Network science

Genoveva Vargas-Solar
http://www.vargas-solar.com/big-data-analytics
French Council of Scientific Research, LIG & LAFMIA Labs

Montevideo, 22nd November – 4th December, 2015

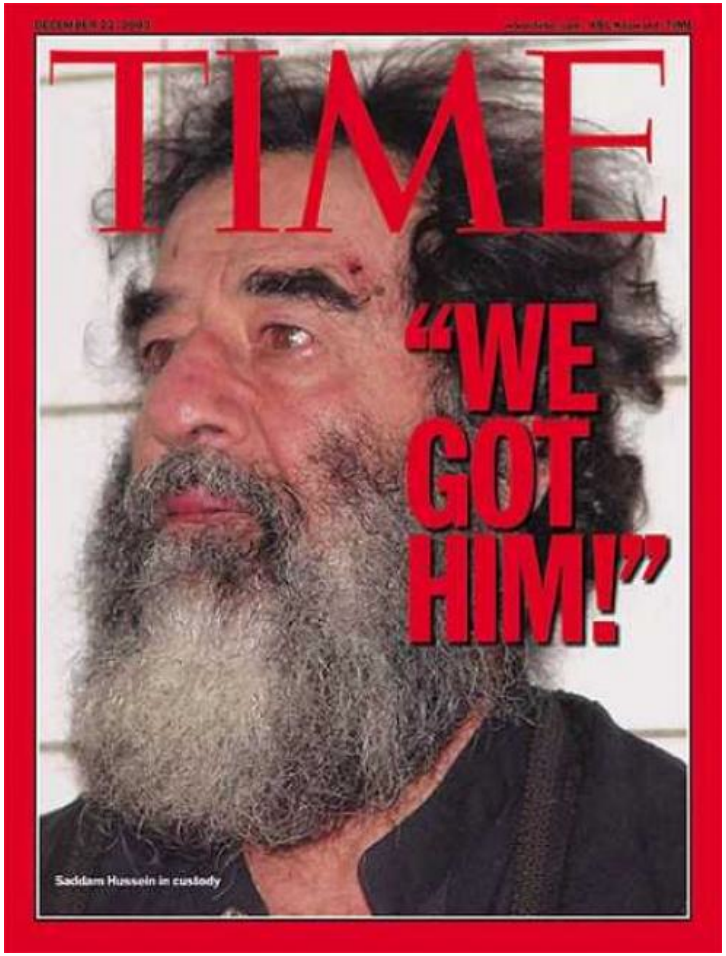
The study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena

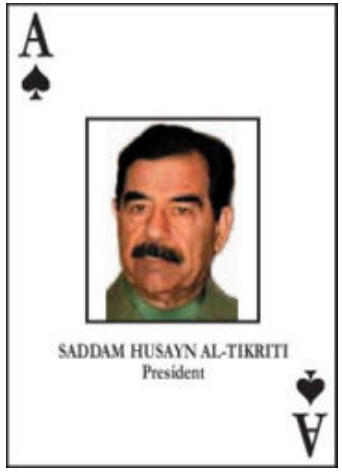
Network science fields

The field draws on theories and methods including

- g**raph theory** from **mathematics**
- statistical mechanics from **physics**
- data mining and information visualization from **computer scien**ce
- inferential modelling from **statistics**
- **social** structure from sociology

A SIMPLE STORY (1) The fate of Saddam and network science



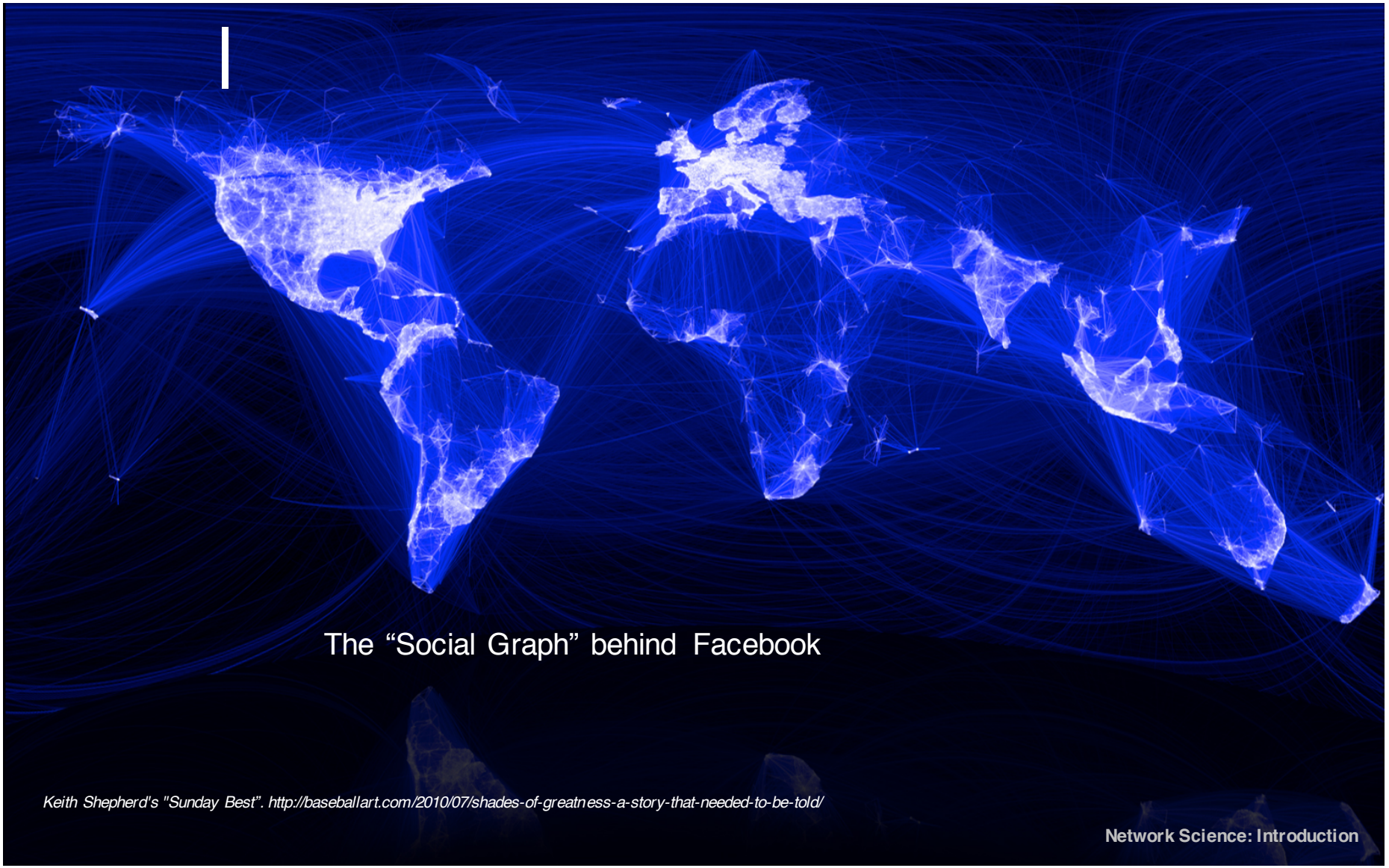


The faith of Saddamm Hussein and network science

The capture of Saddam Hussein:

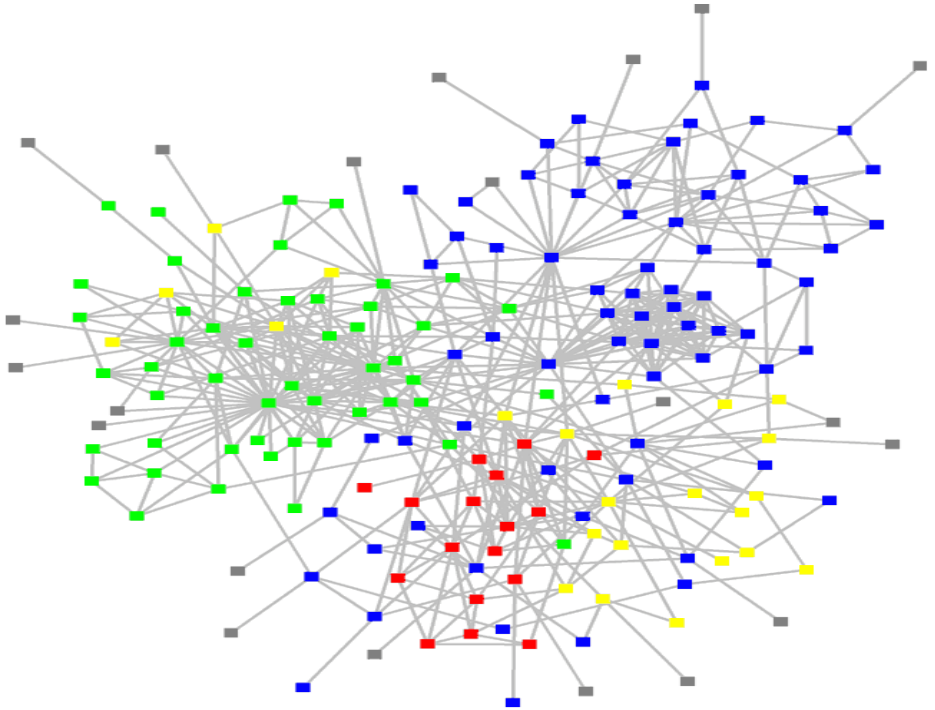
- Shows the strong **predictive power** of networks
- Underlies the need to obtain **accurate maps of the net**works; and the often heroic difficulties we encounter during the mapping process.
- demonstrates the **remarkable stability of these networks**:
  - the capture of Hussein was not based on fresh intelligence, but rather on his pre-invasion social links, unearthed from old photos stacked in his family album.
- Shows that the **choice of network** we focus on makes a huge difference:
  - the hierarchical tree, that captured the official organization of the Iraqi government, was of no use when it came to Saddam Hussein's whereabouts

Behind each complex system there is a network, that defines the interactions between the component



Keith Shepherd's "Sunday Best". http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

STRUCTURE OF AN ORGANIZATION

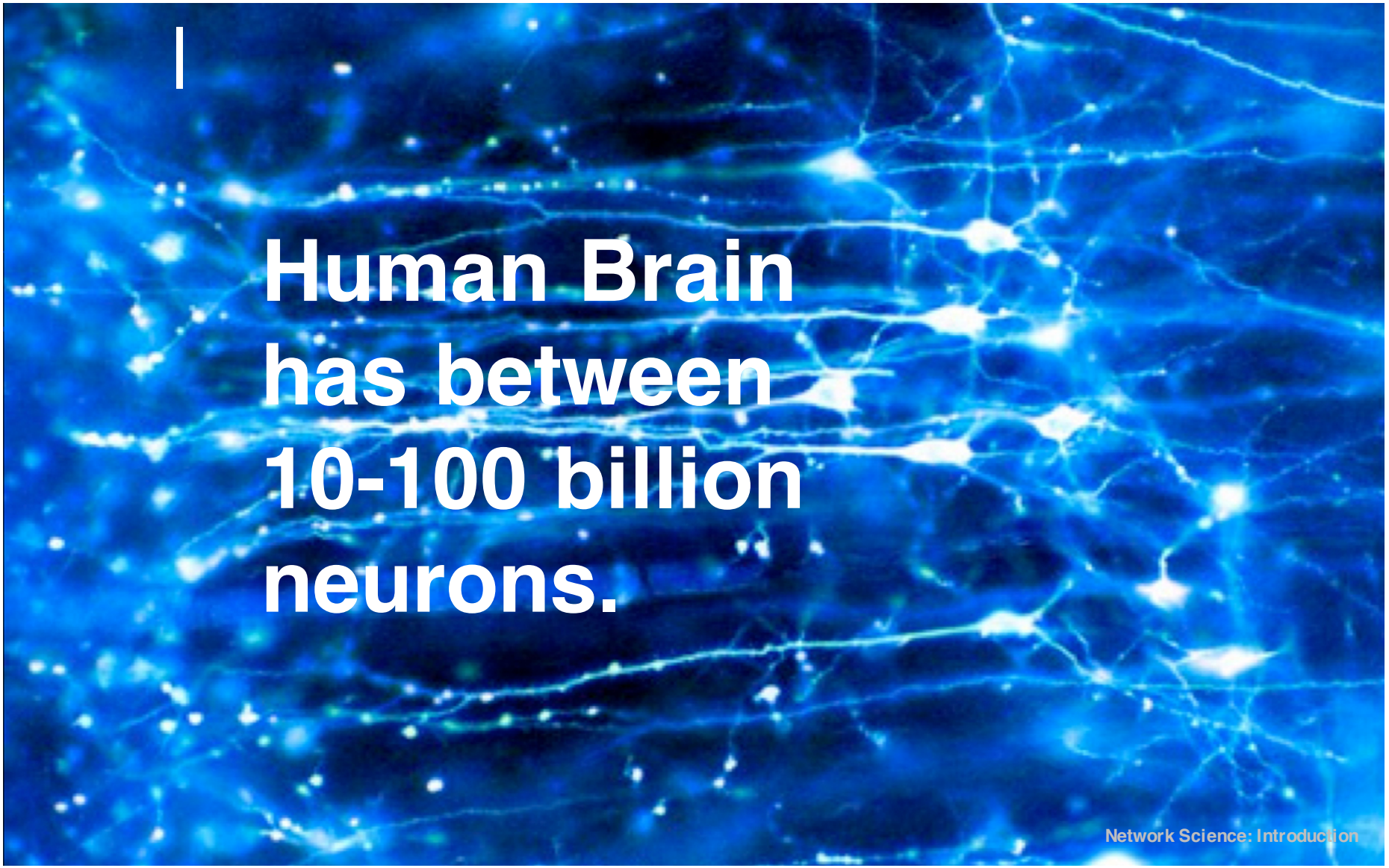


www.orgnet.com

: departments

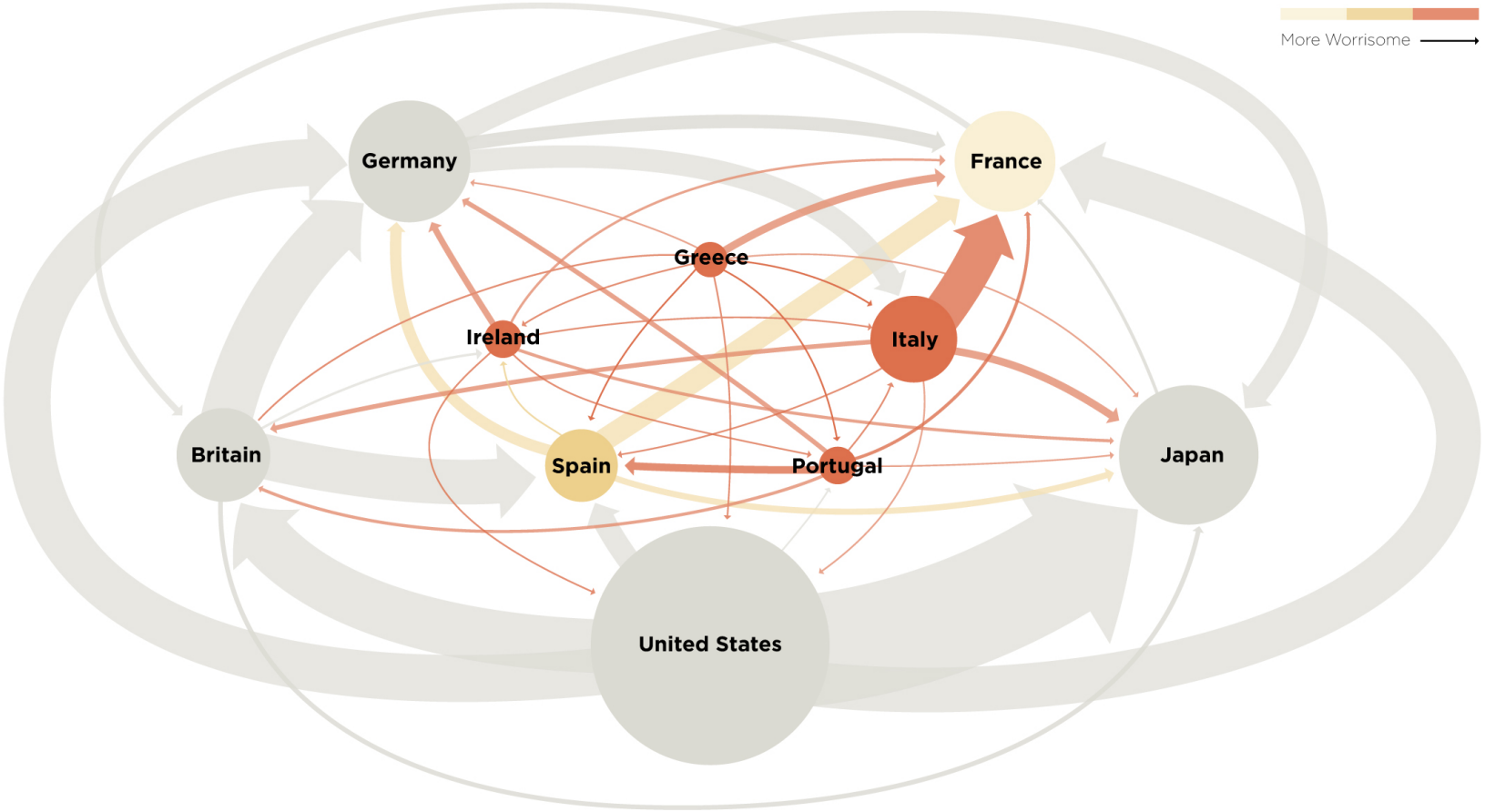
: consultants

: external experts


Human Brain
has between
10-100 billion
neurons.

The subtle financial networks

The not so subtle financial networks: 2011



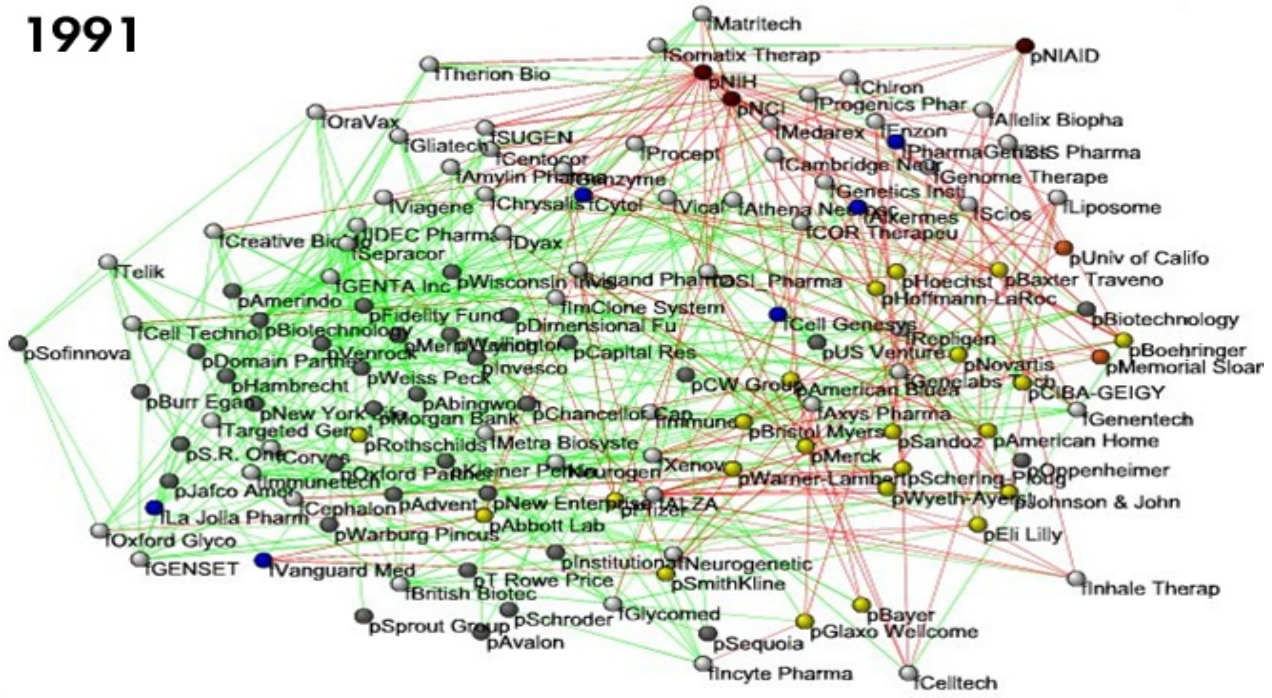
BUSINESS TIES IN US BIOTECH-INDUSTRY

Nodes:

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

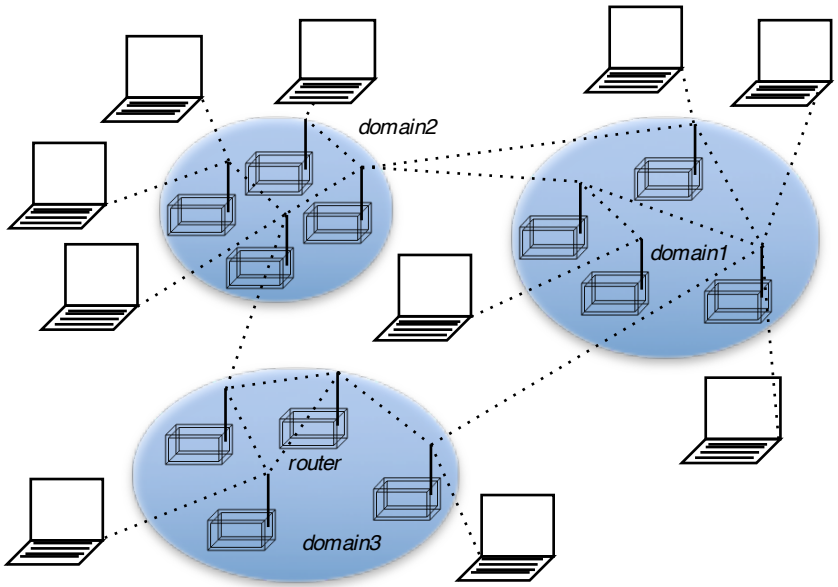
Links:

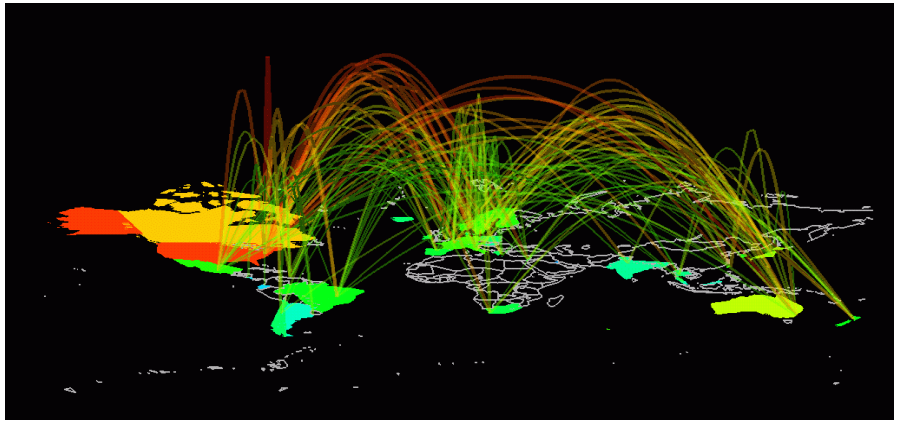
- Collaborations
- Financial
- R&D

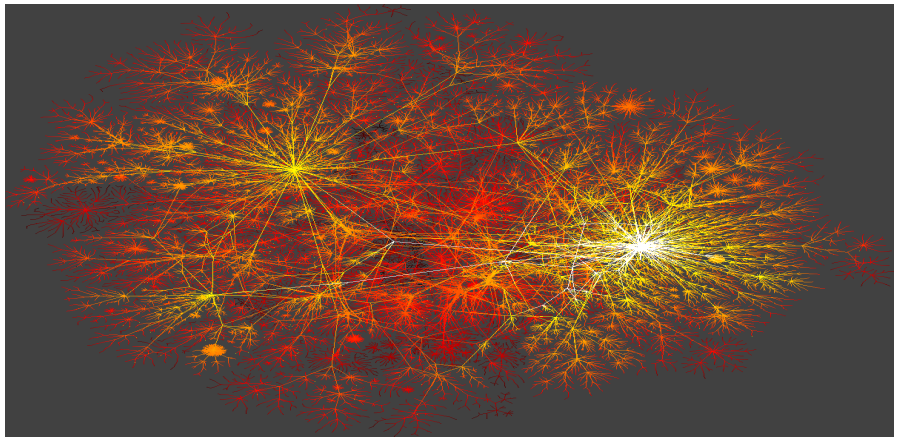


http://ecclectic.ss.uci.edu/~drwhite/Movie

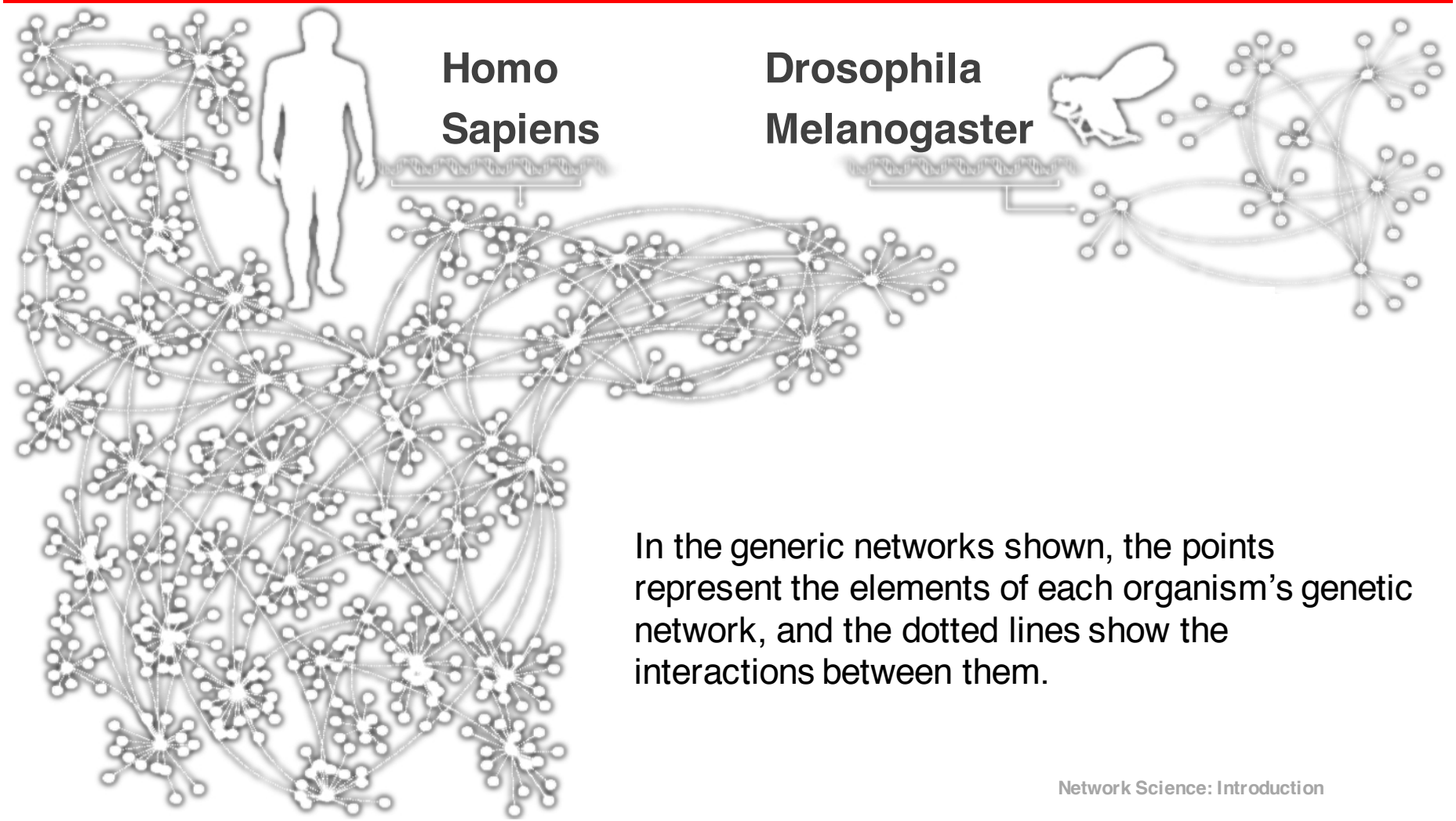
INTERNET





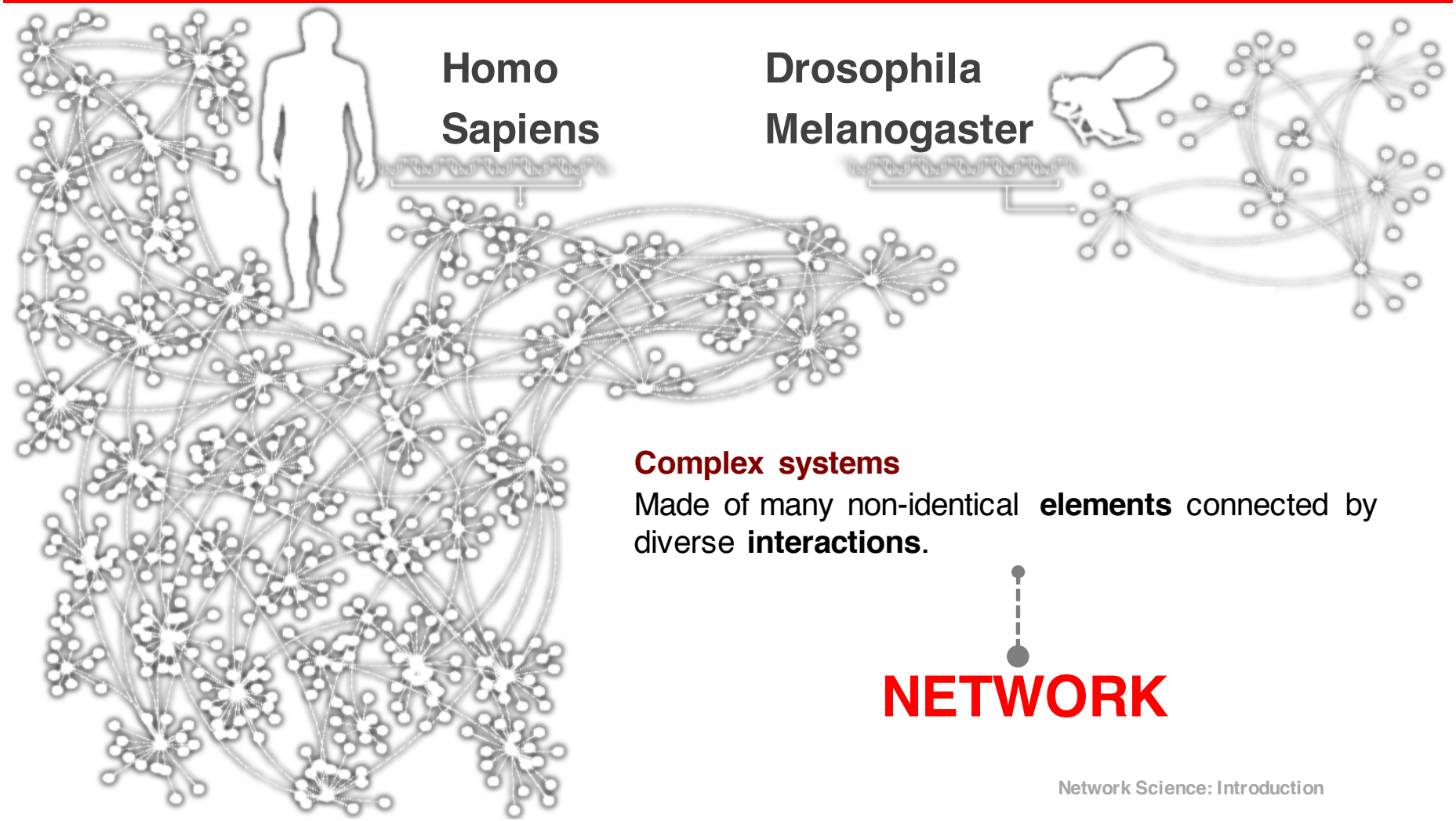


HUMANS GENES



In the generic networks shown, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them.

HUMANS GENES



Complex systems

Made of many non-identical **elements** connected by diverse **interactions**.

NETWORK

THE ROLE OF NETWORKS

Behind each system studied in complexity there is an intricate wiring diagram, or a **network**, that defines the interactions between the component.

We will never understand complex system unless we map out and understand the networks behind them.

Section 5

TWO FORCES HELPED THE EMERGENCE OF NETWORK SCIENCE

THE HISTORY OF NETWORK ANALYSIS

Graph theory: 1735, Euler

Social Network Research: 1930s, Moreno

Communication networks/internet: 1960s

Ecological Networks: May, 1979.

Section 6

THE CHARACTERISTICS OF NETWORK SCIENCE

THE CHARACTERISTICS OF NETWORK SCIENCE

Interdisciplinary

Empirical

Quantitative and Mathematical

Computational

THE CHARACTERISTICS OF NETWORK SCIENCE

Interdisciplinary

Empirical, data driven

Quantitative and Mathematical

Computational

THE CHARACTERISTICS OF NETWORK SCIENCE

Interdisciplinary

Empirical

Quantitative and Mathematical

Computational

THE CHARACTERISTICS OF NETWORK SCIENCE

Interdisciplinary

Empirical

Quantitative and Mathematical

Computational

Section 7

THE IMPACT OF NETWORK SCIENCE

ECONOMIC IMPACT

Google

Market Cap(2010 Jan 1):
$189 billion

Cisco Systems

networking gear Market cap (Jan 1, 2919):
$112 billion

Facebook

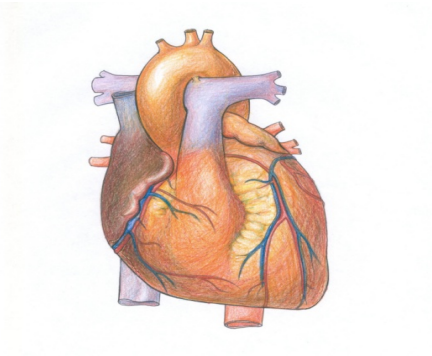
market cap:
$50 billion

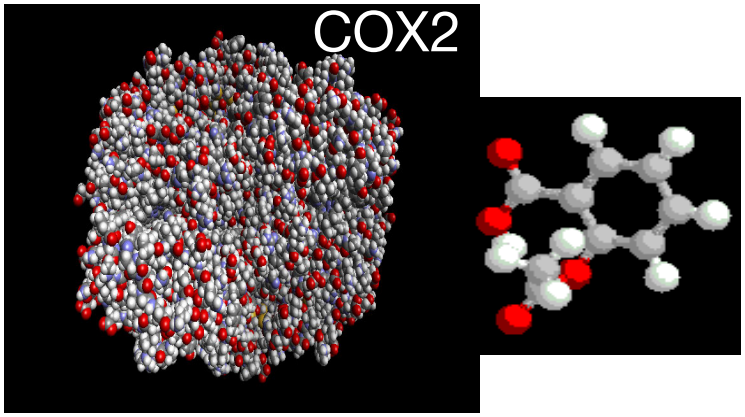
www.bizjournals.com/austin/news/2010/11/15/facebooks... - Cached

DRUG DESIGN, METABOLIC ENGINEERING:

Reduces
Inflammation
Fever
Pain

Prevents
Heart attack
Stroke



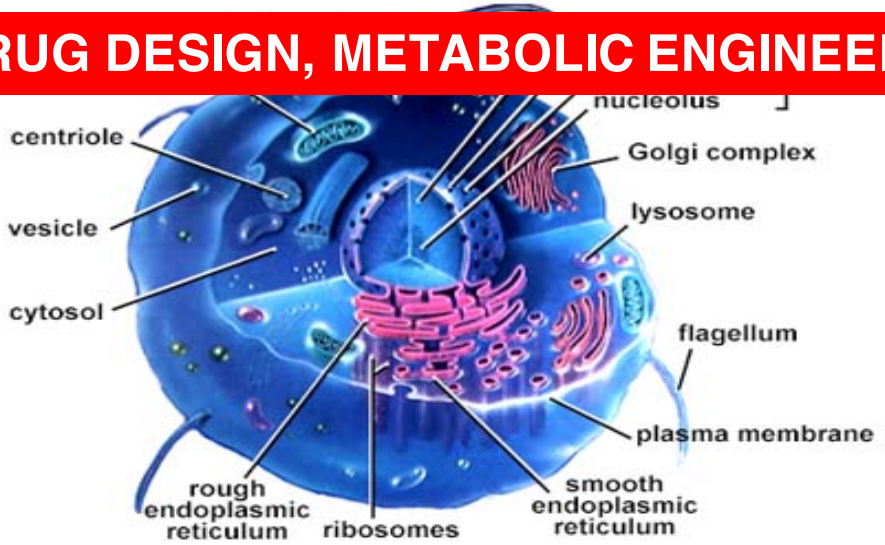


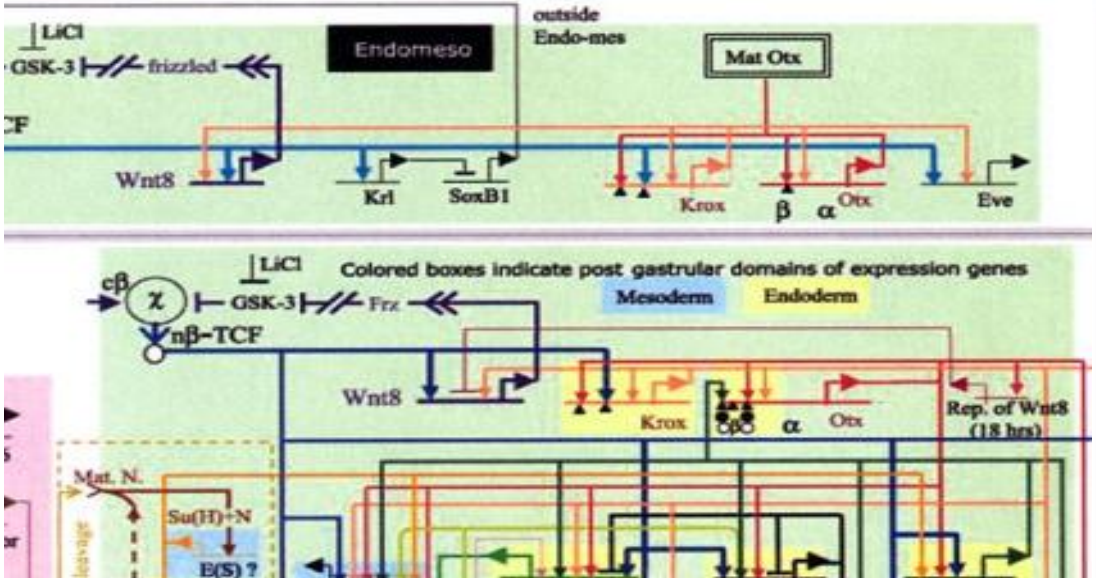
Reduces the risk of
Alzheimer's Disease

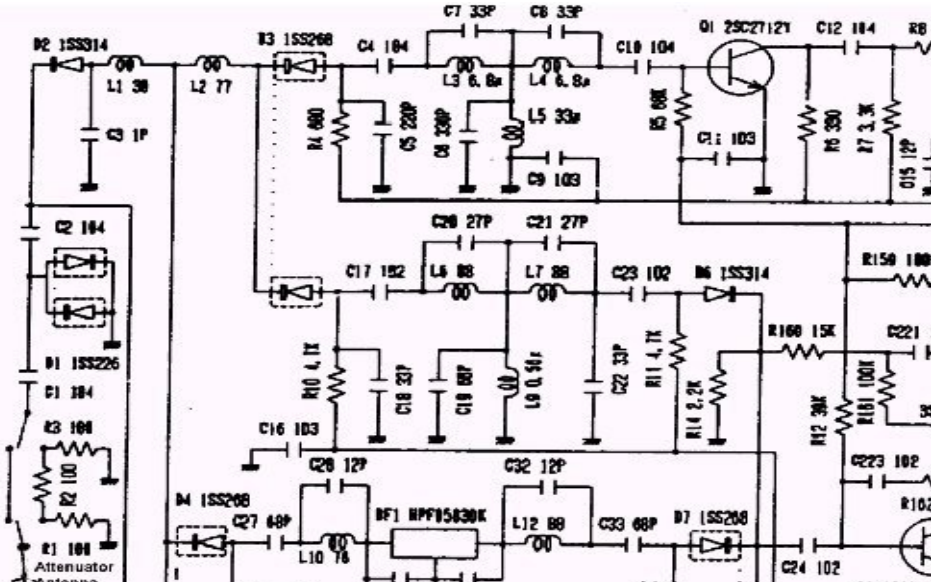
Reduces the risk of
breast cancer
ovarian cancers
colorectal cancer

Causes
Bleeding
Ulcer

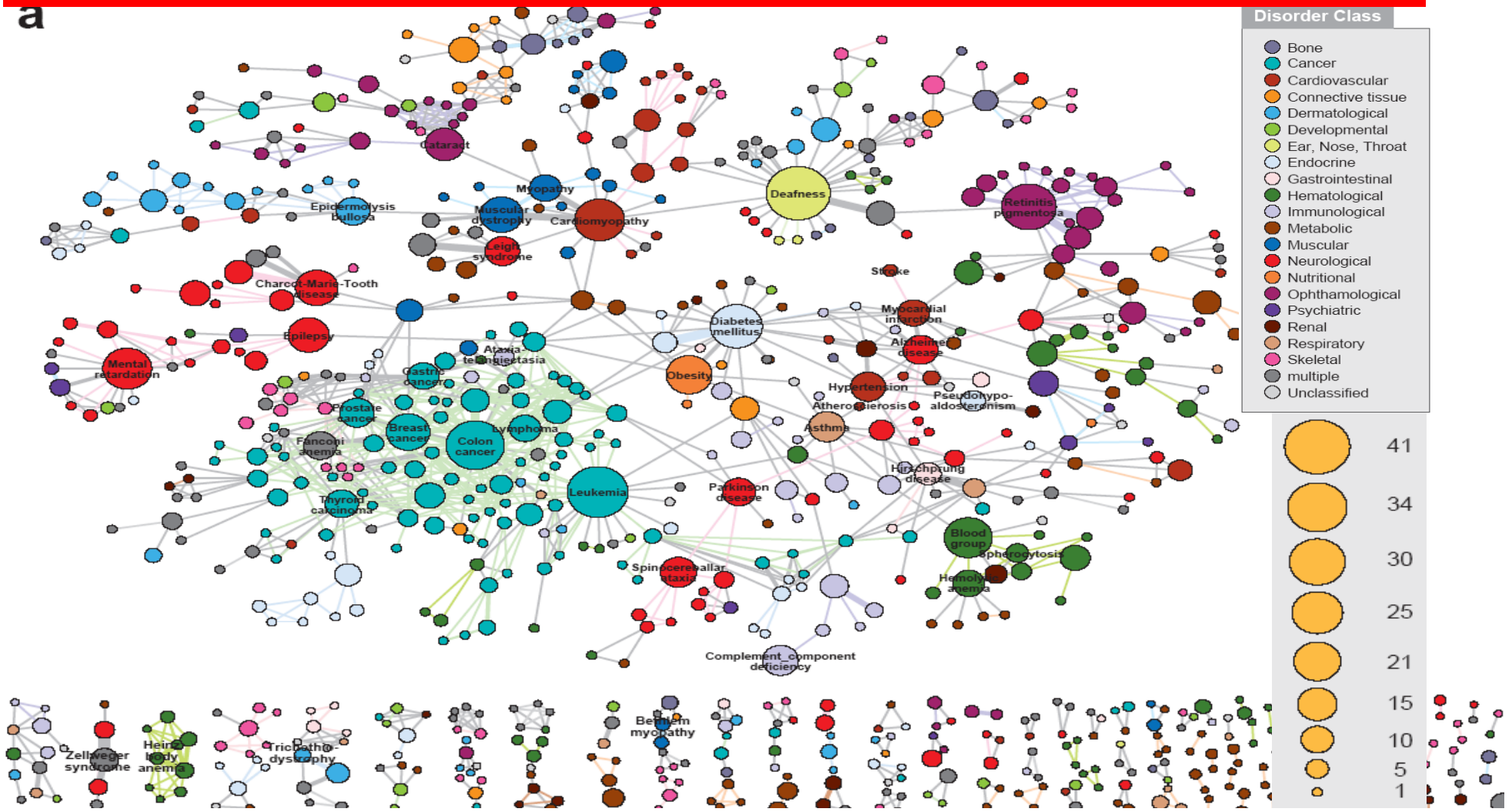
DRUG DESIGN, METABOLIC ENGINEERING:





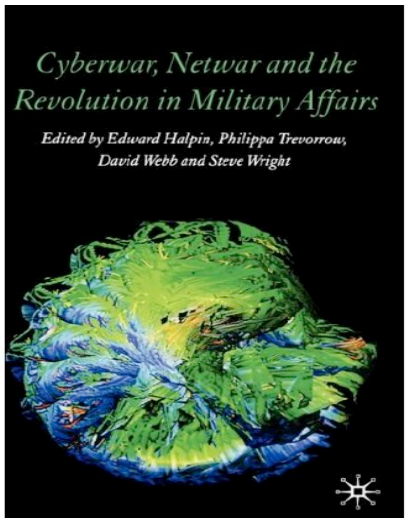


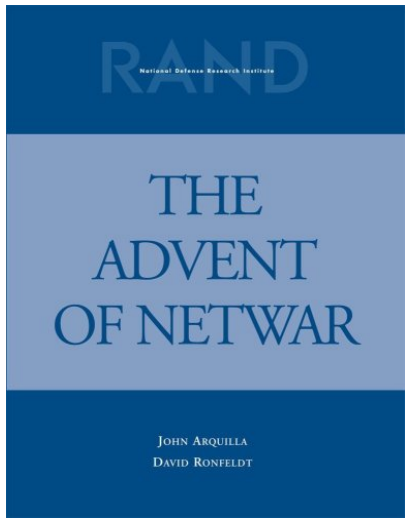
HUMAN DISEASE NETWORK



Network Biology/Network Medicine

FIGHTING TERRORISM AND MILITARY

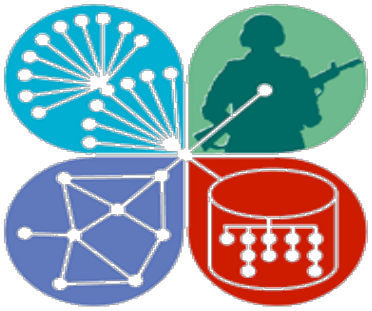




http://www.slate.com/id/2245232

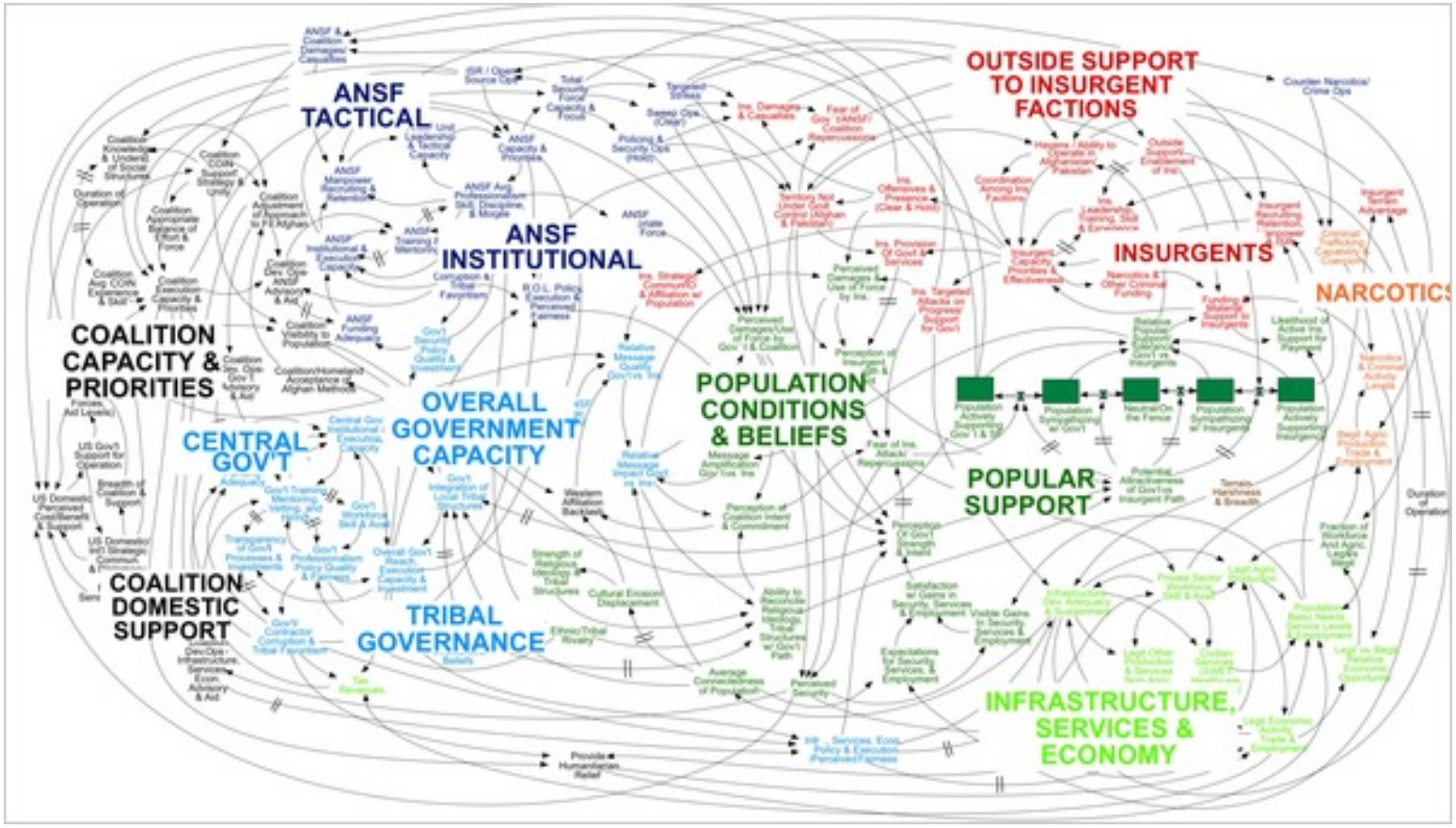
FIGHTING TERRORISM AND MILITARY

Network Science Center
West Point

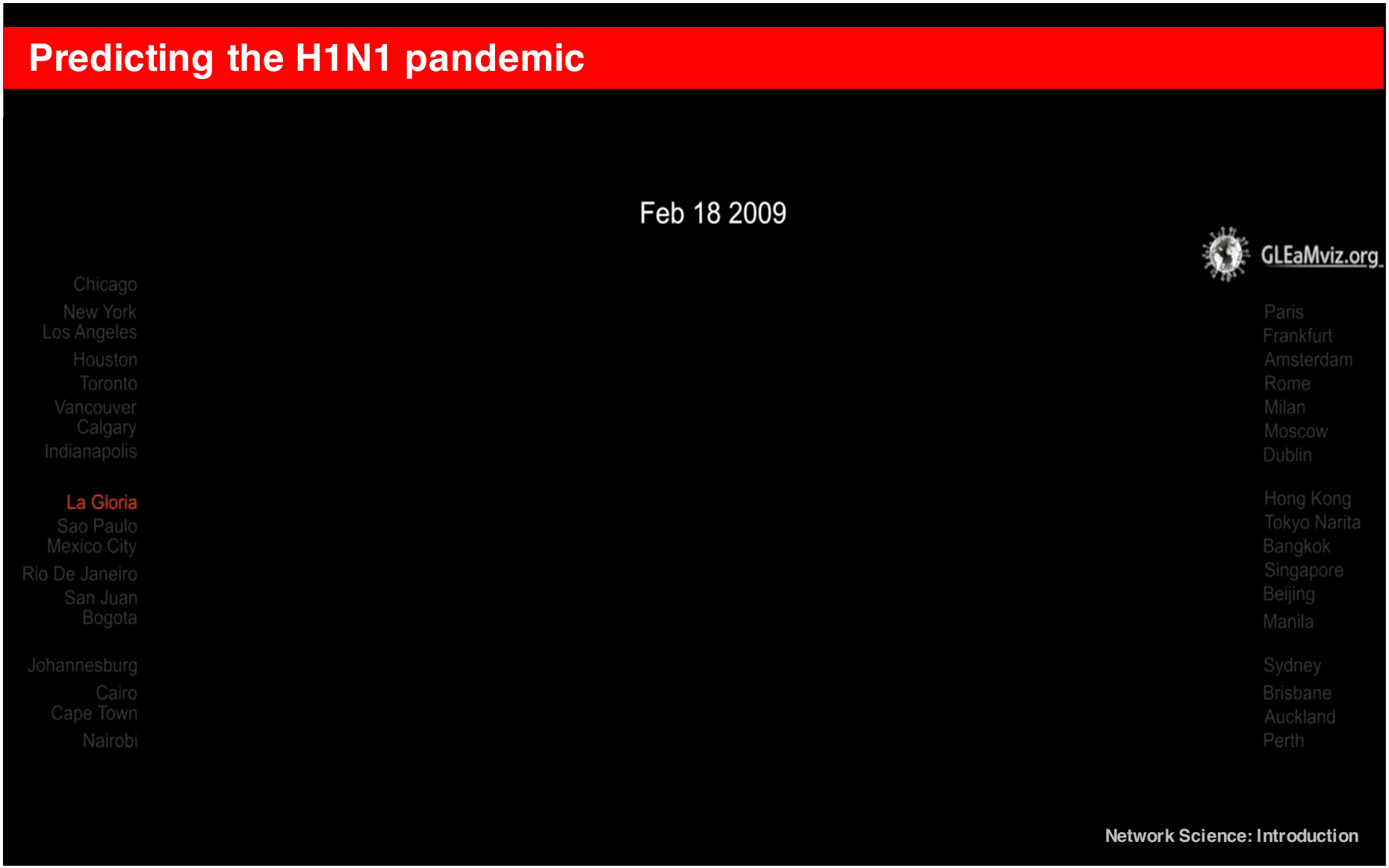


http://www.ns-cta.org/ns-cta-blog/

The network behind a military engagement

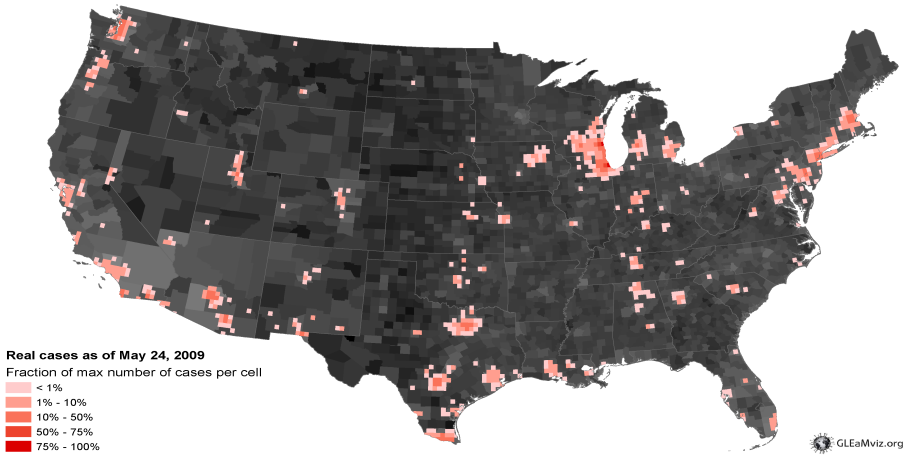


Predicting the H1N1 pandemic

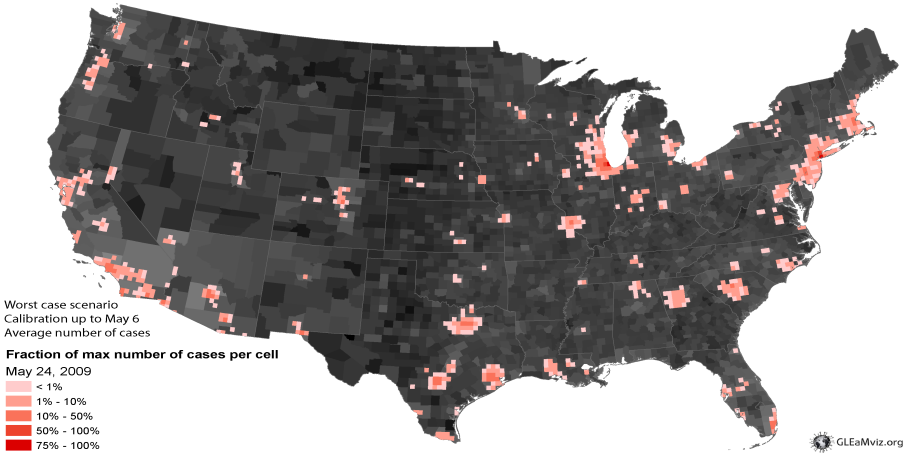


EPIDEMIC FORECAST | Predicting the H1N1 pandemic

Real

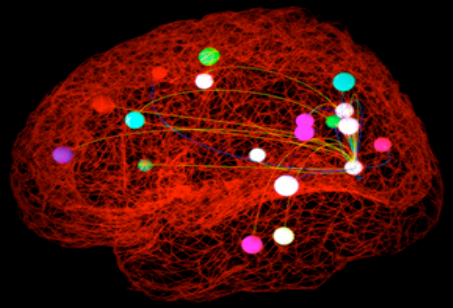


Projected

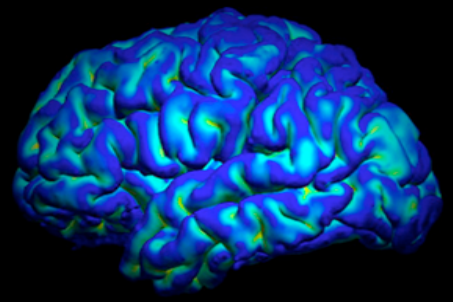


BRAIN RESEARCH

In September 2010 the National Institutes of Health awarded $40 million to researchers at Harvard, Washington University in St. Louis, the University of Minnesota and UCLA, to develop the technologies that could systematically map out brain circuits.

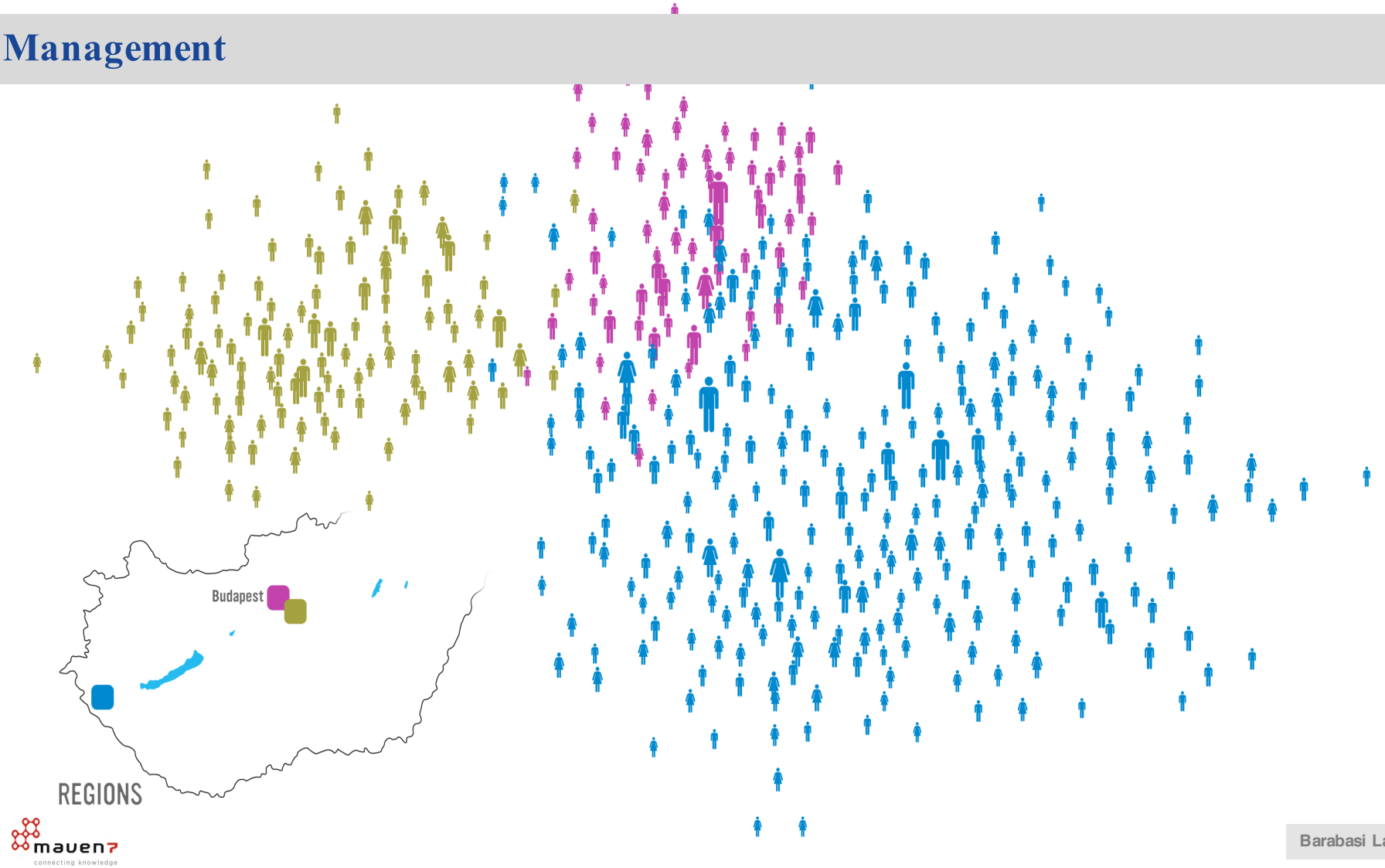


The Human Connectome Project (HCP) with the ambitious goal to construct a map of the complete structural and functional neural connections in vivo within and across individuals.



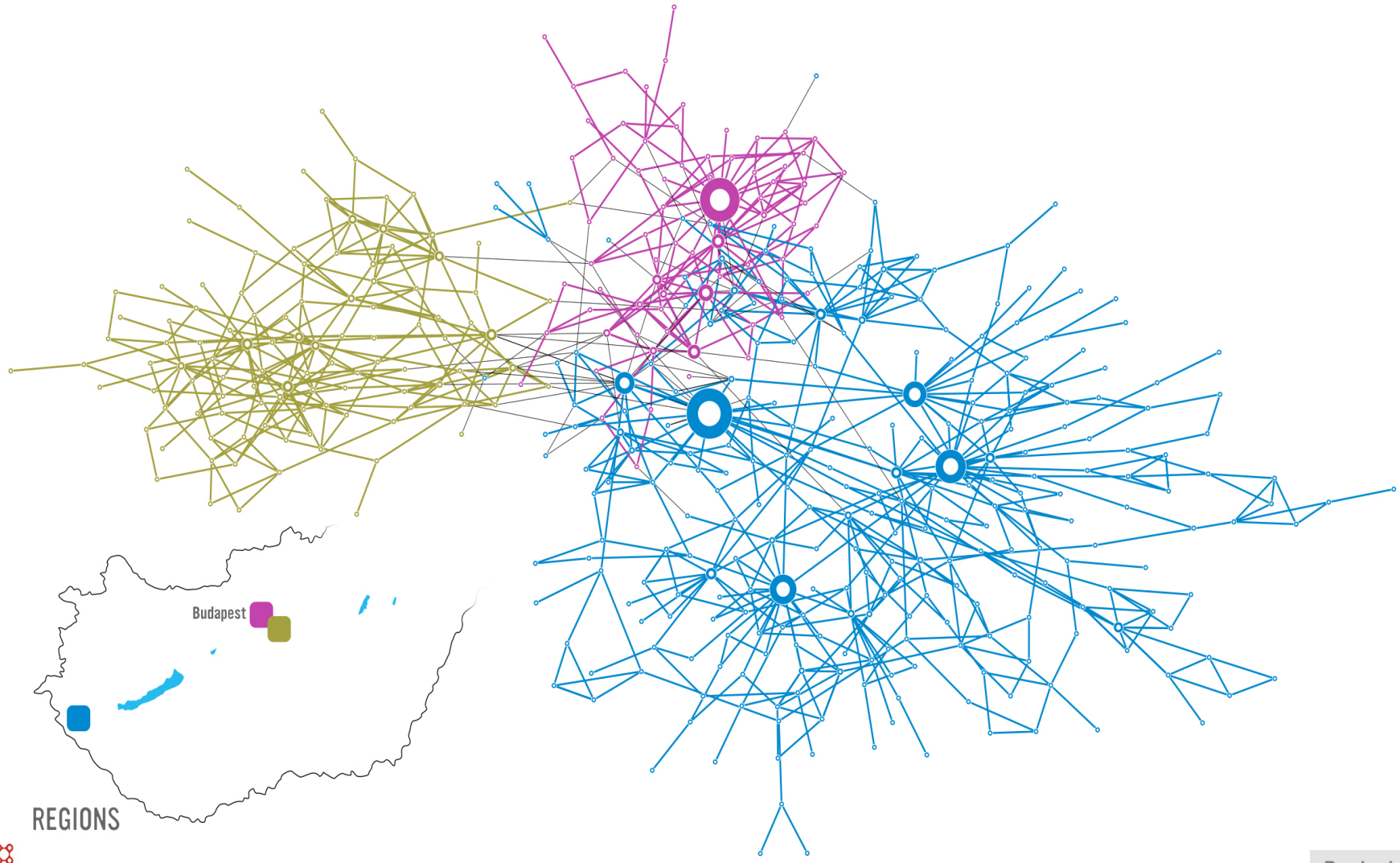
http://www.humanconnectomeproject.org/overview/

Management



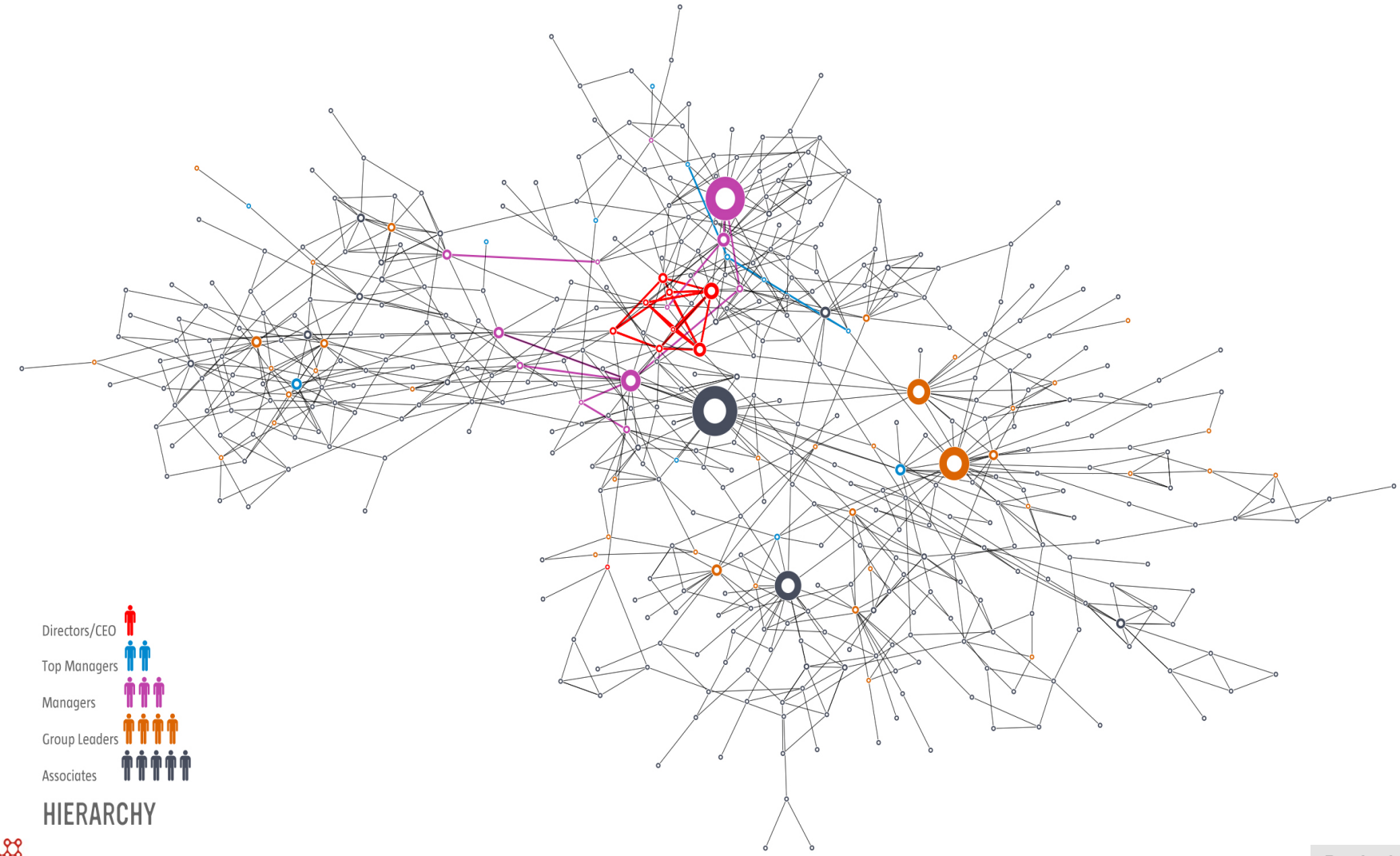
mavenz
connecting knowledge

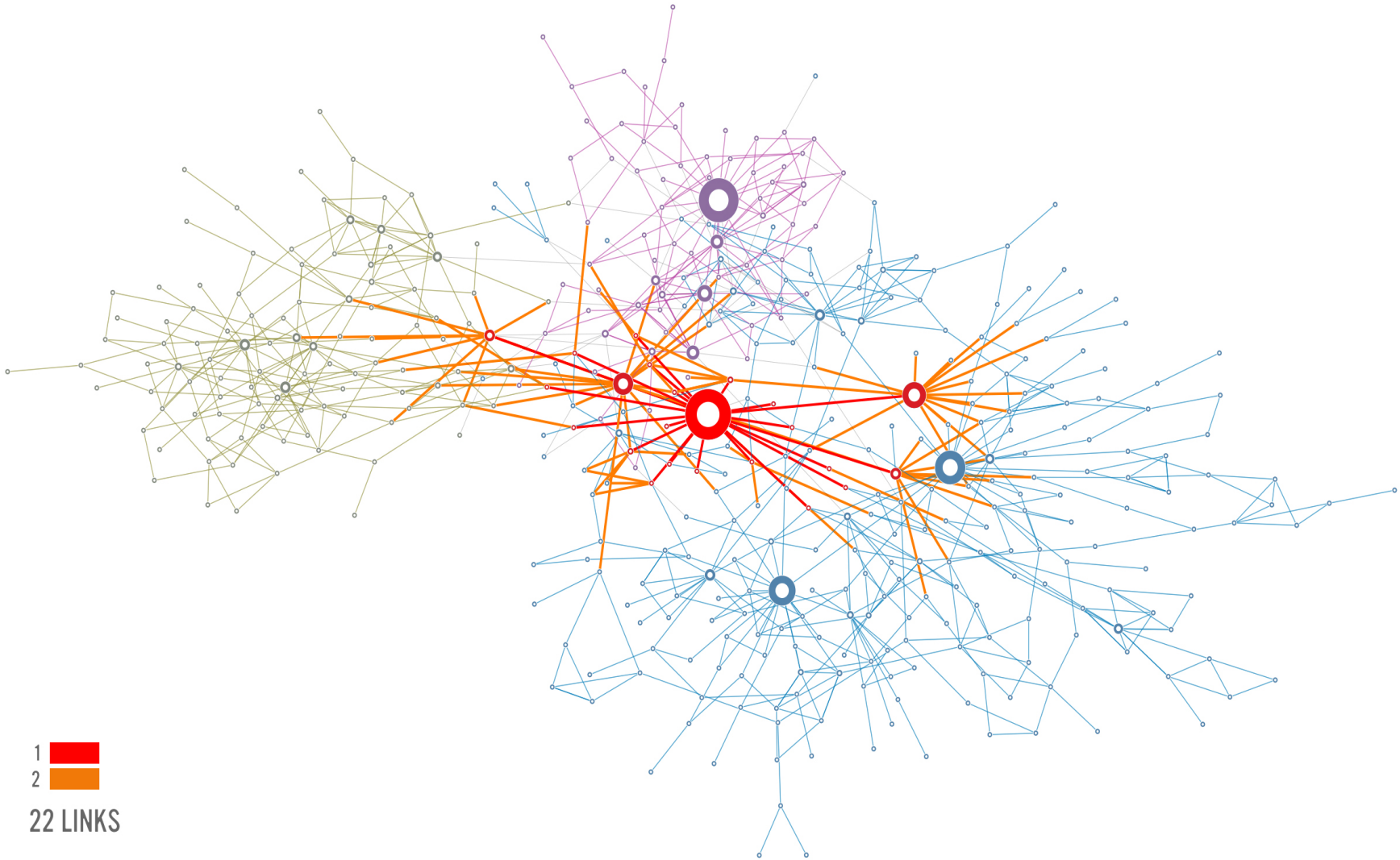
Barabasi Lab



mauen7
connecting knowledge

Barabasi Lab


Directors/CEO
Top Managers
Managers
Group Leaders
Associates
HIERARCHY
maven7
connecting knowledge
Barabasi Lab


1
2
22 LINKS

MOST IMPORTANT | Networks Really Matter

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

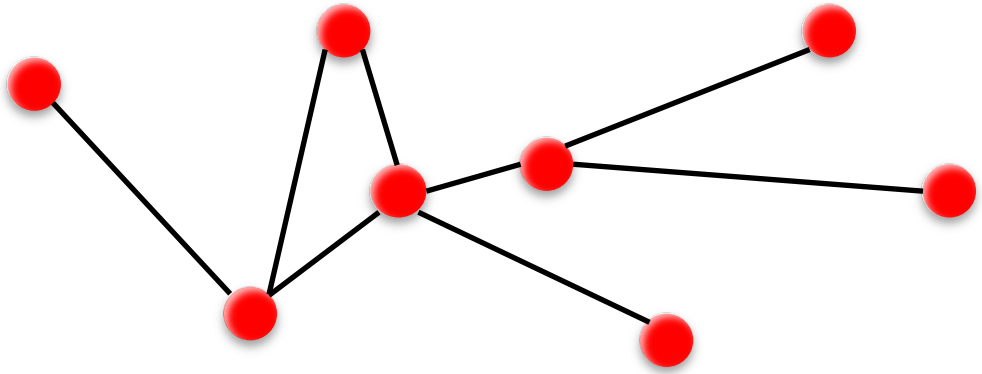
If you were to understand the WWW structure, searchability, etc, **hopeless without invoking the Web's topology.**

If you want to understand human diseases, **it is hopeless without considering the wiring diagram of the cell**.

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)

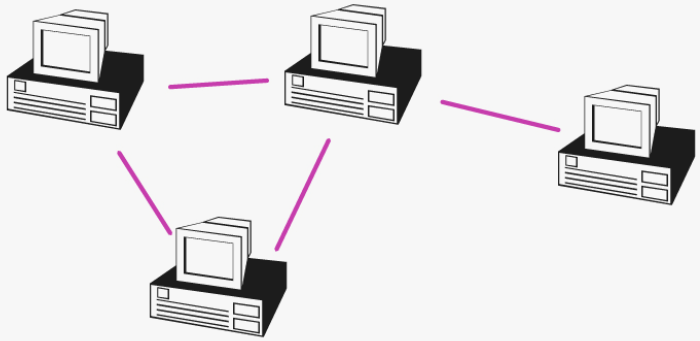
graph: mathematical representation of a network

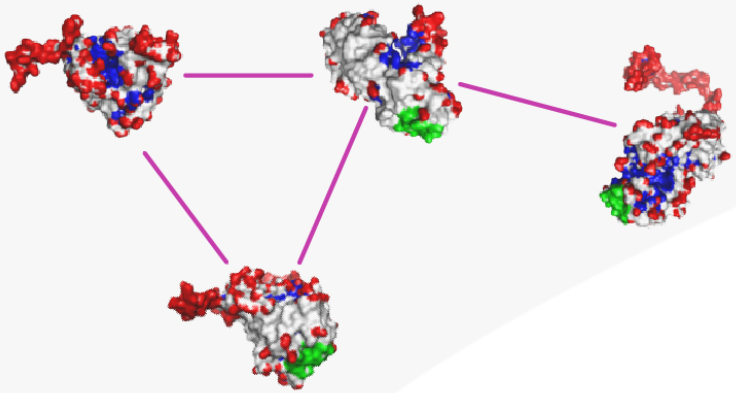
- web graph,
- social graph (a Facebook term)

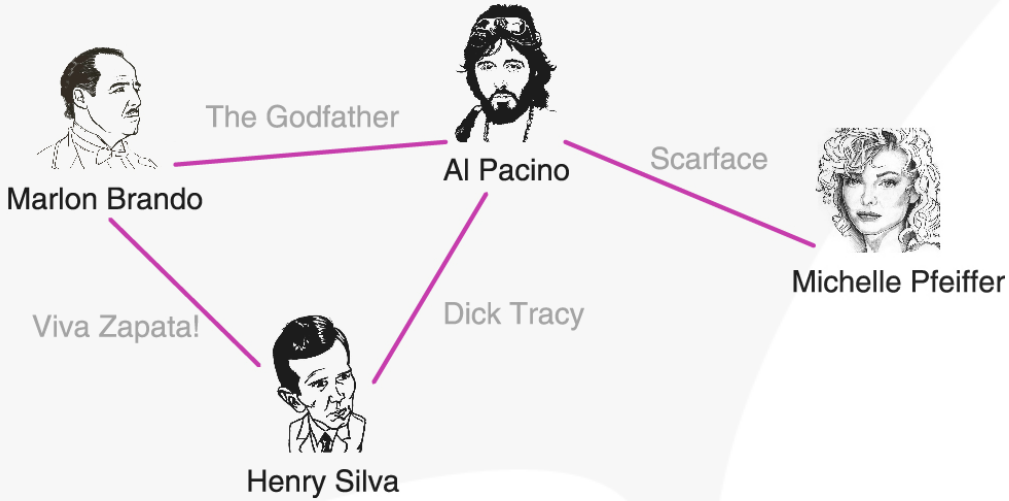
Language: (Graph, vertex, edge)

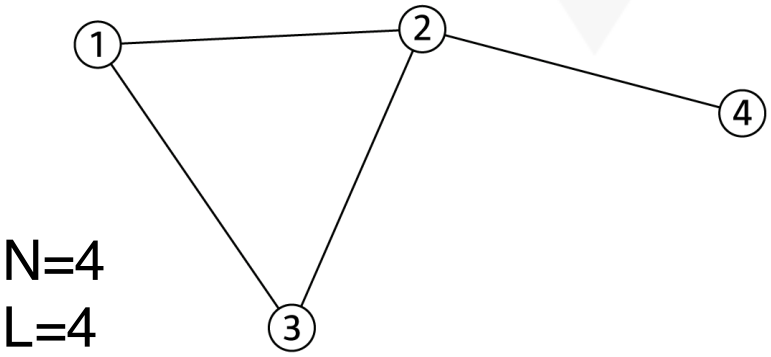
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably.

A COMMON LANGUAGE









N=4
L=4

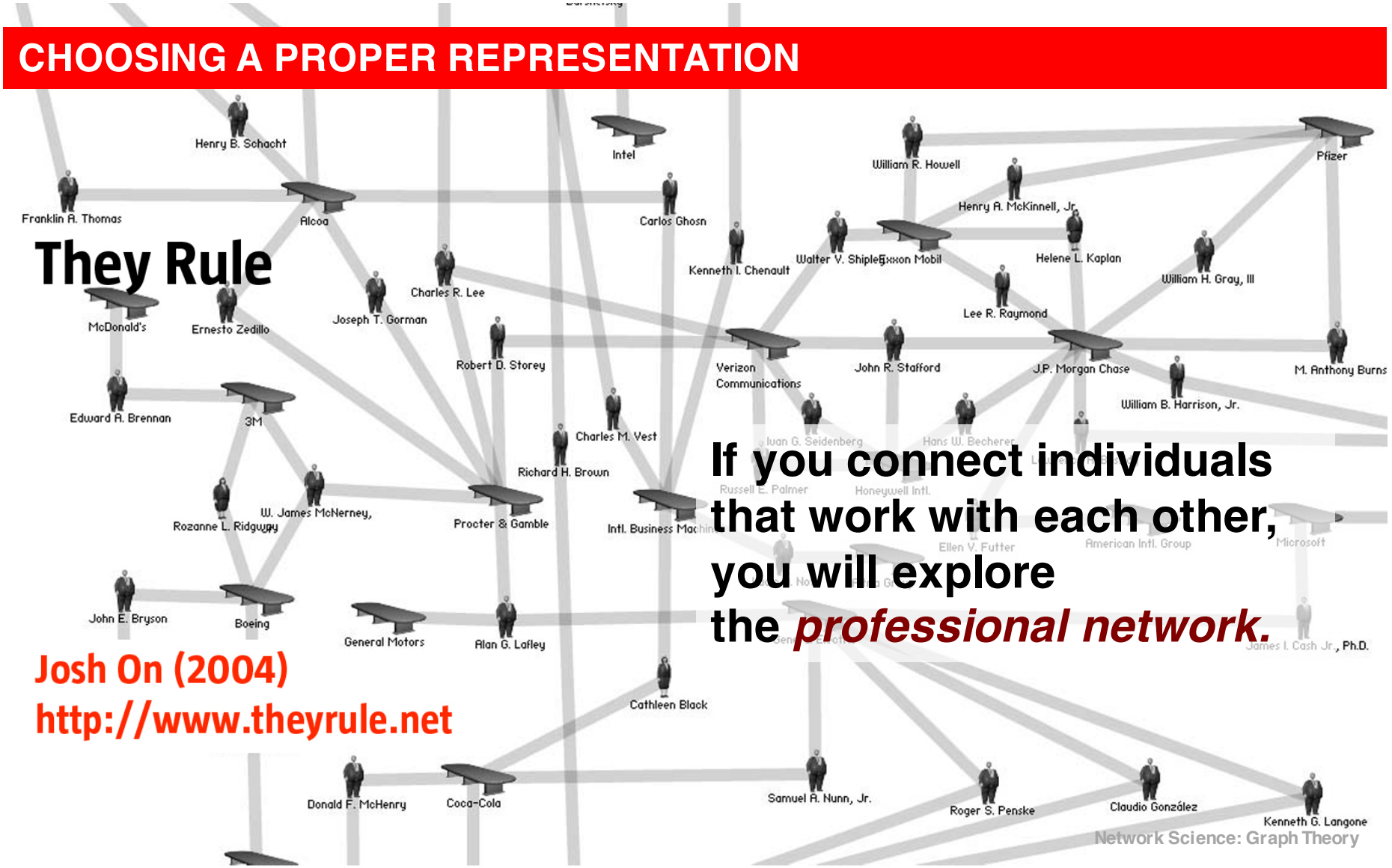
CHOOSING A PROPER REPRESENTATION

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

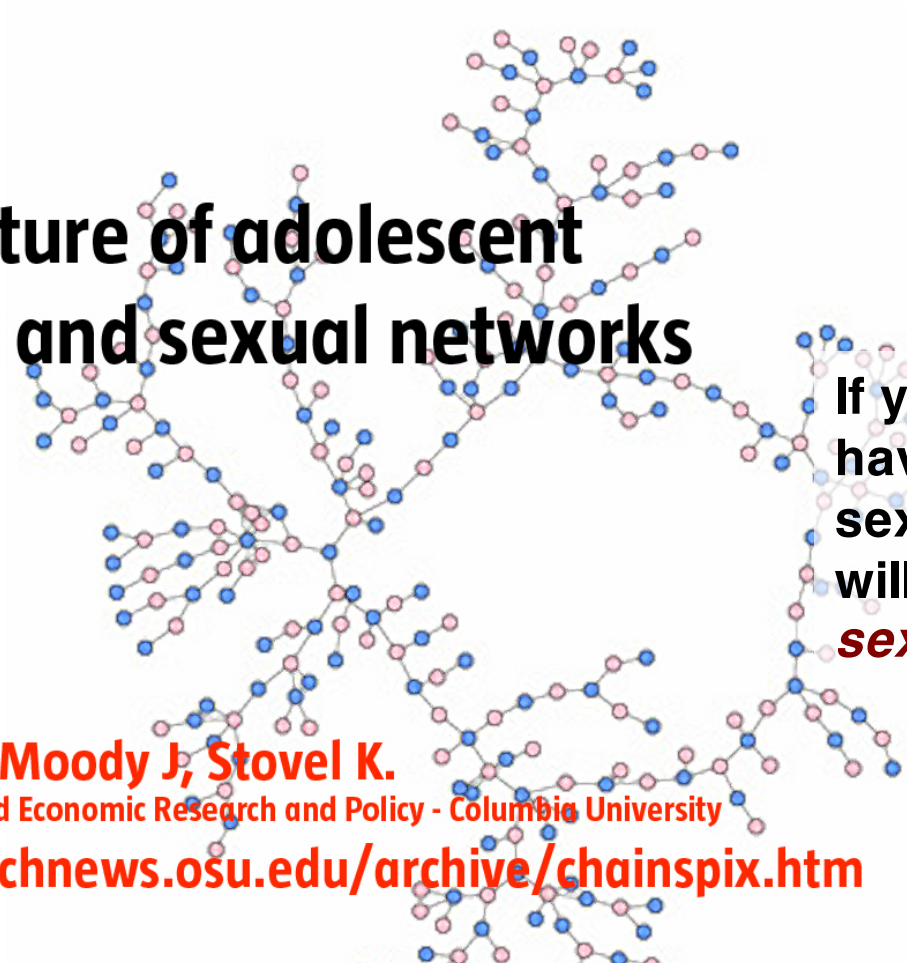
In some cases there is a unique, unambiguous representation.
In other cases, the representation is by no means unique.

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



The structure of adolescent romantic and sexual networks

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

Bearman PS, Moody J, Stovel K.
Institute for Social and Economic Research and Policy - Columbia University
http://researchnews.osu.edu/archive/chainspix.htm

CHOOSING A PROPER REPRESENTATION

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

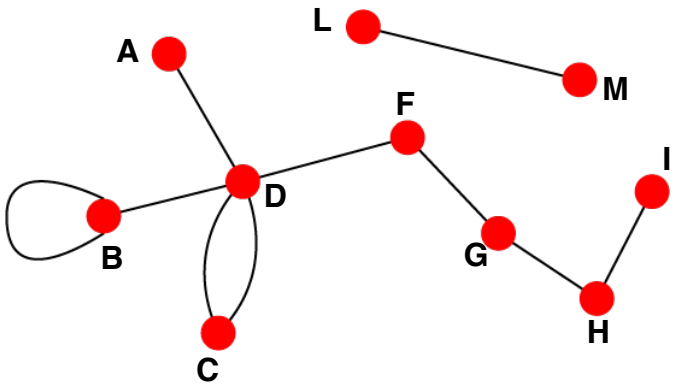
It is a network, nevertheless.

UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:

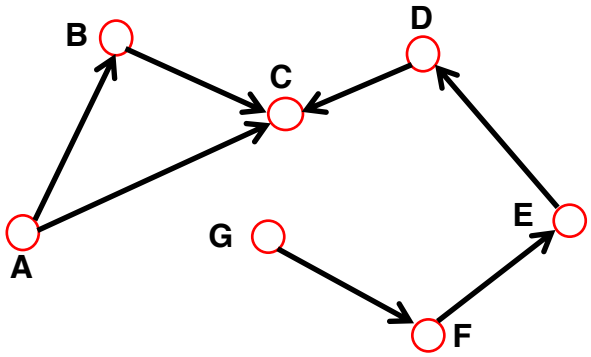


Undirected links :
coauthorship links
Actor network
protein interactions

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

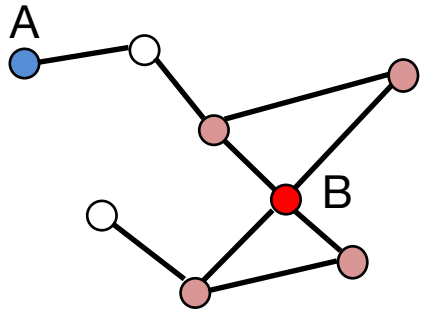
Reference Networks

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

Degree, Average Degree and Degree Distribution

NODE DEGREES

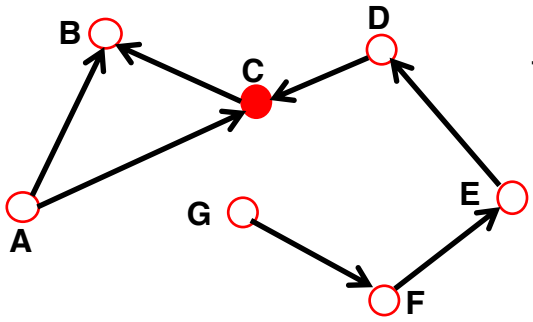
Undirected



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:

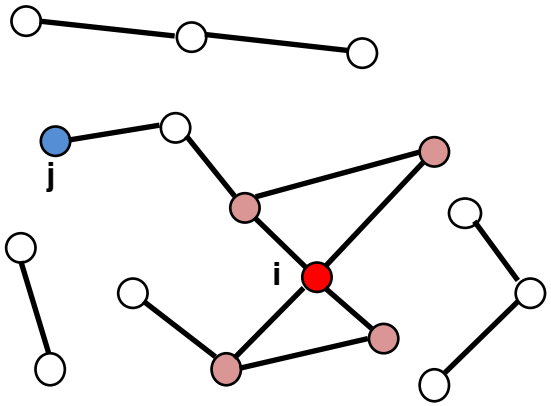
$$p_x = \frac{1}{N} \sum_i \delta_{x,x_i}$$

where p_x follows

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

AVERAGE DEGREE

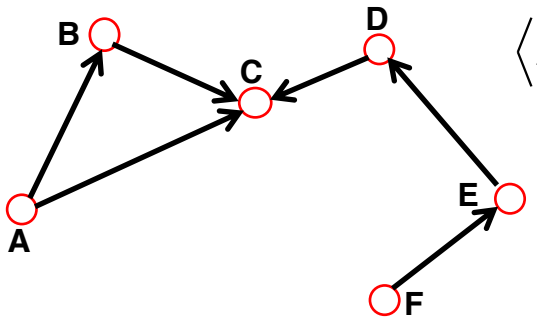
Undirected



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

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	6.33
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-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	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

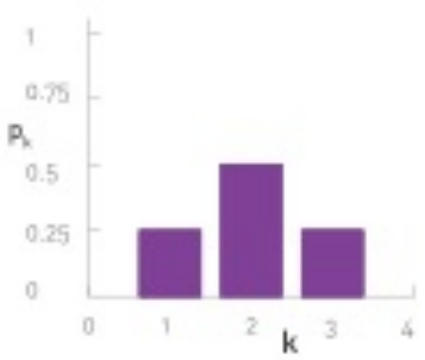
DEGREE DISTRIBUTION

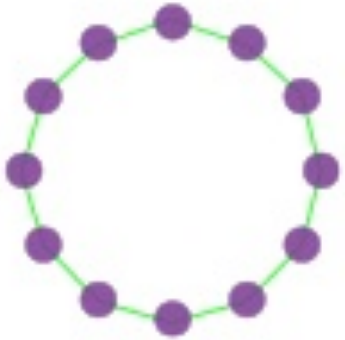
Degree distribution

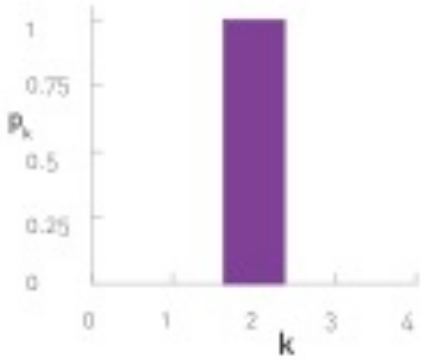
P(k): probability that a

randomly chosen node

has degree *k*







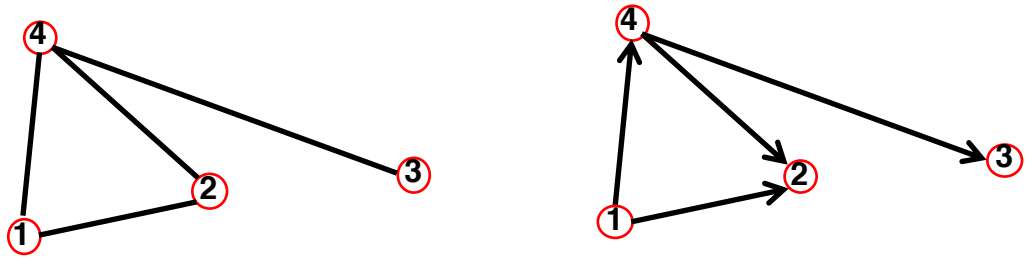
N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot

Section 2.4

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*

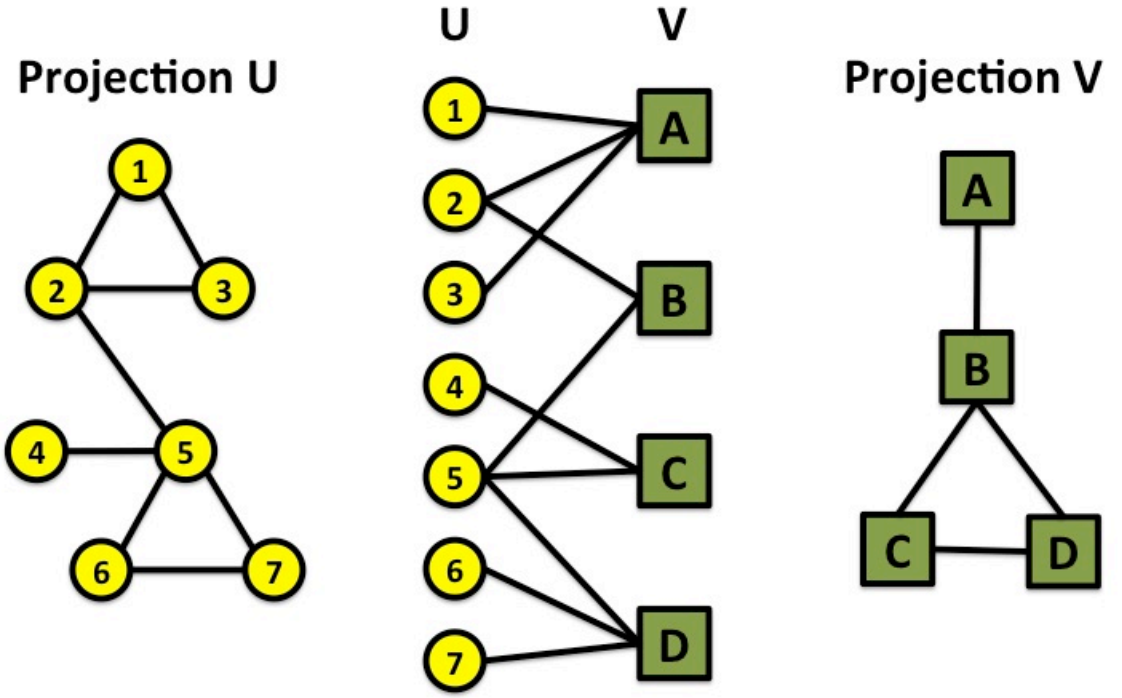
$A_{ij} = 0$ if there is no link pointing from *j* to *i*.

Section 2.7

BIPARTITE NETWORKS

BIPARTITE GRAPHS

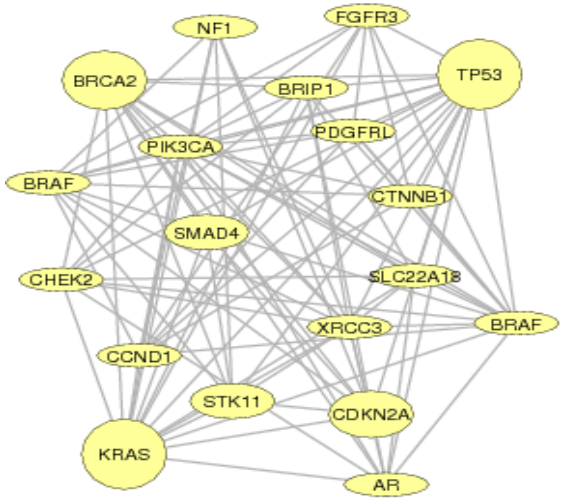
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



Examples:

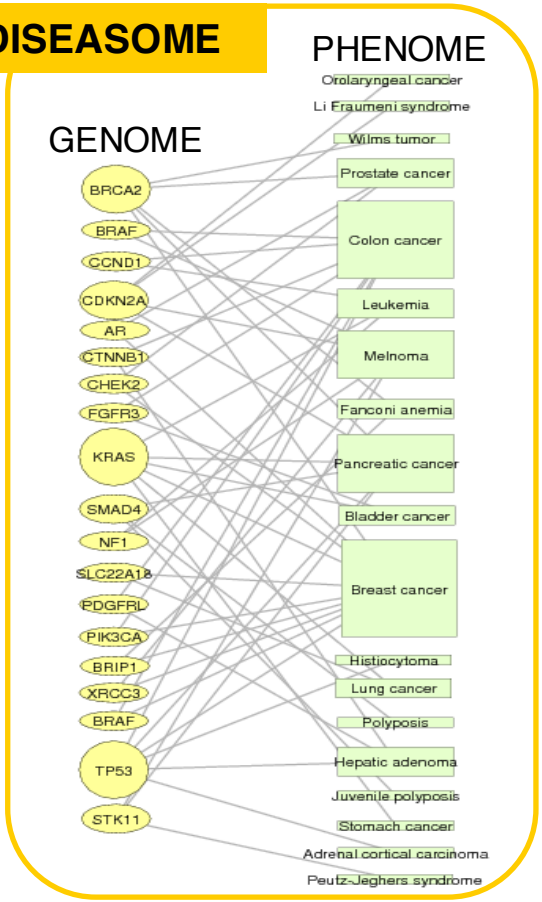
Hollywood actor network
Collaboration networks
Disease network (diseasome)

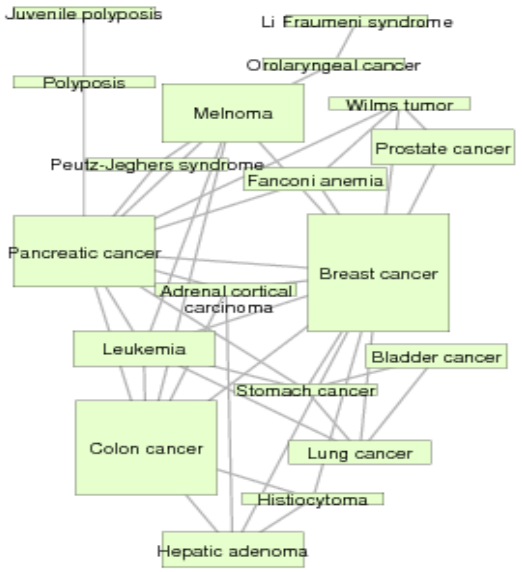
GENE NETWORK – DISEASE NETWORK



Gene network

DISEASOME

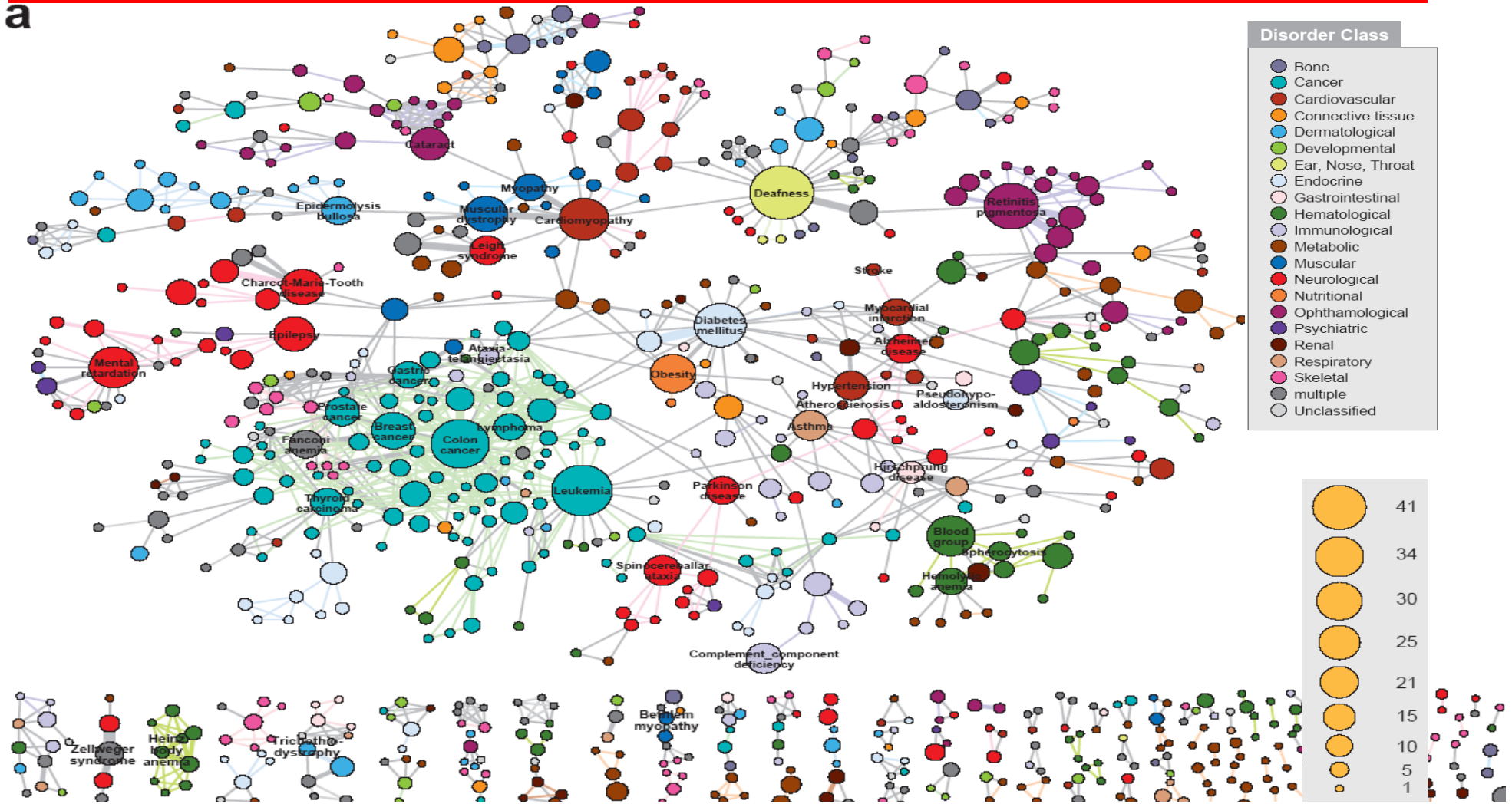




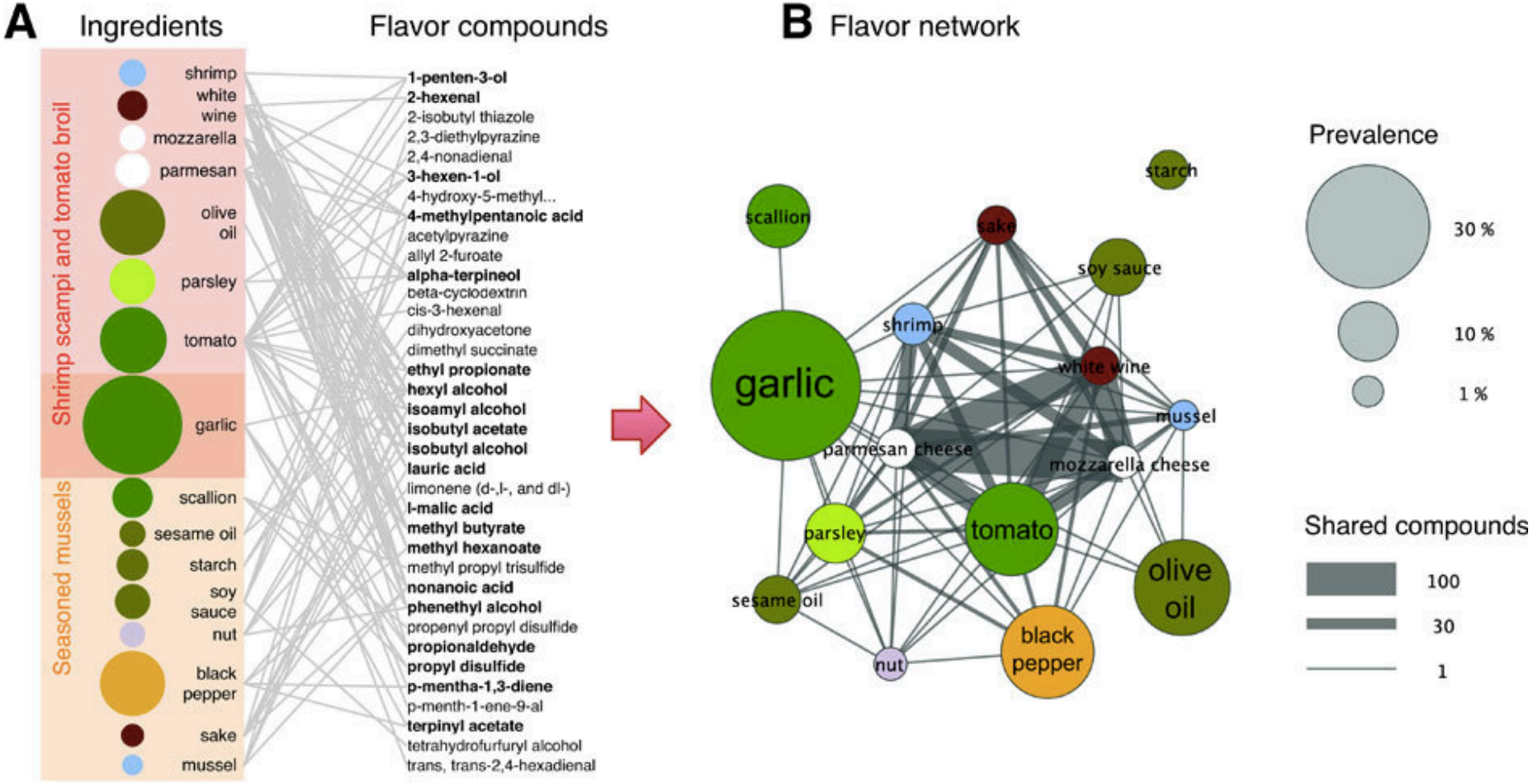
Disease network

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

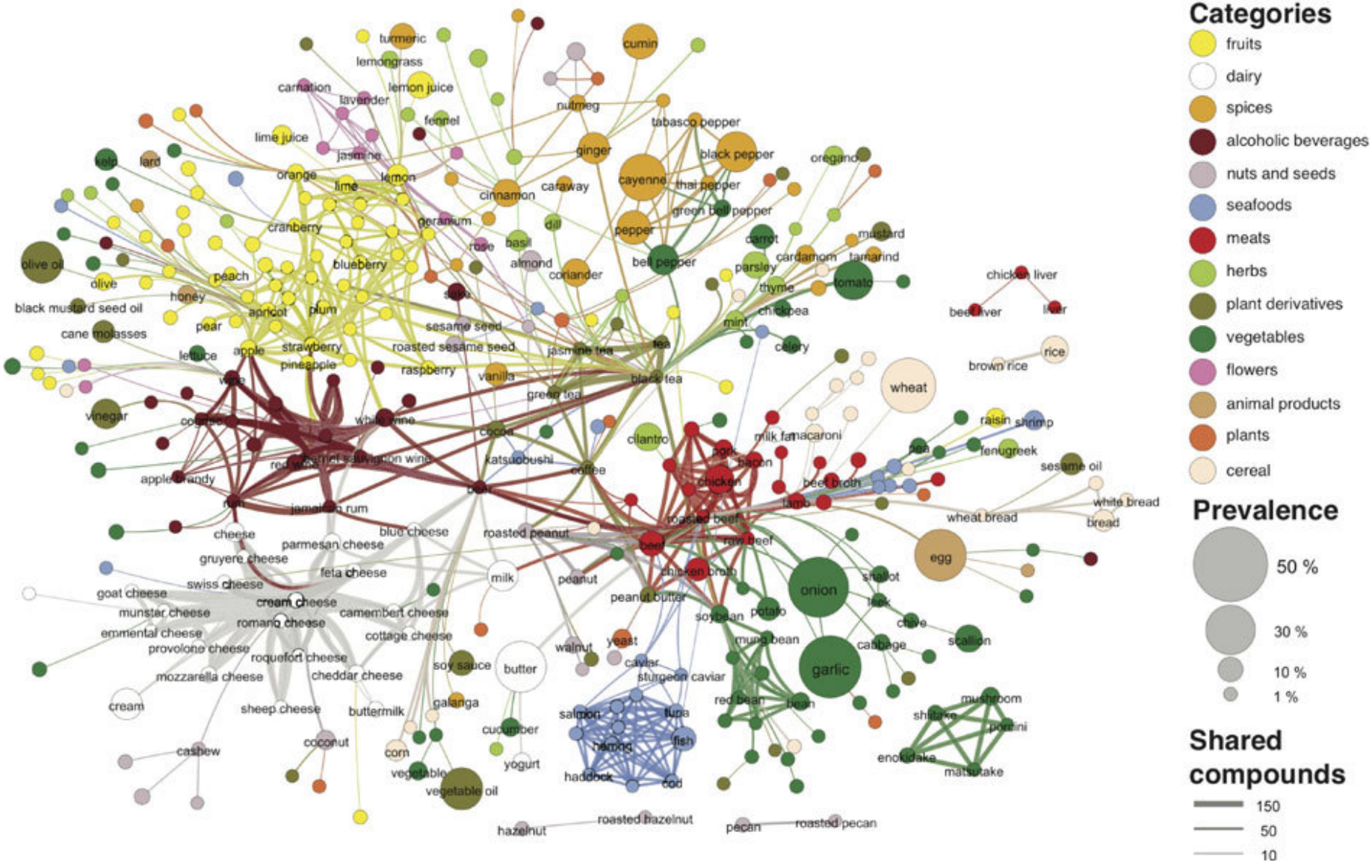
HUMAN DISEASE NETWORK



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

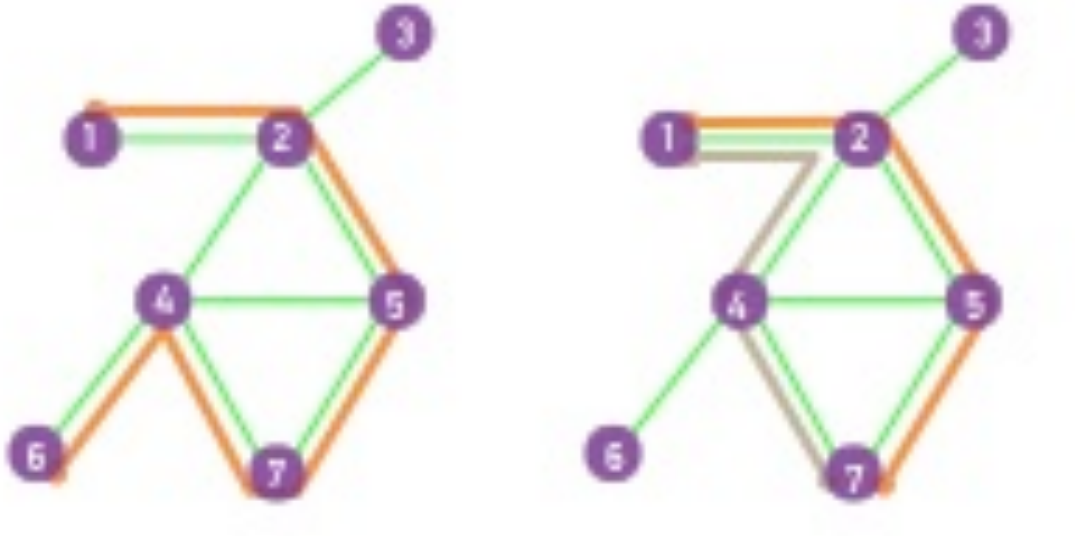
PATHOLOGY

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

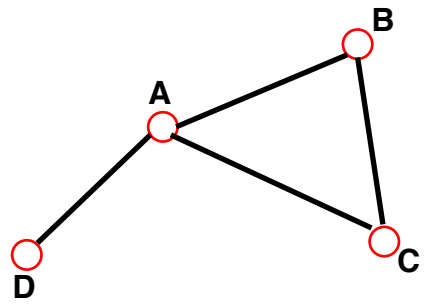
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



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

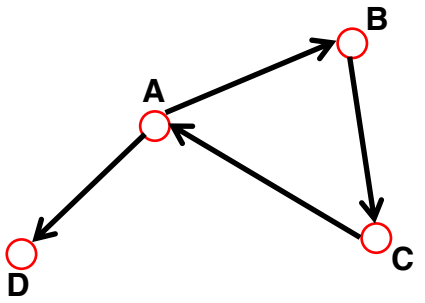
DISTANCE IN A GRAPH

Shortest Path, Geodesic Path



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

*If the two nodes are disconnected, the distance is infinity.



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).

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*

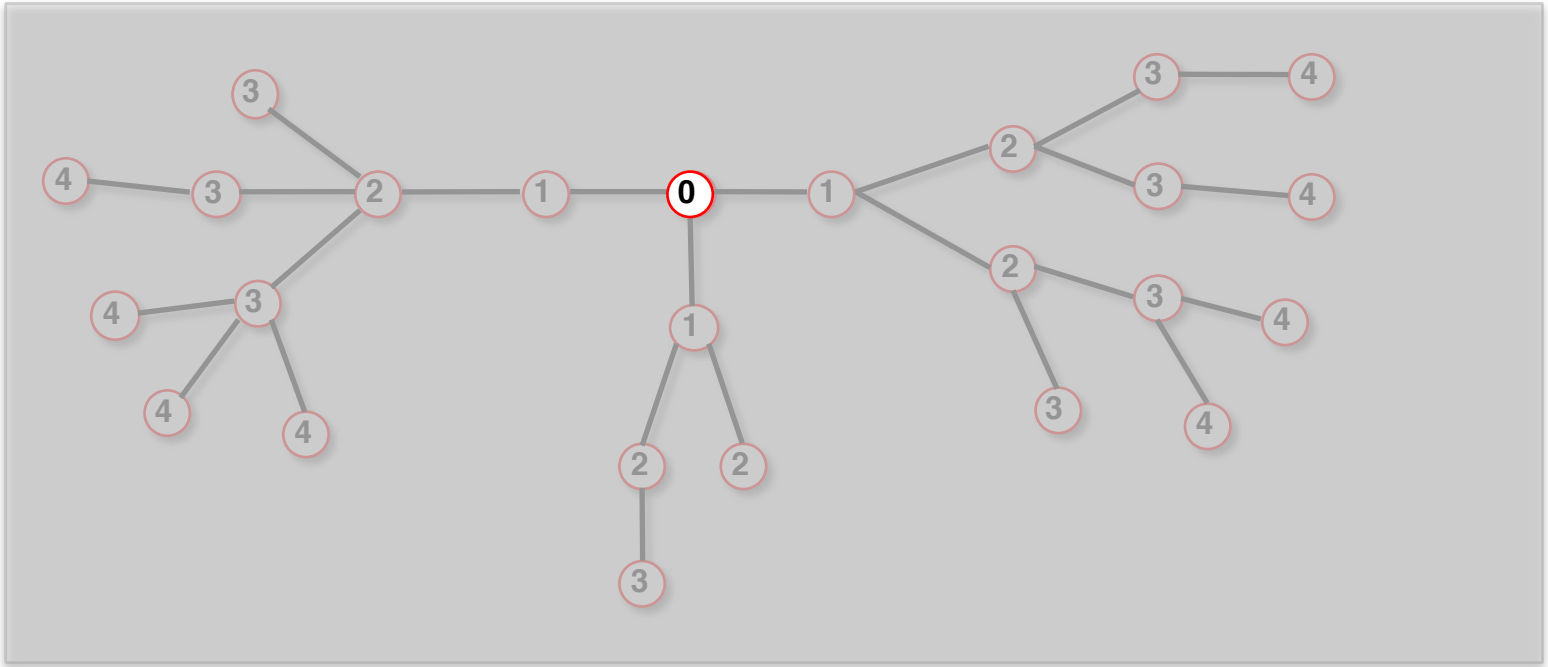
$$N_{ij}^{(n)} = [A^n]_{ij}$$

**holds for both directed and undirected networks.*

FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

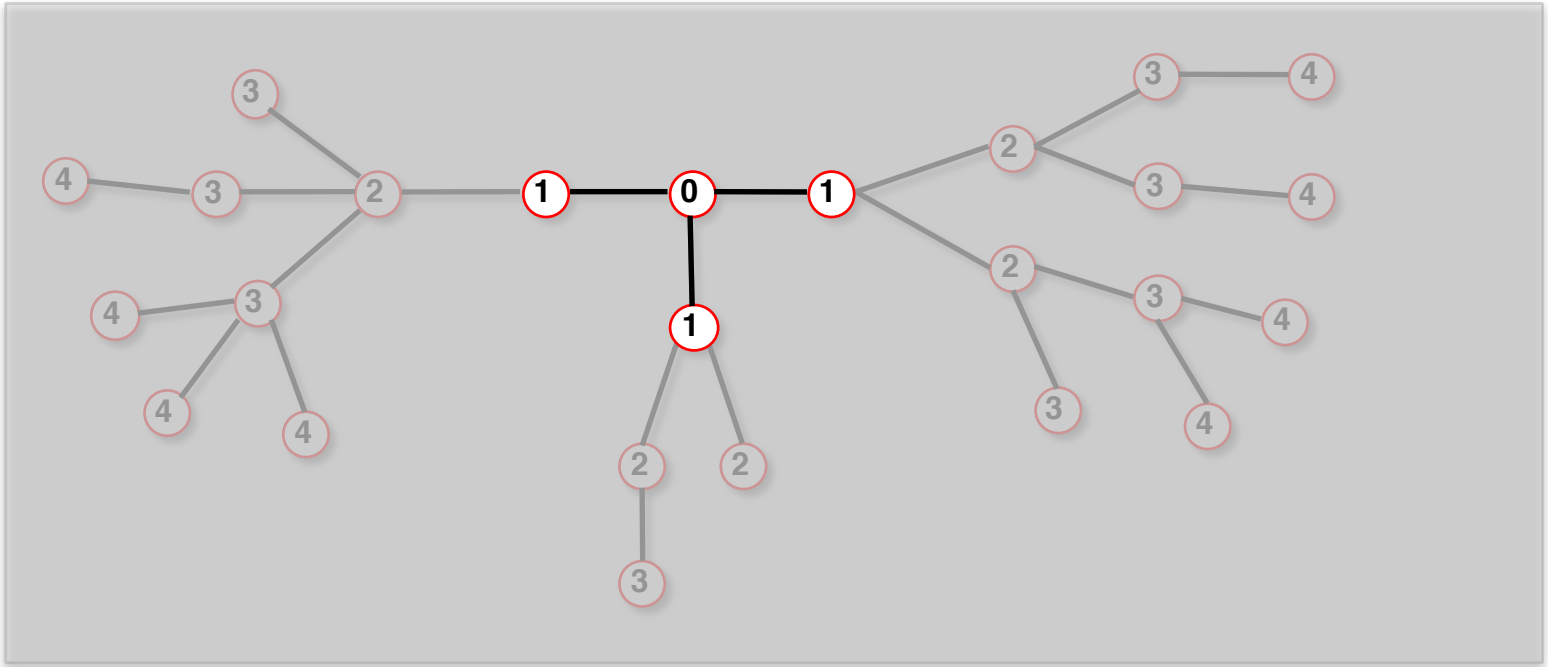
1. Start at 0.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

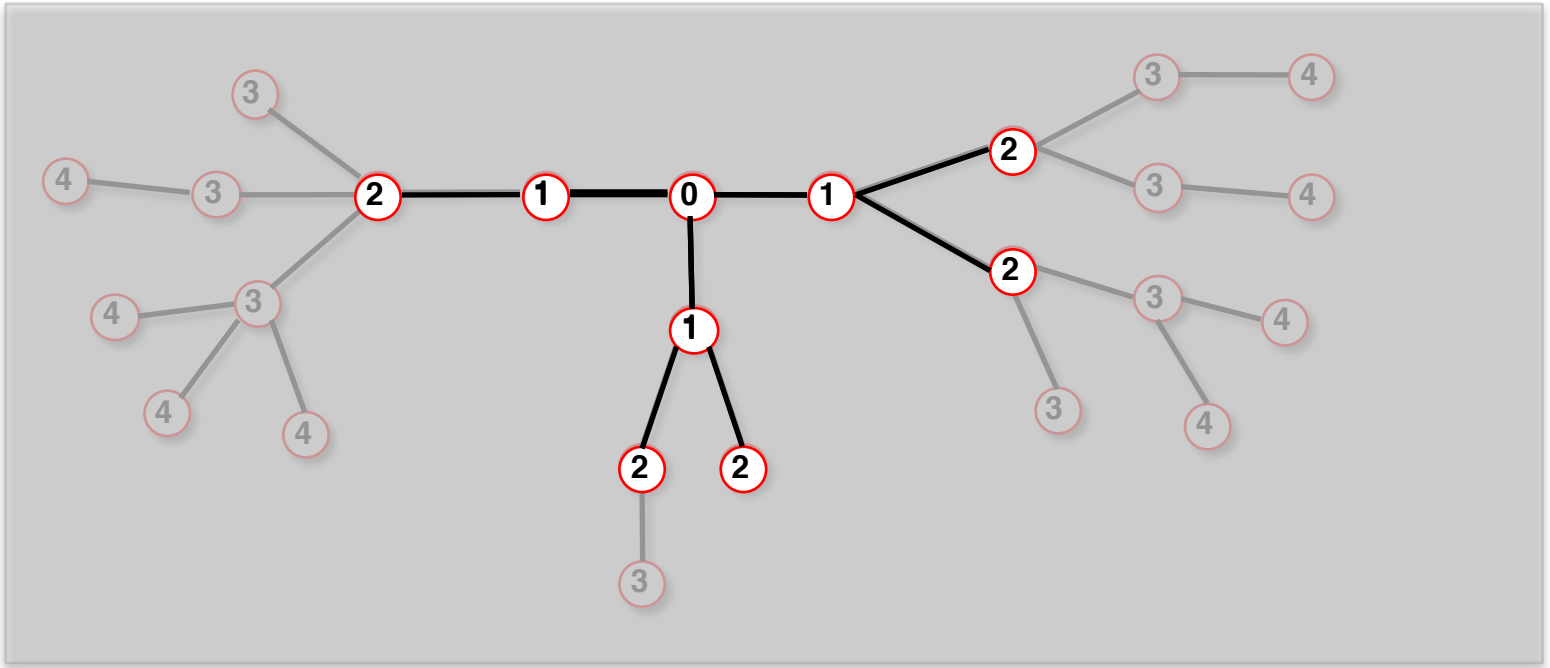
1. Start at 0.
2. Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

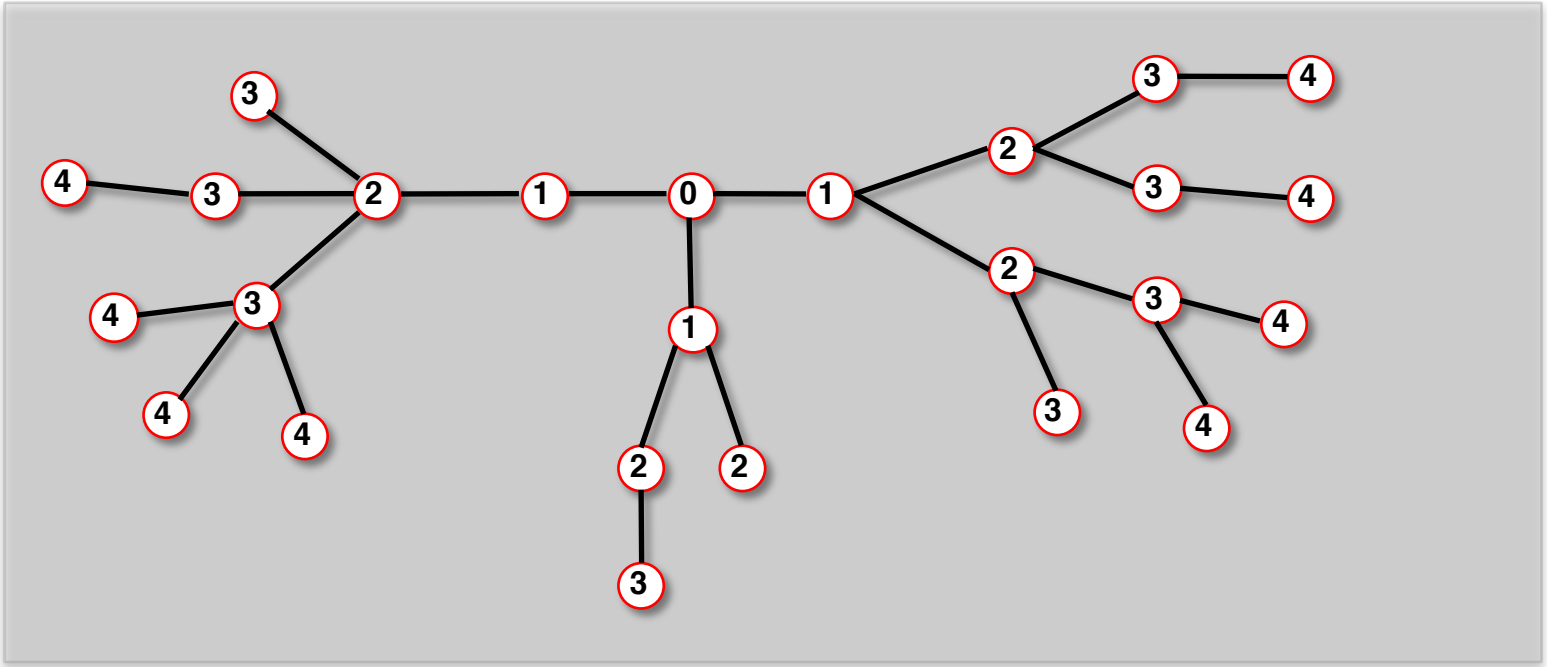
1. Start at 0.
2. Find the nodes adjacent to 0. Mark them as at distance 1. Put them in a queue.
3. Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

1. Repeat until you find node 4 or there are no more nodes in the queue.
2. The distance between 0 and 4 is the label of 4 or, if 4 does not have a label, infinity.



NETWORK DIAMETER AND AVERAGE DISTANCE

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

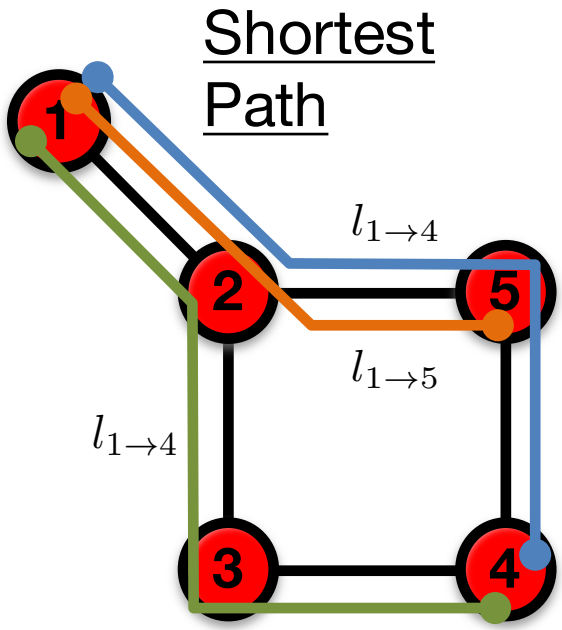
Average path length/distance, <*d*>, for a connected graph:

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

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

PATHOLOGY: summary



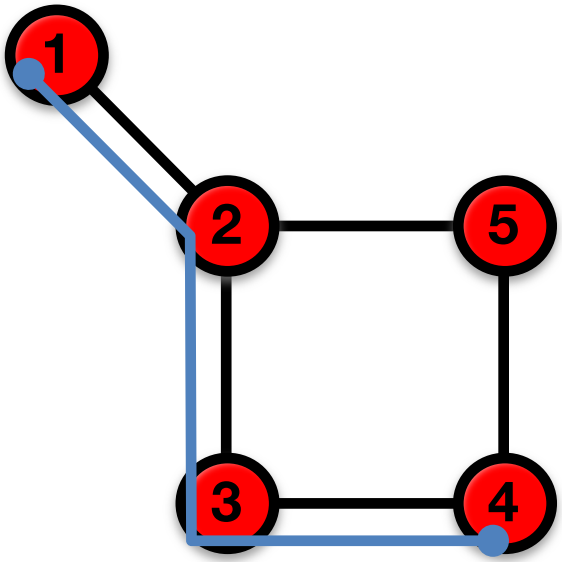
$$l_{1\to4} = 3$$

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

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

PATHOLOGY: summary

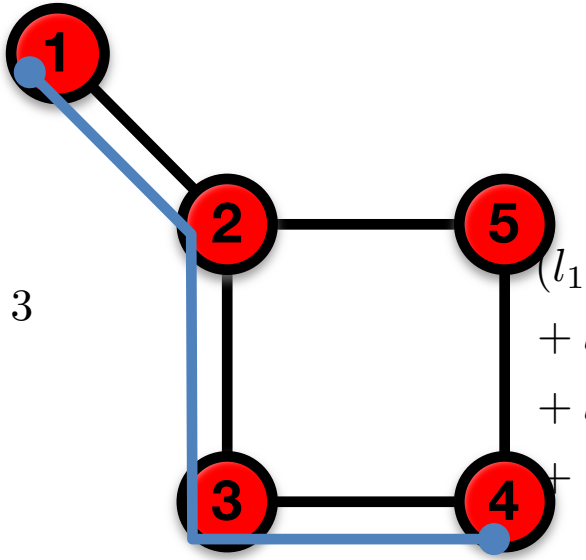
Diameter



$l_{1\to 4} = 3$

The longest shortest path in a graph

Average Path Length

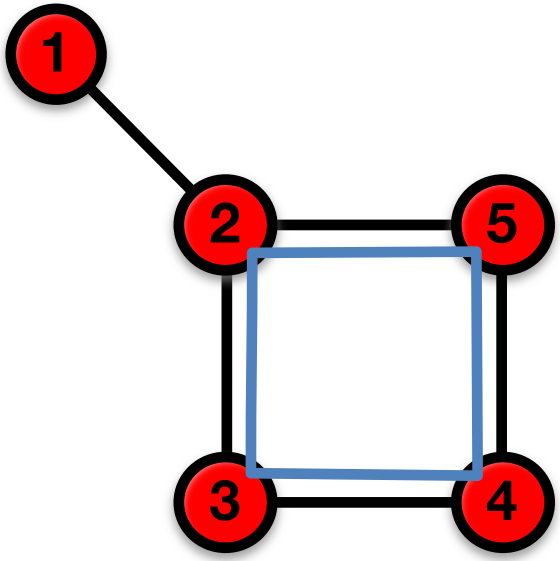


$$(l_{1\to 2} + l_{1\to 3} + l_{1\to 4} + \\ + l_{1\to 5} + l_{2\to 3} + l_{2\to 4} + \\ + l_{2\to 5} + l_{3\to 4} + l_{3\to 5} + \\ + l_{4\to 5})\,/10 = 1.6$$

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

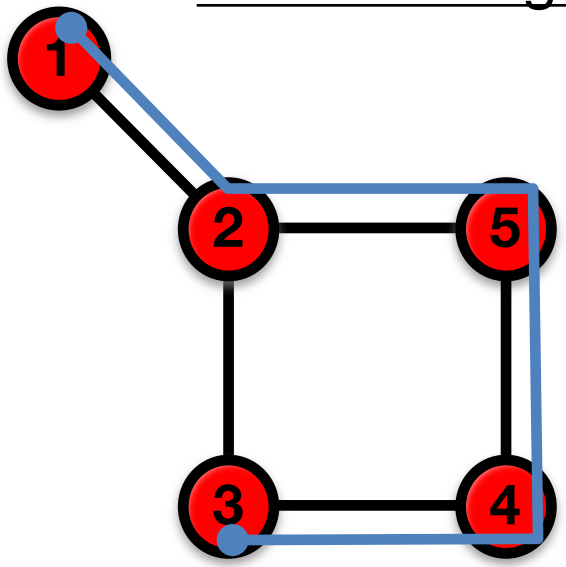
PATHOLOGY: summary

Cycle



A path with the same start and end node.

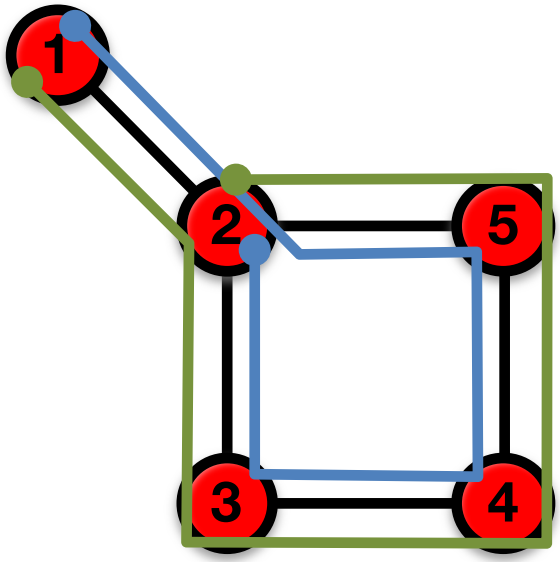
Self-avoiding Path



A path that does not intersect itself.

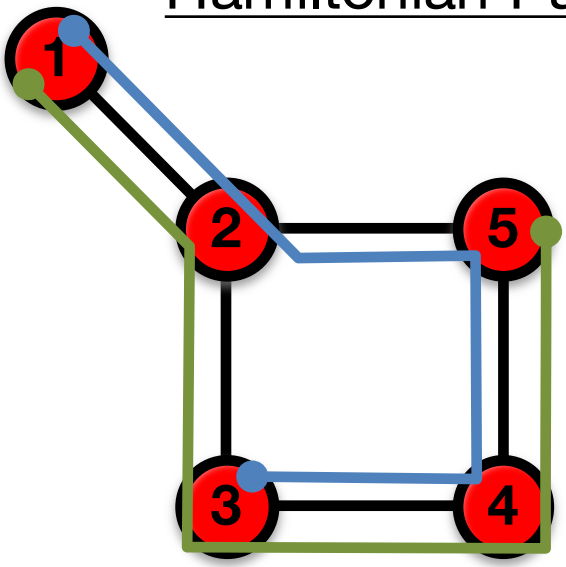
PATHOLOGY: summary

Eulerian Path



A path that traverses each link exactly once.

Hamiltonian Path

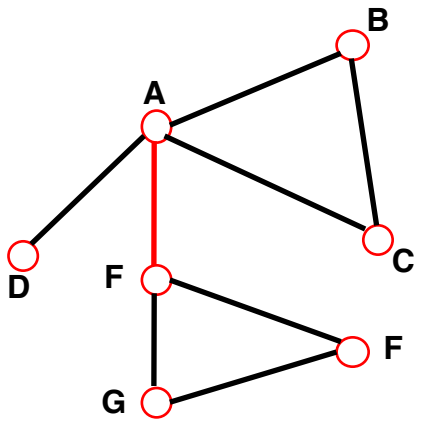


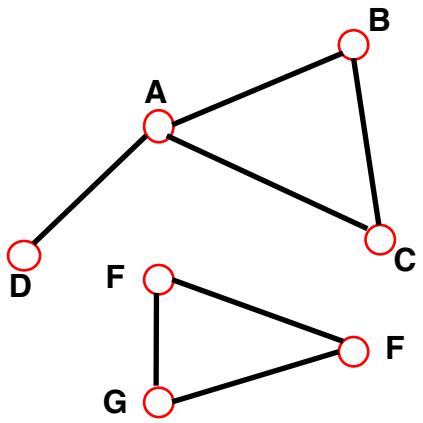
A path that visits each node exactly once.

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**

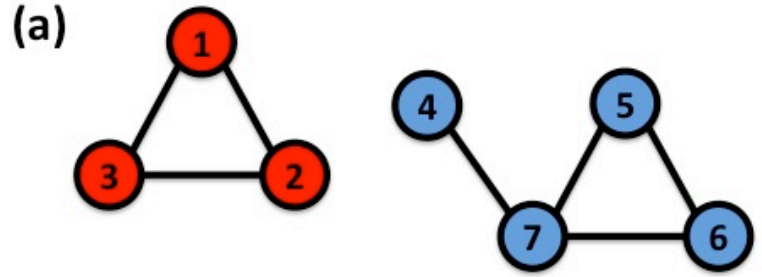
Bridge: if we erase it, the graph becomes disconnected.

CONNECTIVITY OF UNDIRECTED GRAPHS

Adjacency Matrix

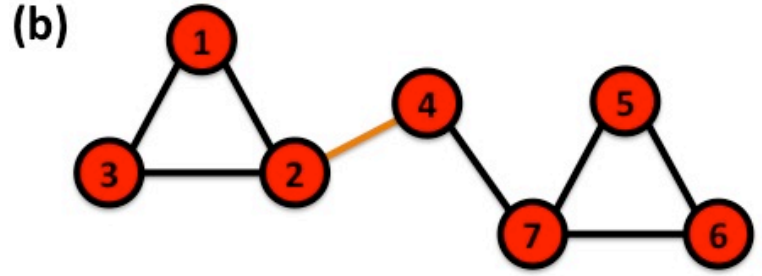
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

(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)



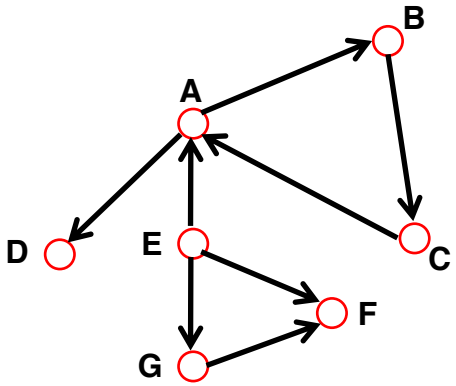
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

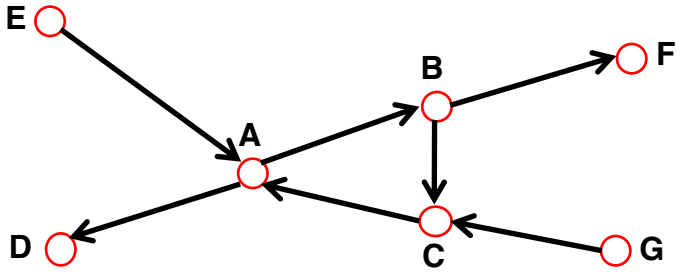
CONNECTIVITY OF DIRECTED GRAPHS

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.

FINDING THE CONNECTED COMPONENTS OF A NETWORK

1. Start from a randomly chosen node i and perform a BFS (BOX 2.5). Label all nodes reached this way with $n = 1$.

2. If the total number of labeled nodes equals N, then the network is connected. If the number of labeled nodes is smaller than N, the network consists of several components. To identify them, proceed to step 3.

3. Increase the label $n \rightarrow n + 1$. Choose an unmarked node j, label it with n. Use BFS to find all nodes reachable from j, label them all with n. Return to step 2.

Clustering coefficient

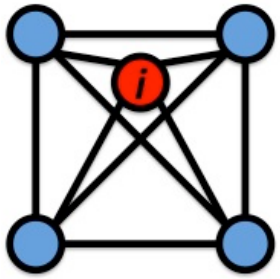
CLUSTERING COEFFICIENT

Clustering coefficient:

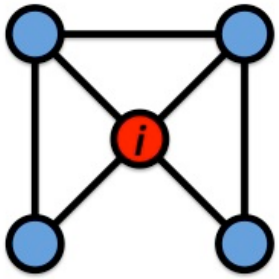
what fraction of your neighbors are connected?

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

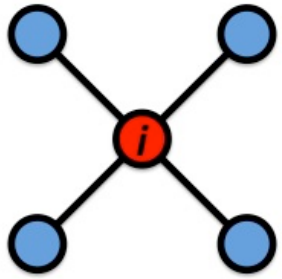
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



$C_i = 1$



$C_i = 1/2$



$C_i = 0$

Watts & Strogatz, Nature 1998.

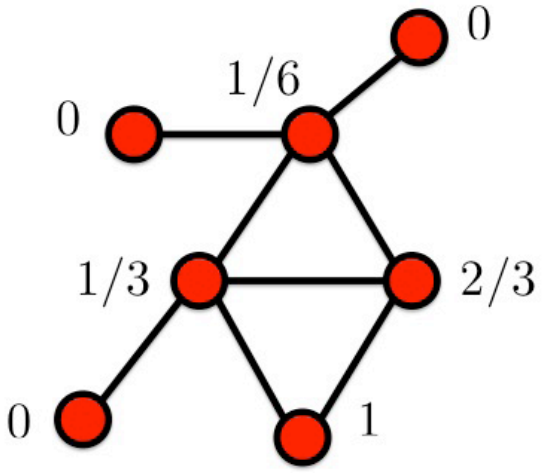
CLUSTERING COEFFICIENT

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.

Section 11

summary

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

Degree distribution: **P(k)**

Path length:

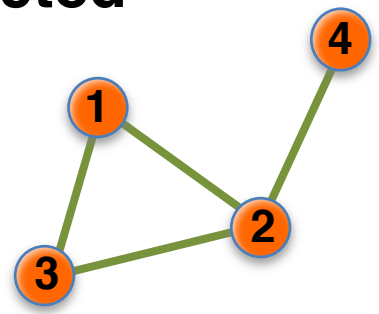
<d>

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

Clustering coefficient:

GRAPHOLOGY 1

Undirected



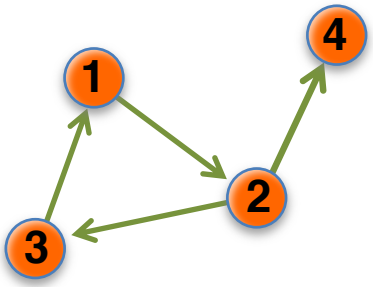
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions

Directed



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

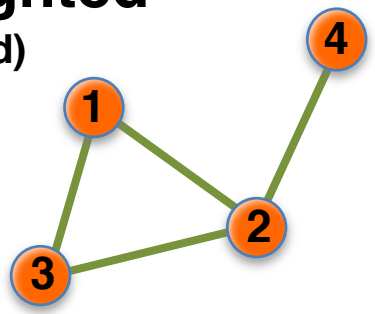
$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

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

WWW, citation networks

GRAPHOLOGY 2

Unweighted
(undirected)



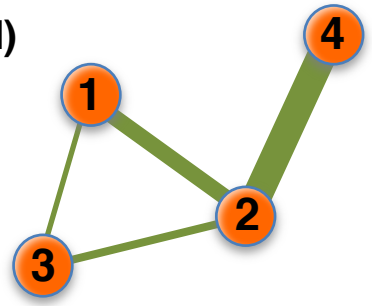
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

protein-protein interactions, www

Weighted
(undirected)



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

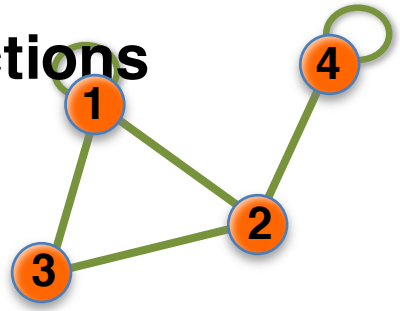
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

Call Graph, metabolic networks

GRAPHOLOGY 3

Self-interactions



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

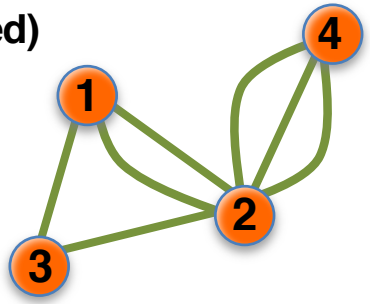
$A_{ii} \neq 0$ $\qquad$ $A_{ij} = A_{ji}$

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

Protein interaction network, www

Multigraph

(undirected)



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

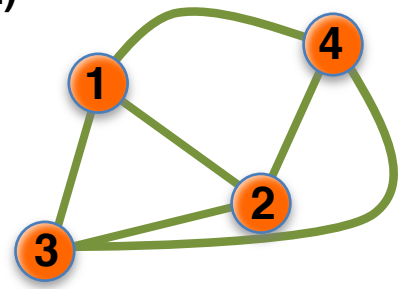
$A_{ii} = 0$ $\qquad$ $A_{ij} = A_{ji}$

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

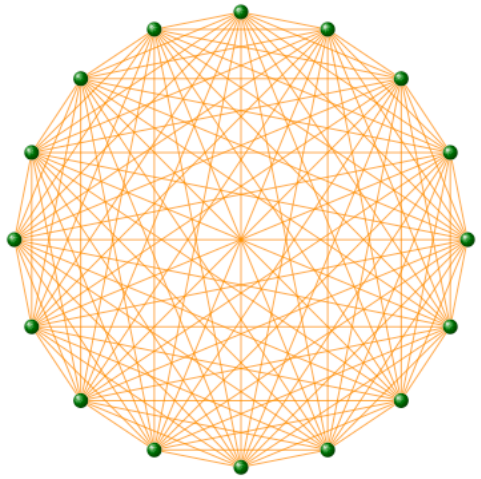
Social networks, collaboration networks

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.

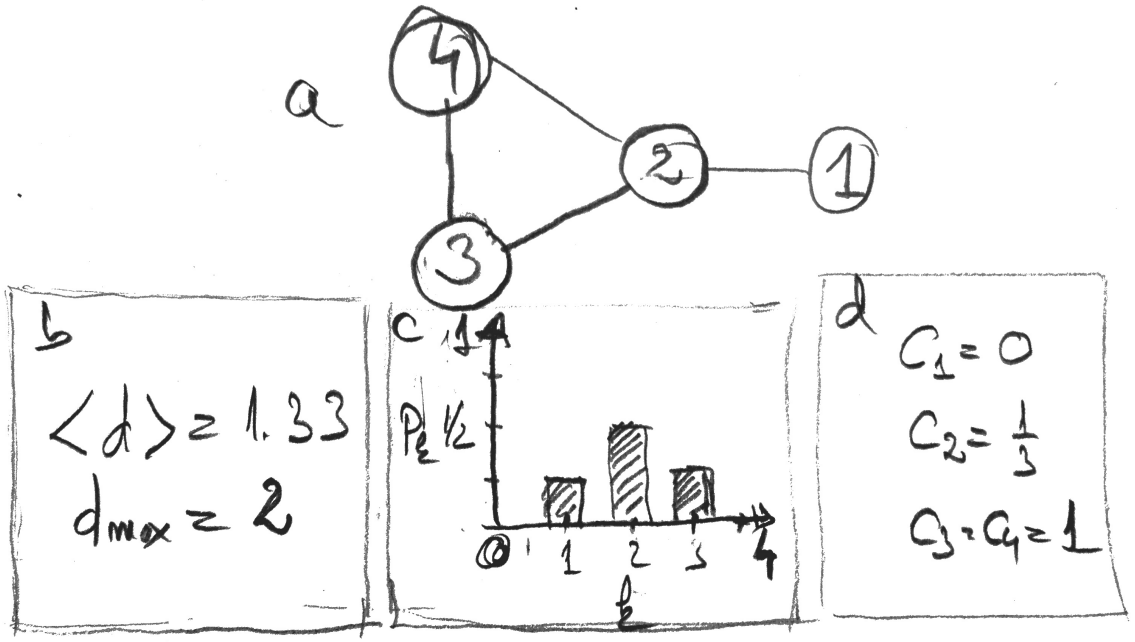
Mobile phone calls > directed, weighted.

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

Network properties

- Density: The density D of a network is defined as a ratio of the number of edges E to the number of possible edges,
- Size: The size of a network can refer to the number of nodes N or, less commonly, the number of edges E which can range from N-1 (a tree) to E_{max} (a complete graph).
- Average degree: The degree k of a node is the number of edges connected to it. Closely related to the density of a network is the average degree, <k> = \tfrac{2E}{N}. In the ER random graph model, we can compute <k> = p(N-1) where p is the probability of two nodes being connected.
- Average path length: Average path length is calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs. This shows us, on average, the number of steps it takes to get from one member of the network to another.
- Diameter of a network: As another means of measuring network graphs, we can define the diameter of a network as the longest of all the calculated shortest paths in a network. It is the shortest distance between the two most distant nodes in the network. In other words, once the shortest path length from every node to all other nodes is calculated, the diameter is the longest of all the calculated path lengths. The diameter is representative of the linear size of a network.
- Clustering coefficient
- Connectedness: The way in which a network is connected plays a large part into how networks are analyzed and interpreted.
  - Clique/Complete Graph: a completely connected network, where all nodes are connected to every other node. These networks are symmetric in that all nodes have in-links and out-links from all others.
  - Giant Component: A single connected component which contains most of the nodes in the network.
  - Weakly Connected Component: A collection of nodes in which there exists a path from any node to any other, ignoring directionality of the edges.
  - Strongly Connected Component: A collection of nodes in which there exists a directed path from any node to any other.
- Node centrality: Centrality indices produce rankings which seek to identify the most important nodes in a network model. Different centrality indices encode different contexts for the word "importance."
  - The betweenness centrality, for example, considers a node highly important if it form bridges between many other nodes.
  - The eigenvalue centrality, in contrast, considers a node highly important if many other highly important nodes link to it. Hundreds of such measures have been proposed in the literature.
- Node influence: In graph theory and network analysis, node influence metrics are measures that rank or quantify the influence of every node (also called vertex) within a graph. They are related to centrality indices. Applications include measuring the influence of each person in a social network, understanding the role of infrastructure nodes in transportation networks, the Internet, or urban networks, and the participation of a given node in disease dynamics.

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE



A. Degree distribution: p_k

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

C. Clustering coefficient:

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

Network models

Erdős–Rényi Random Graph model

- Used for generating random graphs in which edges are set between nodes with equal probabilities
  - prove the existence of graphs satisfying various properties, or
  - provide a rigorous definition of what it means for a property to hold for almost all graphs.
- Generating an Erdős–Rényi model
  - the number of nodes in the graph generated as N
  - the probability that a link should be formed between any two nodes as p
  - A constant ⟨k⟩ may derived from these two components with the formula
    - $\langle k \rangle = 2 \cdot E / N = p \cdot (N - 1)$, where
    - E is the expected number of edges

`http://igraph.org/r/doc/erdos.renyi.game.html`

Watts-Strogatz Small World model

- A random graph generation model that produces graphs with small-world properties
- An initial lattice structure is used to generate a Watts-Strogatz model.
  - Each node in the network is initially linked to its <k> closest neighbours
  - Another parameter is specified as the rewiring probability:
    - Each edge has a probability p that it will be rewired to the graph as a random edge.
    - The expected number of rewired links in the model is pE = pN<k>/2.

http://www.mathworks.com/help/matlab/math/build-watts-strogatz-small-world-graph-model.html

Barabási–Albert (BA) Preferential Attachment model

- Random network model used to demonstrate a preferential attachment
  - "rich-get-richer" effect
  - An edge is most likely to attach to nodes with higher degrees
- The network begins with an initial network of m0 nodes
  - $m0 \geq 2$
  - the degree of each node in the initial network should be at least 1,
  - otherwise it will always remain disconnected from the rest of the network.
- New nodes are added to the network one at a time.
  - Each new node is connected to m existing nodes
  - With a probability that is proportional to the number of links that the existing nodes already have

Barabási–Albert (BA) Preferential Attachment model

- Random network model used to demonstrate a preferential attachment

Some remarks

- Heavily linked nodes ("hubs") tend to quickly accumulate even more links,
- Nodes with only a few links are unlikely to be chosen as the destination for a new link.
- New nodes have a "preference" to attach themselves to the already heavily linked nodes.

  - Each new node is connected to m existing nodes
  - With a probability that is proportional to the number of links that the existing nodes already have

Network analysis

Network analysis

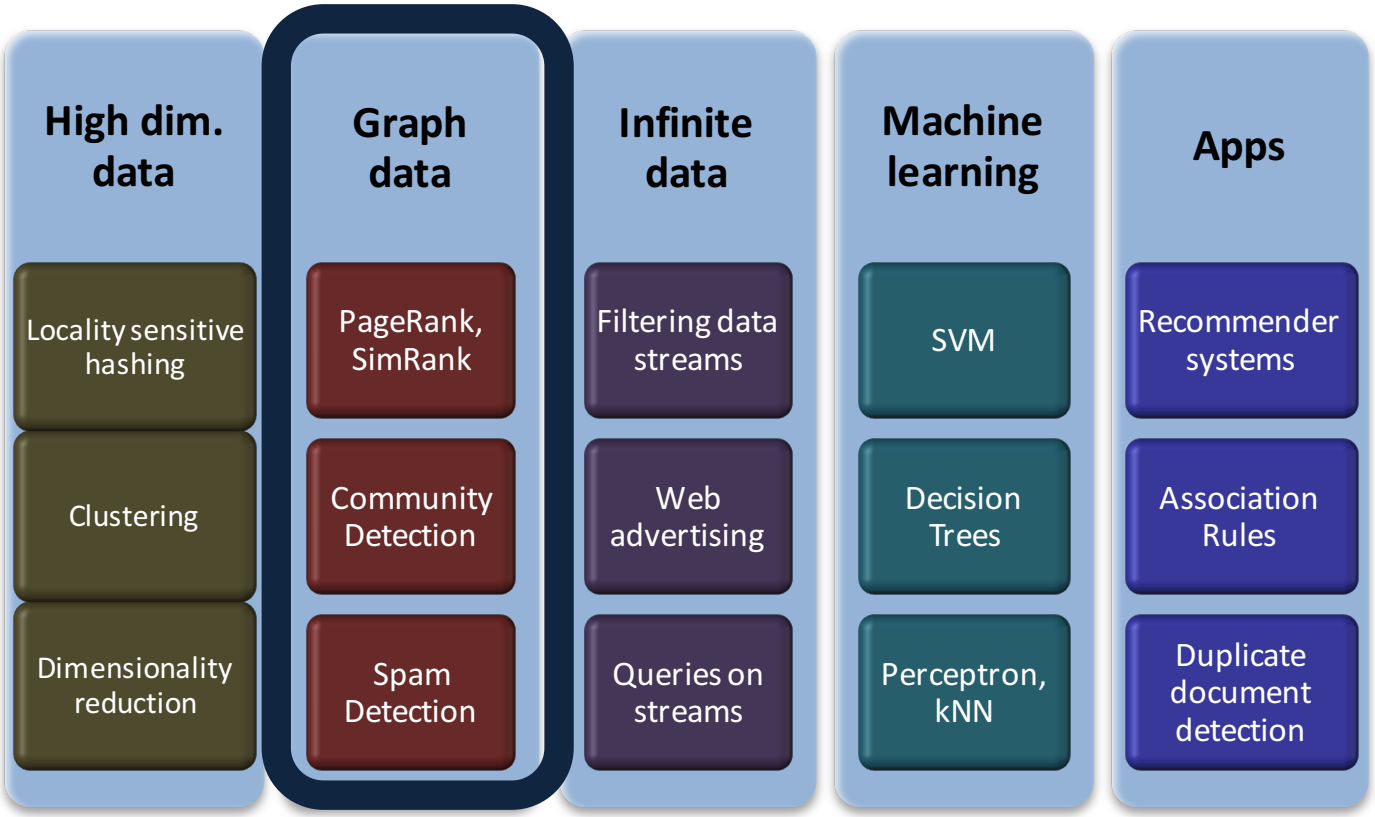
- Social network analysis
  - Examines the structure of relationships between social entities
  - Entities are often people, but may also be groups, organizations nation states, web sites, scholarly publications
- Dynamic network analysis:
  - examines the shifting structure of relationships among different classes of entities in complex socio-technical systems effects
  - reflects social stability and changes such as the emergence of new groups, topics, and leaders

Network analysis

- Biological network analysis
  - closely related to social network analysis
  - focusing on local patterns in the network
    - network motifs are small sub-graphs that are over-represented in the network.
    - analysis of biological networks has led to the development of network medicine
- Link analysis
  - Exploring associations between objects.
  - examining the addresses of suspects and victims, the telephone numbers they have dialled and financial transactions that they have partaken in during a given timeframe, and the familial relationships between these subjects as a part of police investigation.
  - Link analysis here provides the crucial relationships and associations between very many objects of different types that are not apparent from isolated pieces of information
  - Pandemic analysis, Web link analysis, Page Rank, ..

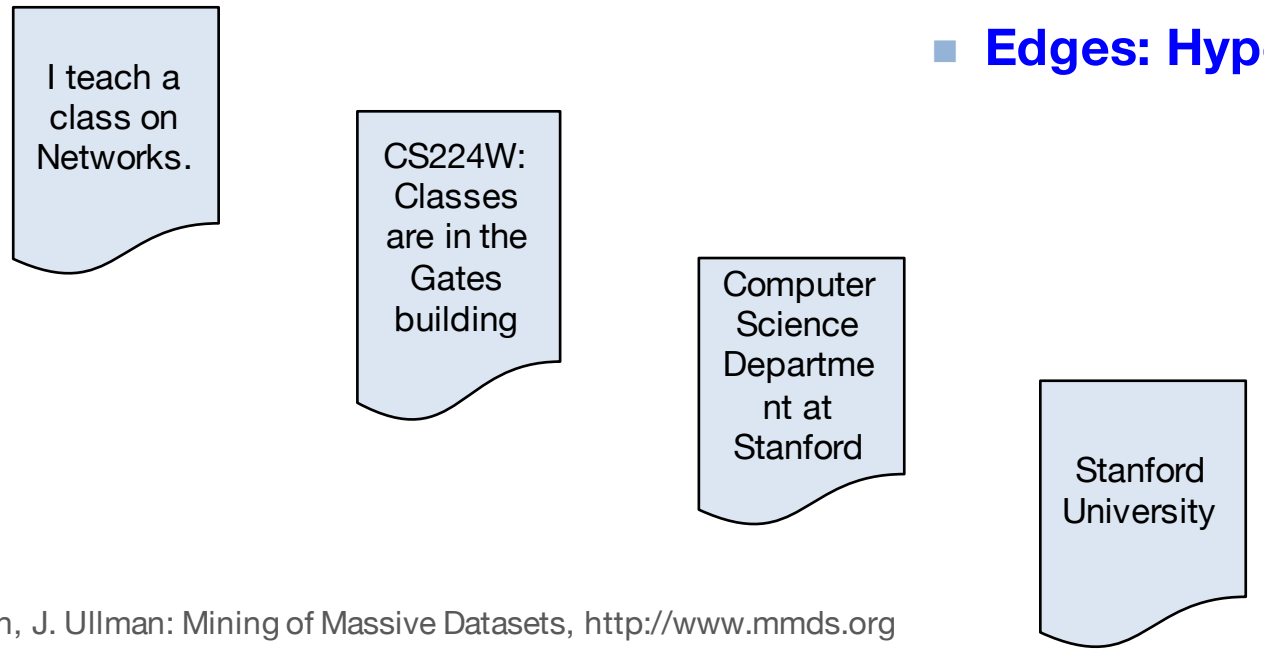
Analysis of large graphs

Graph data


High dim. data
Locality sensitive hashing
Clustering
Dimensionality reduction
Graph data
PageRank, SimRank
Community Detection
Spam Detection
Infinite data
Filtering data streams
Web advertising
Queries on streams
Machine learning
SVM
Decision Trees
Perceptron, kNN
Apps
Recommender systems
Association Rules
Duplicate document detection

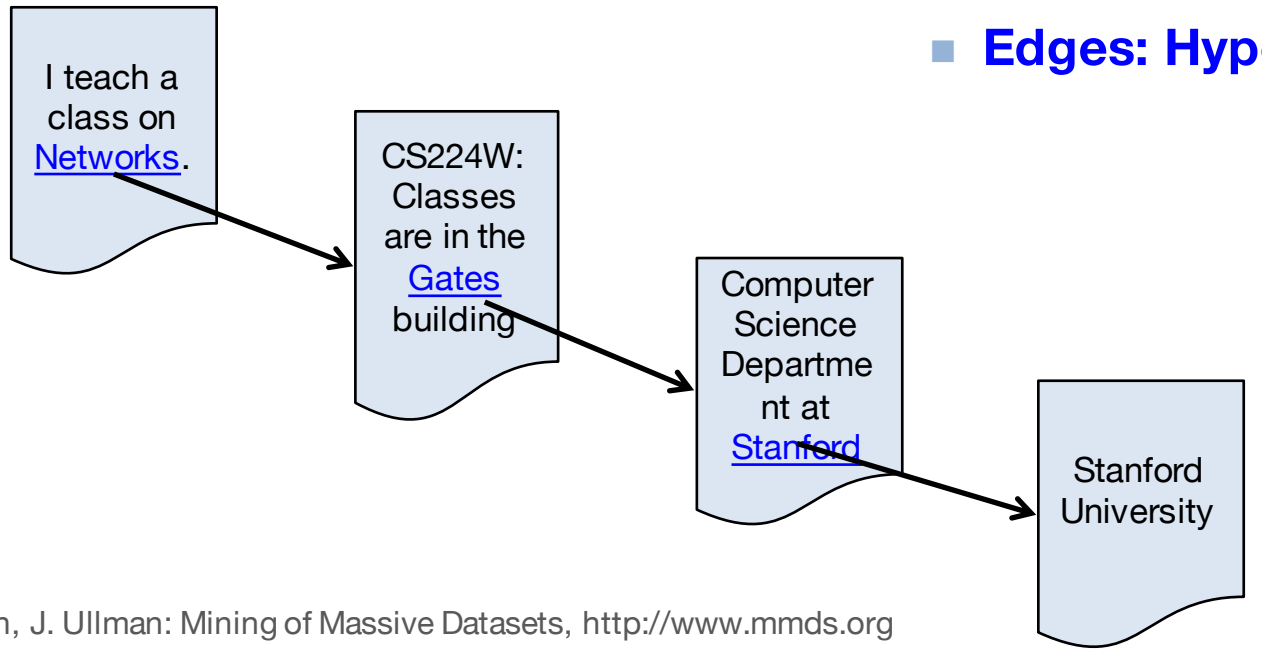
Web as a Graph

- **Web as a directed graph:**
  - **Nodes: Webpages**
  - **Edges: Hyperlinks**

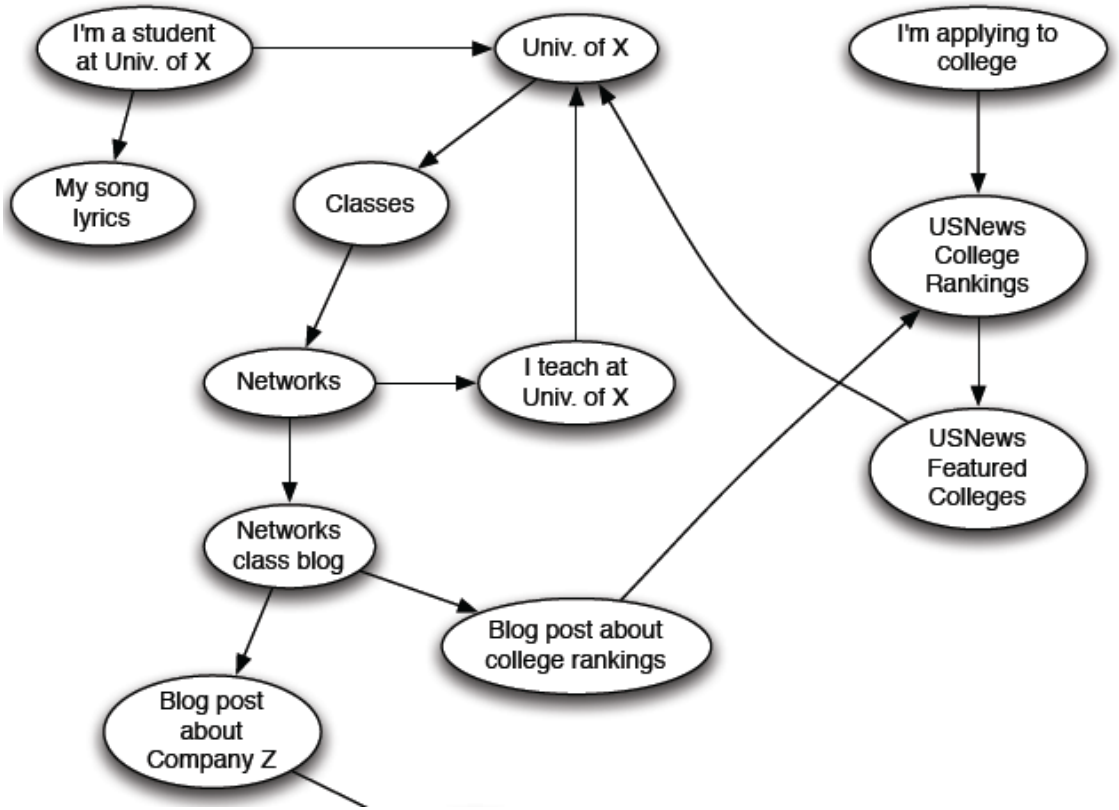


Web as a Graph

- **Web as a directed graph:**
  - **Nodes: Webpages**
  - **Edges: Hyperlinks**



Web as a Directed Graph



Broad Question

- **How to organize the Web?**
- **First try:** Human curated
  Web directories
  - Yahoo, DMOZ, LookSmart
- **Second try: Web Search**
  - **Information Retrieval** investigates:
    Find relevant docs in a small
    and trusted set
    - Newspaper articles, Patents, etc.
  - **But:** Web is **huge**, full of untrusted documents, random things, web spam, etc.

Web Search: 2 Challenges

2 challenges of web search:

- **(1) Web contains many sources of information Who to "trust"?**
  - **Trick:** Trustworthy pages may point to each other!

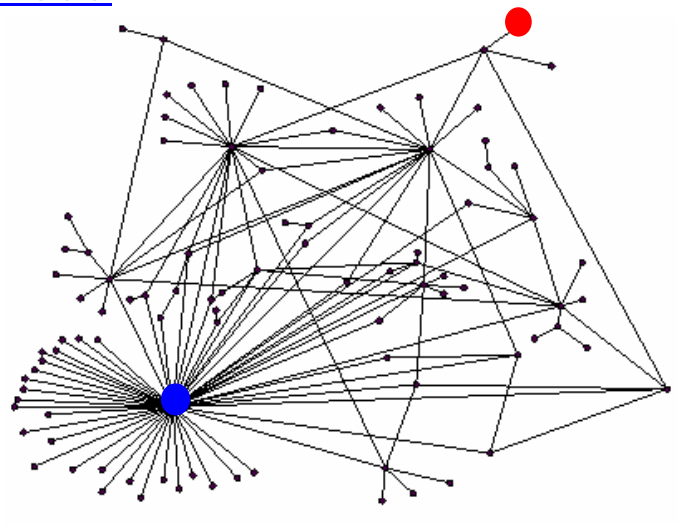
- **(2) What is the "best" answer to query "newspaper"?**
  - No single right answer
  - **Trick:** Pages that actually know about newspapers might all be pointing to many newspapers

Ranking Nodes on the Graph

- **All web pages are not equally "important"**

  www.joe-schmoe.com vs. www.stanford.edu

- There is large diversity in the web-graph node connectivity. **Let's rank the pages by the link structure!**



Link Analysis Algorithms

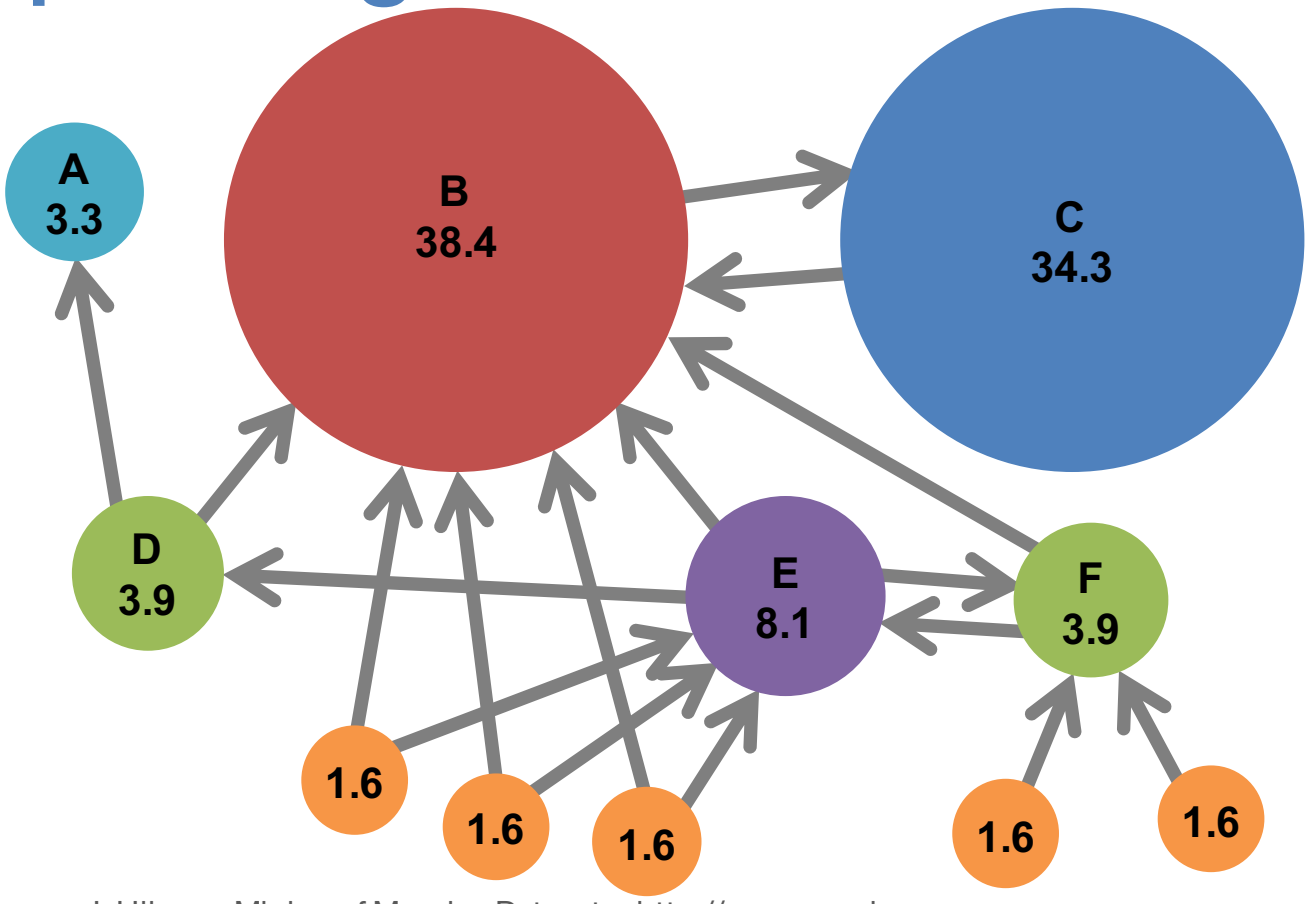
- **Link Analysis approaches** for computing **importances** of nodes in a graph:
  - Page Rank
  - Topic-Specific (Personalized) Page Rank
  - Web Spam Detection Algorithms

Page rank

Links as Votes

- **Idea: Links as votes**
  - **Page is more important if it has more links**
    - In-coming links? Out-going links?
- **Think of in-links as votes:**
  - www.stanford.edu has 23,400 in-links
  - www.joe-schmoe.com has 1 in-link
- **Are all in-links are equal?**
  - **Links from important pages count more**
  - Recursive question!

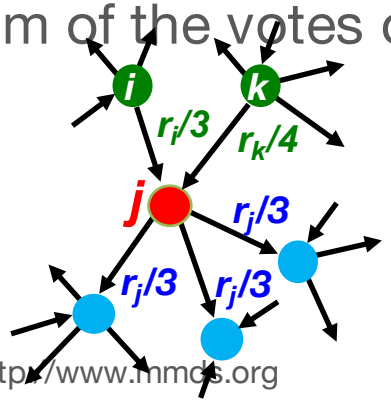
Example: PageRank Scores



Simple Recursive Formulation

- Each link's vote is proportional to the **importance** of its source page
- If page ***j*** with importance r_j has ***n*** out-links, each link gets r_j / ***n*** votes
- Page ***j***'s own importance is the sum of the votes on its in-links

$r_j = r_i/3 + r_k/4$



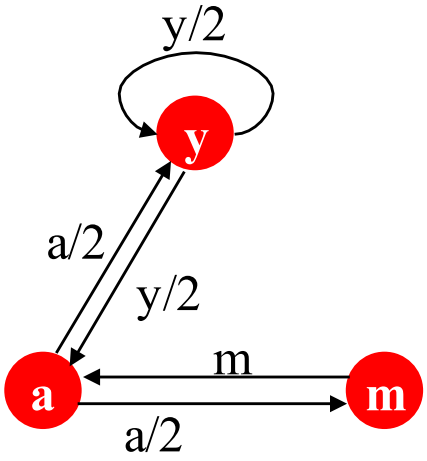
PageRank: The "Flow" Model

- A "vote" from an important page is worth more
- A page is important if it is pointed to by other important pages
- Define a "rank" r_j for page j

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … out-degree of node i

The web in 1839



"Flow" equations:

$r_y = r_y/2 + r_a/2$

$r_a = r_y/2 + r_m$

$r_m = r_a/2$

PageRank: Three Questions

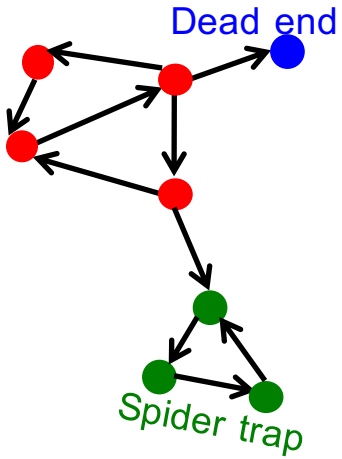
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

- **Does this converge?**
- **Does it converge to what we want?**
- **Are results reasonable?**

PageRank: Problems

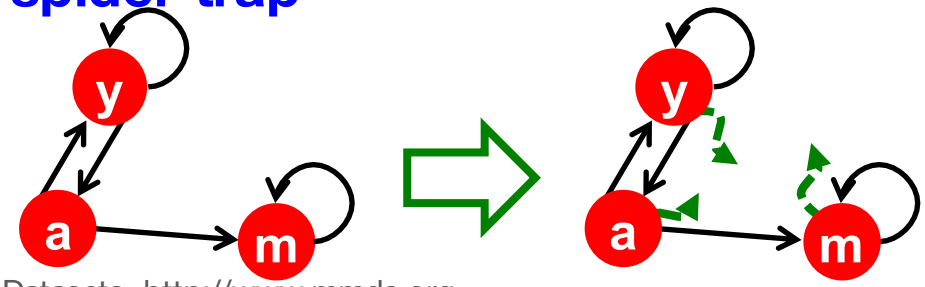
2 problems:

- **(1)** Some pages are **dead ends** (have no out-links)
  - Random walk has "nowhere" to go to
  - Such pages cause importance to "leak out"
- **(2) Spider traps:** (all out-links are within the group)
  - Random walked gets "stuck" in a trap
  - And eventually spider traps absorb all importance



Solution: Teleports!

- **The Google solution for spider traps: At each time step, the random surfer has two options**
  - With prob. β, follow a link at random
  - With prob. **1-β**, jump to some random page
  - Common values for β are in the range 0.8 to 0.9
- **Surfer will teleport out of spider trap within a few time steps**



Some Problems with Page Rank

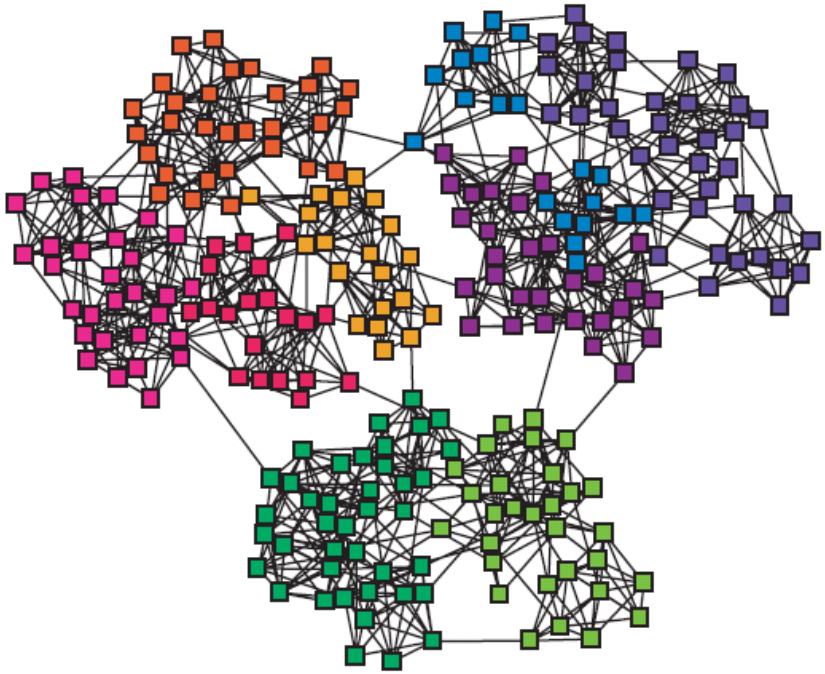
- **Measures generic popularity of a page**
  - Biased against topic-specific authorities
  - **Solution:** Topic-Specific PageRank (**next**)
- **Uses a single measure of importance**
  - Other models of importance
  - **Solution:** Hubs-and-Authorities
- **Susceptible to Link spam**
  - Artificial link topographies created in order to boost page rank
  - **Solution:** TrustRank

Challenge: implement a map reduce page rank algorithm

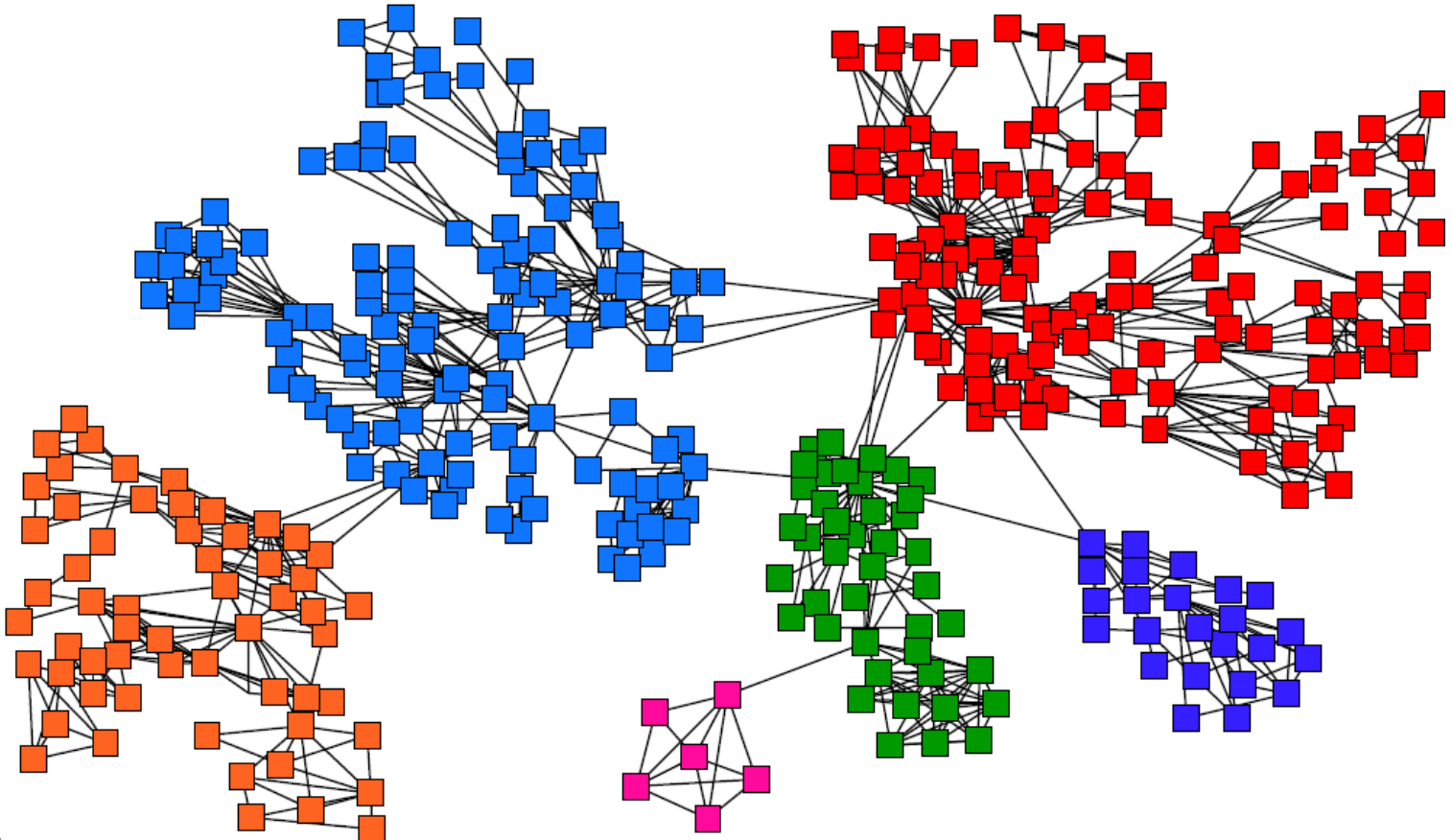
Community detection

Networks & Communities

- We often think of networks being organized into **modules, cluster, commur**



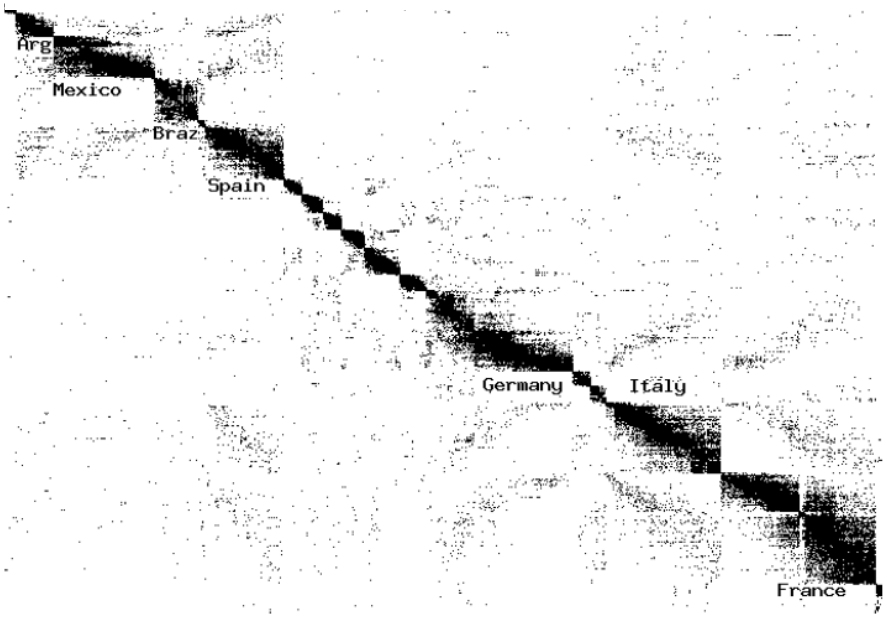
Goal: Find Densely Linked Clusters



Movies and Actors

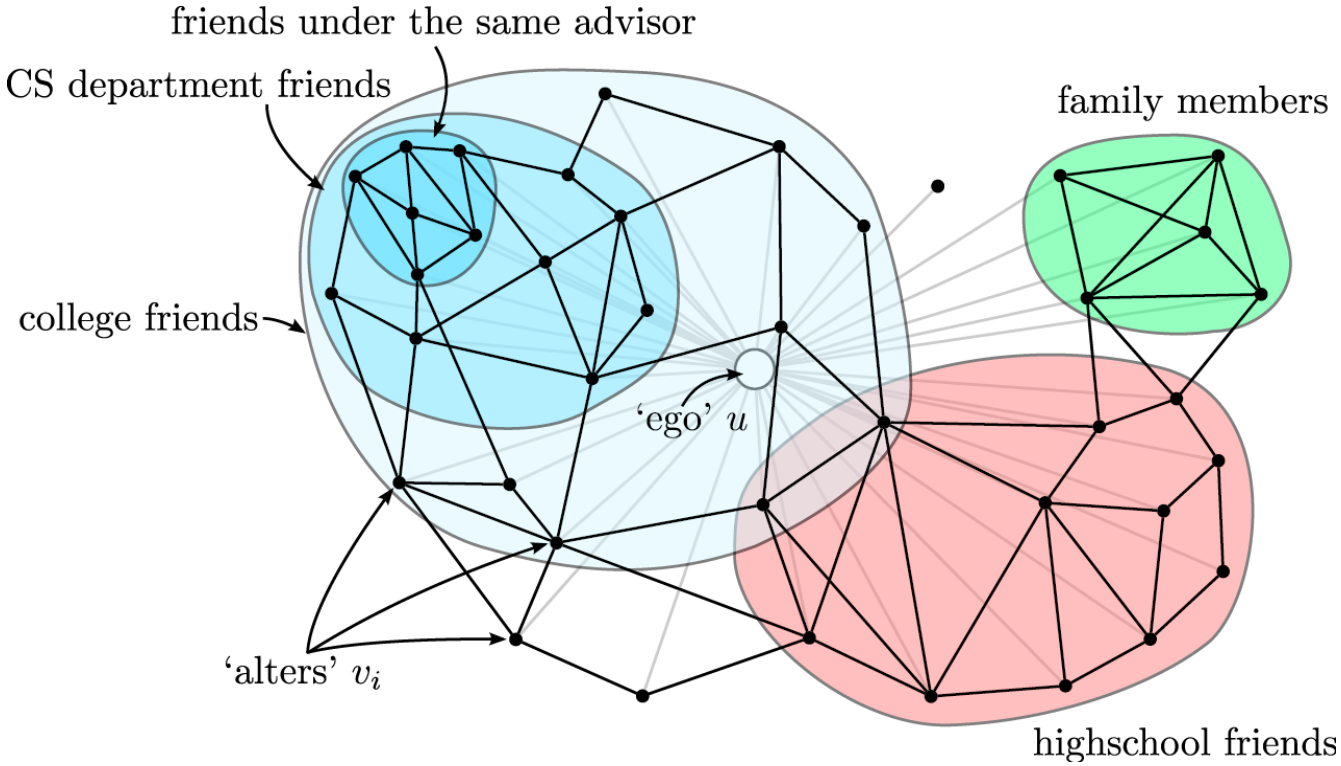
- **Clusters in Movies-to-Actors graph:**

[Andersen, Lang: Communities from seed sets, 2006]



Twitter & Facebook

- **Discovering social circles, circles of trust:**

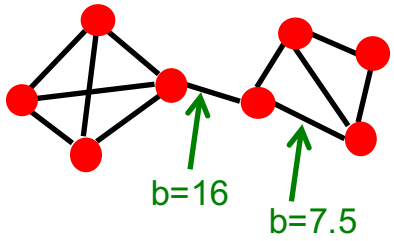


[McAuley, Leskovec: Discovering social circles in ego networks, 2012]

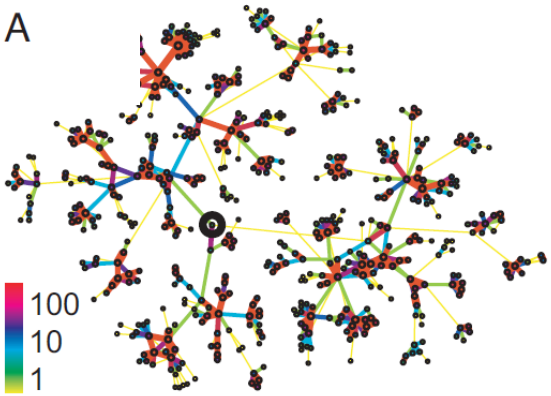
How to find communities?

Method 1: Strength of Weak Ties

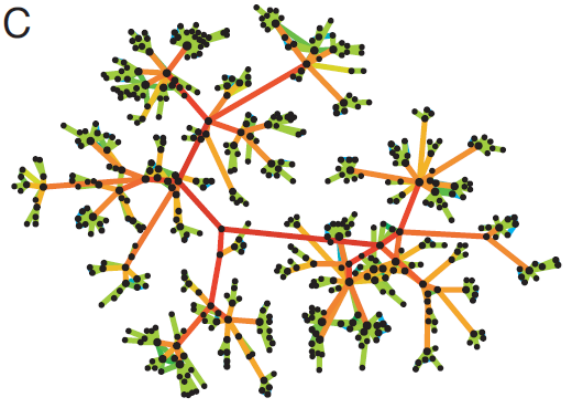
- **Edge betweenness:** Number of shortest paths passing over the edge



- **Intuition:**



Edge strengths (call volume) in a real network

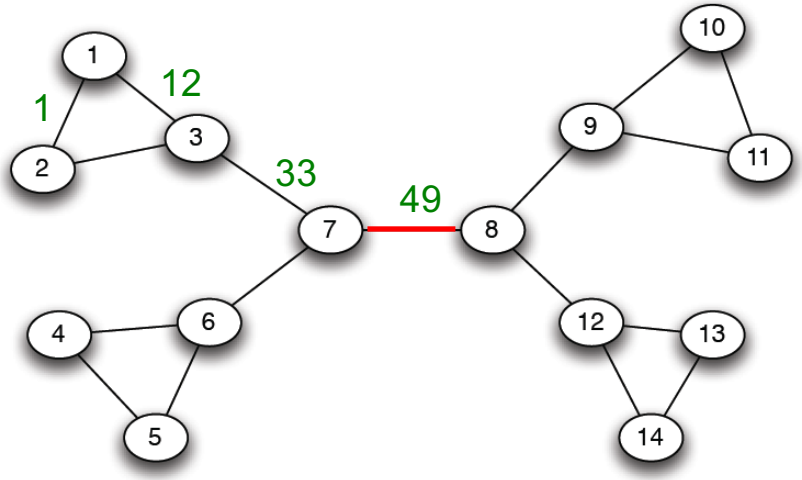


Edge betweenness in a real network

Method 1: Girvan-Newman

- Divisive hierarchical clustering based on the notion of edge **betweenness**:
  - **Number of shortest paths passing through the edge**
- **Girvan-Newman Algorithm:**
  - **Undirected unweighted networks**
  - **Repeat until no edges are left:**
    - Calculate betweenness of edges
    - Remove edges with highest betweenness
  - Connected components are communities
  - Gives a hierarchical decomposition of the network

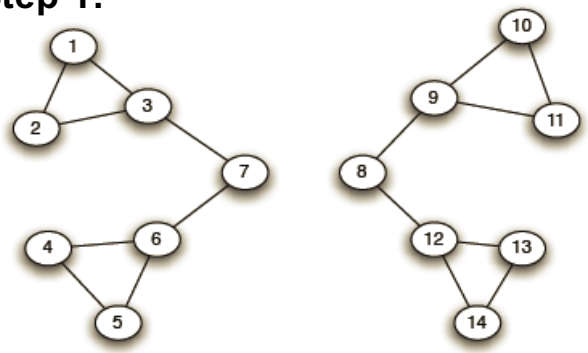
Girvan-Newman: Example



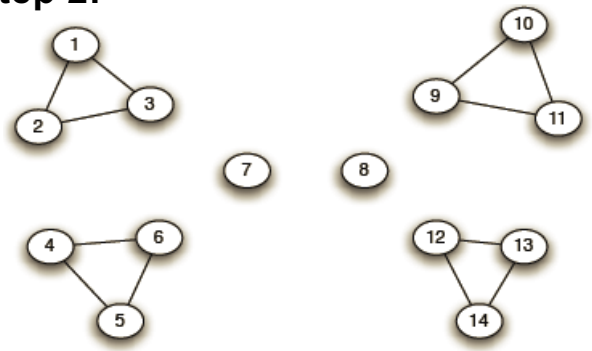
Need to re-compute betweenness at every step

Girvan-Newman: Example

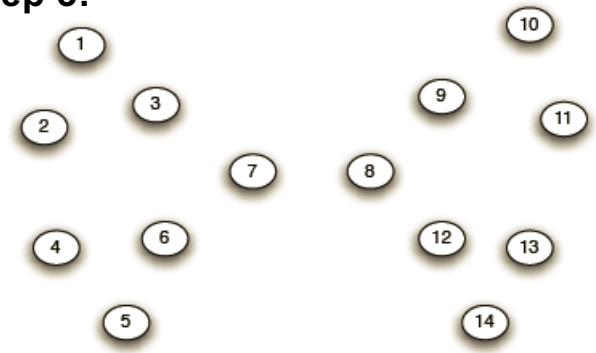
Step 1:



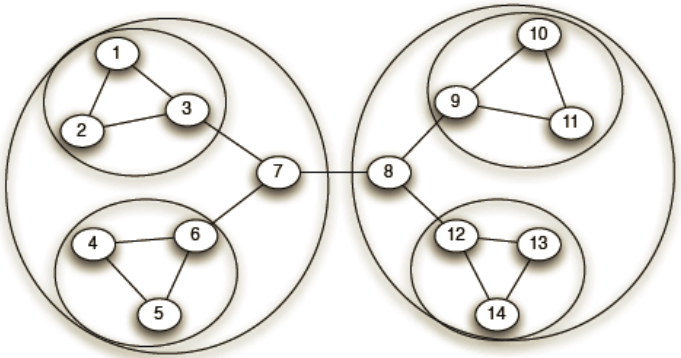
Step 2:



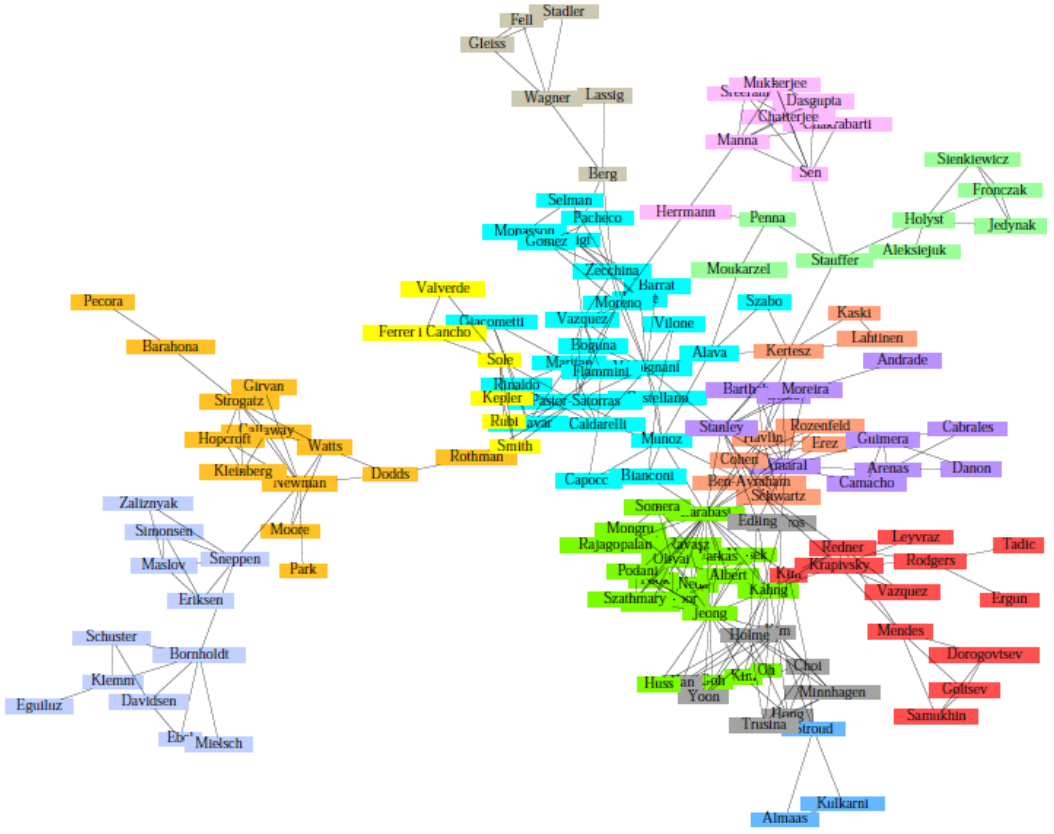
Step 3:



Hierarchical network decomposition:



Girvan-Newman: Results



Communities in physics collaborations

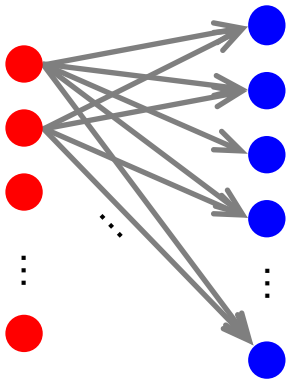
J. Leskovec, A. Rajaraman, J. Ullman: M

We need to resolve 2 questions

1. **How to compute betweenness?**
2. **How to select the number of clusters?**

Trawling

- **Searching for small communities in the Web graph**
- **What is the signature of a community / discussion in a Web graph?**



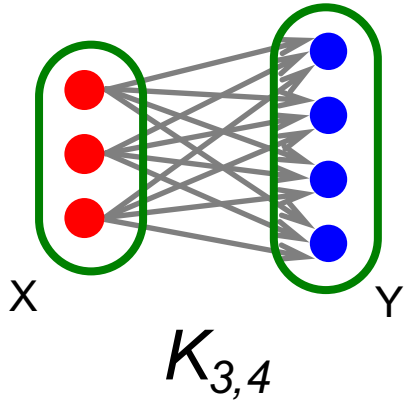
Dense 2-layer graph

Use this to define "topics":
What the same people on the left talk about on the right
Remember HITS!

Intuition: Many people all talking about the same things

Searching for Small Communities

- **A more well-defined problem:**
  Enumerate complete bipartite subgraphs $K_{s,t}$
  - Where $K_{s,t}$: s nodes on the "left" where each links to the same t other nodes on the "right"



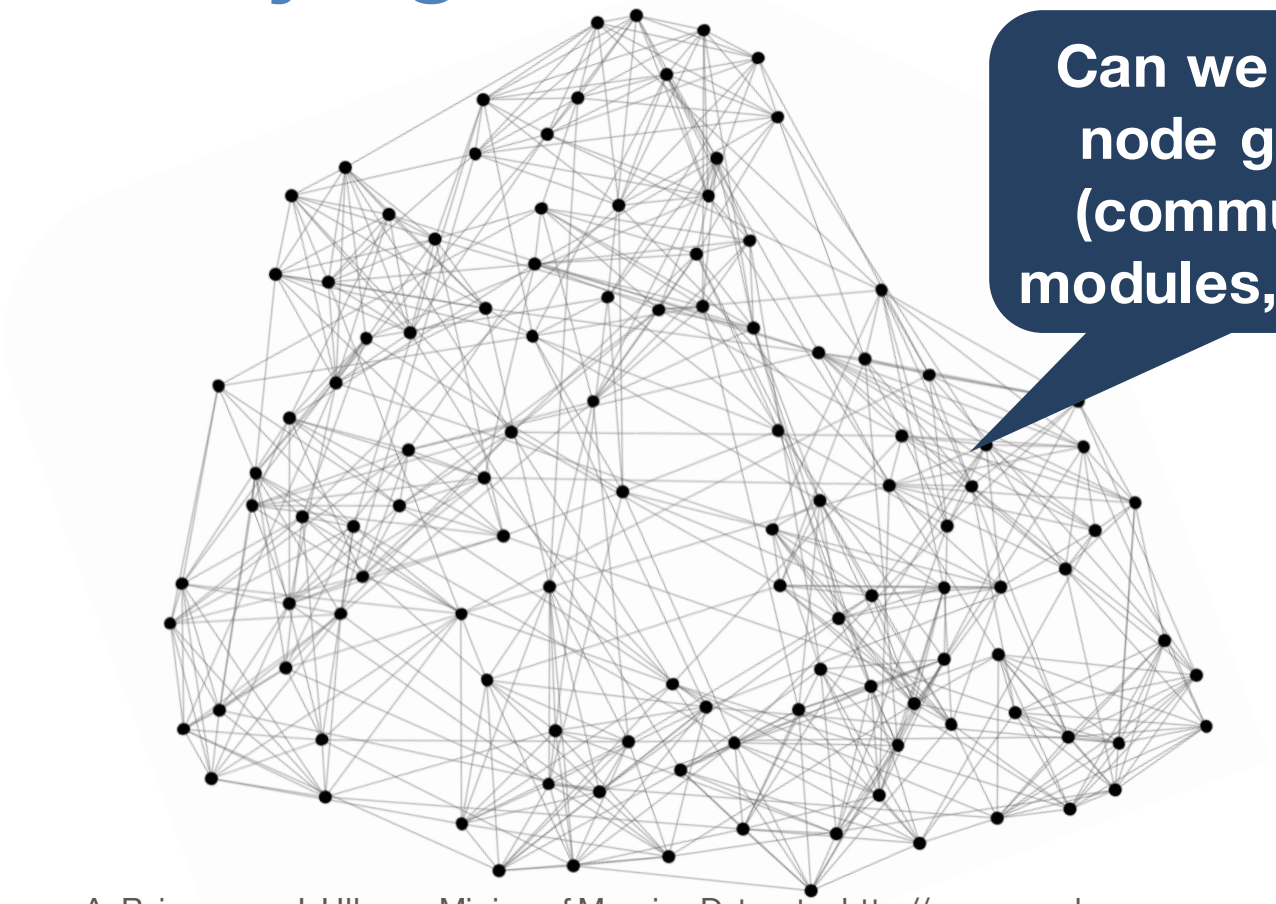
$K_{3,4}$

|X| = s = 3
|Y| = t = 4

Fully connected

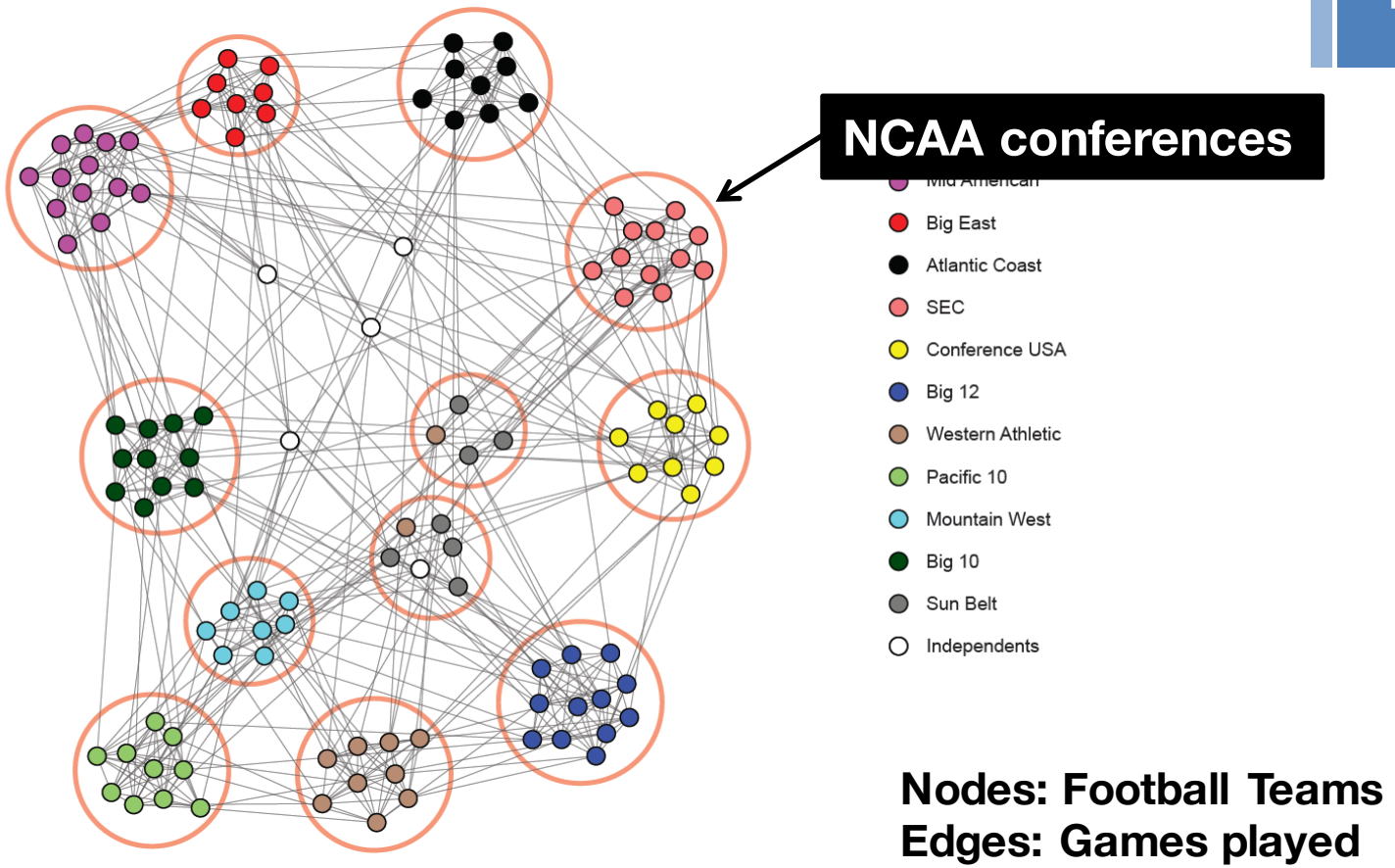
Overlapping communities

Identifying Communities



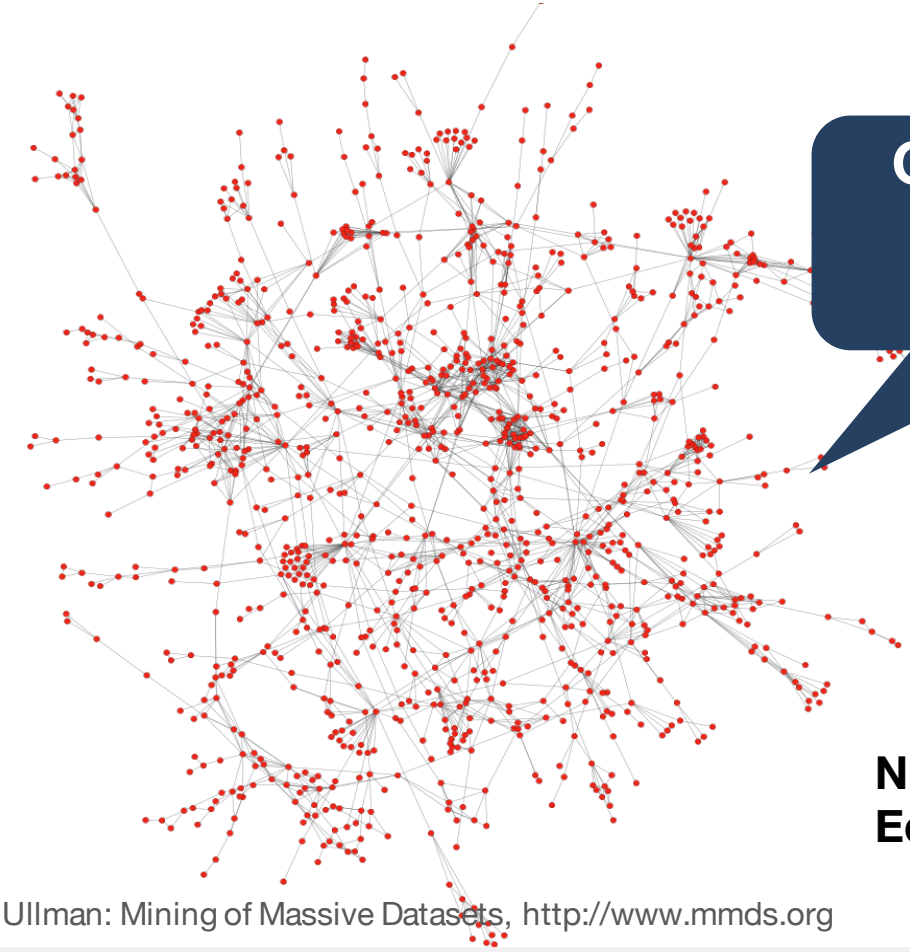
Nodes: Football Teams
Edges: Games played

NCAA Football Network



Nodes: Football Teams
Edges: Games played

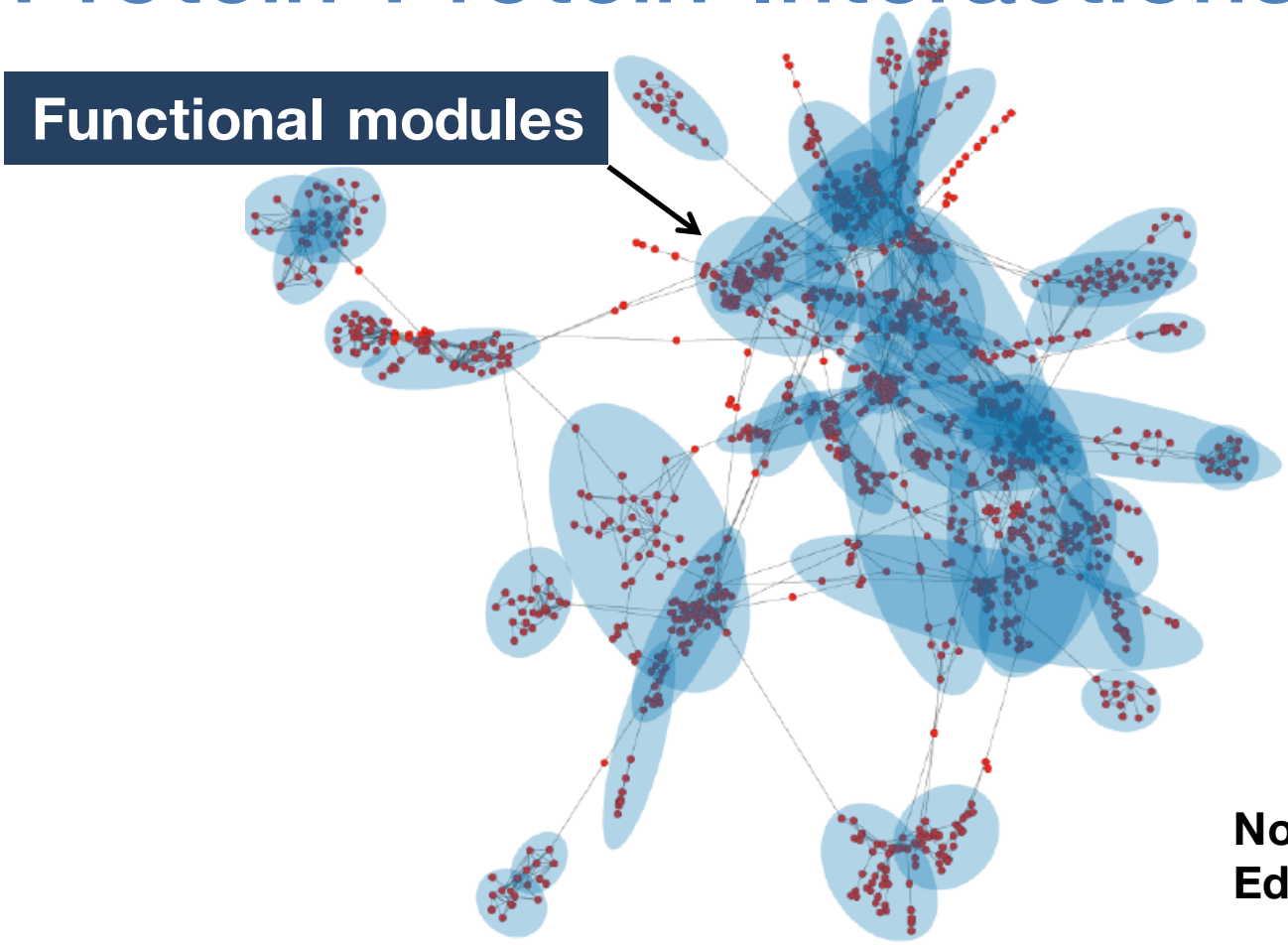
Protein-Protein Interactions



Can we identify functional modules?

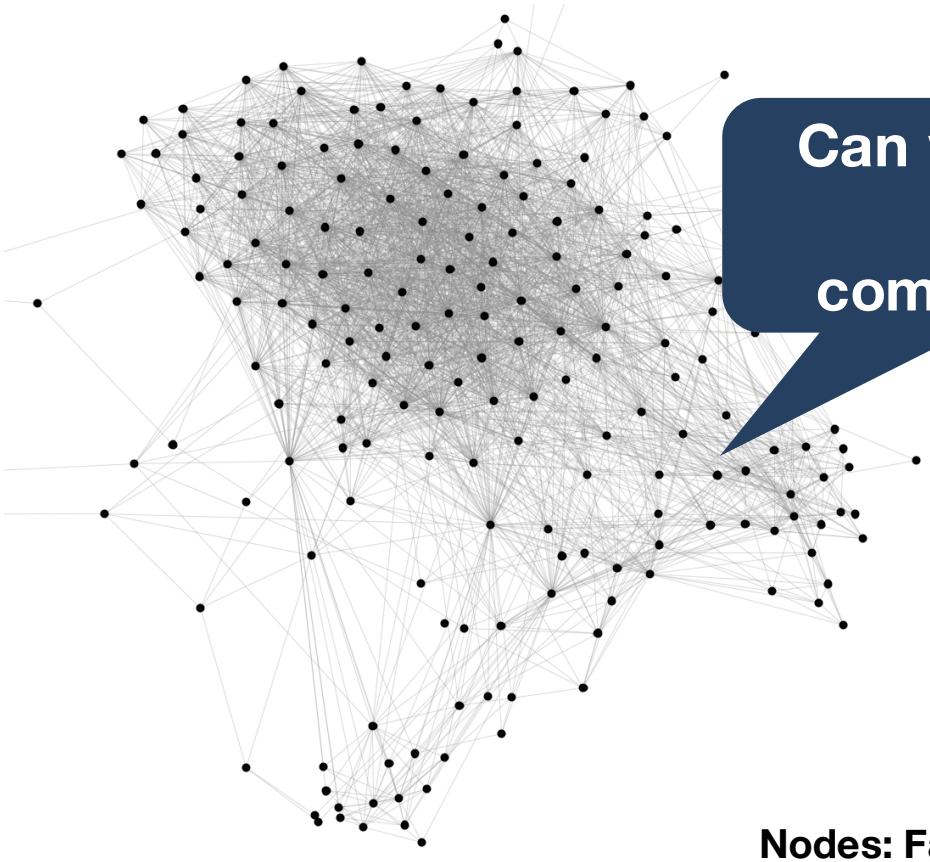
Nodes: Proteins
Edges: Physical interactions

Protein-Protein Interactions



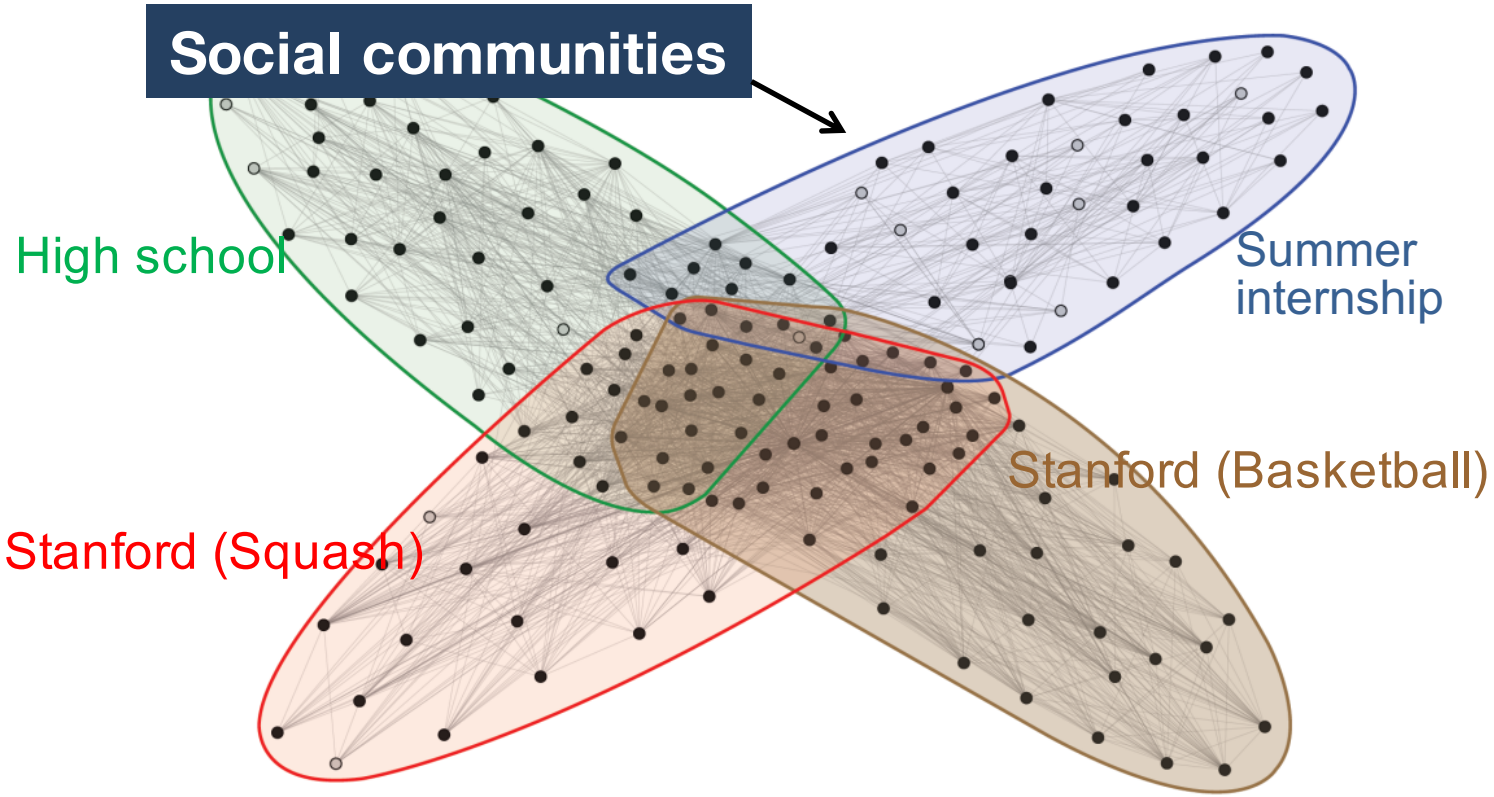
Nodes: Proteins
Edges: Physical interactions

Facebook Network



Nodes: Facebook Users
Edges: Friendships

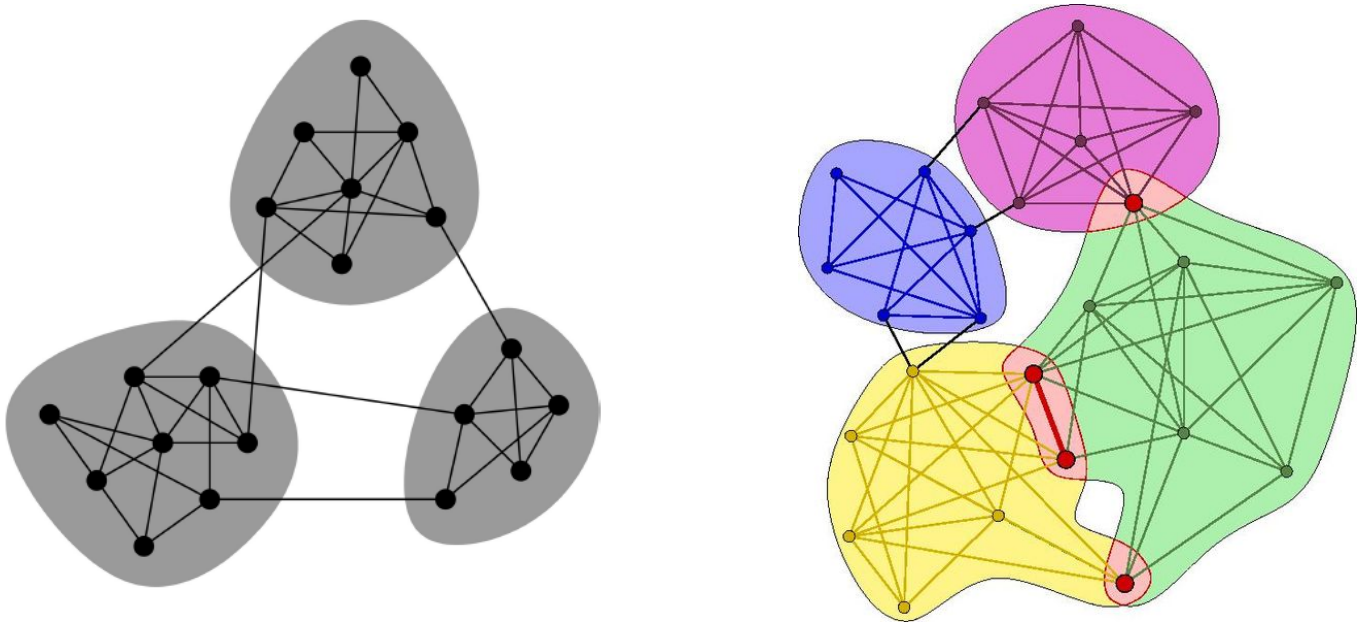
Facebook Network



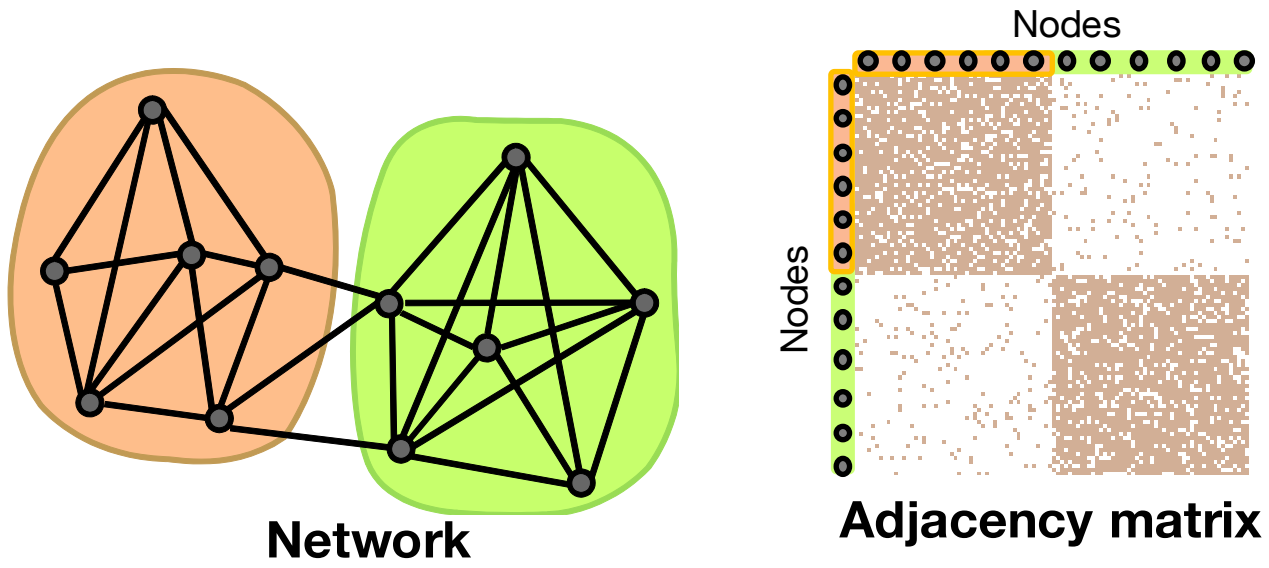
Nodes: Facebook Users
Edges: Friendships

Overlapping Communities

- **Non-overlapping vs. overlapping communities**

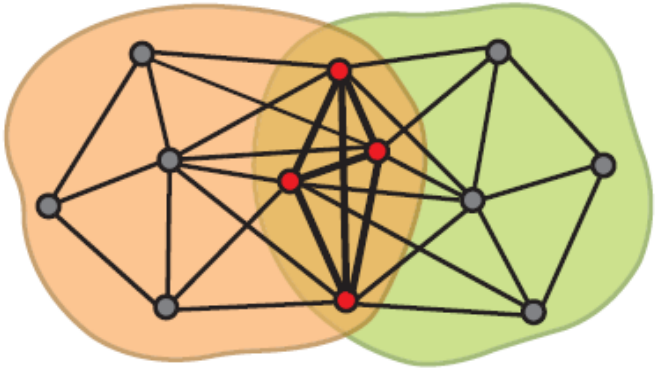


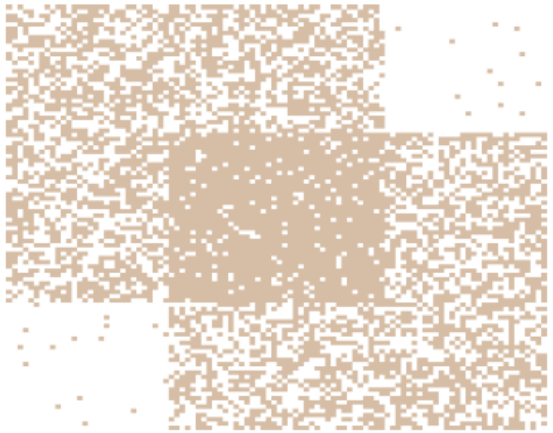
Non-overlapping Communities



Communities as Tiles!

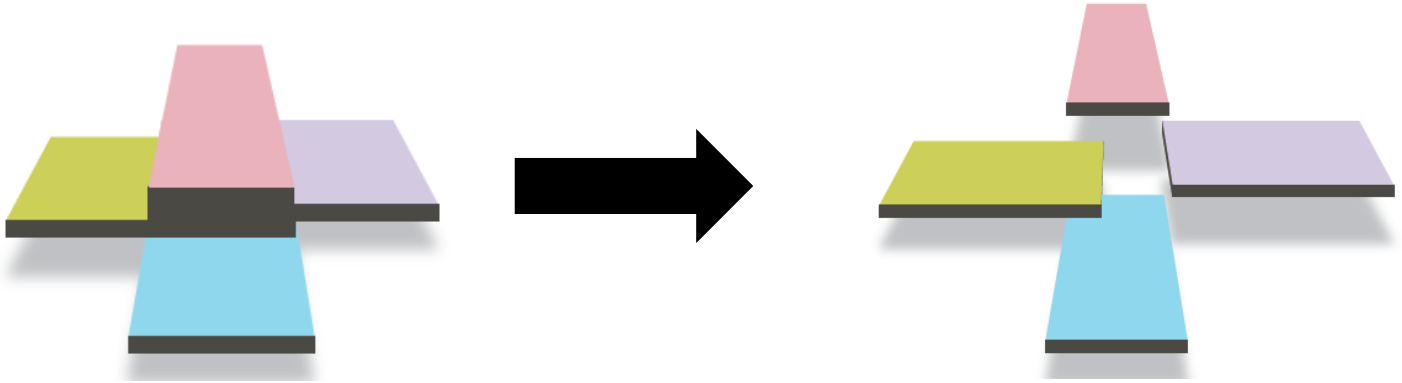
- **What is the structure of community overlaps:**
  Edge density in the overlaps is higher!





Communities as "tiles"

Recap so far…

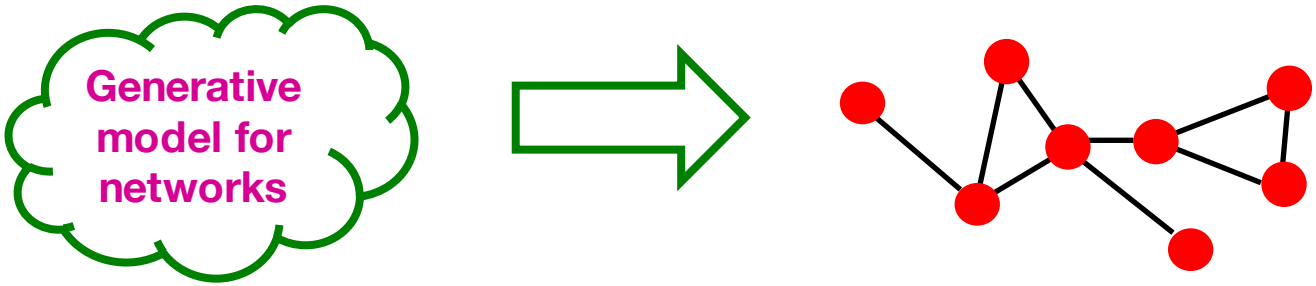


Communities in a network

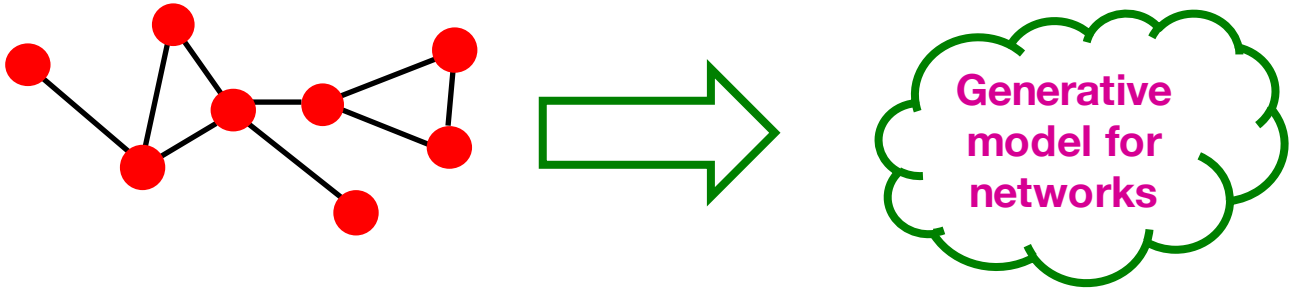
This is what we want!

Plan of attack

- **1)** Given a model, we generate the network:

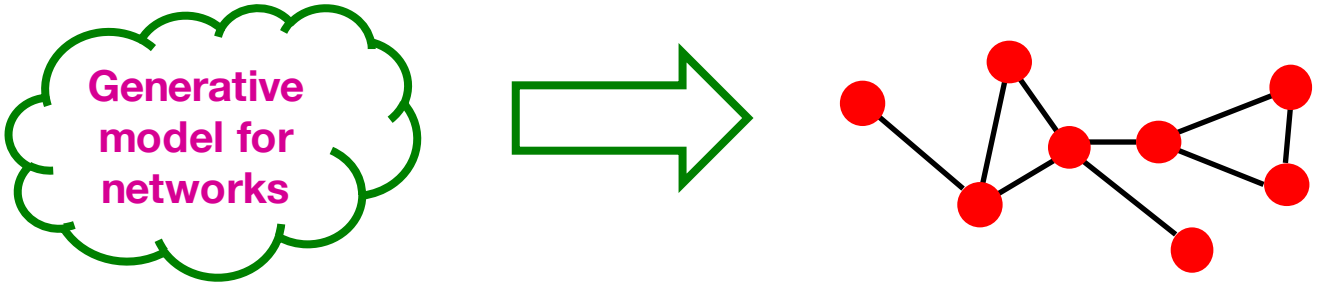


- **2)** Given a network, find the "best" model



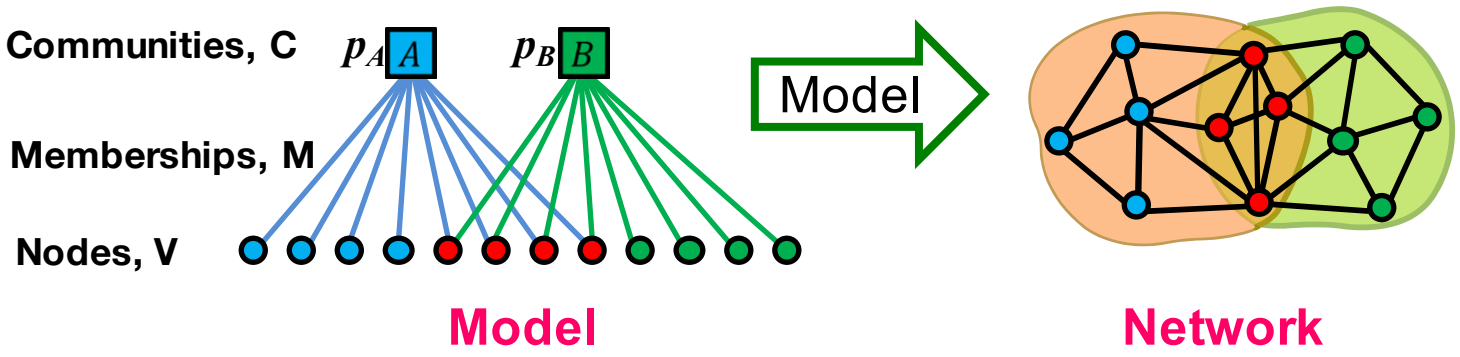
Model of networks

- **Goal: Define a model that can generate networks**
  - The model will have a set of "parameters" that we will later want to estimate (and detect communities)



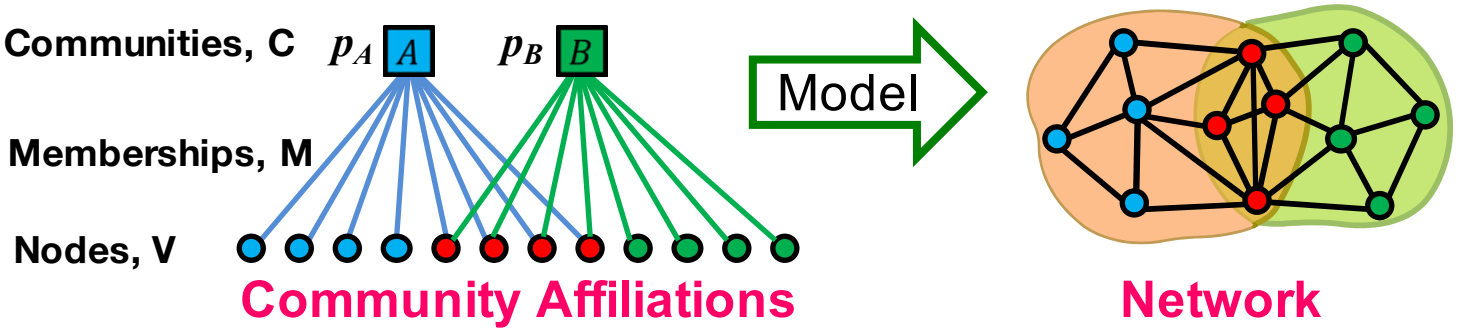
- **Q: Given a set of nodes, how do communities "generate" edges of the network?**

Community-Affiliation Graph



- **Generative model B(V, C, M, $\{p_c\}$) for graphs:**
  - Nodes **V**, Communities **C**, Memberships **M**
  - Each community **c** has a single probability $\mathbf{p_c}$
  - Later we fit the model to networks to detect communities

AGM: Generative Process



- **AGM generates the links: For each**
  - For each pair of nodes in community A, we connect them with prob. p_A
  - **The overall edge probability is:**

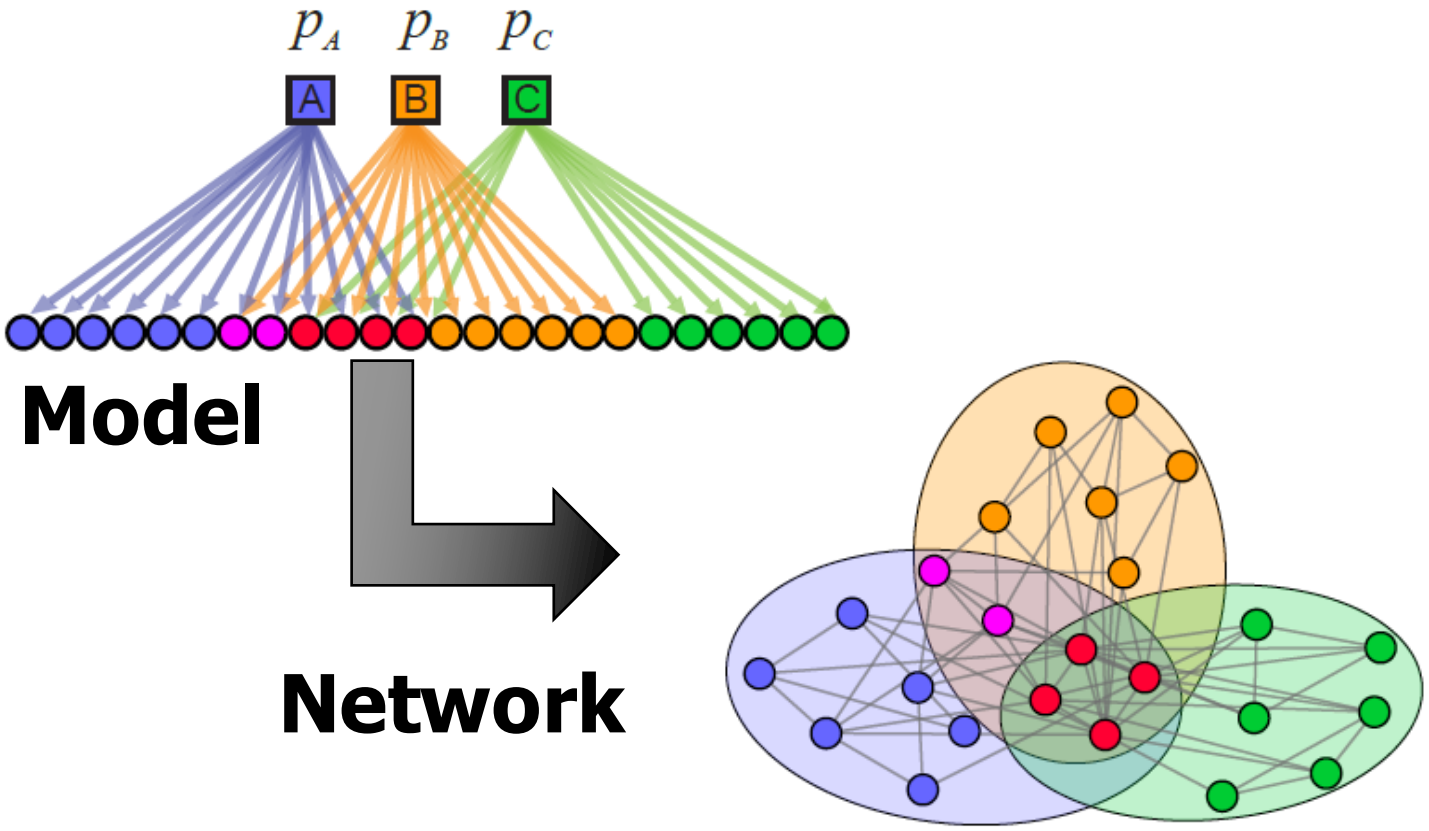
$$P(u,v) = 1 - \prod_{c \in M_u \cap M_v} (1 - p_c)$$

If u, v share no communities: $P(u,v) = \varepsilon$

M_u … set of communities node u belongs to

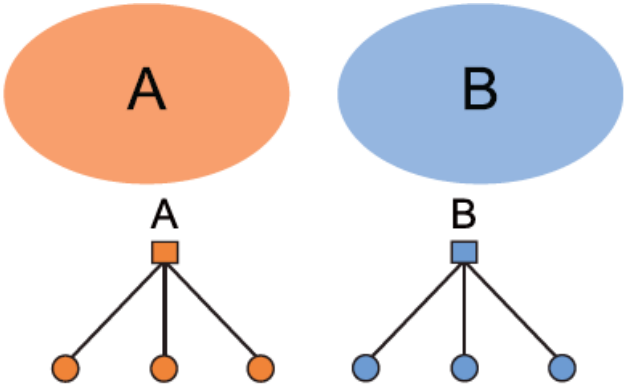
Think of this as an "OR" function: If at least 1 community says "YES" we create an edge

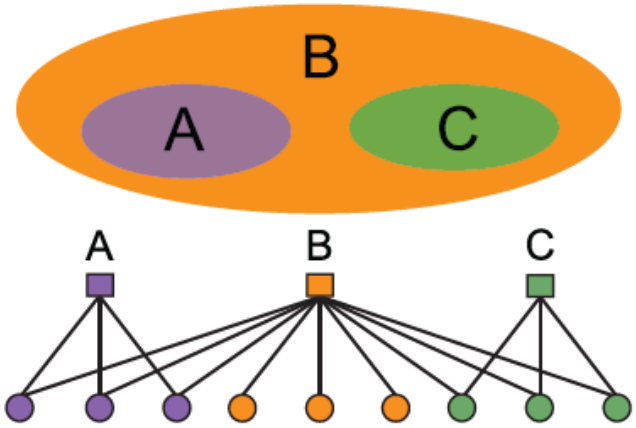
Recap: AGM networks

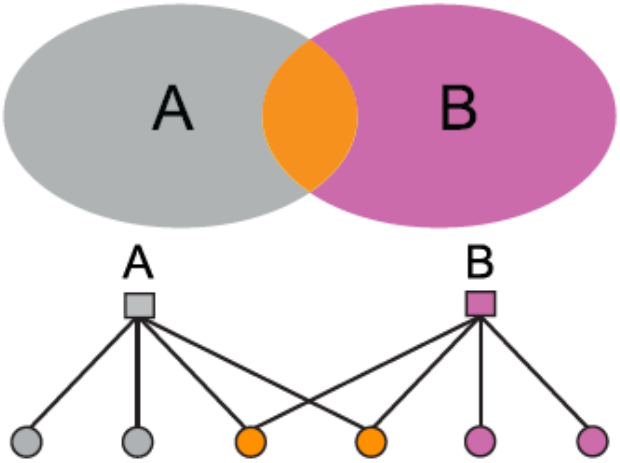


AGM: Flexibility

- **AGM can express a variety of community structures:**
  Non-overlapping, Overlapping, Nested







More details at…

- Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach by J. Yang, J. Leskovec. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2013.
- Detecting Cohesive and 2-mode Communities in Directed and Undirected Networks by J. Yang, J. McAuley, J. Leskovec. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2014.
- Community Detection in Networks with Node Attributes by J. Yang, J. McAuley, J. Leskovec. *IEEE International Conference On Data Mining (ICDM)*, 2013.

Let's go for it !