

# Applied Data Analysis (CS401)



Lecture 12  
Handling  
networks  
2018/12/13

Robert West



# Announcements

- Projects (“milestone 3”) due on Sun 12/16, 23:59
- Looking for semester or MS projects?
  - We have some [openings](#)
  - Cave canem: we’re very strict in selecting students
- Tomorrow’s lab session:
  - Project office hours
  - Talk: Christian Sinobas, KiWi



Give us feedback on this lecture here:

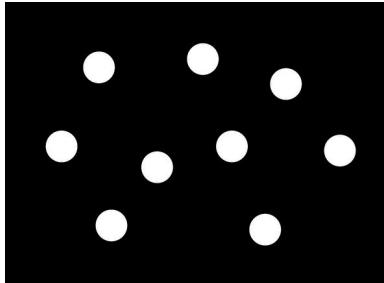
<https://go.epfl.ch/ada2018-lec12-feedback>

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- ...

# Beyond flat tables

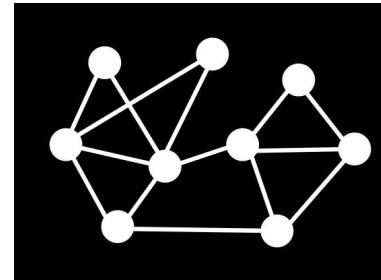
People

<b>id</b>	<b>name</b>	<b>age</b>
1	Bob	36
2	Willy	32
...	...	...

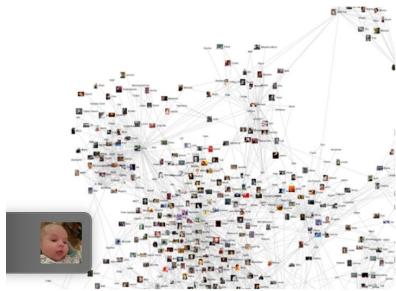


Marriages

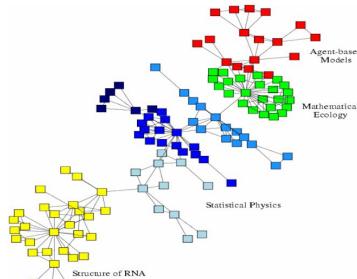
<b>husband_id</b>	<b>wife_id</b>
1	34
2	5
2	87
...	...



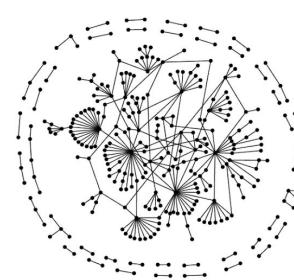
# Examples



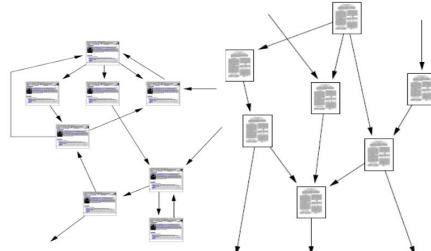
Social networks



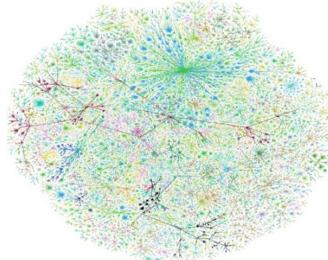
Economic networks



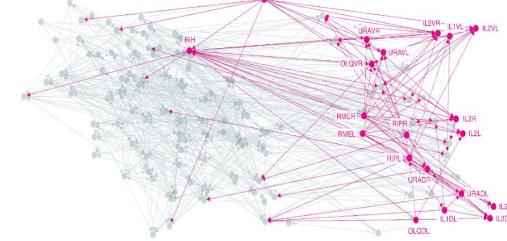
Communication graphs



Information networks:  
Web & citations

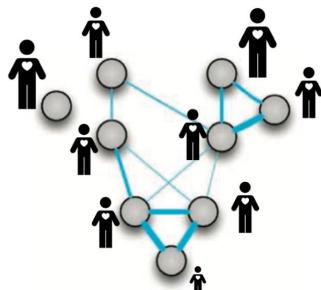


Internet

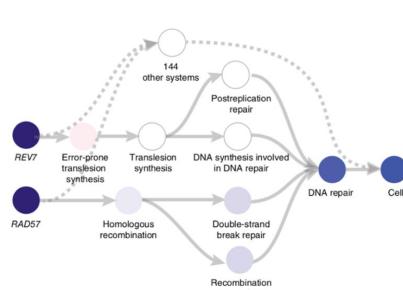


Networks of neurons

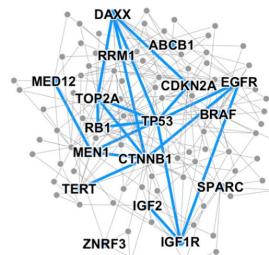
# Examples



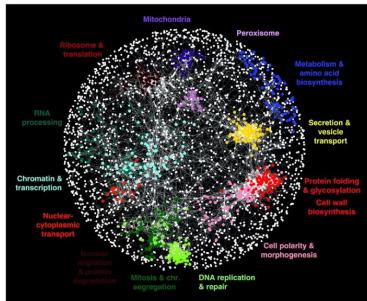
## Patient networks



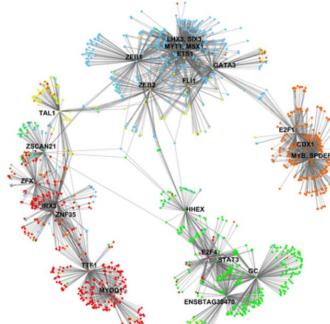
## Hierarchies of cell systems



## Disease pathways



# Genetic interaction networks



# Gene co-expression networks



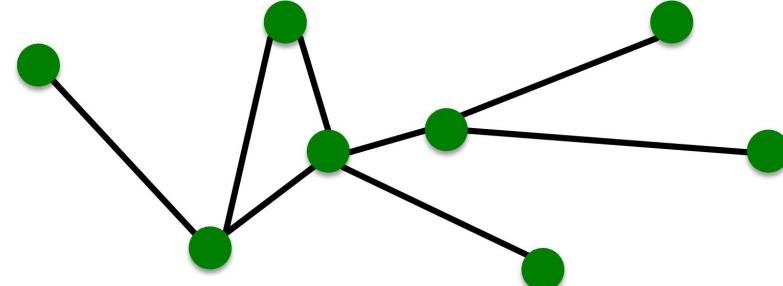
## Cell-cell similarity networks

# Networks as graphs

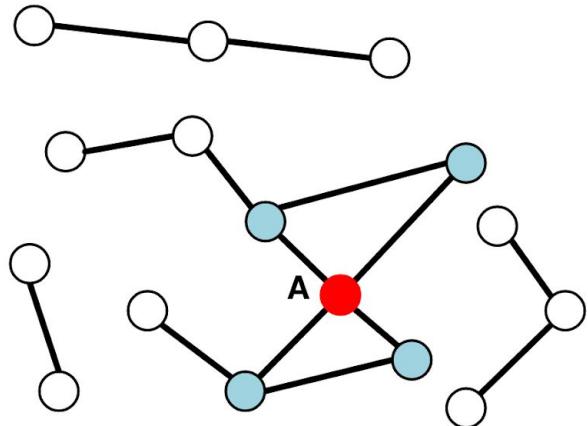
- **Network:** a real-world system of dependent variables, e.g.,
  - WWW is a network of hyperlinked documents
  - Society is a network of individuals linked by family, friendship, professional ties
  - Metabolic network is sum of all chemical reactions in a cell
- **Graph:** mathematical abstraction for describing networks
- In practice, “network” and “graph” are used interchangeably
- You can make a graph out of almost anything (e.g., connect all people whose name starts with the same letter), so must ask:  
Does this graph correspond to a meaningful network?

# Most basic type: undirected graphs

- Entities:  
nodes/vertices  $V$  
- Relationships/interactions:  
edges/links/arcs  $E$  
- Entire system:  
network/graph  $G = (V, E)$



# Node degree



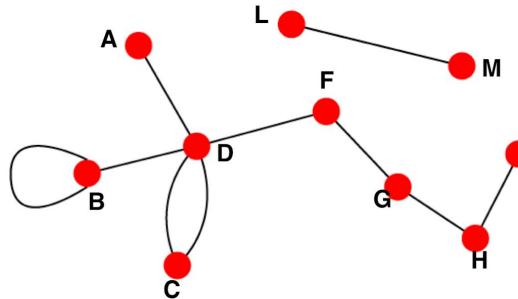
**Node degree,  $k_i$ :** the number of edges adjacent to node  $i$

$$k_A = 4$$

# Types of graphs: undirected vs. directed

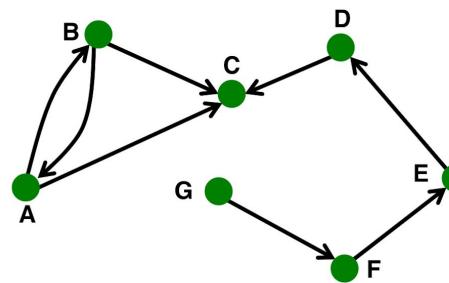
## Undirected

- Links: undirected  
(symmetrical, reciprocal)



## Directed

- Links: directed  
(arcs)



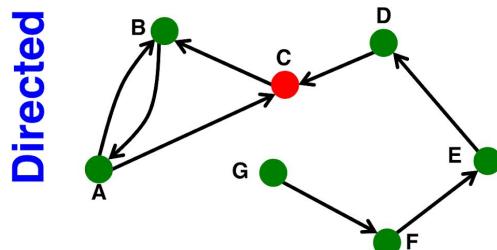
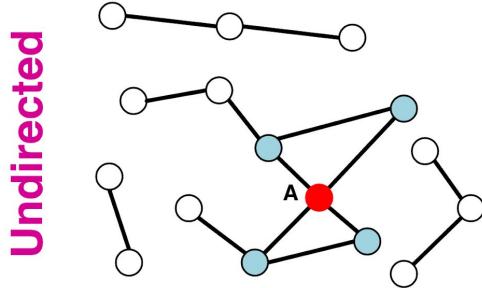
## Examples:

- Collaborations
- Friendship on Facebook

## Examples:

- Phone calls
- Following on Twitter

# Average node degree



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

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

$$k_A = 4$$

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

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.

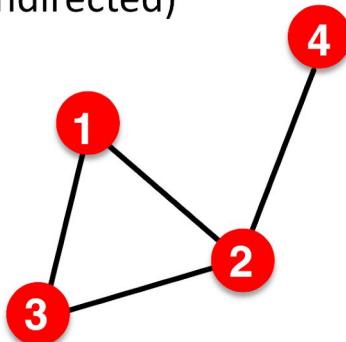
$$k_C^{in} = 2 \quad k_C^{out} = 1 \quad k_C = 3$$

$$\bar{k} =$$

# Types of graphs: weighted

- **Unweighted**

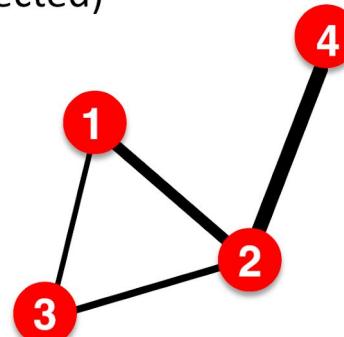
(undirected)



**Examples:** Friendship, Hyperlink

- **Weighted**

(undirected)



**Examples:** Collaboration, Internet, Roads

# Types of graphs: bipartite

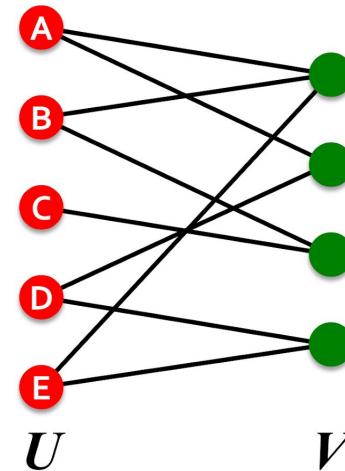
- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets  $U$  and  $V$  such that every link connects a node in  $U$  to one in  $V$ ; that is,  $U$  and  $V$  are **independent sets**

- **Examples:**

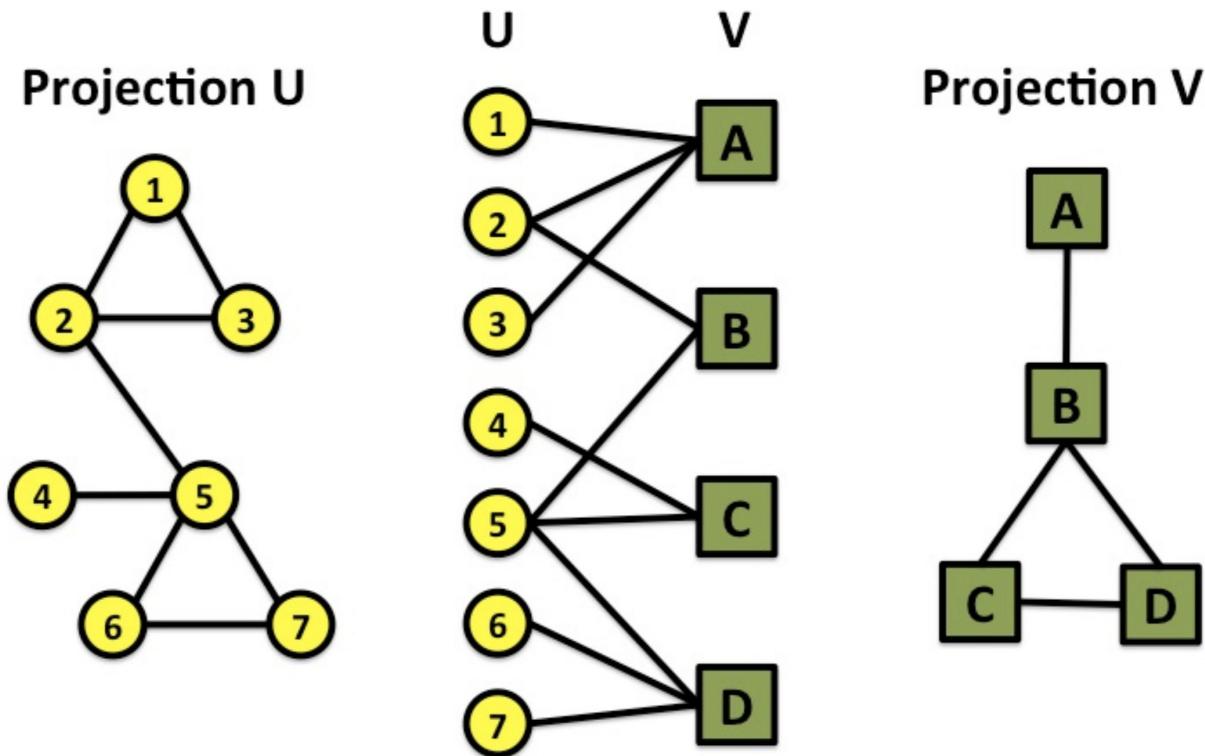
- Authors-to-Papers (they authored)
- Actors-to-Movies (they appeared in)
- Users-to-Movies (they rated)
- Recipes-to-Ingredients (they contain)

- **“Folded” networks:**

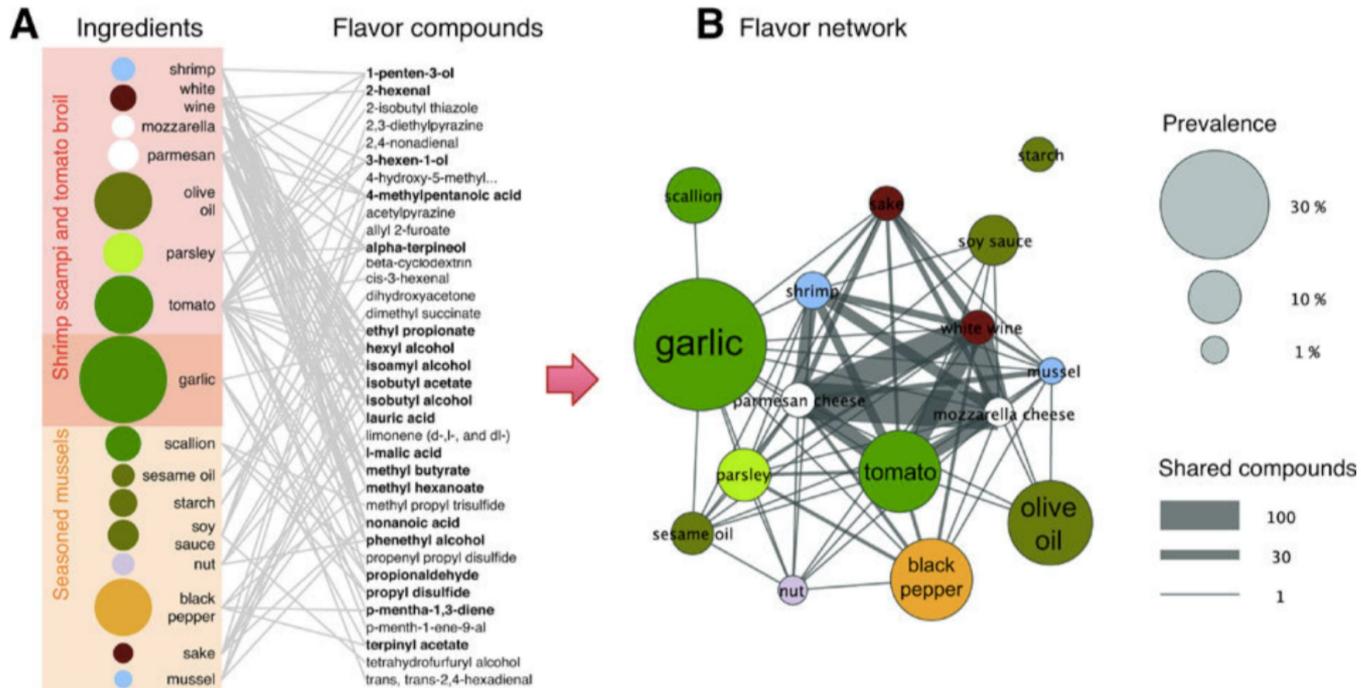
- Author collaboration networks
- Movie co-rating networks



# Projections of bipartite graph



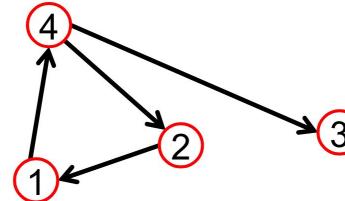
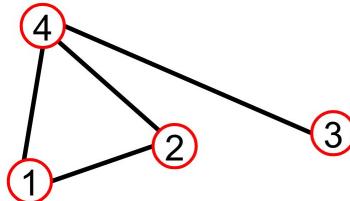
# Example: flavor networks



Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási  
*Flavor network and the principles of food pairing*, Scientific Reports 196, (2011).

Network Science: Graph Theory

# Representing graphs on computers: adjacency matrix



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

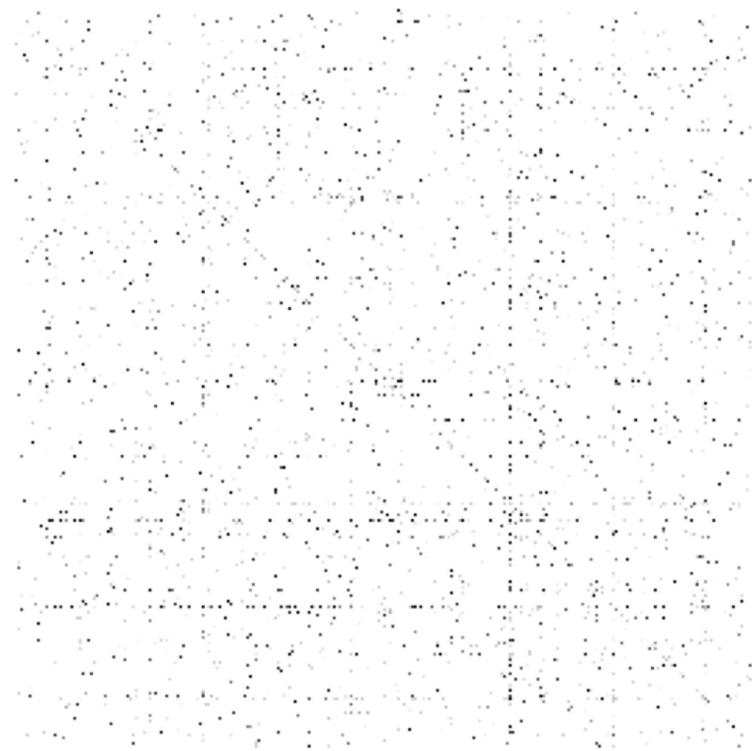
$A_{ij} = 0$  otherwise

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

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

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

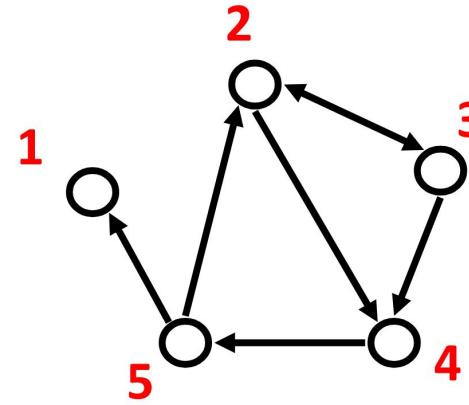
# Adjacency matrix usually sparse



# Representing graphs on computers: edge list

- Represent graph as a set of edges:

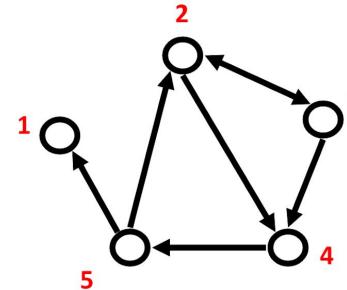
- (2, 3)
- (2, 4)
- (3, 2)
- (3, 4)
- (4, 5)
- (5, 2)
- (5, 1)



# Representing graphs on computers: adjacency list

## ■ Adjacency list:

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

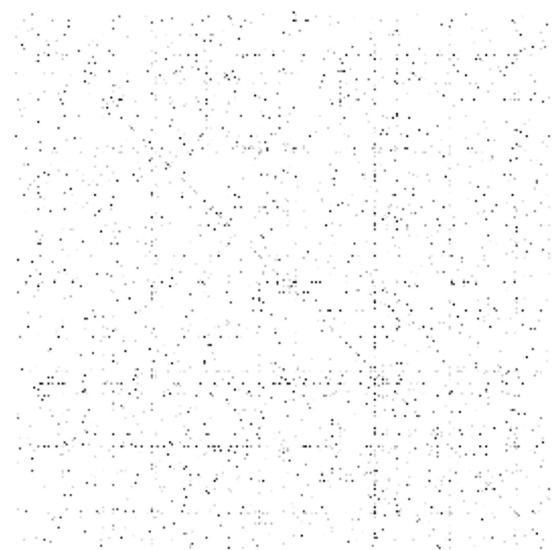


# Properties of real-world networks

- Real networks are different from arbitrary graphs
- Real networks tend to share certain properties
- Remarkable, given the diversity of networks
  - Information networks (e.g., Web graph, knowledge graphs)
  - Social networks (e.g., Facebook, sexual networks)
  - Biological networks
  - ...

# Properties of real-world networks: sparsity

- Every node connected to only small fraction of all other nodes
- i.e.,  $k_i \ll N$
- Often bounded by a constant
  - e.g., social networks: [Dunbar's number](#) (cognitive limit to the number of people with whom one can maintain stable social relationships; allegedly 150)



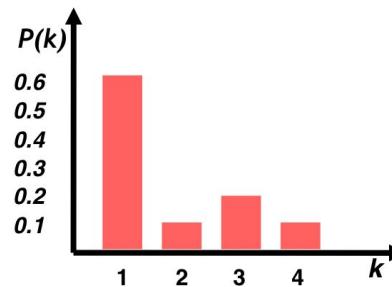
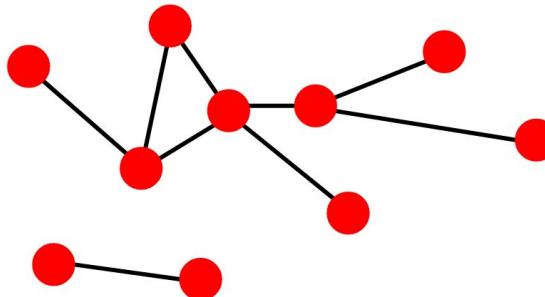
# Properties of real-world networks: degree distribution

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

$$N_k = \# \text{ nodes with degree } k$$

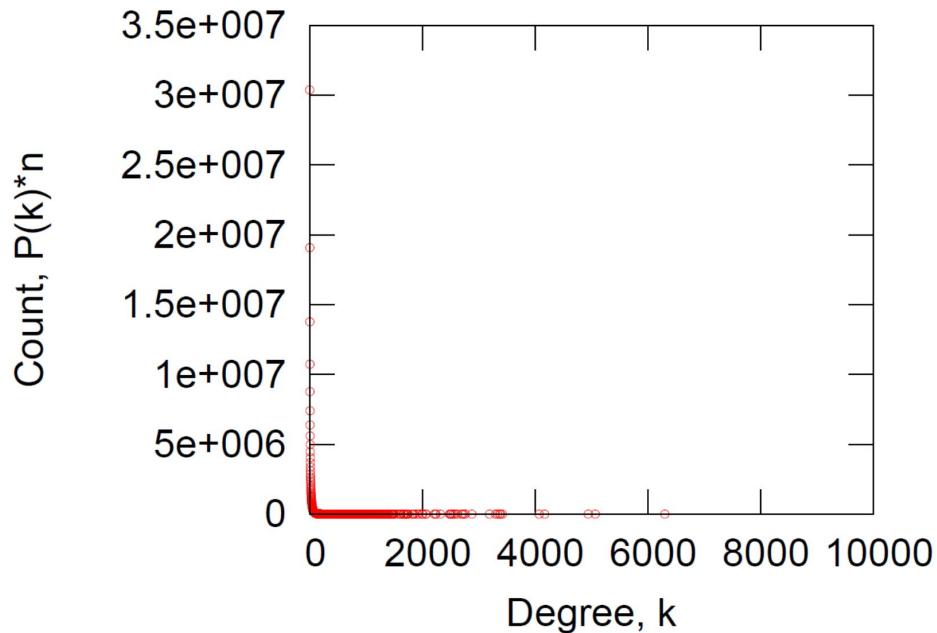
- Normalized histogram:

$$P(k) = N_k / N \rightarrow \text{plot}$$

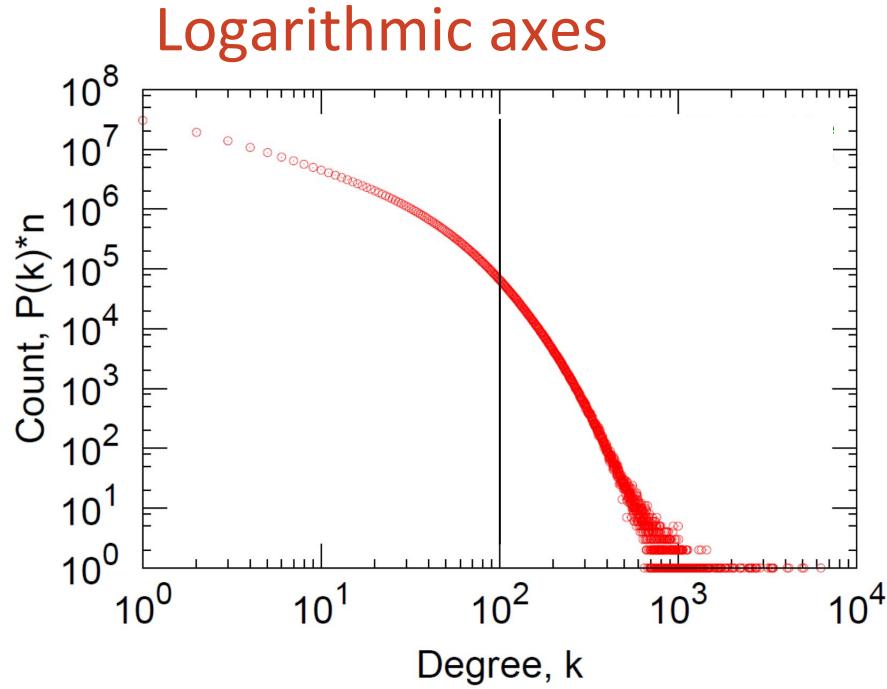


# Properties of real-world networks: degree distribution

Linear axes

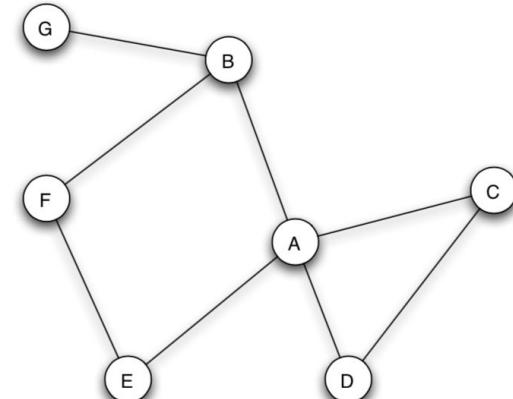


Logarithmic axes



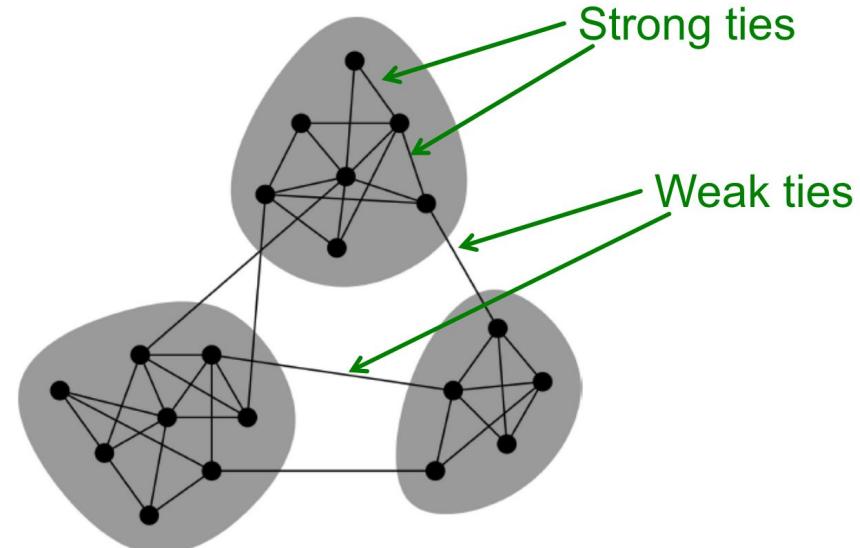
# Properties of real-world networks: triadic closure

- “A friend of my friend is my friend”
- Measured via clustering coefficient  $C_i$  of node  $i$ :  
$$C_i = \frac{\text{\#edges among neighbors of } i}{\text{\#potential edges among neighbors of } i}$$
- In real networks, nodes have large clustering coefficients



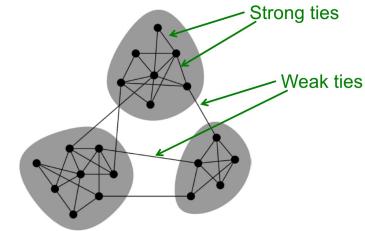
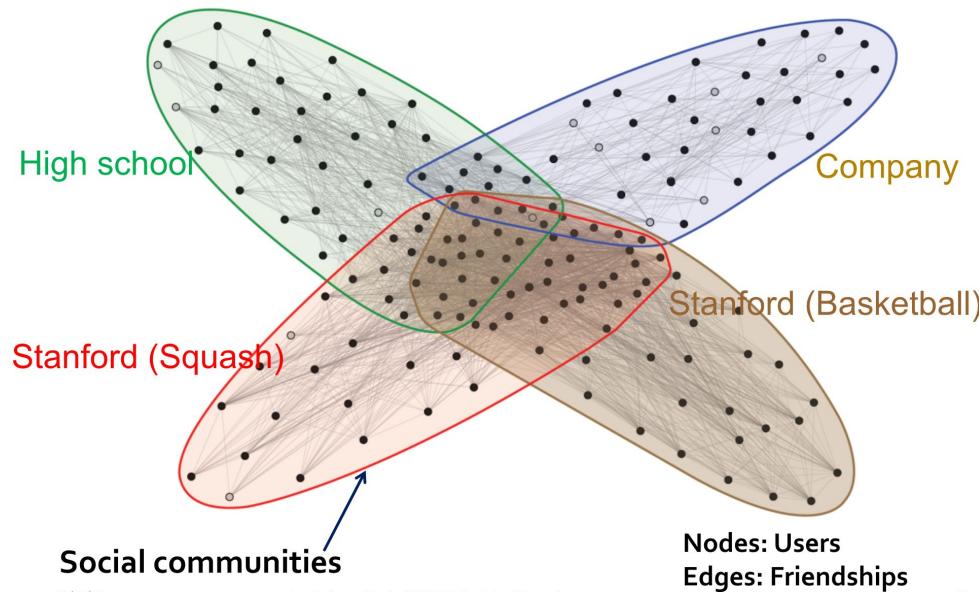
# Properties of real-world networks: community structure

- Triadic closure makes real networks cluster into locally dense “communities”
- Communities connected via “weak ties”
- “Strength of weak ties” (Granovetter)
- “Structural holes”



# Properties of real-world networks: community structure

- In real life, communities are often not disjoint,  
But overlapping:

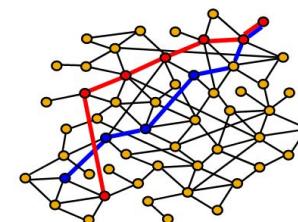


# Properties of real-world networks: average shortest-path length

- What is the typical shortest path length between any two people?
  - Experiment on the global friendship network
    - Can't measure, need to probe explicitly
- Small-world experiment [Milgram '67]
  - Picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - Ask them to get a letter to a stock-broker in Boston by passing it through friends
- How many steps did it take?

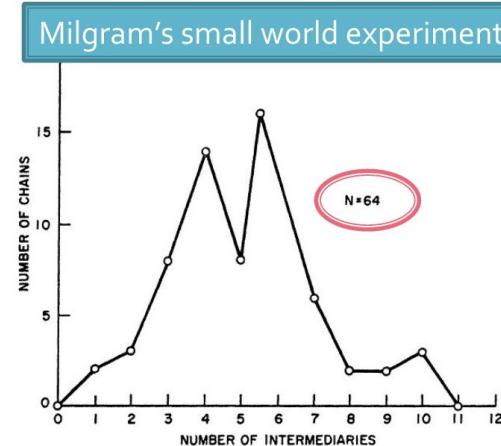


[[Movie](#)]



# Properties of real-world networks: average shortest-path length

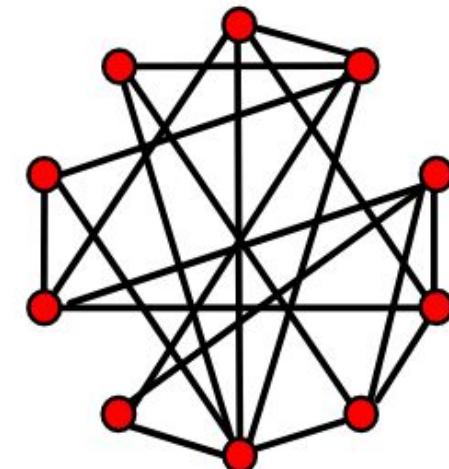
- **64 chains completed:**  
(i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus  
**“6 degrees of separation”**
- **Further observations:**



- People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7
- People from the Boston area have even closer paths: 4.4

# Properties of real-world networks: navigability

- For decades, people focused on the fact that short paths exist in social networks
- But this is true even in random graphs  
(e.g., [Erdős-Rényi model](#))
- The truly amazing fact is not that short paths exist, but that they are discoverable via greedy decentralized routing (as in Milgram's experiment)
- Intrigued? Read Jon Kleinberg:  
[The Small-World Problem: An Algorithmic Perspective](#)



# Don't believe me?

- Play a game and see for yourself how well you can navigate an a-priori unknown network:
  - [Wikispeedia.net](#)

## Wikispeedia

This game is easy and fun:

- You are given two Wikipedia articles\* (or you choose two yourself).
- Starting from the first article, your goal is to reach the second one, exclusively by following links in the articles you encounter. (For the articles you are given this is always possible.)
- Links you can take are colored like [this](#).
- Of course, it's more fun if you try to be as quick as possible...
- Next to wasting some precious time and learning interesting yet useless Wikipedia facts, you're also providing Bob ([west@cs.mcgill.ca](mailto:west@cs.mcgill.ca)) with data for his [research project](#).

\* The articles have been borrowed from the 4,600-article CD version of Wikipedia available at [schools-wikipedia.org](http://schools-wikipedia.org) (version of 2007).

**Let's see if I can find a path from Banjul to Yellow River**

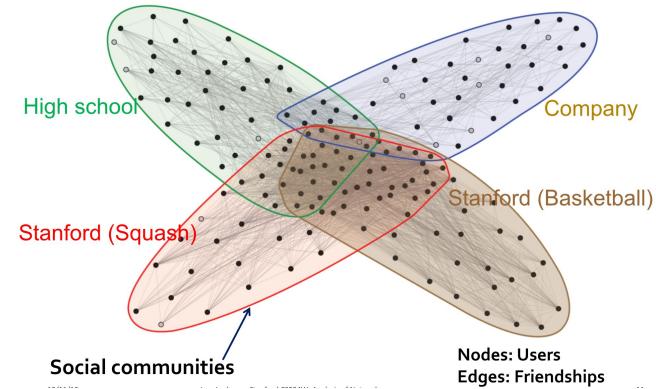
Go!

Gimme another one!

(Click on the target word if you don't know what it means...)

# Properties of real-world networks: homophily/heterophily

- “Birds of a feather flock together”
- Especially in social networks
- Big confound and cause of debate:  
Influence vs. homophily
- E.g., obese people’s friends are more  
likely to also be obese
  - Influence: I copy eating behavior of those around me [[video](#)]
  - Homophily: people with similar eating behavior prone to  
become friends



# Properties of real-world networks: summary

Real-world networks (across many types)

- are sparsely connected,
- but some nodes are much more connected than most others (i.e., skewed degree distribution);
- form locally dense clusters via triadic closure,
- which leads to community structure;
- have short paths between random node pairs (partly due to “hubs” [skewed degree distribution!]),
- and the short paths are easily discoverable.

# How to measure “importance” of a node?

- Formalized via *centrality measures*
- Map each node  $i$  to a scalar value  $C(i)$  capturing its importance in the overall network

# Degree centrality

- Simplest centrality measure
- Many neighbors → important node
  - $C(i)$  = number of neighbors of  $i$
- Very brittle, easy to “rig”
  - E.g., Twitter: scam account, followed by 100,000 other scam accounts

# Closeness centrality

- Farness( $i$ ) = average (or cumulative) distance to  $i$  from other nodes

- $C(i) = 1 / \text{Farness}(i)$

- Reciprocal to turn farness into closeness

- Only defined for connected graphs (otherwise  $d$  infinite)

- Under closeness centrality, nodes that are easy to reach from anywhere in the network are considered important

- Variant: harmonic centrality: switch sum and reciprocal

- $C(i) = \text{average reciprocal distance of } i \text{ to other nodes}$

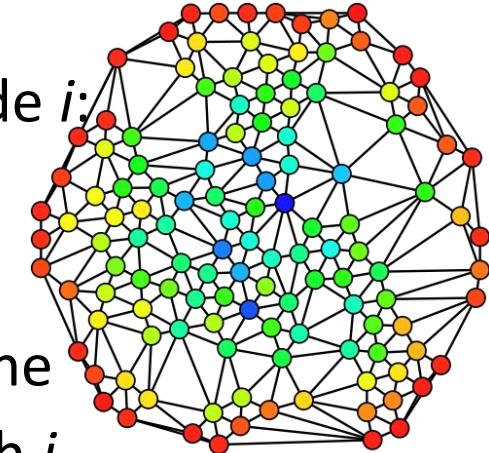
- Defined even for disconnected graphs  
(define  $d$  of disconnected nodes as 0)

$$C(x) = \frac{1}{\sum_y d(y, x)}$$

$$H(x) = \sum_{y \neq x} \frac{1}{d(y, x)}$$

# Betweenness centrality

- $C(i)$  = fraction of all shortest paths in the network that pass through node  $i$
- Computation of betweenness centrality of node  $i$ :
  - For each pair of vertices  $(s, t)$ , compute all shortest paths between them.
  - For each pair of vertices  $(s, t)$ , determine the fraction of shortest paths that pass through  $i$ .
  - Average this fraction over all pairs of vertices  $(s, t)$ .
- Expensive to compute



# Katz centrality

- Generalization of degree centrality
- Degree centrality counts only number of direct neighbors (i.e., neighbors at distance 1)
- Katz centrality also counts neighbors at distances 2, 3, ...
- More precisely, number of paths from other nodes to  $i$ , giving less weight to larger distances:

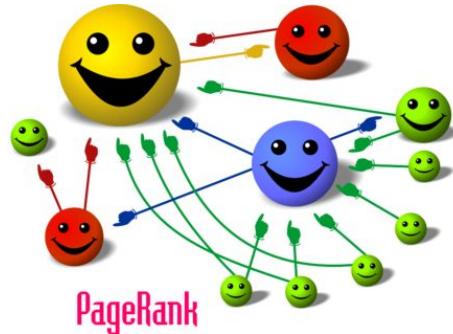
$$C(i) = \sum_{k=1}^{\infty} \sum_{j=1}^N \alpha^k (A^k)_{ji}$$

$k$ -th power of adjacency matrix  $A$  contains number of length- $k$  paths for each node pair

- More robust than degree centrality

# PageRank centrality

- Recursive definition: my centrality  $x_i$  is high if I receive inlinks from many other central nodes:



$$x_i = \sum_j a_{ji} \frac{x_j}{L(j)}$$
$$L(j) = \sum_i a_{ji}$$

$a_{ji}$ : entry  $(j, i)$  of adjacency matrix  $A$   
(1 if  $j$  links to  $i$ , else 0)  
 $L(j)$ : degree (number of neighbors) of  $j$

- Some extra tweaks to make it work with any graph (e.g., disconnected)

# PageRank centrality

$$x_i = \sum_j a_{ji} \frac{x_j}{L(j)}$$

- Matrix notation:  $x = M x$   
(where  $M$  is computed from adjacency matrix  $A$ )
- Do you recognize this?
  - $x$  is eigenvector of  $M$  with eigenvalue 1 → we're in linear-algebra land (home sweet home)
  - $x$  is the steady state the Markov chain induced by the network:  $x_i$  is fraction of time a random walker will spend in node  $i$

# PageRank centrality

- The technology that made Google huge
  - “Page” in PageRank for Larry Page
  - MapReduce was invented to compute PageRank on full Google Web crawl
  - “The \$25,000,000,000 eigenvector” [[link](#)]
- Bottom line:  
Pay attention in your linear algebra class



Give us feedback on this lecture here:

<https://go.epfl.ch/ada2018-lec12-feedback>

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- ...

## Credits

- Some slides borrowed from [CS224W class](#)  
(Jure Leskovec, Stanford)