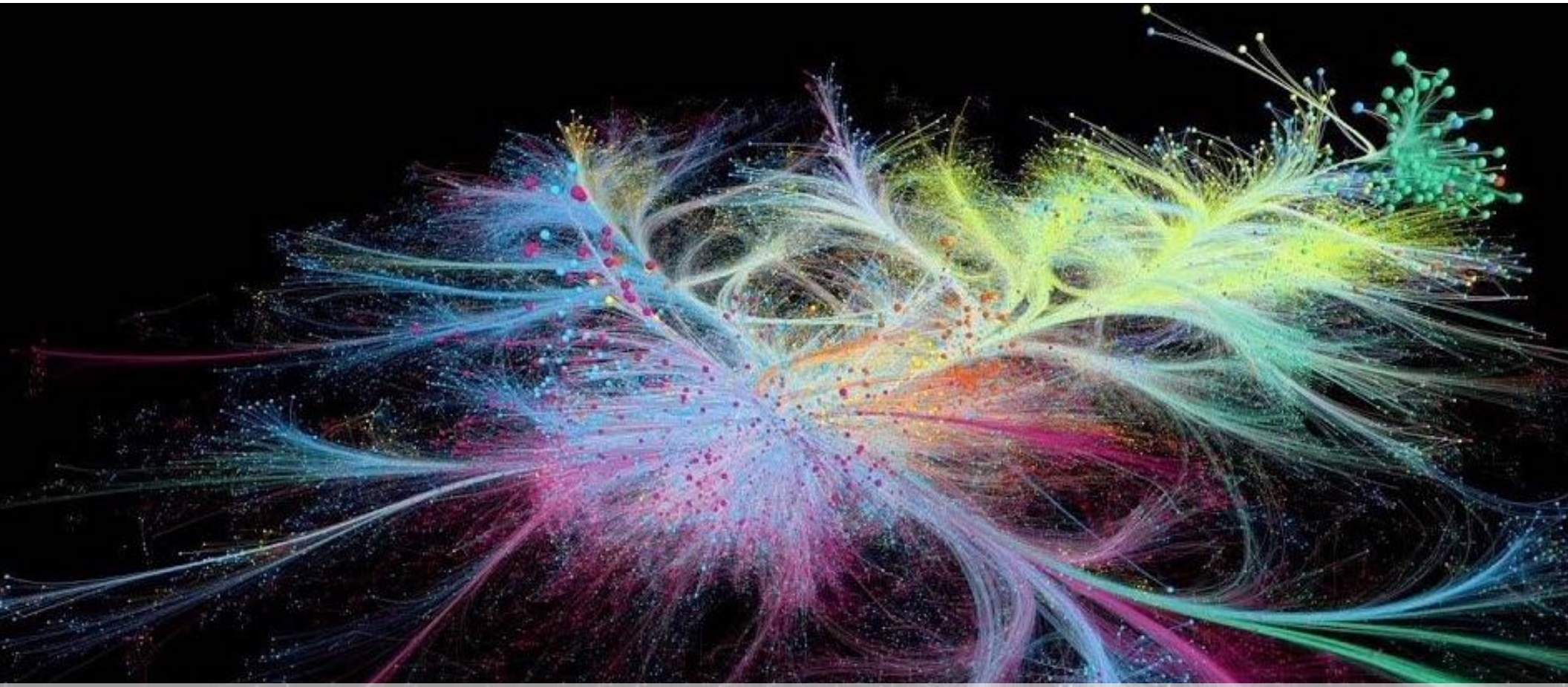


CPSC 572/672: Fundamentals of Social Network Analysis and Data Mining

Admin

Assessment: Project proposal Sept. 27th

Presentation

2 minutes (time limit will be enforced)
4 slides

Discuss:

- What are your nodes and links?
- Describe your dataset. How will you get it?
- Expected size of the network (number of nodes, number of links)
- What questions do you plan to ask? These may change during the course of the class.
- Why do we care about the network you plan to study?

Written component

1 page
Written summary of the details in your presentation.

Project Matchmaking: Week 1

CPSC 572/672 Project Matchmaking

File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago

Share

D8

	A	B	C	D	E	F	G	H	I
1	**Name**	**572/672**	**Skills I have**	**Skills I want to develop**	**Datasets or questions I am interested in**	**Partner**			
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

https://docs.google.com/spreadsheets/d/1vAOhbuaQooo_O5cRZru4QlqgwzTncoXp9URFmTqvfkM/edit?usp=sharing

Project Proposal: [My Social Network Project Title]

Author 1, Author 2, [Author 3 if three]

1. Nodes and Links

Describe in plain language what the node and links are in your social network. This only needs to be a sentence or two.

2. Dataset

Tell us about your dataset. What is the subject matter? Where are you getting it? Most important: How are you getting it and how long will that take you? Describe any expected challenges, need for permissions, which tools you will use, what format the data is in, any expected data cleaning. Note: You may not use data that is already in network format.

3. Expected size of the network

State the number of nodes and number of links. Tell us why – is this everything available? Are you taking a subset? If a subset, why, and how do you plan to define it?

4. Questions you plan to ask and why we care

This is the part to really practice your skills at something akin to a scientific abstract (minus the results). Motivate us – what is so important and/or interesting about your social network? What research questions do you plan to pursue? Do you have any preliminary ideas about how you want to pursue these questions (e.g. community detection)?

Page limit: 1 page
You can use this however you see fit, there is no limit per section. You will all present a variety of networks that require more or less info in each section. You also do not to fill the whole page – succinct is good, just as long as you cover each component as described above.

BUDDY GROUP CONTRACT **DATE: _____**

EXPECTATIONS FROM TEAM MEMBERS

(e.g., Attend all meetings – Bring cinnamon rolls after missing a meeting, Complete project task as promised – Kicked out of team if not completed 3 times, Be open to contributions and ideas from all team members, etc.)

Expectation	**Consequence if expectation not met**
We'll meet as a team every week at: **_____ for _____ hrs and at:** **_____ for _____ hrs**	(One to Two team meetings a week is recommended. Your group may choose more or less. Consequences for missing a few team meetings should be less severe than missing many team meetings.)
Be on time and prepared for team meetings. (Both in and out of class time)	
Follow through on commitments made to the team. (These are likely to be minimal as your projects are individual, but may include committing to giving thoughtful feedback on members' work.)	
Contribute to the team voluntarily. (These contributions may be ideas, questions, code, organizing meetings, managing code repository, creating charts for the report, etc.)	
Welcome and invite contributions by other team members.	*(How will you deal with a team member that consistently dismisses/discourages ideas from other team members or a single team member? It may be helpful to assign a single team member that manages discussions.)*
How you define the line between support and plagiarism. (helpful to play out a scenario, e.g. a buddy group member is stuck with a block of code)	
Each team member will talk for **at least _____ minutes and** **at most _____ minutes** **during a 1-hour meeting.**	*(Some team members have a tendency to dominate discussions and others hesitate to contribute. Set some guidelines on what is expected and how you, respectfully, will make team members aware that they need to contribute more/less.)*
(Think of any other team behaviours that hindered or helped the team performance and formalize this here.)	

If you've read and agreed with this contract, add your name here:

_____ _____

_____ _____

Questions?

Graph Theory

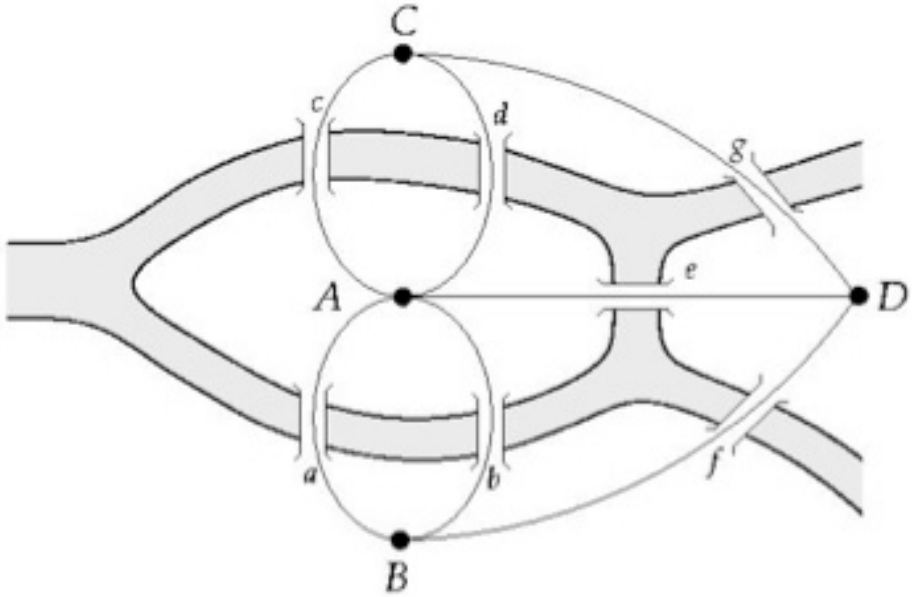
Section 1

The Bridges of Königsberg

THE BRIDGES OF KÖNIGSBERG

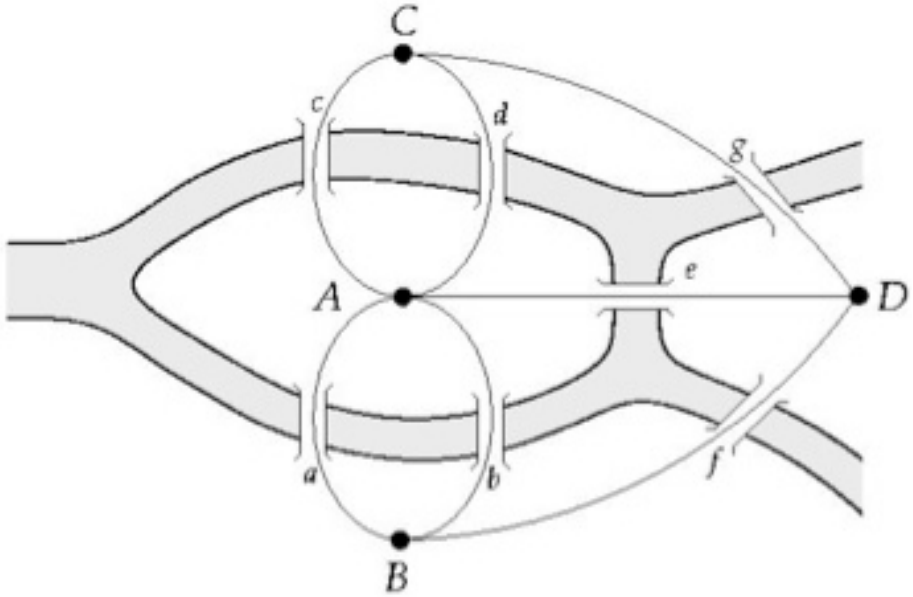
Can one walk across the seven bridges and never cross the same bridge twice?

THE BRIDGES OF KÖNIGSBERG



Can one walk across the seven bridges and never cross the same bridge twice?

THE BRIDGES OF KÖNIGSBERG



Can one walk across the seven bridges and never cross the same bridge twice?

1735: **Euler's theorem:**

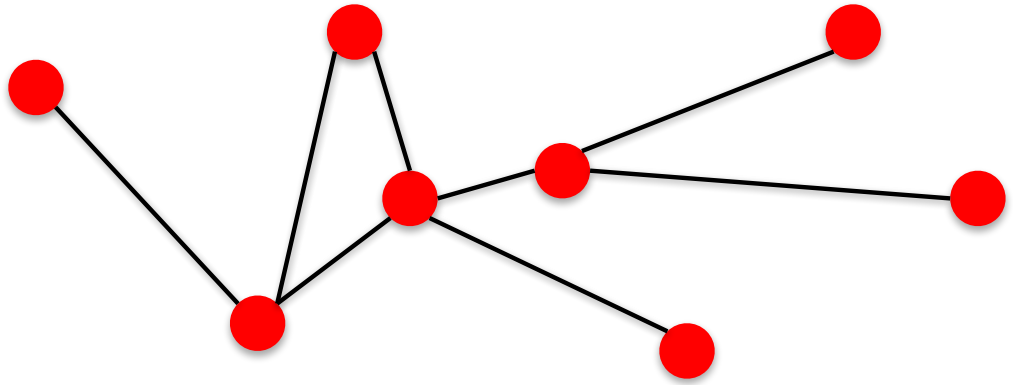
(a) If a graph has more than two nodes of odd degree, there is no path.

(b) If a graph is connected and has no odd degree nodes, it has at least one path.

Section 2

Networks and graphs

COMPONENTS OF A COMPLEX SYSTEM



- **components**: nodes, vertices N
- **interactions**: links, edges L
- **system**: network, graph (N,L)

NETWORKS OR GRAPHS?

network often refers to real systems

- www,
- social network
- metabolic network.

Language: (Network, node, link)

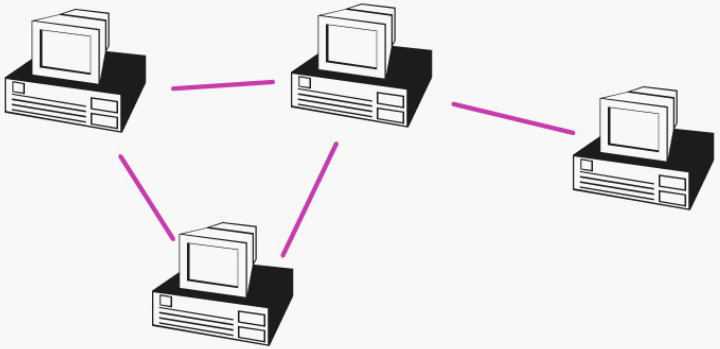
graph: mathematical representation of a network

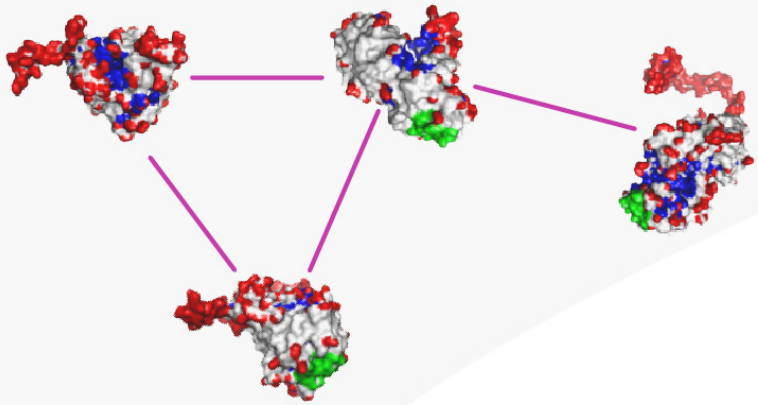
- web graph,
- social graph (a Facebook term)

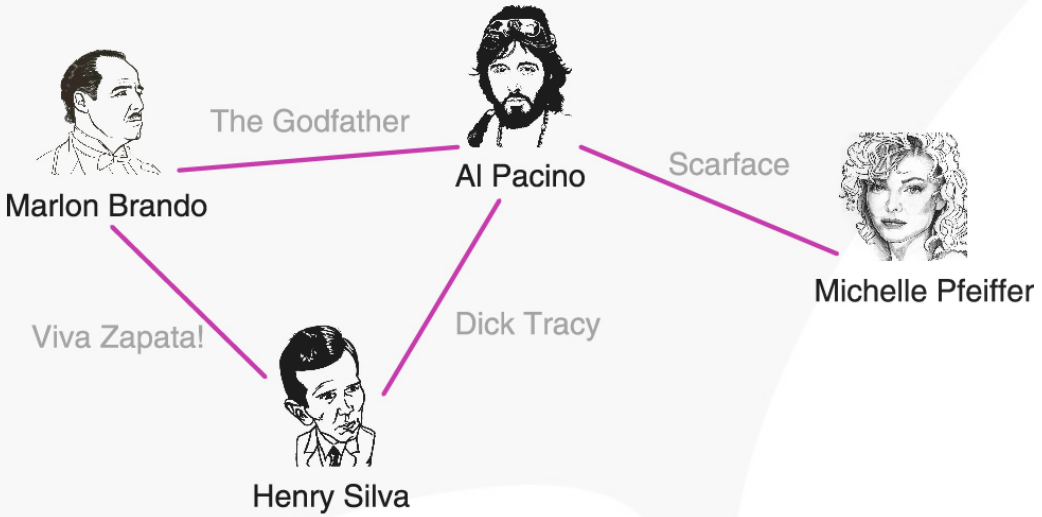
Language: (Graph, vertex, edge)

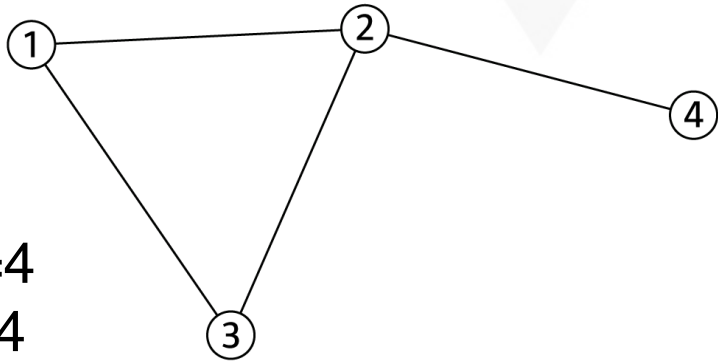
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably.

A COMMON LANGUAGE









N=4
L=4

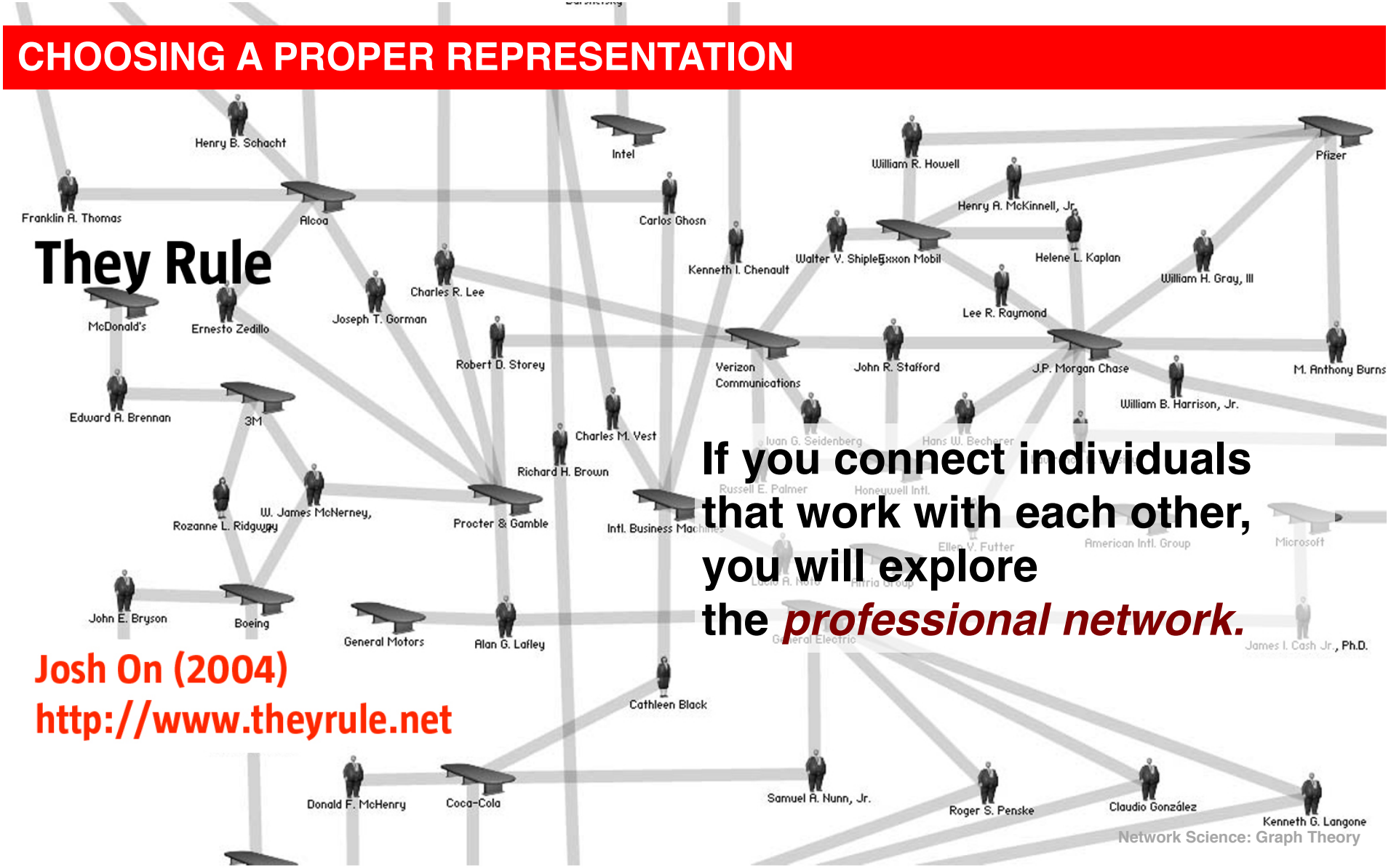
CHOOSING A PROPER REPRESENTATION

The choice of the proper network representation determines our ability to use network theory successfully.

In some cases there is a unique, unambiguous representation.
In other cases, the representation is by no means unique.

For example, the way we assign the links between a group of individuals will determine the nature of the question we can study.

CHOOSING A PROPER REPRESENTATION



CHOOSING A PROPER REPRESENTATION

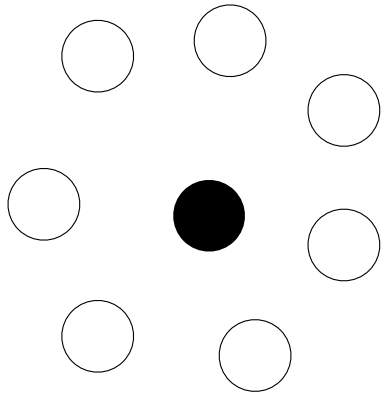
If you connect individuals based on their first name (*all Peters connected to each other*), you will be exploring what?

It is a network, nevertheless.

Activity: Your social network

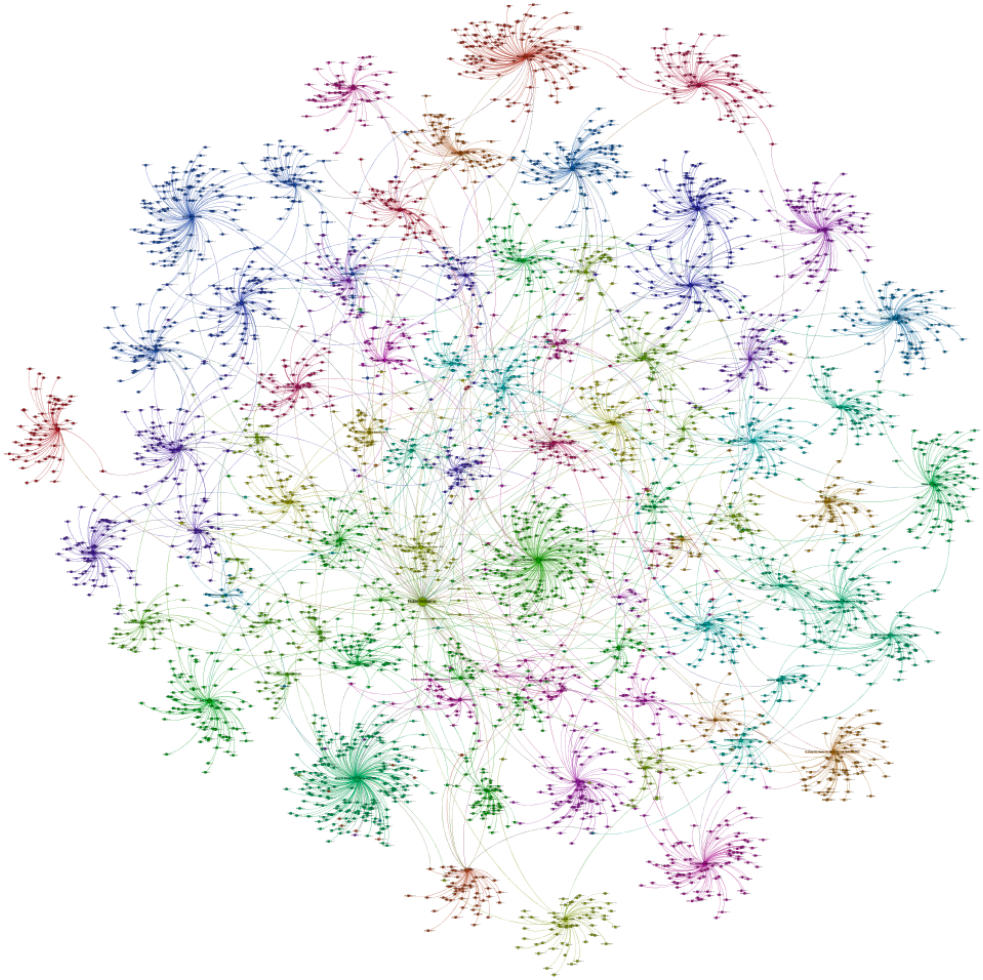
1. Write down a list of the last 10-20 people you interacted with
2. Next to this list, write down the way you know each person (e.g. family, through class, via a hobby…)

3. Re-write that list of 10-20 people in a big circle, with your name in the centre
4. If two people know each other, draw a line between them



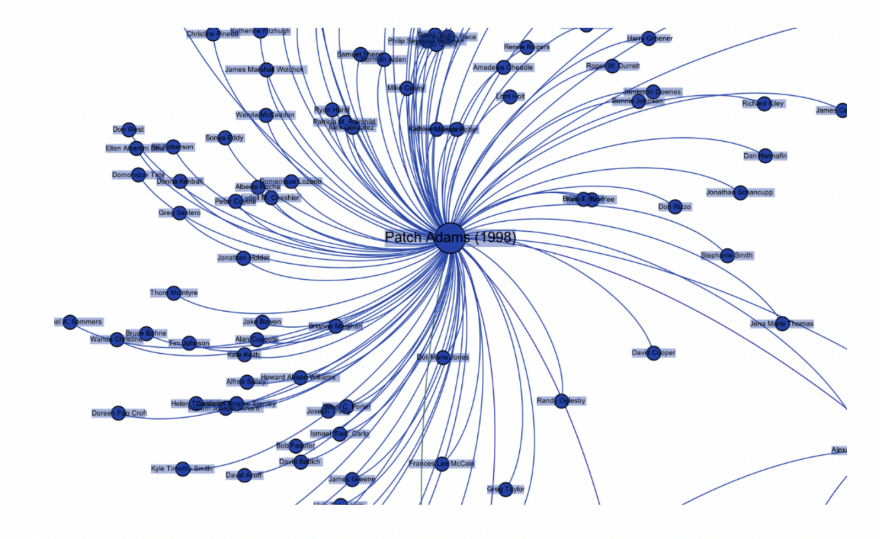
5. What do you see?

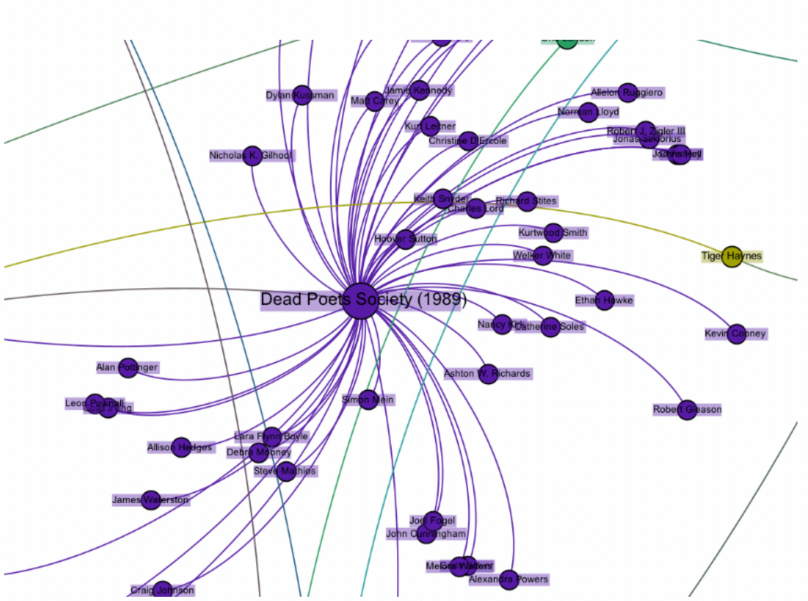
Example: Robin Williams



http://allthingsgraphed.com/2014/08/15/the-lives-he-touched-network-of-robin-williams/

Example: Robin Williams





Top Connectors

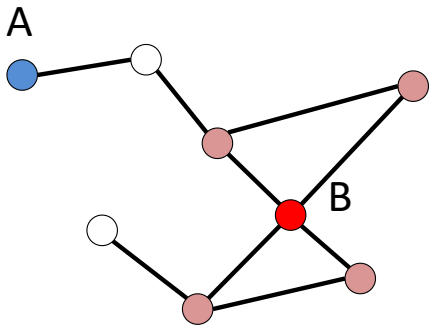
Through his career, Robin Williams has worked with certain actors/actresses on different movies or shows several times. Here are some of the actors and actresses who most frequently worked with him:

Actor/Actress	Movies/TV	
Frank Welker	Jumanji (voices), In Search of Dr. Seuss (voices and extra), A Wish for Things That Work (voice of Santa Claus), Pac Preview Party (himself), Aladdin (voice of Abu), Aladdin on Ice (voice of Abu)	
Adam Bryant	Death to Smoochy (extra), Mrs. Doubtfire (man in restroom), Bicentennial Man (humanoid head), The Fisher King (radio engineer), Awakenings (librarian), Being Human (neighbor)	
Billy Crystal	Freedom: A History of Us (union soldier), Hamlet (First Gravedigger), Father's Day (Jack), In Search of Dr. Seuss (voice), Deconstructing Harry (Larry/Devil)	
Pam Dawber	Pac Preview Party (voice of Mindy), Mork and Mindy (Mindy), The Crazy Ones (Lily)	

Degree, Average Degree, and Degree Distribution

NODE DEGREES

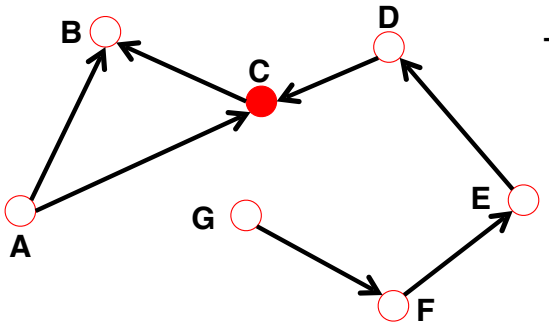
Undirected



Node degree: the number of links connected to the node.

$$k_A = 1 \qquad k_B = 4$$

Directed



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

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

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

Source: a node with $k^{in} = 0$; Sink: a node with $k^{out} = 0$.

A BIT OF STATISTICS

BRIEF STATISTICS REVIEW

Four key quantities characterize a sample of N values $x_1, \ldots, x_N$:

Average (mean):

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The n^{th} moment:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

Standard deviation:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

Distribution of x:

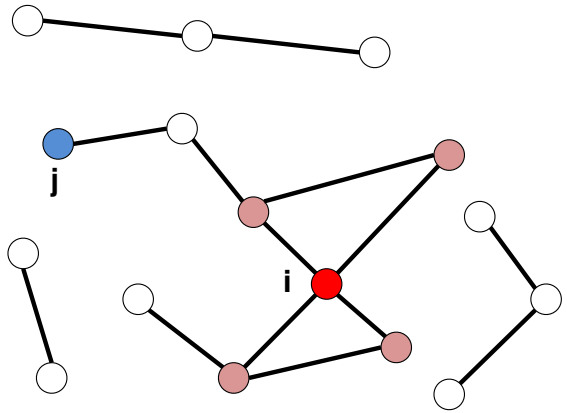
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

where p_x follows

$$\sum_i p_x = 1 \left(\int p_x\, dx = 1\right)$$

AVERAGE DEGREE

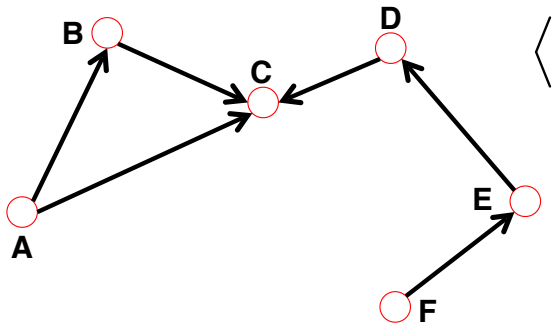
Undirected



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – the number of nodes in the graph

Directed



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

$$\langle k \rangle \equiv \frac{L}{N}$$

Average Degree

Network	Nodes	Links	Directed / Undirected	N	L	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

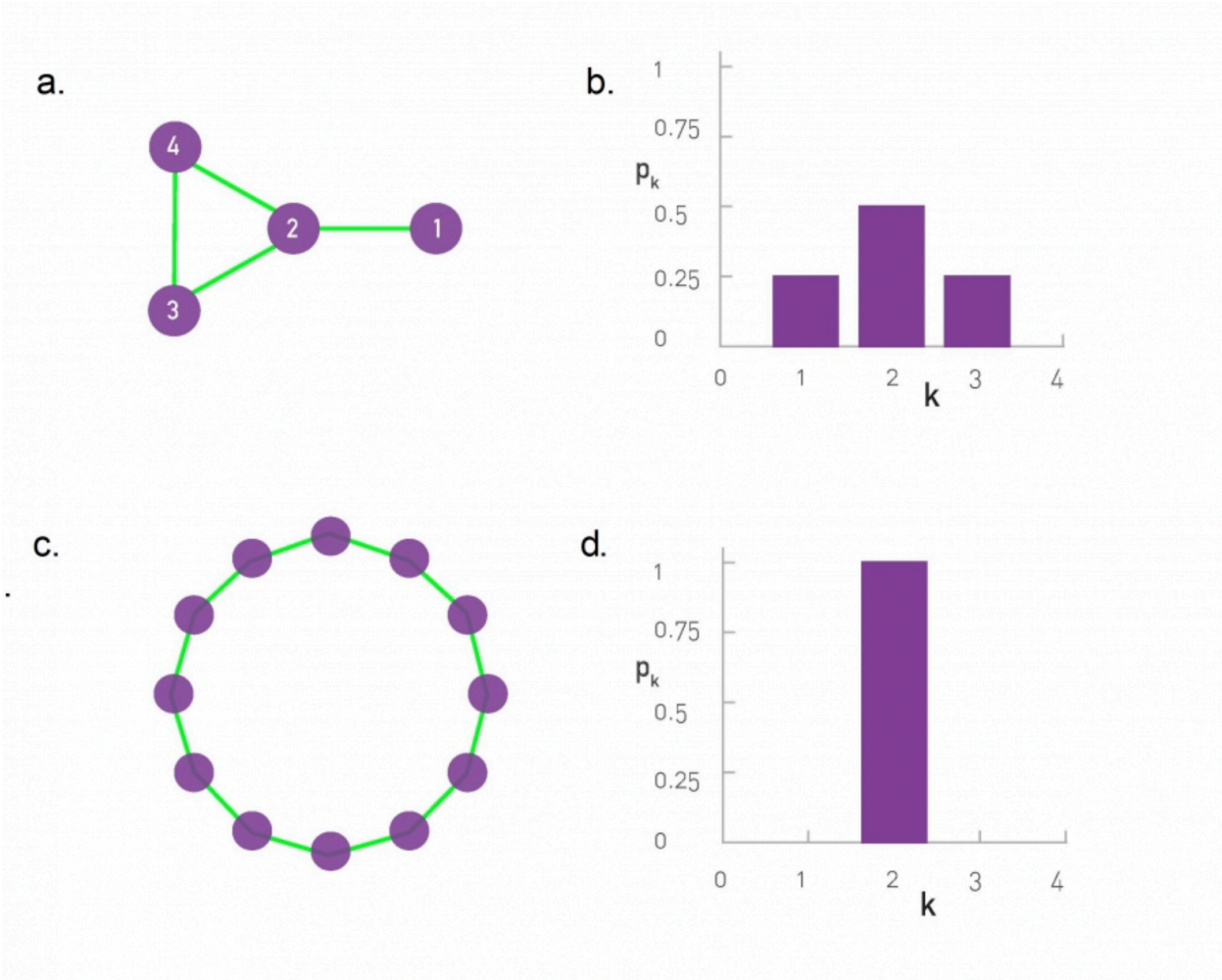
DEGREE DISTRIBUTION

Degree distribution

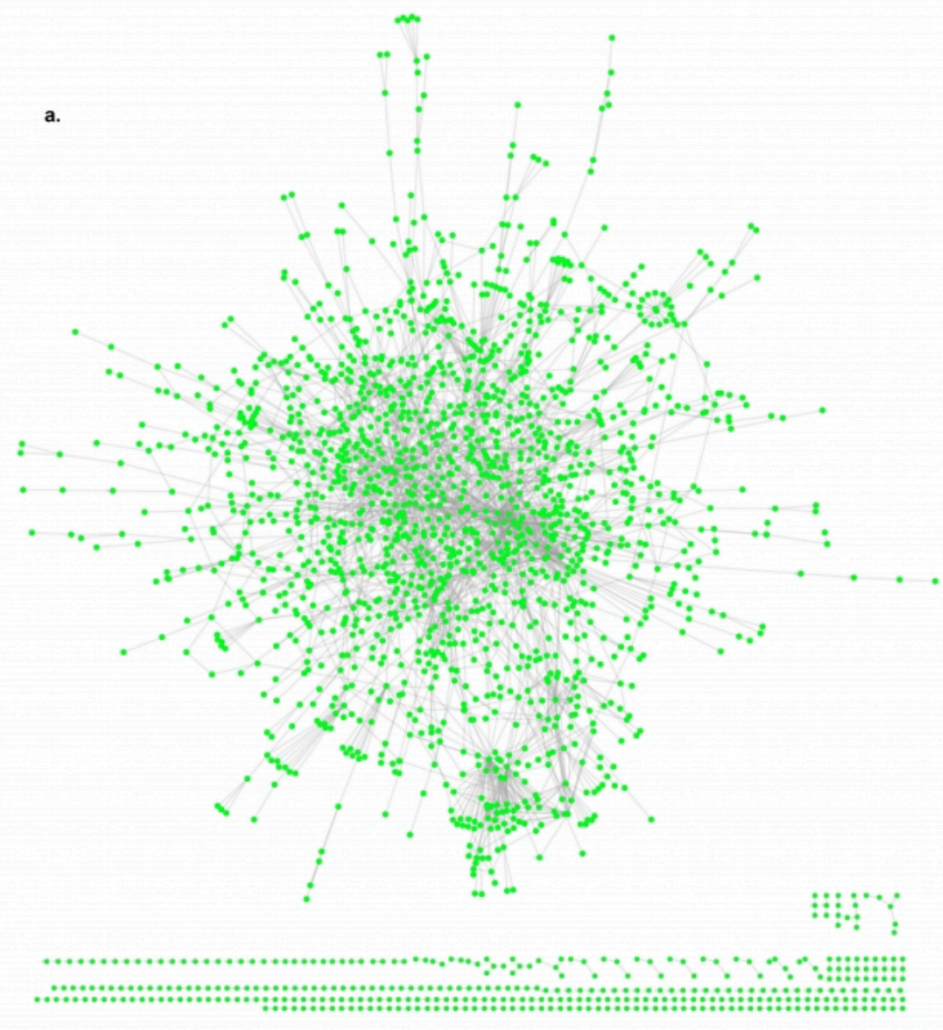
P(k): probability that a randomly chosen node has degree *k*

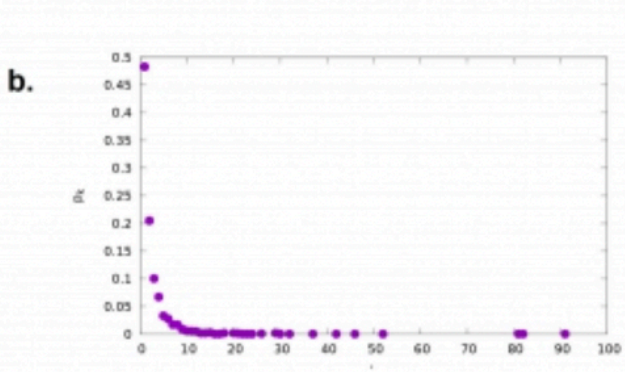
N_k = # nodes with degree k

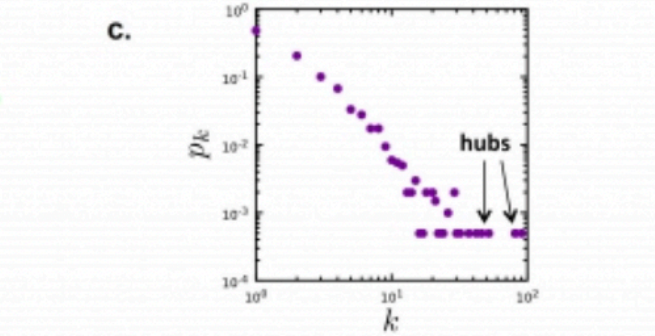
$P(k) = N_k / N$ → plot



DEGREE DISTRIBUTION

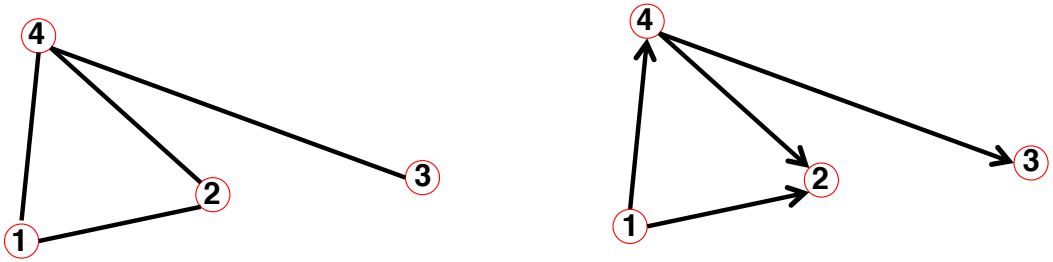






Adjacency matrix

ADJACENCY MATRIX



A_{ij}=1 if there is a link between node *i* and *j*

A_{ij}=0 if nodes *i* and *j* are not connected to each other.

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

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

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

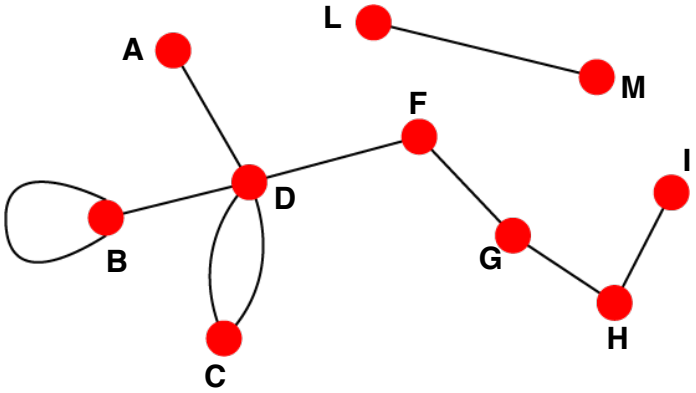
$A_{ij} = 0$ if there is no link pointing from *j* to *i*.

UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:

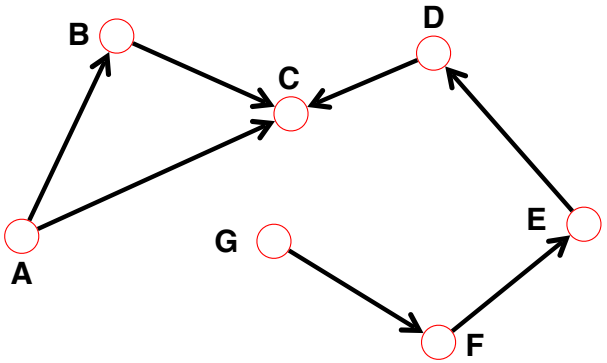


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:

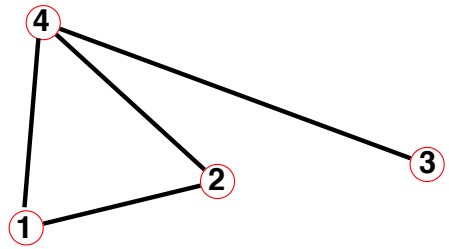


An undirected link is the superposition of two opposite directed links.

Directed links :
URLs on the www
phone calls
metabolic reactions

ADJACENCY MATRIX AND NODE DEGREES

Undirected



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

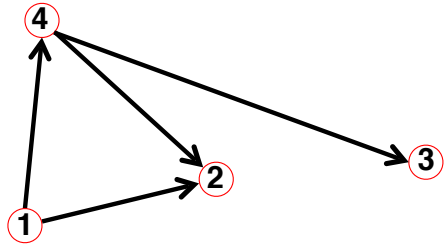
$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

$$k_i = \sum_{j=1}^{N} A_{ij}$$

$$k_j = \sum_{i=1}^{N} A_{ij}$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij} A_{ij}$$

Directed



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

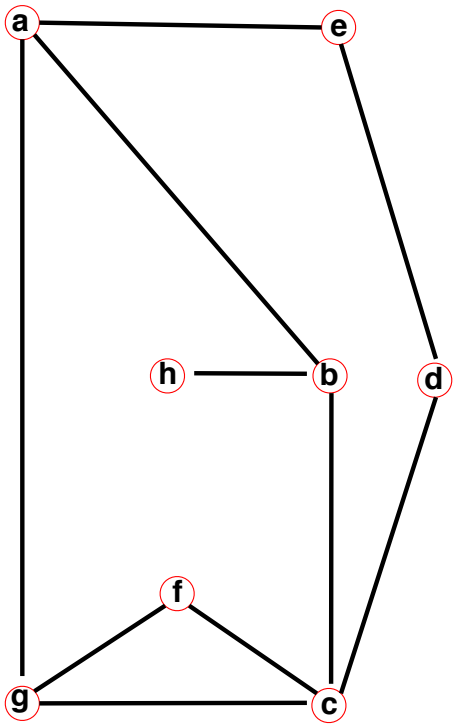
$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j} A_{ij}$$

ADJACENCY MATRIX



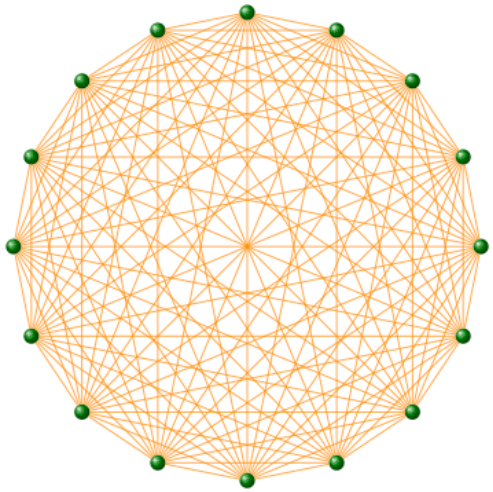
	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	0	0	0
h	0	1	0	0	0	0	0	0

Real networks are sparse

COMPLETE GRAPH

The maximum number of links a network of N nodes can have is:

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



A graph with degree $L=L_{max}$ is called a complete graph, and its average degree is **<k>=N-1**

REAL NETWORKS ARE SPARSE

Most networks observed in real systems are sparse:

$$L \ll L_{max}$$

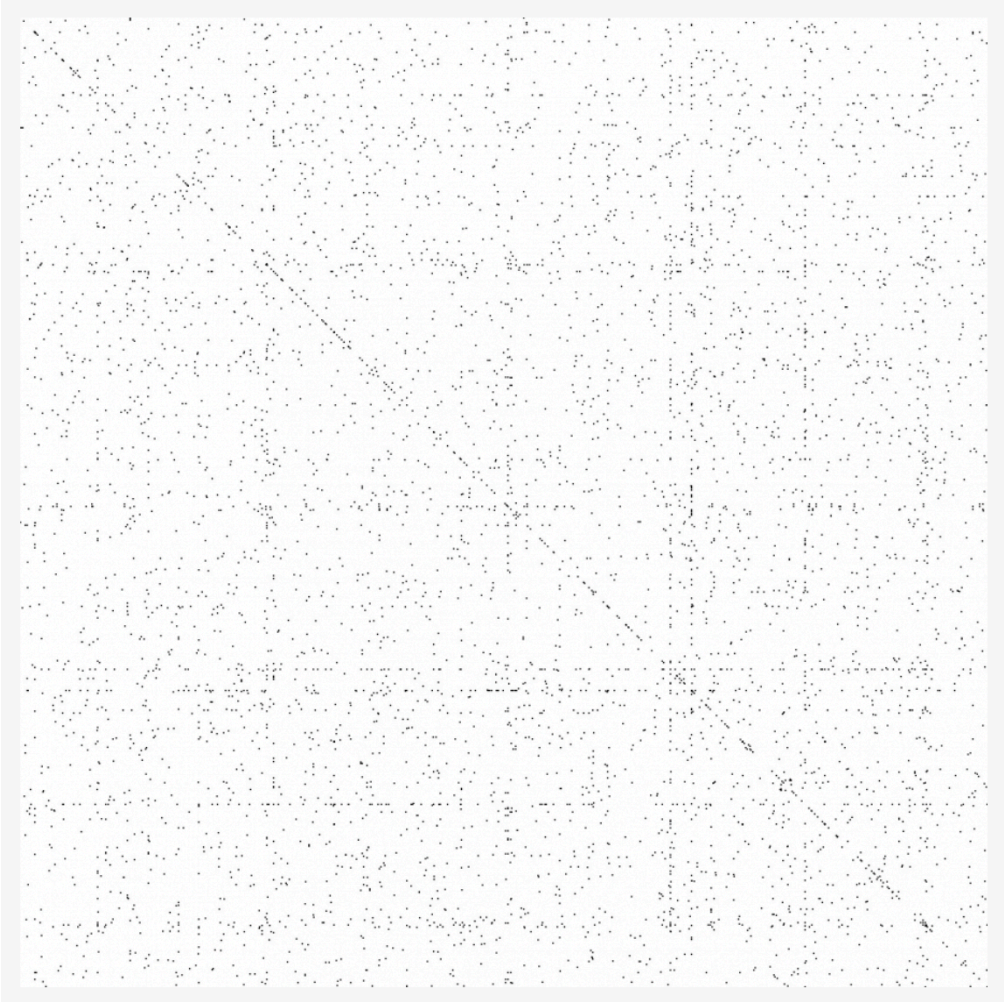
or

$$\langle k \rangle \ll N-1.$$

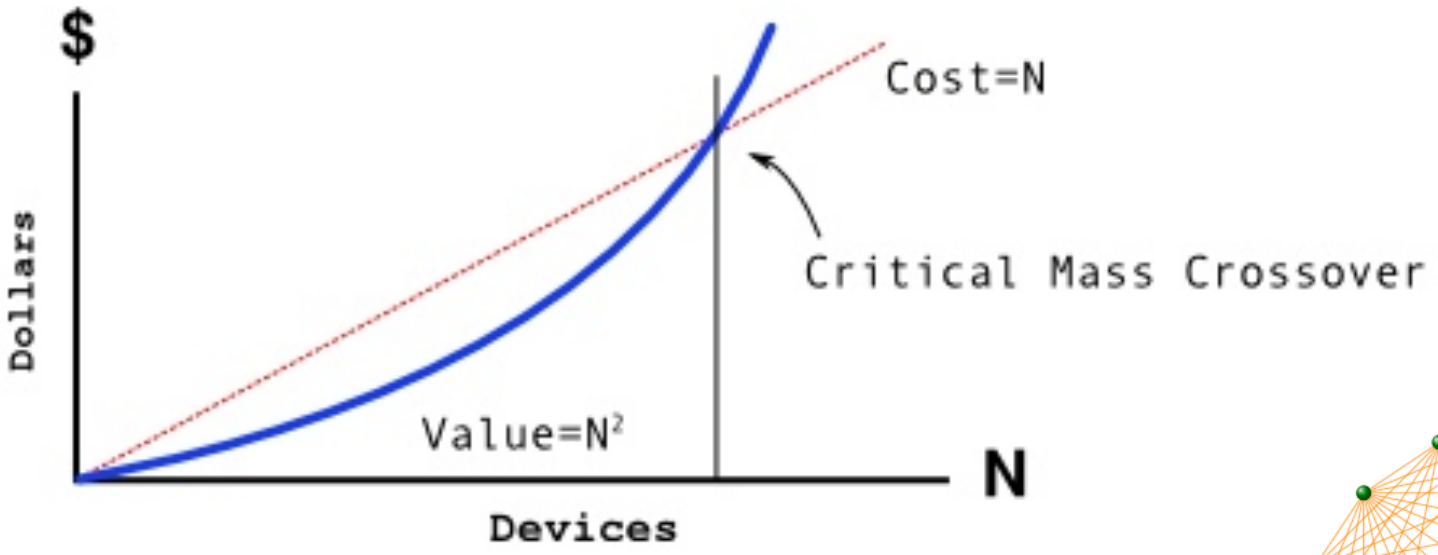
WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	$\langle k \rangle=4.51$
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^7$	$\langle k \rangle=2.39$
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	$\langle k \rangle=3.9$
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	$\langle k \rangle=28.78$

(Source: Albert, Barabasi, RMP2002)

ADJACENCY MATRICES ARE SPARSE

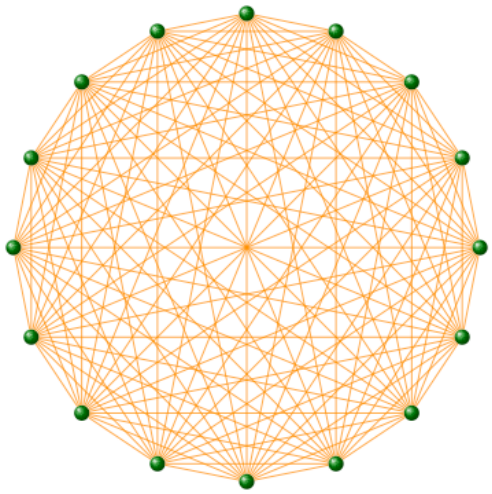


METCALFE'S LAW



The maximum number of links a network of N nodes can have is:

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



Activity: What does the degree distribution of your social network look like?

Is your social network sparse?

Section 2.6

WEIGHTED AND UNWEIGHTED NETWORKS

EXAMPLE: SCIENCE COLLABORATION NETWORK

- Nodes: scientists
- Links: joint publications
- Weights: number of joint pubs.

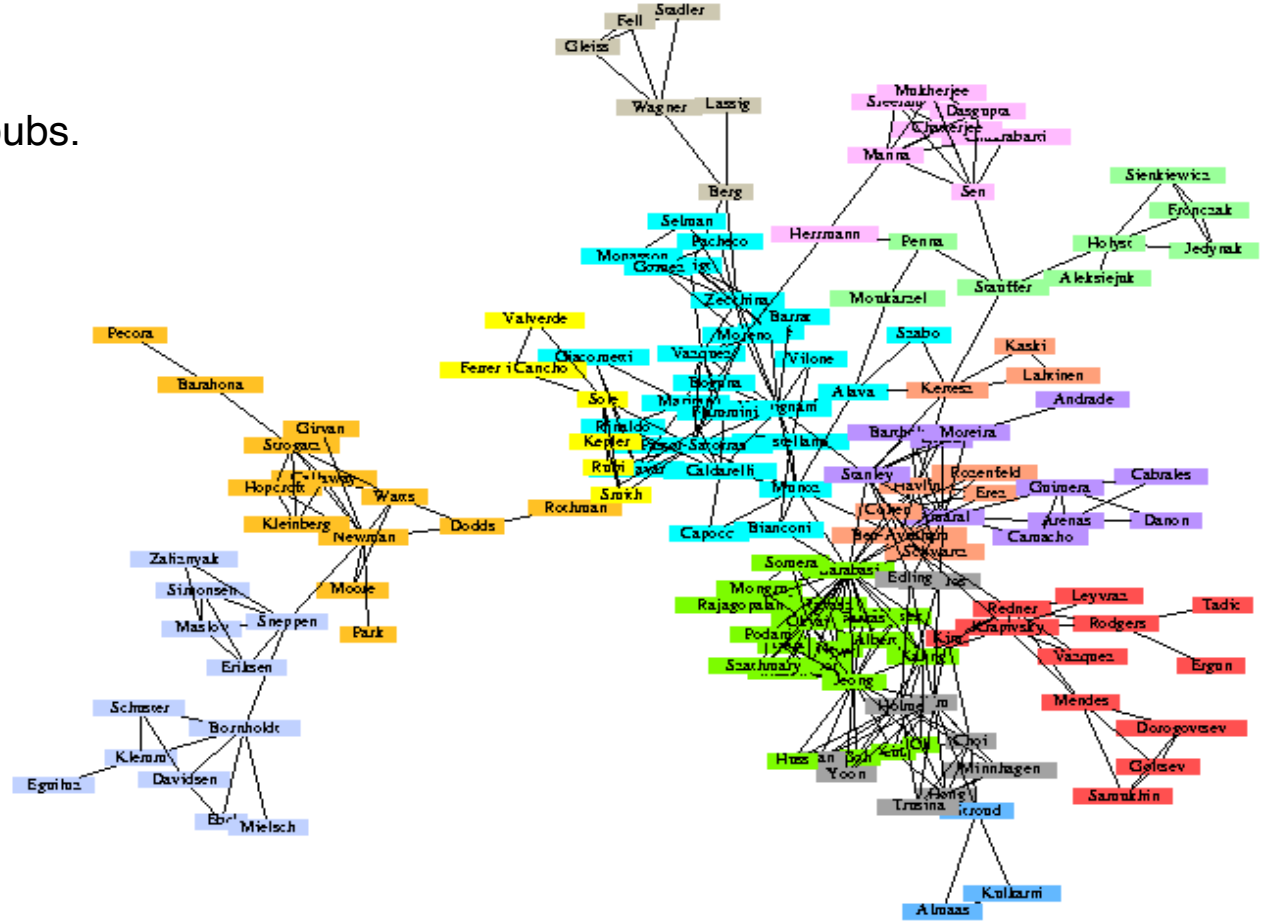
$$w_{ij} = \Sigma_k \frac{\delta_i^k \delta_j^k}{n_k - 1}$$

i, j: authors

k: paper

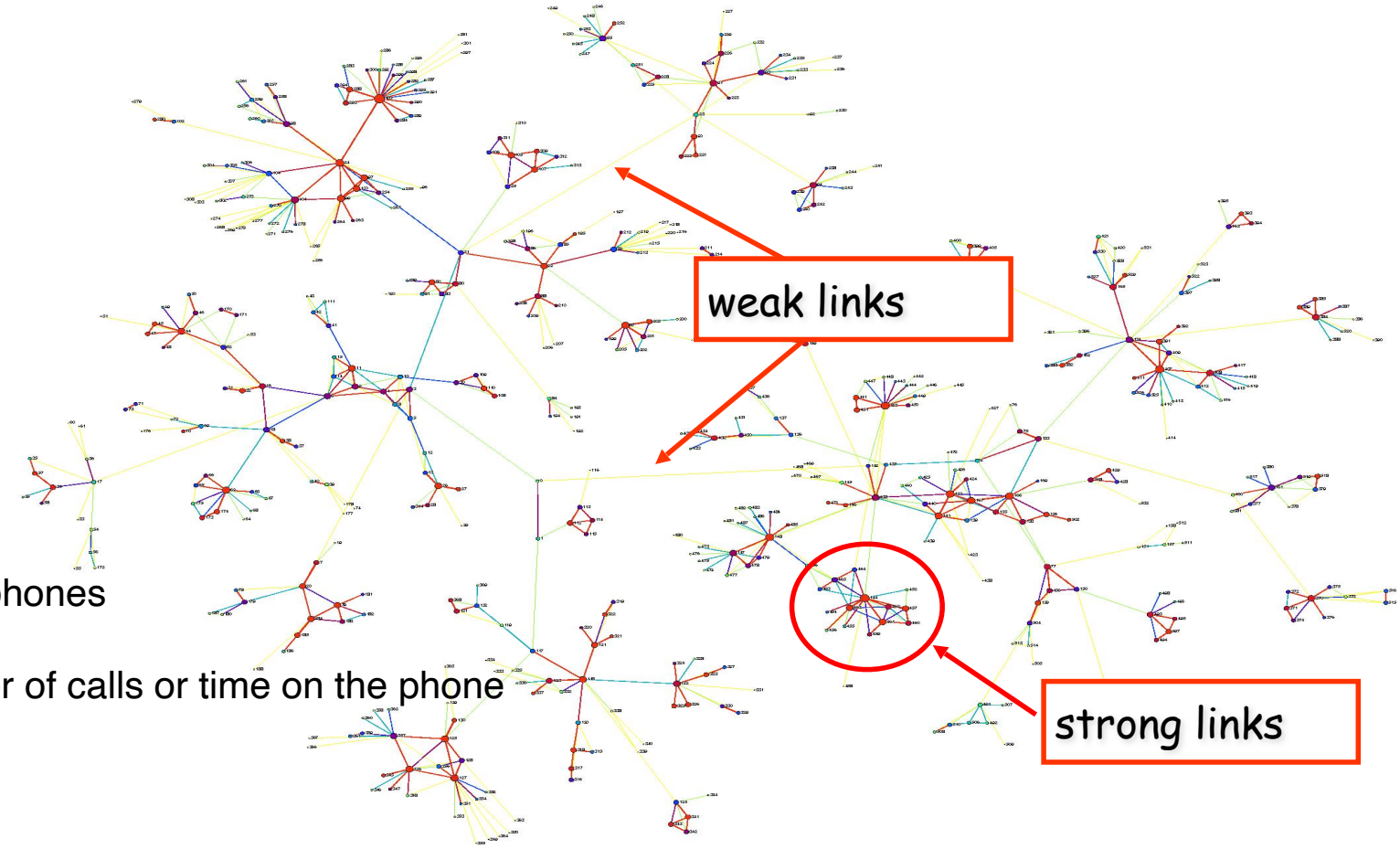
n_k: number of authors

δ_i^k=1 if author i contributed to paper k



Newman and Girvan, cond-mat/0308217 (2003)

EXAMPLE: MOBILE CALL GRAPH

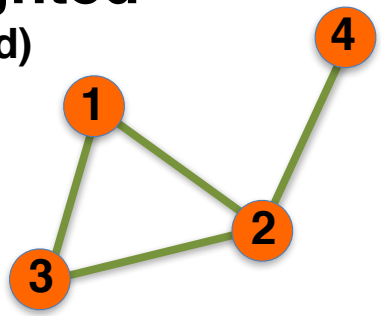


- Nodes: mobile phones
- Links: calls
- Weights: number of calls or time on the phone

Onella et al, PNAS 2007

GRAPHOLOGY 2

Unweighted
(undirected)



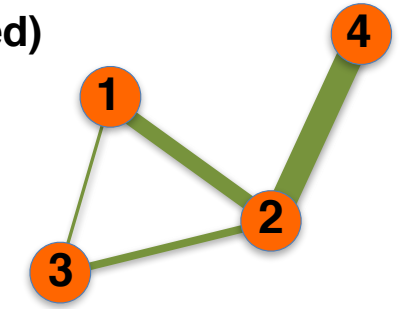
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

$$L = \frac{1}{2}\sum_{i,\,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

Protein-protein interactions, www

Weighted
(undirected)



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

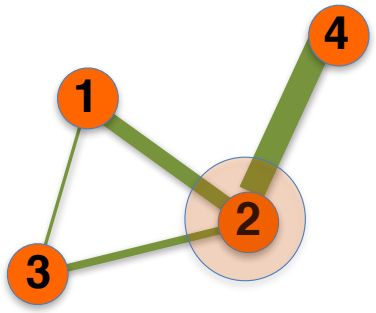
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

Call Graph, metabolic networks

a_{ij} and w_{ij}

In the literature we often use a double notation: A_{ij} and w_{ij} (somewhat redundant)



Adjacency Matrix (A_{ij})

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

Weight Matrix (W_{ij})

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

Node Strength (*weighted degree*) s:

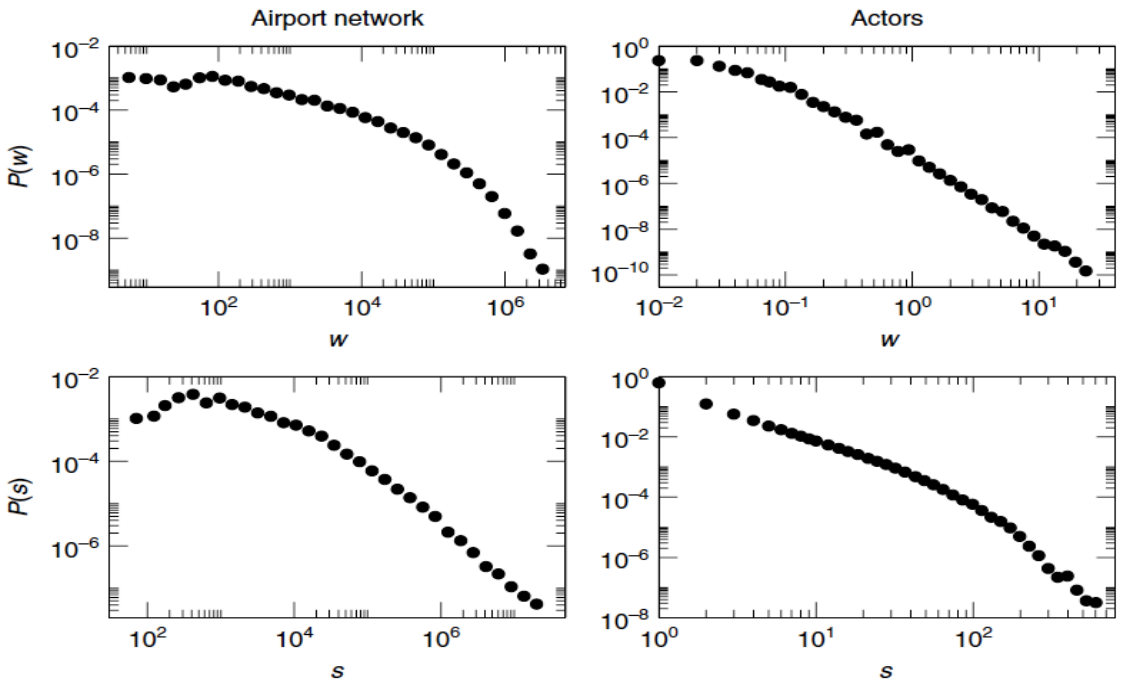
$$s_i = \sum_{j=1}^{N} a_{ij} w_{ij} = \sum_{j=1}^{N} w_{ij}$$

$$s_2 = \sum_{j=1}^{N} w_{2j} = w_{21} + w_{23} + w_{24}$$

EMPIRICAL FINDING 1: P(s) AND P(w) ARE FAT TAILED

Strength distribution *P(s):*
probability that a randomly chosen node has strength *s*

Weight distribution ***P(w)*:**
probability that a randomly chosen link has weight ***w***

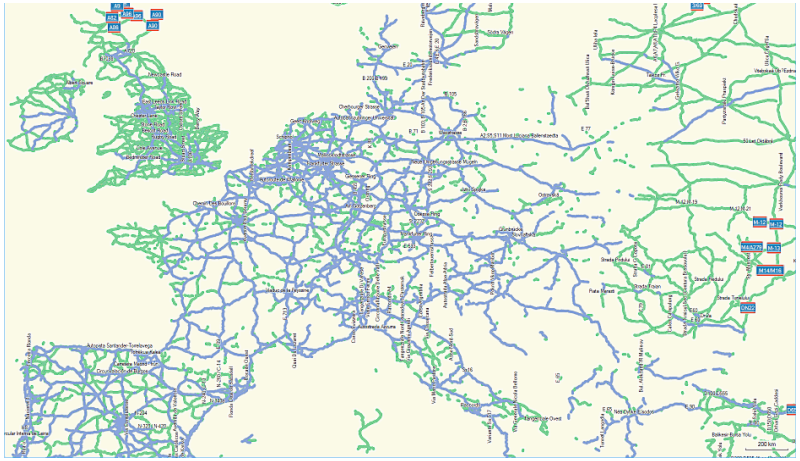


In most real systems P(s) and P(w) are fat tailed.

THINK CAREFULLY: WHAT DO YOUR WEIGHTS MEAN?

The way to appropriately handle and interpret the extra information from edge weights is sensitive to what the weights are describing.

Eg path lengths (and any metric exploiting this concept!)



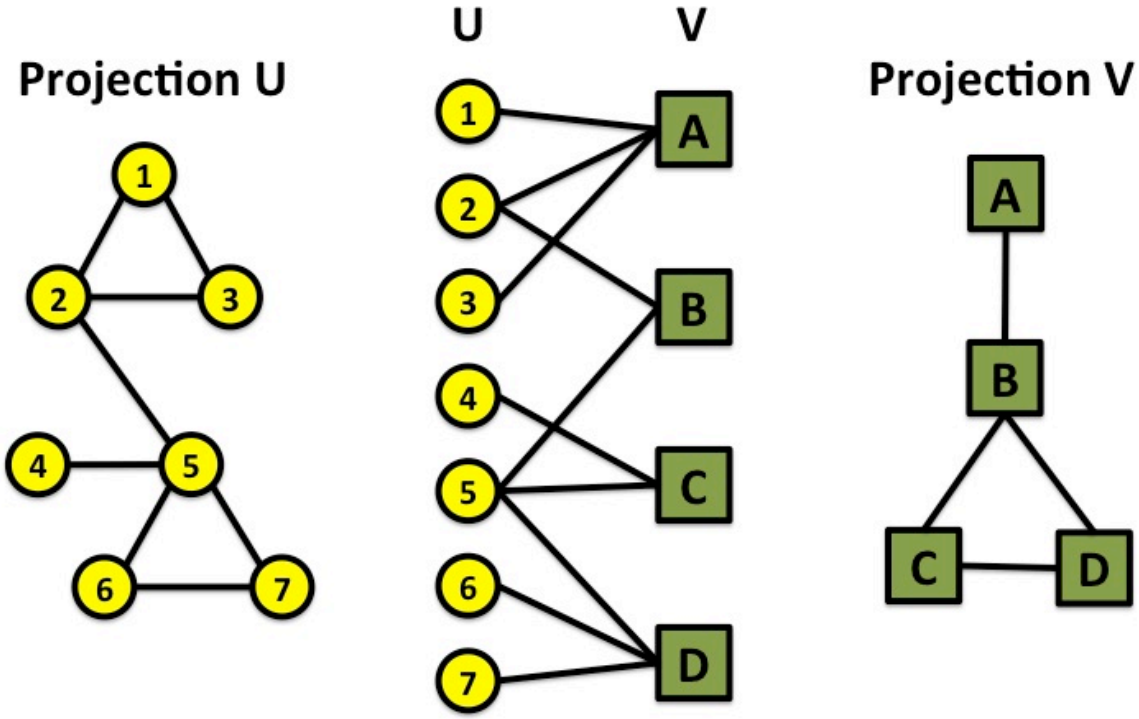
Higher weights can represent higher distances (or other ‘costs’). Eg road networks, spatial networks…
-> Want high strengths to correspond to long path lengths.

Higher weights can represent higher strength of connections/interactions. Eg fMRI, metabolic...
-> Want high strengths to correspond to short paths.

BIPARTITE NETWORKS

BIPARTITE GRAPHS

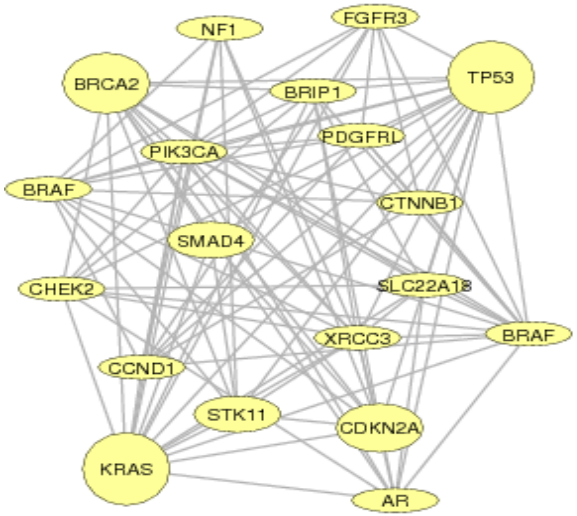
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



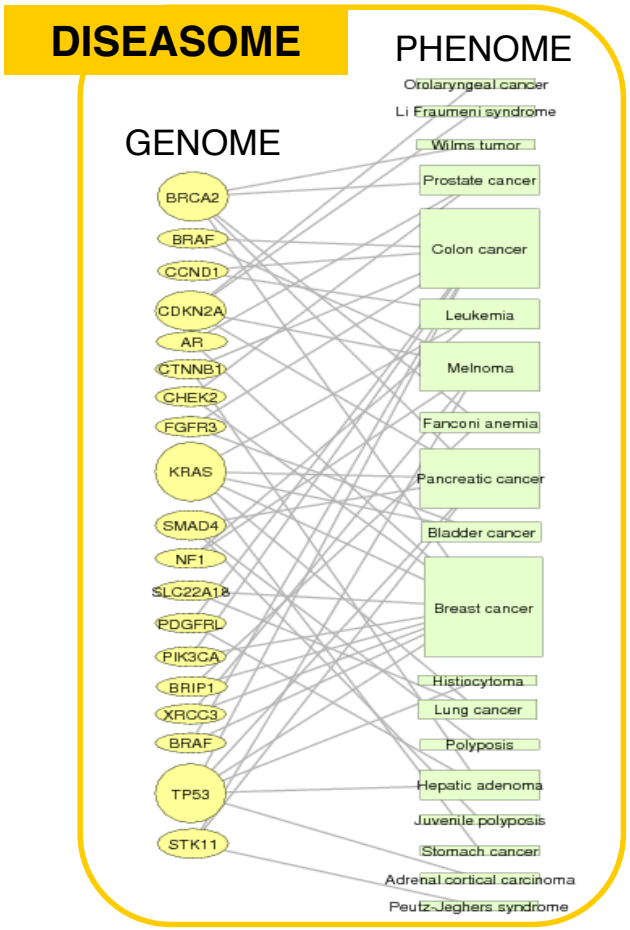
Examples:

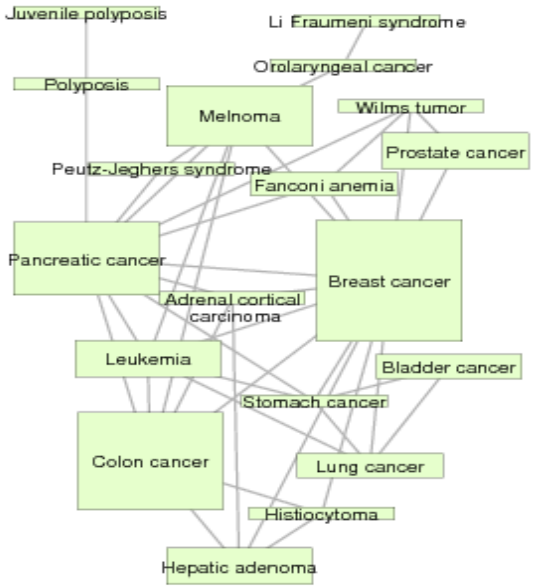
Hollywood actor network
Collaboration networks
Disease network (diseasome)

GENE NETWORK – DISEASE NETWORK



Gene network

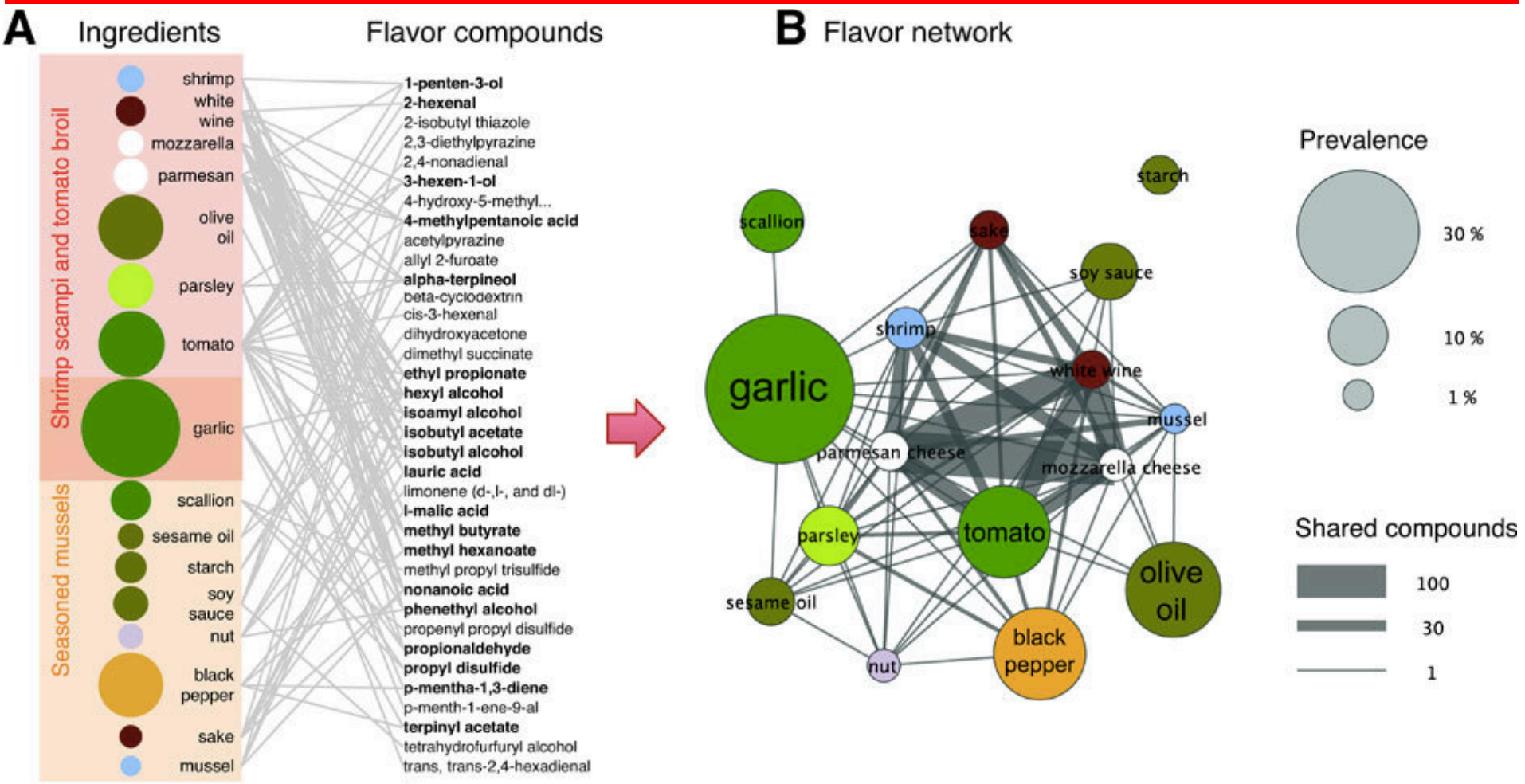




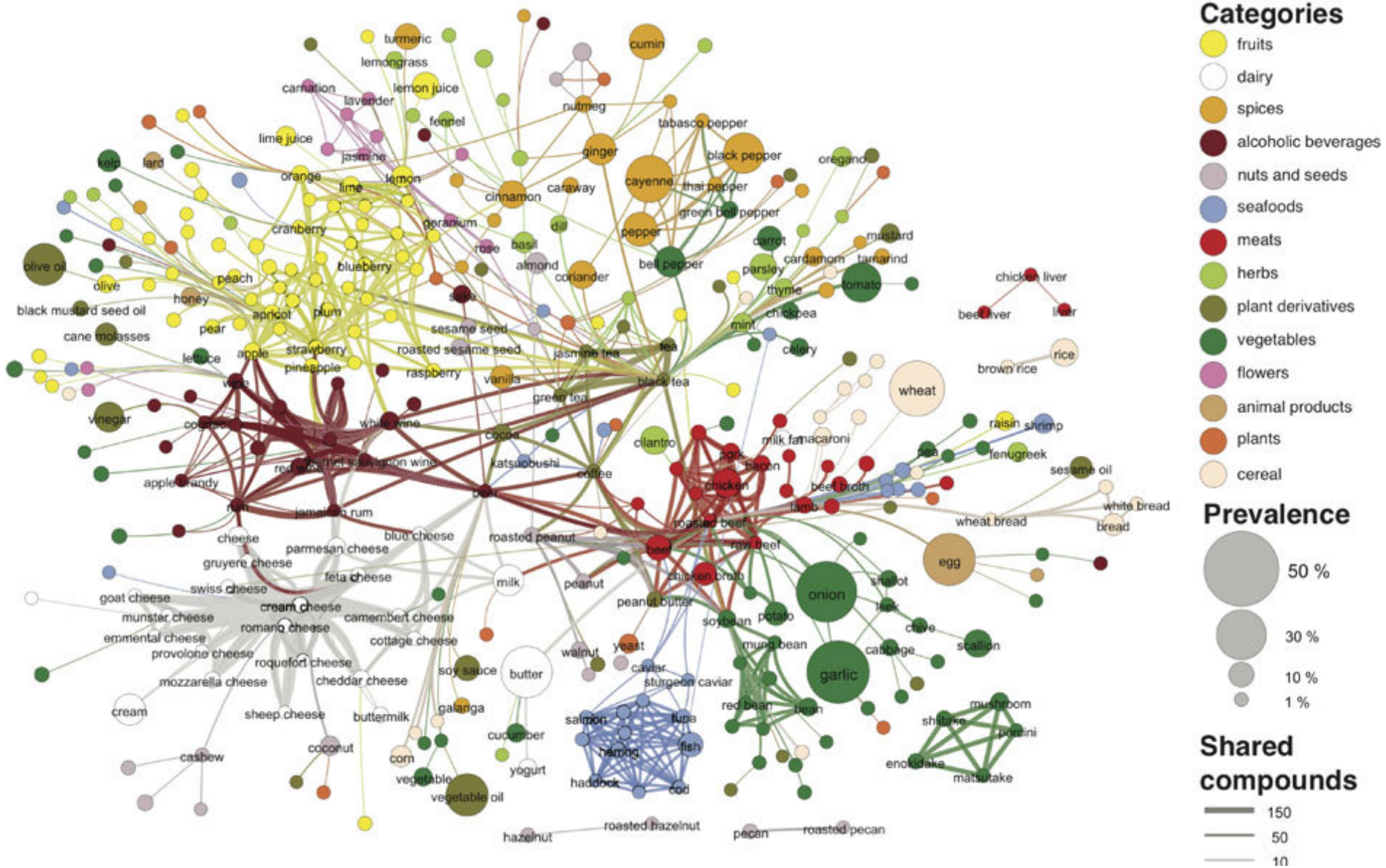
Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Ingredient-Flavor Bipartite Network



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



Categories

- fruits
- dairy
- spices
- alcoholic beverages
- nuts and seeds
- seafoods
- meats
- herbs
- plant derivatives
- vegetables
- flowers
- animal products
- plants
- cereal

Prevalence

50 %

30 %

10 %

1 %

Shared compounds

150

50

10

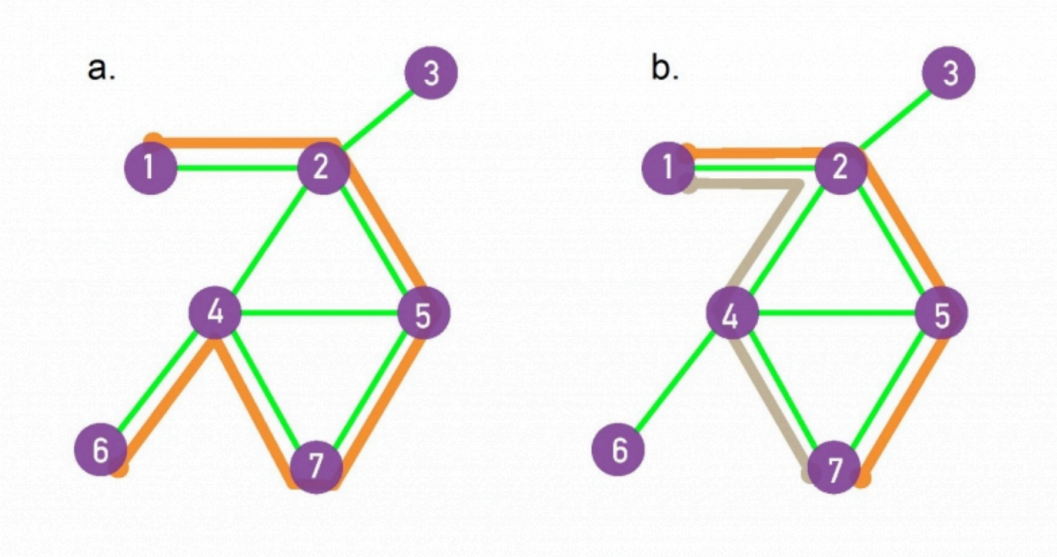
PATHOLOGY

PATHS

A *path is* a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

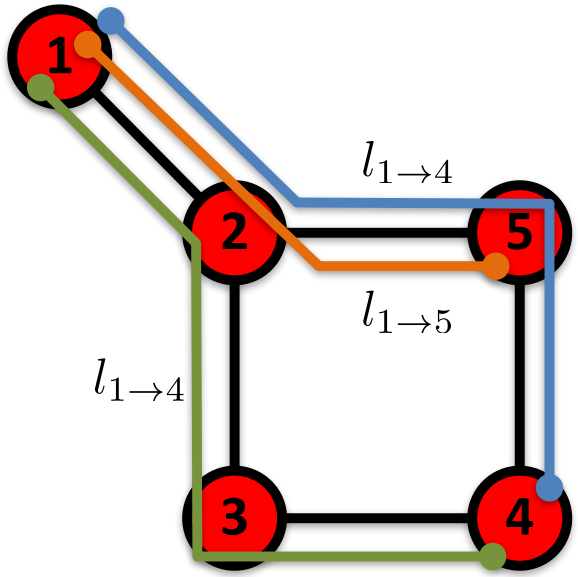
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



- In a directed network, the path can follow only the direction of an arrow.

PATHOLOGY

Shortest Path



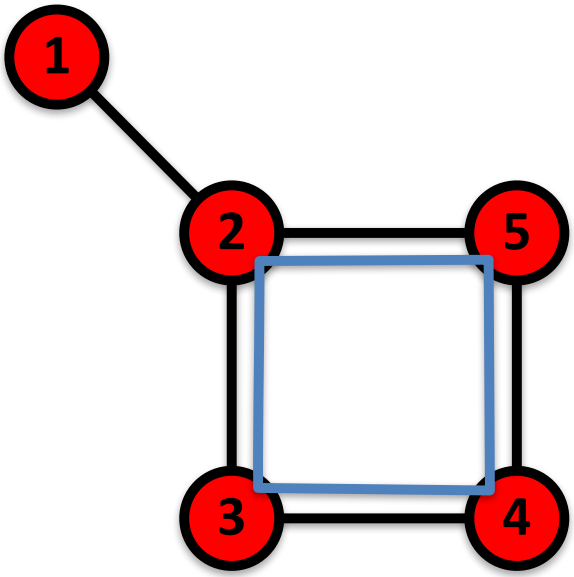
$$l_{1\to4} = 3$$

$$l_{1\to5} = 2$$

The path with the shortest length between two nodes (distance).

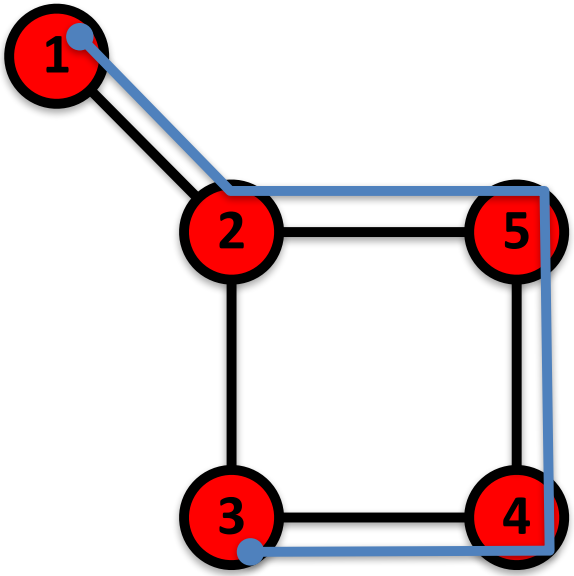
PATHOLOGY

Cycle



A path with the same start and end node.

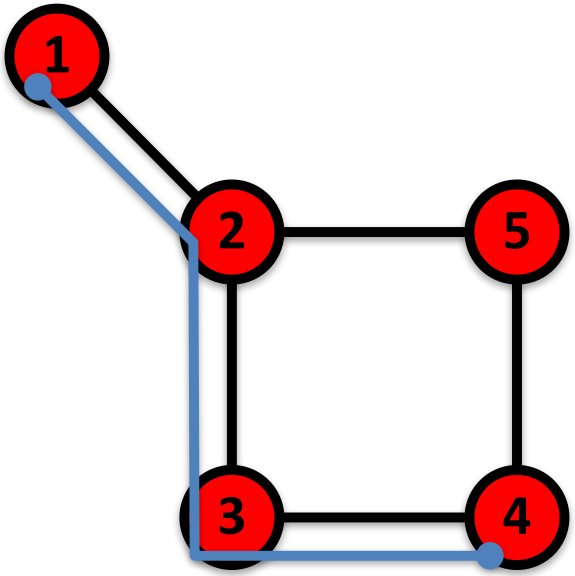
Self-avoiding Path



A path that does not intersect itself.

PATHOLOGY

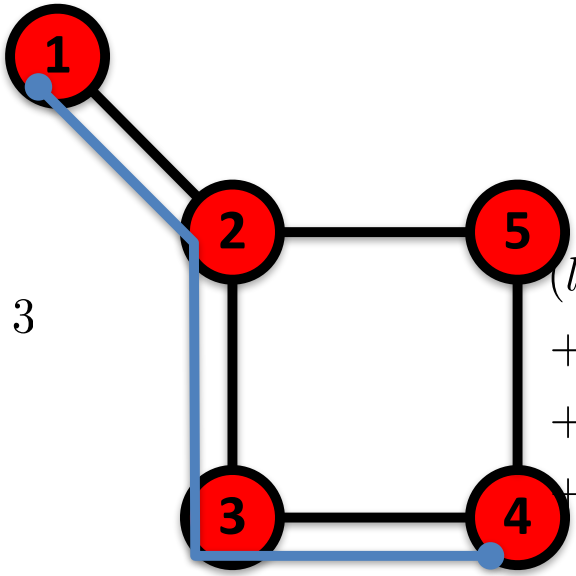
Diameter



$l_{1\to 4} = 3$

The longest shortest path in a graph

Average Path Length



$$(l_{1\to 2} + l_{1\to 3} + l_{1\to 4} + l_{1\to 5} + l_{2\to 3} + l_{2\to 4} + l_{2\to 5} + l_{3\to 4} + l_{3\to 5} + l_{4\to 5})/10 = 1.6$$

The average of the shortest paths for all pairs of nodes.

NETWORK DIAMETER AND AVERAGE DISTANCE

Diameter: $\boldsymbol{d_{max}}$ the maximum distance between any pair of nodes in the graph.

Average path length/distance, <d>, for a connected graph:

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,\, j \neq i} d_{ij}$$

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

In an *undirected graph* $d_{ij} = d_{ji}$, *so* we only need to count them once:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,\, j > i} d_{ij}$$

N_{ij}, number of paths between any two nodes *i* and *j*:

***Length n=1*:** If there is a link between *i* and *j*, then $A_{ij}=1$ and $A_{ij}=0$ otherwise.

Length n=2: If there is a path of length two between *i* and *j*, then $A_{ik}A_{kj}=1$, and $A_{ik}A_{kj}=0$ otherwise.
The number of paths of length 2:

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik}A_{kj} = [A^2]_{ij}$$

Length n: In general, if there is a path of length *n* between *i* and *j*, then $A_{ik}\ldots A_{lj}=1$ and $A_{ik}\ldots A_{lj}=0$ otherwise.
The number of paths of length *n* between *i* and *j* is*

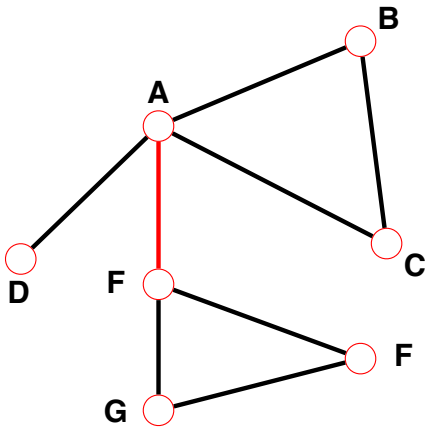
$$N_{ij}^{(n)} = [A^n]_{ij}$$

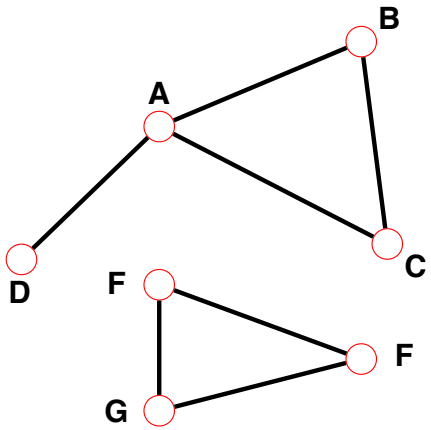
**holds for both directed and undirected networks.*

CONNECTEDNESS

CONNECTIVITY OF UNDIRECTED GRAPHS

Connected (undirected) graph: any two vertices can be joined by a path.
A disconnected graph is made up by two or more connected components.





Largest Component:
Giant Component

The rest: **Isolates**

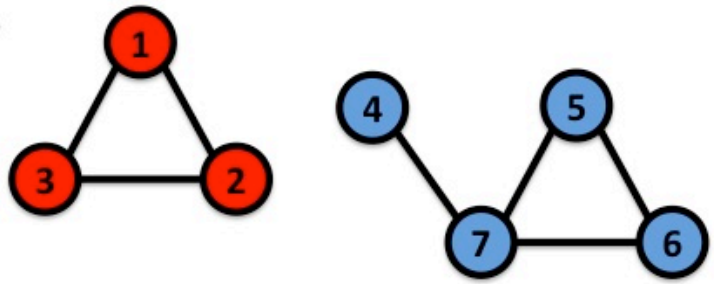
Bridge: if we erase it, the graph becomes disconnected.

CONNECTIVITY OF UNDIRECTED GRAPHS

Adjacency Matrix

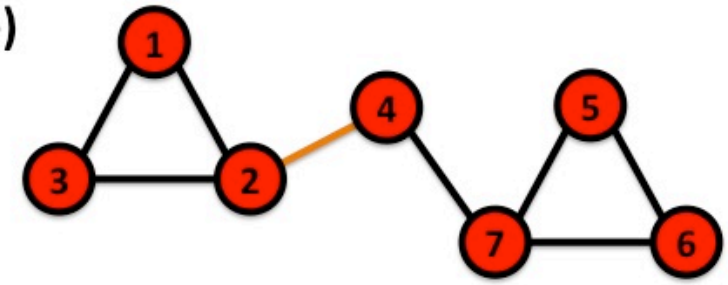
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

(a)



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

(b)



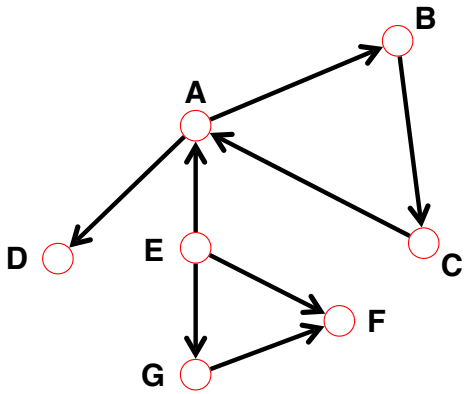
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

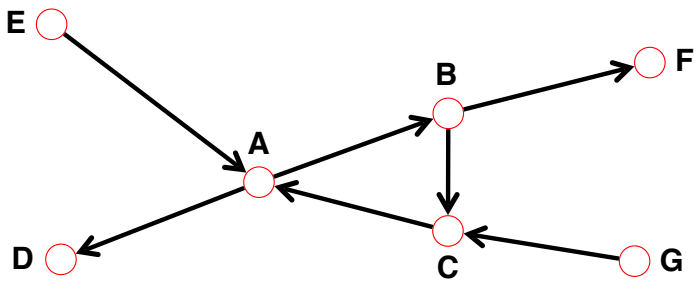
CONNECTIVITY OF DIRECTED GRAPHS

Strongly connected directed graph: has a path from each node to every other node and vice versa (e.g. AB path and BA path).

Weakly connected directed graph: it is connected if we disregard the edge directions.

Strongly connected components can be identified, but not every node is part of a nontrivial strongly connected component.





In-component: nodes that can reach the scc,

Out-component: nodes that can be reached from the scc.

Clustering coefficient

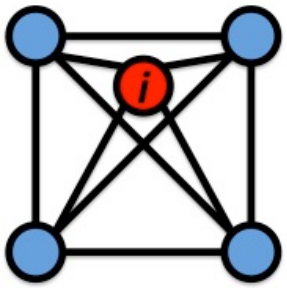
CLUSTERING COEFFICIENT

Clustering coefficient:

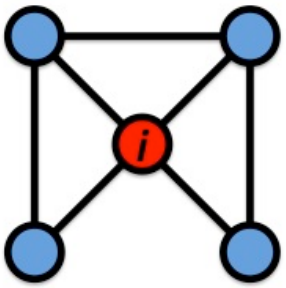
what fraction of your neighbors are connected?

- Node i with degree k_i
- C_i in [0,1]

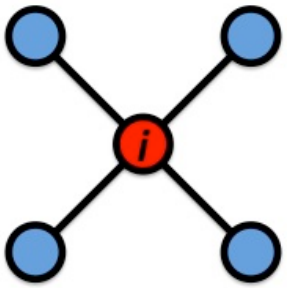
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



$C_i = 1$



$C_i = 1/2$



$C_i = 0$

Watts & Strogatz, Nature 1998.

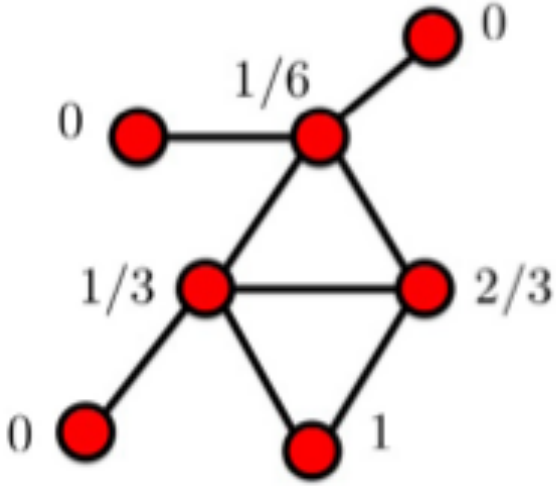
CLUSTERING COEFFICIENT

Clustering coefficient:

what fraction of your neighbors are connected?

- Node i with degree k_i
- C_i in [0,1]

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



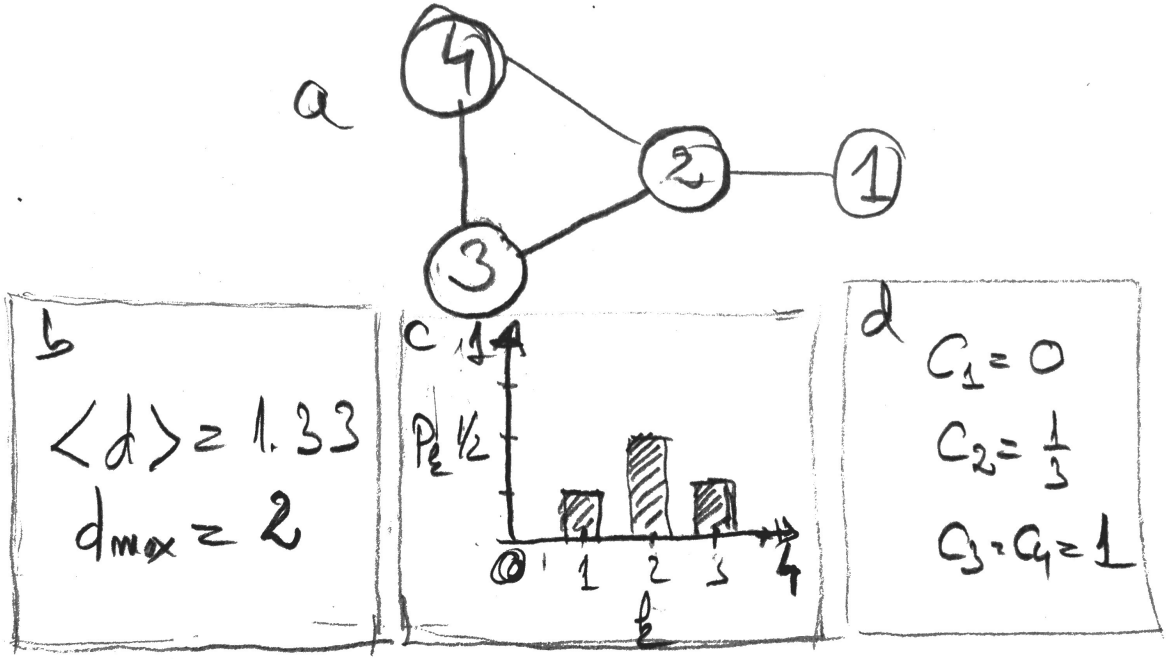
$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C = \frac{3}{8} = 0.375$$

Watts & Strogatz, Nature 1998.

Summary

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE



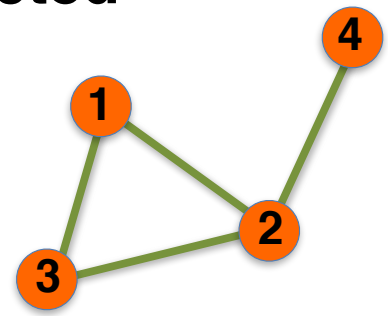
A. Degree distribution: p_k

B. Path length: $\langle d \rangle$

C. Clustering coefficient: $C_i = \frac{2e_i}{k_i(k_i-1)}$

GRAPHOLOGY *1*

Undirected



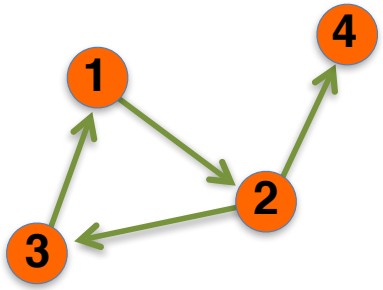
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

Actor network, protein-protein interactions

Directed



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

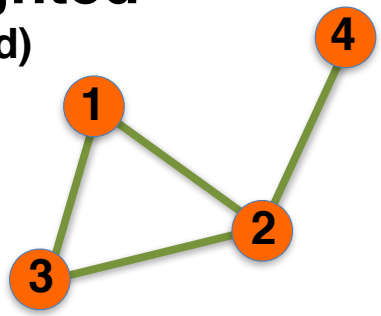
$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{L}{N}$$

WWW, citation networks

GRAPHOLOGY 2

Unweighted
(undirected)



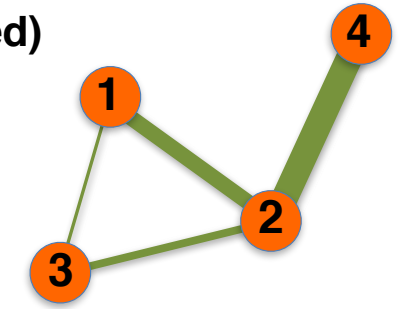
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

$$L = \frac{1}{2}\sum_{i,\,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

protein-protein interactions, www

Weighted
(undirected)



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

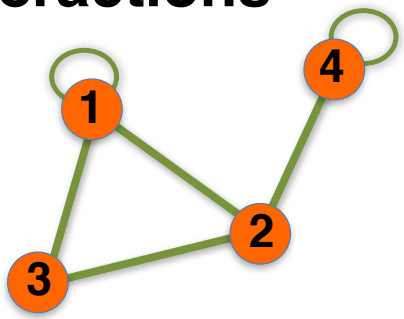
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

Call Graph, metabolic networks

GRAPHOLOGY 3

Self-interactions



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

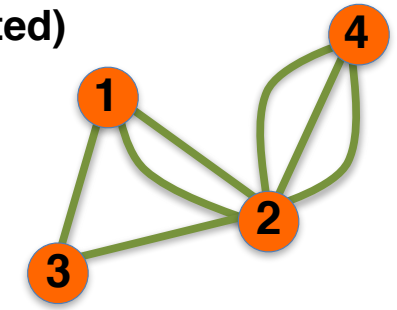
$A_{ii} \neq 0$ $\qquad$ $A_{ij} = A_{ji}$

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

Protein interaction network, www

Multigraph

(undirected)



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

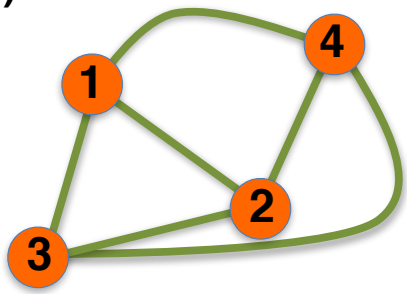
$A_{ii} = 0$ $\qquad$ $A_{ij} = A_{ji}$

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

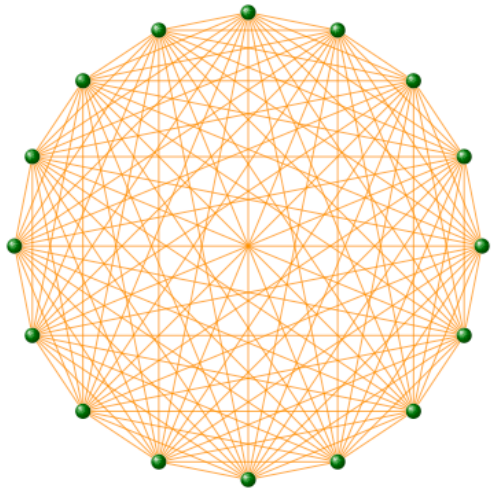
Social networks, collaboration networks

Complete Graph

(undirected)



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



$$A_{ii} = 0 \qquad A_{i \neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \qquad < k >= N - 1$$

Actor network, protein-protein interactions

GRAPHOLOGY: Real networks can have multiple characteristics

WWW > **d**irected multigraph with self-interactions

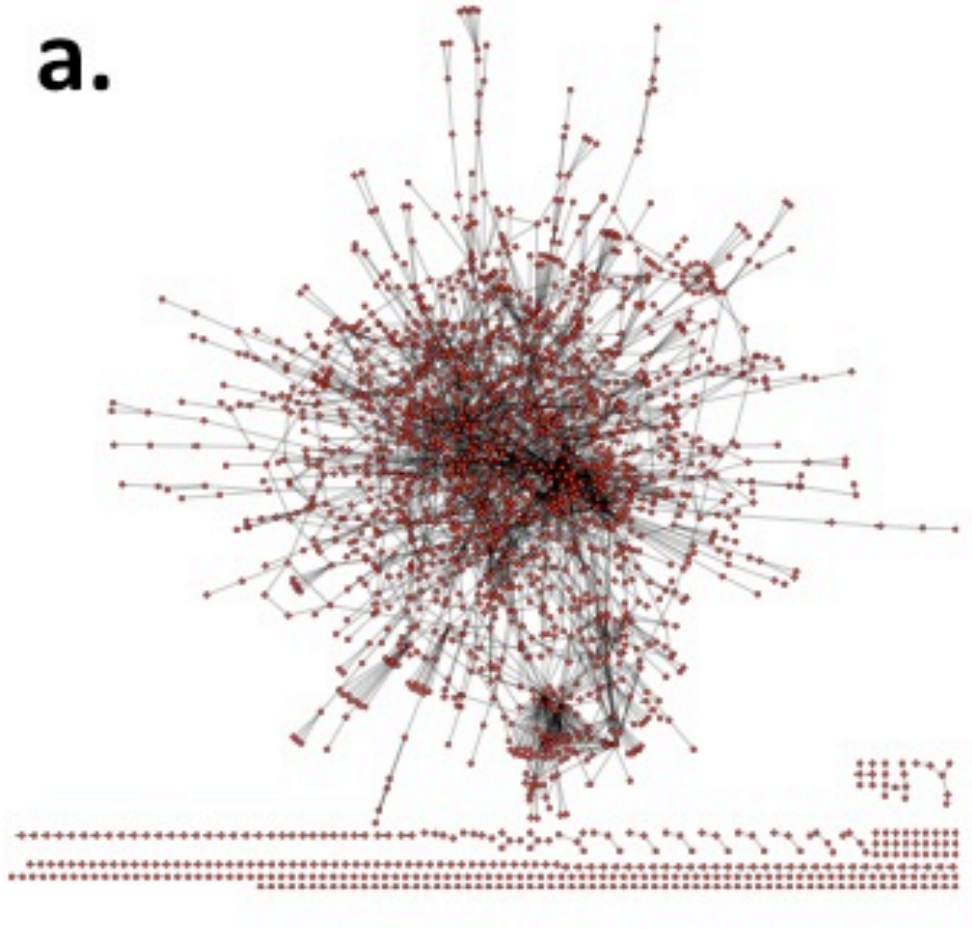
Protein Interactions > **u**ndirected unweighted with self-interactions

Collaboration network > **u**ndirected multigraph or weighted.

Mobile phone calls > directed, weighted.

Facebook Friendship links > **u**ndirected, unweighted.

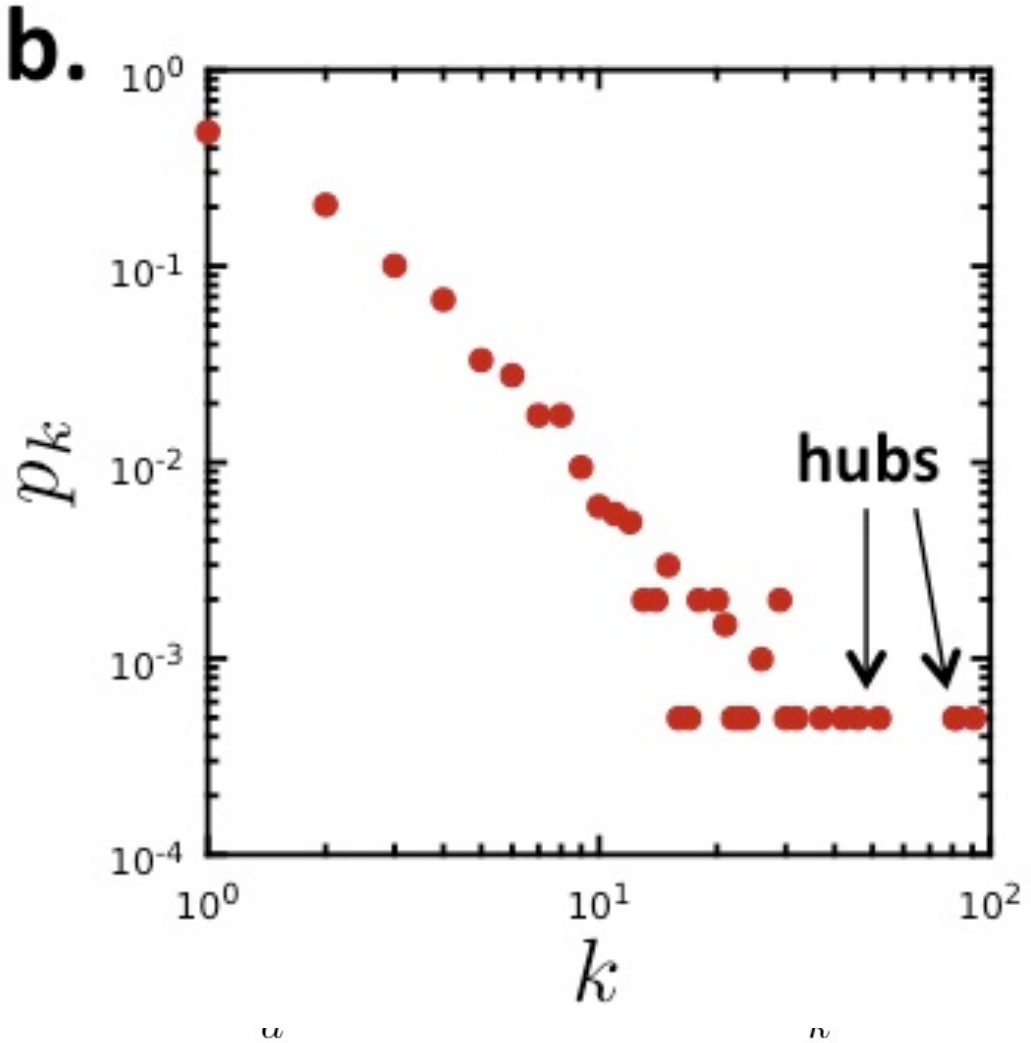
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Undirected network
N=2,018 proteins as nodes
L=2,930 binding interactions as links.
Average degree <k>=2.90.

Not connected: 185 components
the largest (giant component) 1,647 nodes

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

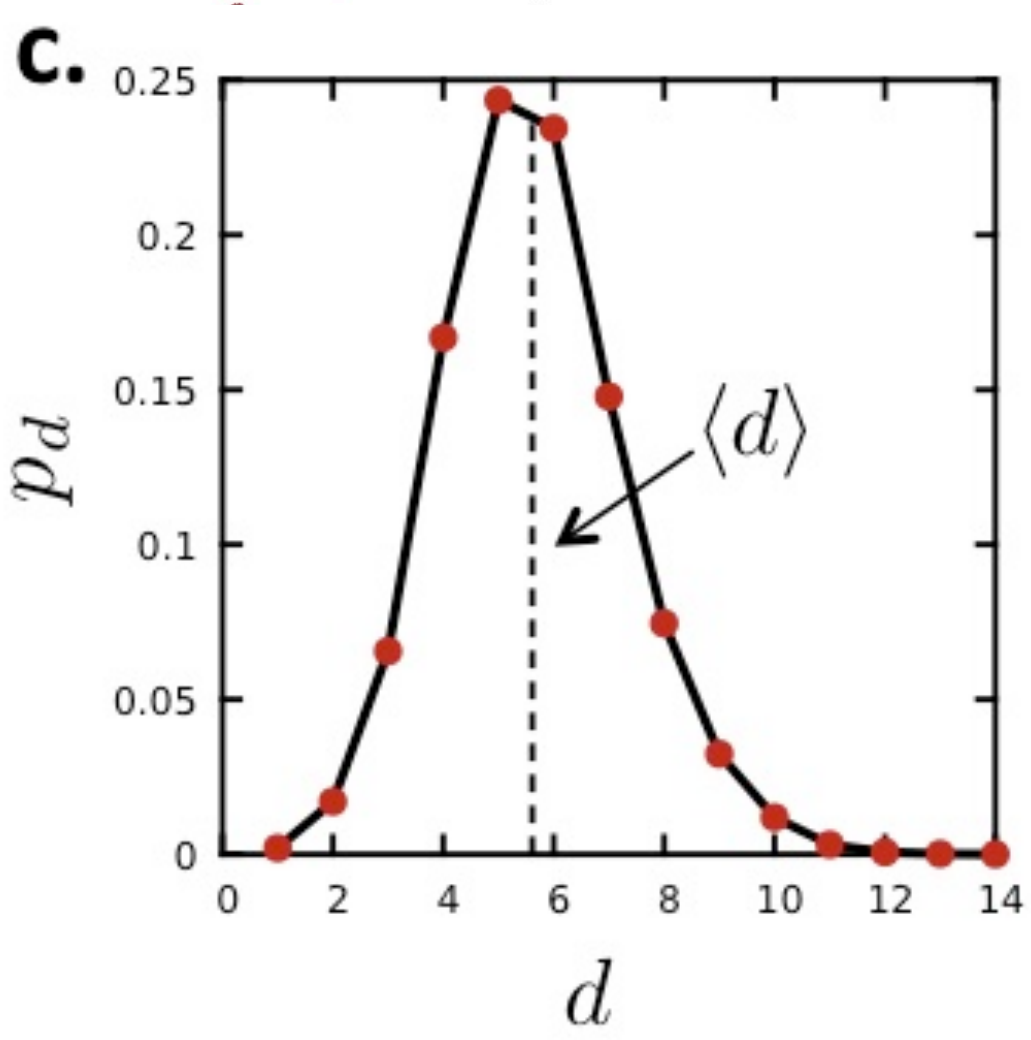


p_k is the probability that a node has degree **k**.

N_k = # nodes with degree k

$p_k = N_k / N$

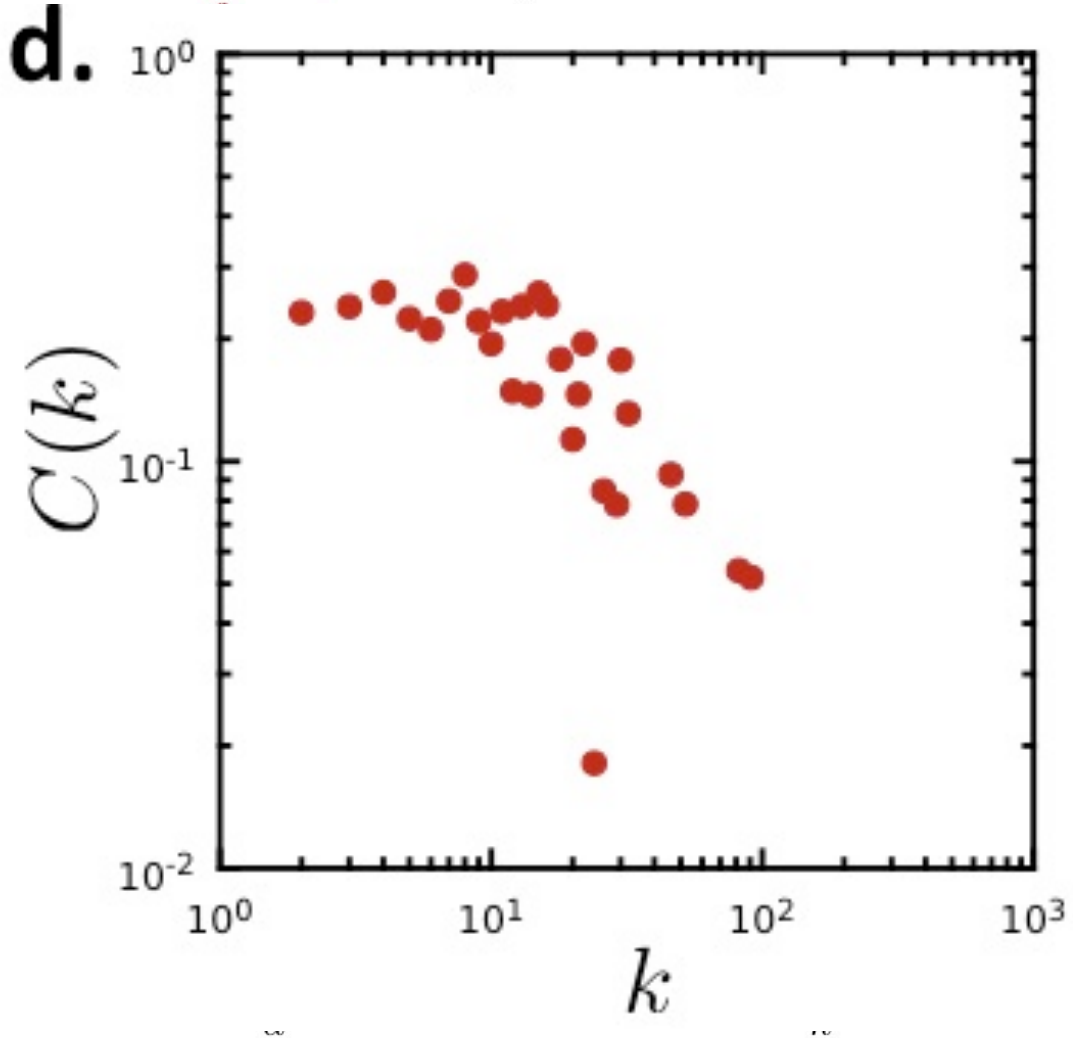
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

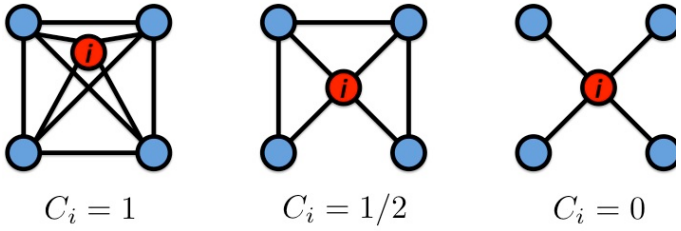


d_{max}=14

<d>=5.61

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK





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

<C>=0.12

Questions?

Reminder...

Assessment: Project proposal Sept. 27th

Presentation

2 minutes (time limit will be enforced)
4 slides

Discuss:

- What are your nodes and links?
- Describe your dataset. How will you get it?
- Expected size of the network (number of nodes, number of links)
- What questions do you plan to ask? These may change during the course of the class.
- Why do we care about the network you plan to study?

Written component

1 page
Written summary of the details in your presentation.

Project Matchmaking: Week 1

CPSC 572/672 Project Matchmaking

File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago

Share

D8

	A	B	C	D	E	F	G	H	I
1	**Name**	**572/672**	**Skills I have**	**Skills I want to develop**	**Datasets or questions I am interested in**	**Partner**			
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

https://docs.google.com/spreadsheets/d/1vAOhbuaQooo_O5cRZru4QlqgwzTncoXp9URFmTqvfkM/edit?usp=sharing

Project Proposal: [My Social Network Project Title]

Author 1, Author 2, [Author 3 if three]

1. Nodes and Links

Describe in plain language what the node and links are in your social network. This only needs to be a sentence or two.

2. Dataset

Tell us about your dataset. What is the subject matter? Where are you getting it? Most important: How are you getting it and how long will that take you? Describe any expected challenges, need for permissions, which tools you will use, what format the data is in, any expected data cleaning. Note: You may not use data that is already in network format.

3. Expected size of the network

State the number of nodes and number of links. Tell us why – is this everything available? Are you taking a subset? If a subset, why, and how do you plan to define it?

4. Questions you plan to ask and why we care

This is the part to really practice your skills at something akin to a scientific abstract (minus the results). Motivate us – what is so important and/or interesting about your social network? What research questions do you plan to pursue? Do you have any preliminary ideas about how you want to pursue these questions (e.g. community detection)?

Page limit: 1 page
You can use this however you see fit, there is no limit per section. You will all present a variety of networks that require more or less info in each section. You also do not to fill the whole page – succinct is good, just as long as you cover each component as described above.