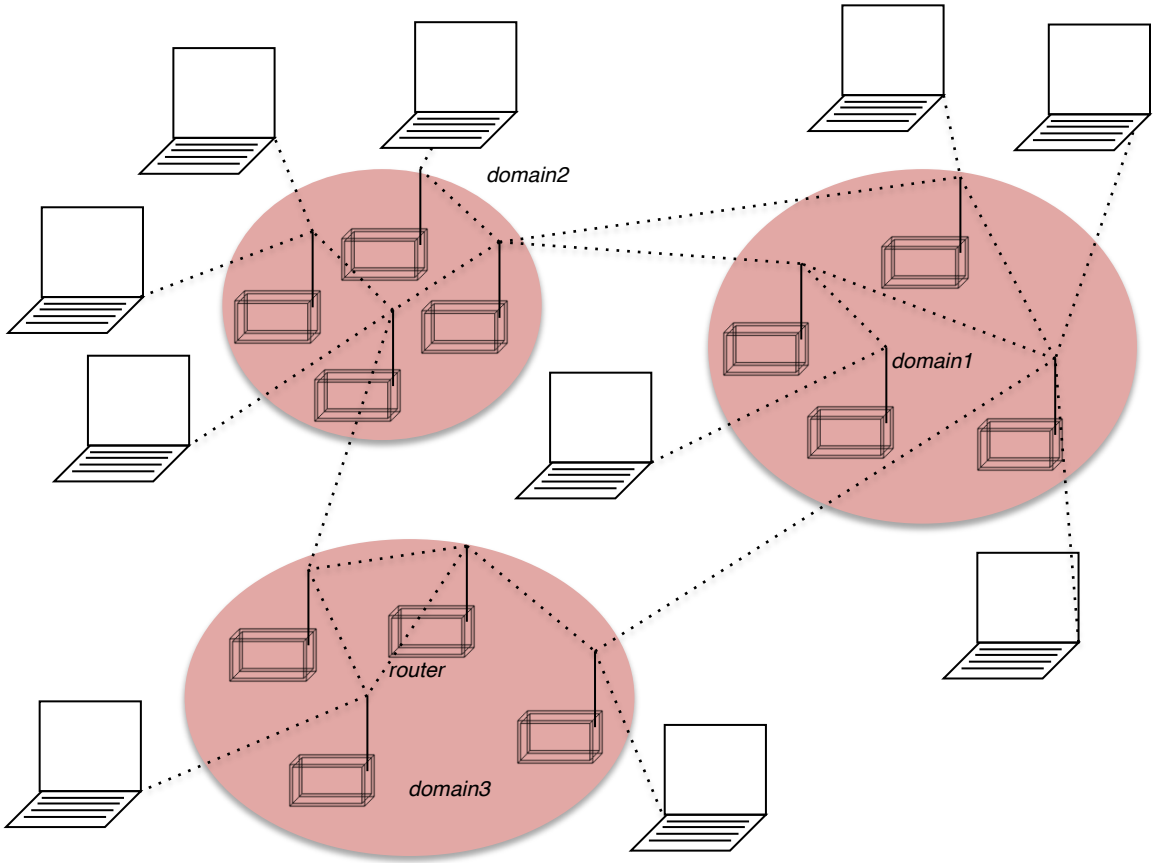
Human cell



Paris
RATP
www.ratp.fr
Légende
Pontoise
Argenteuil
Orry-la-Ville – Coye
Saint-Denis – Université
Aéroport Charles de Gaulle
Mitry – Claye
Cergy
Poissy
St-Germain en-Laye
La Défense
Grande Arche
Bobigny Pablo Picasso
Chelles Gournay
Tournan
Mairie des Lilas
Porte des Lilas
Gallieni
Mairie de Montreuil
Marne-la-Vallée
Parcs Disneyland
Boissy-Saint-Léger
Château de Vincennes
Boulogne Pont de St-Cloud
Pont de Sèvres
Balard
Mairie d'Issy
Châtillon – Montrouge
Villejuif – Louis Aragon
Mairie d'Ivry
Créteil – Préfecture
Versailles – Rive Gauche
Château de Versailles
Saint-Quentin-en-Yvelines
Robinson
Saint-Rémy lès-Chevreuse
Malesherbes
Melun

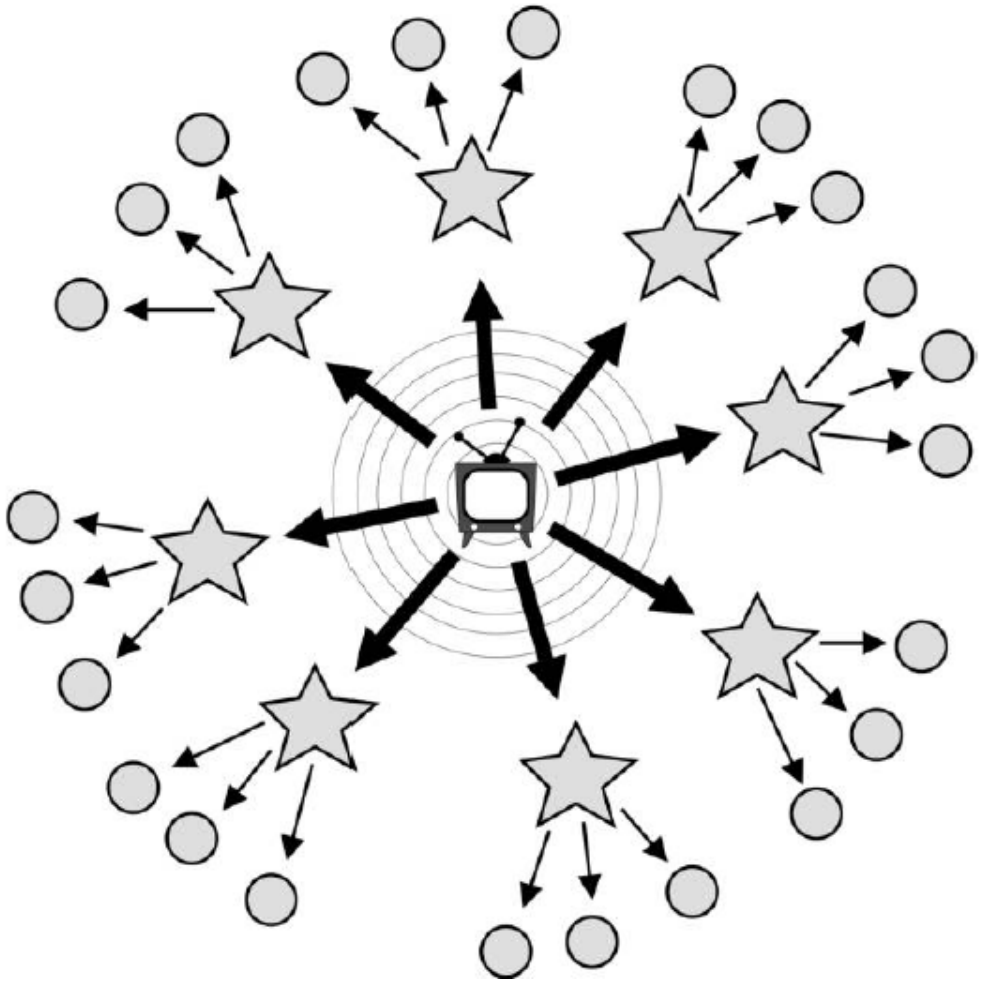
Railroads

Brain

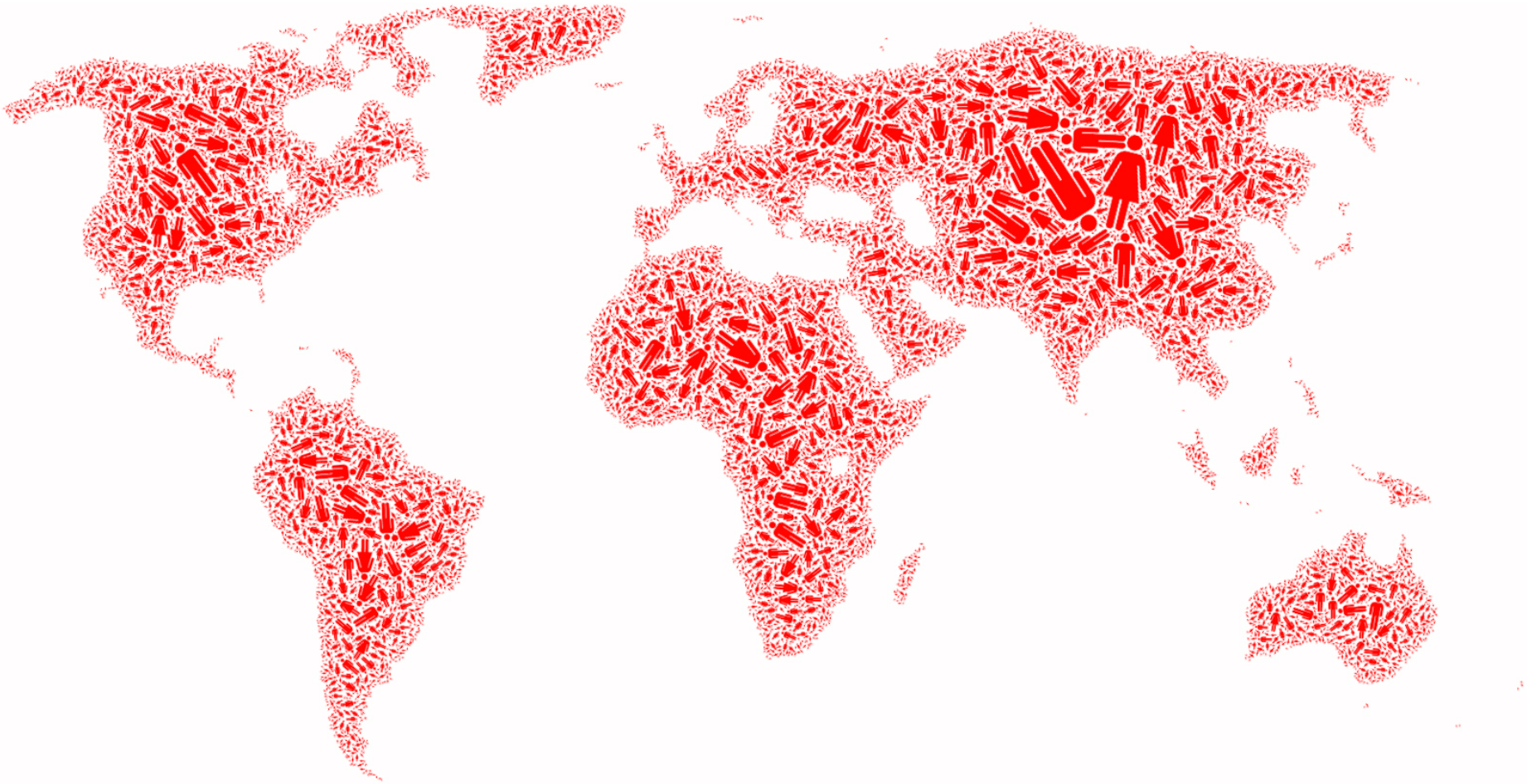


Internet

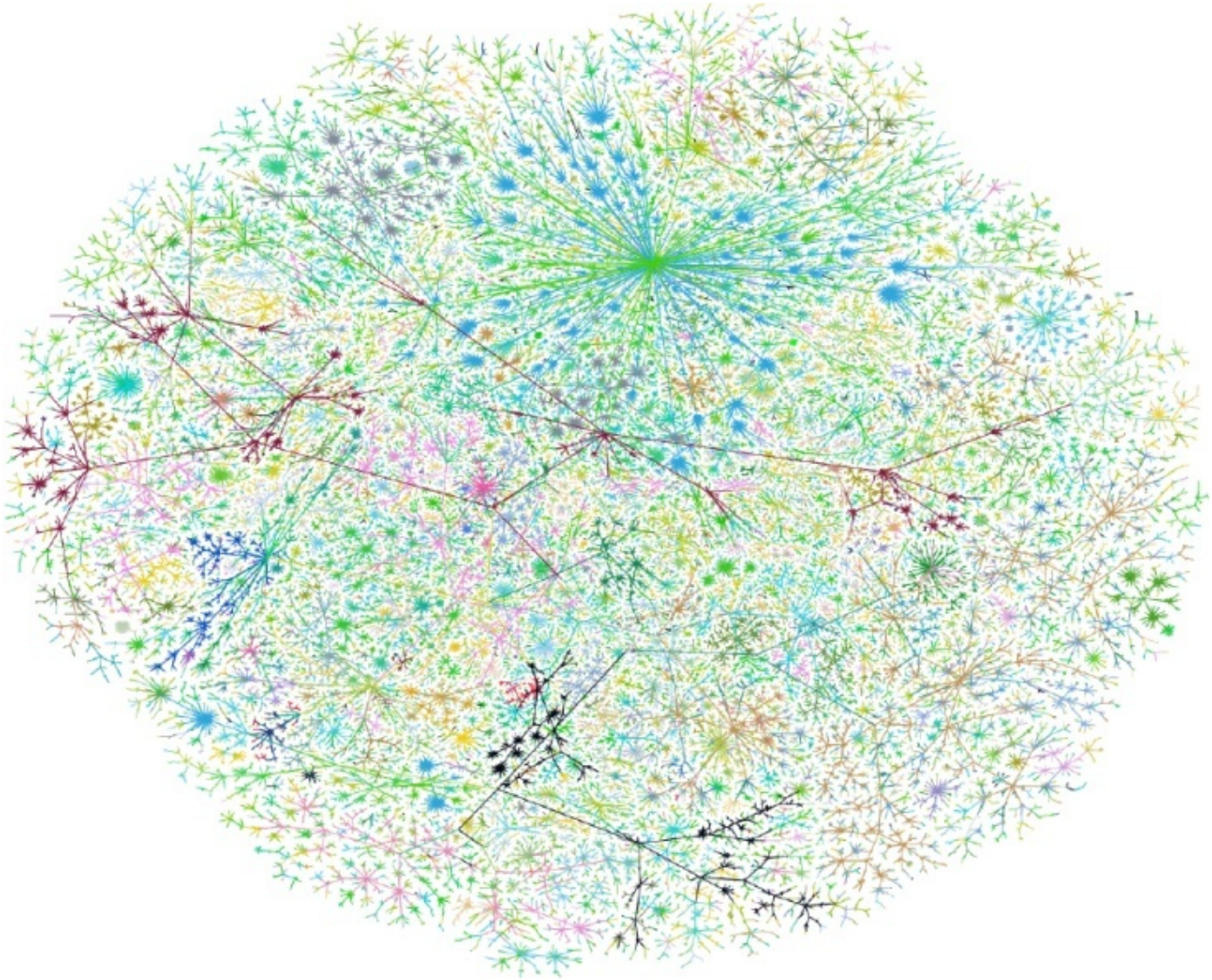
Friends & Family



Media & Information

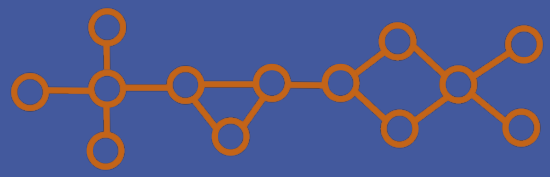


Society

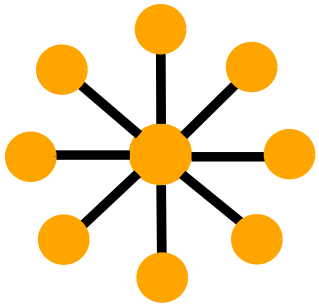


The Network!

Networks



Networks allow to model relationships between entities

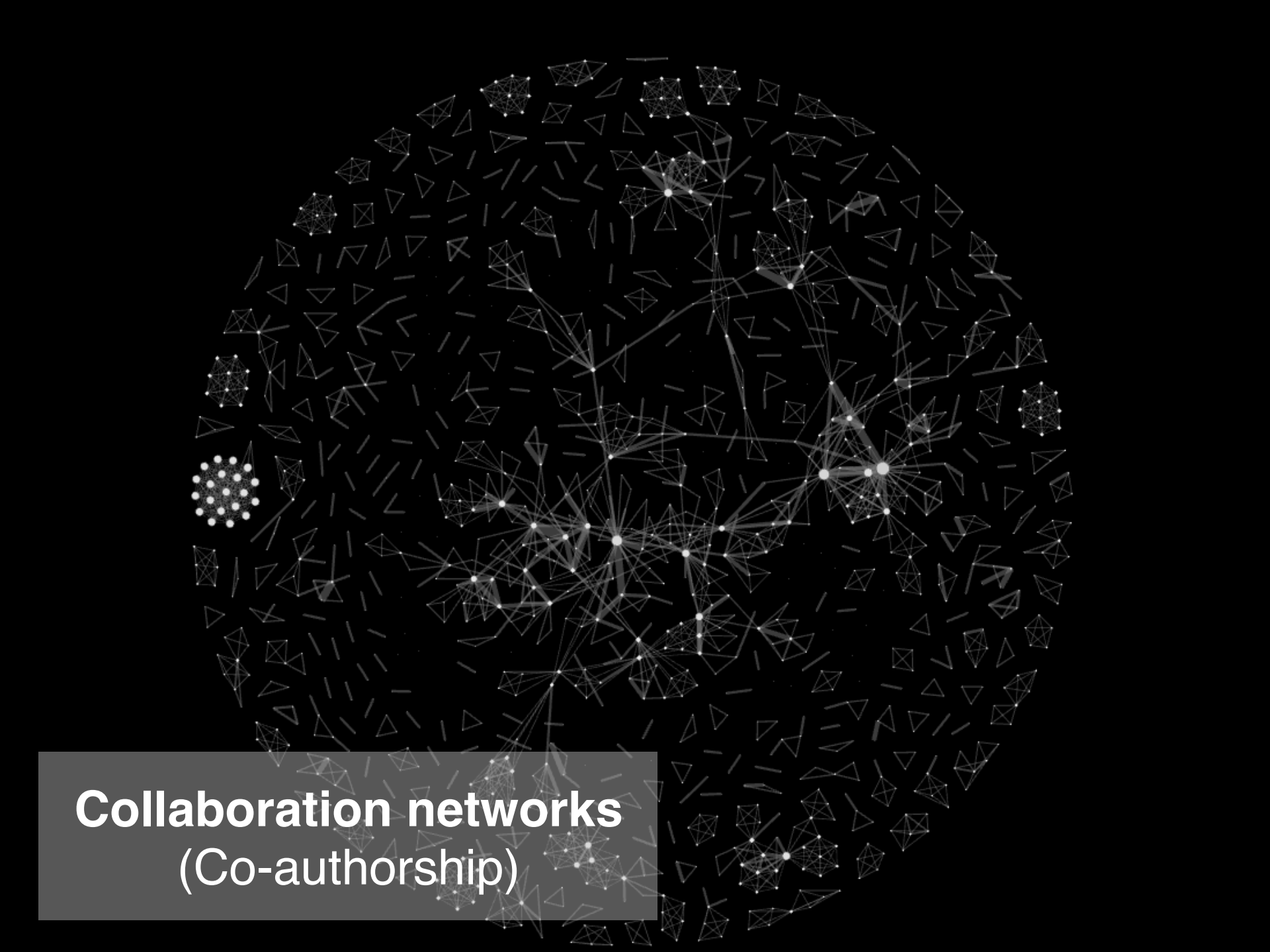


General-purpose language

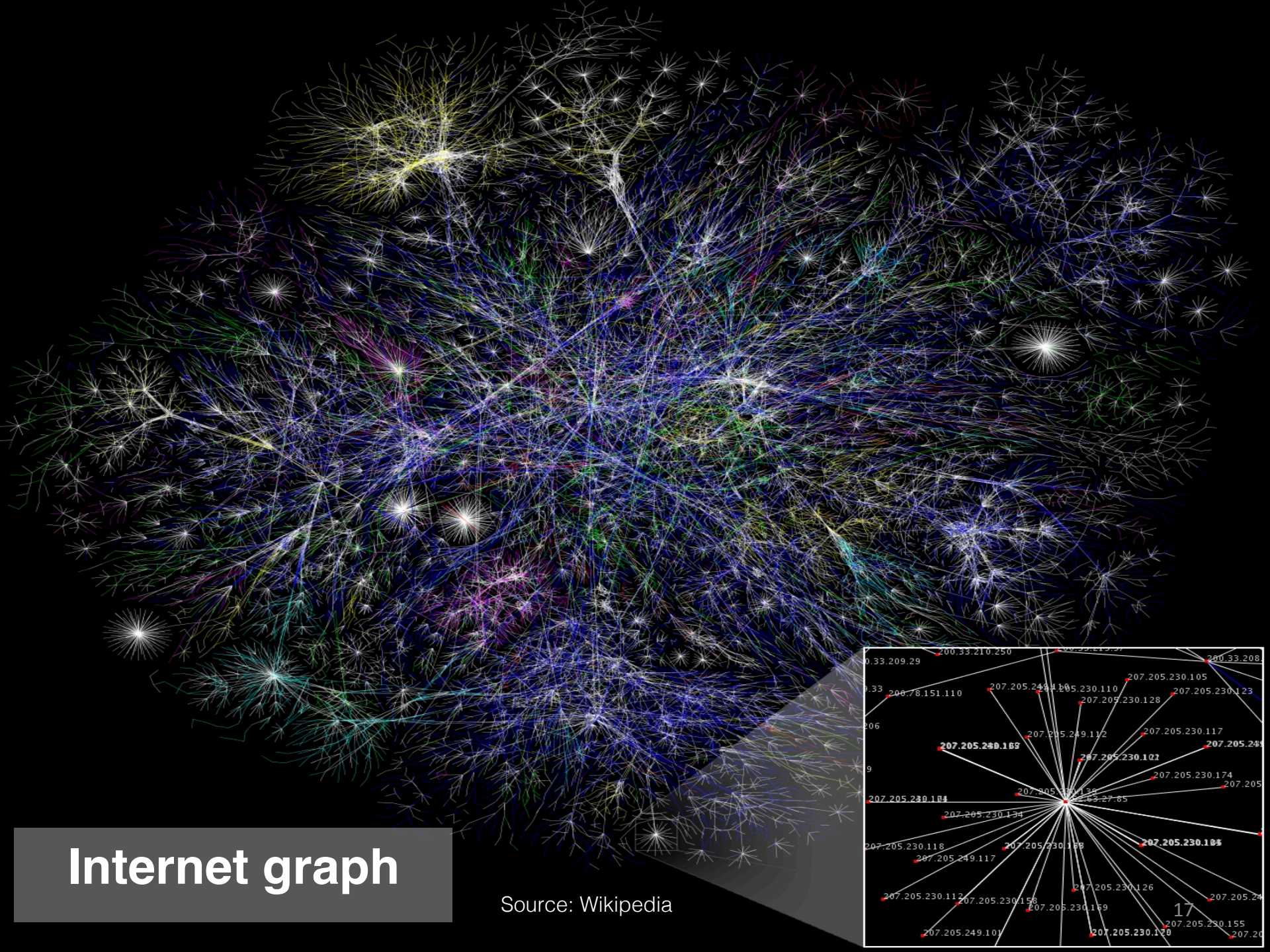
for describing real-world systems

Facebook
2.07 Billion users (Q2 of 2017)

Source: https://www.facebook.com/zuck

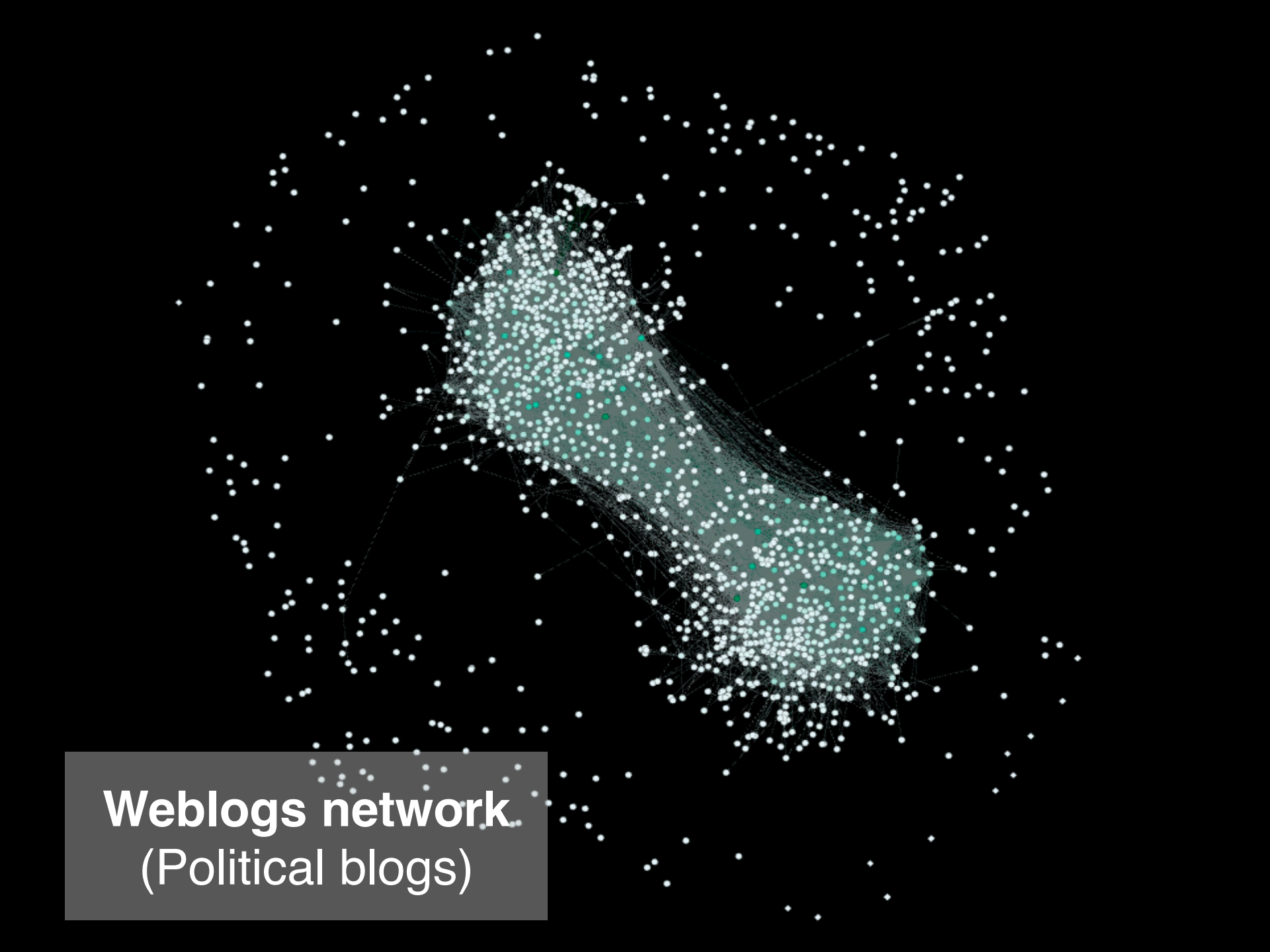


Collaboration networks
(Co-authorship)

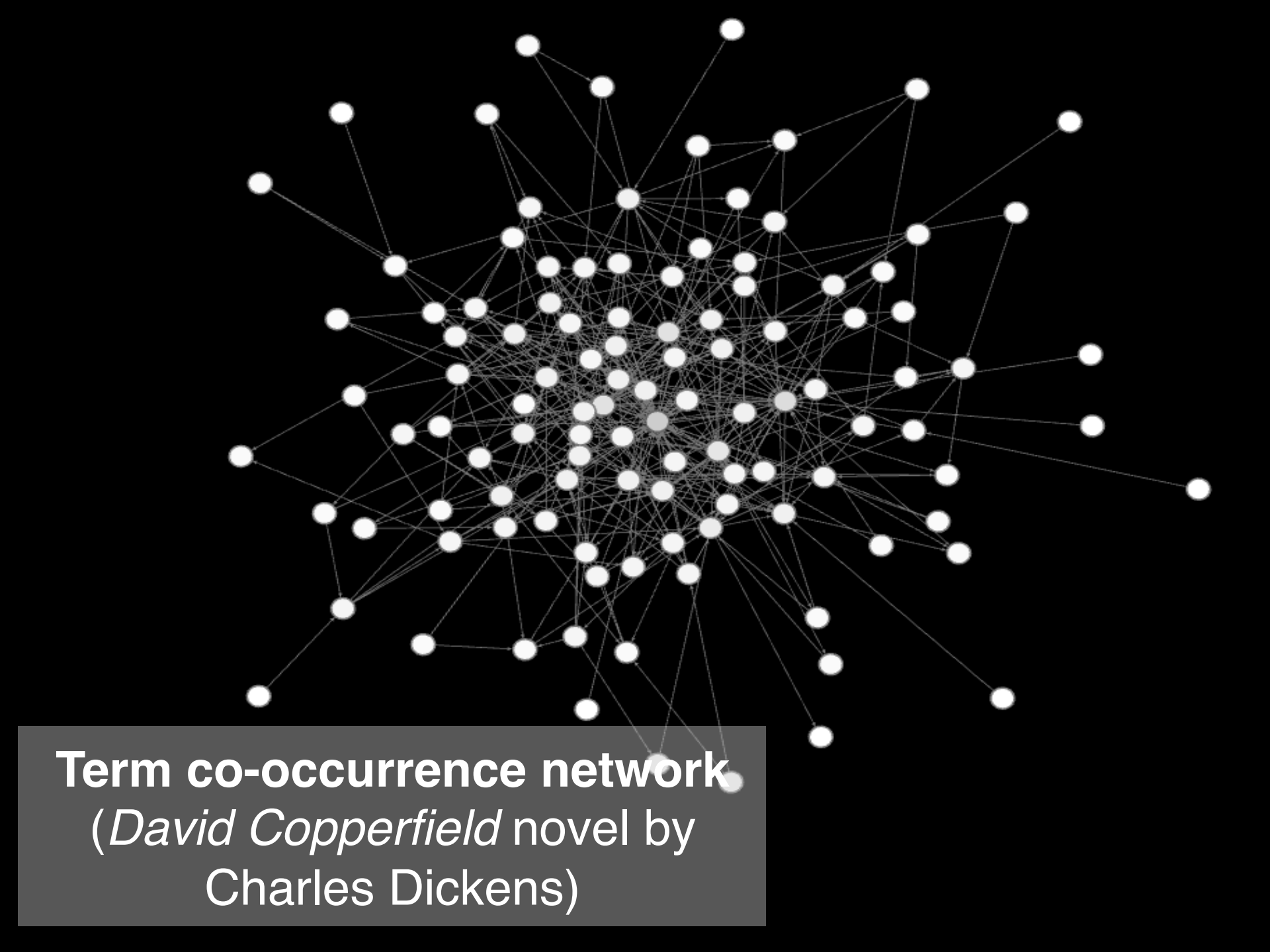


Internet graph

Source: Wikipedia

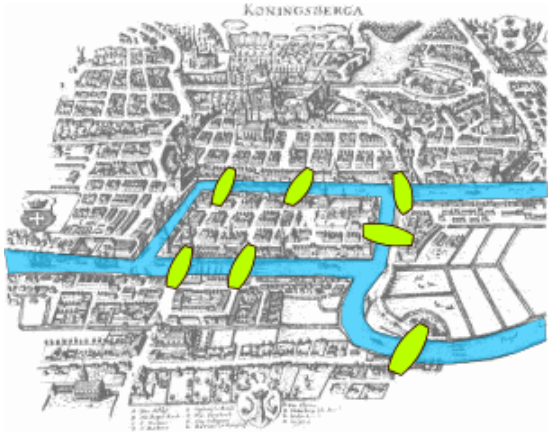


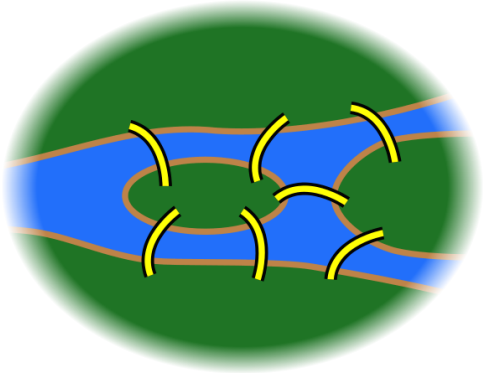
Weblogs network
(Political blogs)

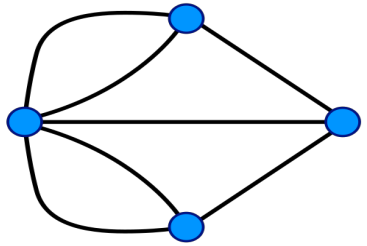


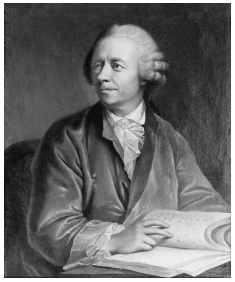
Term co-occurrence network
(*David Copperfield* novel by
Charles Dickens)

Infrastructure Networks







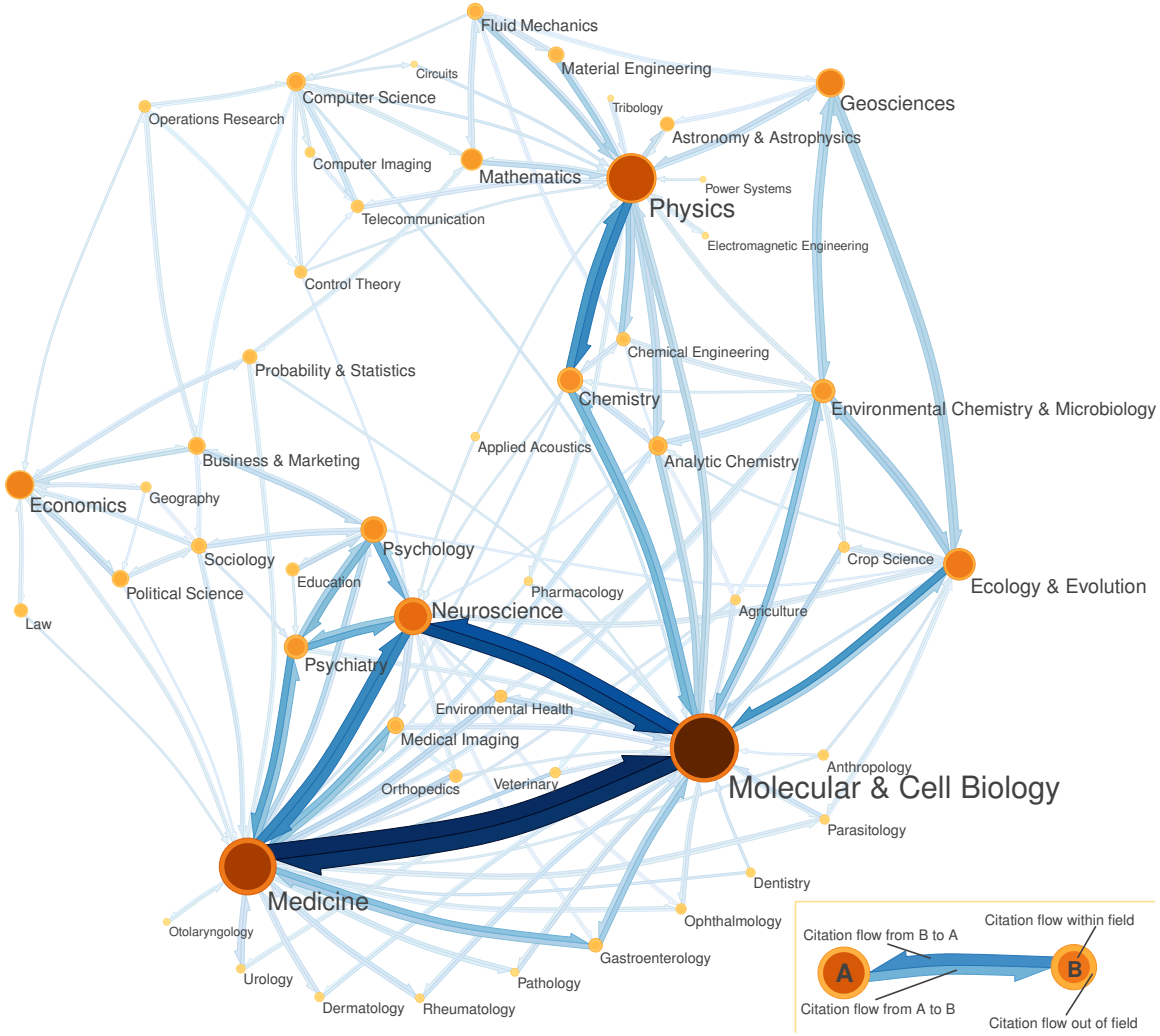


Seven Bridges of Königsberg (Kaliningrad) **[Euler, 1735]**

Devise a walk through the city that would cross each of those bridges once and only once

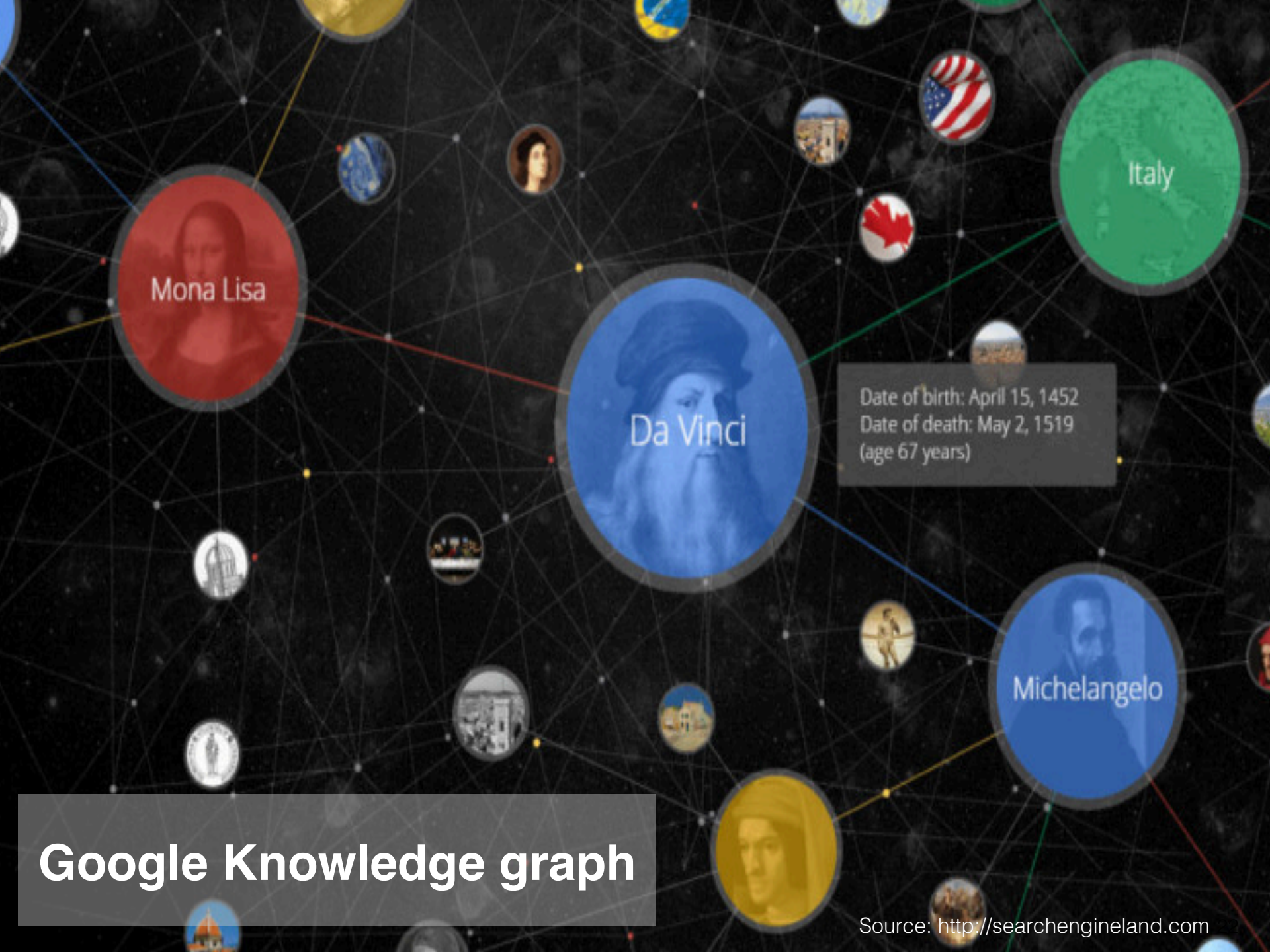
Source: https://en.wikipedia.org/wiki/Seven_Bridges_of_Konigsberg

Information Networks



Citation networks and Map of science

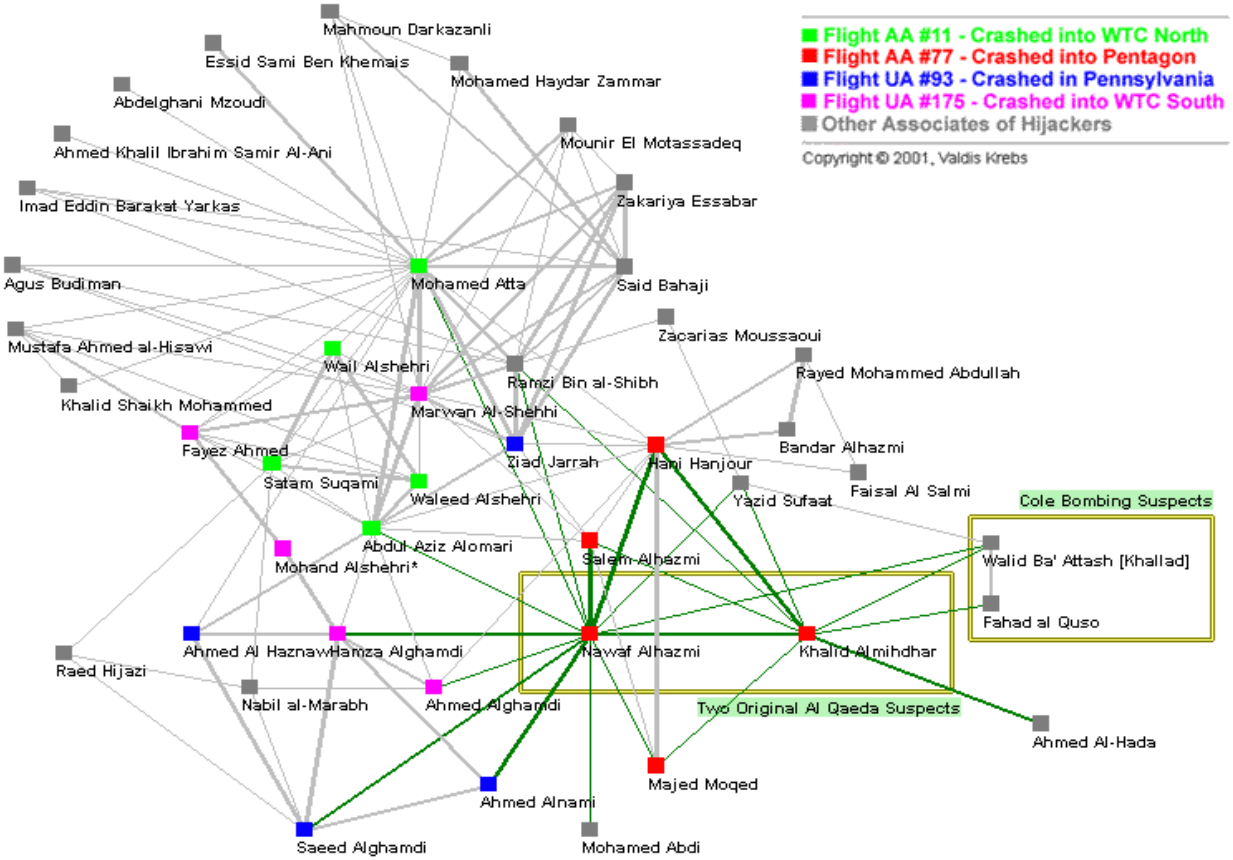
[Rosvall and Bergstrom, 2008]



Google Knowledge graph

Source: http://searchengineland.com

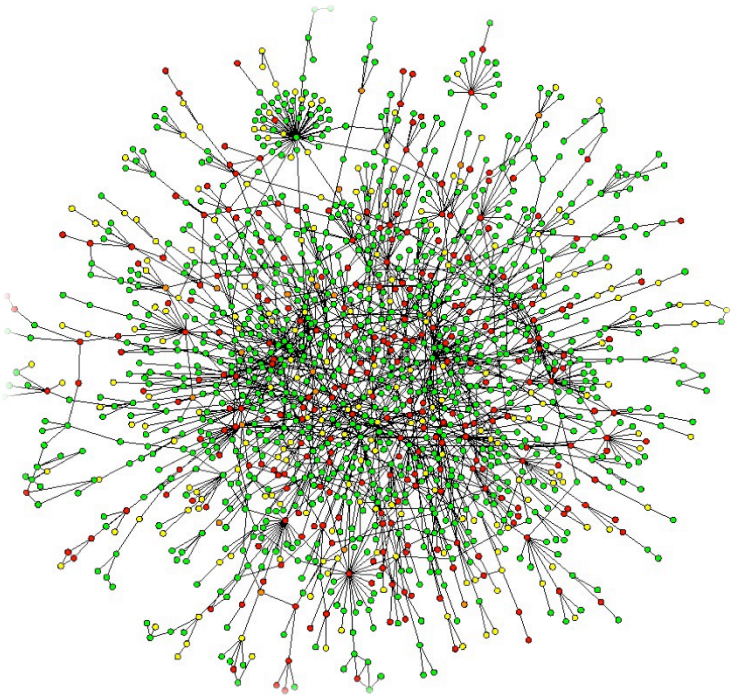
Networks of Organizations

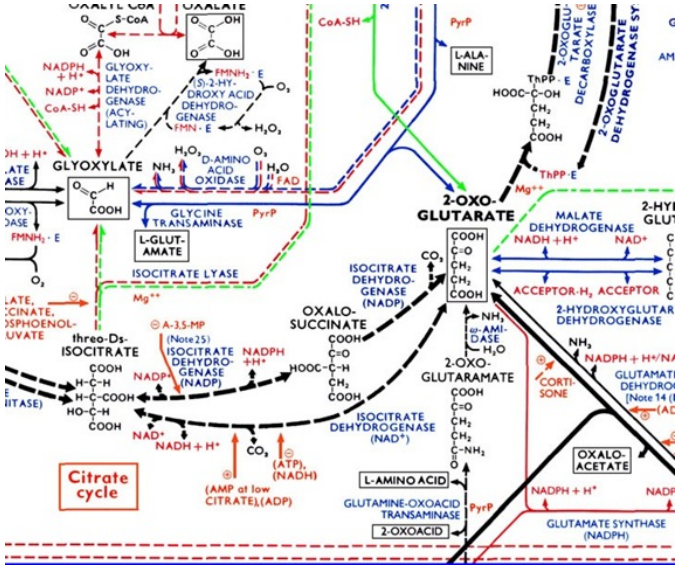


9/11 terrorist network

Source: http://www.orgnet.com/prevent.html

Biological Networks





Protein-Protein Interaction Networks:

Nodes: Proteins

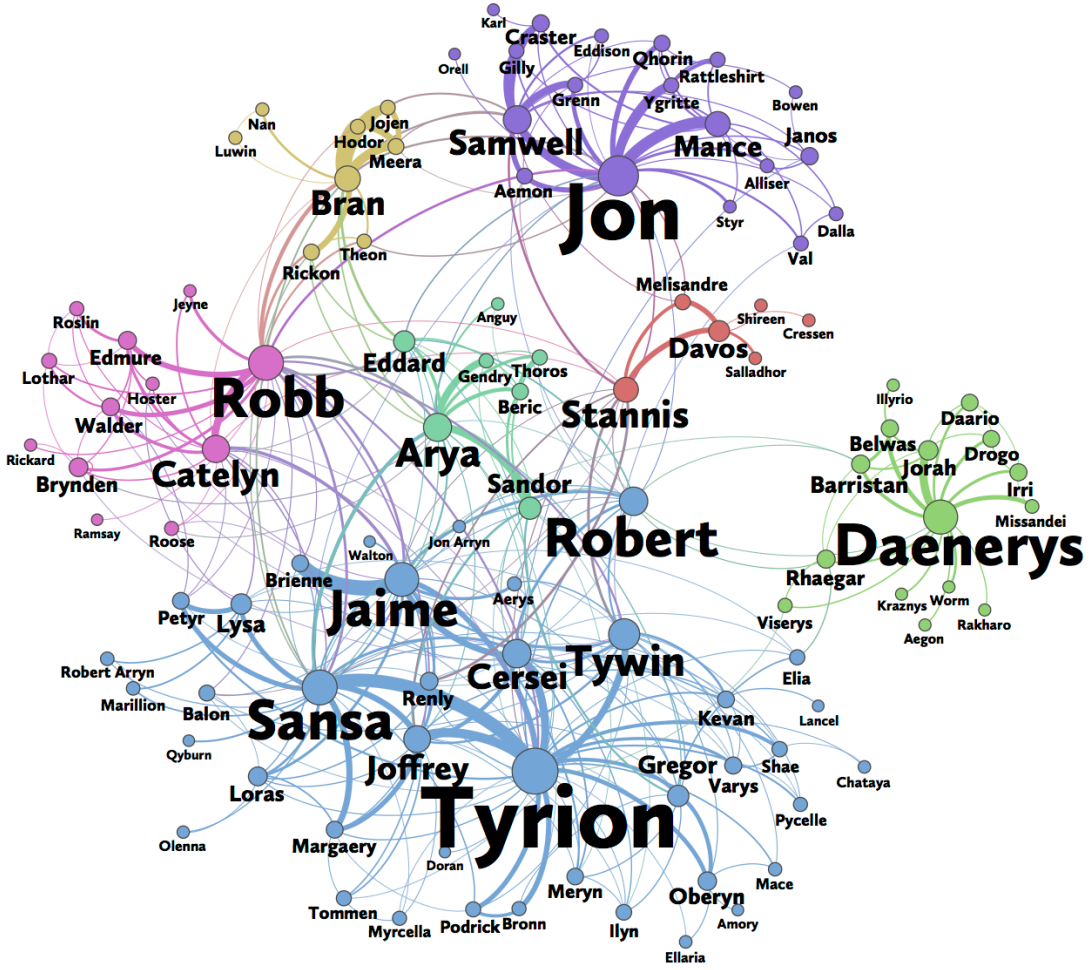
Edges: ‘physical’ interactions

Metabolic networks:

Nodes: Metabolites and enzymes

Edges: Chemical reactions

What Else?



In fact, anything that captures relationships between entities can be modeled as graph (any 3-way join in the DB community)

Why should I care about networks?

Why Graphs? Why Now?

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect
- **Shared vocabulary (representation) between fields**
  - Computer Science, Engineering, Social Sciences, Physics, Economics, Statistics, Biology, …
  - Cross-disciplinary topic
- **Availability of big and rich data**
  - Web/mobile, bio, health, and medical
  - Computational challenges
- **Impact**
  - Social networking and social media, recommender systems, drug design, neuroscience, epidemiology, …

Web – The Lab of Humanity

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

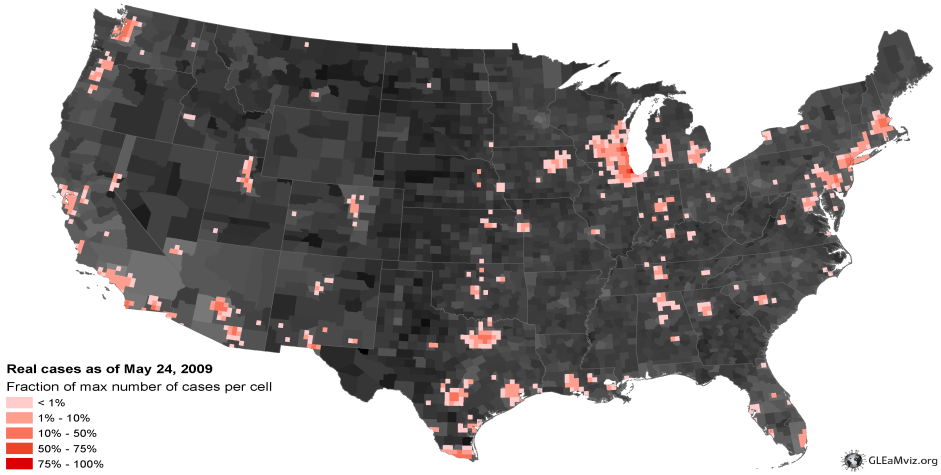
Networks: Economic Impact

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

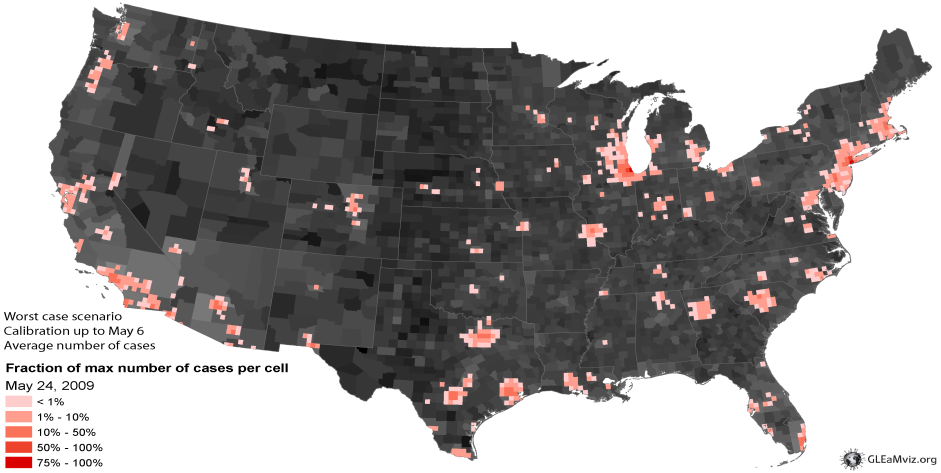
- **Google**
  Market cap: $394 billion
  (1y ago it was 300b)
- **Cisco**
  Market cap: $130 billion
  (1y ago it was 100b)
- **Facebook**
  Market cap: $201 billion
  (1y ago it was 114)

Networks: Healthcare Impact

Predicting epidemics (e.g., the 2009 H1N1 pandemic)



Real



Predicted

What Can we do With Networks?

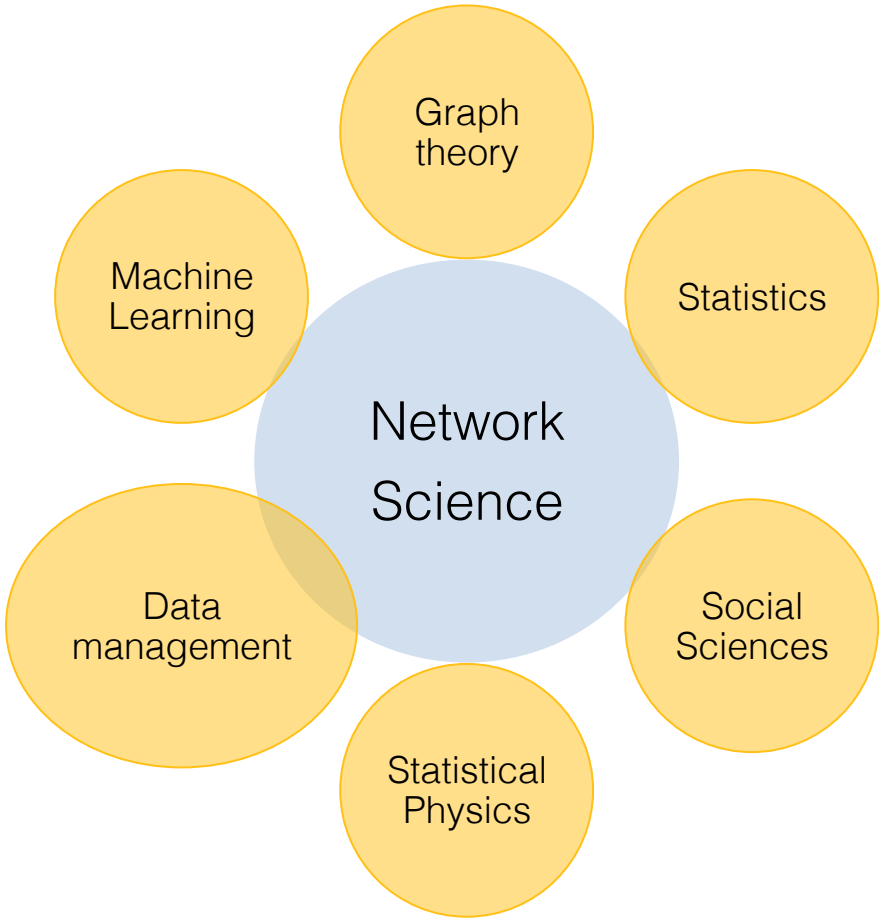
- **Communication networks**
  - Intrusion detection, fraud detection
  - Churn prediction (e.g., telecommunication providers)

- **Social networks**
  - Link prediction, friend recommendation
    - E.g., Facebook, LinkedIn
  - Social circle detection, community detection
  - Social recommendations
  - Identifying influential nodes, information spreading, influence

- **Information networks**
  - Navigational aids

Network Science Analytics

Discovering, analyzing and making sense of graph data



About this course

Reasoning about Networks (1/2)

- **What do we hope to achieve from studying networks?**
  - Patterns and statistical properties of network data
  - Design principles and models
  - Understand why real networks are organized the way they are
    - Predict behavior of networked systems
  - Utilize the extracted knowledge in practical applications

Reasoning about Networks (2/2)

- **How do we reason about networks?**
  - Empirical analysis: 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 (scale and complexity of the underlying networks)
    - Unsupervised vs. supervised algorithms

Software Tools

- We strongly advise to use Python
  - NetworkX library
  - igraph library (also for C++ and R)
- Snap library
  - C++ and Python
- Gephi, Jung and graph-tool for network visualization

Topics in Learning in Networks

Networks or Graphs?

- Network often refers to real systems
  - Web, Social network, Internet Metabolic network

  Language: network, node, link

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

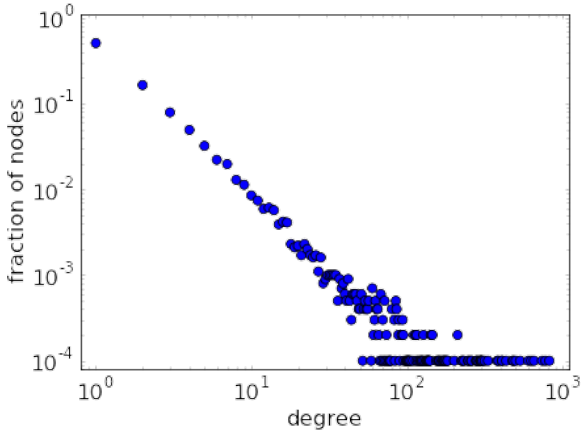
  Language: graph, vertex, edge

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

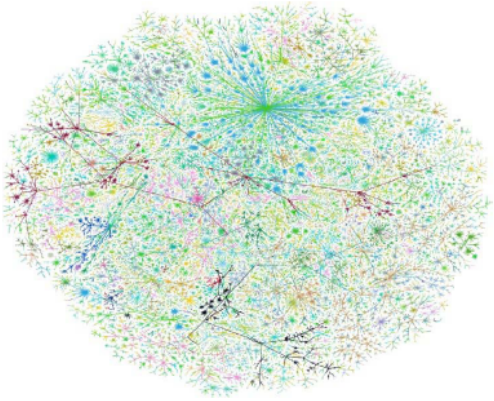
Patterns and Graph Generative Models

In this course

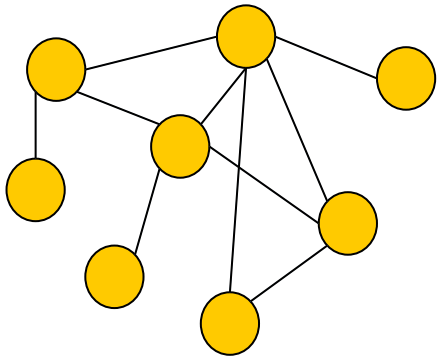
Q1: How does a real-network look like?
Q2: Properties, patterns, deviation from randomness?



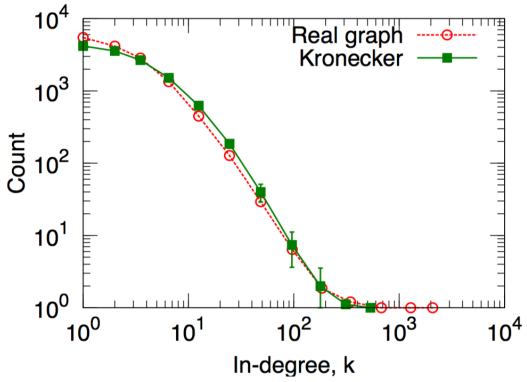
Q3: Can we generate artificial networks that are similar to real ones?



Given a real graph
(e.g., Facebook social graph)



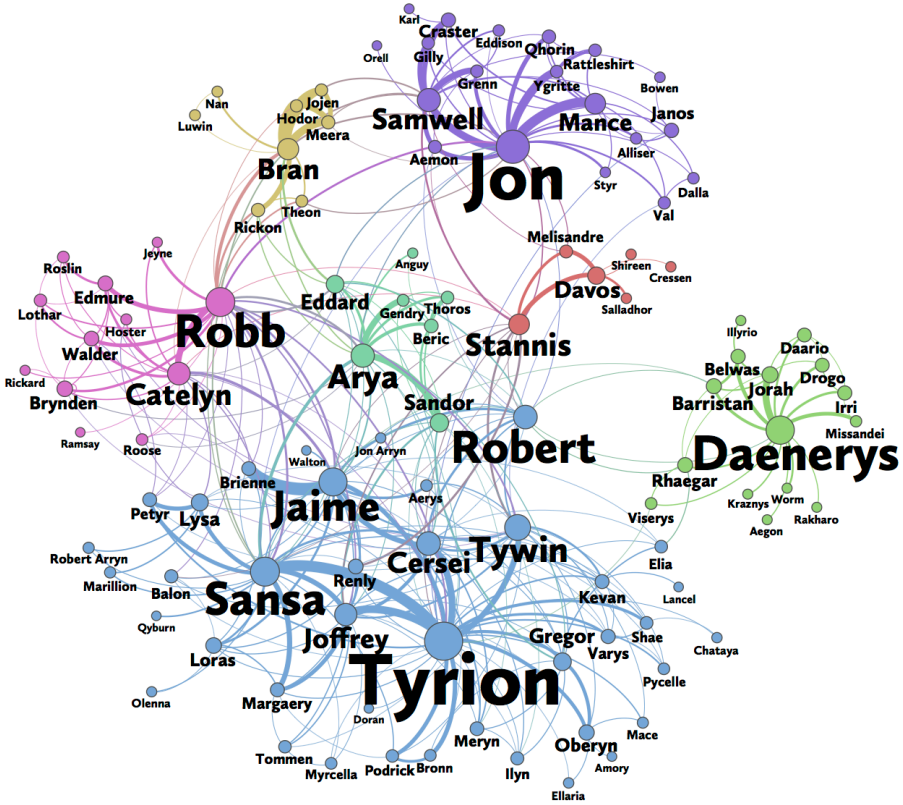
Generate an artificial graph
(e.g., model the formation process)



Fit some properties
(e.g., same degree distribution)

Centrality and Ranking in Networks

In this course



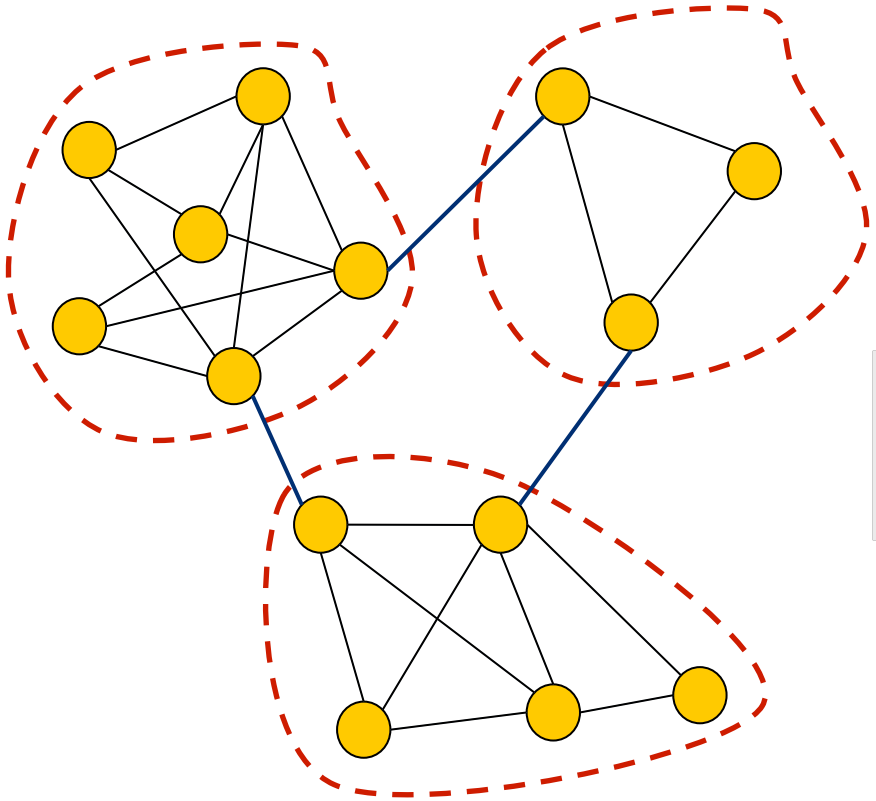
Q: How to determine the importance of a node in the graph?

- Centrality criteria (e.g., degree, closeness, betweenness)
- HITS and PageRank algorithms
- Scalability issues

Source: https://www.macalester.edu/~abeverid/thrones.html

Graph Clustering - Community Detection

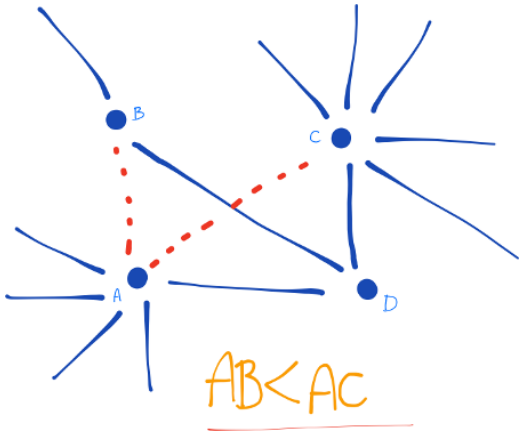
In this course



Example graph with three communities

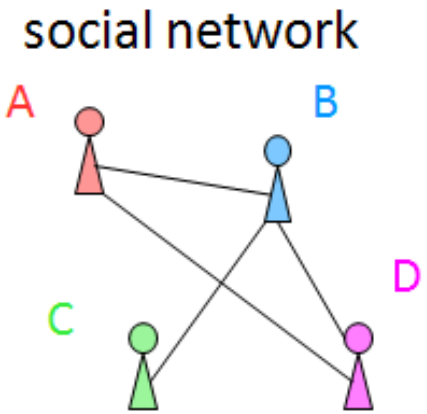
- [Models and definitions] How a community in graphs looks like?
- [Algorithms] How can we extract the inherent communities?
- [Patterns in large networks] What is happening in large-scale networks?

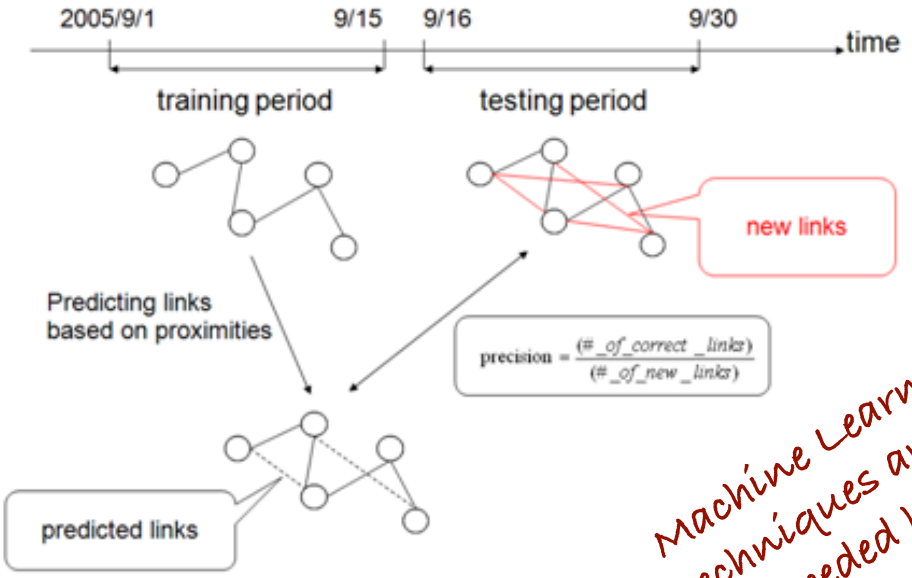
Node Similarity and Link Prediction



- How similar are two nodes in the graph?
- Can we utilize this similarity to predict missing edges (links)?

Applications?



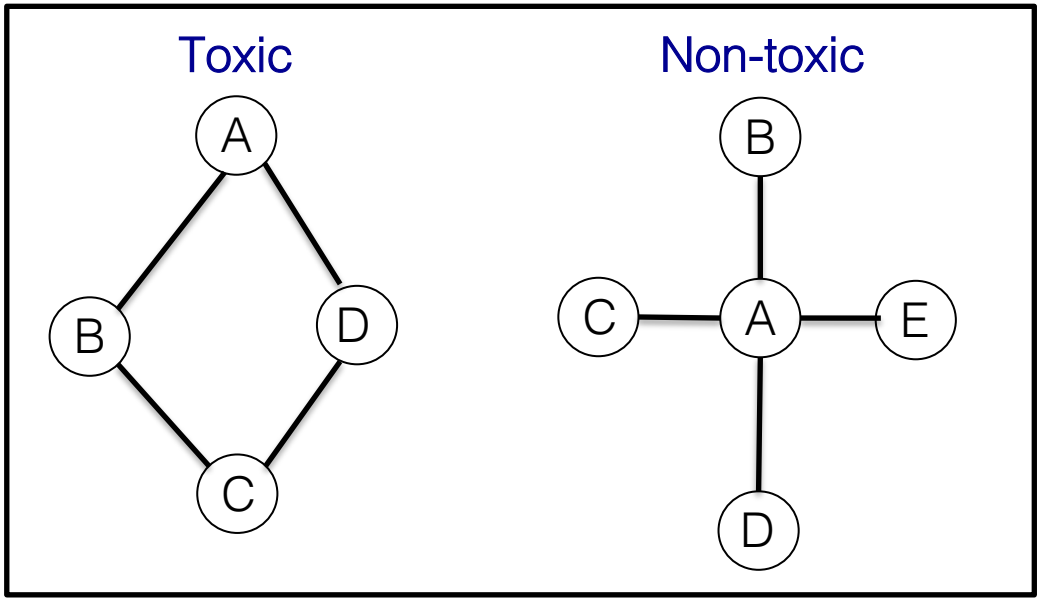


Machine Learning techniques are also needed here!

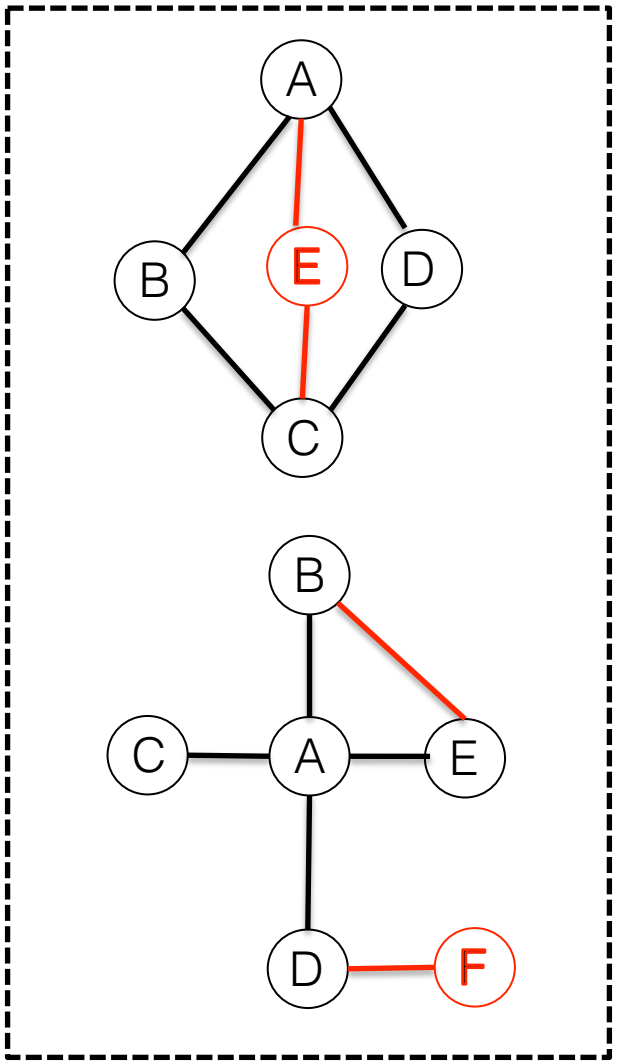
Source: http://be.amazd.com/link-prediction/
http://www.net.c.titech.ac.jp/research.html

Graph Similarity and Classification

Dataset of known molecules

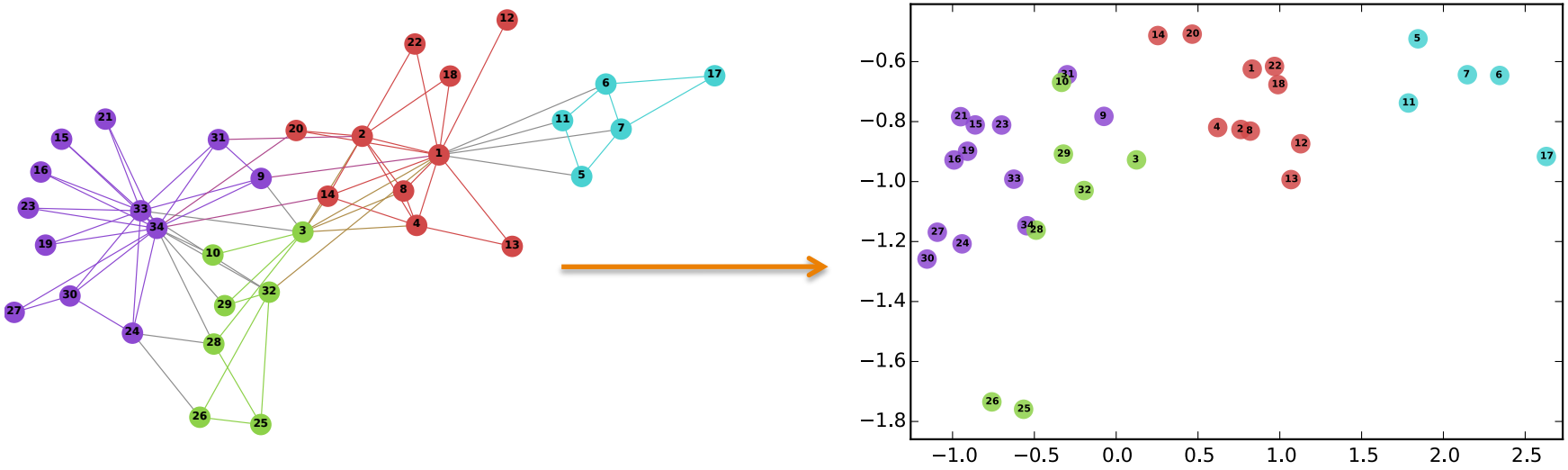


Unknown molecules



Task:
Given a set of molecules that are either toxic or non-toxic
Predict the class of unknown molecules

Representation Learning in Graphs



Input graph

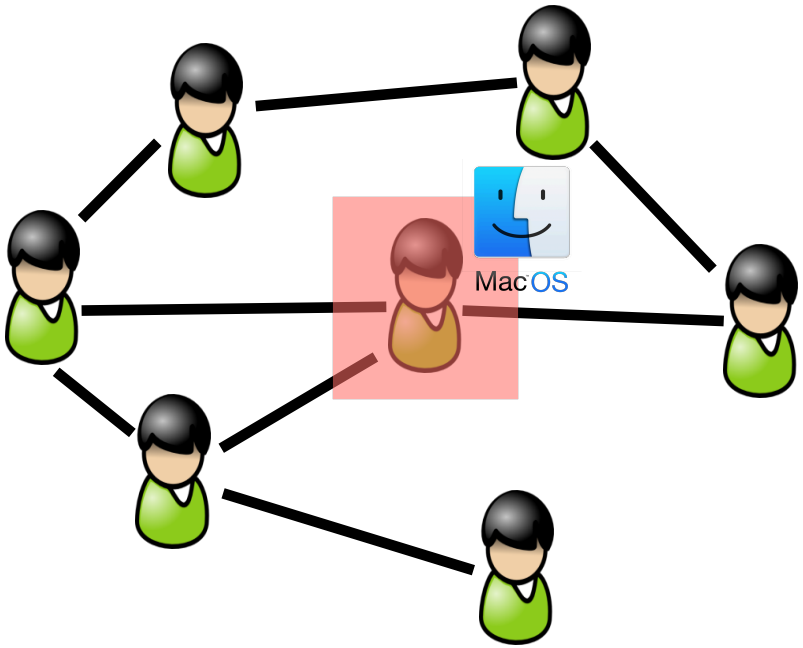
Learn latent representation

Applications: clustering, classification, link prediction, …

Also known as node embedding techniques (e.g., DeepWalk, node2vec)

[Perozzi et al., KDD '14]

Influential Nodes and Influence Maximization



[Viral marketing] How to organize an effective product promotion campaign?

[Opinion dynamics] How do opinions/rumors spread?

[Epidemiology] How do viruses/ diseases propagate?

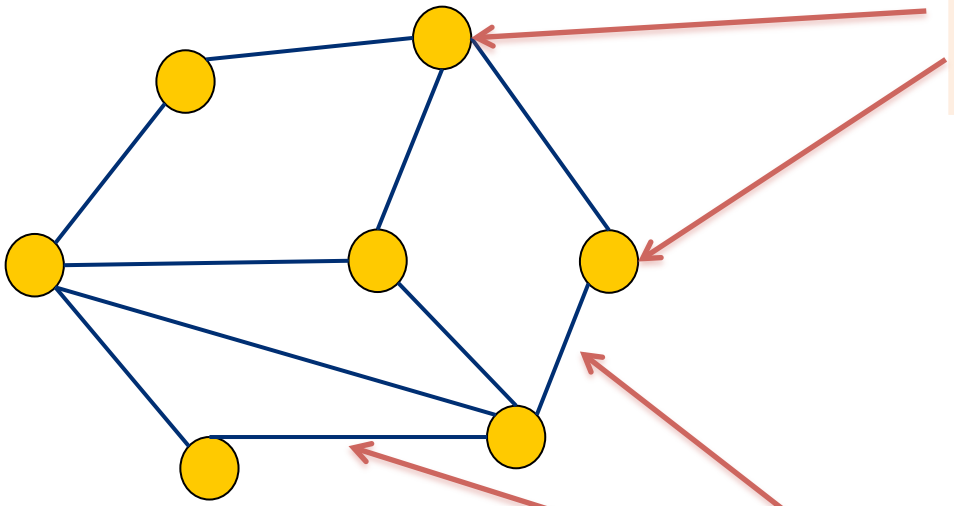
- Detection of influential nodes (spreaders) in networks
- Influence maximization algorithms
- Epidemic processes in networks

[Prakash, Ramakrishnan, KDD '16]

Basic graph-theoretic concepts and definitions

Graphs and Networks

Graphs: modeling dependencies



Nodes (or vertices)
(objects/entities)

Edges (or links)
(interconnections)

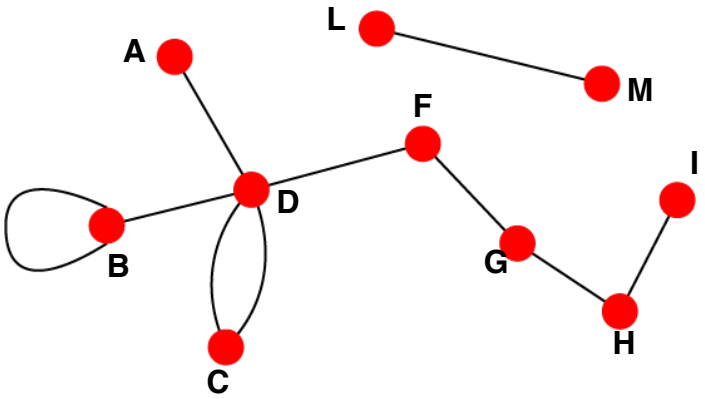
$G = (V, E)$
(network or graph)

$n = |V|$ is the number of nodes
$m = |E|$ is the number of edges

Undirected vs. Directed Networks

Undirected

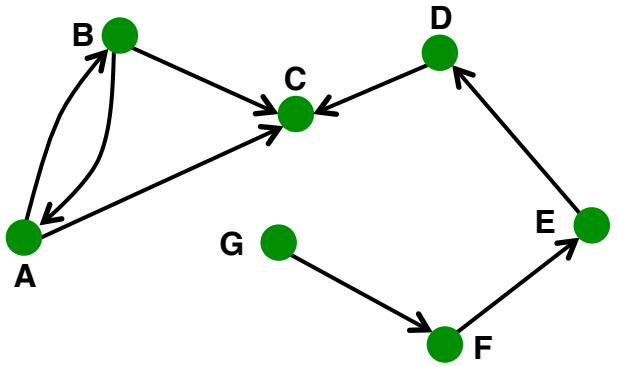
- **Links:** undirected (symmetrical, reciprocal)



- Examples
  - Collaborations
  - Friendship on Facebook

Directed

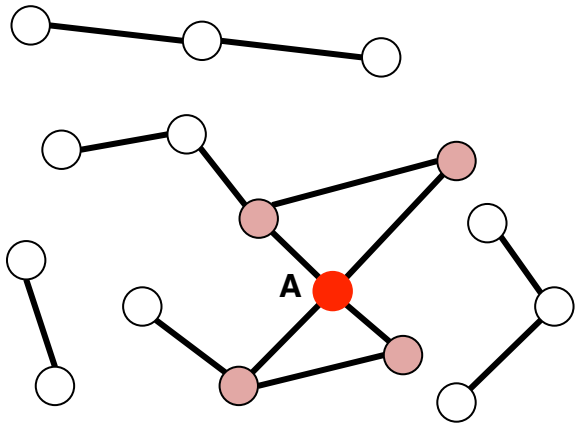
- **Links:** directed (arcs)



- Examples
  - Phone calls
  - Following on Twitter

Node Degree

Undirected



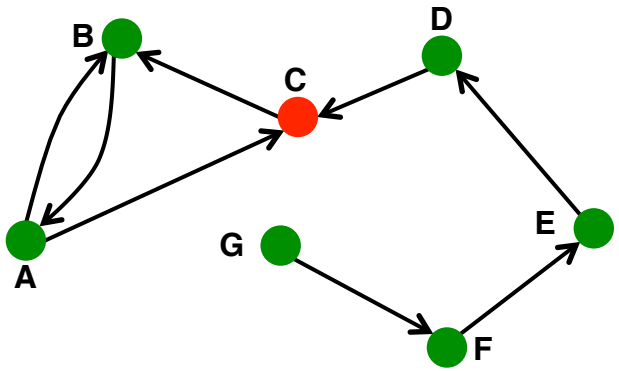
Node degree k_i: the number of edges adjacent to node i

$$k_A = 4$$

Average degree:

$$\bar{k} = \langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2|E|}{n}$$

Directed



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

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

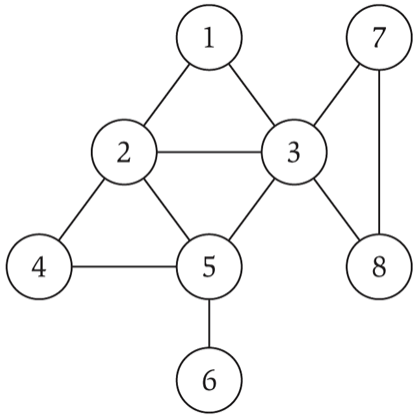
Source: Node with $k^{in} = 0$

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

Average: $\bar{k^{in}} = \bar{k^{out}}$

Graph Representation: Adjacency Matrix

- A graph can be represented by the adjacency matrix A
  - Matrix of size $n \times n$, where $n=|V|$ is the number of nodes
  - $A_{ij} > 0$, if i and j are connected
  - $A_{ij} = 0$, if i and j are not connected
  - In case of **unweighted** graphs, $A_{ij} = 1$, if (i, j) is an edge of the graph
  - Space proportional to n^2



Undirected graph

Node indexing

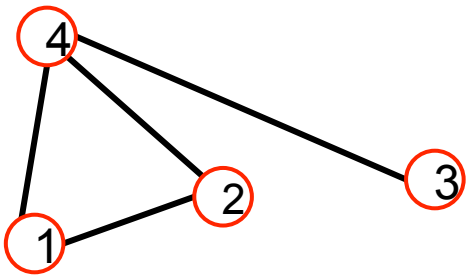
Node indexing

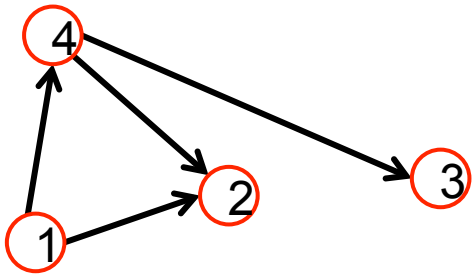
	1	2	3	4	5	6	7	8
1	0	1	1	0	0	0	0	0
2	1	0	1	1	1	0	0	0
3	1	1	0	0	1	0	1	1
4	0	1	0	0	1	0	0	0
5	0	1	1	1	0	1	0	0
6	0	0	0	0	1	0	0	0
7	0	0	1	0	0	0	0	1
8	0	0	1	0	0	0	1	0

Adjacency matrix

More matrix representations in a while

Adjacency Matrix





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

$A_{ij} = 0$ otherwise

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

Undirected graph

Symmetric matrix

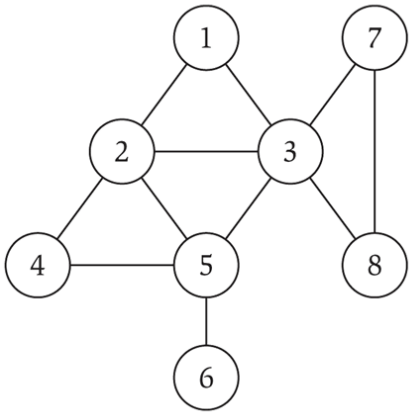
$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

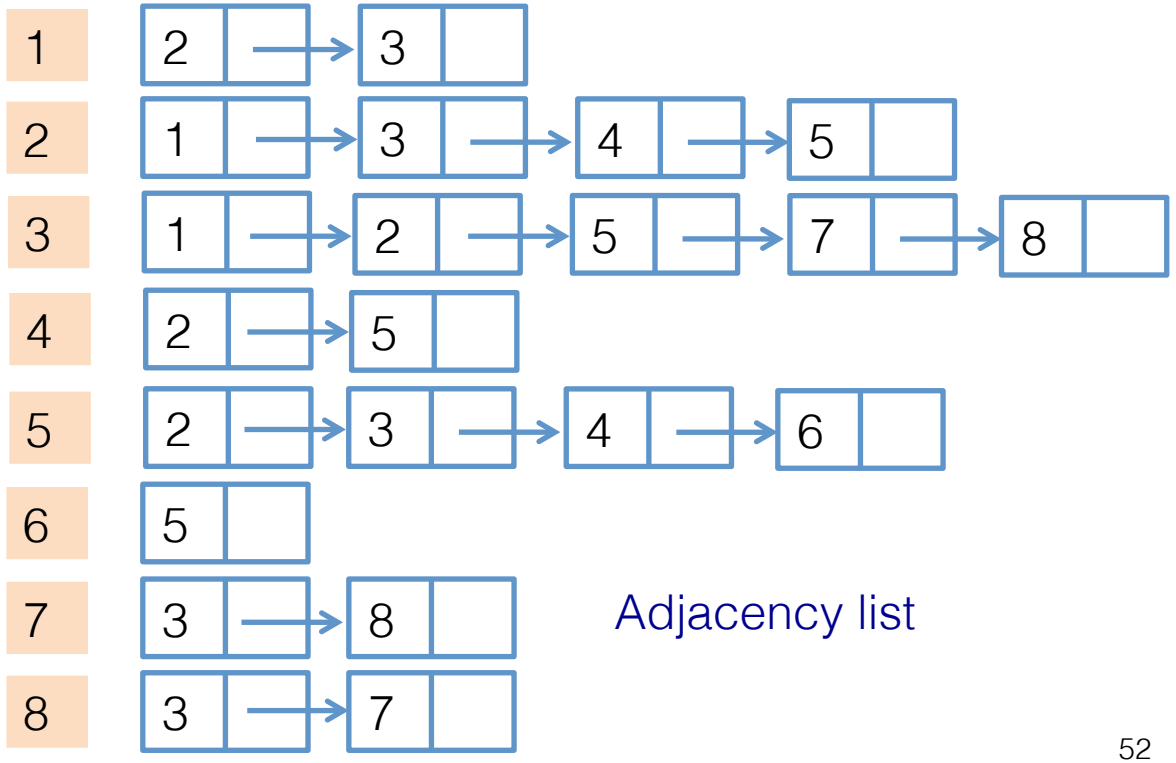
Directed graph

Nonsymmetric matrix

Graph Representation: Adjacency List

- Adjacency lists
  - Representation of a graph with n nodes using an array of n lists of nodes
  - List i contains node j if there is an edge (i, j)
  - A weighted graph can be represented with a list of node/weight pairs
  - Space proportional to $\Theta(m+n)$
  - Checking if (i, j) is an edge takes $O(k_i)$ time

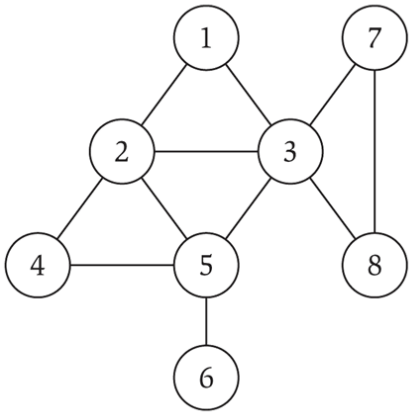




Adjacency list

Graph Representation: Edge List

- Edge list
  - Very simple way to represent a graph
  - List of $|E|$ edges
  - An edge is represented as as pair of vertices (e.g., source, destination)
  - Space proportional to $\Theta(|E|)$
  - Checking if (i, j) is an edge takes $O(|E|)$ time (if the edges appear in no particular order)

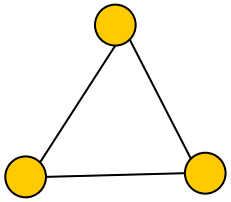


1	2
1	3
2	3
2	4
2	5
3	5
3	7
3	8
4	5
5	6
7	8

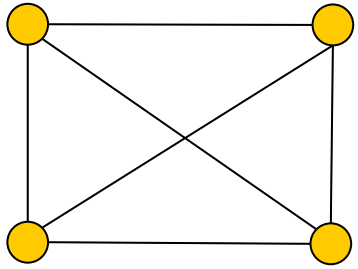
Edge list

Complete Graph

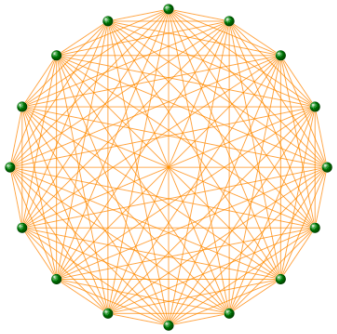
- Definition: A graph $G = (V, E)$ is called complete K_n if every pair of nodes is connected by an edge



Complete graph with 3 nodes: triangle (K_3)



K_4



K_{16}

- Q: What is the number of edges in K_n (complete graph with n nodes)?

$$\binom{n}{2} = \frac{n!}{2!(n-1)!} = \frac{n(n-1)}{2}$$

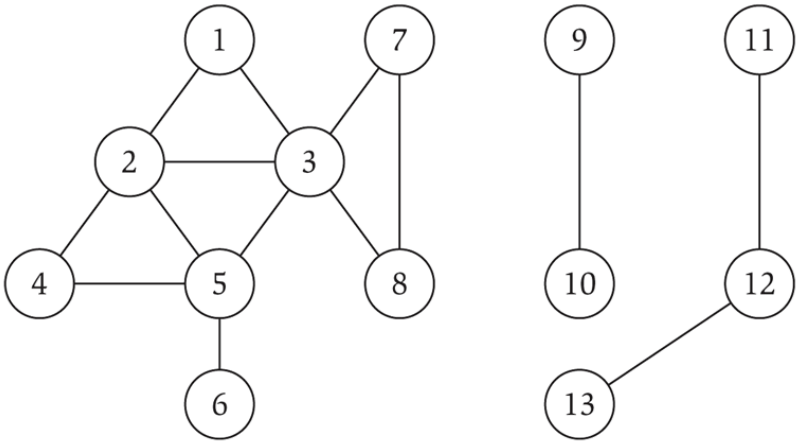
Paths, graph connectivity and distance

Paths and Connectivity in Graphs

- Definition: A path in an undirected graph $G=(V,E)$ is a sequence of nodes $v_1, v_2, \ldots, v_k$ with the property that each consecutive pair v_{i-1}, v_i is joined by an edge in E

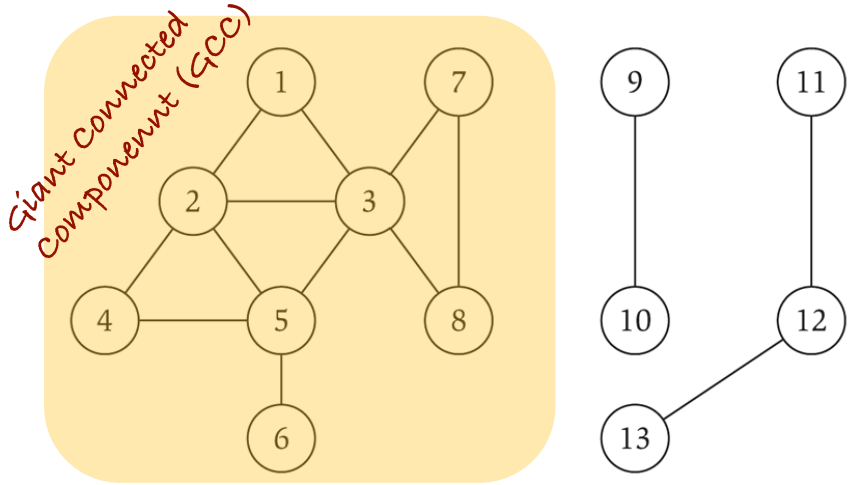
- Definition: An undirected graph is connected if for every pair of nodes u and v, there is a path between u and v

Is this graph connected?



Connected Components

- A connected component is a maximal connected subgraph of a graph G (there is a path between any pair of nodes)
  - Maximal means adding another vertex will ruin connectivity

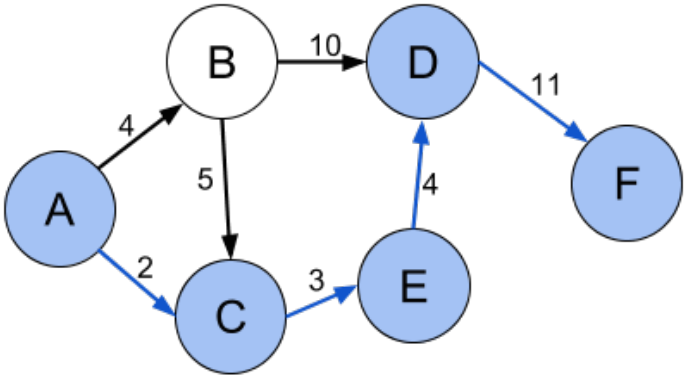


Graph with 3 connected components

Connected component containing node 1: {1, 2, 3, 4, 5, 6, 7, 8}

Shortest Paths

- Definition: find a path between two nodes in a graph, in such a way that the sum of the weights of its constituent edges is minimized
  - Many applications (e.g., road networks, community detection, communications)
- Variants
  - Single-source shortest path problem
  - Single-destination shortest path problem
  - All-pairs shortest path problem



Shortest path (A, C, E, D, F) between vertices A and F in the weighted directed graph

Various algorithms:
- Dijkstra
- Bellman-Ford
  (works with negative edge weights)

See: https://en.wikipedia.org/wiki/Shortest_path_problem

How do we measure and characterize a network?

More properties

Key Network Properties

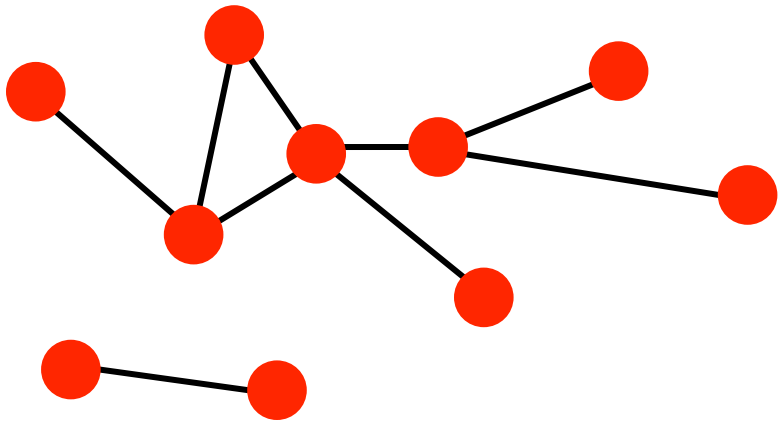
Degree distribution:	$P(k)$
Path length:	h
Clustering coefficient:	C

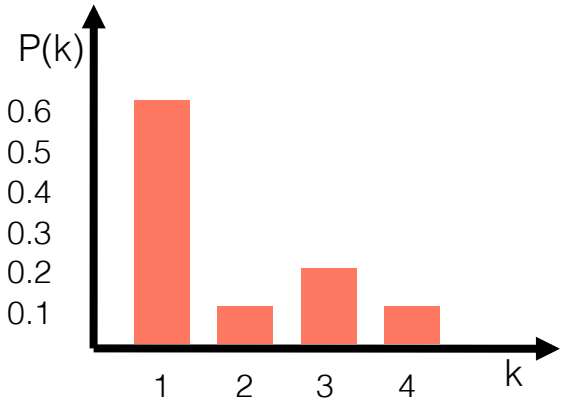
Degree Distribution

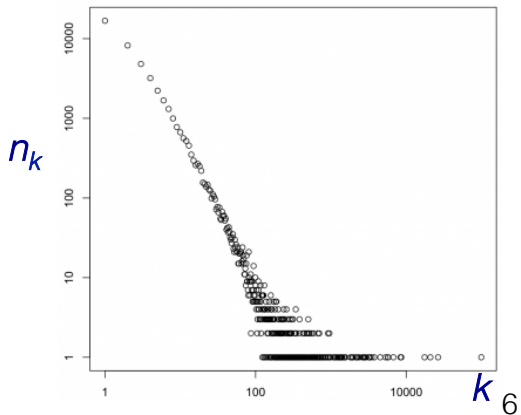
- Degree distribution *P(k)*: Probability that a randomly chosen node has degree *k*

  n_k = # nodes with degree *k*

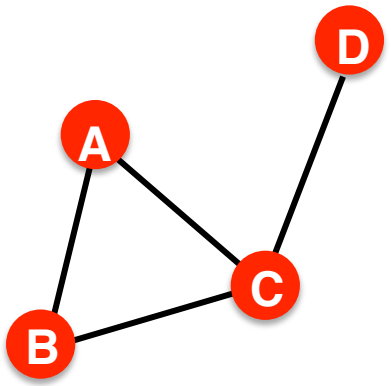
- Normalized histogram:
  $P(k) = n_k / n$ ➔ plot





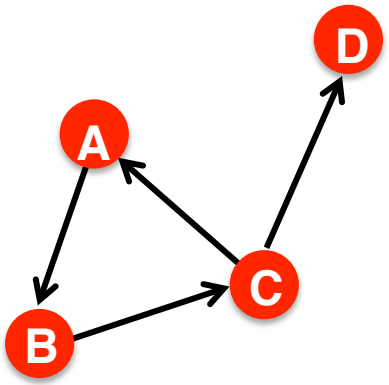


Distance in a Graph



$h_{B,D} = 2$

- Distance (shortest path, geodesic) between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - If the two nodes are disconnected, the distance is usually defined as infinite



$h_{B,C} = 1, h_{C,B} = 2$

- In directed graphs paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{A,C} \neq h_{C,A}$

Network Diameter

- Diameter: the maximum (shortest path) distance between any pair of nodes in a graph

- Average path length for a connected graph (component) or a strongly connected (component of a) directed graph
  - Many times we compute the average only over the connected pairs of nodes (that is, we ignore “infinite” length paths)

$$\bar{h} = \frac{1}{n(n-1)} \sum_{i,j \neq i} h_{ij}$$

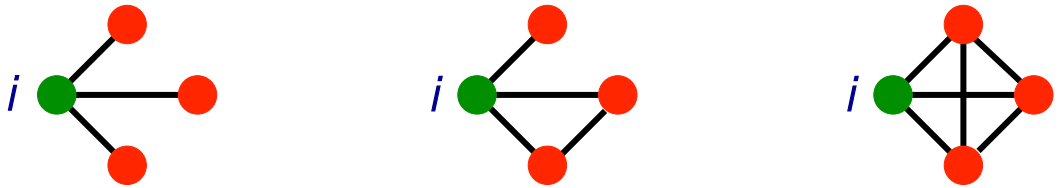
where h_{ij} is the distance from node i to node j

Clustering Coefficient (1/2)

- Clustering coefficient
  - What portion of node i's neighbors are connected?
  - Node *i* with degree k_i
  - $C_i \in [0,1]$

$$C_i = \frac{2e_i}{k_i(k_i-1)}$$

where e_i is the number of edges between the neighbors of node *i*



C_i=0 C_i=1/3 C_i=1

Average clustering coefficient: $C = \frac{1}{|V|}\sum_{i=1}^{|V|} C_i$

Key Network Properties

Degree distribution:	*P(k)*
Path length:	*h*
Clustering coefficient:	*C*

They can inform us about the structure of a network

What is happening in real networks?

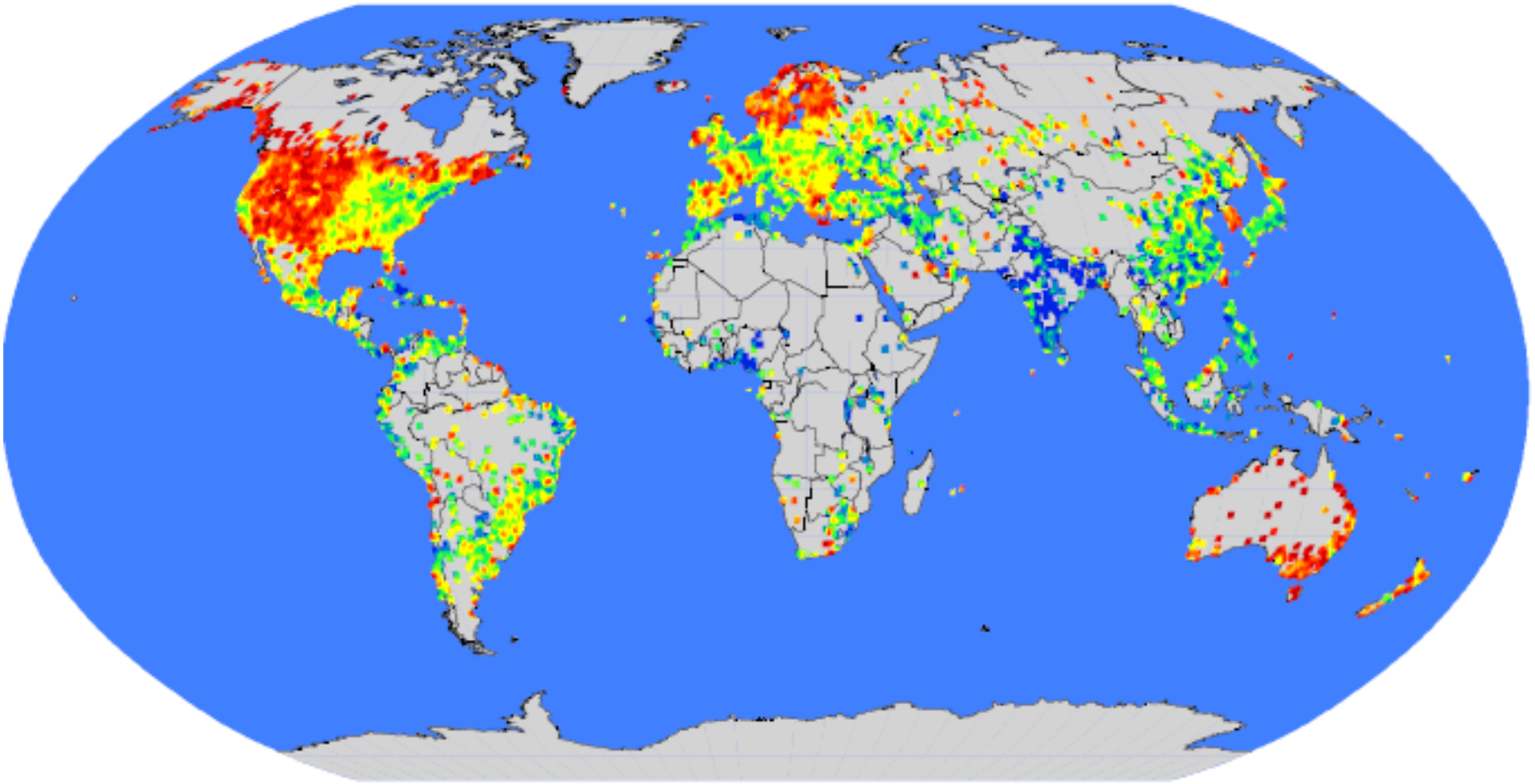
Let’s see the properties of a real network:

The MSN messenger case

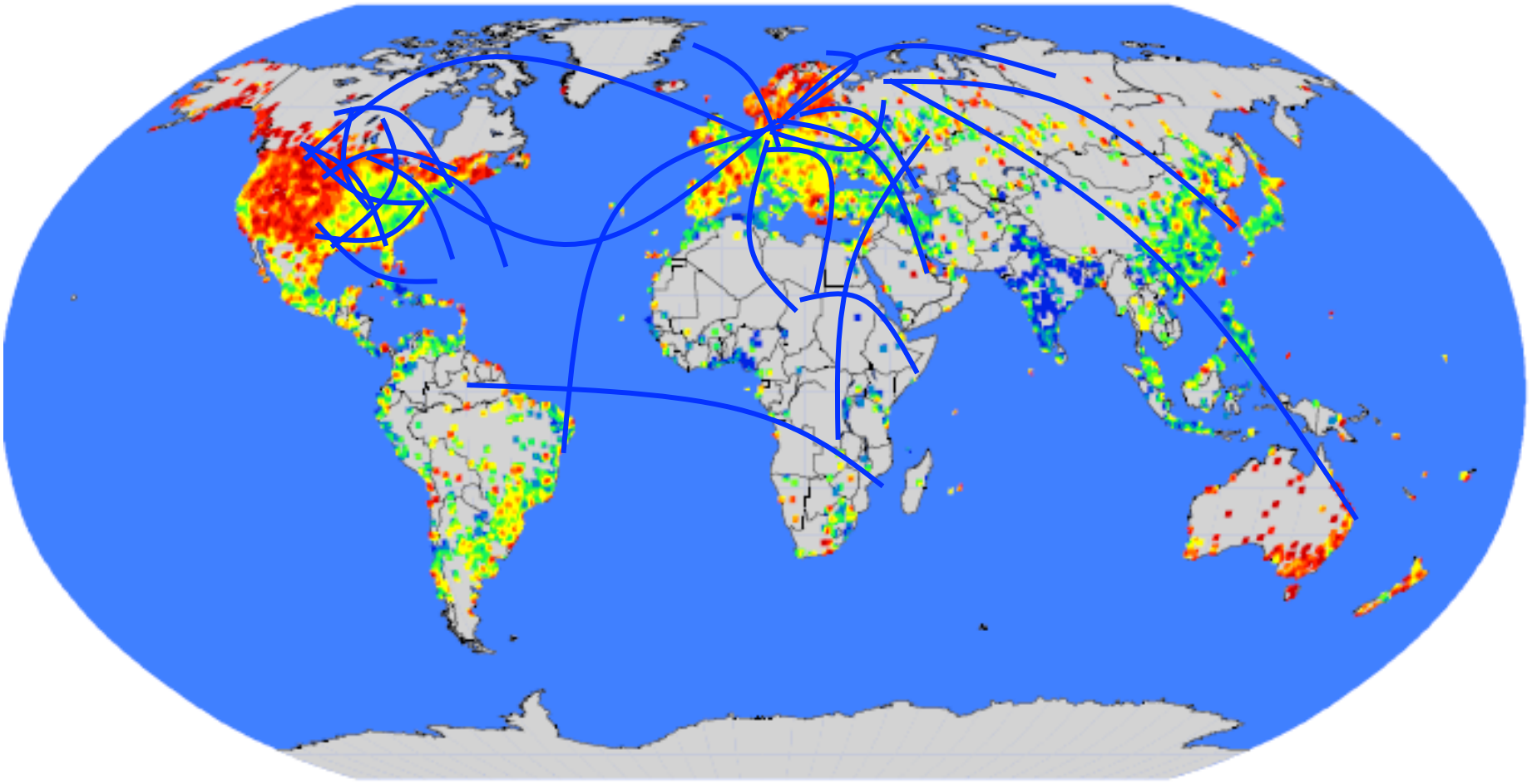
The MSN Messenger

- MSN Messenger activity in June 2006:
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages
  - Now called *Windows Live Messenger*

Communication: Geography

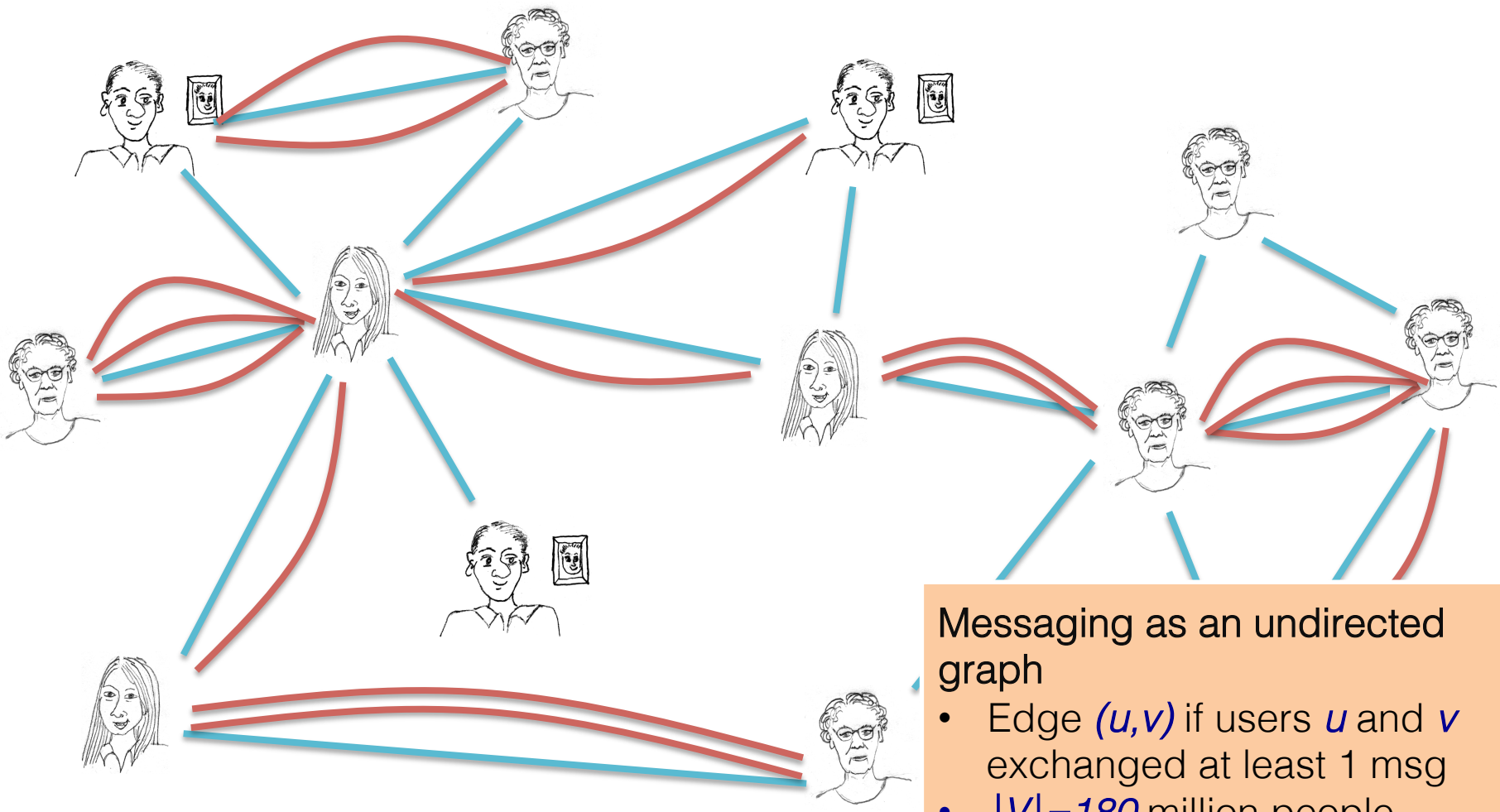


Communication network



Network: 180M people, 1.3B edges

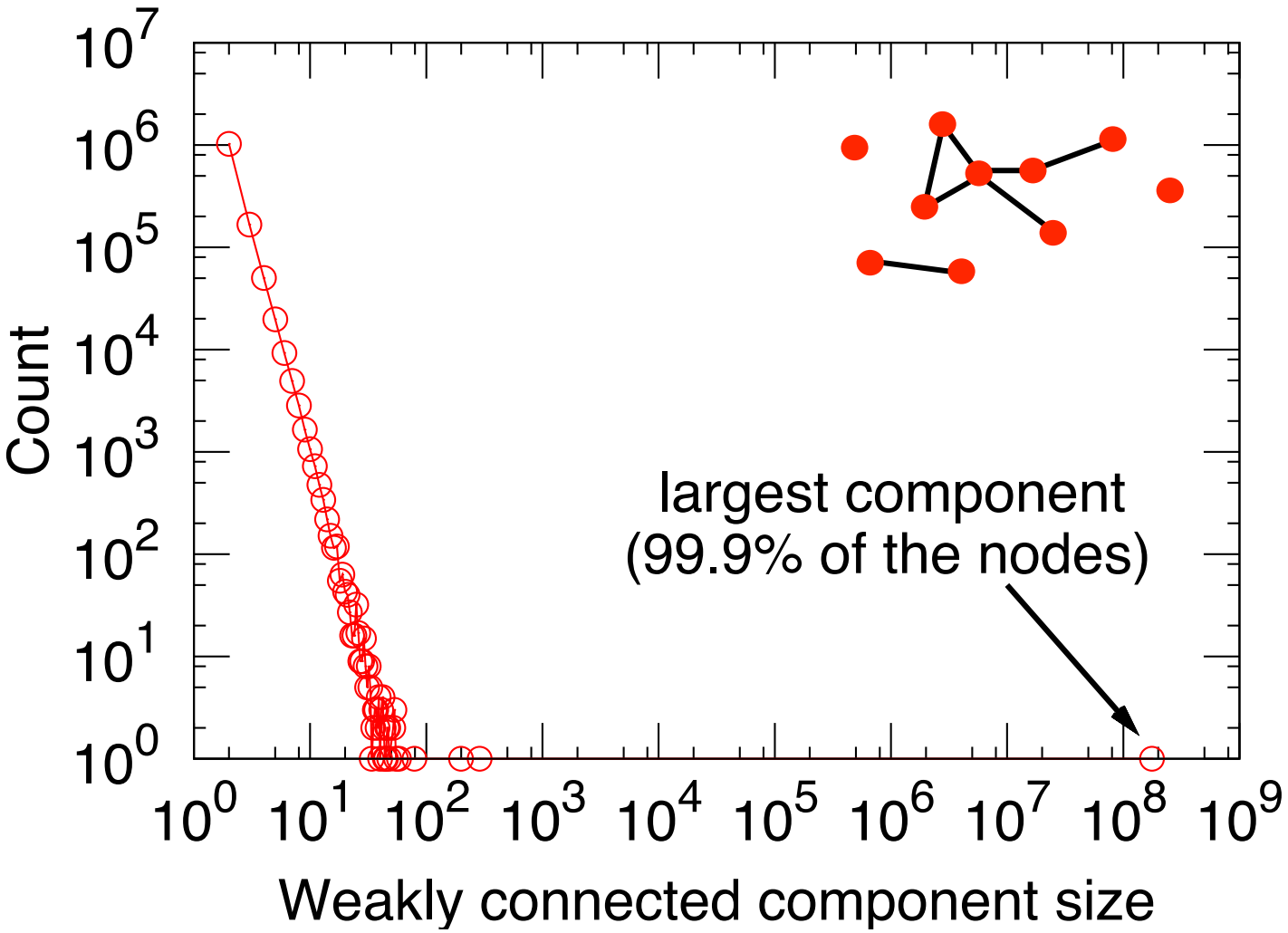
Messaging as a Multigraph



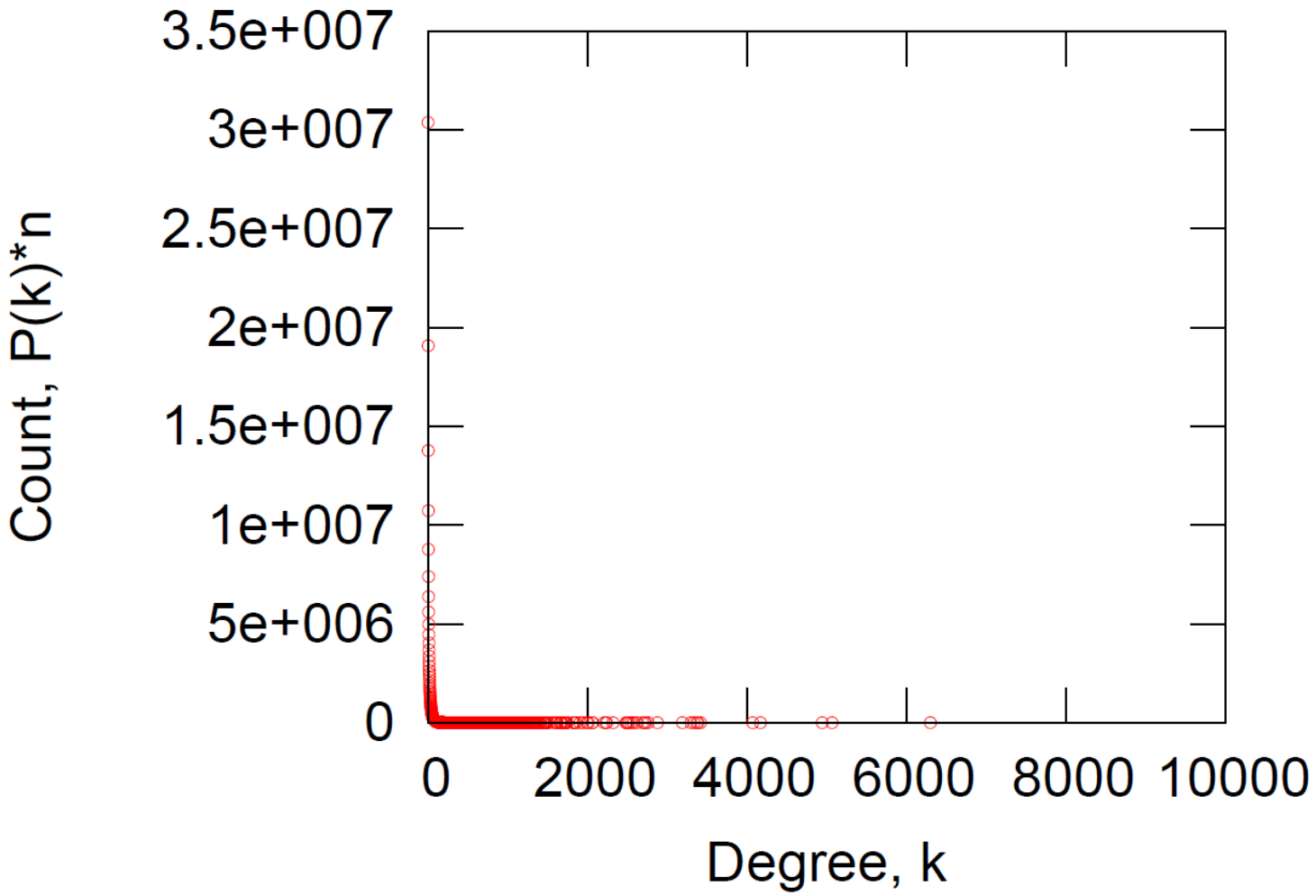
Messaging as an undirected graph

- Edge *(u,v)* if users *u* and *v* exchanged at least 1 msg
- $|V|=180$ million people
- $|E|=1.3$ billion edges

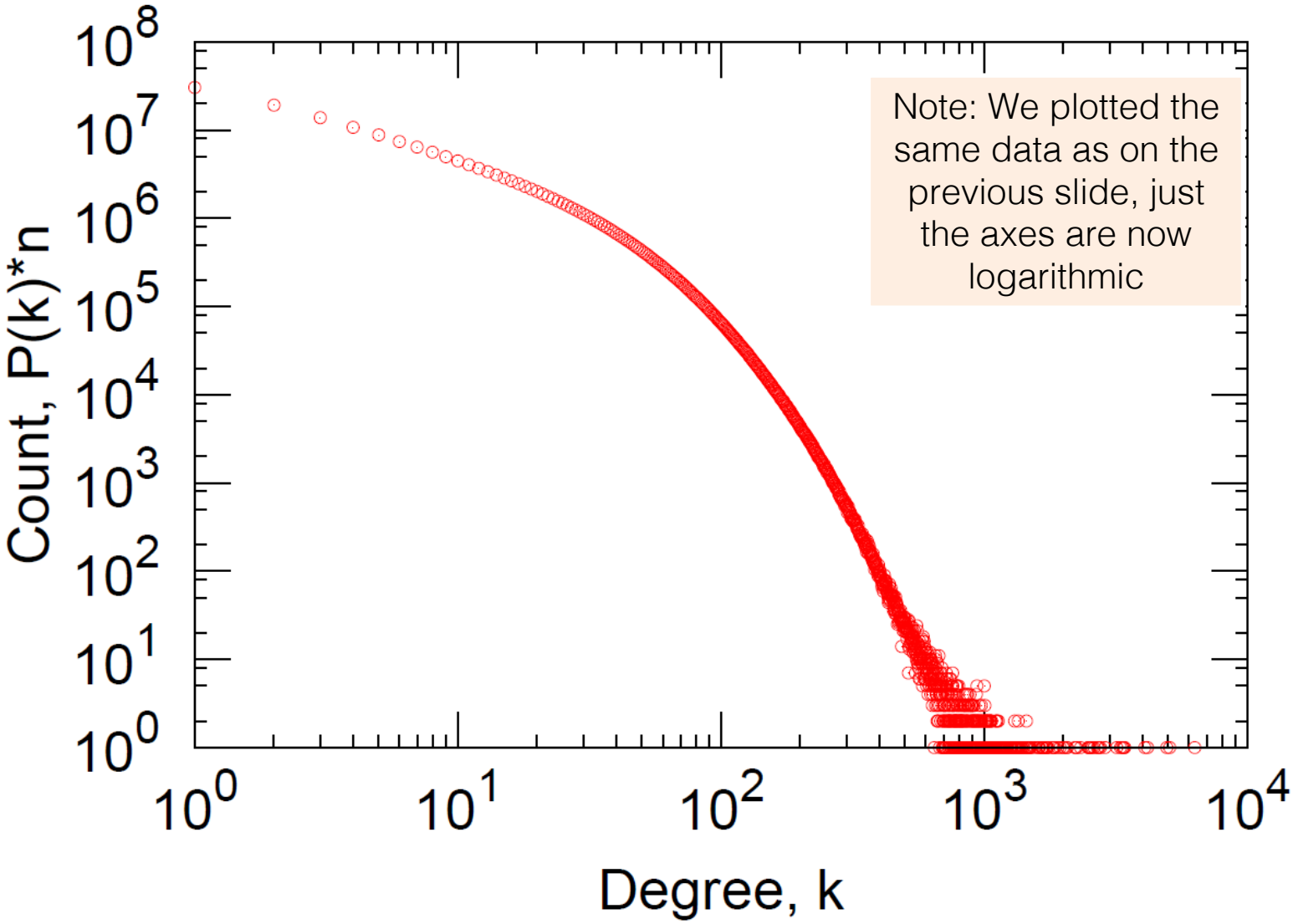
MSN: Connectivity



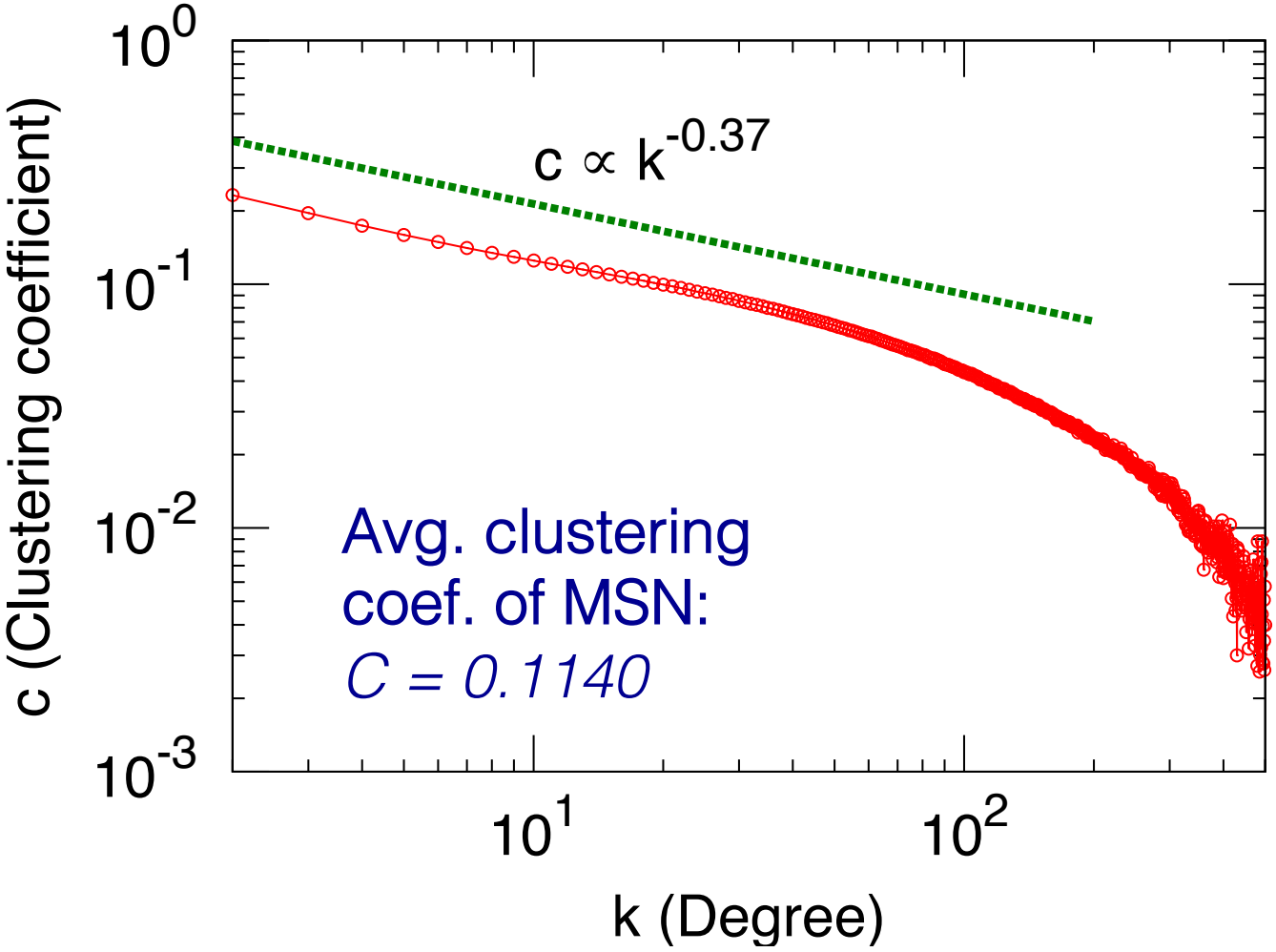
MSN: Degree Distribution



MSN: Log-Log Degree Distribution

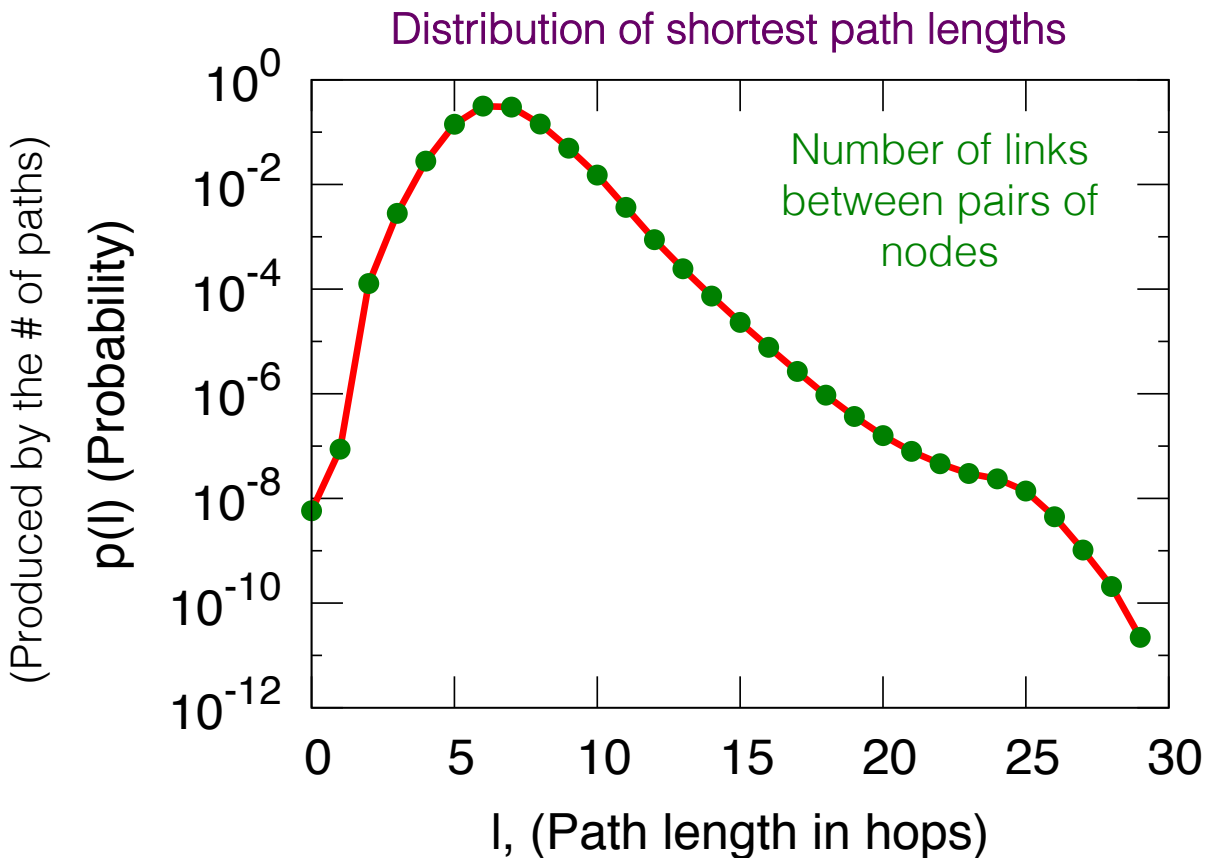


MSN: Clustering



C_k: average C_i of nodes i of degree k: $$C_k = \frac{1}{n_k} \sum_{i:k_i=k}^{n} C_i$$

MSN: Diameter



Avg. path length **6.6**
90% of the nodes can be reached in < 8 hops

nodes as we do BFS out of a random node

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

MSN: Key Network Properties

Degree distribution:	*Heavily skewed* *avg. degree= 14.4*
Path length:	*6.6*
Clustering coefficient:	*0.11*

Are these values “expected”?
Are they “surprising”?

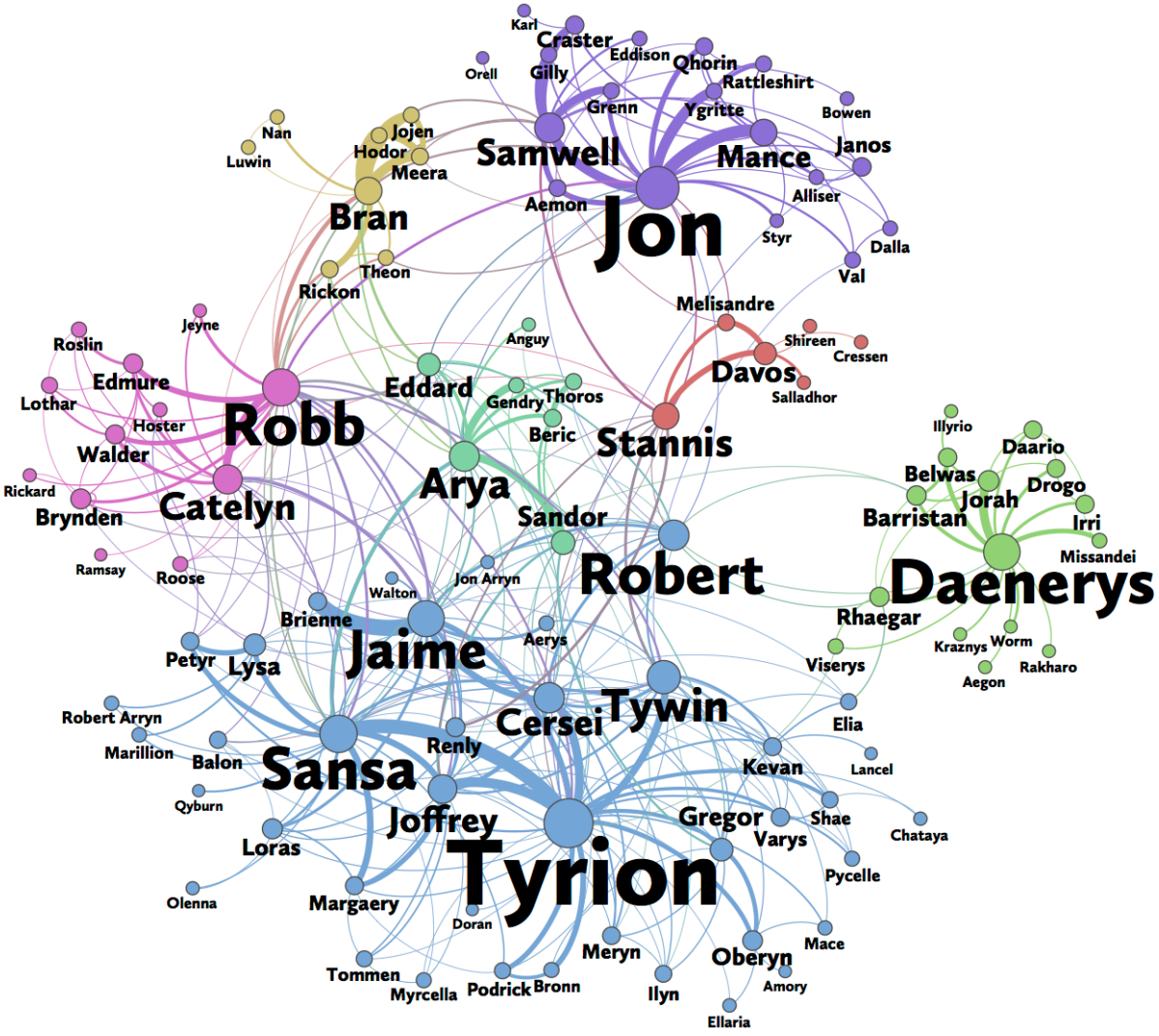
To answer this we need a model to compare against
More on the lab session

Centrality criteria

Centrality in Networks (1/2)

- Determine the relative importance of a node in the network
  - Applications in Social Network Analysis, the Internet, Epidemiology, Urban informatics, …
- What do we mean by centrality?
  - A central node is more important or powerful …
  - Or, more influential ...
  - Or, is more critical due to its location in the graph
- Also, very closely related to the problem of ranking in the context of Web search
  - Each webpage can be considered as a 'user'
  - Each hyperlink is an endorsement relationship
  - Centrality measures provide a query independent link-based score of importance of a web page

Centrality in Networks (2/2)



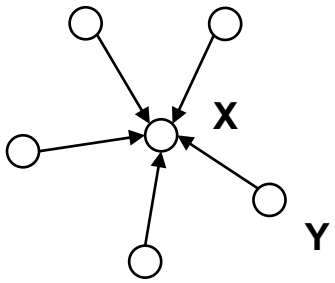
Source: https://www.macalester.edu/~abeverid/thrones.html

Measures of Centrality

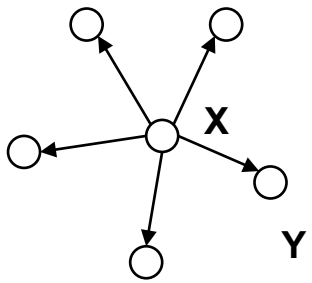
- Various types of centrality criteria:
  - Measures based on distances (e.g., degree, closeness)
  - Measures based on paths (e.g., betweenness, Katz's index)
  - Spectral measures (eigenvector, PageRank, HITS, SALSA, random walks with restarts)
  - Measure based on groups of nodes (e.g., cliques, plexes, cores)
    - Related to the "clustering" structure
    - More on that in another lecture

A First Example

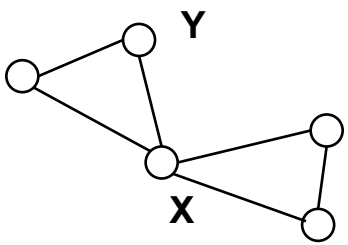
In each of the following networks, X has higher centrality than Y according to a particular measure



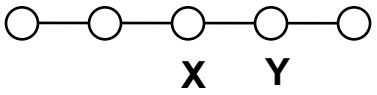
in-degree



out-degree



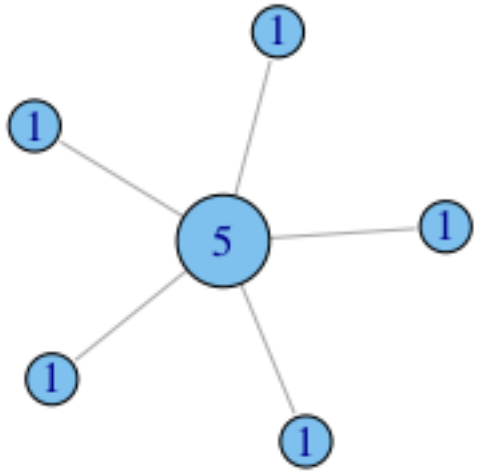
betweenness



closeness

Degree Centrality (1/2)

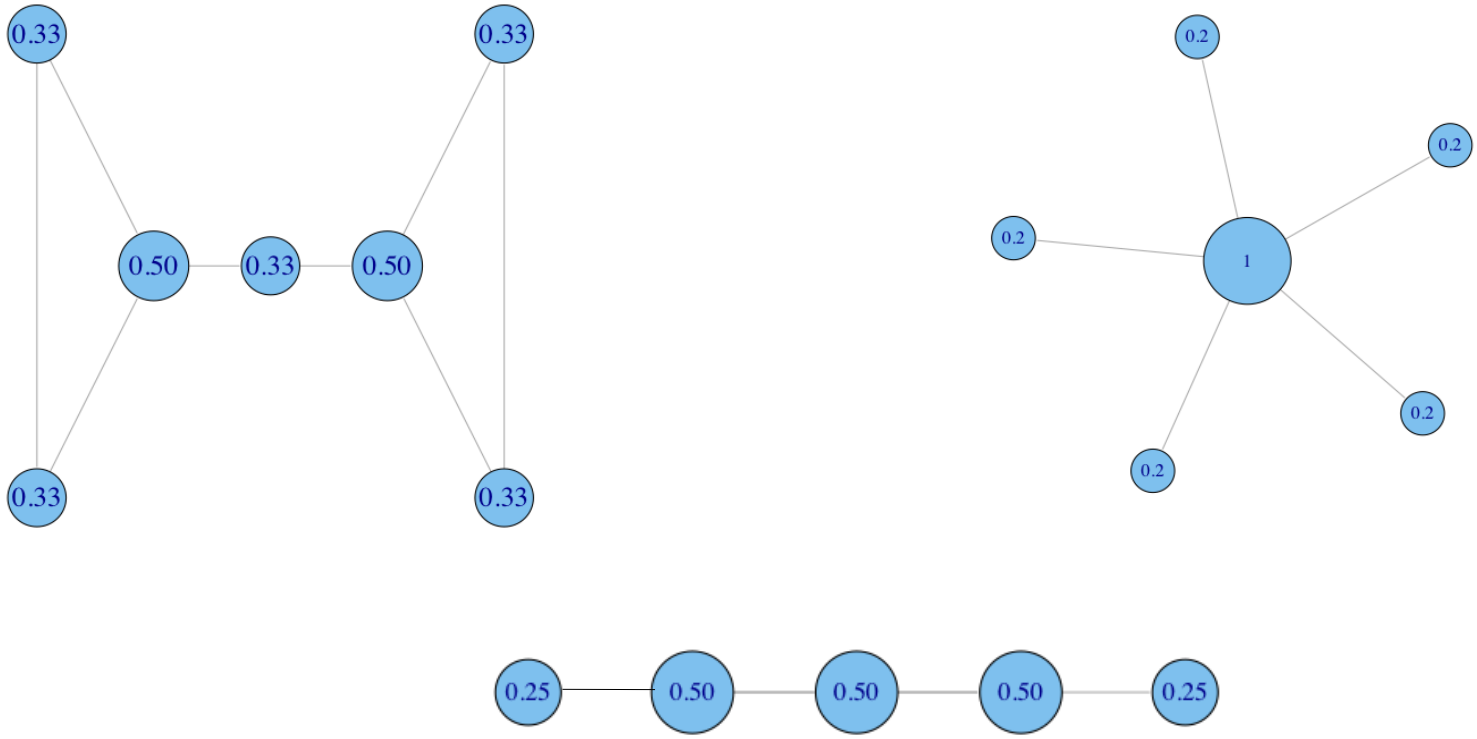
- Idea: A central node is one with many connections



- $C_d\ (i) = k(i)$, where $k(i)$ is the degree of node i

Degree Centrality (2/2)

- Idea: A central node is one with many connections



- Normalized degree centrality: divide by the max possible degree (*n-1*)

Degree Centrality in Directed Graphs

- In directed graphs, we can use the
  - in-degree
  - out-degree
  - Combination of them (e.g., sum of in- and out-degree)

- Which one is more important?
  - In practice, mostly the in-degree is used
  - Why? Think about the Web graph or the Twitter social network

Closeness Centrality

- Motivation: it measures the ability to quickly access or pass information through the graph

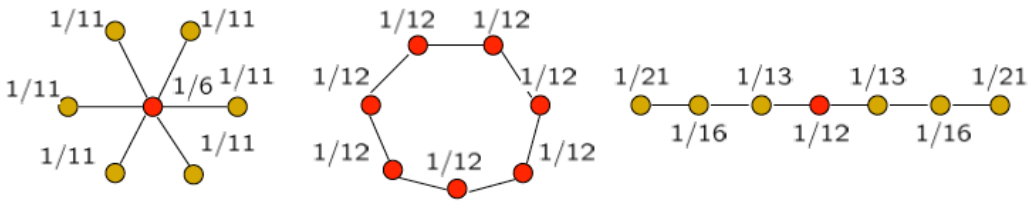
$$C_{cl}(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}$$

values in the range [0,1]

Mean distance from a node to other nodes

d(i, j) is the length of the shortest path between *i* and *j* (geodesic distance)

- The closeness of a node is defined as the inverse of the sum of the shortest path (SP) distances between the node and all other nodes in the graph



Why inverse the distance?

- Nodes with low mean distance should get high score

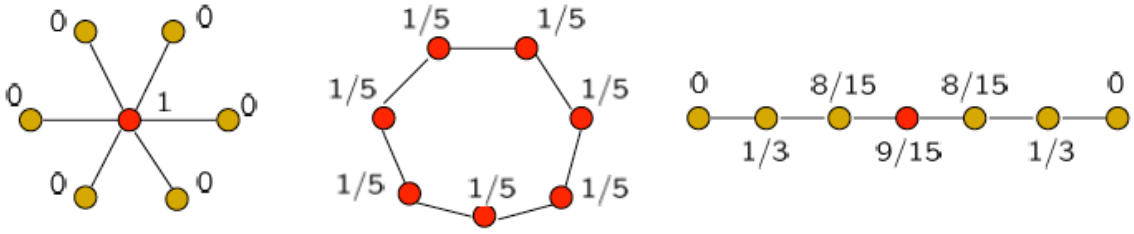
Be close to everybody else
(e.g., influence on other nodes)

Betweenness Centrality

- Motivation: a node is important if it lies in many shortest paths

$$C_{bt}(i) = \sum_{s \neq i \neq t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)}$$

- $\sigma(s, t)$ is the total number of shortest paths from s to t
- $\sigma(s, t|i)$ is the number of shortest paths from s to t that pass through i



Essential nodes in passing information through the network

Oftentimes it is normalized: $\frac{C_{bt}(i)}{\binom{n-1}{2}}$

The PageRank Algorithm

PageRank

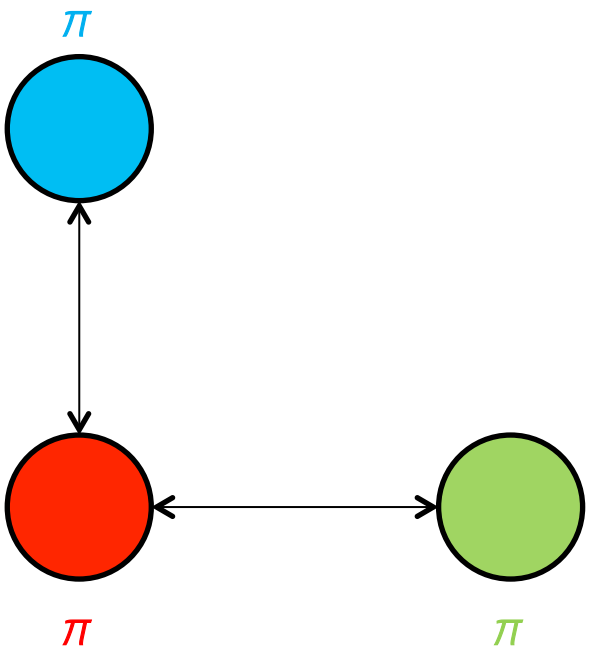
- Good authorities should be pointed by good authorities
  - The value of a node comes from the value of the nodes that point to it

- How do we implement that?
  - Assume that we have a unit of authority to distribute to all nodes
    - Initially, each node gets *1/n* amount of authority
  - Each node distributes its authority value to its neighbors
  - The authority value of each node is the sum of the authority fractions that they collect from their neighbors

$$\pi_v = \sum_{\forall (u,v) \in E} \frac{1}{k_{out}(u)} \pi_u$$

π_v: the PageRank value of node *v*
- Recursive definition

A Simple Example



$\pi + \pi + \pi = 1$

$\pi = \pi + \pi$

$\pi = \frac{1}{2}\pi$

$\pi = \frac{1}{2}\pi$

- Solving the system of equations we get the authority values for the nodes
  - $\pi = \frac{1}{2}$ $\pi = \frac{1}{4}$ $\pi = \frac{1}{4}$

A More Complex Example

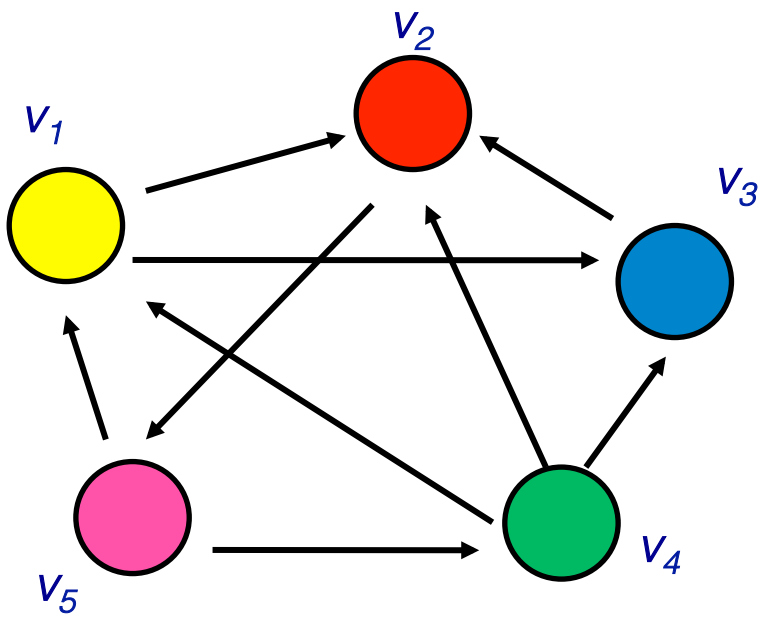
$\pi_1 = 1/3\ \pi_4 + 1/2\ \pi_5$

$\pi_2 = 1/2\ \pi_1 + \pi_3 + 1/3\ \pi_4$

$\pi_3 = 1/2\ \pi_1 + 1/3\ \pi_4$

$\pi_4 = 1/2\ \pi_5$

$\pi_5 = \pi_2$



$$\pi_v = \sum_{\forall (u,v) \in E} \frac{1}{k_{out}(u)} \pi_u$$

Computing PageRank Weights

- A simple way to compute the weights is by iteratively updating the weights

Initialize all PageRank weights to *1/n*

Repeat:

$$\pi_v = \sum_{\forall (u,v) \in E} \frac{1}{k_{out}(u)} \pi_u$$

Until the weights do not change

This process converges

Random Walks on Graphs

- The PageRank algorithm defines a random walk on the graph
- Random walk
  - Start from a node chosen uniformly at random with probability *1/n*
  - Pick one of the outgoing edges uniformly at random
  - Move to the destination of the edge
  - Repeat

The PageRank of node *v* is the probability that the random walk is at node *v* after a very large number of steps

The Random Surfer model

- Users wander on the web, following links

Random Walk - Example

- Q: What is the probability p_i^t of being at node i after t steps?

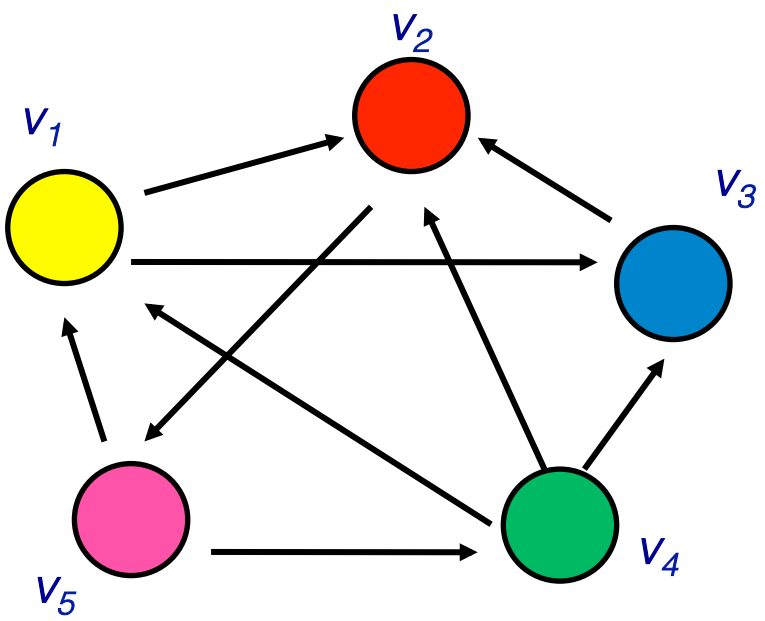
$p_1^0 = 1/5$ $\quad p_1^t = 1/3\, p_4^{t-1} + 1/2\, p_5^{t-1}$

$p_2^0 = 1/5$ $\quad p_2^t = 1/2\, p_1^{t-1} + p_3^{t-1} + 1/3\, p_4^{t-1}$

$p_3^0 = 1/5$ $\quad p_3^t = 1/2\, p_1^{t-1} + 1/3\, p_4^{t-1}$

$p_4^0 = 1/5$ $\quad p_4^t = 1/2\, p_5^{t-1}$

$p_5^0 = 1/5$ $\quad p_5^t = p_2^{t-1}$

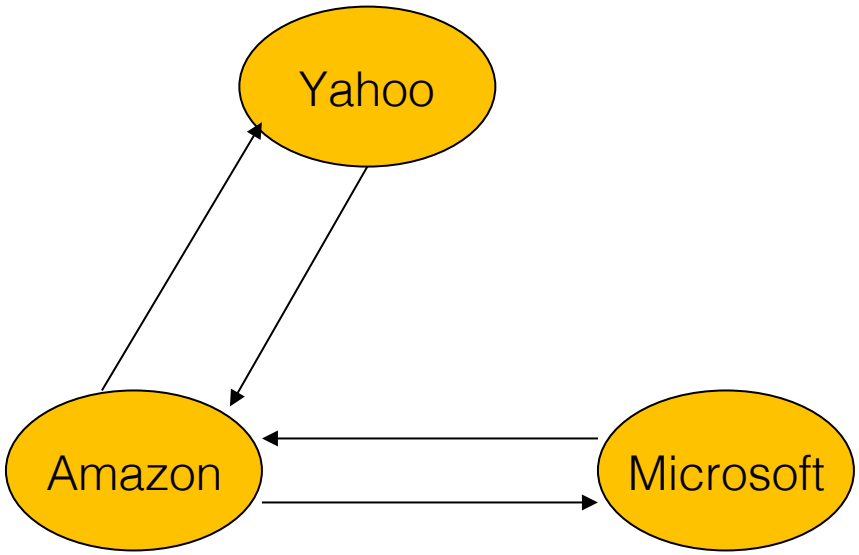


Random Walks

- Random walks on graphs correspond to Markov Chains
  - The set of states *S* is the set of nodes of the graph *G*
  - The transition probability matrix is the probability that we follow an edge from one node to another

$$P(i, j) = \frac{1}{k_{out}(i)}$$

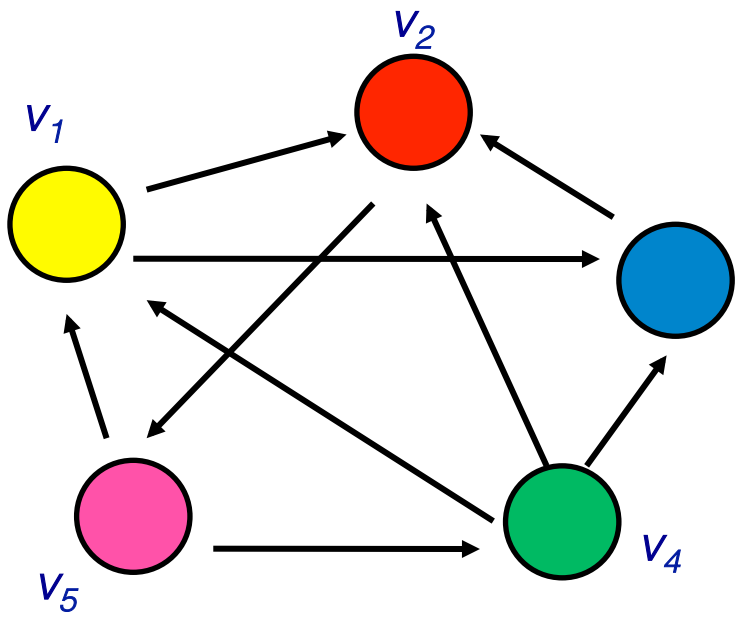
the transition matrix



An Example

the adjacency matrix

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



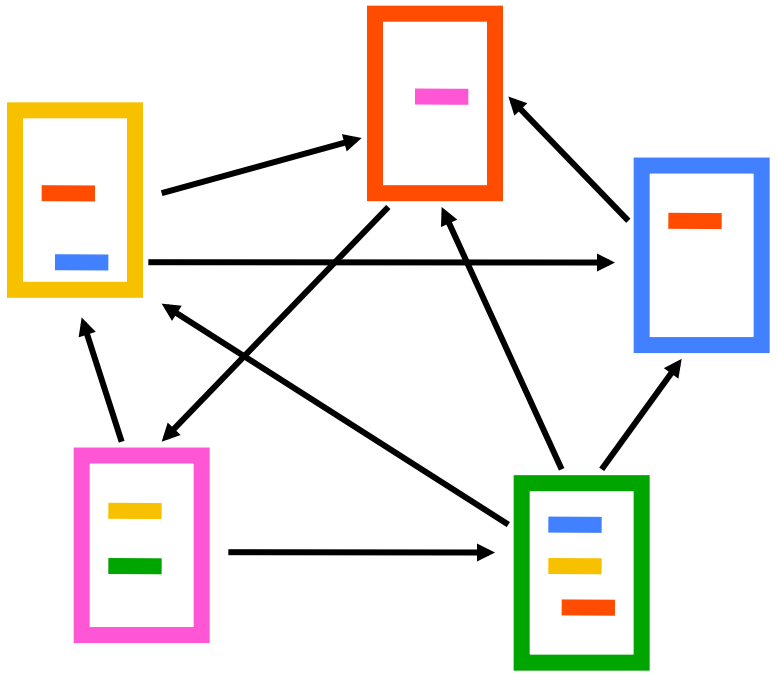
the transition matrix

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

The PageRank Random Walk

- Make the adjacency matrix stochastic and run a random walk

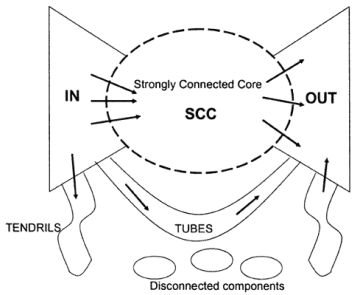
$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

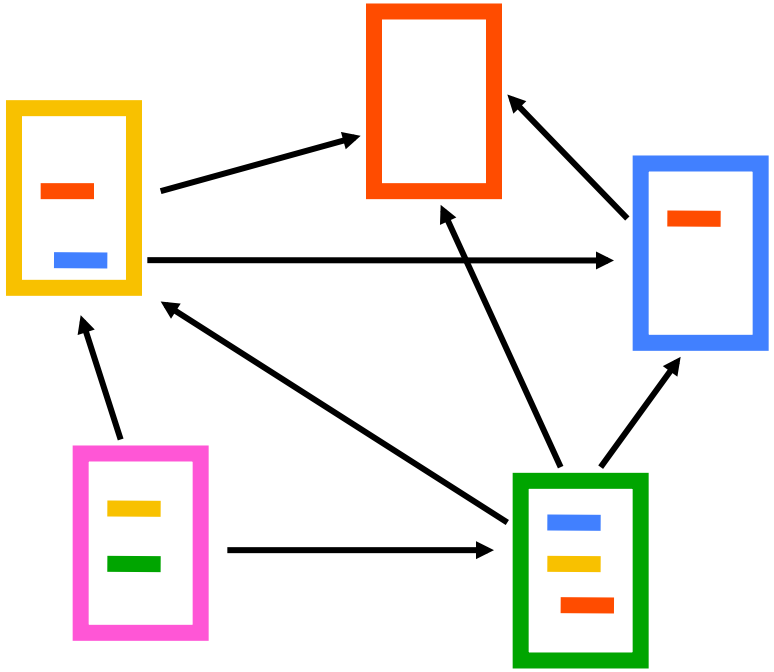


Sink Nodes (1/2)

- What about sink nodes?
  - What is happening when the random walk moves to a node without any outgoing links?

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

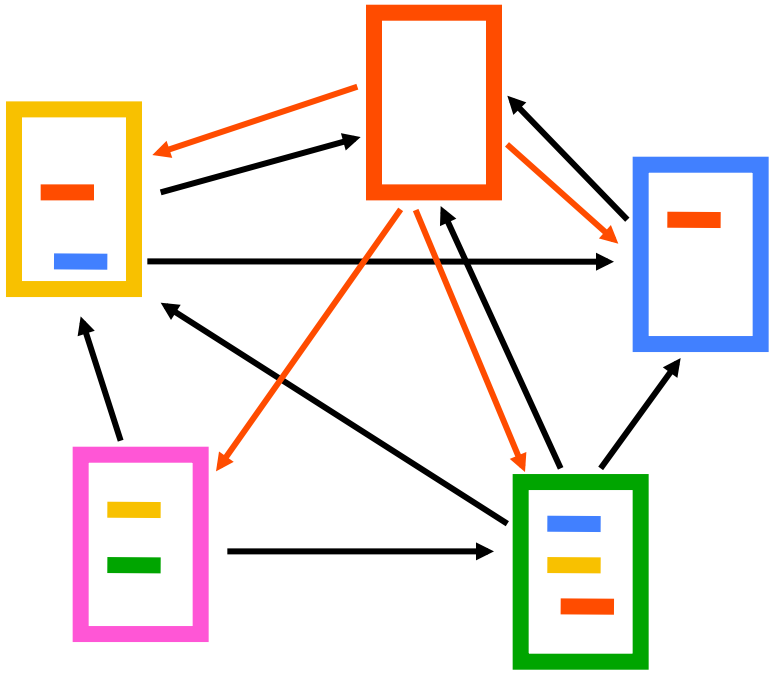




Sink Nodes (2/2)

- Replace these row vectors with a vector v
  - Typically, the uniform vector

$$P' = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

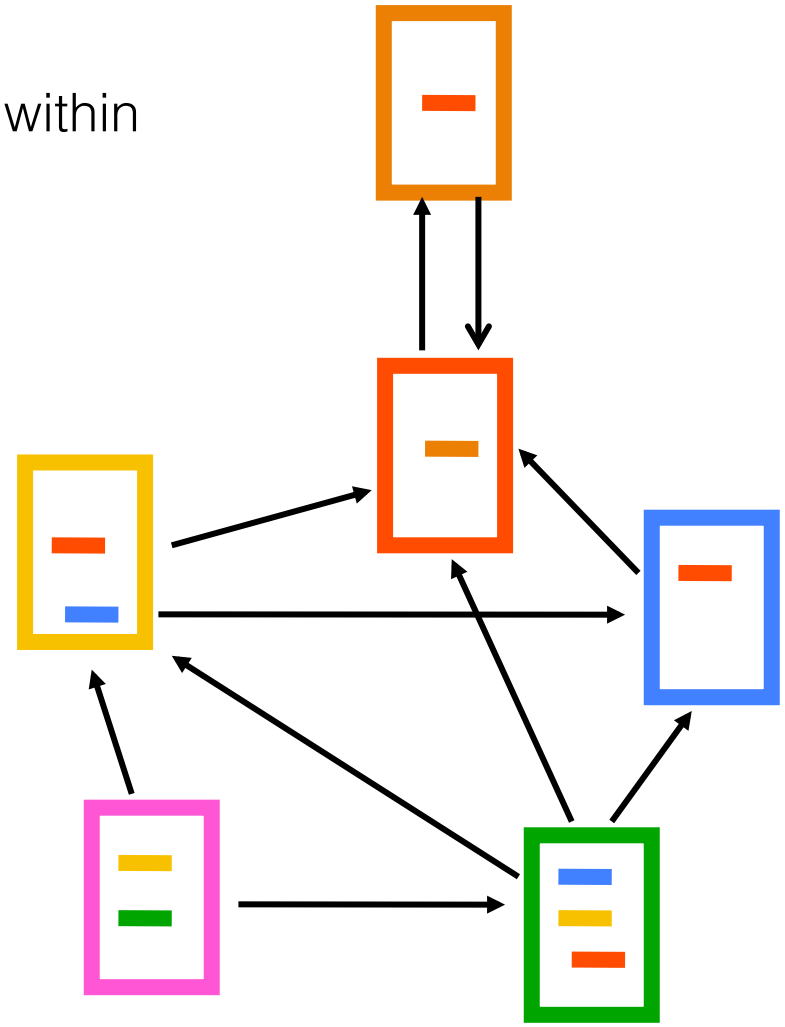


We can jump to any other node

$$P' = P + dv^T \qquad d = \begin{cases} 1 & \text{if } i \text{ is sink} \\ 0 & \text{otherwise} \end{cases}$$

Spider Traps

- What about loops?
  - Spider traps: all out-links are within the group



Solution: Random Teleports

- Add a random jump to any other node (vector *v*) with prob *1-α*
  - Typically, to a uniform vector
- Restarts after *1/(1-α)* steps in expectation
  - Guarantees convergence

$$P'' = \alpha \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

With prob. **α** follow a random outgoing link

With prob. **1-α** jump to a random node

Surfer will teleport out of spider trap within a few time steps

P'' = αP' + (1-α)uv^T, where *u* is the vector of all 1s

Google's PageRank Algorithm

- The Random Surfer model: at each time step, the random surfer has two options
  - With probability α follow a link at random
  - With probability $1-\alpha$ jump to a random page

- Rank according to the stationary distribution

$$\pi_v = \alpha \sum_{\forall (u,v) \in E} \frac{1}{k_{out}(u)} \pi_u \quad + \quad (1-\alpha)\frac{1}{n}$$

Typically, $\alpha = 0.85$

- We repeat this computation until convergence

[Page and Brin '98]

Community Structure

Properties of Real-World Networks

- Degree distribution
- Small-world phenomenon
- High clustering coefficient
- Triangle power-law
- Eigenvalue power-law

Static graphs

- Densification power-law
- Shrinking diameter

Time-evolving graphs

and many more …

In this part:
- Community structure
- Modularity-based community detection algorithms

Communities

- The notion of community structure captures the tendency of nodes to be organized into modules (communities, clusters, groups)
  - Members within a community are more similar among each other
- Typically, the communities in graphs correspond to densely connected entities (nodes)
- Set of nodes with more/better/stronger connections between its members, than to the rest of the network

- Why is this happening?
  - Individuals are typically organized into social groups (e.g., family, associations, profession)
  - Web pages can form groups according to their topic
  - …

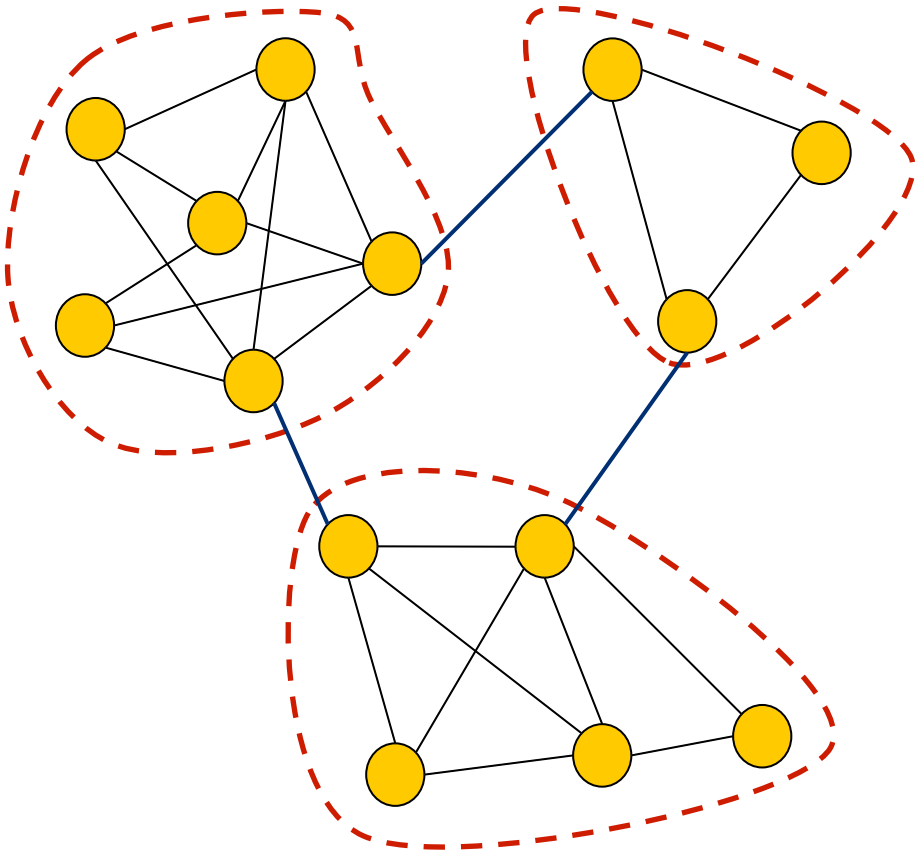
Definition of Communities

- How a community in graphs looks like?
- The property of community structure is difficult to be defined
  - There is no universal definition of the problem
  - It depends heavily on the application domain and the properties of the graph under consideration
- Most widely used notion/definition of communities is based on the number of edges within a group (density) compared to the number of edges between different groups

A community corresponds to a group of nodes with more intra-cluster edges than inter-clusters edges

[Newman '03], [Newman and Girvan '04], [Schaeffer '07], [Fortunato '10], [Danon et al. '05], [Coscia et al. 11]

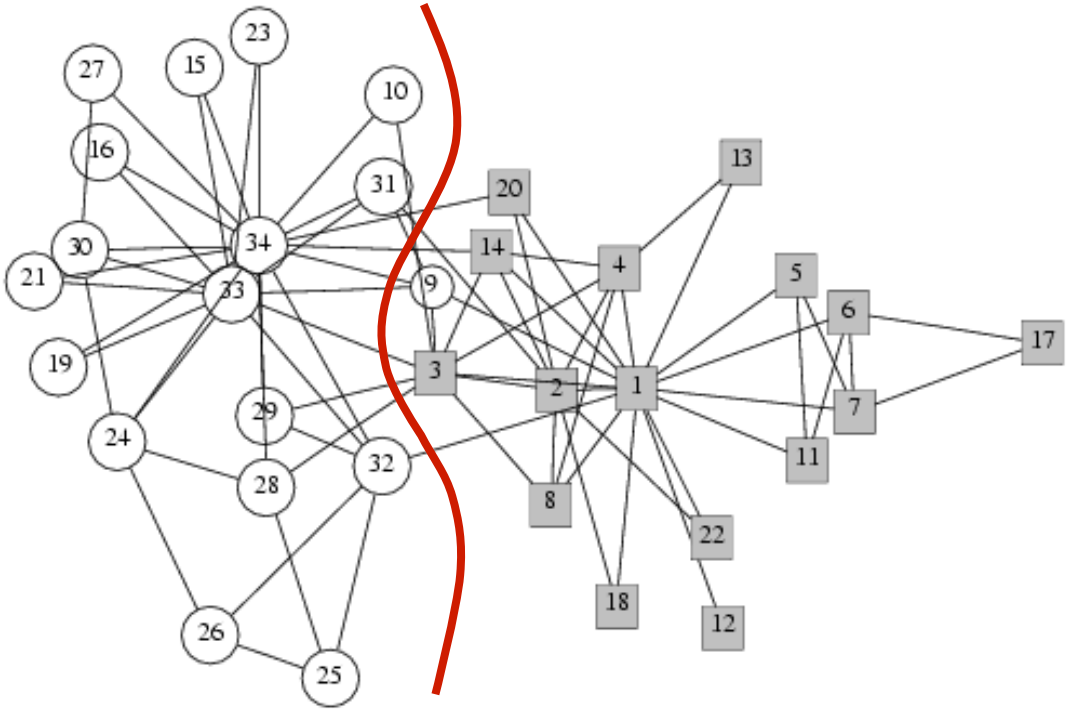
Community Structure



What leads to such a conceptual picture?

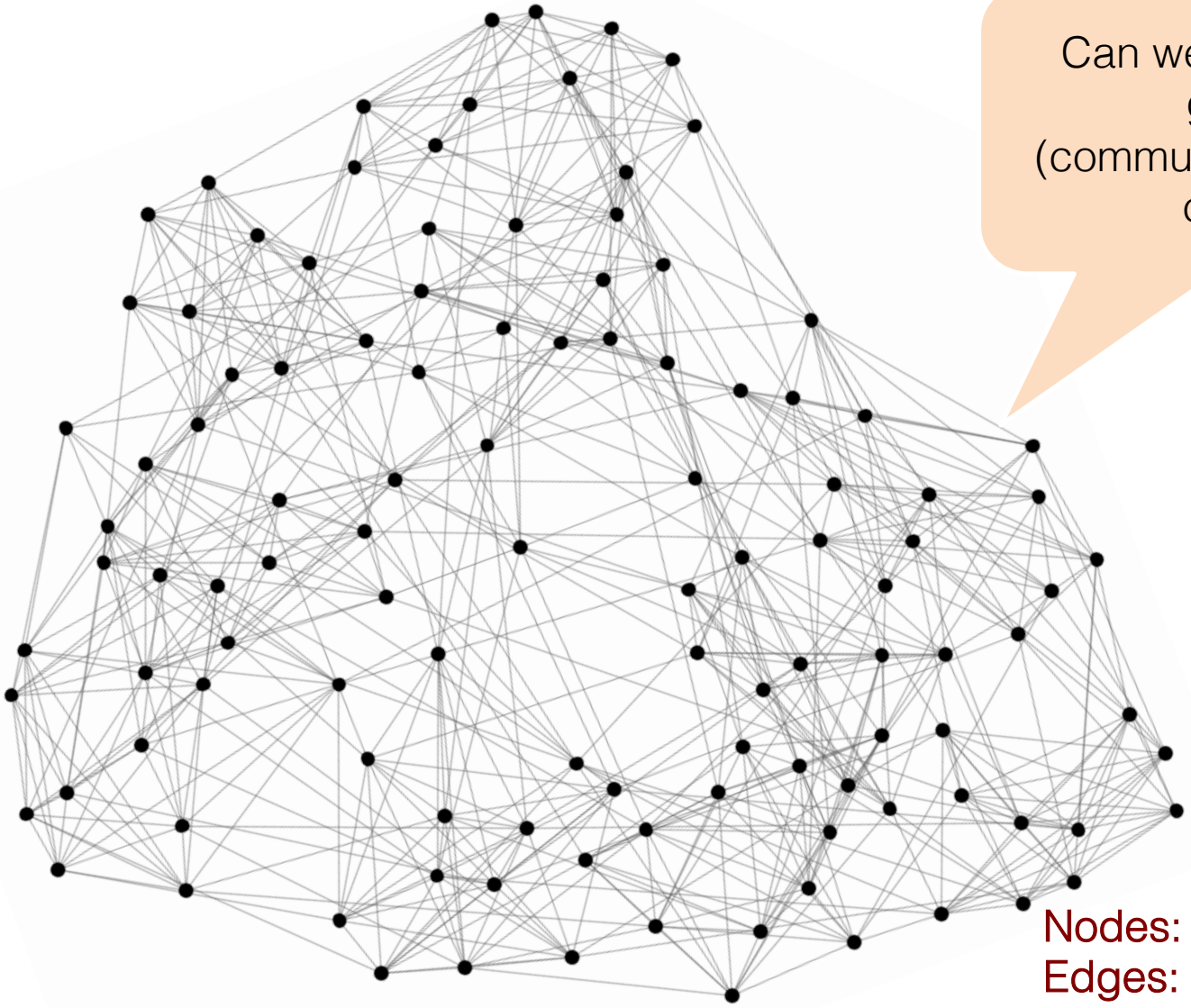
Network Communities

Social Network Data



- Zachary's Karate club network
  - Observe social ties and rivalries in a university karate club
  - During his observation, conflicts led the group to split
  - Split could be explained by a minimum cut in the network

NCAA Football Network (1/2)

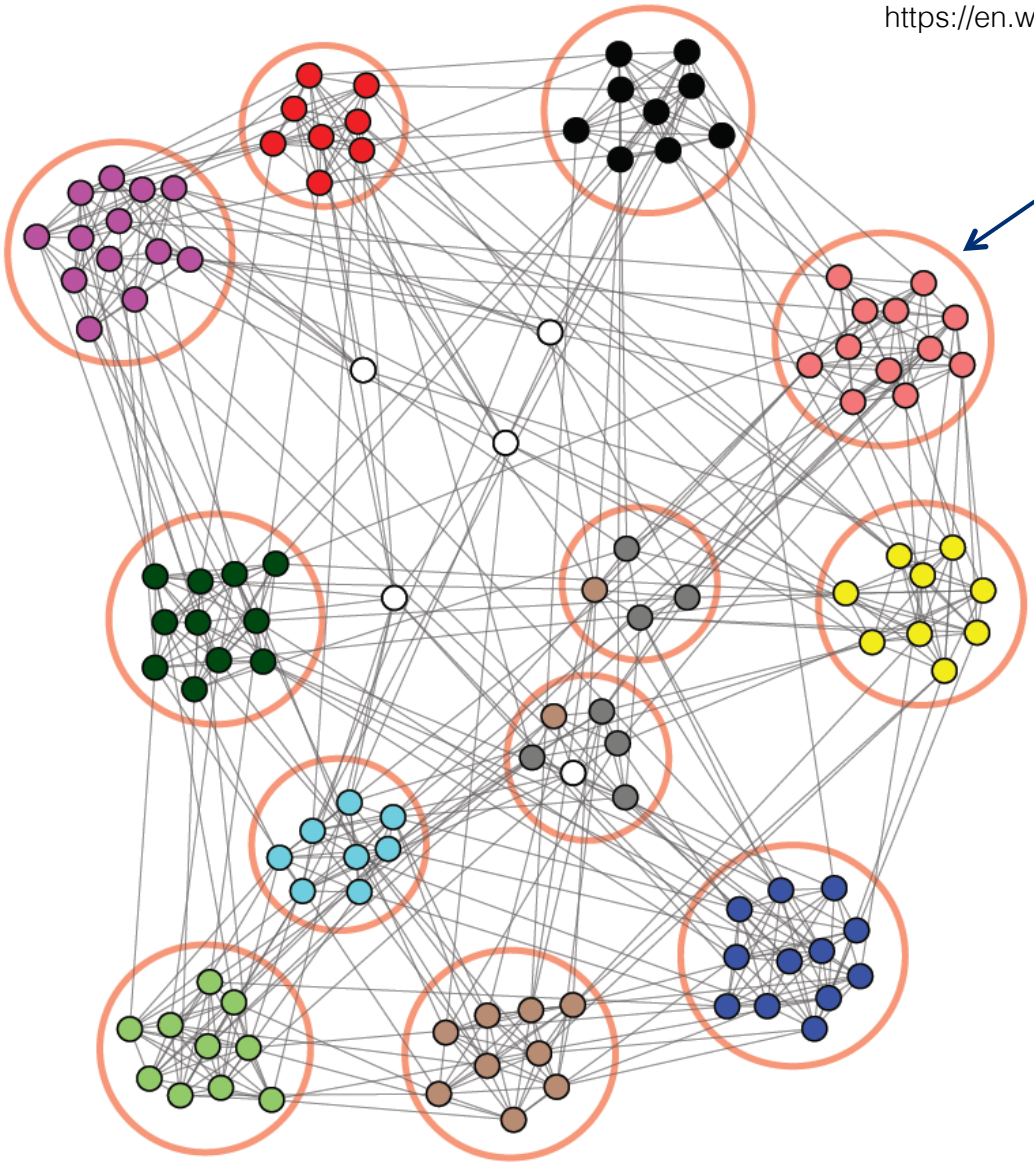


Can we identify node groups? (communities, modules, clusters)

Nodes: Teams
Edges: Games played

NCAA Football Network (2/2)

https://en.wikipedia.org/wiki/List_of_NCAA_conferences

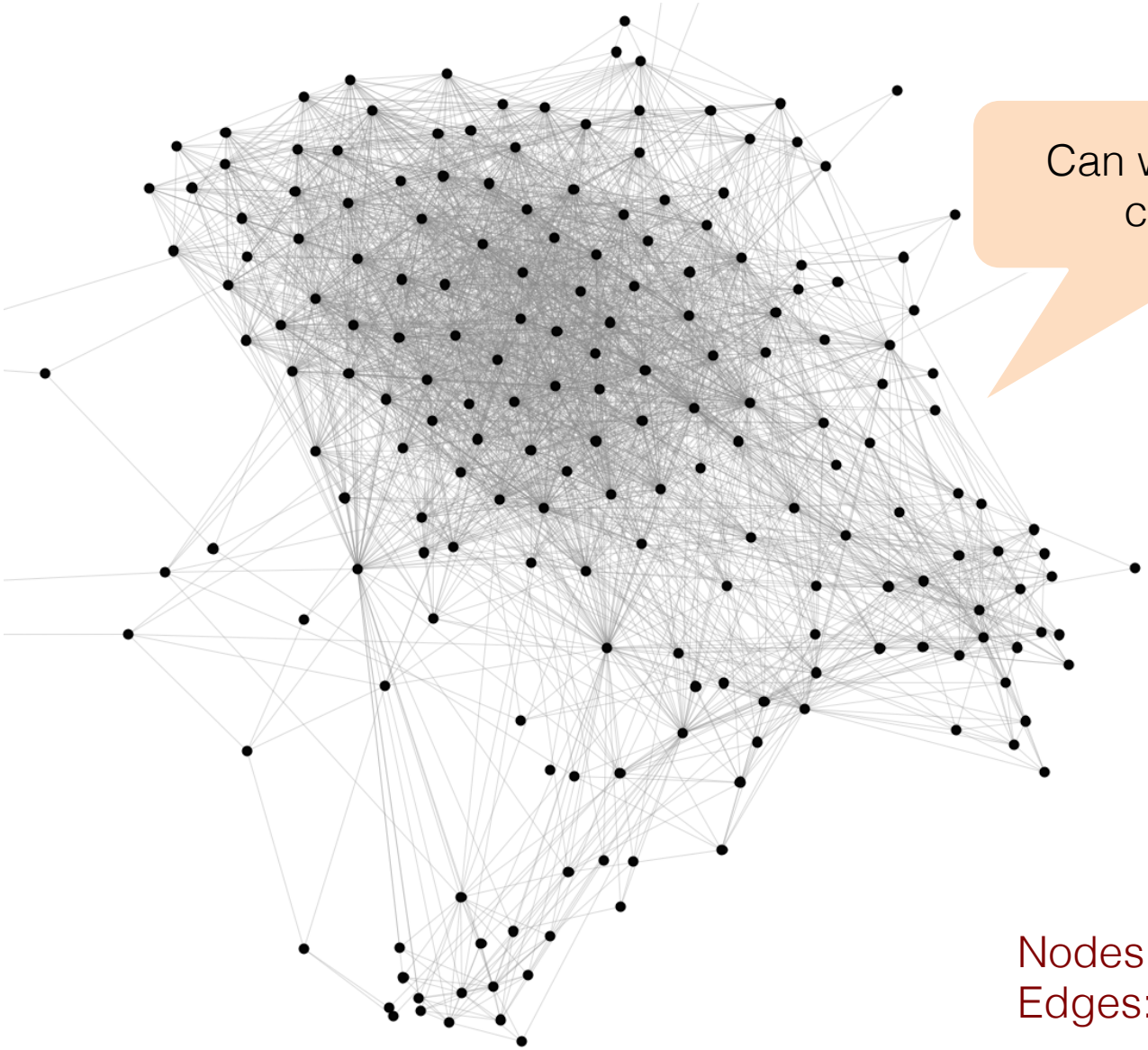


NCAA conferences

- Mid American
- Big East
- Atlantic Coast
- SEC
- Conference USA
- Big 12
- Western Athletic
- Pacific 10
- Mountain West
- Big 10
- Sun Belt
- Independents

Nodes: Teams
Edges: Games played

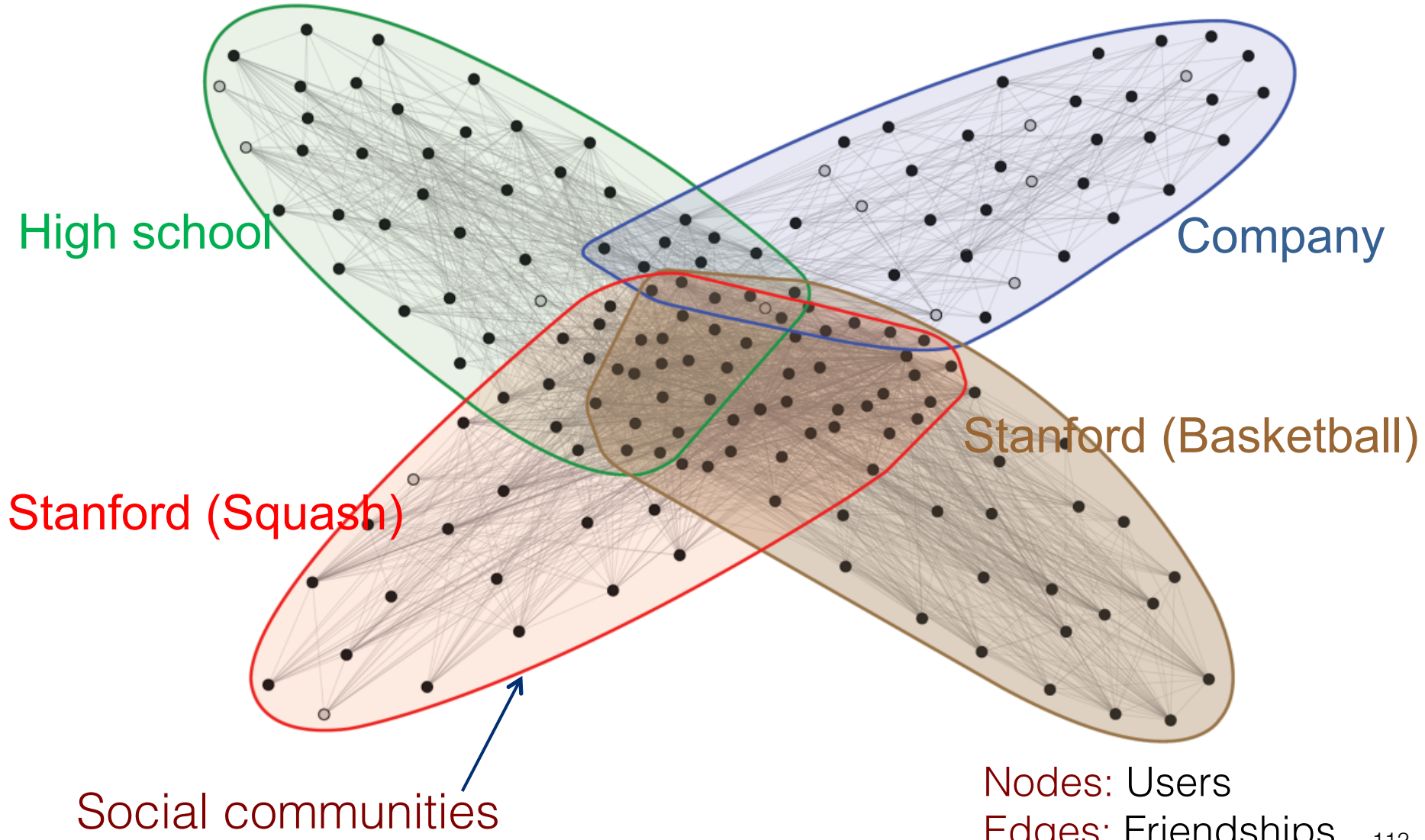
Facebook Ego-network



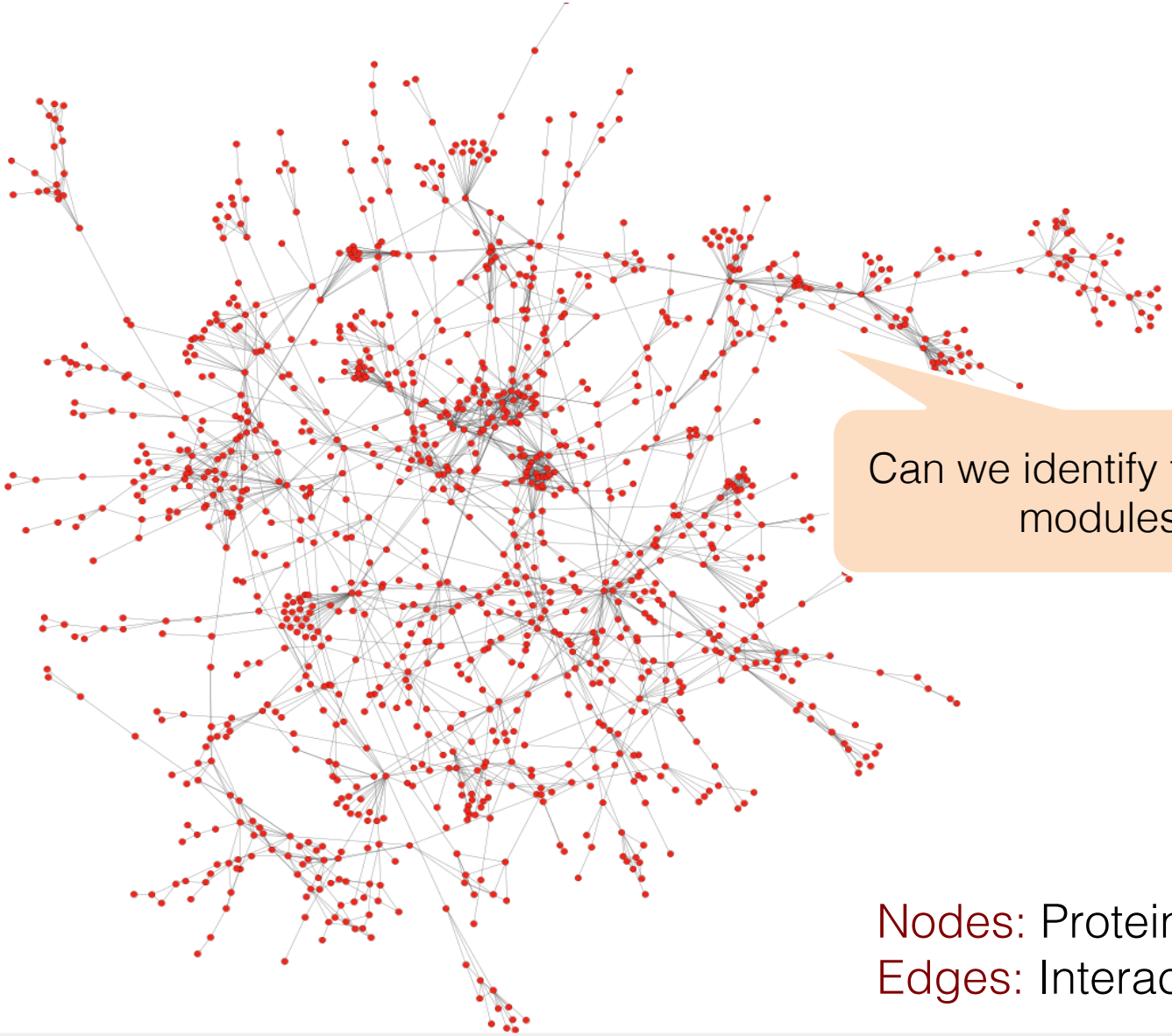
Can we identify social communities?

Nodes: Users
Edges: Friendships

Facebook Ego-network



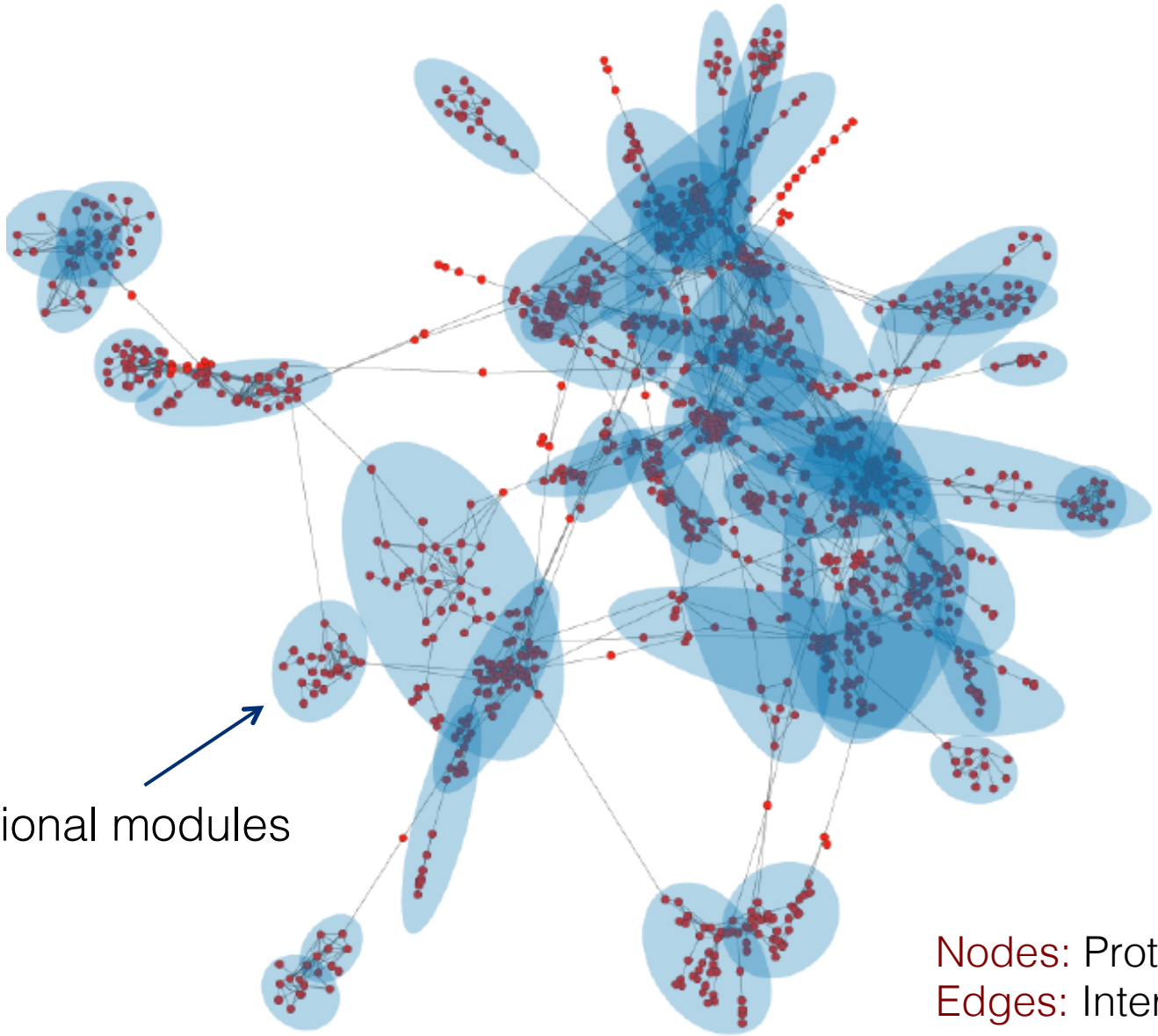
Protein-Protein Interactions



Can we identify functional modules?

Nodes: Proteins
Edges: Interactions

Protein-Protein Interactions



Functional modules

Nodes: Proteins
Edges: Interactions

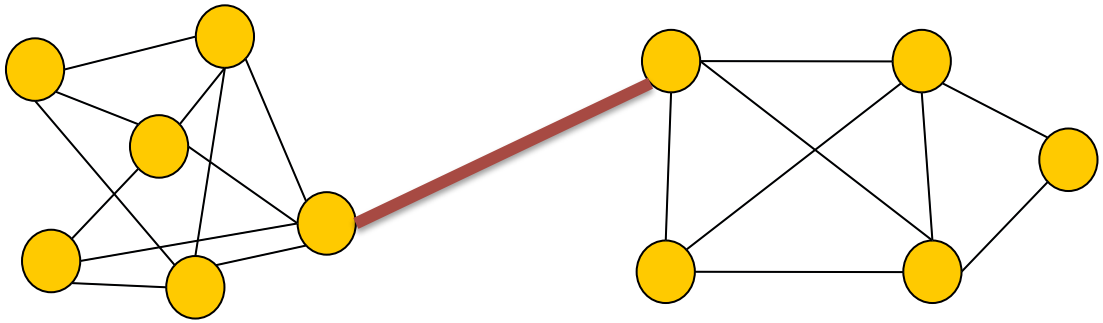
Community Detection

How to find communities?

Girvan-Newman's Method

- Divisive hierarchical clustering based on the notion of edge betweenness centrality
  - Number of shortest paths passing through the edge

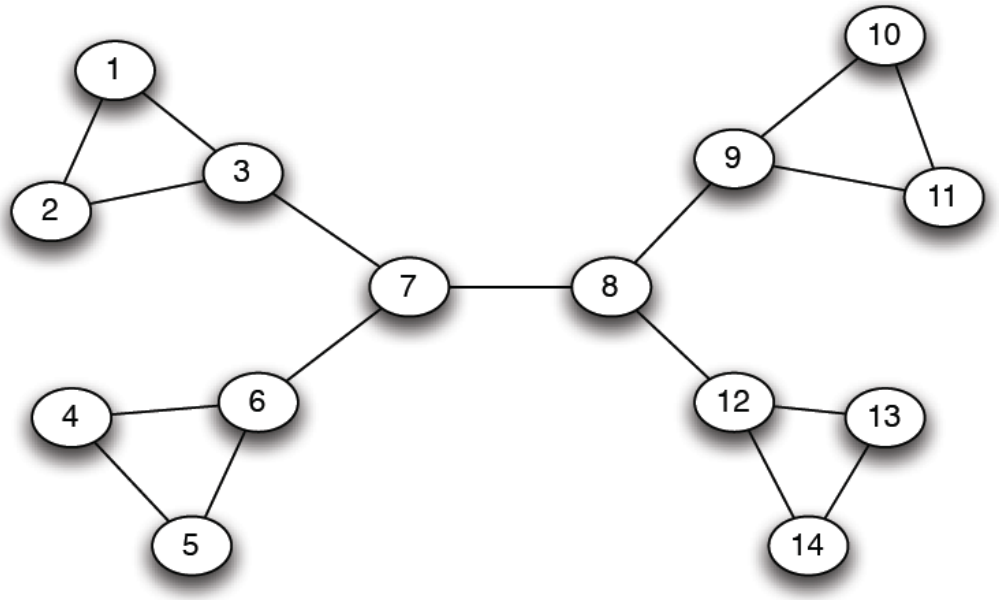
- Algorithm
  1. Calculate the betweenness centrality of all edges in the graph
  2. Remove the edge with the highest betweenness score
  3. Recalculate betweenness for all edges affected by the removal
  4. Repeat step 2 until no edges remain



Try to identify the edges of the graph that are most between other vertices
- Responsible for connecting many node pairs

[Girvan and Newman '02]

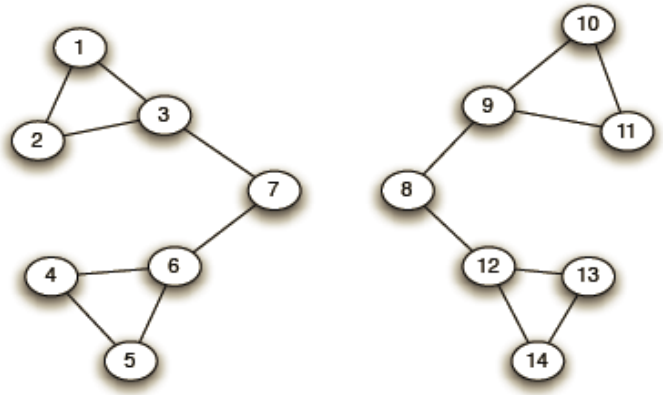
Girvan-Newman Algorithm - Example



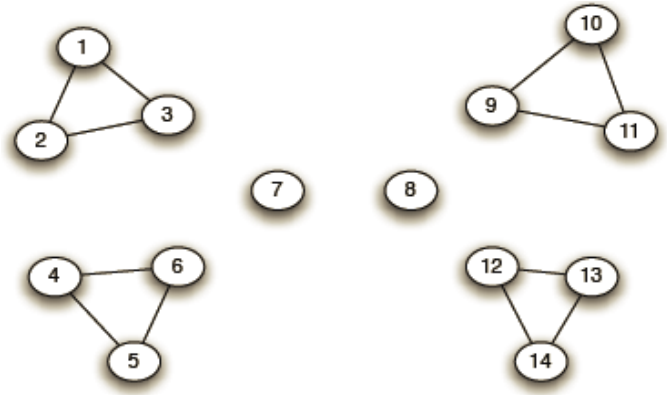
Edge with the highest betweenness centrality score?

Girvan-Newman Algorithm - Example

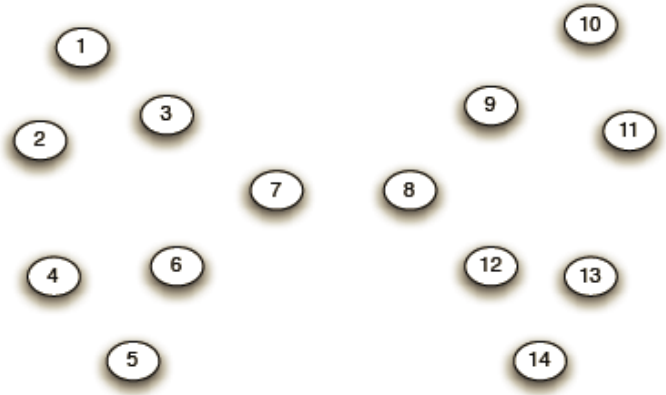
Step 1:



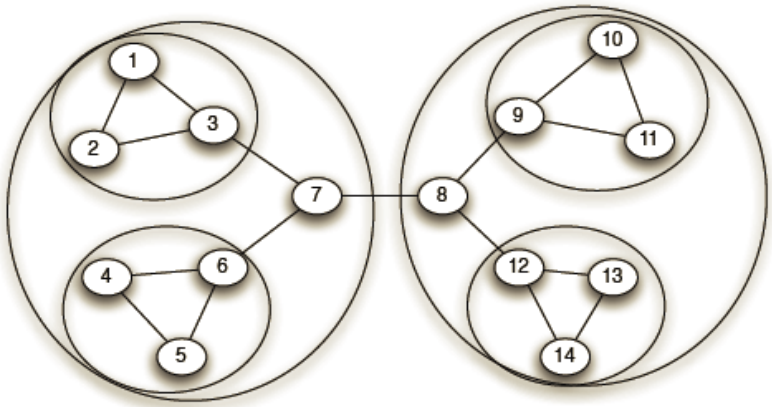
Step 2:



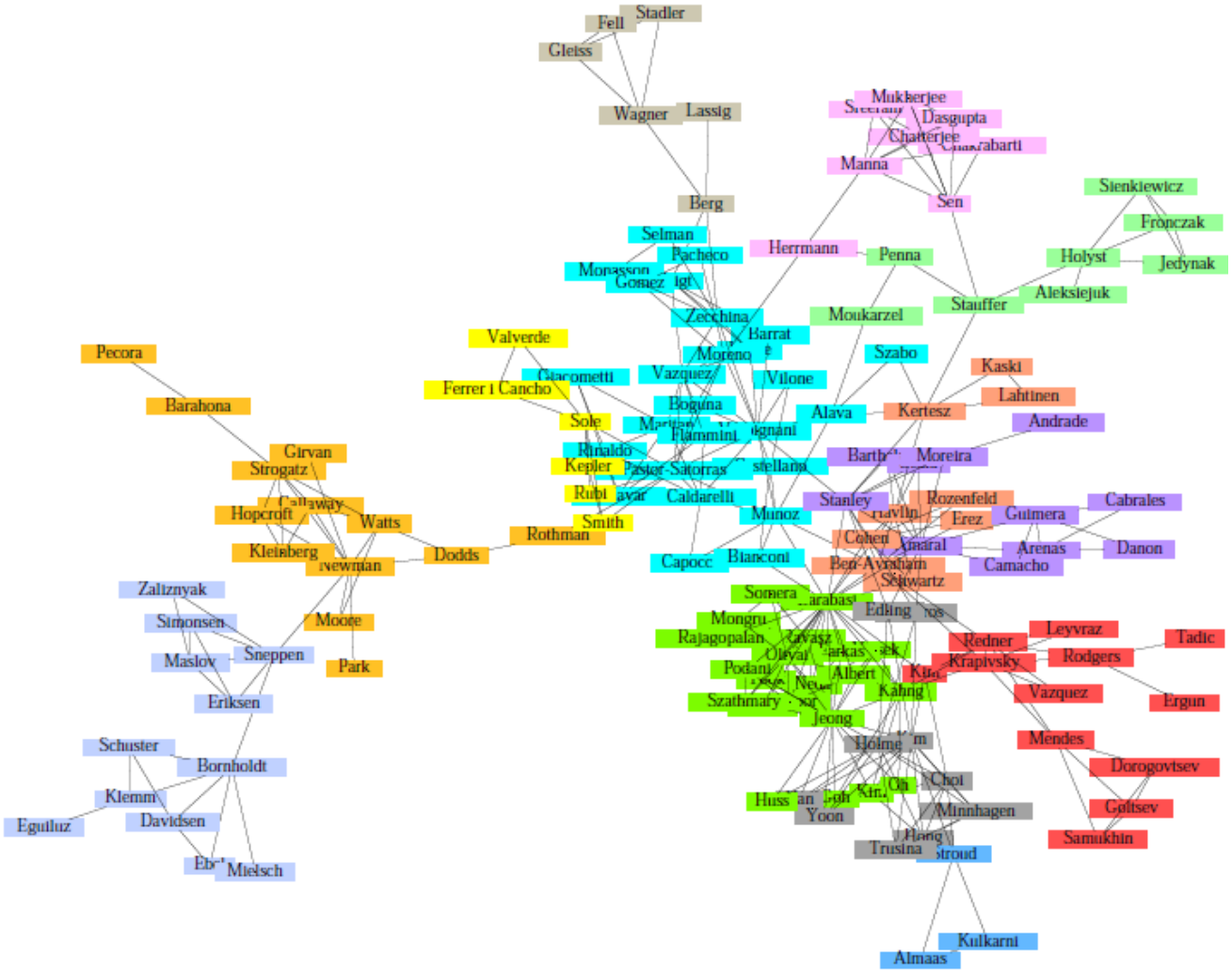
Step 3:



Hierarchical network decomposition:

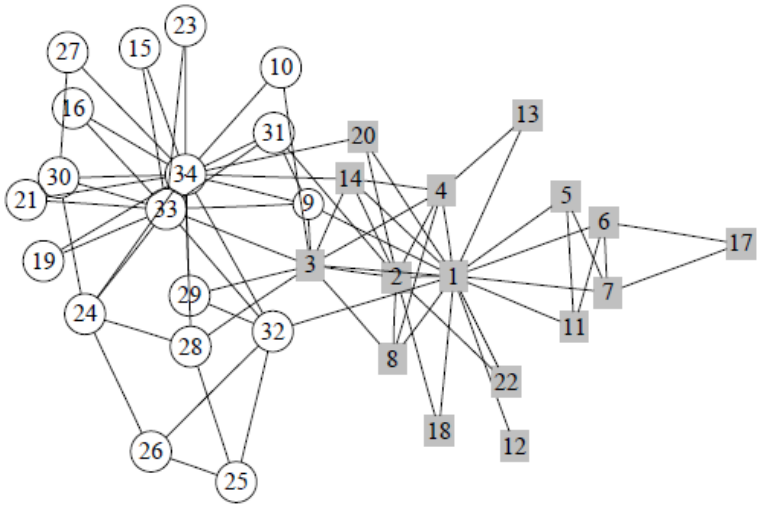


Scientific Collaboration Network

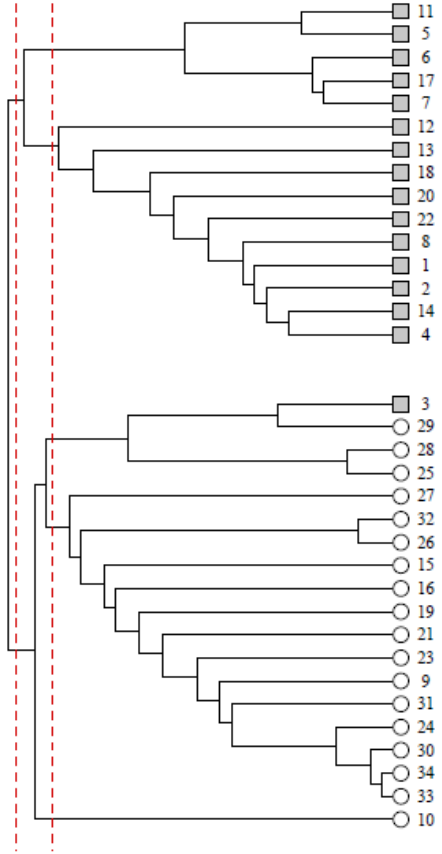


Communities in physics collaborations

Zachary's Karate Club



Zachary's karate club



Dendrogram

- Which of the divisions is the most useful (or optimal)?
  - Need to define metrics of quality of the community structure

How to Evaluate the Quality of Communities

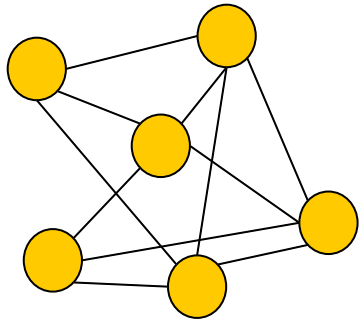
- Typically, two criteria
  - Internal connectivity (intra-community edges)
  - External connectivity (inter-community edges)

- Q: Is there any other way to distinguish groups of nodes with good community structure?

- Random graphs are not expected to present inherent community structure
- Idea: Compare the number of edges that lie within a community to the expected one in case of random graphs with the same degree distribution
  - Modularity measure

Modularity: Main Idea

- Modularity function Q
- Initially introduced as a measure for assessing the strength of communities
  - $Q = \sum_{c \in C}$ (number of edges within community c) – (expected number of edges within community c)
- What is the expected number of edges?
- Consider a configuration model
  - Random graph model with the same degree distribution
  - Let P_{ij} = probability of an edge between nodes i and j with degrees k_i and k_j respectively
  - Then $P_{ij} = k_i k_j / 2m$, where $m = |E| = \frac{1}{2} \sum_i k_i$



[Newman and Girvan '04], [Newman '06]

Formal Definition of Modularity

- Modularity Q

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

where

- A is the adjacency matrix
- k_i, k_j are the degrees of nodes i and j respectively
- m is the number of edges in the graph
- C_i is the community of node i
- δ(.) is the Kronecker function: 1 if both nodes i and j belong to the same community ($C_i = C_j$), 0 otherwise

[Newman and Girvan '04], [Newman '06]

Properties of Modularity

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

- Larger modularity Q indicates better communities (more than random intra-cluster density)
  - The community structure is better if the number of internal edges exceed the expected number
  - Q in the range of 0.3 - 0.7 means significant community structure
- Modularity value is always $-1 < Q < 1$
- It can also take negative values
  - E.g., if each node is a community itself
  - No partitions with positive modularity → No community structure
  - Partitions with large negative modularity → Existence of subgraphs with small internal number of edges and large number of inter-community edges

[Newman and Girvan '04], [Newman '06], [Fortunato '10]

Back to the Girvan-Newman Algorithm

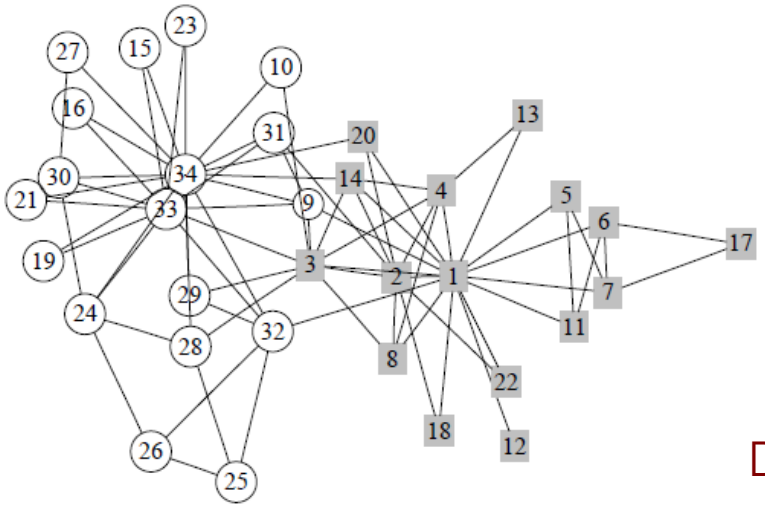
- Basic steps:
  1. Compute betweenness centrality for all edges in the graph
  2. Find and remove the edge with the highest score
  3. Recalculate betweenness centrality score for the remaining edges
  4. Go to step 2

- How do we know if the produced communities are of good quality in order to stop the algorithm?
  - The output of the algorithm is in the form of a dendrogram
  - Use modularity as a criterion to cut the dendrogram and terminate the algorithm (Q ~= 0.3-0.7 indicates good partitions)

- Complexity: $O(m^2n)$ (or $O(n^3)$ in sparse graphs)

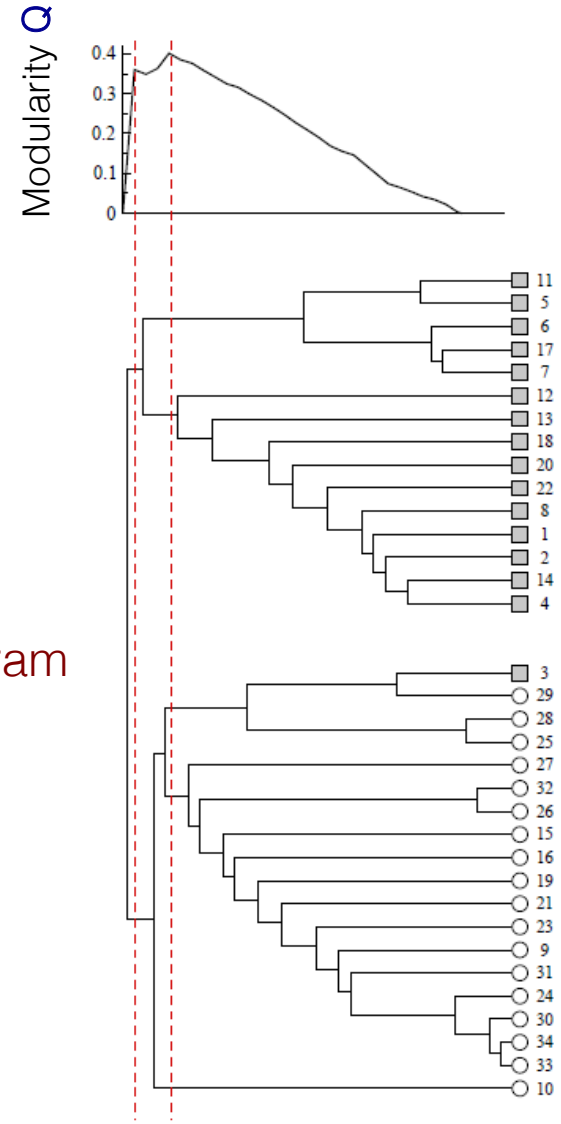
[Newman and Girvan '04], [Girvan and Newman '02]

Zachary's Karate Club – Modularity



Zachary's karate club

Dendrogram



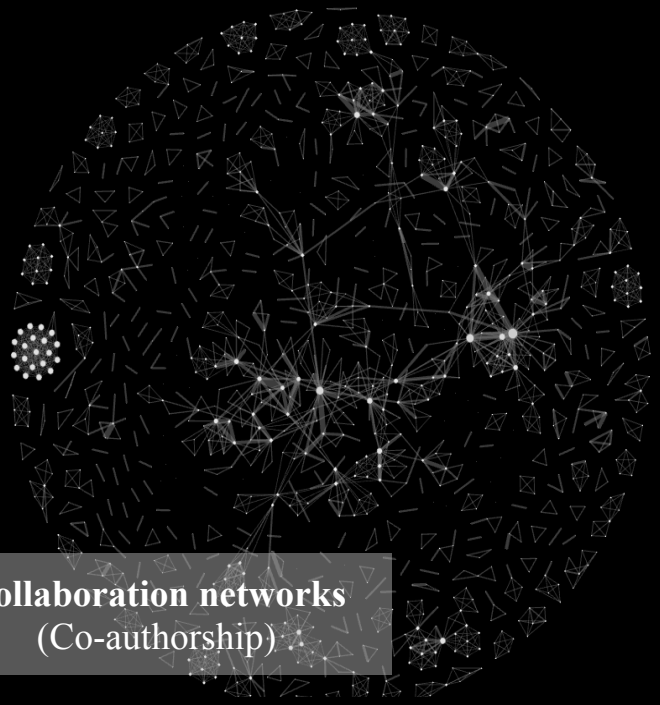
Applications of Modularity

- Modularity can be applied:
  - As quality function in clustering algorithms
  - As evaluation measure for comparison of different partitions or algorithms
  - As criterion for reducing the size of a graph
    - Size reduction preserving modularity [Arenas et al. '07]
  - As a community detection algorithm itself
    - Modularity optimization

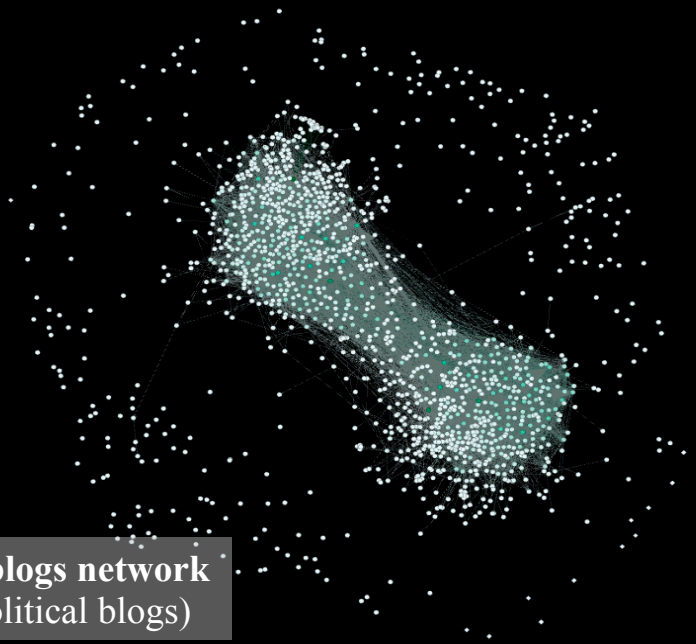
Summary

Online Social Networks

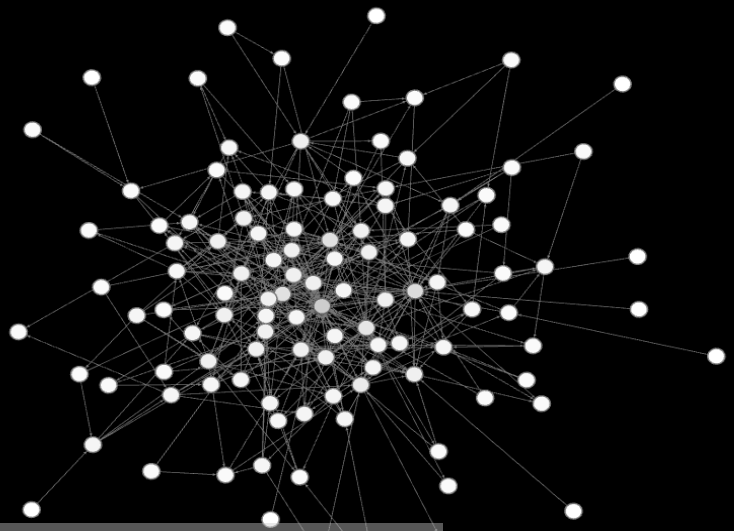
Source: https://www.facebook.com/zuck



Collaboration networks
(Co-authorship)



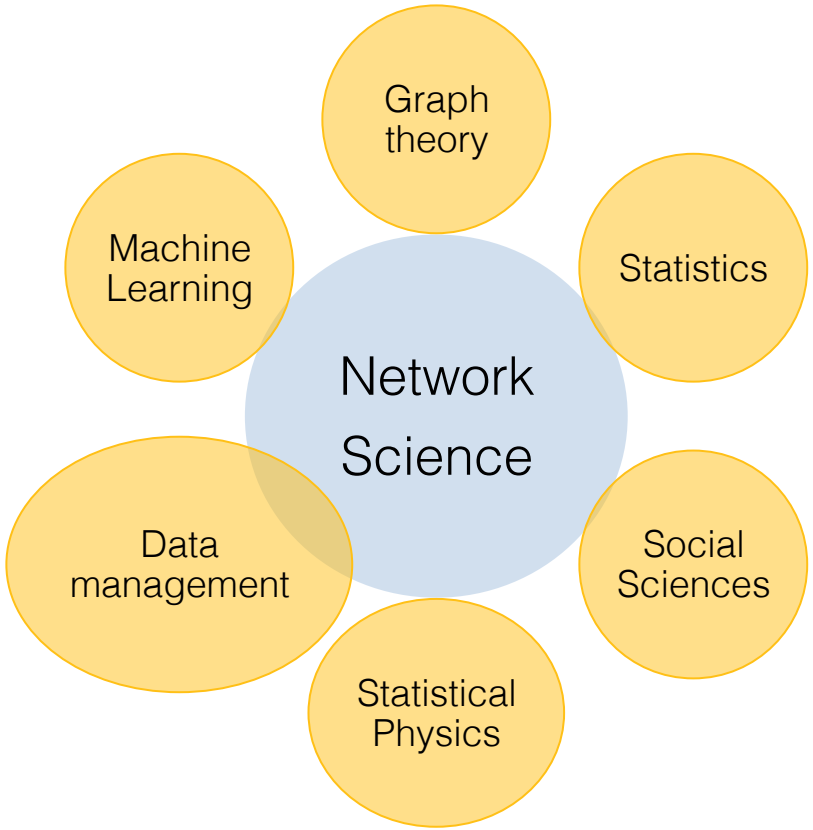
Weblogs network
(Political blogs)



Term co-occurrence network
(*David Copperfield* novel by Charles Dickens)

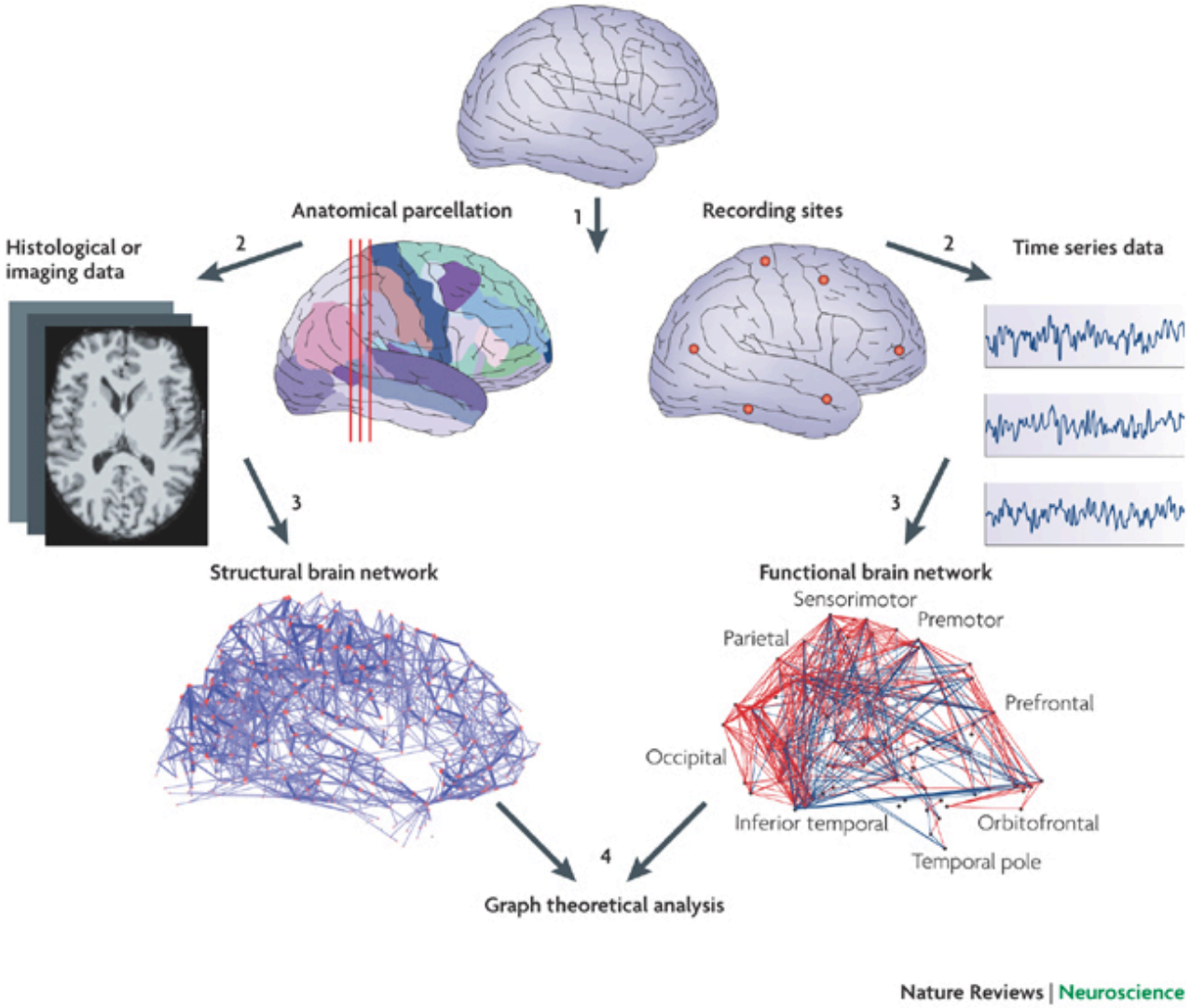
Network Science

Discovering, analyzing and making sense of network data



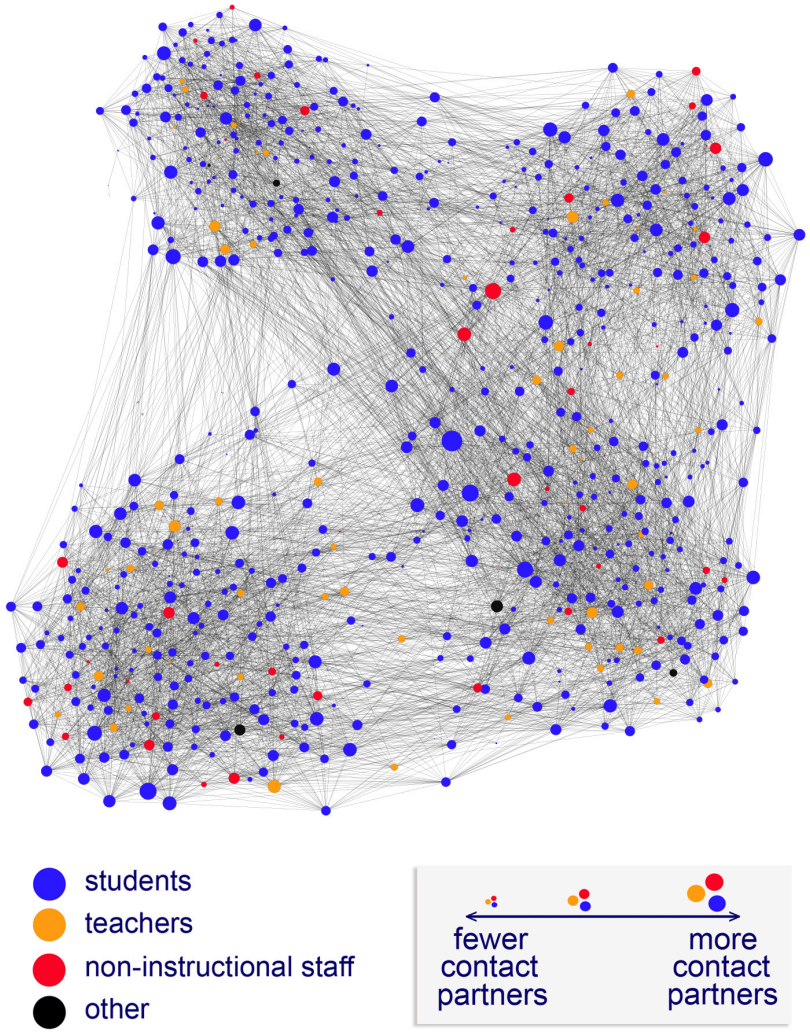
Plethora of applications

Complex Brain Networks



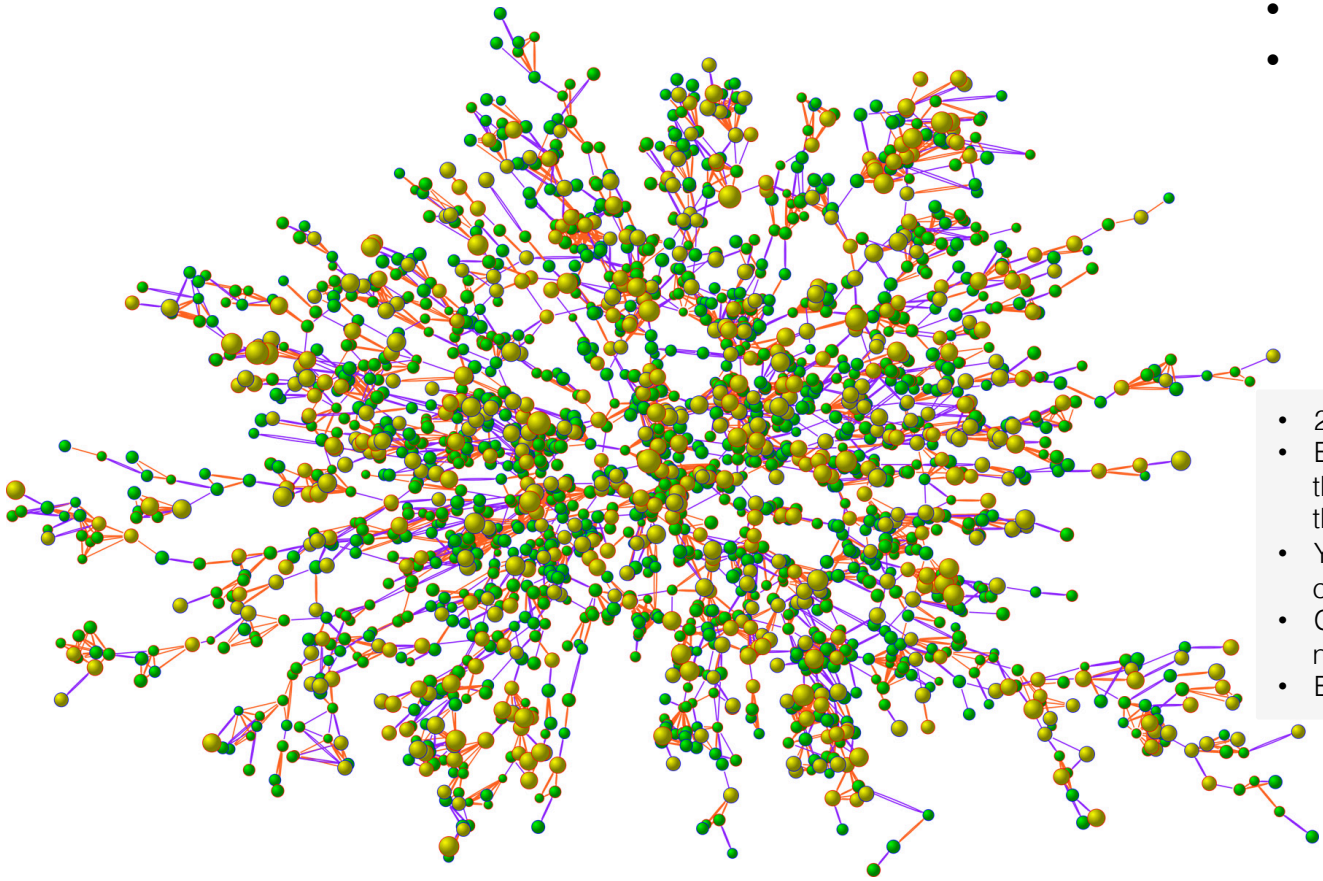
[Bullmore and Sporns, Nature Reviews Neuroscience '09]

Epidemiology – Disease Network



[Smieszekand and Salathé, BMC Medicine ‘13]

The Spread of Obesity



- 5124 key subjects
- 1971 – 2003
- increase in obesity during this period cannot be explained by genetics

- 2,200 people
- Each node represents one person, and the size of each circle is proportional to that person's body-mass index (BMI)
- Yellow circles indicate people who are considered medically obese
- Green circles indicate people who are not obese
- Edges indicate family and friendship ties

Obesity is socially contagious

[Christakis and Fowler, N Engl J Med '00]

Thank You!