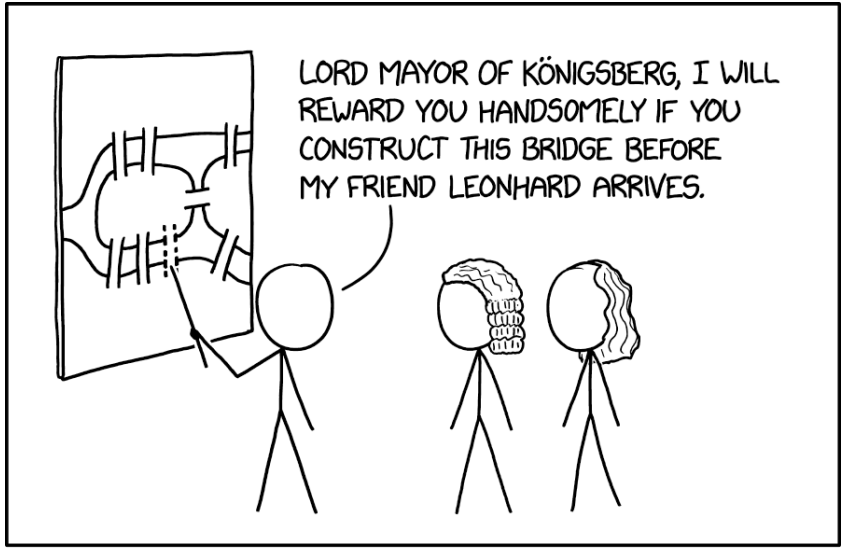
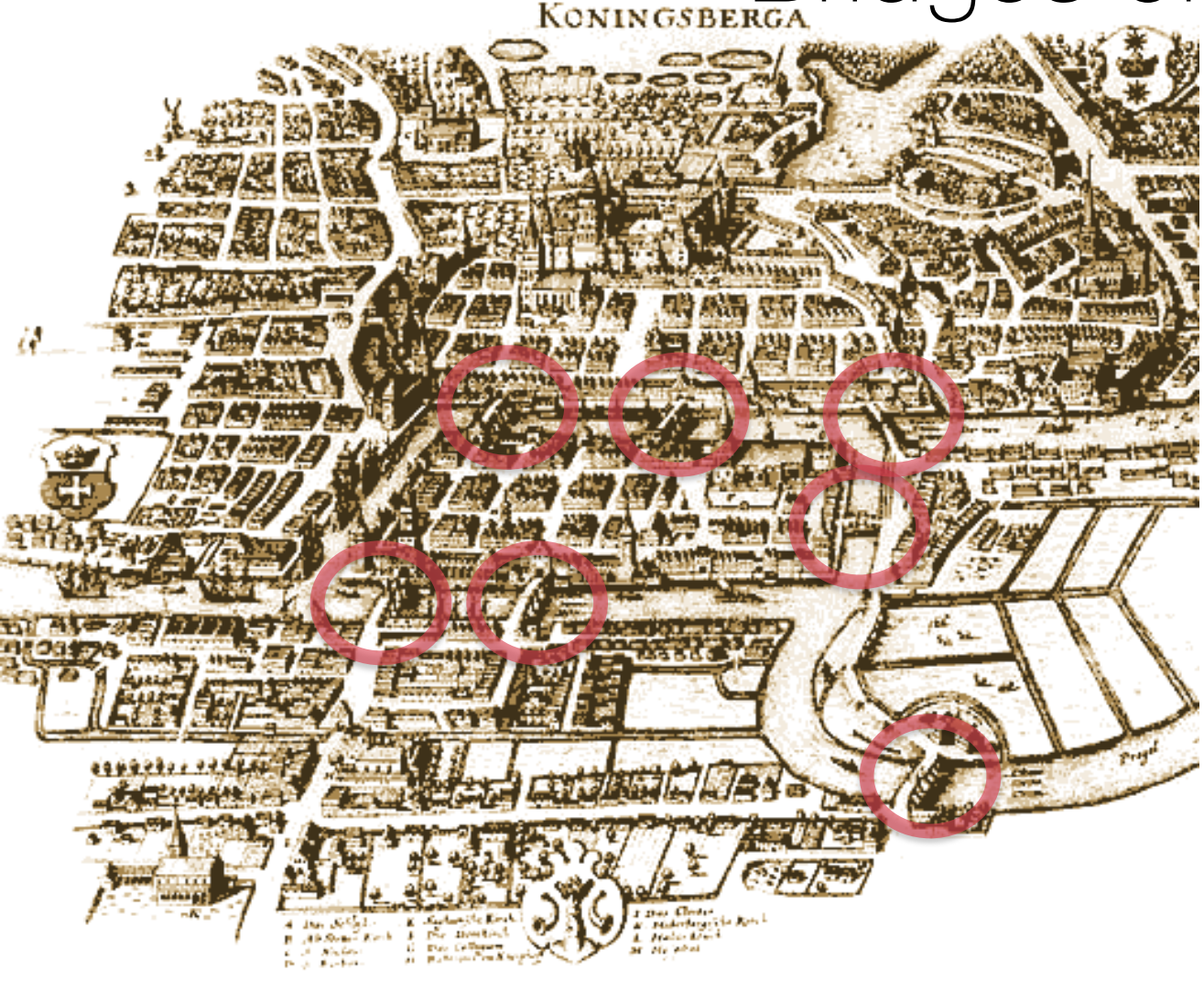
DS8104: Network Science

Class 2: First Connections a.k.a. graph theory 101



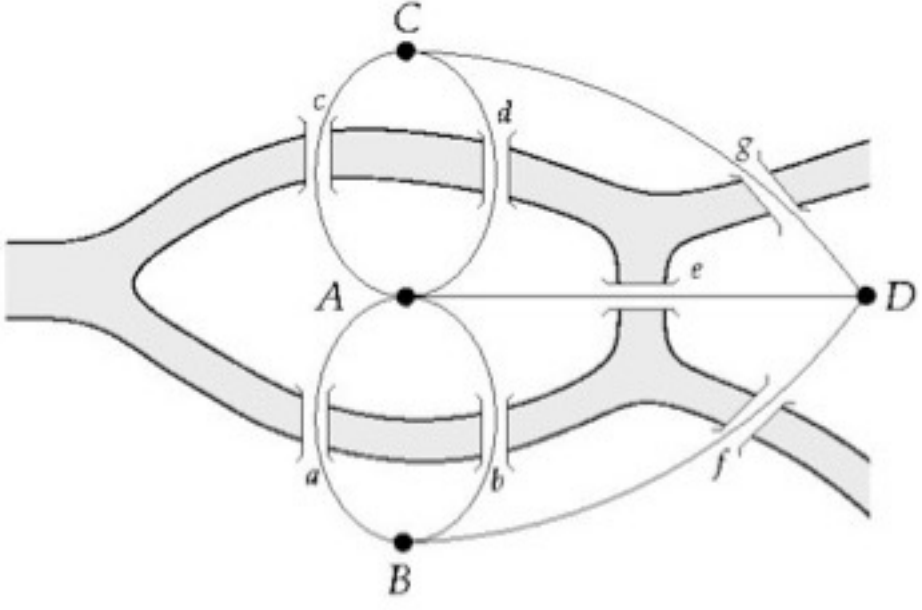
www.alexandergates.net

Bridges of Konigsberg



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

Bridges of Konigsberg



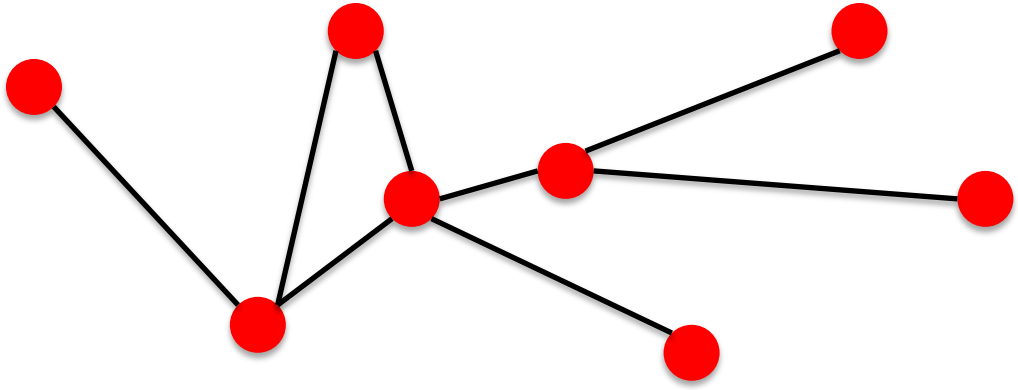
Can one walk across the seven bridges and never cross the same bridge twice?

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.

Basic building blocks



components: nodes, vertices, elements N

relationships: links, edges, interactions L

system: network, graph G:(N,L)

network often refers to real systems

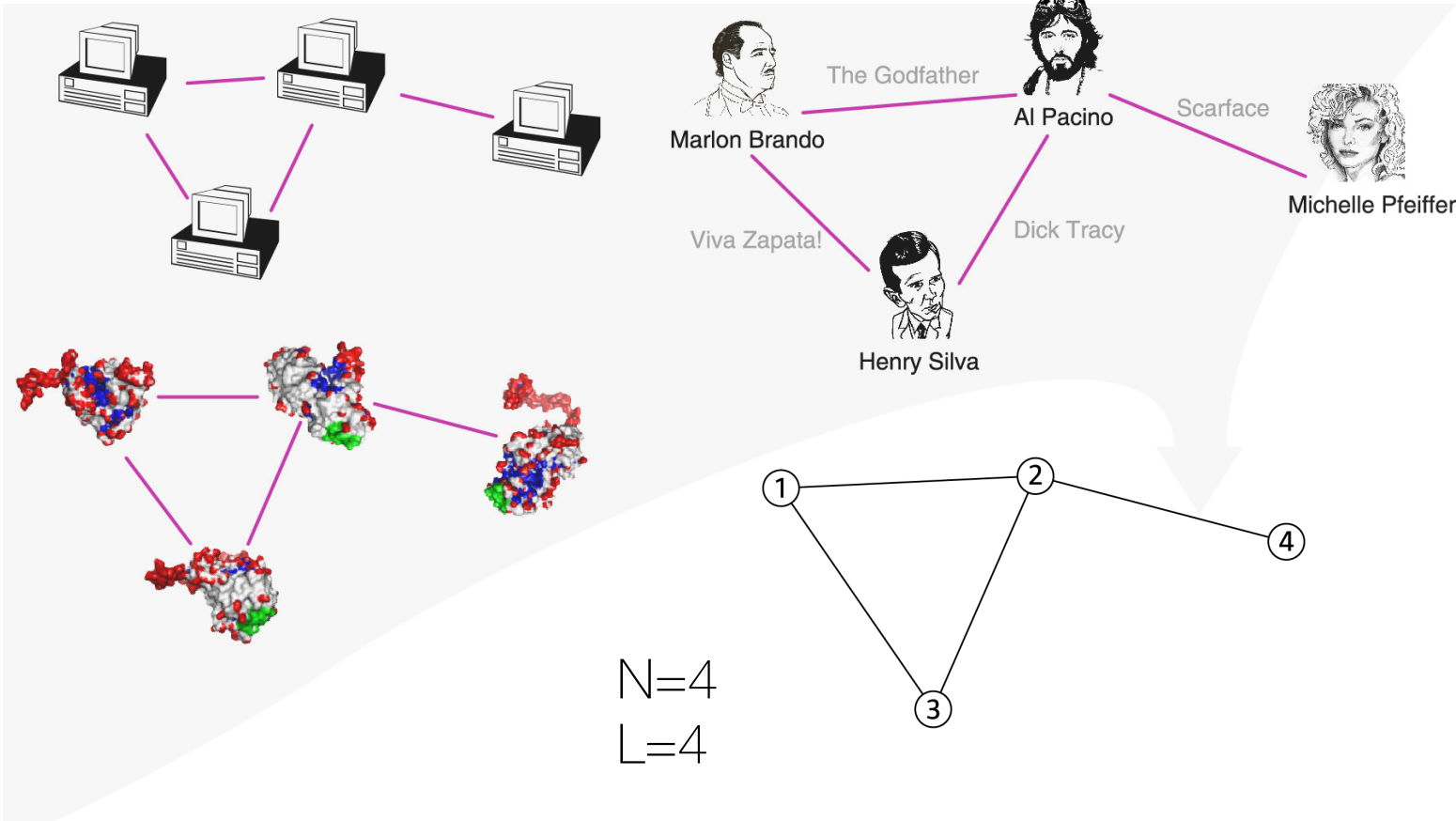
Language: (Network, node, link)

graph refers to the mathematical representation of a network

Language: (Graph, vertex, edge)

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

Networks provide a general framework to capture the pattern underlying a system's relationships



Whats my network?

One aspect of network science is to choose the appropriate set of nodes and relationships.

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.

PERSPECTIVE

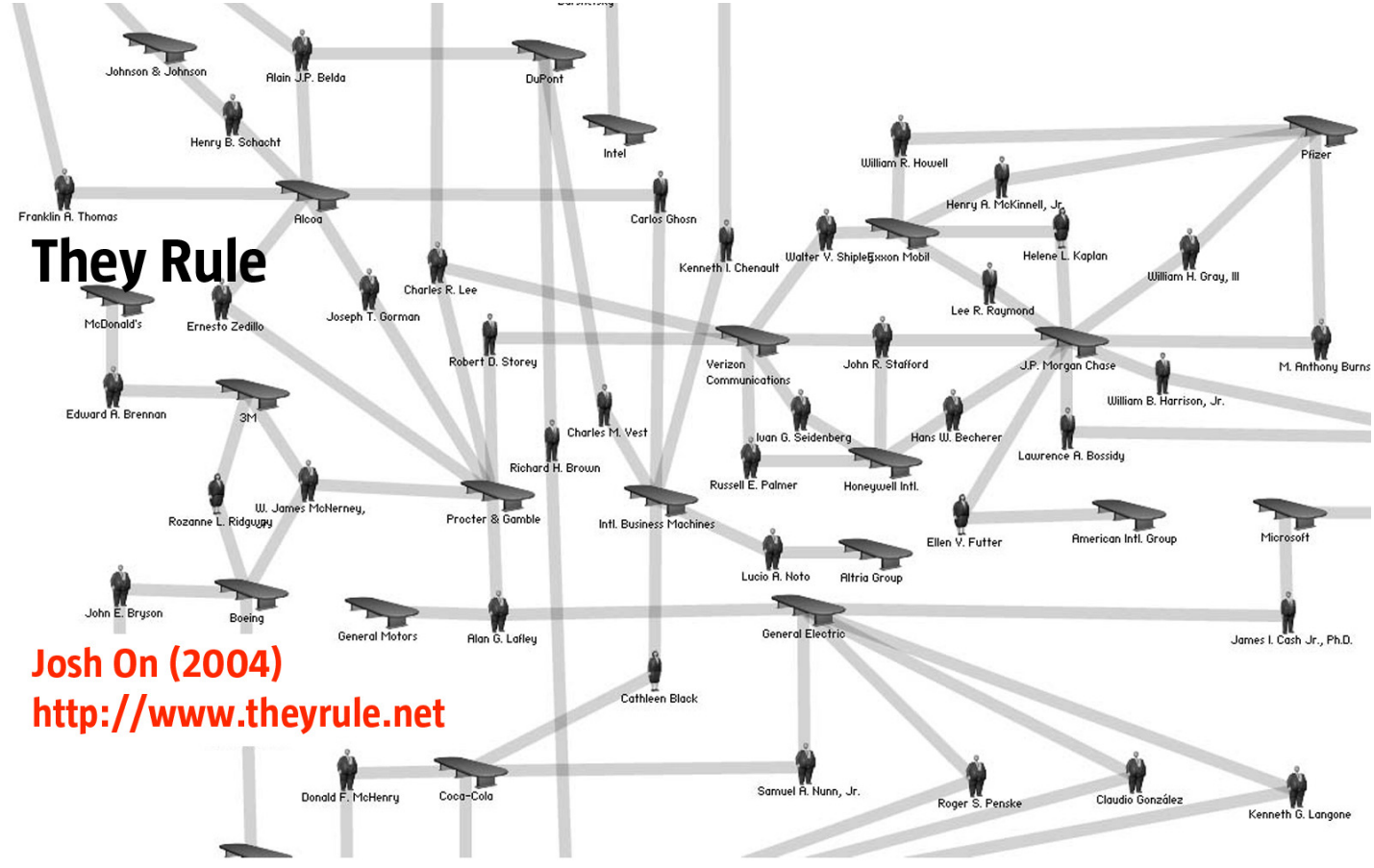
Revisiting the Foundations of Network Analysis

Carter T. Butts

Network analysis has emerged as a powerful way of studying phenomena as diverse as interpersonal interaction, connections among neurons, and the structure of the Internet. Appropriate use of network analysis depends, however, on choosing the right network representation for the problem at hand.

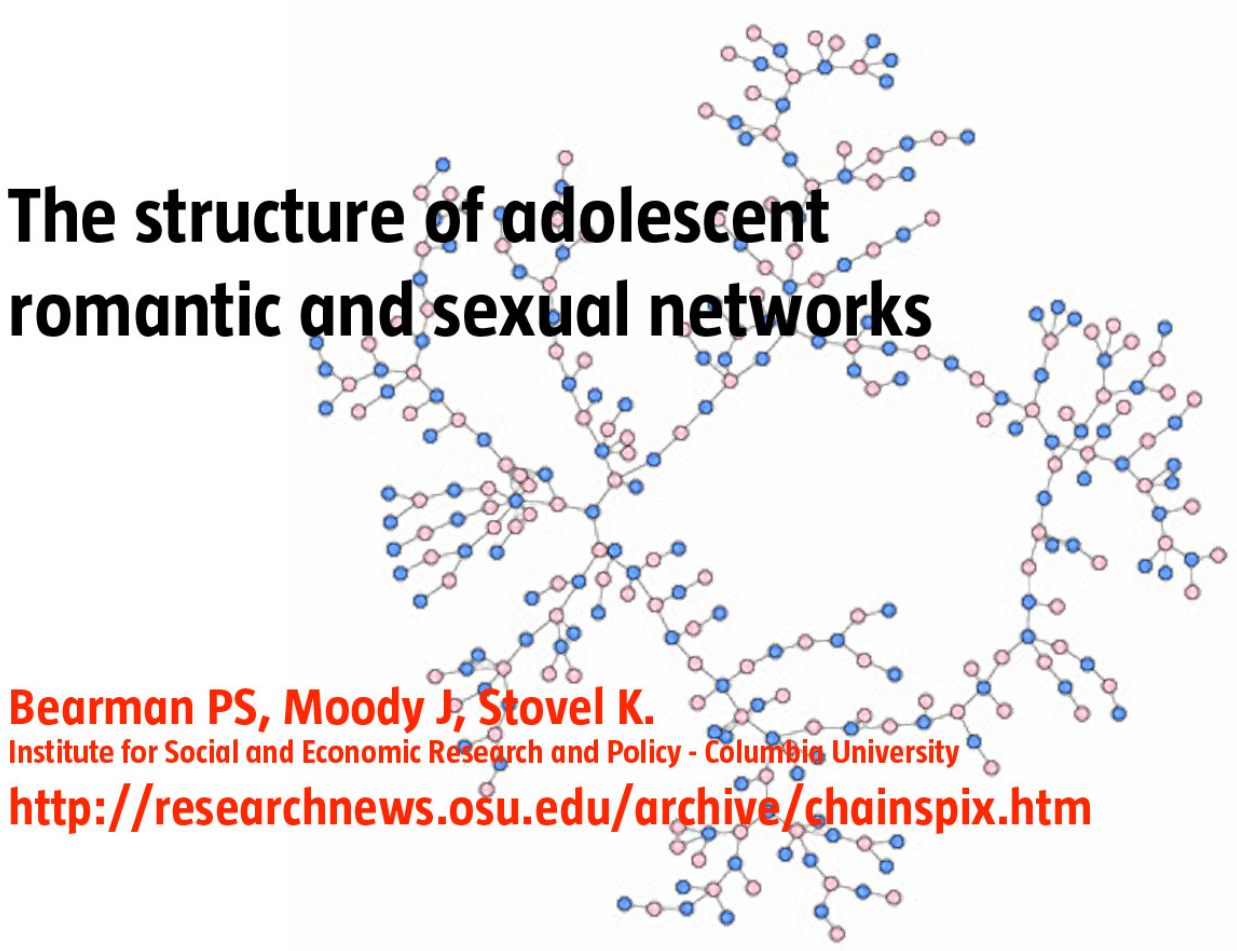
24 JULY 2009 VOL 325 **SCIENCE** www.sciencemag.org

One reading this week provides a deeper perspective on this issue.



If you connect individuals that work with each other,
you will explore their **professional network.**

If you connect those that have a romantic and sexual relationship, you will be exploring their **sexual network**.



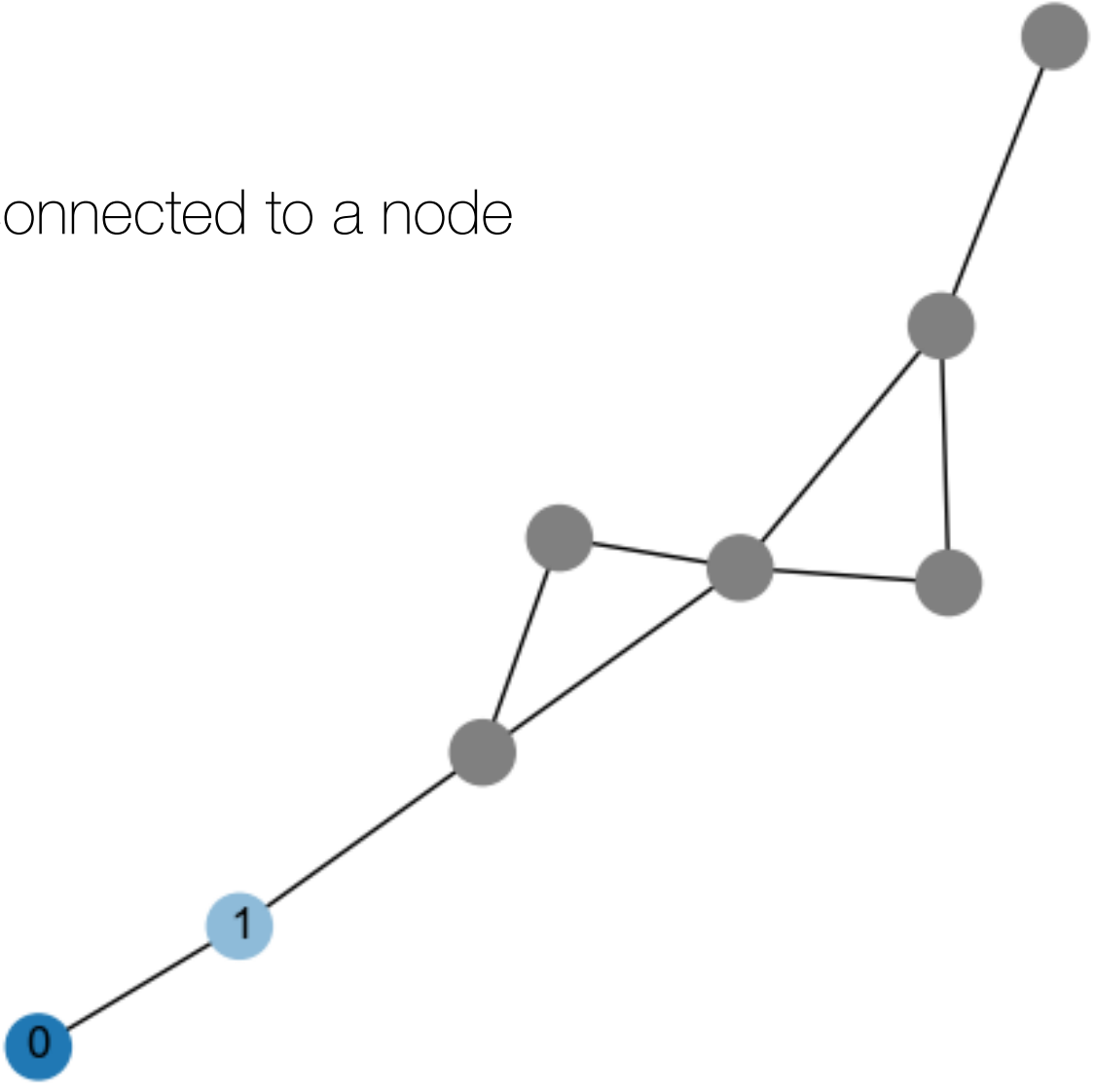
If you connect individuals based on their first name (all Peters connected to each other), you will be exploring what?

But this is a network, nevertheless.

Node degree

the number of links connected to a node

$$k_0 = 1$$

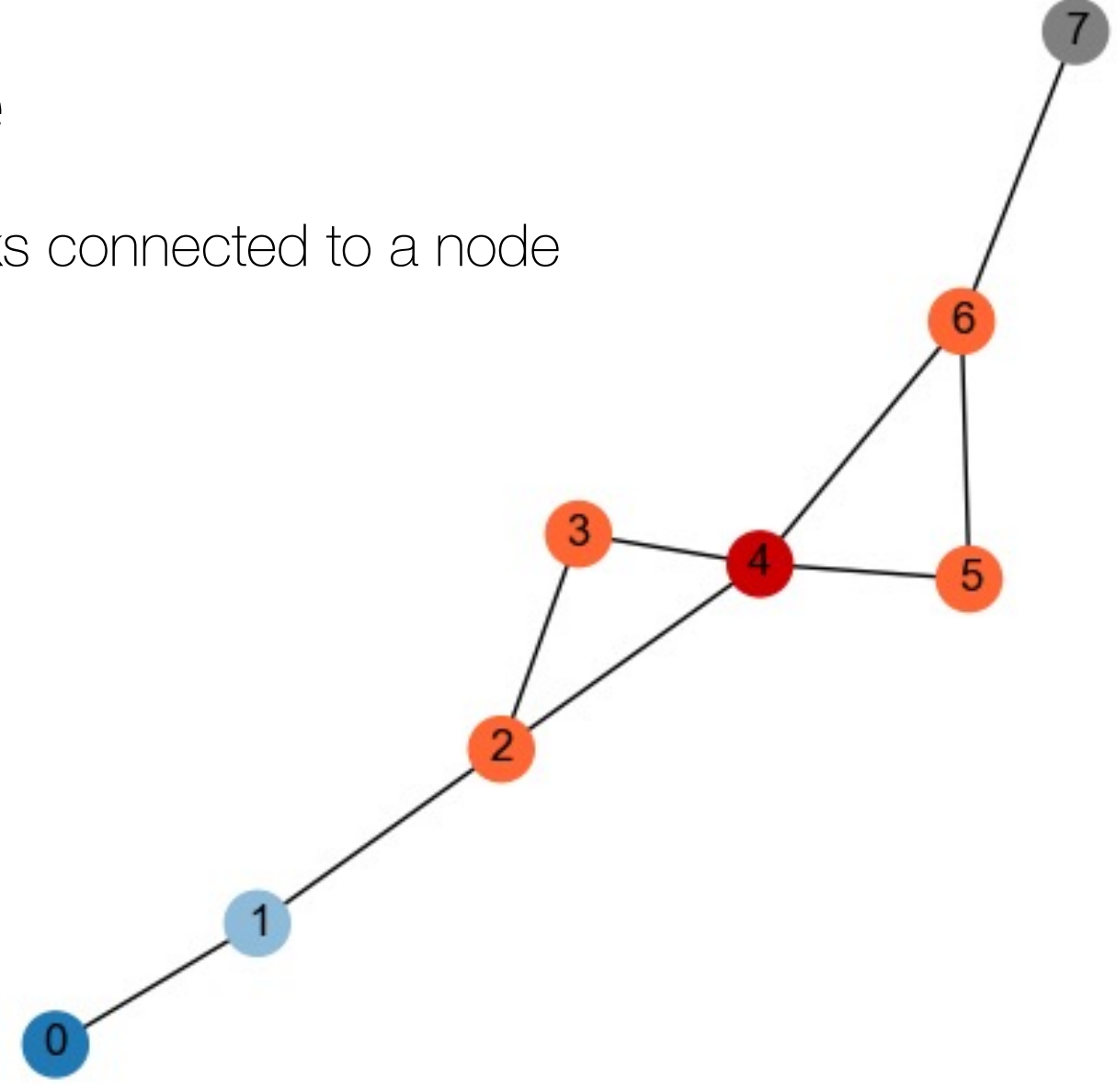


Node degree

the number of links connected to a node

$$k_0 = 1$$

$$k_4 = 4$$

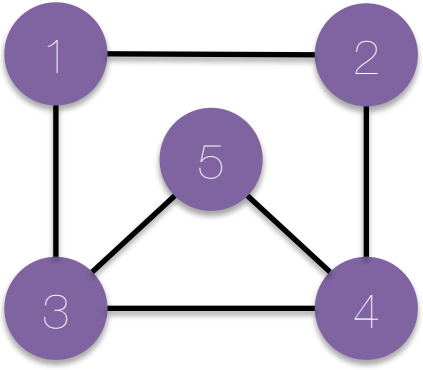


Average degree

$$\langle \boldsymbol{k} \rangle \equiv \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{k}_{\boldsymbol{i}}$$

$$\langle k \rangle \equiv \frac{2L}{N}$$

Graph Representations



Adjacency Matrix

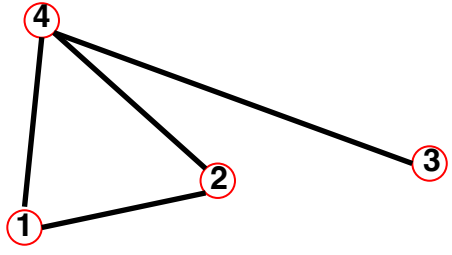
	0	1	2	3	4
0	0	1	1	0	0
1	1	0	0	1	1
2	1	0	0	0	1
3	0	1	0	0	1
4	0	1	1	1	0

source	target
0	1
0	2
1	3
1	4
2	4
3	4

Edge List

Adjacency List

```
node : list of neighbors
{
        "0": [1, 2],
         "1": [0, 3, 4],
         "2": [0, 4],
         "3": [1, 4],
         "4": [1, 2, 3]
}
```



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

$$A_{ij} = A_{ji}$$

$$A_{ii} = 0$$

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

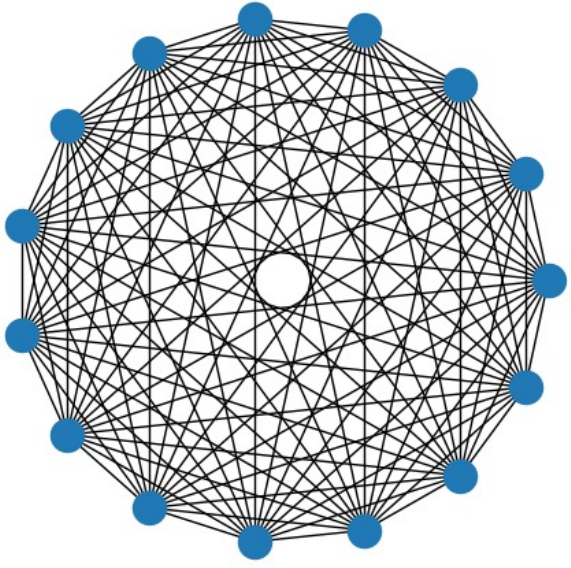
$$k_j = \sum_{i=1}^{N} A_{ij}$$

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

Dense graphs

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

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



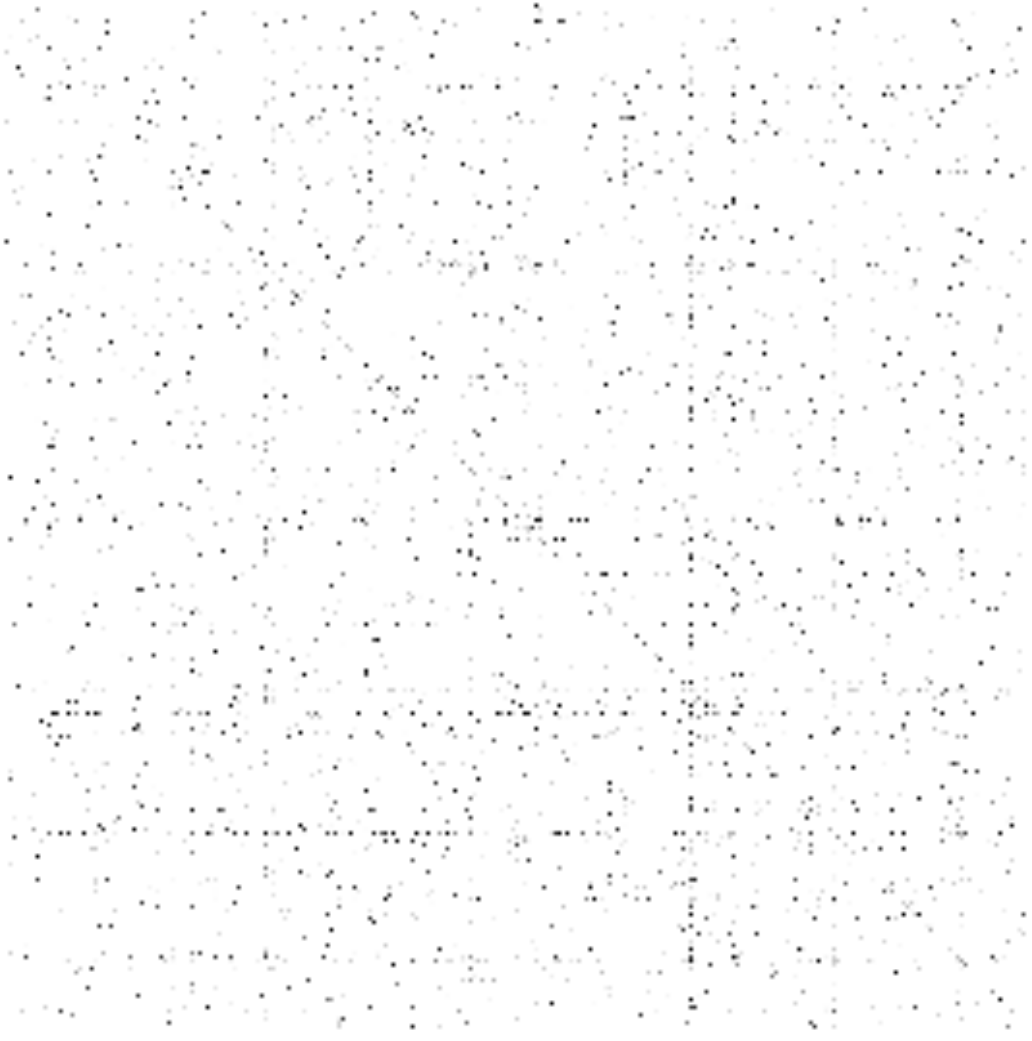
A graph with links $L=L_{max}$ is called a **complete graph**

$$< k >= k_i = \mathrm{N} - 1$$

Most observed networks have far fewer links (they are **sparse**)

$$< k > \ll \text{N - 1}$$

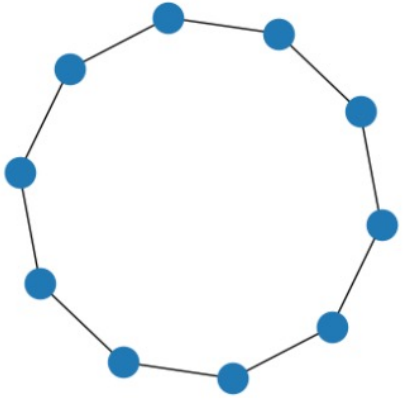
Network	N	L	$< k >$
UVA Facebook	17196	789321	91.8
Protien Interaction	2018	2930	2.9
IMDB Co-acting	702388	29397908	83.71
Powergrid	4941	6594	2.67
E coli. metabolism	1039	4741	9.13

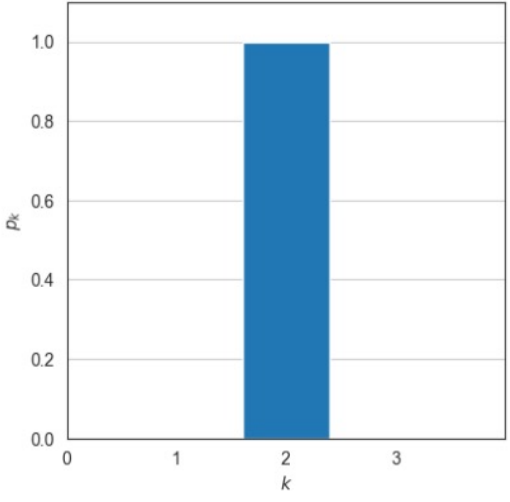


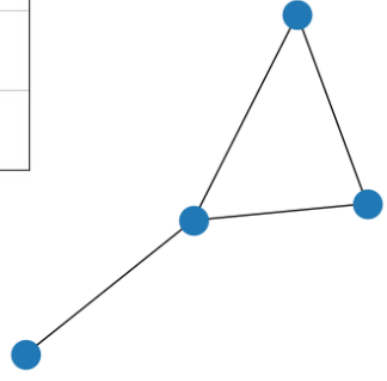
Real networks are sparse

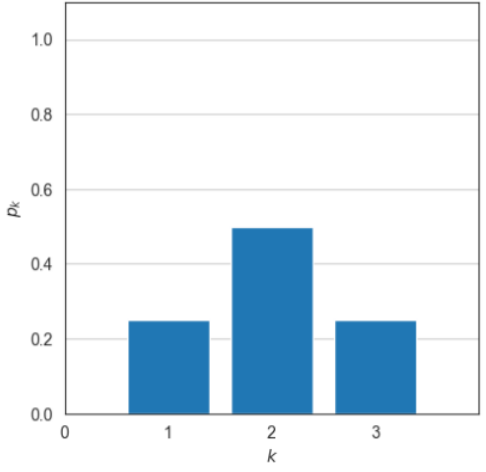
Degree distribution

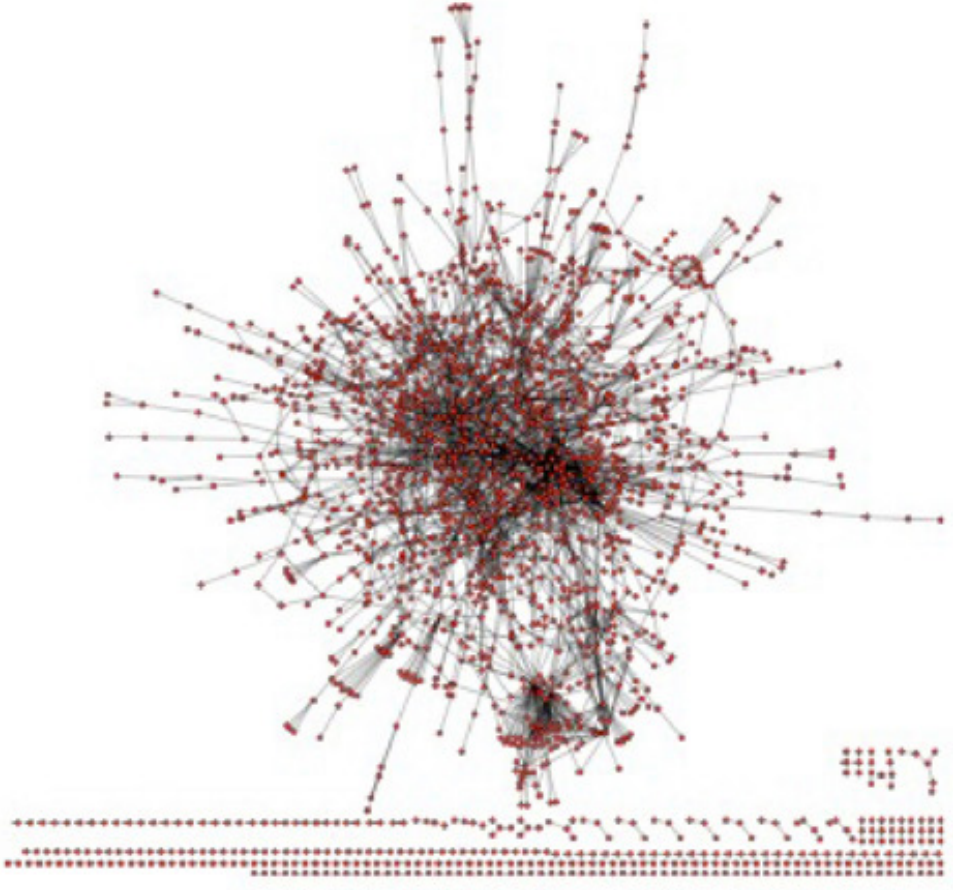
P(k): probability that a randomly chosen node has degree k

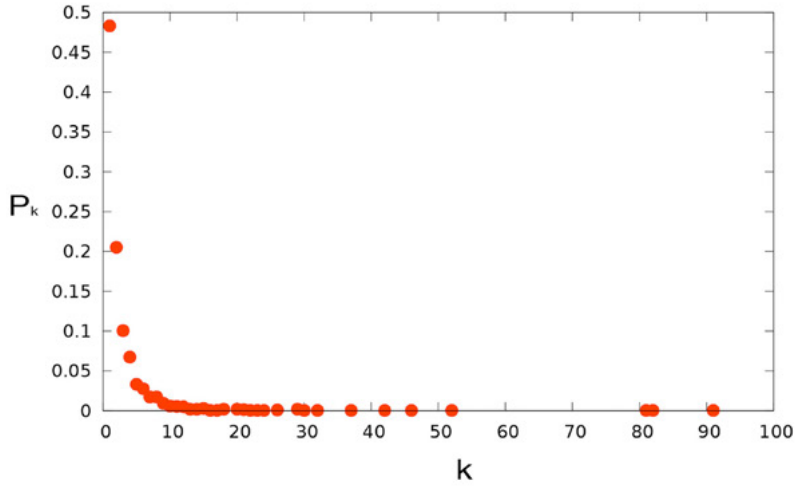


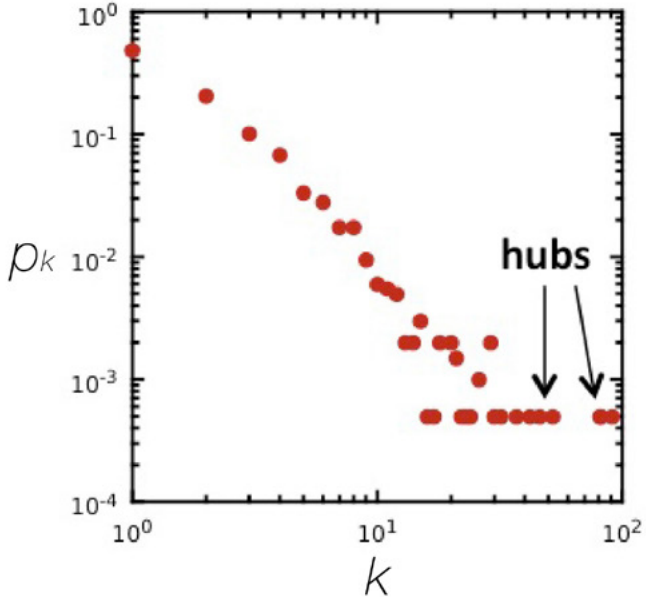









0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Pk
0
10
20
30
40
50
60
70
80
90
100
k


10^0
10^-1
10^-2
10^-3
10^-4
pk
hubs
10^0
10^1
10^2
k

Graph Types

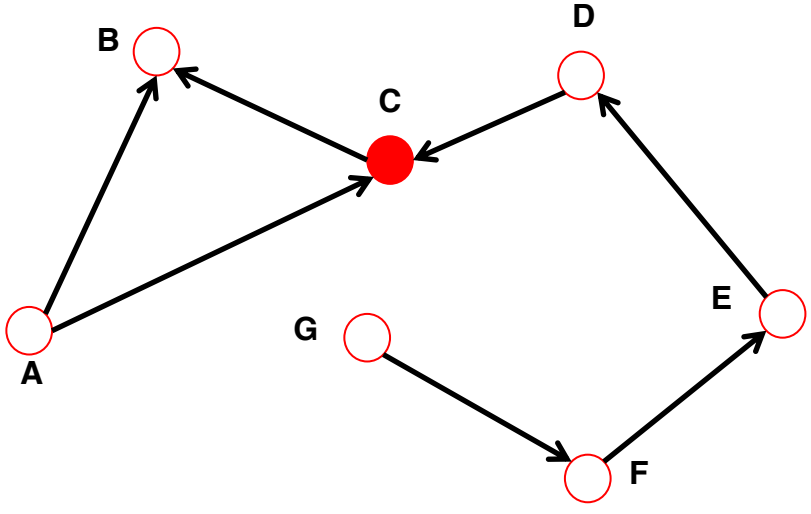
Directed graphs

So far the relationships we have considered are bi-directional (a.k.a. symmetric)

But many relationships are uni-directional

Alex follows Beyonce
(but Beyonce doesn't follow Alex….yet)

In directed networks we can define an **in-degree** and **out-degree**. The (total) degree is the sum of in- and out-degree.



$$k_C^{in} = 2$$

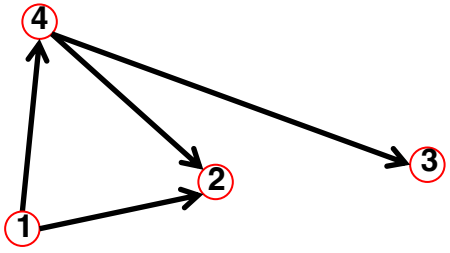
$$k_C^{out} = 1$$

$$k_C = 3$$

$$\langle k^{in} \rangle = \langle k^{out} \rangle$$

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

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



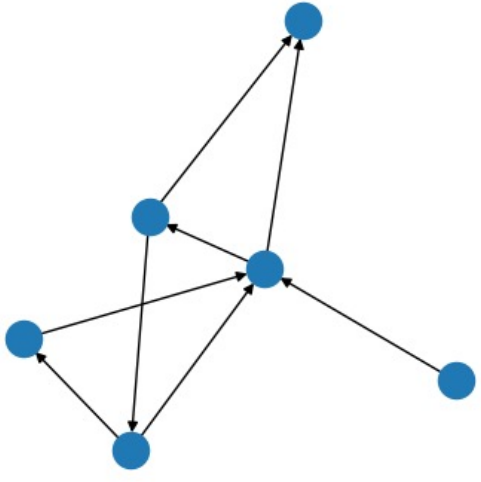
$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

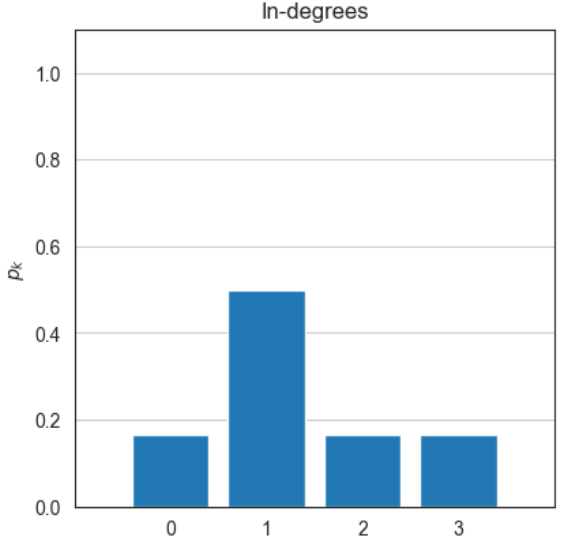
$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

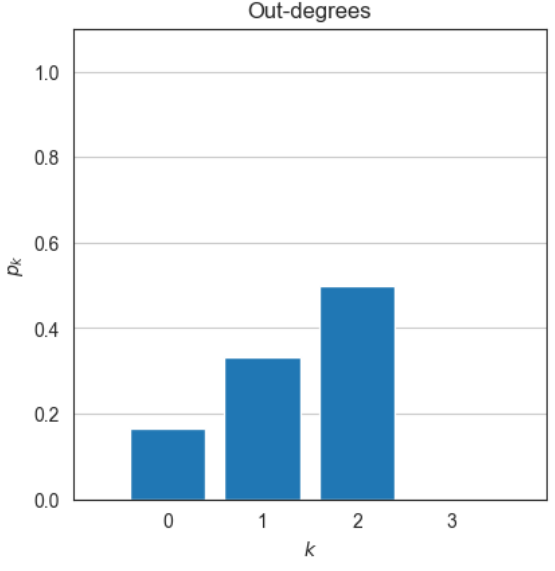
$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

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

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





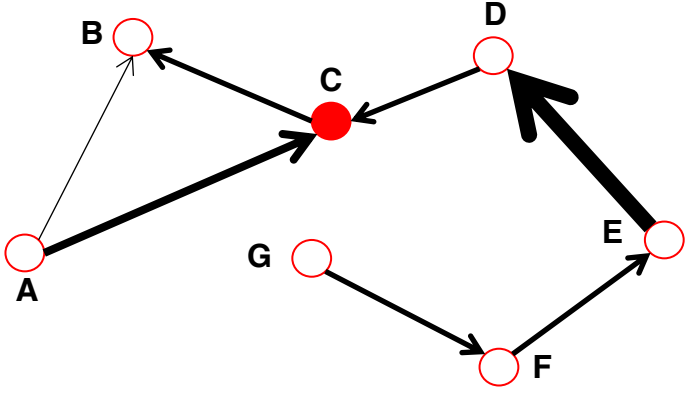


Weighted graphs

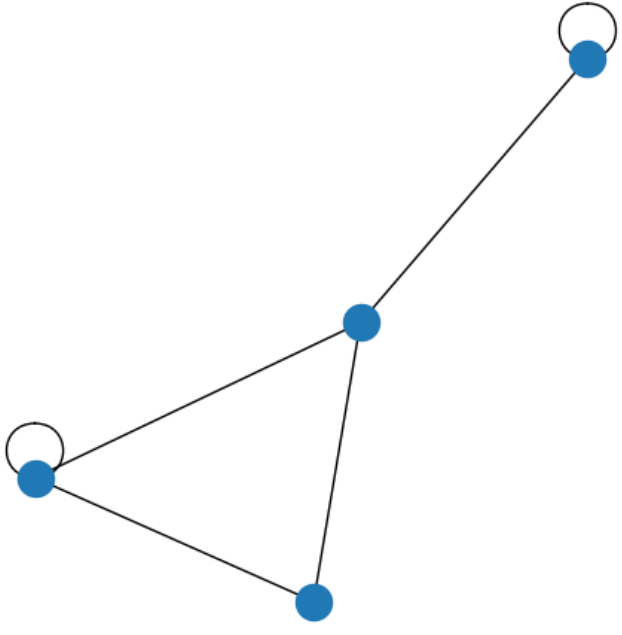
Sometimes, not all edges are equivalent

In many applications, edges carry additional information by an **edge weight**

	Unweighted	Weighted
Adjancy value	$A_{ij} = 1$	$A_{ij} \in \mathbb{R}$



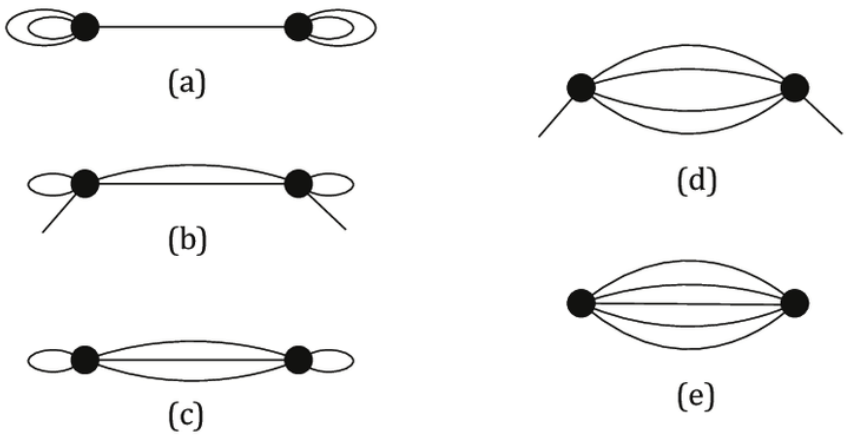
Self interactions

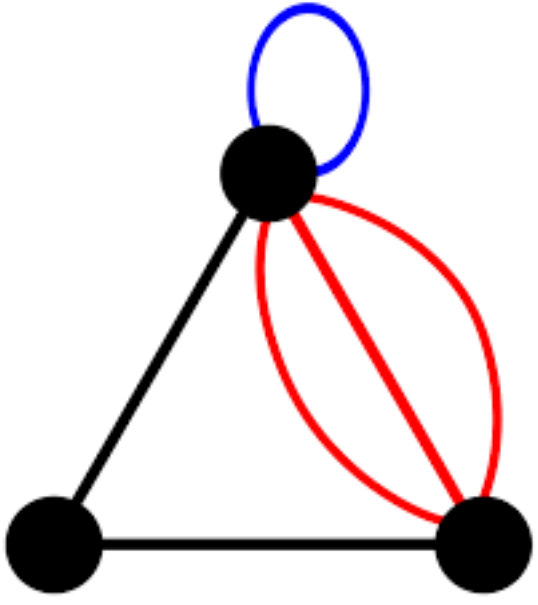


$$A_{ii} \neq 0$$

	0	1	2	3
0	1	1	1	0
1	1	0	1	0
2	1	1	0	1
3	0	0	1	1

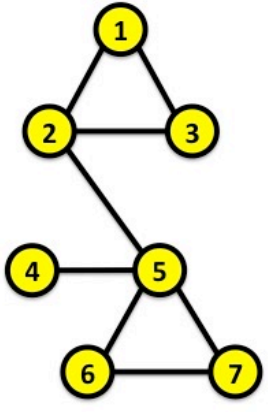
Multigraphs

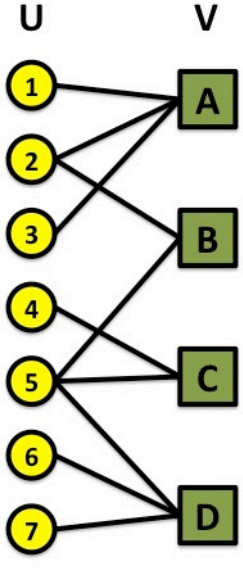


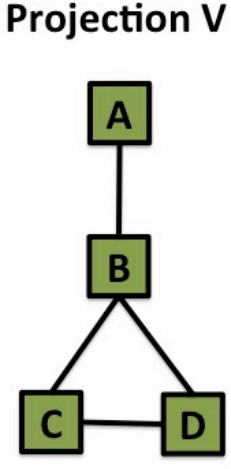


Bipartite Graphs

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.



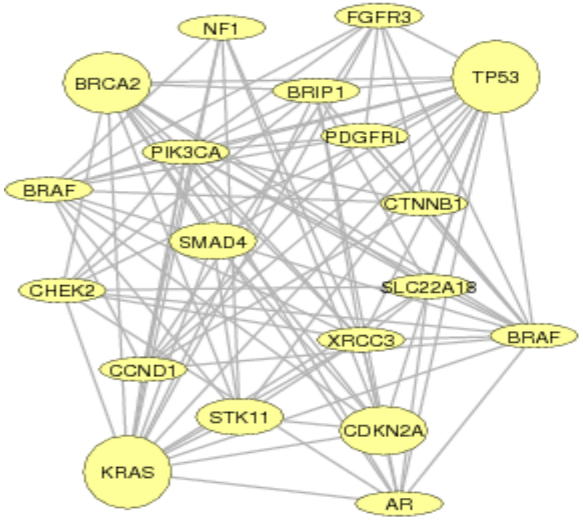




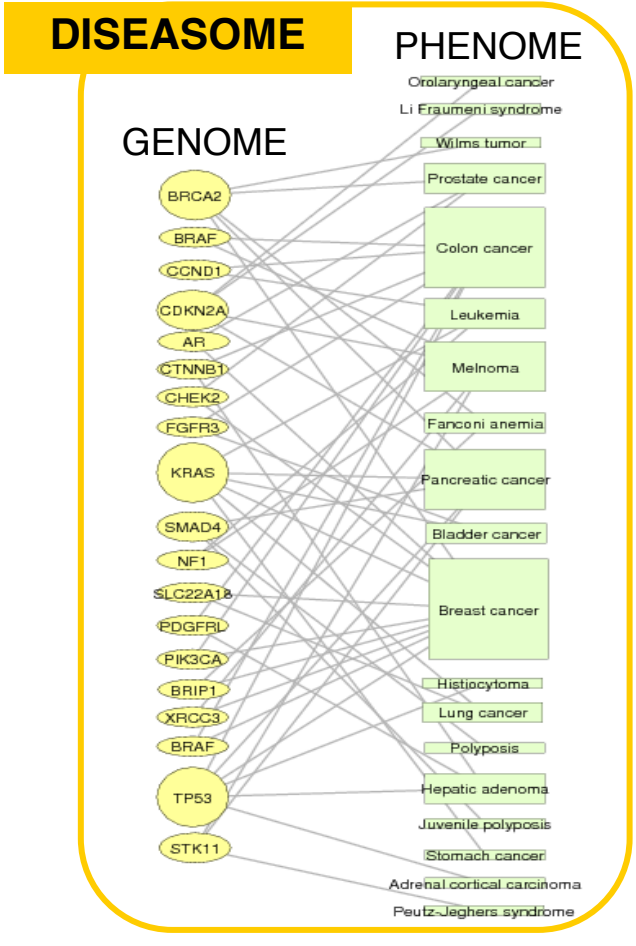
Examples:

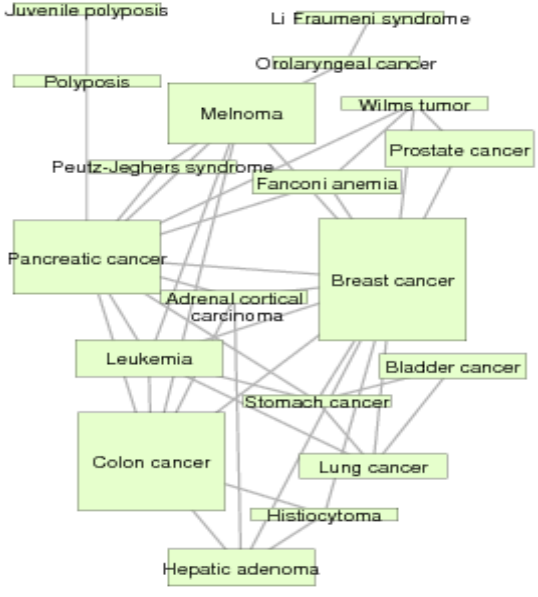
Hollywood actor network
Collaboration networks
Disease network (diseasome)

BB^T $B = M_{u,v}$ $B^T B$



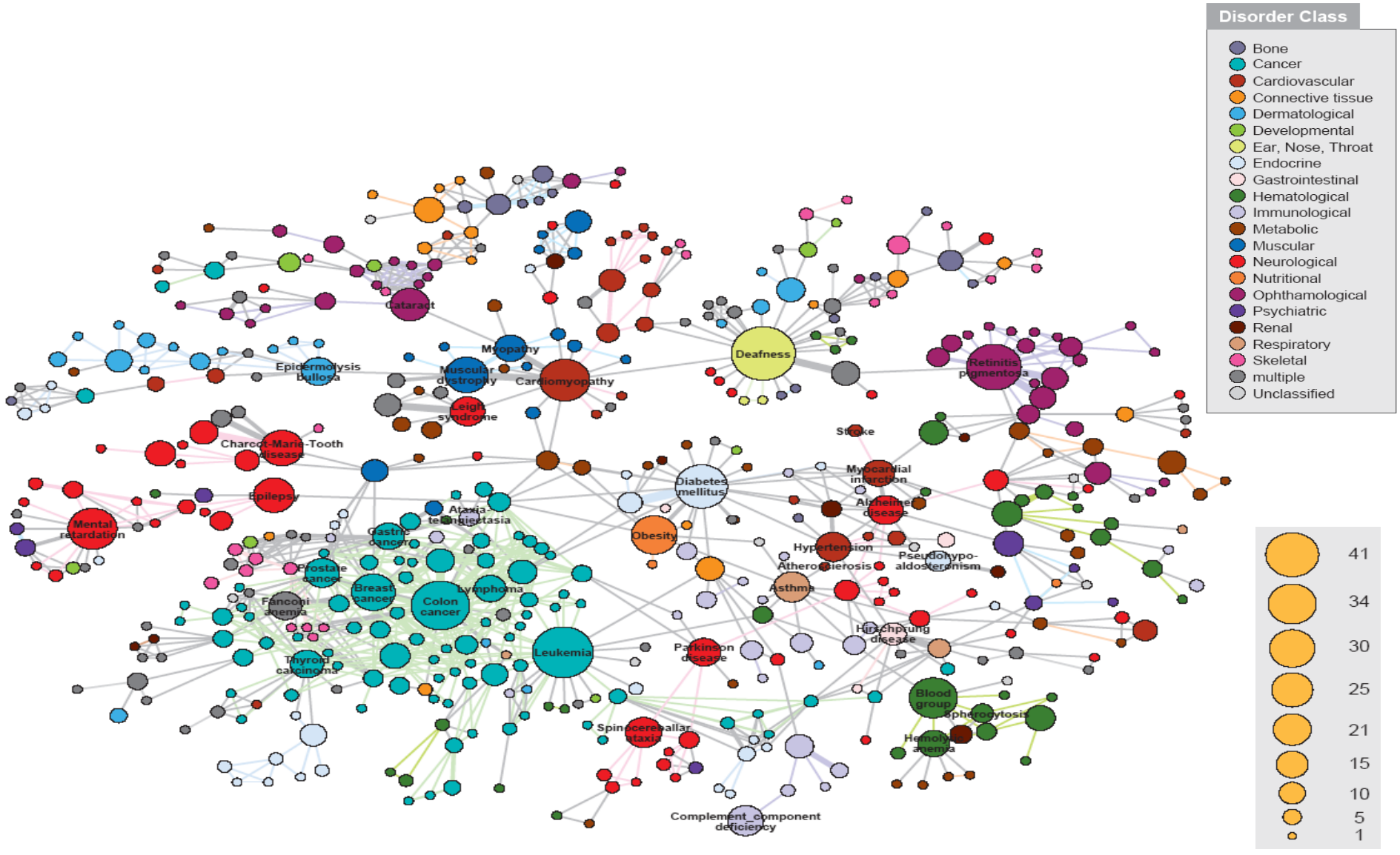
Gene network

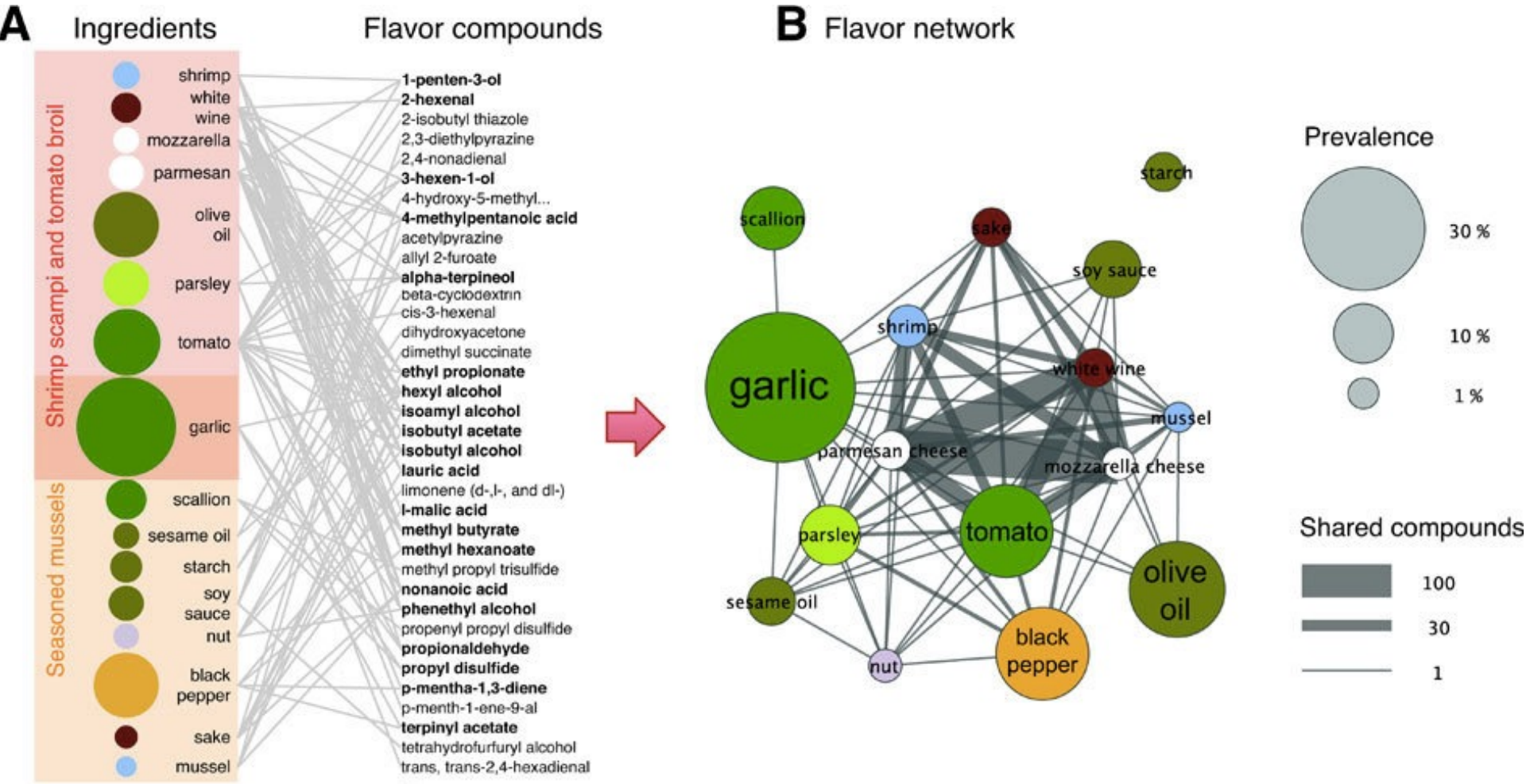




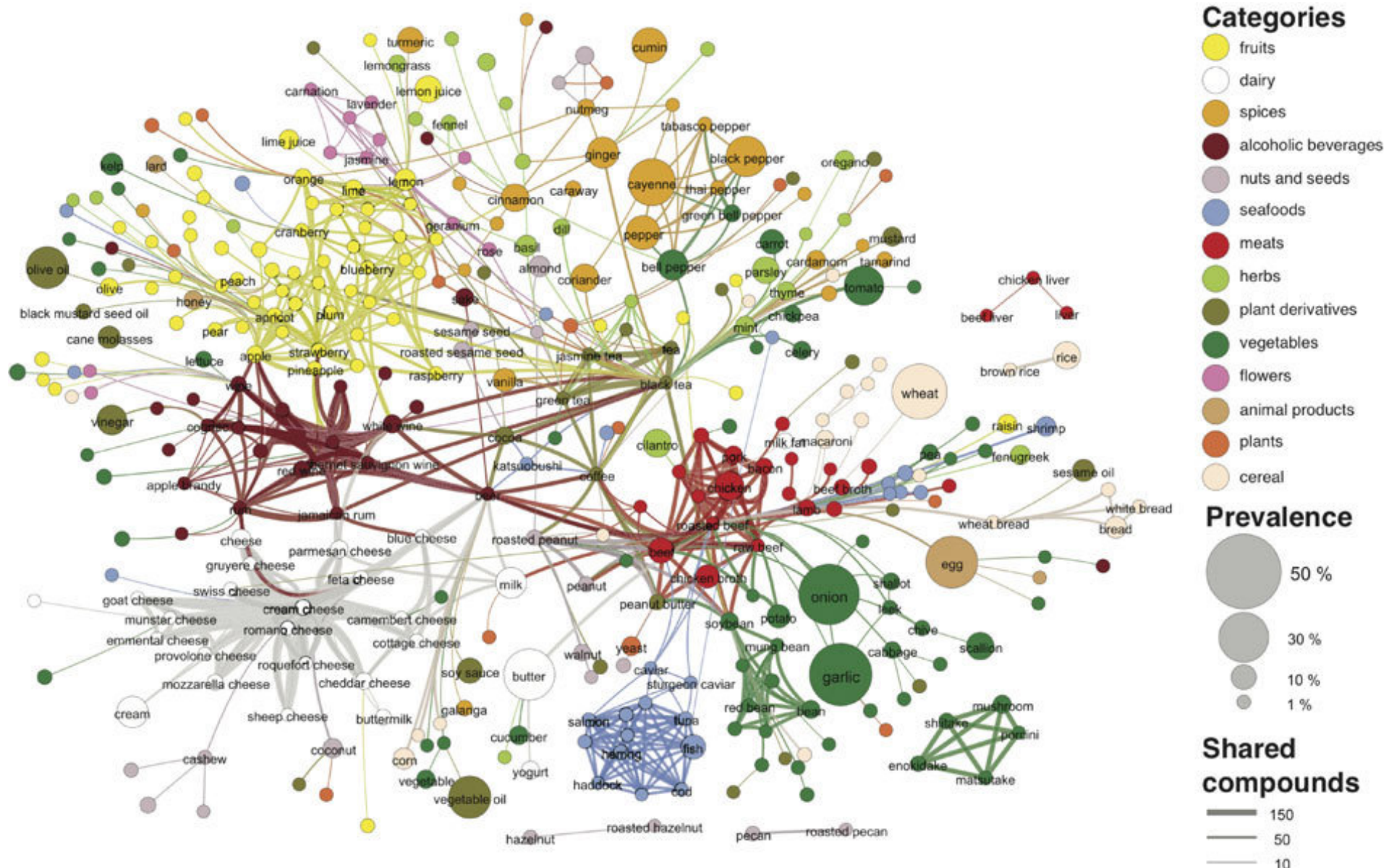
Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási (2007) The human disease network. *PNAS* 104


Disorder Class
Bone
Cancer
Cardiovascular
Connective tissue
Dermatological
Developmental
Ear, Nose, Throat
Endocrine
Gastrointestinal
Hematological
Immunological
Metabolic
Muscular
Neurological
Nutritional
Ophthamological
Psychiatric
Renal
Respiratory
Skeletal
multiple
Unclassified
41
34
30
25
21
15
10
5
1
Cataract
Epidermolysis bullosa
Myopathy
Muscular dystrophy
Cardiomyopathy
Leigh syndrome
Deafness
Retinitis pigmentosa
Charcot-Marie-Tooth disease
Epilepsy
Mental retardation
Ataxia-telangiectasia
Gastric cancer
Prostate cancer
Breast cancer
Colon cancer
Lymphoma
Fanconi anemia
Thyroid carcinoma
Leukemia
Diabetes mellitus
Obesity
Stroke
Myocardial infarction
Alzheimer disease
Hypertension
Atherosclerosis
Pseudohypo-aldosteronism
Asthma
Hirschprung disease
Parkinson disease
Blood group
Spherocytosis
Hemolytic anemia
Spinocerebellar ataxia
Complement_component deficiency



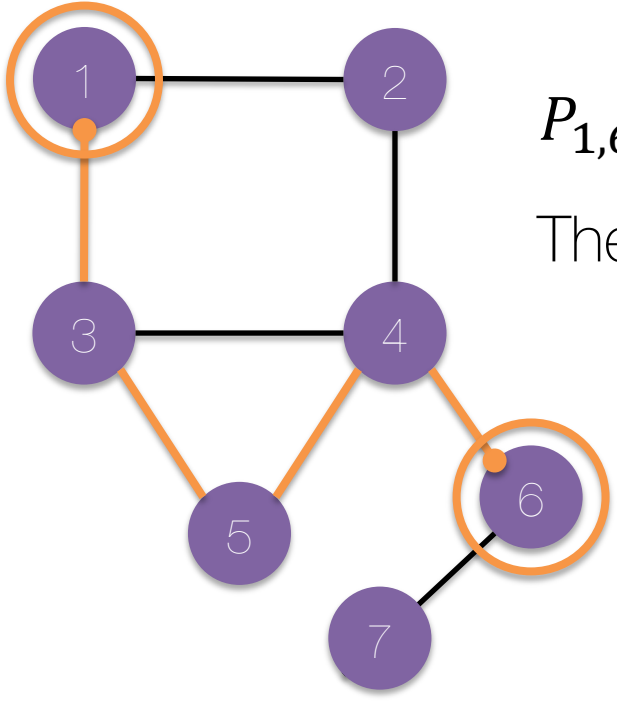
Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási (2011) Flavor network and the principles of food pairing , *Scientific Reports* 196


turmeric
lemongrass
cumin
carnation
lemon juice
lavender
nutmeg
fennel
tabasco pepper
lime juice
ginger
black pepper
jasmine
kelp
lard
orange
lime
lemon
cayenne
thai pepper
oregano
caraway
cinnamon
green bell pepper
cranberry
geranium
dill
pepper
carrot
mustard
rose
basil
olive oil
blueberry
almond
bell pepper
cardamom
tamarind
parsley
coriander
peach
olive
thyme
tomato
chicken liver
honey
black mustard seed oil
apricot
plum
chickpea
beef liver
liver
pear
sesame seed
mint
cane molasses
lettuce
apple
strawberry
roasted sesame seed
jasmine tea
tea
celery
rice
pineapple
raspberry
vanilla
brown rice
wheat
green tea
black tea
vinegar
white wine
cocoa
raisin
shrimp
cilantro
milk
macaroni
pork
fenugreek
red wine
cabernet sauvignon wine
katsuobushi
bacon
sesame oil
apple brandy
coffee
chicken
beef broth
rum
jamaican rum
lamb
white bread
wheat bread
bread
cheese
blue cheese
roasted peanut
roasted beef
parmesan cheese
beef
raw beef
egg
gruyere cheese
feta cheese
milk
peanut
chicken broth
shallot
swiss cheese
onion
goat cheese
peanut butter
leek
cream cheese
camembert cheese
soybean
potato
munster cheese
romano cheese
chive
emmental cheese
cottage cheese
mung bean
scallion
provolone cheese
walnut
yeast
cabbage
roquefort cheese
caviar
soy sauce
butter
garlic
mozzarella cheese
cheddar cheese
sturgeon caviar
red bean
bean
cream
sheep cheese
buttermilk
galanga
mushroom
salmon
tuna
shiitake
cucumber
porcini
herring
fish
cashew
coconut
corn
yogurt
enokidake
vegetable
haddock
cod
matsutake
vegetable oil
hazelnut
roasted hazelnut
pecan
roasted pecan
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

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

$P_{0,n} = (0,1,2,\ldots,n)$ $P_{0,n} = ((0,1),(1,2),\ldots,(n-1,n))$

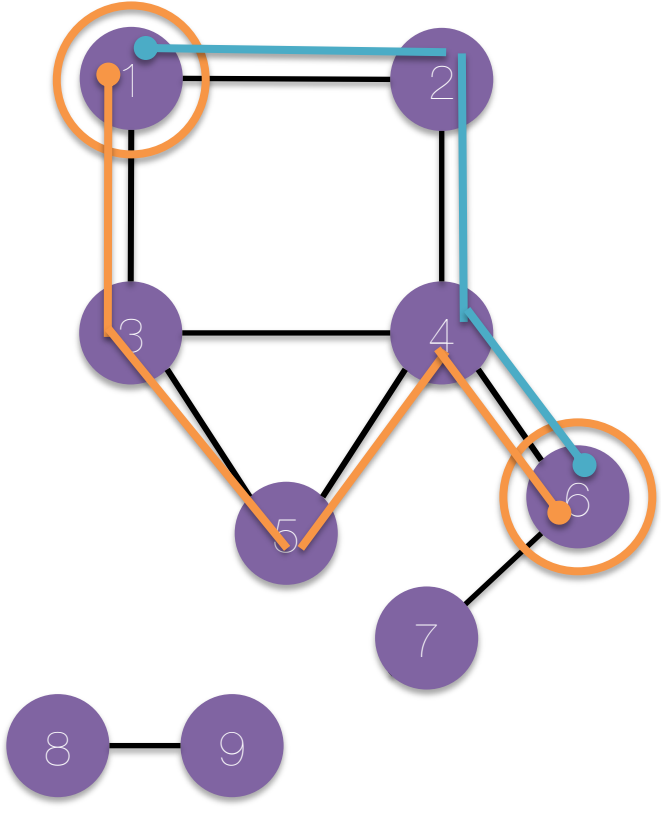


$P_{1,6} = (1,3,5,4,6)$

The **path length** is the number of links.

Pathology

There are typically many paths between any two nodes



$P_{1,6} = (1,3,5,4,6)$ $P_{1,6} = (1,2,4,6)$

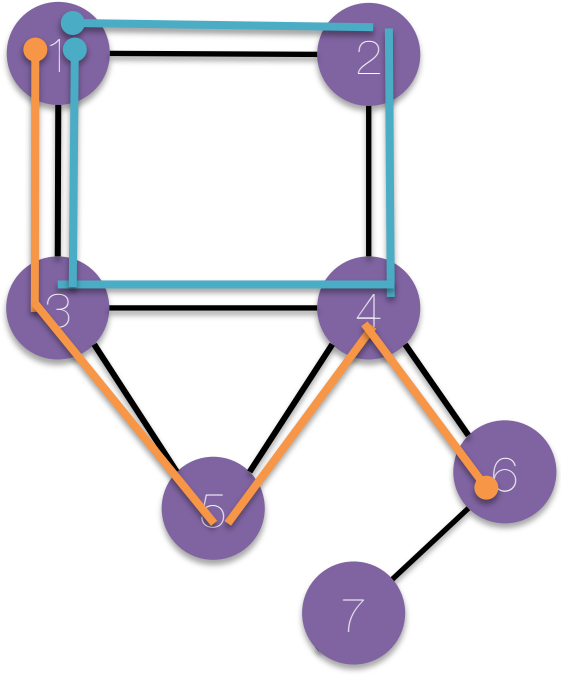
The **shortest path** is the path with the fewest number of links.

The **distance** between two nodes is the length of the shortest path.

If no paths exist between two nodes, the distance is infinite.

Pathology

There are typically many paths between any two nodes

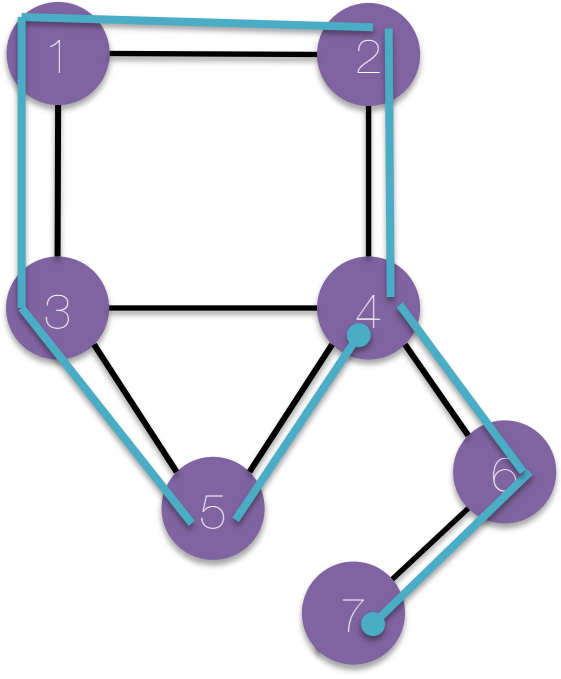


$P_{1,1} = (1,2,4,3,1)$

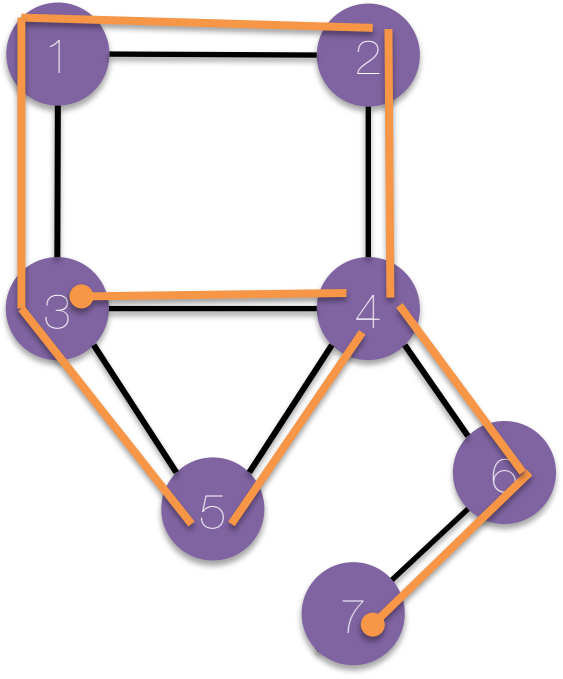
A **cycle** is a path that starts and stops at the same node.

A **self-avoiding path** is a path that does not intersect itself.

Pathology

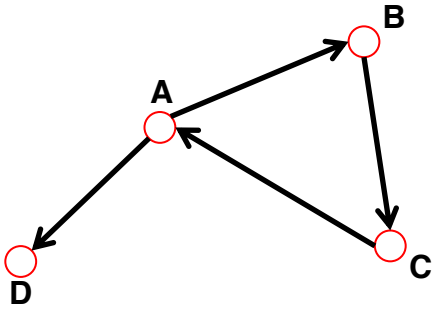


A **Hamilton path** is a path that touches each node exactly once.



A **Euclidean path** is a path that touches each link exactly once.

Pathology



In directed graphs each path needs to follow the direction of the arrows.

Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

Pathology

The graph **diameter** is the maximum distance between any two nodes in the graph (maximum shortest path length).

The **average distance** in a graph:

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

Pathology

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

N_{ij}^{n} number of paths between any two nodes i and j

N_{ij}^{1} Is there a path of length 1 between nodes i and j

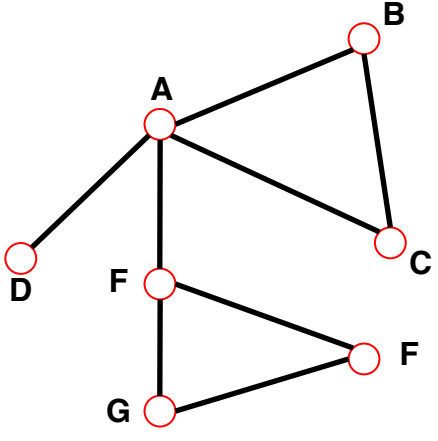
$$N_{ij}^{1} = A_{ij}$$

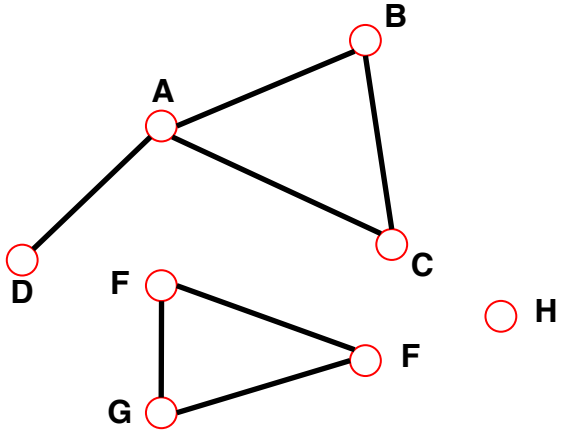
N_{ij}^{2} If there is a path of length 2 between nodes i and j, then there is a node k where $A_{ik}A_{kj} = 1$

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

Connectedness

Connected graph:
(undirected) any two vertices can be joined by a path.

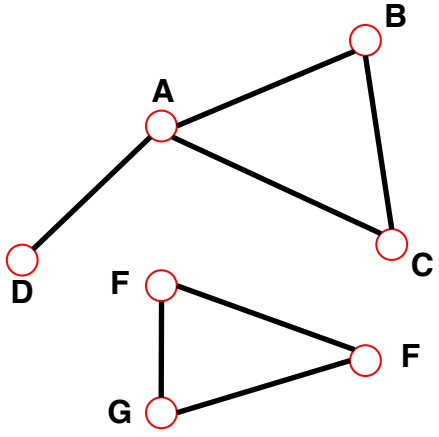




Disconnected graph: is made up of two or more connected graphs (aka connected components).

Connectedness

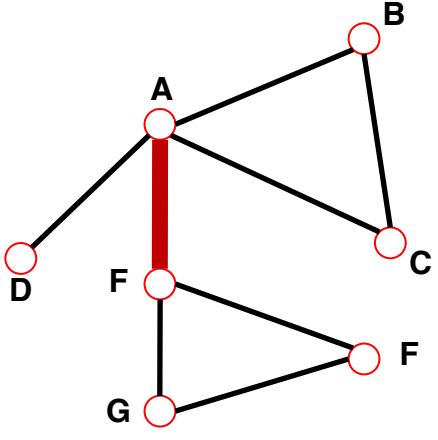
Largest connected component
(a.k.a. giant component) is the one with the most nodes



Isolates
are the rest of the components

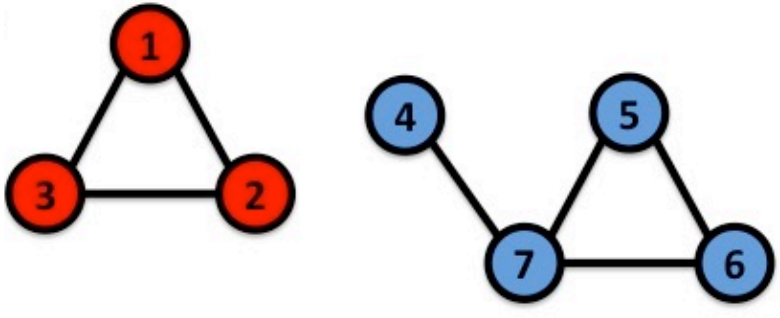
Connectedness

Bridge: an edge whose removal breaks the graph in disconnected components

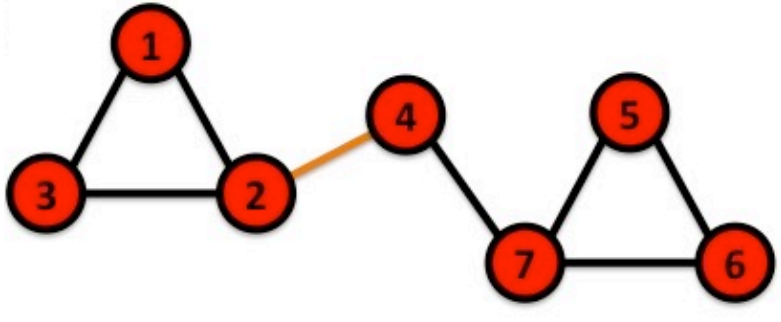


Connectedness

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:



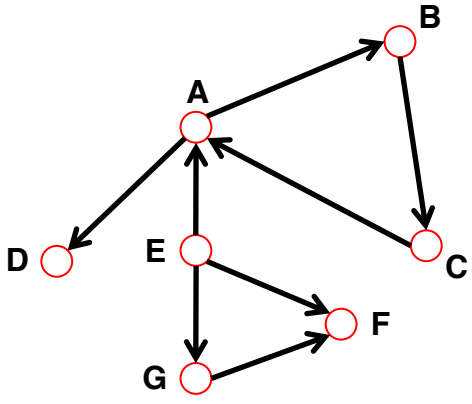
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$



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

Connectedness

Strongly connected digraph:
has a directed path between every pair of nodes (and vise versa)

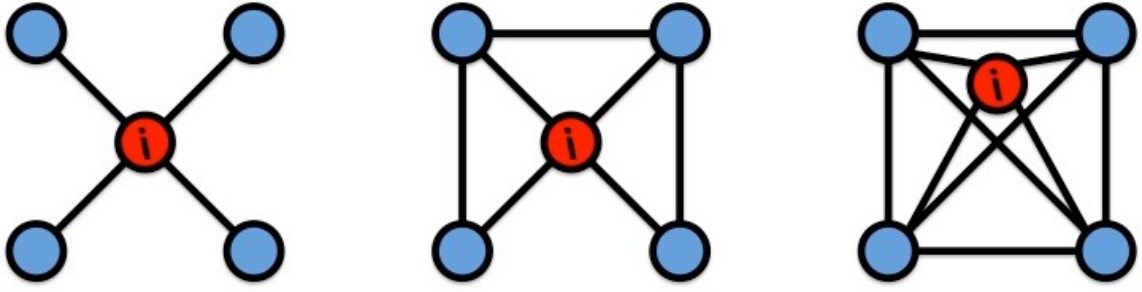


Strongly connected compent:
a subgraph with a path between every pair of nodes (and vise versa)

Weakly connected digraph:
connected if we disregard the edge direction

Clustering

What fraction of your neighbors are also connected to each-other?



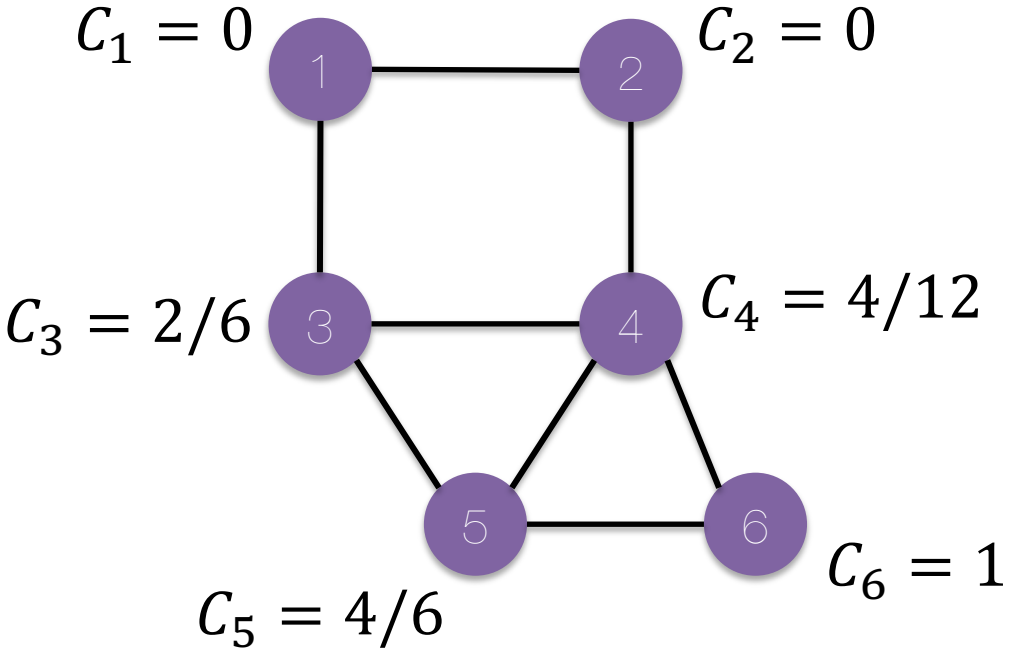
$C_i = 0$ $C_i = 1/2$ $C_i = 1$

How many **triangles** does node a participate in? $\tau(i)$

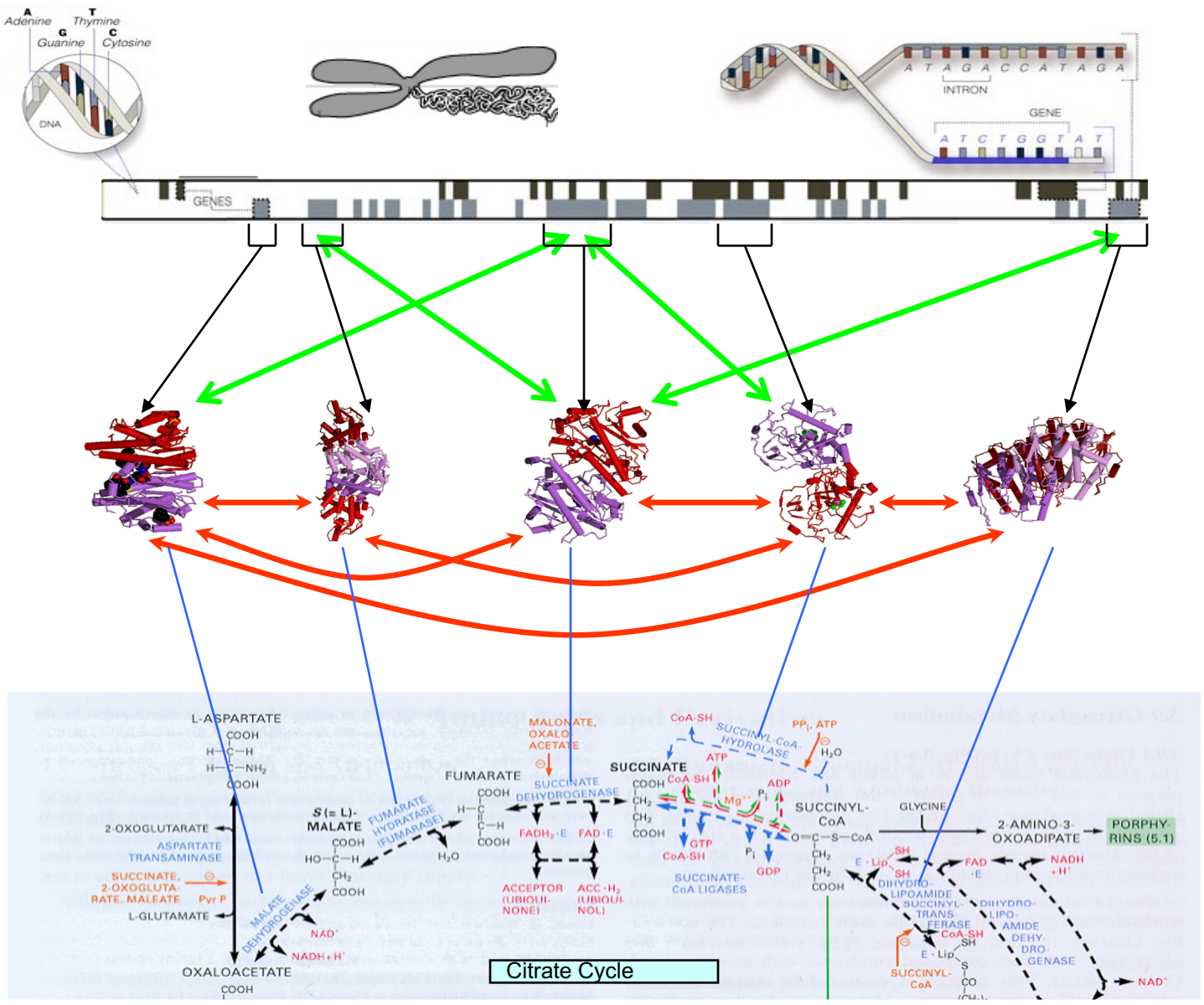
Clustering coefficient the fraction of possible triangles in which a node participates:

$$C_i = \frac{\tau(i)}{\tau_{max}(i)} \qquad C_i = \frac{2\tau(i)}{k_i(k_i - 1)}$$

Clustering



$< C > = .389$



GENOME

protein-gene interactions

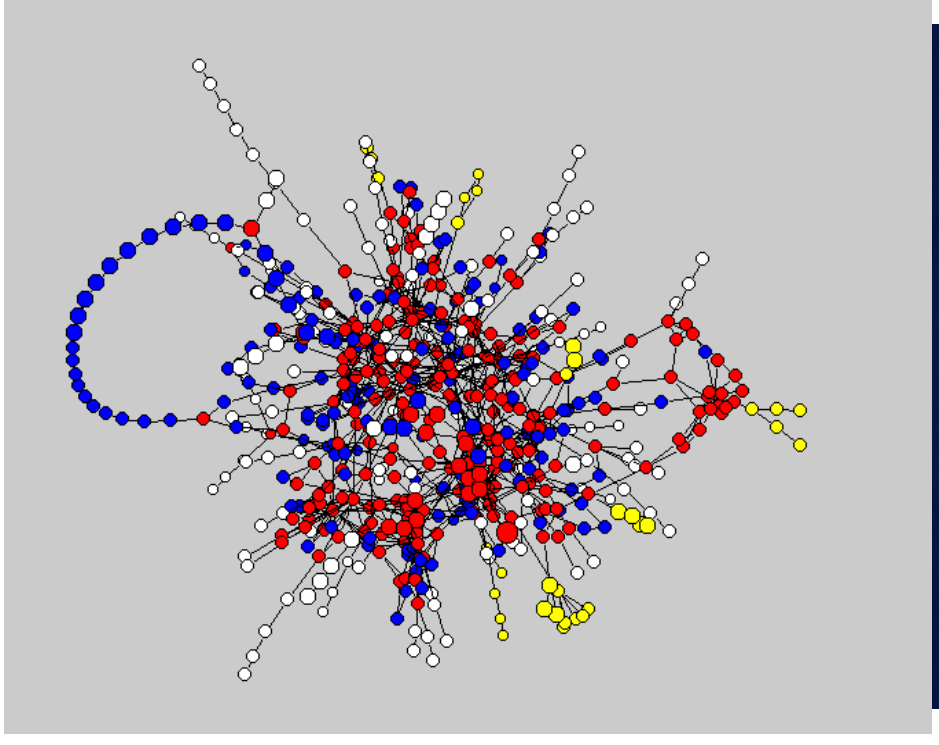
PROTEOME

protein-protein interactions

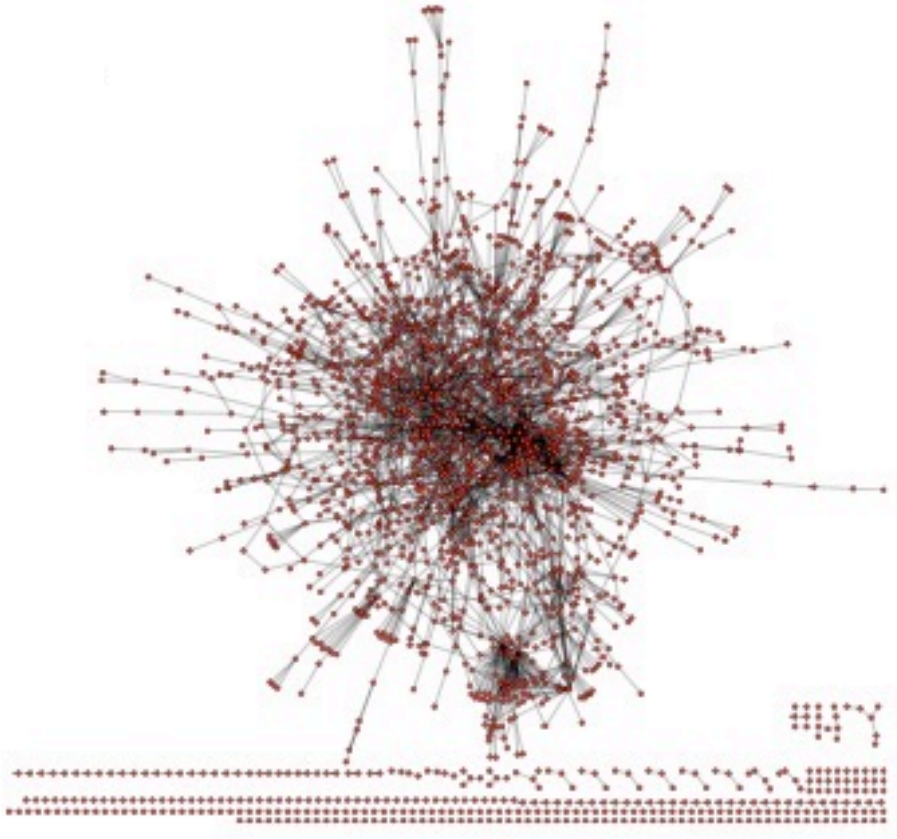
METABOLISM

Bio-chemical reactions

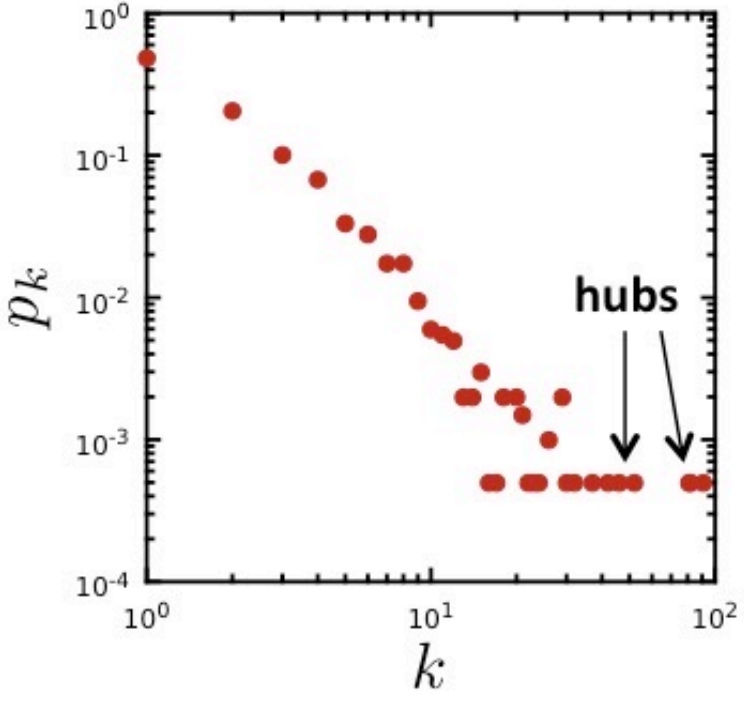
Metabolic Network



Protein Interactions



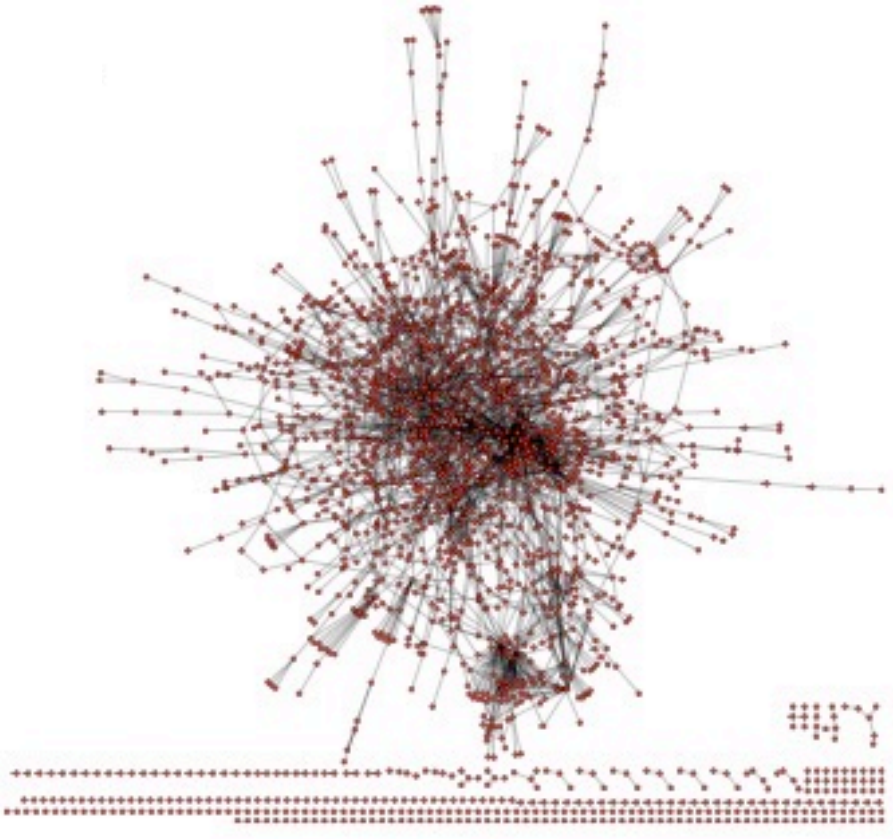
A few hubs connect to ~10% of the network



Not connected: 185 components

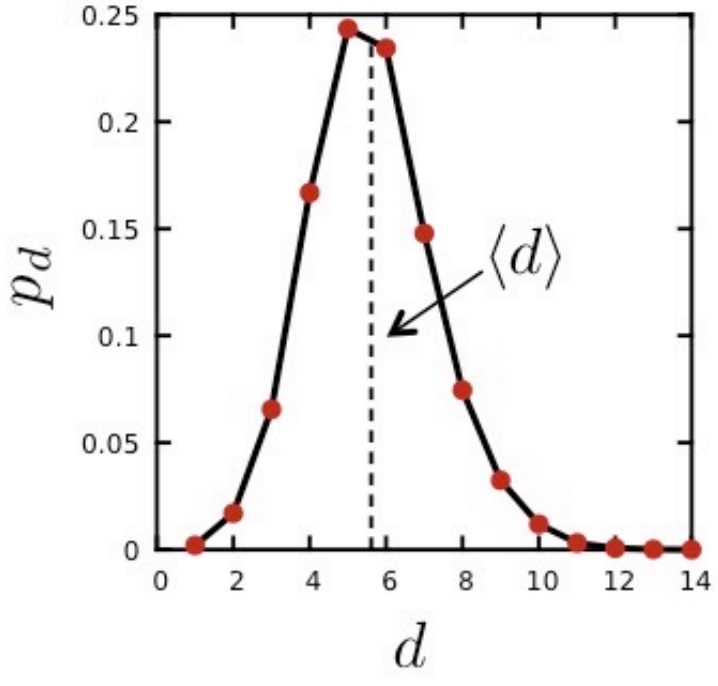
the largest (giant component)
1,647 nodes

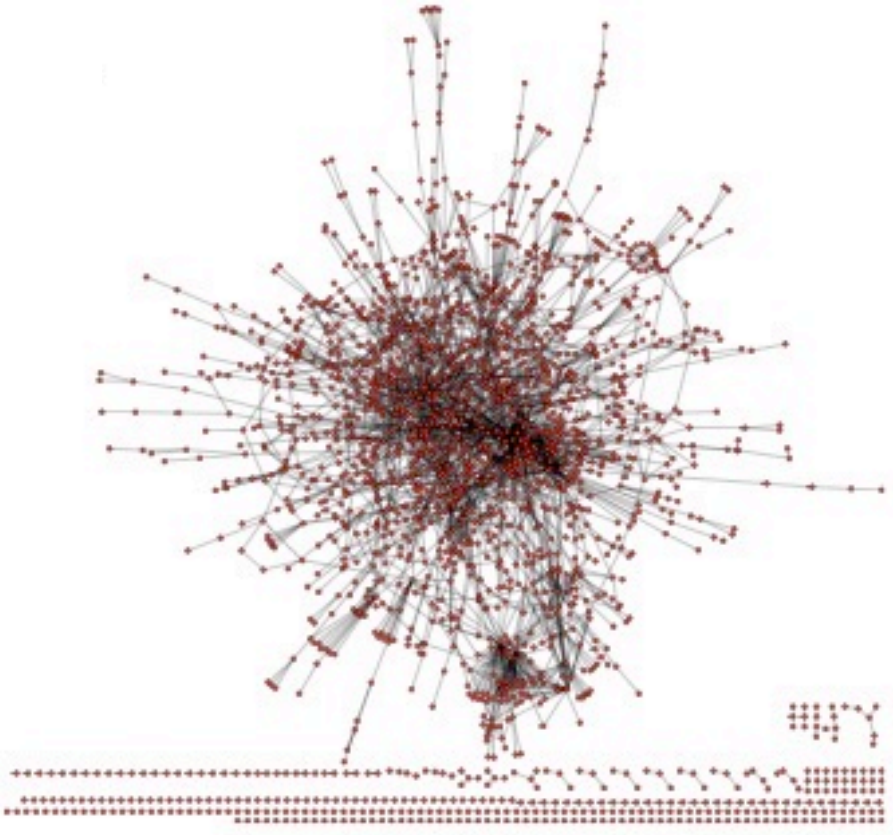
Most nodes have only 1-2 connections <k> = 2.9



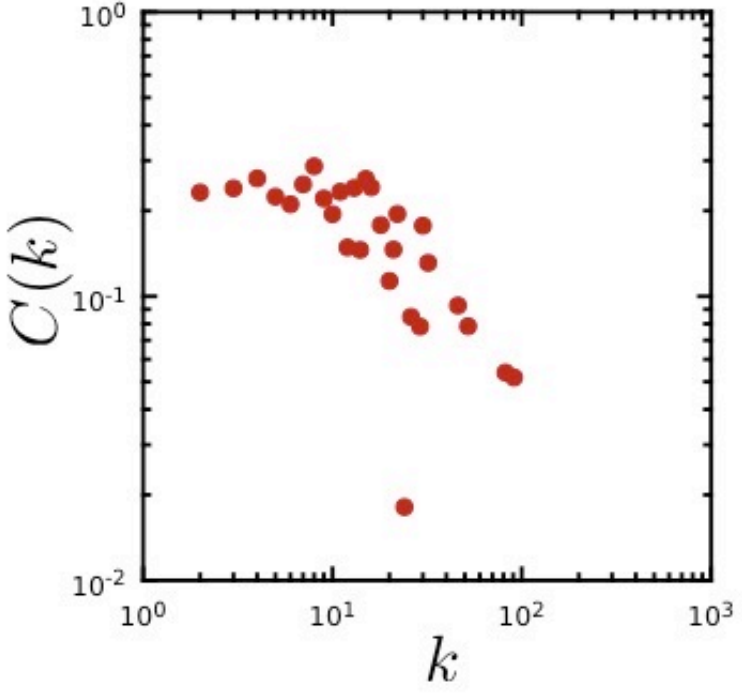
The network's diameter is 14

<d> = 5.6





<C> = 0.12

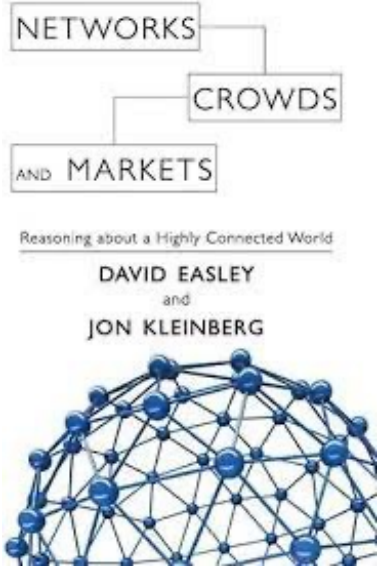


We can explore how the average clustering coefficient changes with node degree

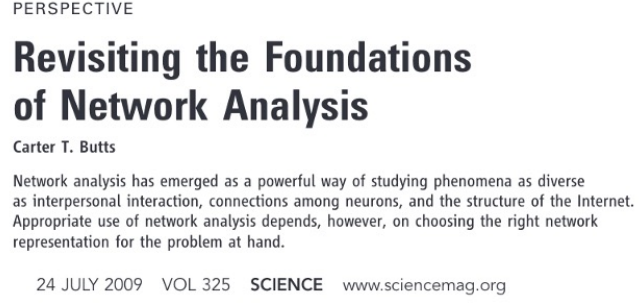
Weekly Class Readings

Readings

- Chapters 1 & 2 in Easley & Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world.* The full book is also here.
- Butts, Carter T. (2009) "Revisiting the foundations of network analysis." *Science* **325**, no. 5939: 414-416.



Chapters 1 & 2



PERSPECTIVE

Revisiting the Foundations of Network Analysis

Carter T. Butts

Network analysis has emerged as a powerful way of studying phenomena as diverse as interpersonal interaction, connections among neurons, and the structure of the Internet. Appropriate use of network analysis depends, however, on choosing the right network representation for the problem at hand.

24 JULY 2009 VOL 325 **SCIENCE** www.sciencemag.org

https://uvads8104.github.io/content/

Weekly Class Readings

Next week (1/31 & 2/2):

Readings

- Milgram, Stanley. (1967) " The small world problem." *Psychology today* **2**, no. 1: 60-67.
- Watts, Duncan J., and Steven H. Strogatz. (1998) " Collective dynamics of 'small world' networks." *Nature* **393**, 6684: 440-442.
- Fosdick, Bailey K., Daniel B. Larremore, Joel Nishimura, and Johan Ugander. (2018) " Configuring random graph models with fixed degree sequences." *Siam Review* **60**, no. 2: 315-355.
- Milo, Ron, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. (2002) " Network motifs: simple building blocks of complex networks." *Science* **298**, no. 5594: 824-827.

https://uvads8104.github.io/content/

Discussion moderation

1. Brief summary
   i. Less than <5 min only!
   ii. Slides are OK, but not necessary
   iii. Main Thesis? / Which figure shows this?

2. Contextualize
   i. What class ideas does the article connect with / draw upon?

3. Questions for discussion
   i. Do you agree with the findings?
   ii. Are all results consistent with hypothesis?
   iii. Anything not clear / vague?
   iv. Any problems? Were all methods correctly used?
   v. What would you do differently?
   vi. Be provocative! (but be respectful!)