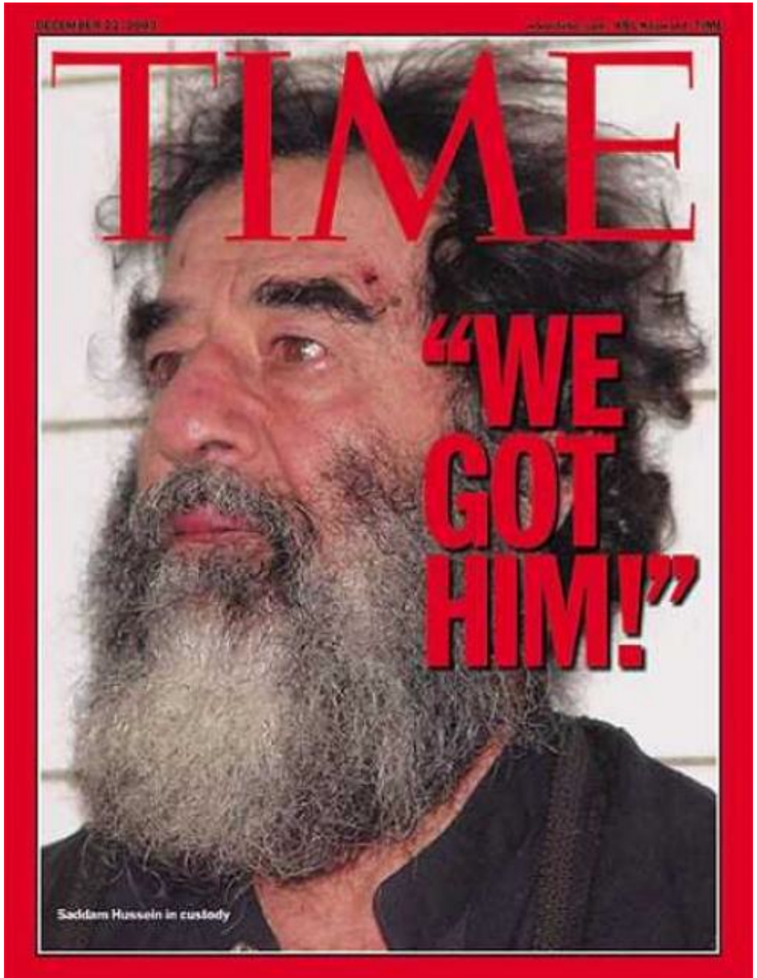
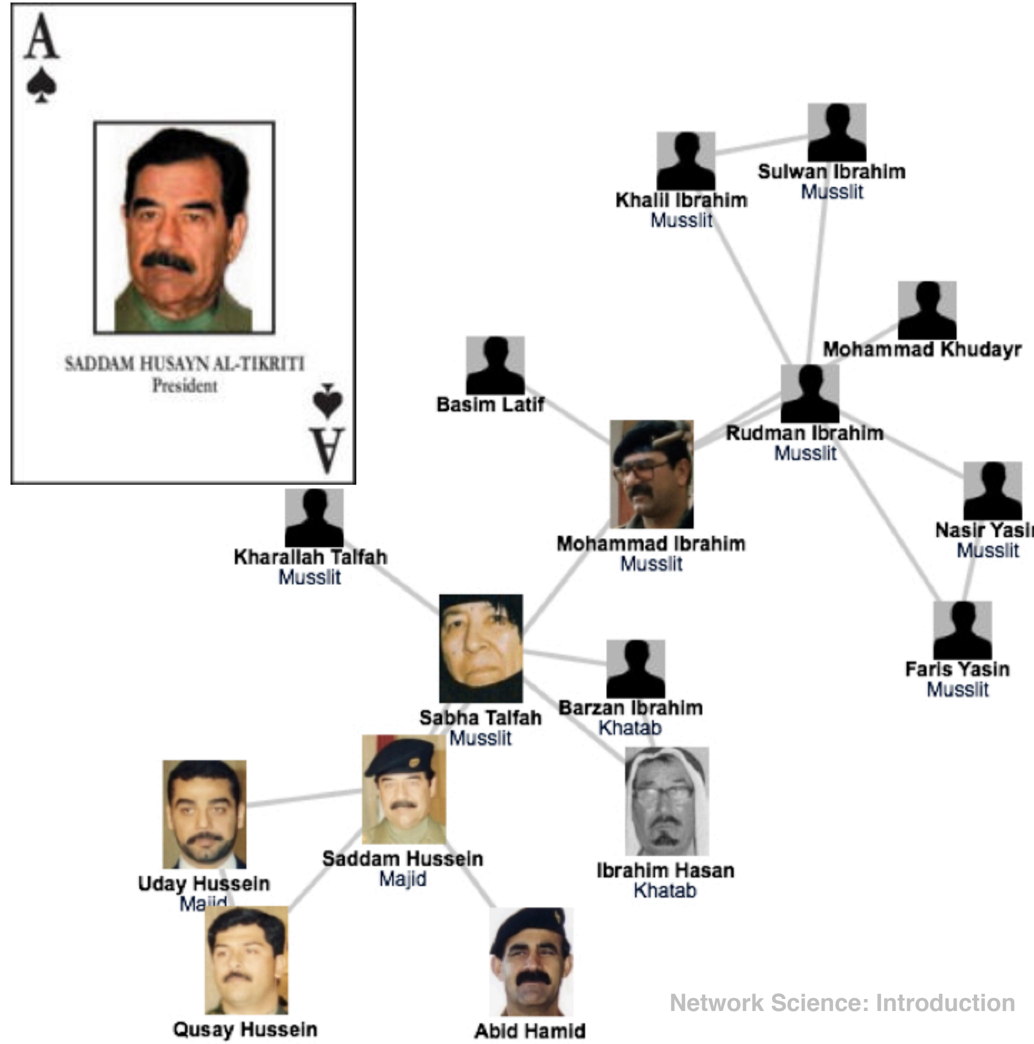
FROM SADDAM HUSSEIN TO NETWORK THEORY

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





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

The capture of Saddam Hussein:

→shows the strong predictive power of networks.

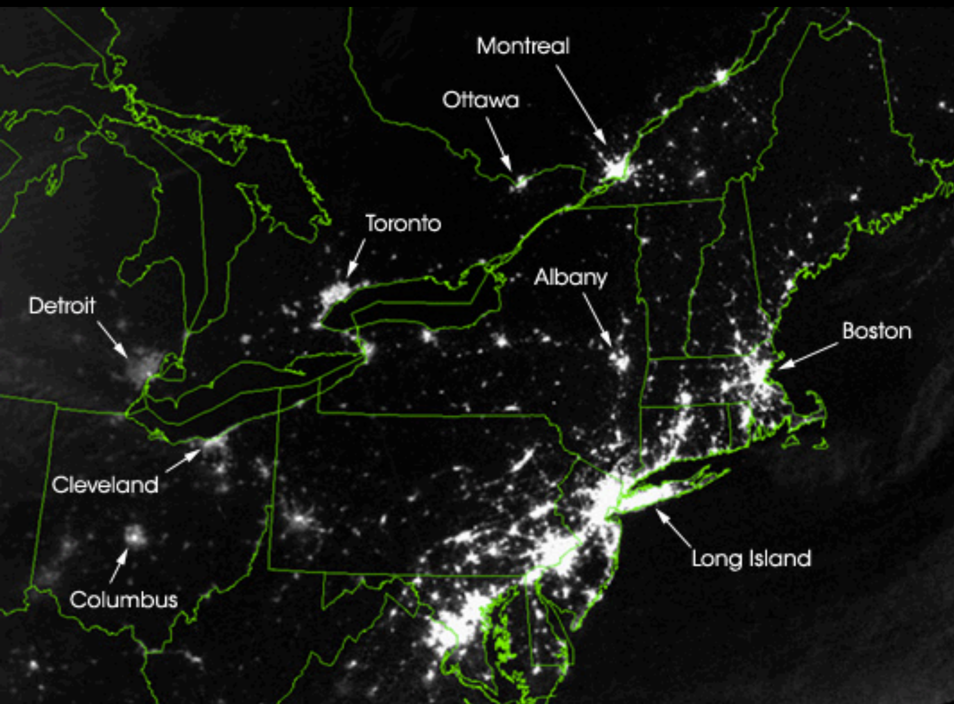
→ underlies the need to obtain accurate maps of the networks we aim to study; and the often heroic difficulties we encounter during the mapping process.

→ 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.

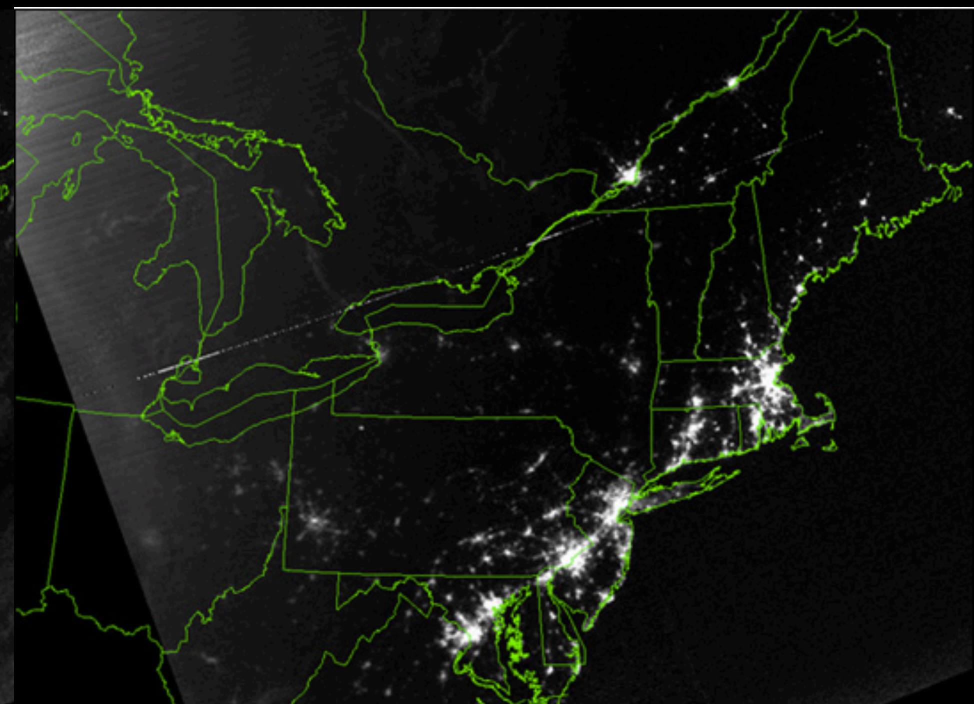
→ shows that the choice of network we focus on makes a huge difference: the hierarchical tree, that captured the official organization of the Iraqi government, was of no use when it came to Saddam Hussein's whereabouts.

VULNERABILITY DUE TO INTERCONNECTIVITY

A SIMPLE STORY (2): August 15, 2003 blackout.



August 14, 2003: 9:29pm EDT
20 hours before



August 15, 2003: 9:14pm EDT
7 hours after

A SIMPLE STORY (2): August 15, 2003 blackout.

An important theme of this class:

→ we must understand how network structure affects the robustness of a complex system.

→ develop quantitative tools to assess the interplay between network structure and the dynamical processes on the networks, and their impact on failures.

→We will learn that failures reality failures follow reproducible laws, that can be quantified and even predicted using the tools of network science.

NETWORKS AT THE HEART OF COMPLEX SYSTEMS

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
–adjective

1. composed of many interconnected parts; compound; composite: a complex highway system.
2. characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.
3. so complicated or intricate as to be hard to understand or deal with: a complex problem.

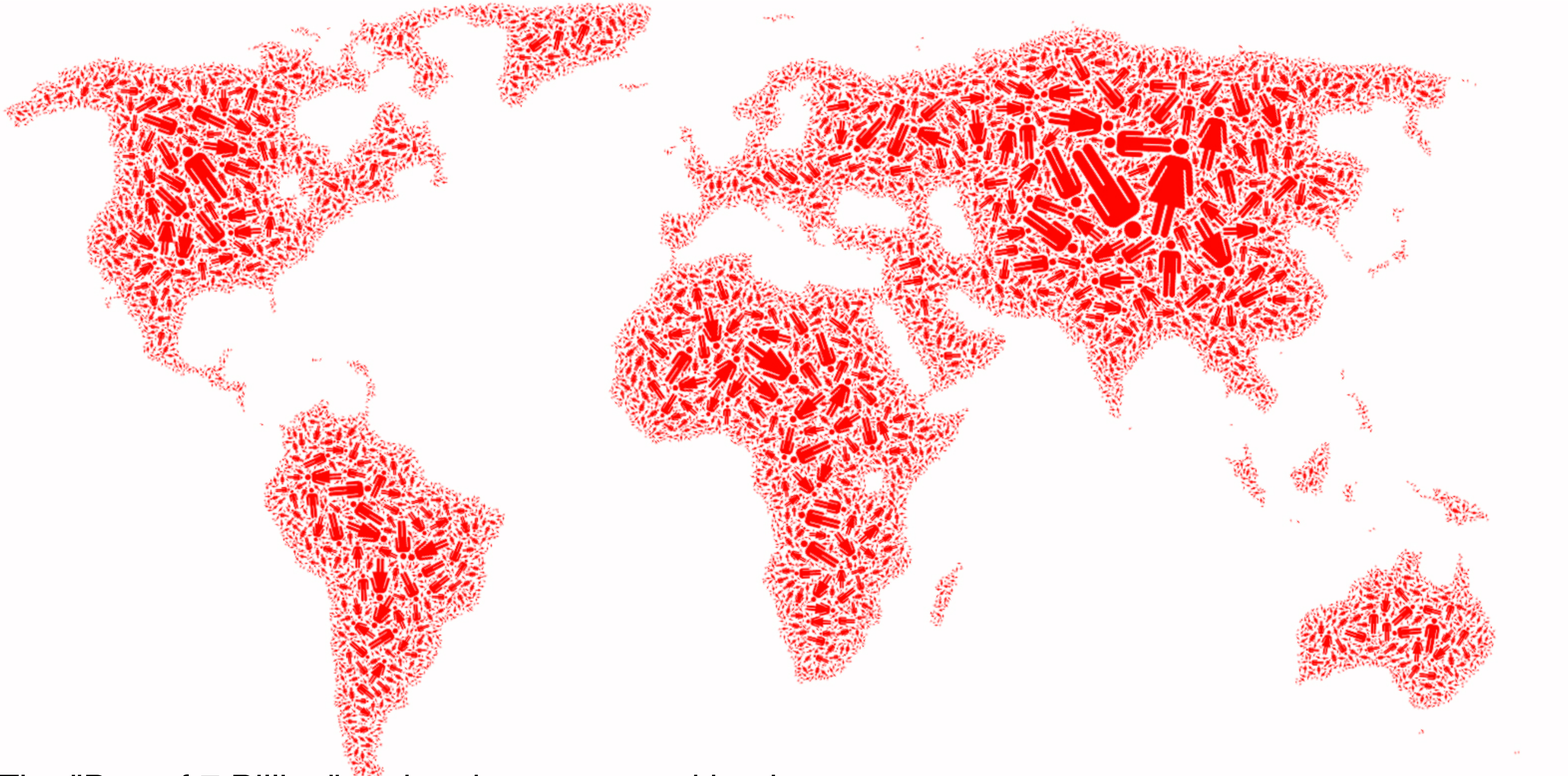
Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

Complexity

Factoid:



The "Day of 7 Billion" has been targeted by the United States Census Bureau to be in July 2012.

http://en.wikipedia.org/wiki/World_population

Human Brain has between 10-100 billion neurons.

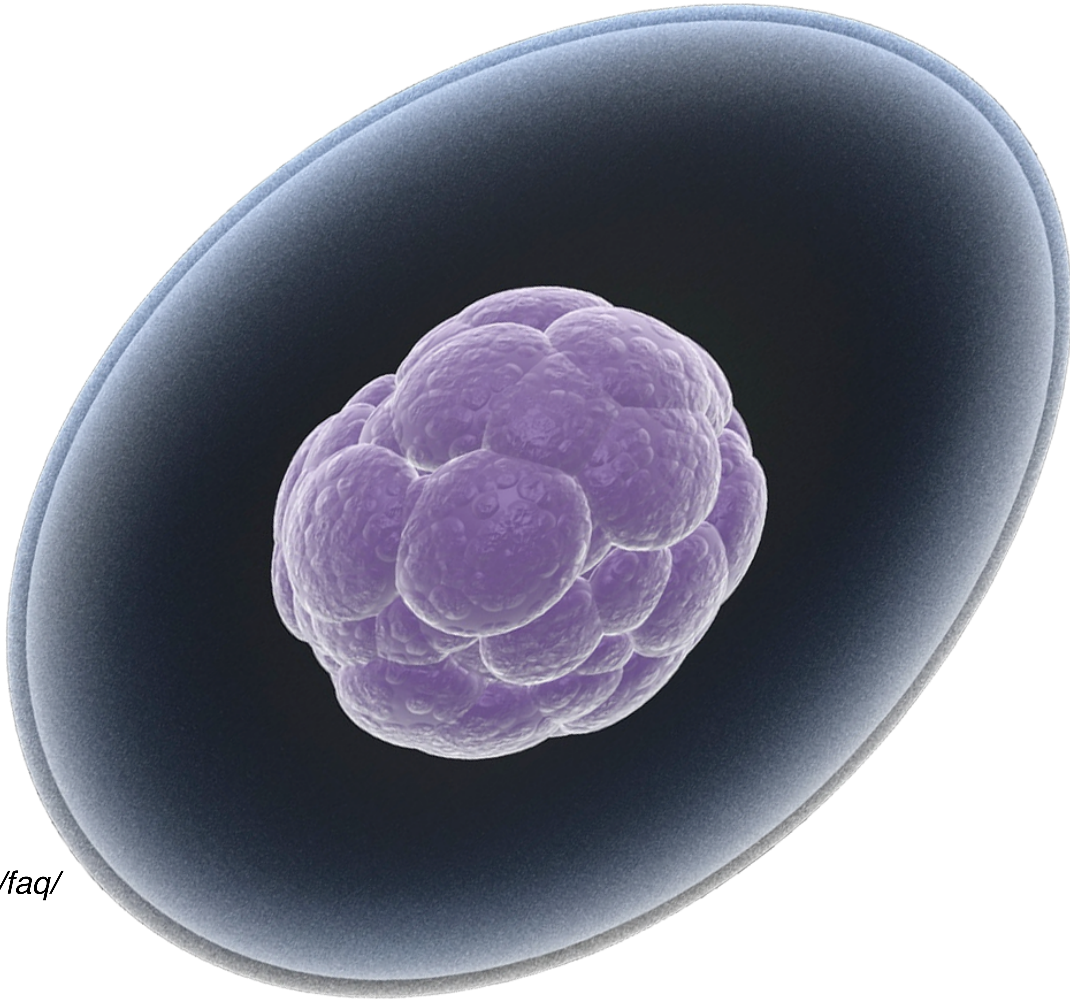
ECONOMY | **Factoid:**

The world economy produced goods and services worth almost $55 trillion in 2005.
(http://siteresources.worldbank.org/ICPINT/Resources/ICPreportprelim.pdf)

Factoid:

How Many Genes are in the Human Genome?

23,299



http://www.ornl.gov/sci/techresources/Human_Genome/faq/genenumber.shtml


"I think the next century will be the century of complexity."
Stephen Hawking
January 23, 2000`

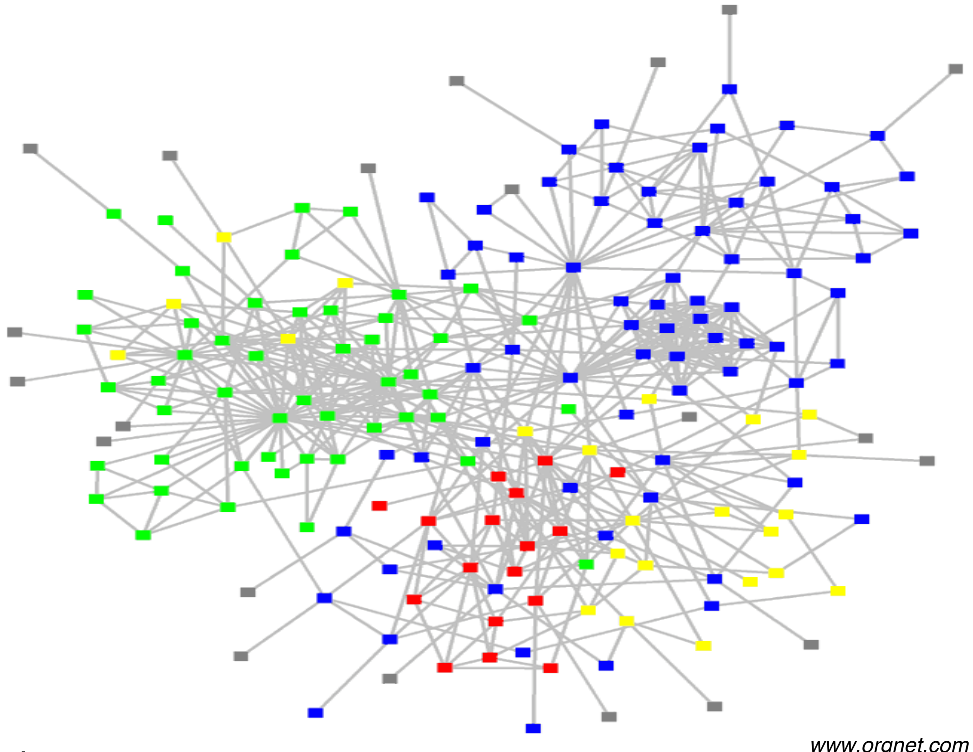
Behind each complex system there is a **network**, that defines the interactions between the component.

SOCIETY | Factoid:

The “Social Graph” behind Facebook

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

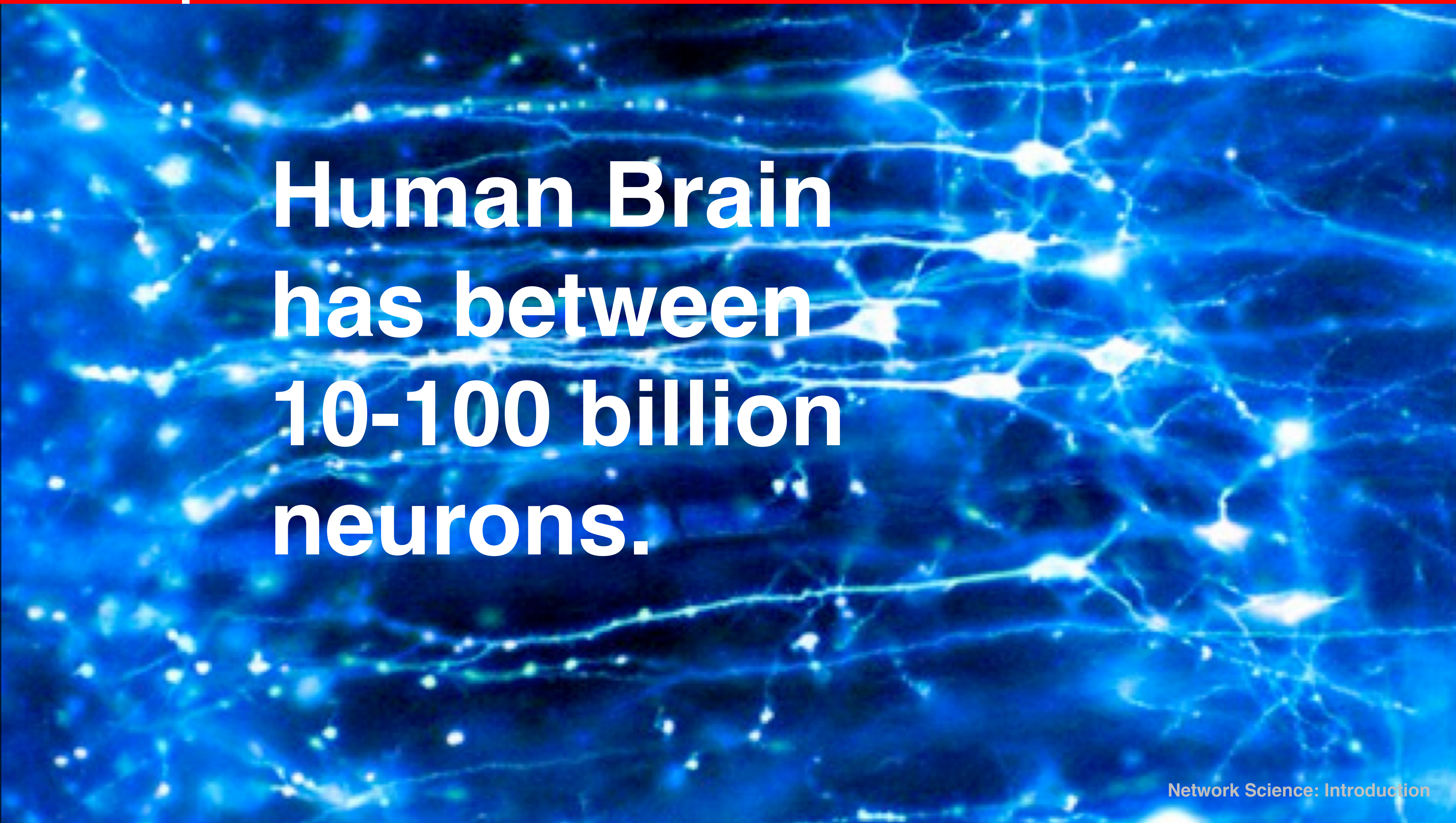
STRUCTURE OF AN ORGANIZATION



: departments

: consultants

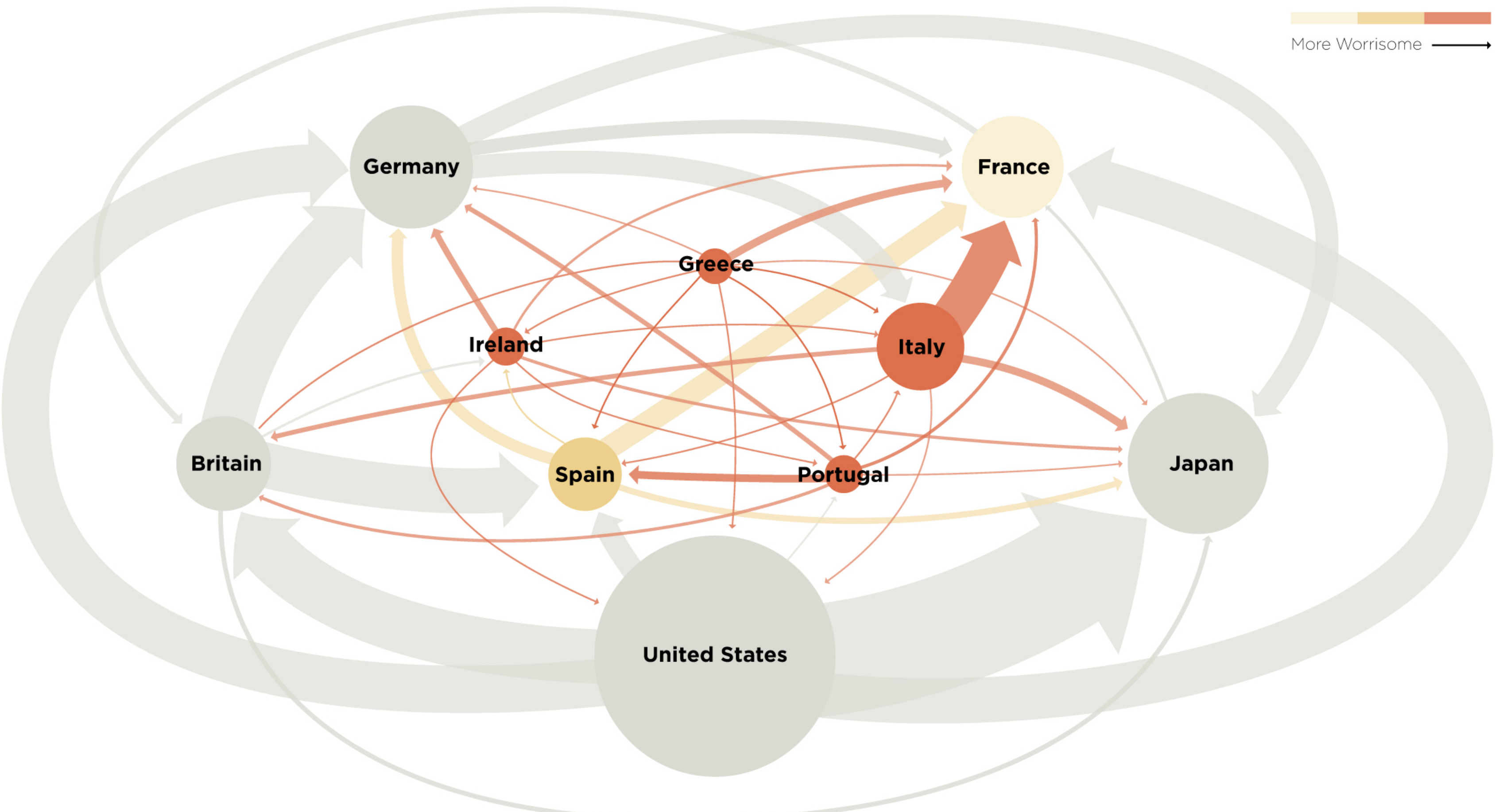
: external experts



Human Brain has between 10-100 billion neurons.

The subtle financial networks

The not so subtle financial networks: 2011



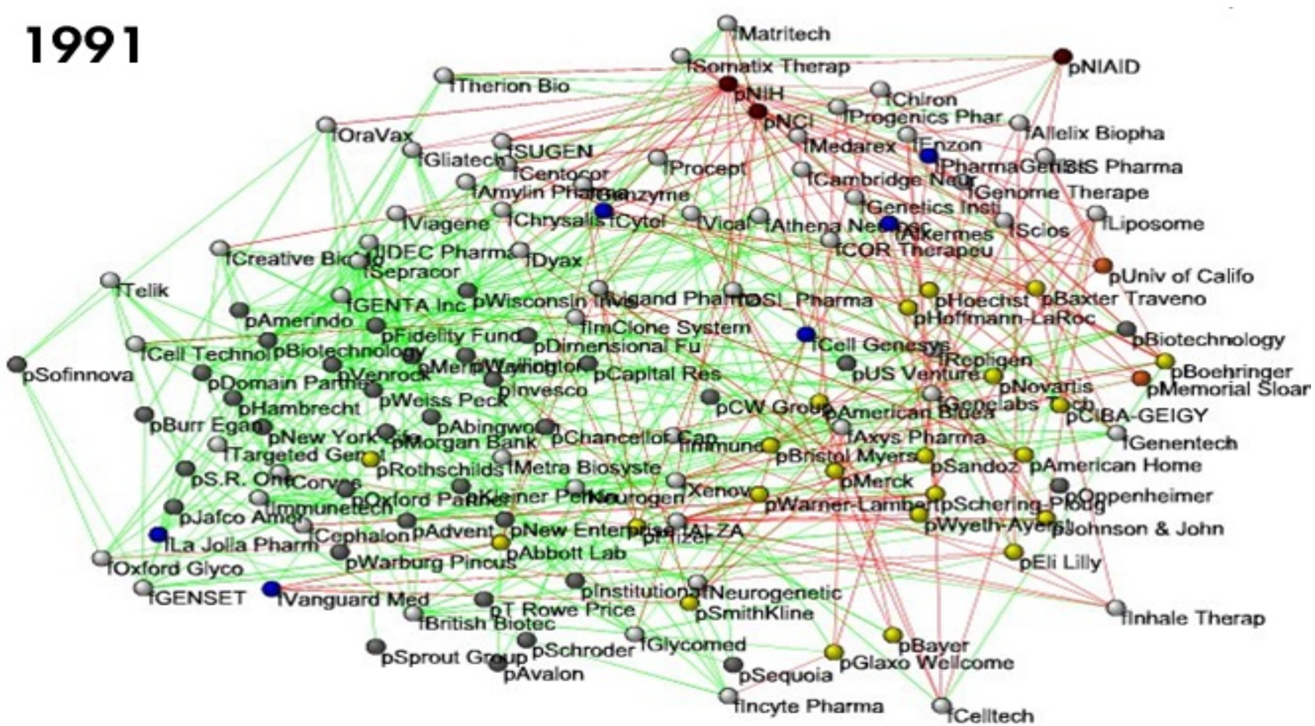
BUSINESS TIES IN US BIOTECH-INDUSTRY

Nodes:

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

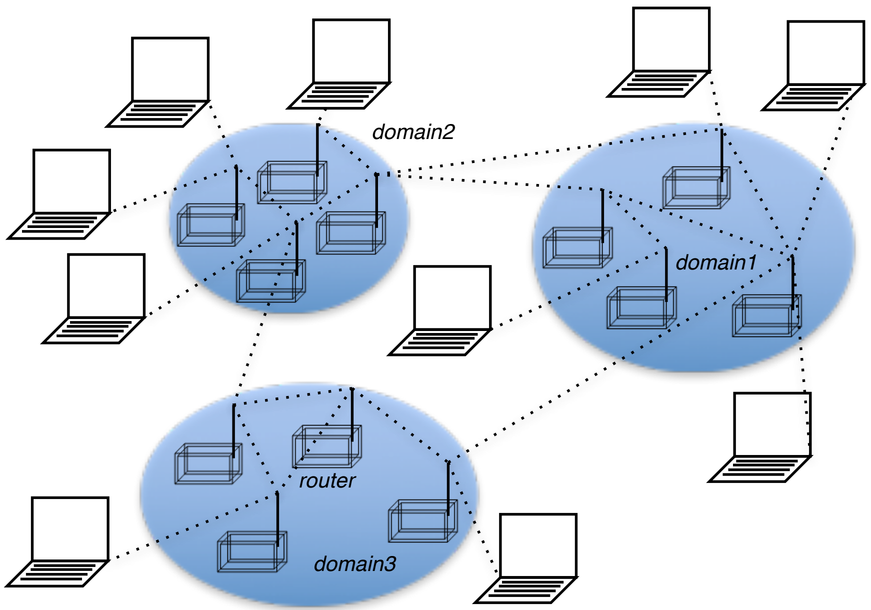
Links:

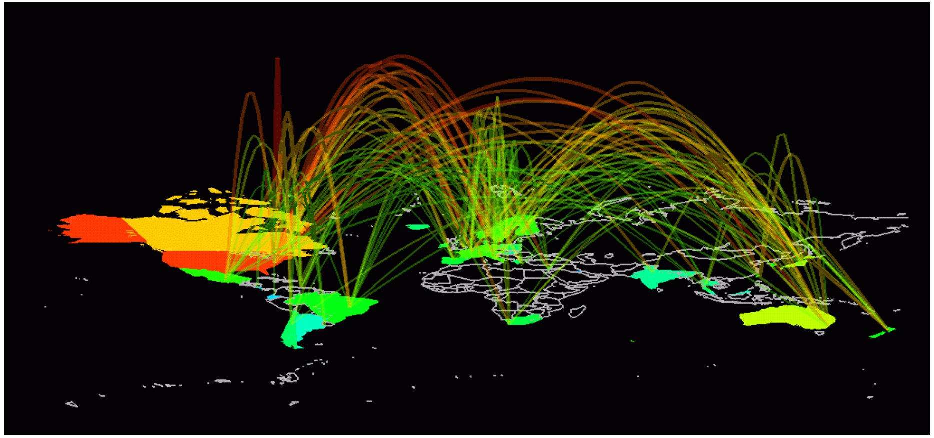
- Collaborations
- Financial
- R&D

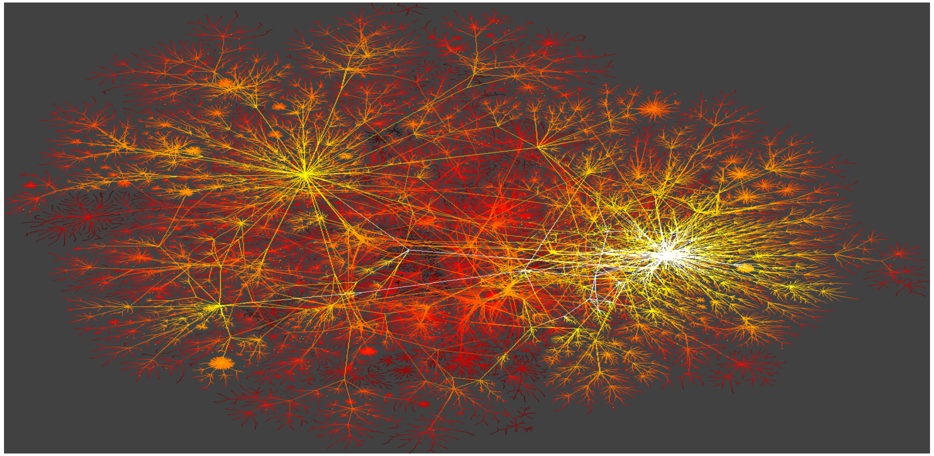


http://ecclectic.ss.uci.edu/~drwhite/Movie

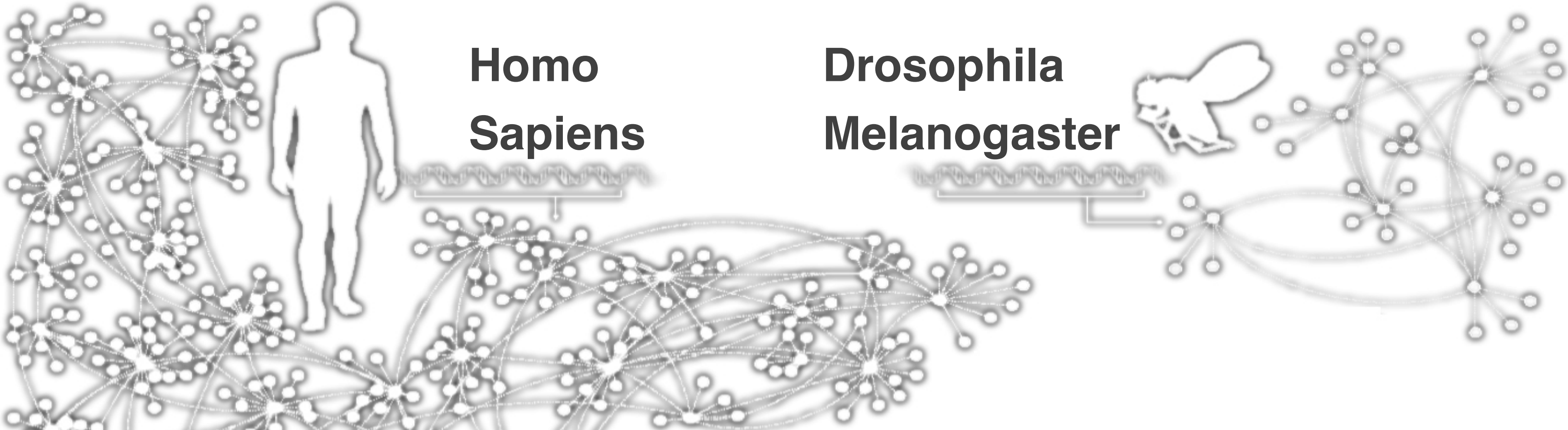
INTERNET

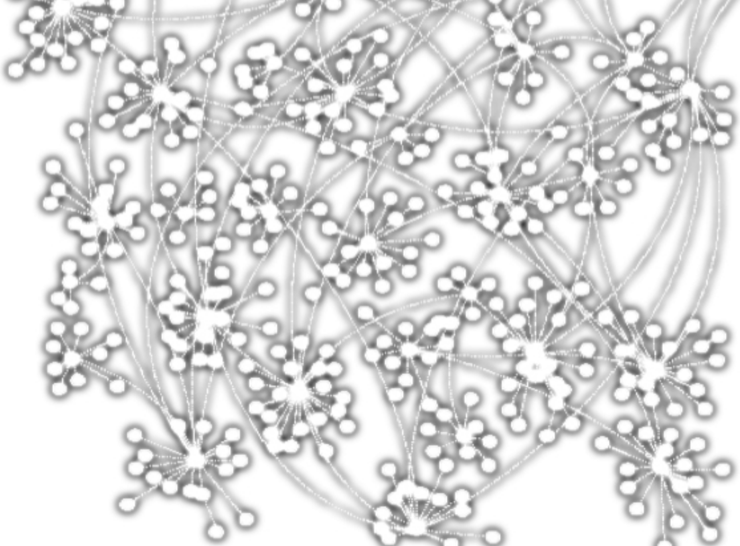






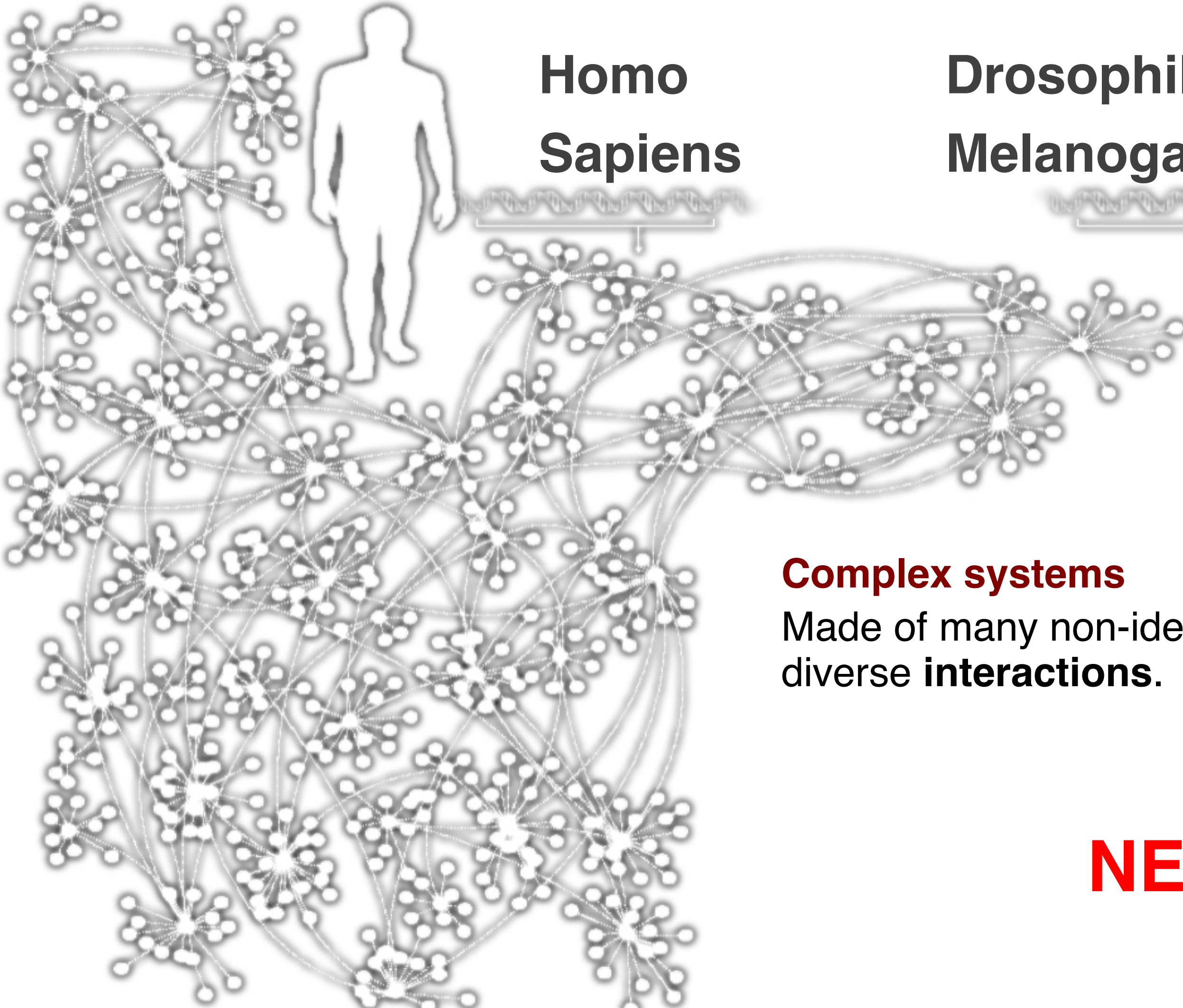
HUMANS GENES





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



Homo Sapiens

Drosophila Melanogaster

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.

TWO FORCES HELPED THE EMERGENCE OF NETWORK SCIENCE

THE HISTORY OF NETWORK ANALYSIS

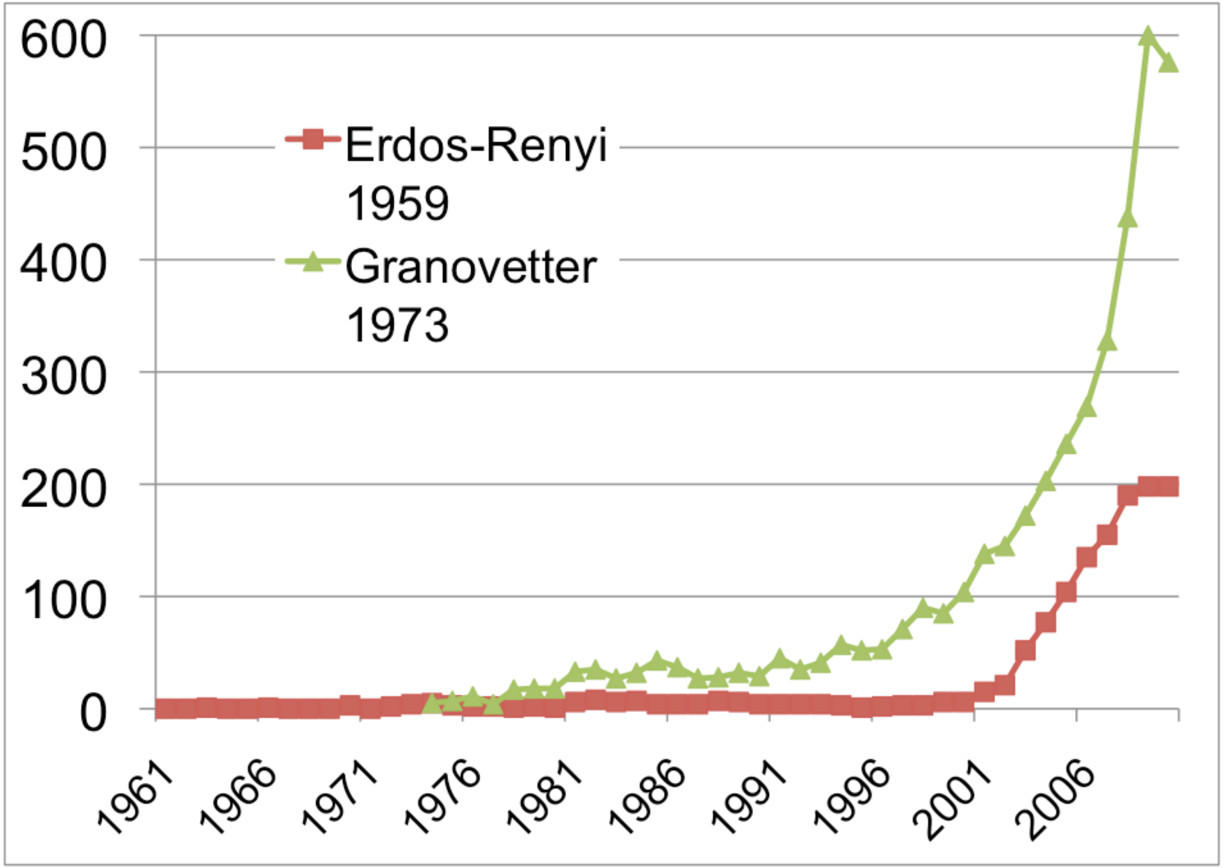
Graph theory: 1735, Euler

Social Network Research: 1930s, Moreno

Communication networks/internet: 1960s

Ecological Networks: May, 1979.

THE HISTORY OF NETWORK ANALYSIS



The emergence of network maps:

Movie Actor Network, 1998;
World Wide Web, 1999.
C elegans neural wiring diagram 1990
Citation Network, 1998
Metabolic Network, 2000;
PPI network, 2001

The universality of network characteristics:

The architecture of networks emerging in various domains of science, nature, and technology are more similar to each other than one would have expected.

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

Interdisciplinary

Empirical

Quantitative and Mathematical

Computational

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

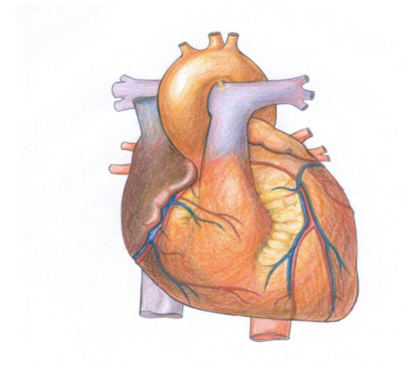
Facebook
market cap:
$50 billion

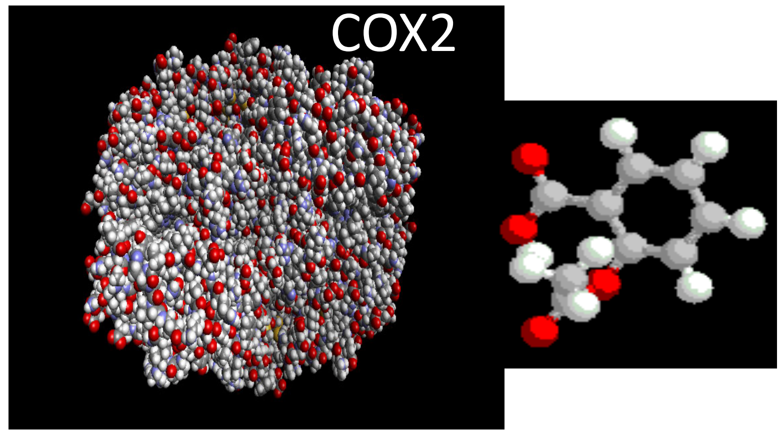
www.bizjournals.com/austin/news/2010/11/15/facebooks... - Cached

DRUG DESIGN, METABOLIC ENGINEERING:

Reduces
Inflammation
Fever
Pain

Prevents
Heart attack
Stroke



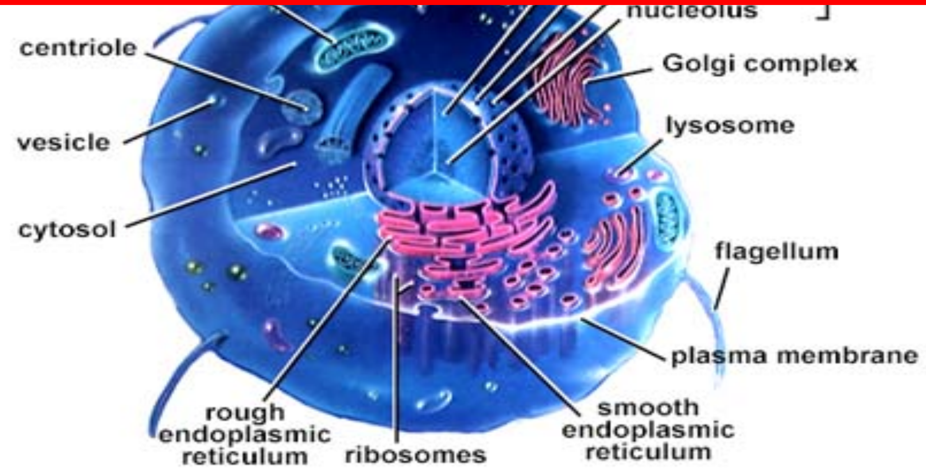


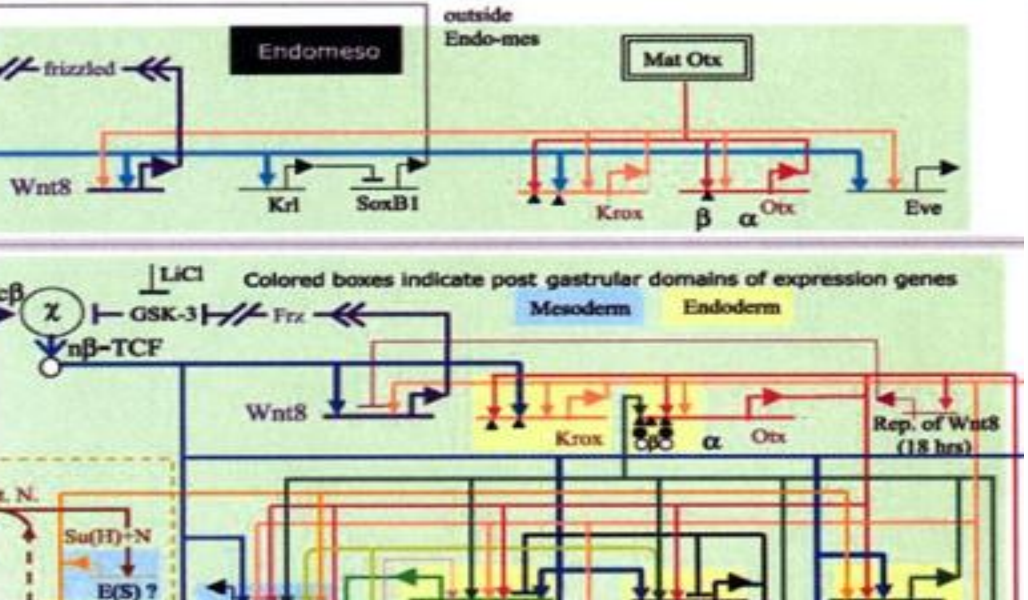
Reduces the risk of
Alzheimer's Disease

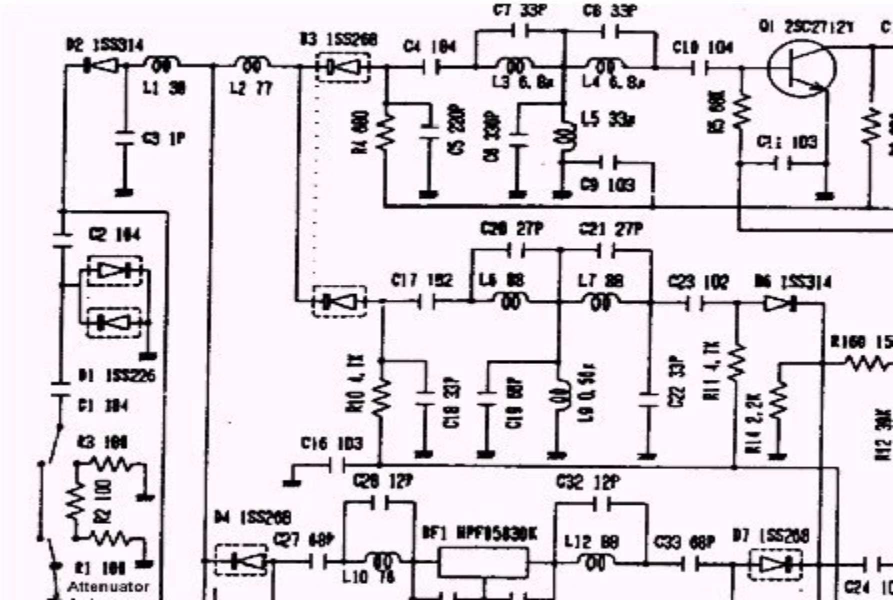
Reduces the risk of
breast cancer
ovarian cancers
colorectal cancer

Causes
Bleeding
Ulcer

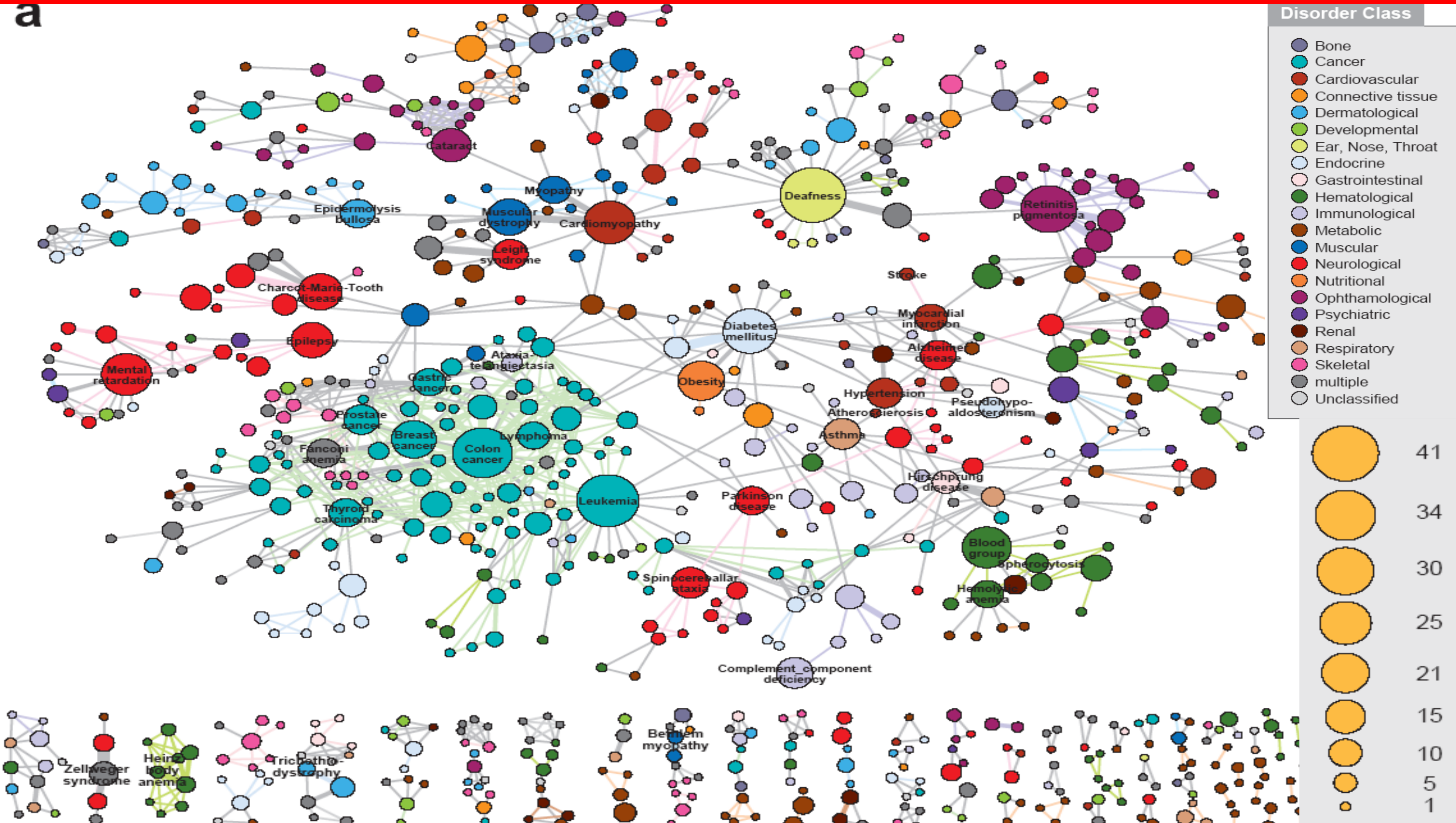
DRUG DESIGN, METABOLIC ENGINEERING:





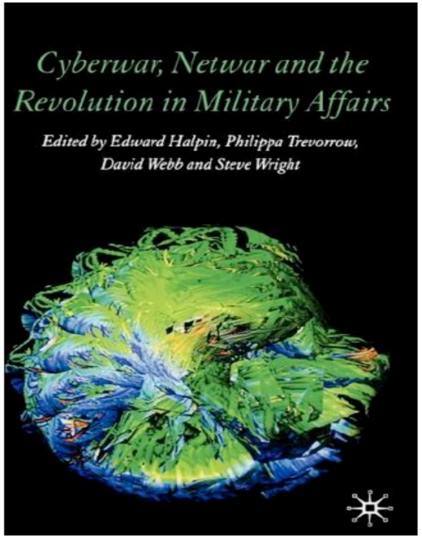


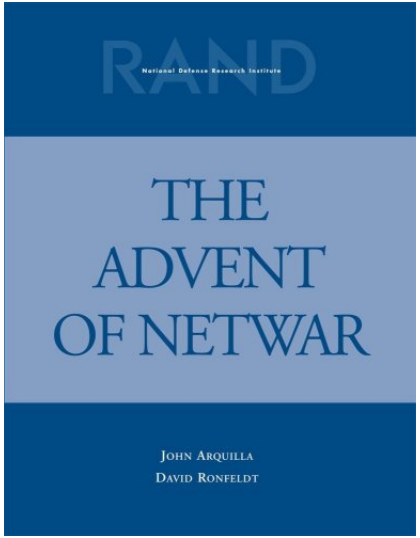
HUMAN DISEASE NETWORK



Network Biology/Network Medicine

FIGHTING TERRORISM AND MILITARY





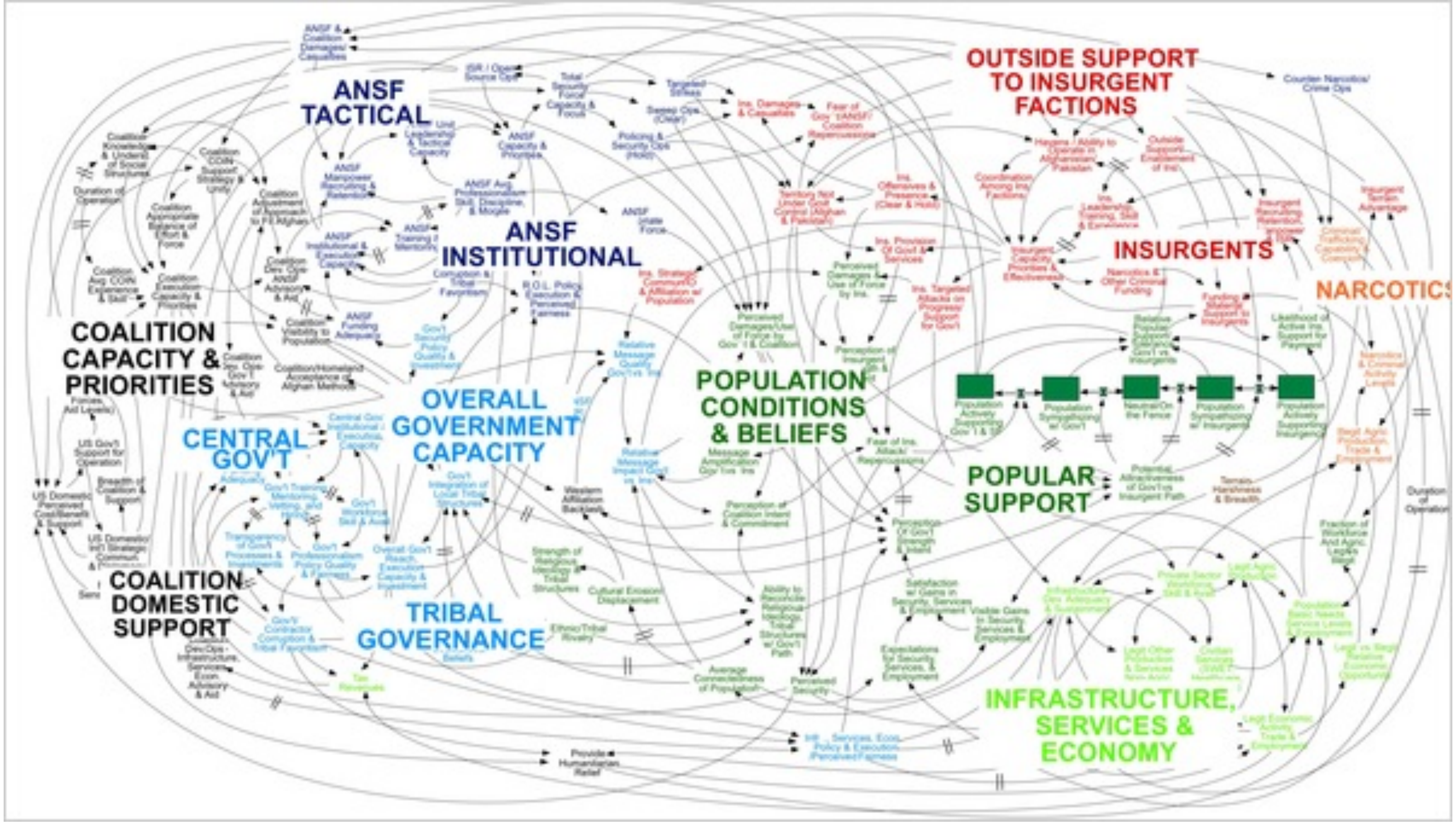
http://www.slate.com/id/2245232

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

Feb 18 2009

GLEaMviz.org

Chicago
New York
Los Angeles
Houston
Toronto
Vancouver
Calgary
Indianapolis

La Gloria
Sao Paulo
Mexico City
Rio De Janeiro
San Juan
Bogota

Johannesburg
Cairo
Cape Town
Nairobi

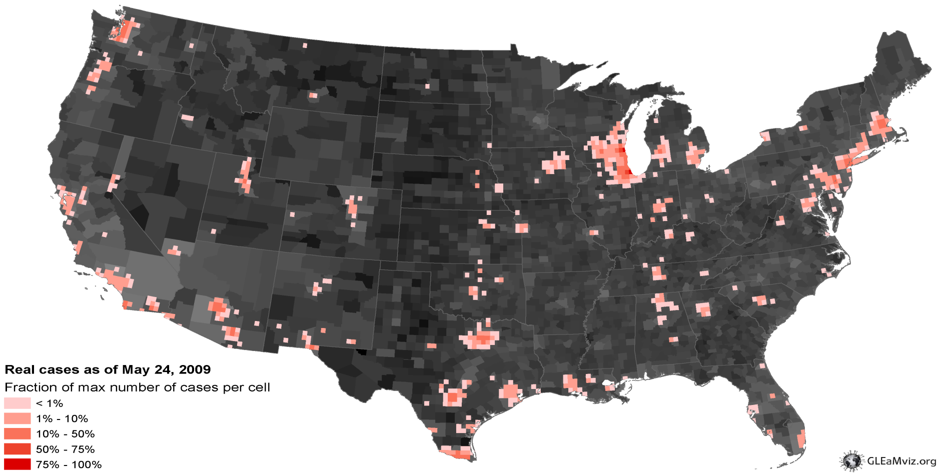
Paris
Frankfurt
Amsterdam
Rome
Milan
Moscow
Dublin

Hong Kong
Tokyo Narita
Bangkok
Singapore
Beijing
Manila

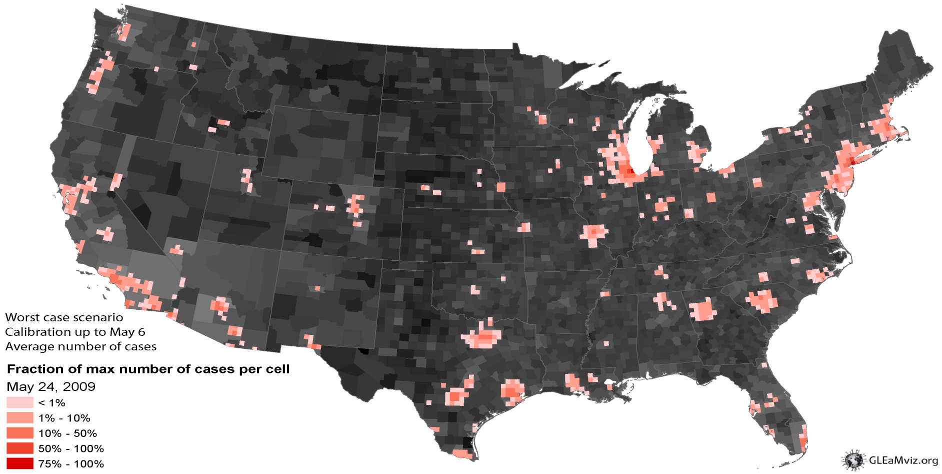
Sydney
Brisbane
Auckland
Perth

Predicting the H1N1 pandemic

Real

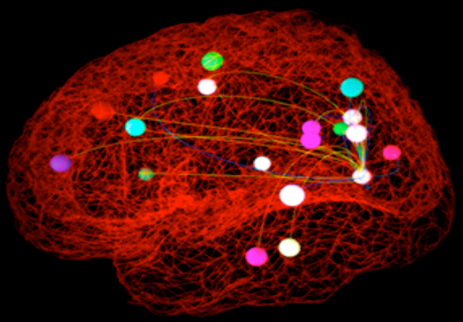


Projected

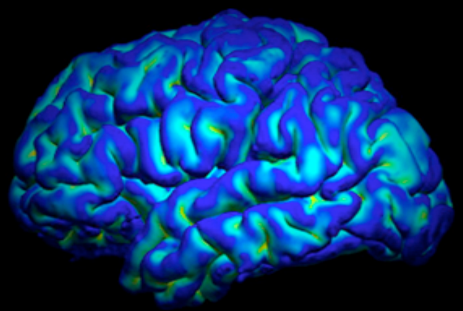


BRAIN RESEARCH

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.

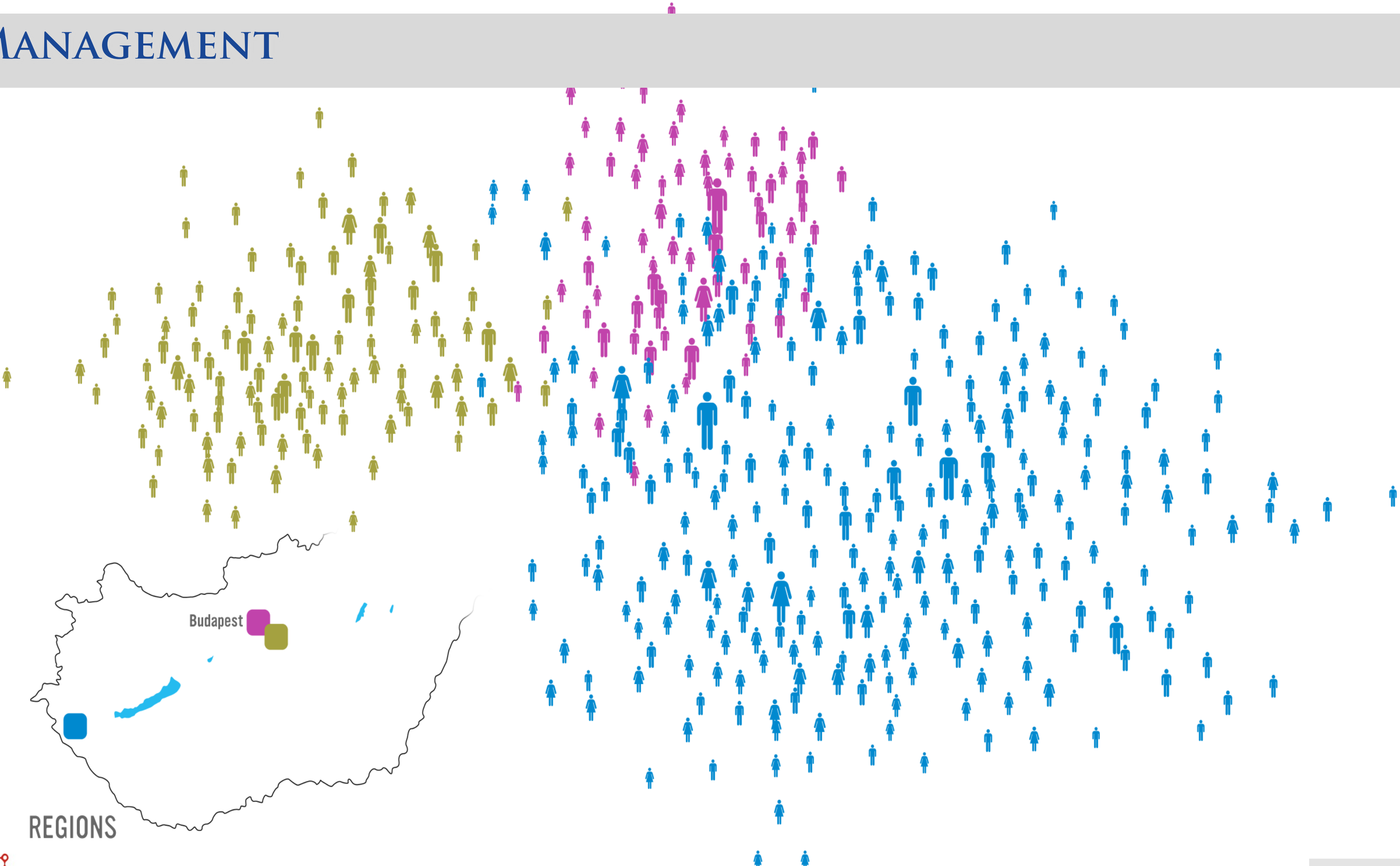


The Human Connectome Project (HCP) with the ambitious goal to construct a map of the complete structural and functional neural connections in vivo within and across individuals.



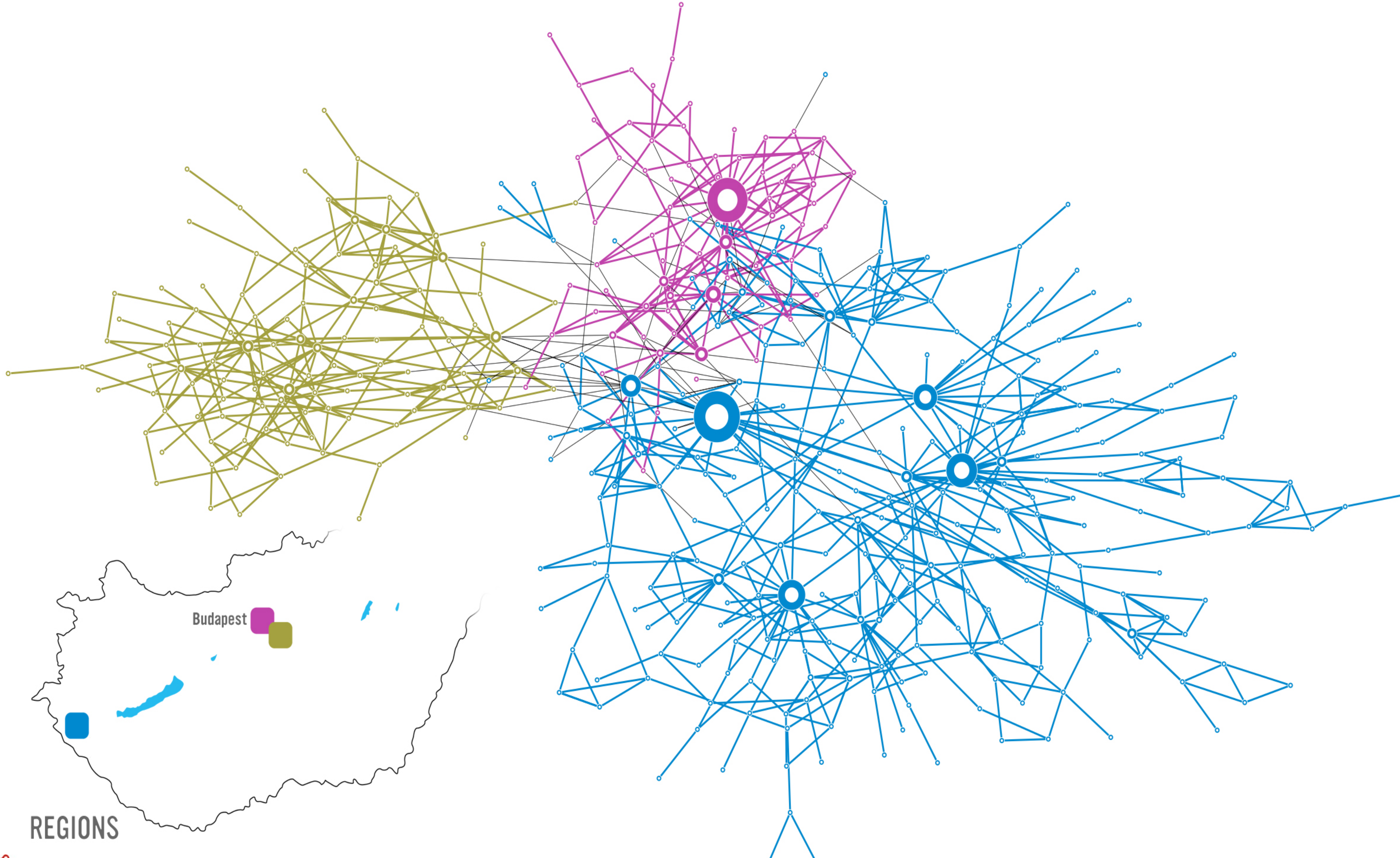
http://www.humanconnectomeproject.org/overview/

Management



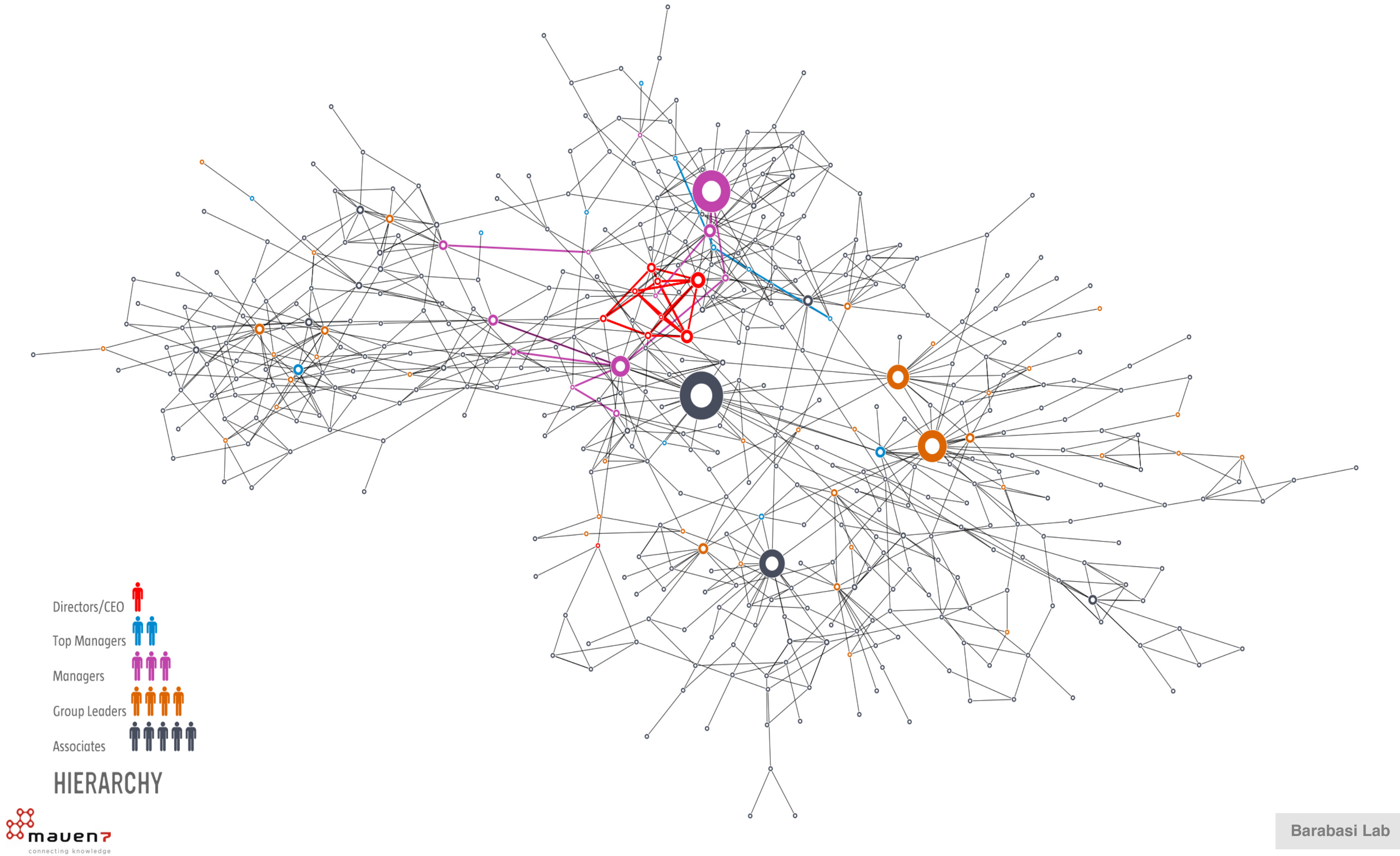
maven7
connecting knowledge

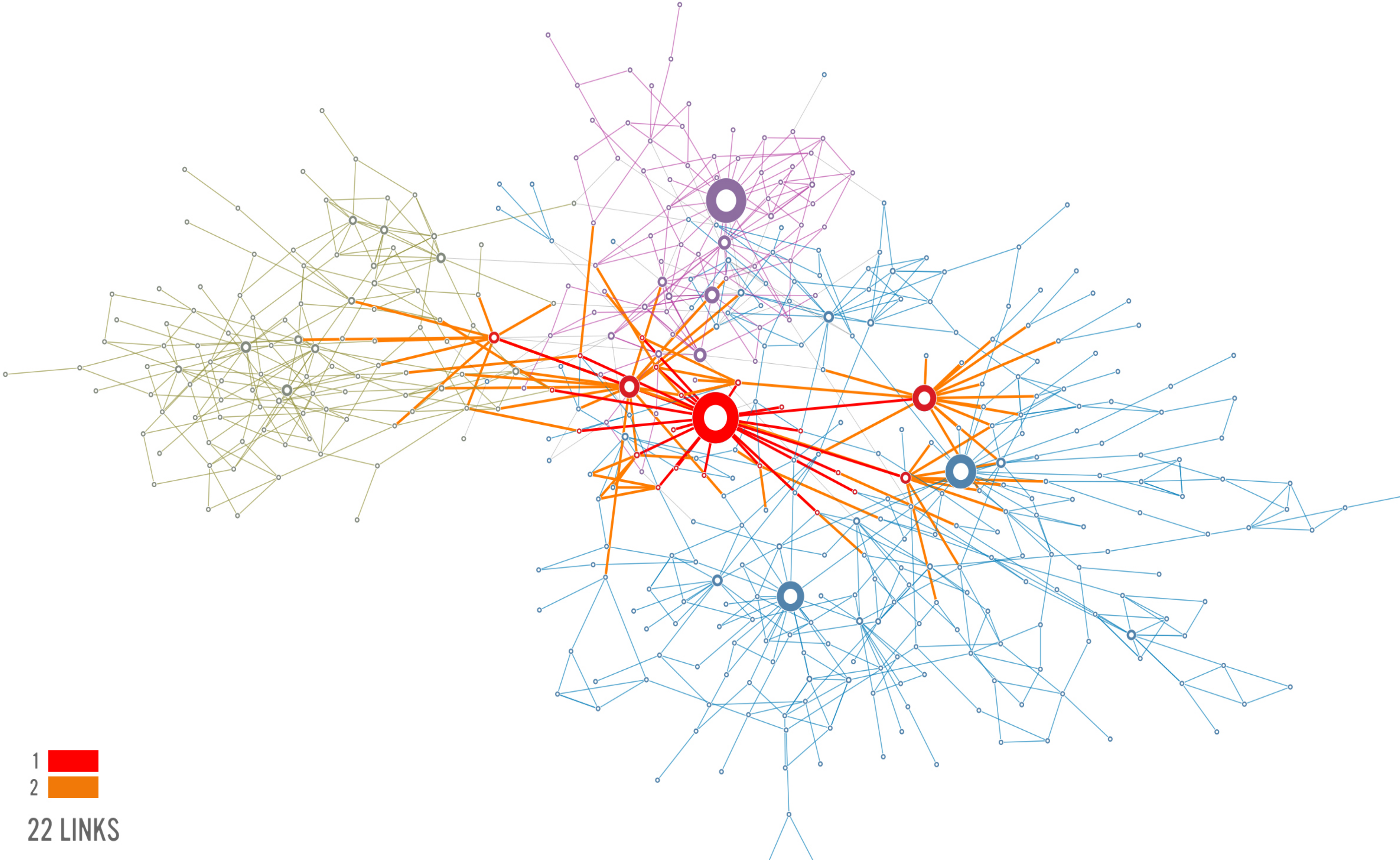
Barabasi Lab



maven7
connecting knowledge

Barabasi Lab


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


1
2
22 LINKS

SCIENTIFIC IMPACT

Complex systems and networks.

•**Science**:

Special Issue for the 10 year anniversary of Barabas i& Albert 1999 paper.

Original papers:

• 1998: Watts-Strogatz paper in the most cited **Nature** publication from 1998; highlighted by ISI as one of the ten most cited papers in physics in the decade after its publication.

•1999: Barabasi and Albert paper is the most cited **Science** paper in 1999;highlighted by ISI as one of the ten most cited papers in physics in the decade after its publication.

•2001: Pastor -Satorras and Vespignani is one of the two most cited papers among the papers published in 2001 by **Physical Review Letters**.

•2002: Girvan-Newman is the most cited paper in 2002 **Proceedings of the National Academy of Sciences**.

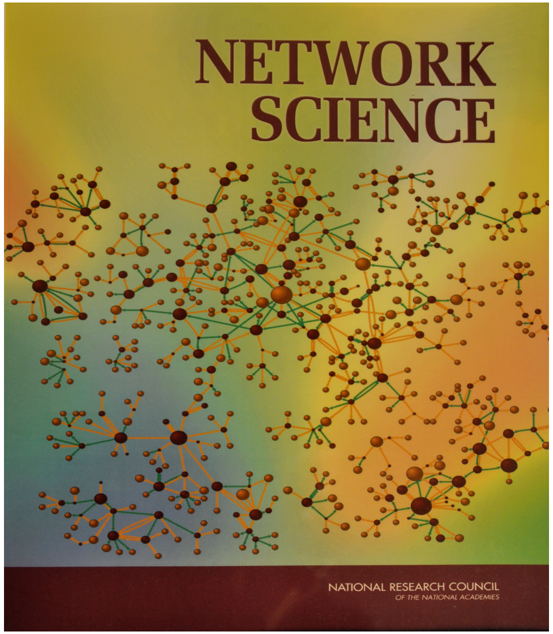
REVIEWS:

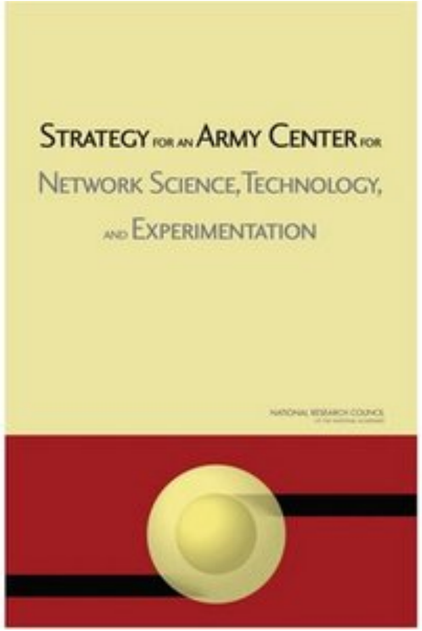
•The first review of network science by Albert and Barabasi, 2001) is the second most cited paper published in **Reviews of Modern Physics**, the highest impact factor physics journal, published since 1929. The most cited is *Chandaseklar*'s 1944 review on solar processes, but it will be surpassed by the end of 2012 by Albert *et al.*

•The SIAM review of Newman on network science is the most cited paper of any **SIAM journal**.

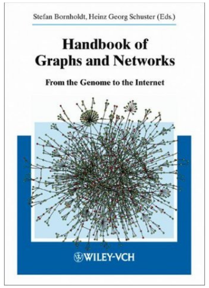
•BIOLOGY: "Network Biology", by Barabasi and Oltvai (2004) , is the second most cited paper in the history of **Nature Reviews Genetics**, the top review journal in genetics.

National Research Council

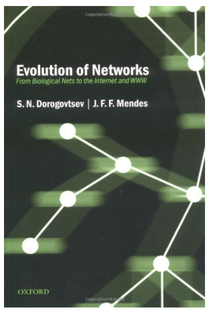




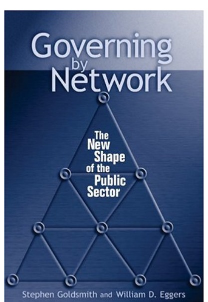
BOOKS



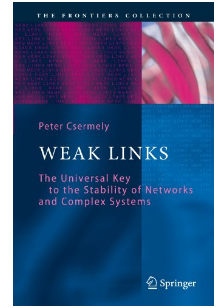
Handbook of Graphs and Networks: From the Genome to the Internet (Wiley-VCH, 2003).



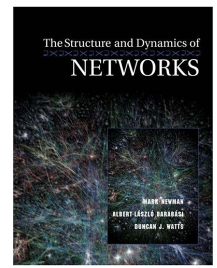
S. N. Dorogovtsev and J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW (Oxford University Press, 2003).



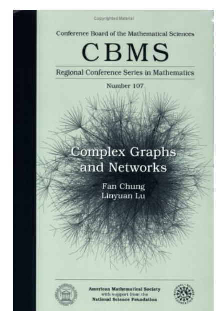
S. Goldsmith, W. D. Eggers, Governing by Network: The New Shape of the Public Sector (Brookings Institution Press, 2004).



P. Csermely, Weak Links: The Universal Key to the Stability of Networks and Complex Systems (The Frontiers Collection) (Springer, 2006), rst edn.

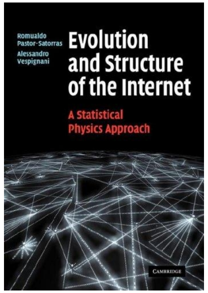


M. Newman, A.-L. Barabasi, D. J. Watts, The Structure and Dynamics of Networks: (Princeton Studies in Complexity) (Princeton University Press, 2006), rst edn.

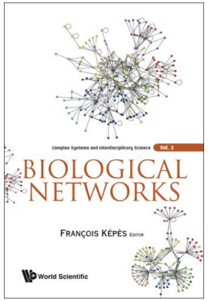


L. L. F. Chung, Complex Graphs and Networks (CBMS Regional Conference Series in Mathematics) (American Mathematical Society, 2006).

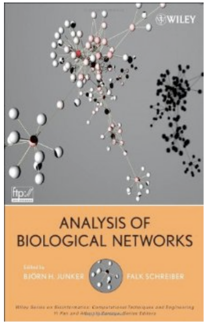
BOOKS



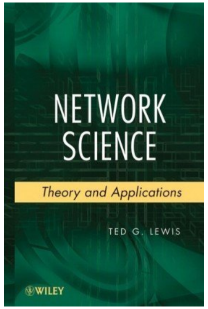
R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet: A Statistical Physics Approach (Cambridge University Press, 2007), rst edn.



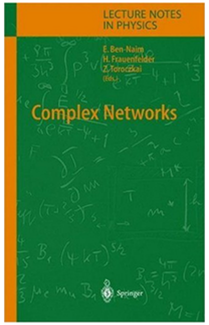
F. Kopos, Biological Networks (Complex Systems and Interdisciplinary Science) (World Scientic Publishing Company, 2007), rst edn.



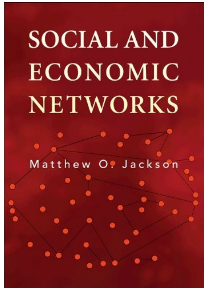
B. H. Junker, F. Schreiber, Analysis of Biological Networks (Wiley Series in Bioinformatics) (Wiley-Interscience, 2008).



T. G. Lewis, Network Science: Theory and Applications (Wiley, 2009).

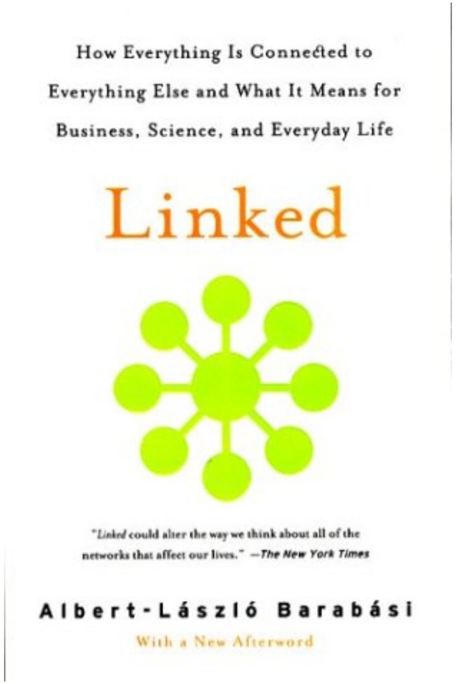


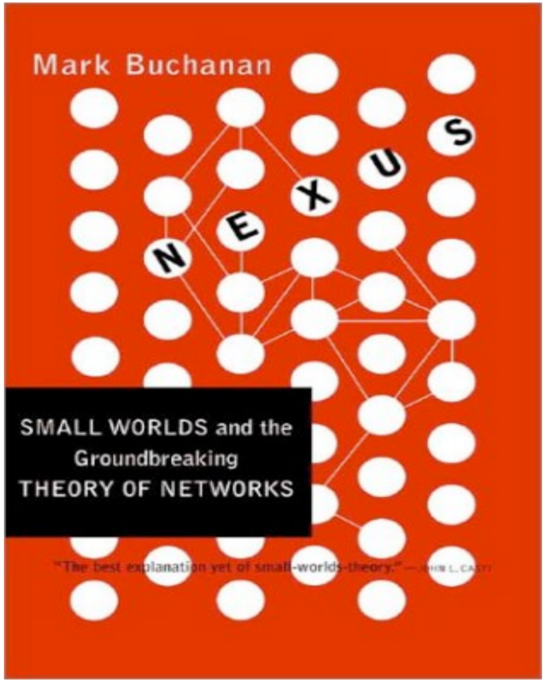
E. Ben Naim, H. Frauenfelder, Z.Torotzai, Complex Networks (Lecture Notes in Physics) (Springer, 2010), rst edn.

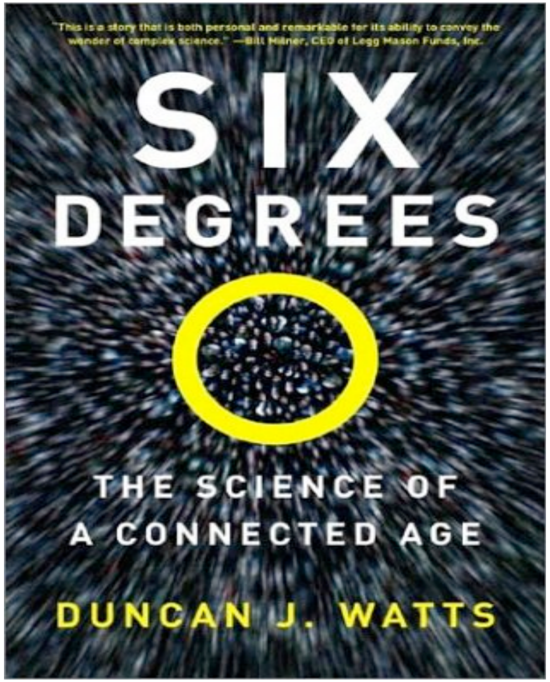


M. O. Jackson, Social and Economic Networks (Princeton University Press, 2010).

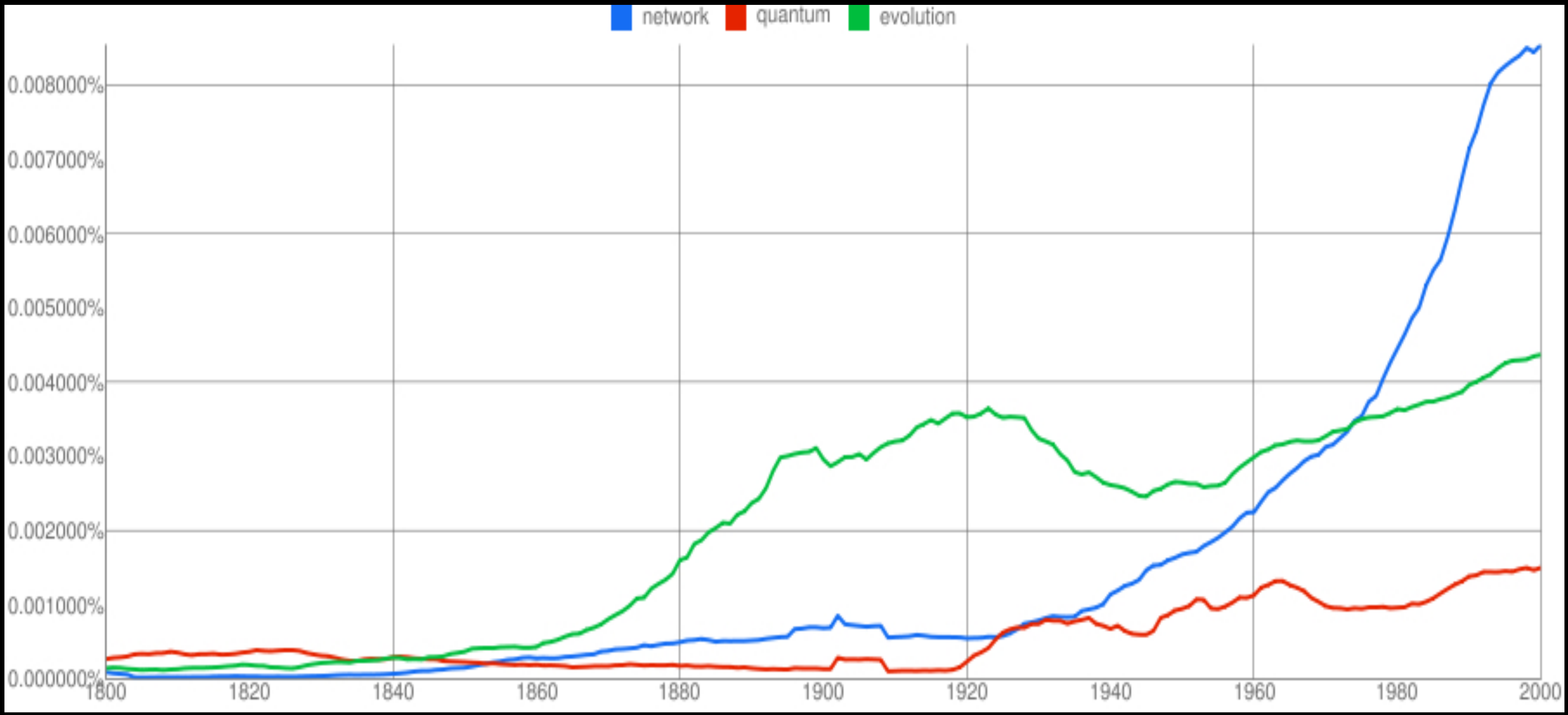
GENERAL AUDIENCE







Networks Awareness



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**.

Network Science

an interactive textbook

barabasi.com/**NetworkScienceBook/**

facebook.com/**NetworkScienceBook**

Download PDF

Download ibooks

Download SLIDES

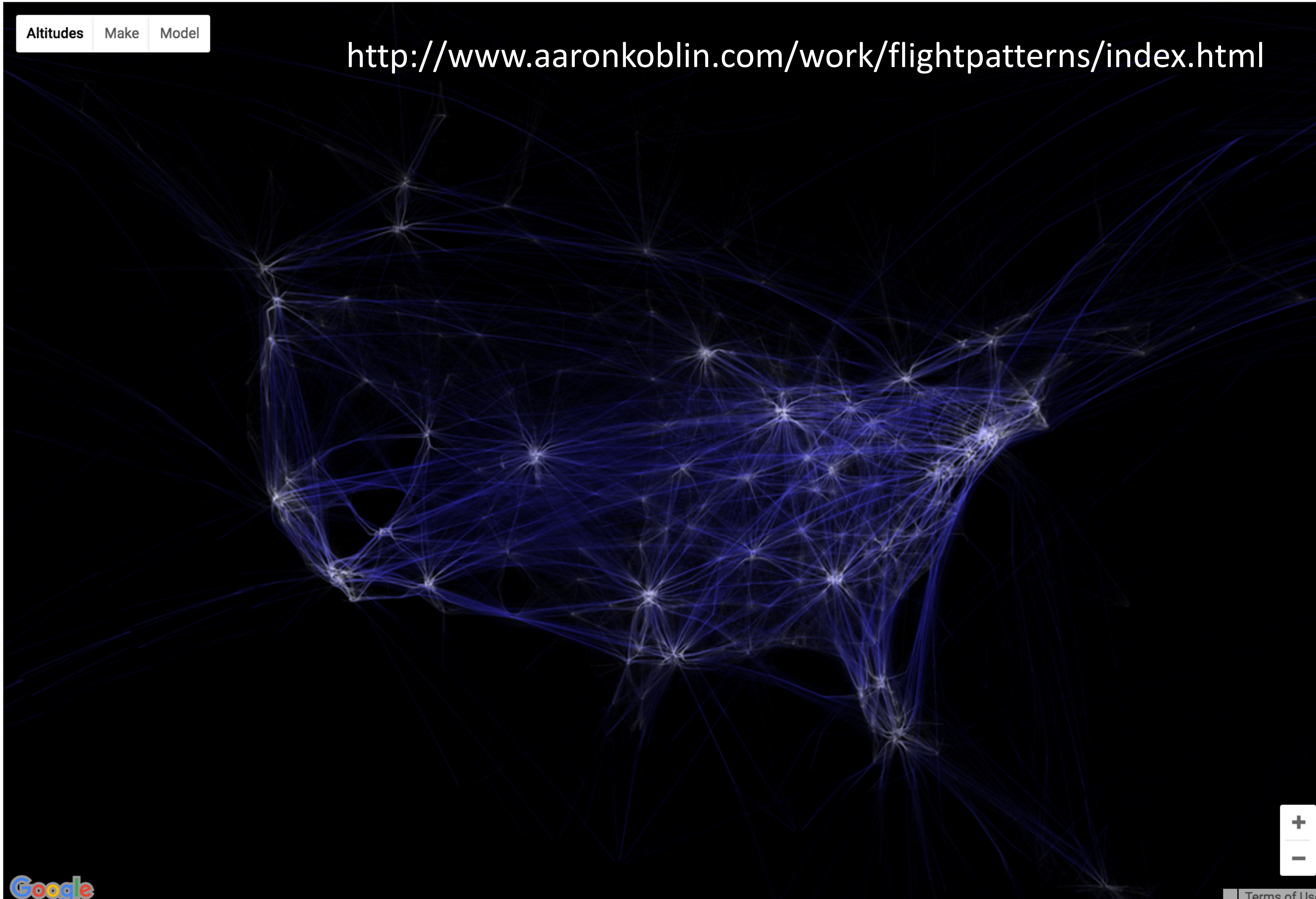
WHY NETWORK SCIENCE FOR SOCIAL SYSTEMS

- The study social systems is not new but we have a lot of new data
- Understand the emergence of collective behavior and form of organizations stemming from individuals behavior of individuals
- Find regularities and patterns on large scale social phenomena
- Find ”universal” laws in different social systems

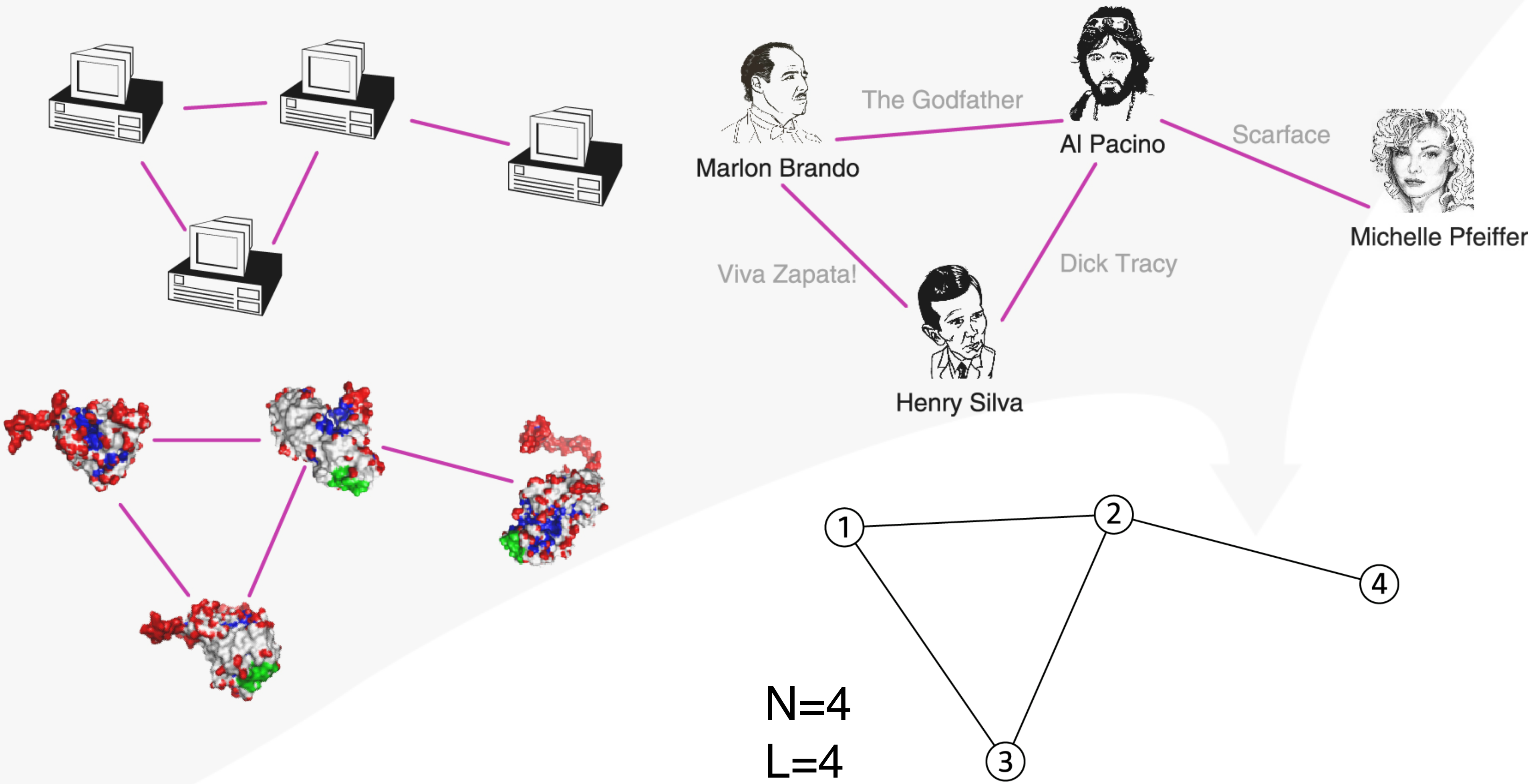
MODELING OF SOCIAL SYSTEMS: LIMITS

- You cannot approximate a human with a particle
- Social systems are neither closed nor isolated
- It is very hard to study the system without perturbing it

- We need big numbers of individuals (large populations)
- Interdisciplinary approach!


Altitudes
Make
Model
http://www.aaronkoblin.com/work/flightpatterns/index.html
Google
Terms of Use

A COMMON LANGUAGE



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.

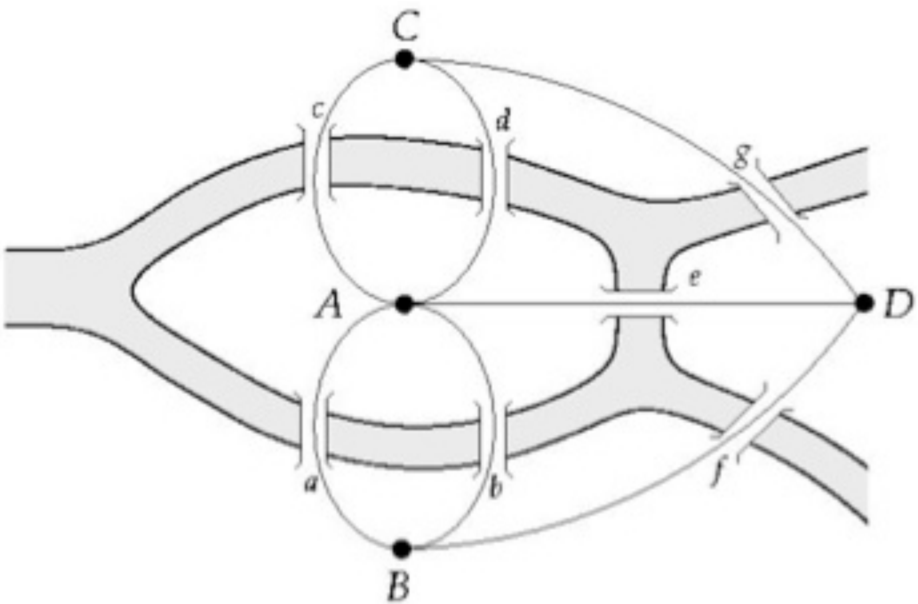
For example, the way we assign the links between a group of individuals will determine the nature of the question we can study.

The Bridges of Konigsberg

THE BRIDGES OF KONIGSBERG

CAN ONE WALK ACROSS THE SEVEN BRIDGES AND NEVER CROSS THE SAME BRIDGE TWICE?

THE BRIDGES OF KONIGSBERG



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

1735: Euler's theorem:

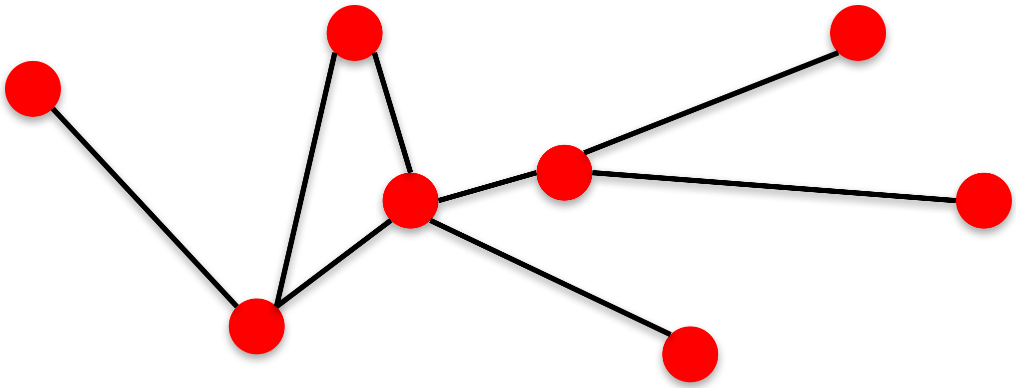
(a) If a graph has more than two nodes of odd degree, there is no path.

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

Networks and graphs

http://barabasi.com/networksciencebook/

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)

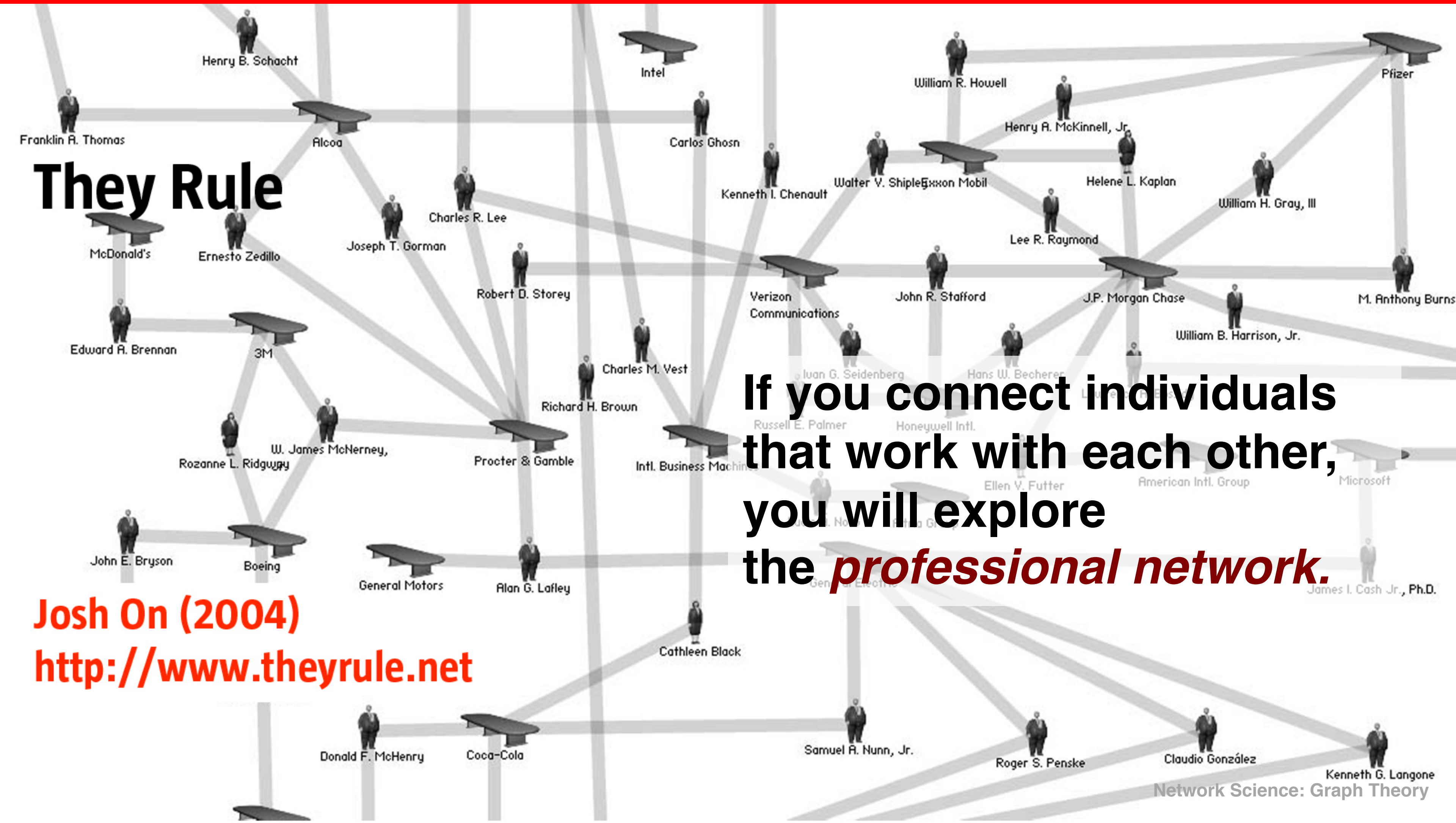
graph: mathematical representation of a network

•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.

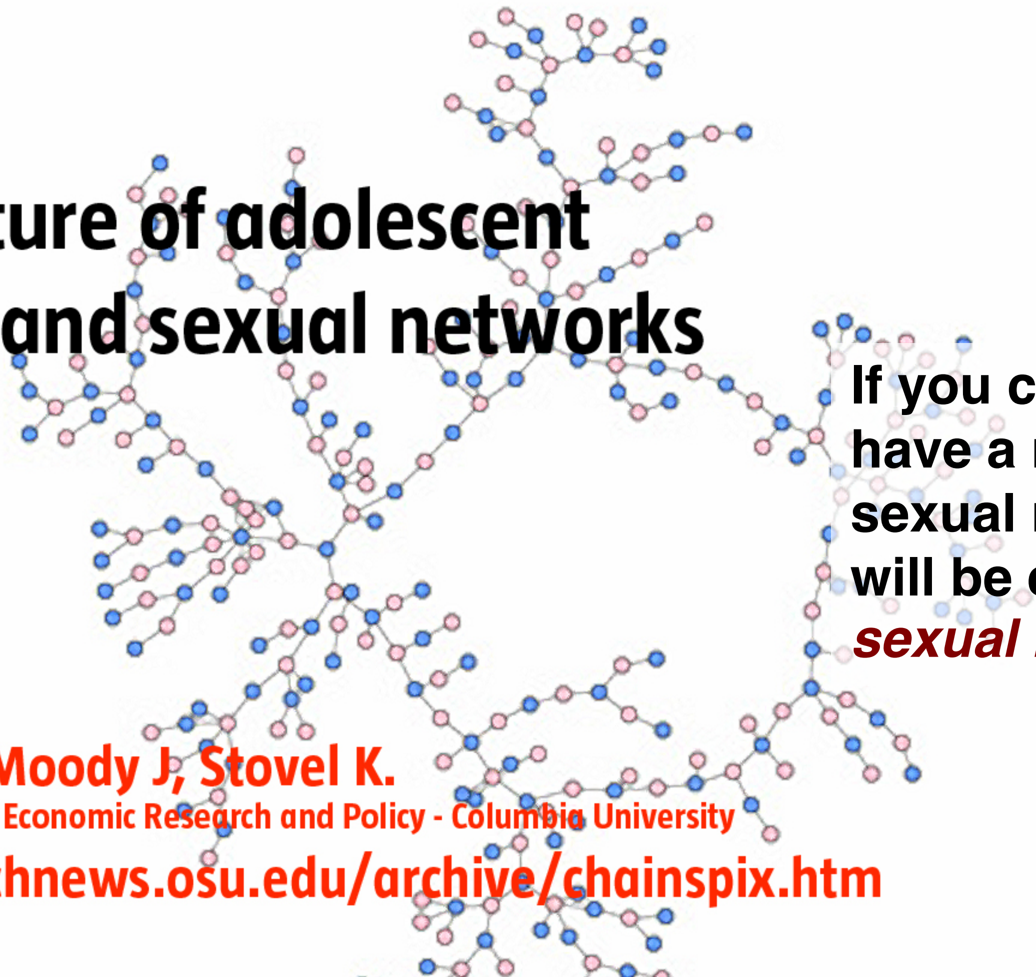
CHOOSING A PROPER REPRESENTATION



They Rule

Josh On (2004)
http://www.theyrule.net

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



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?

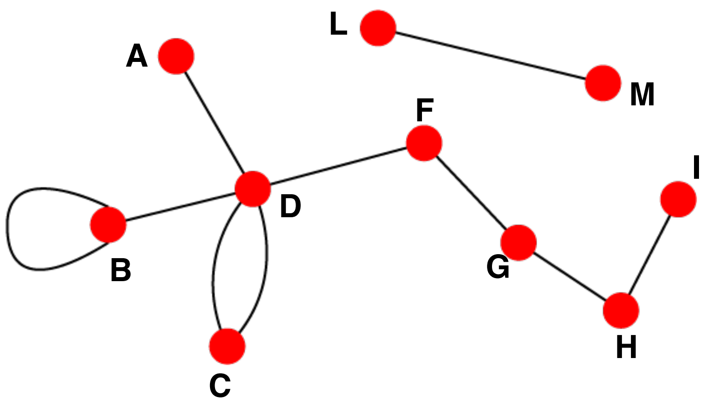
It is a network, nevertheless.

UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:



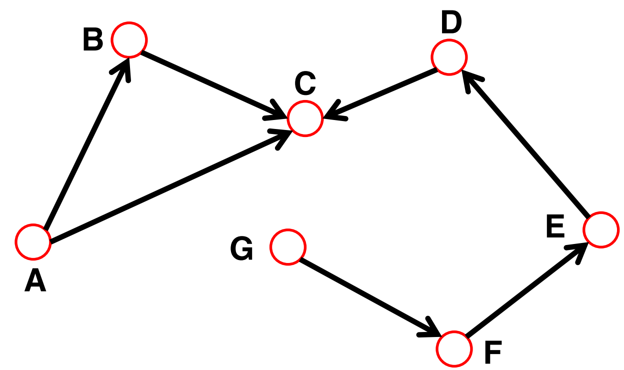
Undirected links :

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

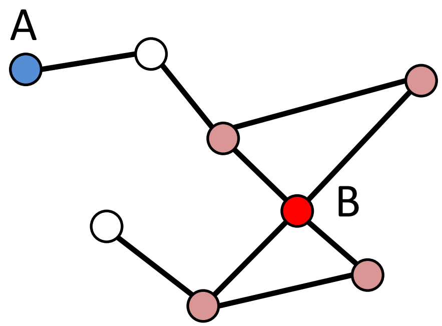
REFERENCE NETWORKS

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

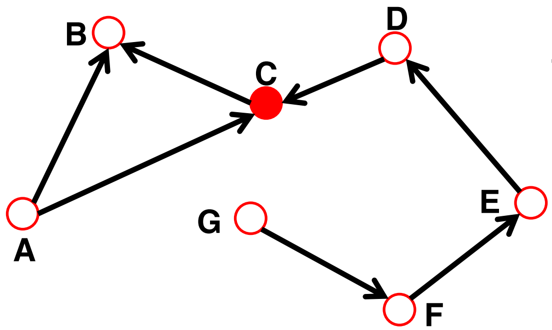
Undirected



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.

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

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

A BIT OF STATISTICS

BRIEF STATISTICS REVIEW

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

Average (mean):

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

The n^{th} moment:

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

Standard deviation:

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

Distribution of x:

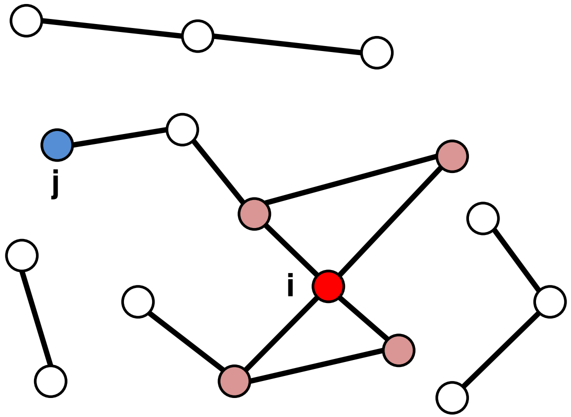
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

where p_x follows

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

AVERAGE DEGREE

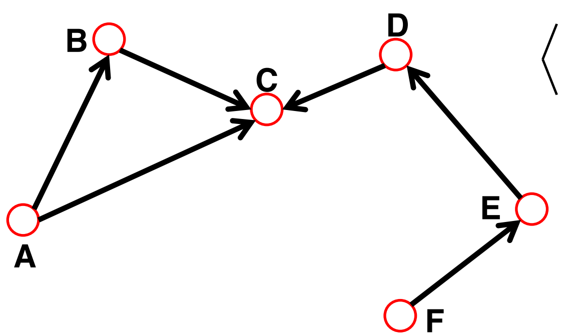
Undirected



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

N – the number of nodes in the graph

Directed



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

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

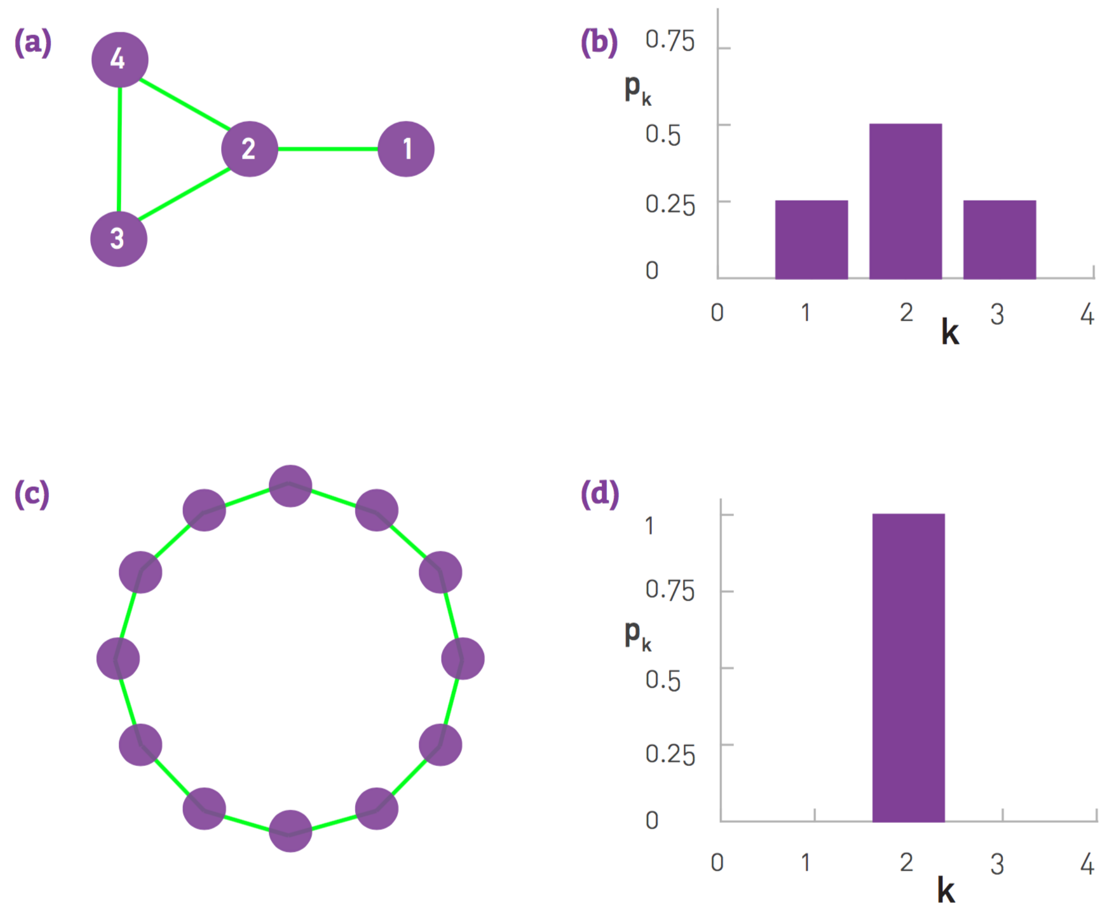
Average Degree

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
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

DEGREE DISTRIBUTION

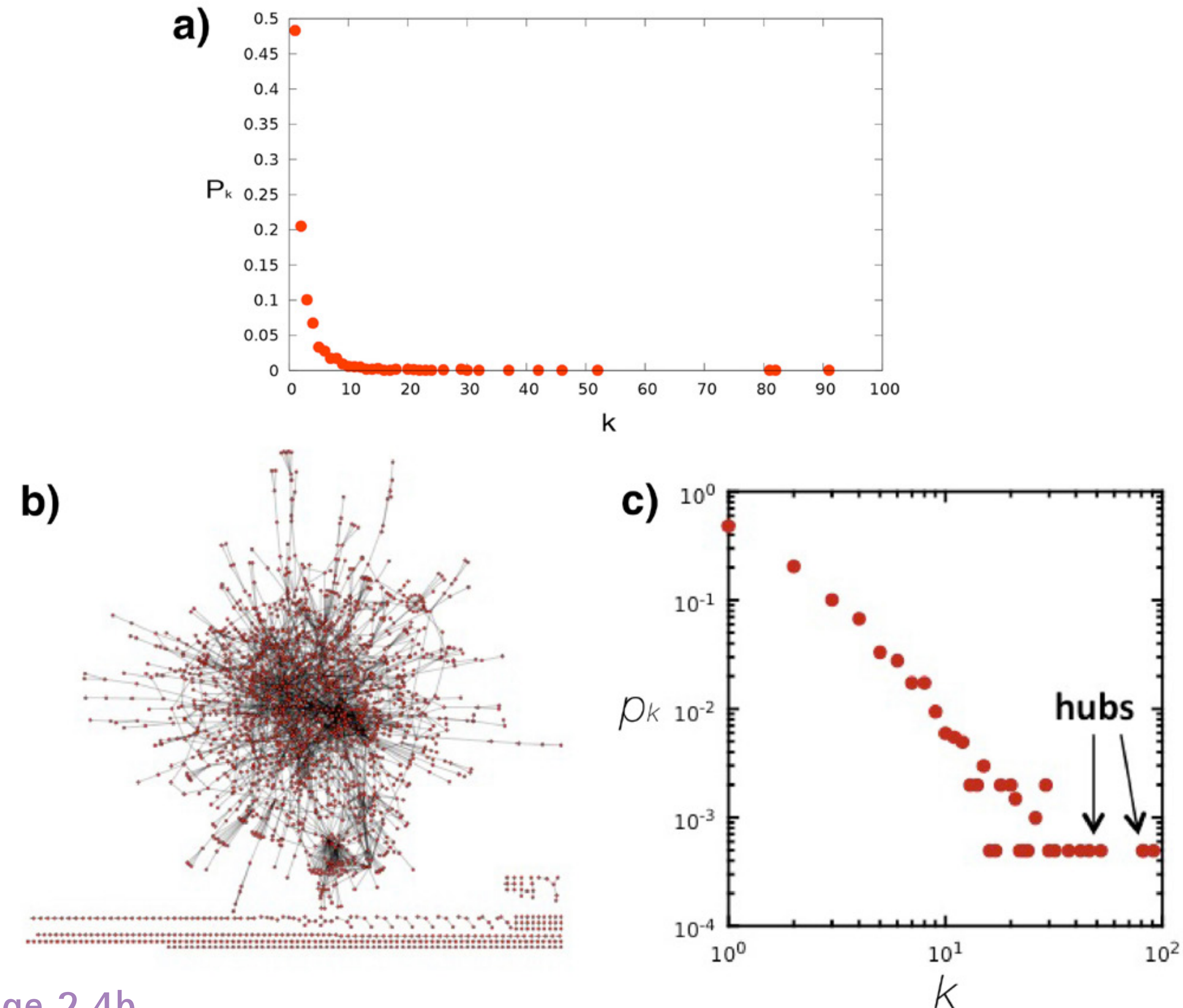


Image 2.4b

DEGREE DISTRIBUTION

Discrete Representation: $\mathbf{p_k}$ is the probability that a node has degree **k**.

Continuum Description: **p(k)** is the pdf of the degrees, where

$$\int_{k_1}^{k_2} p(k)dk$$

represents the probability that a node's degree is between $\mathbf{k_1}$ and $\mathbf{k_2}$.

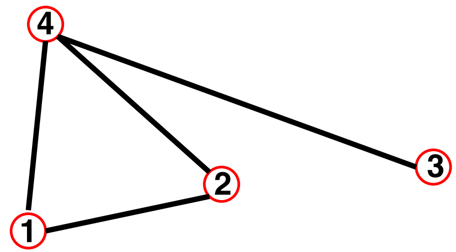
Normalization condition:

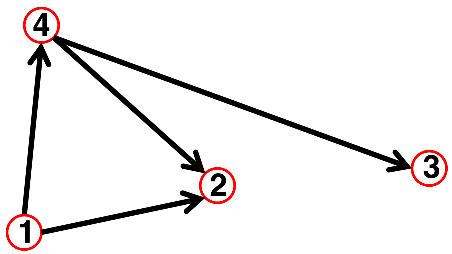
$$\sum_{0}^{\infty} p_k = 1 \qquad \int_{K_{\min}}^{\infty} p(k)dk = 1$$

where K_{min} is the minimal degree in the network.

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.

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

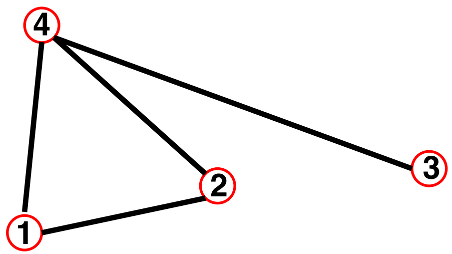
Note that for a directed graph (right) the matrix is not symmetric.

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

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

ADJACENCY MATRIX AND NODE DEGREES

Undirected



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

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

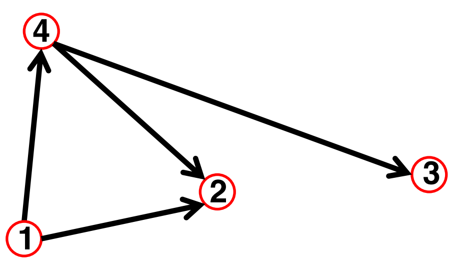
$$A_{ii} = 0$$

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

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

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

Directed



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

$$A_{ij} \neq A_{ji}$$

$$A_{ii} = 0$$

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

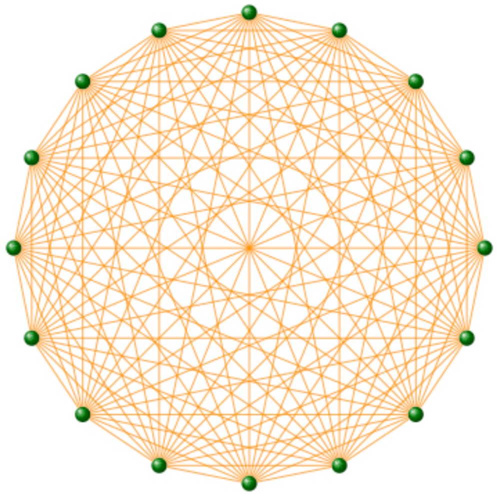
$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

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

Real networks are sparse

COMPLETE GRAPH

The maximum number of links a network of N nodes can have is: $L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$



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

REAL NETWORKS ARE SPARSE

Most networks observed in real systems are sparse:

$$L \ll L_{max}$$

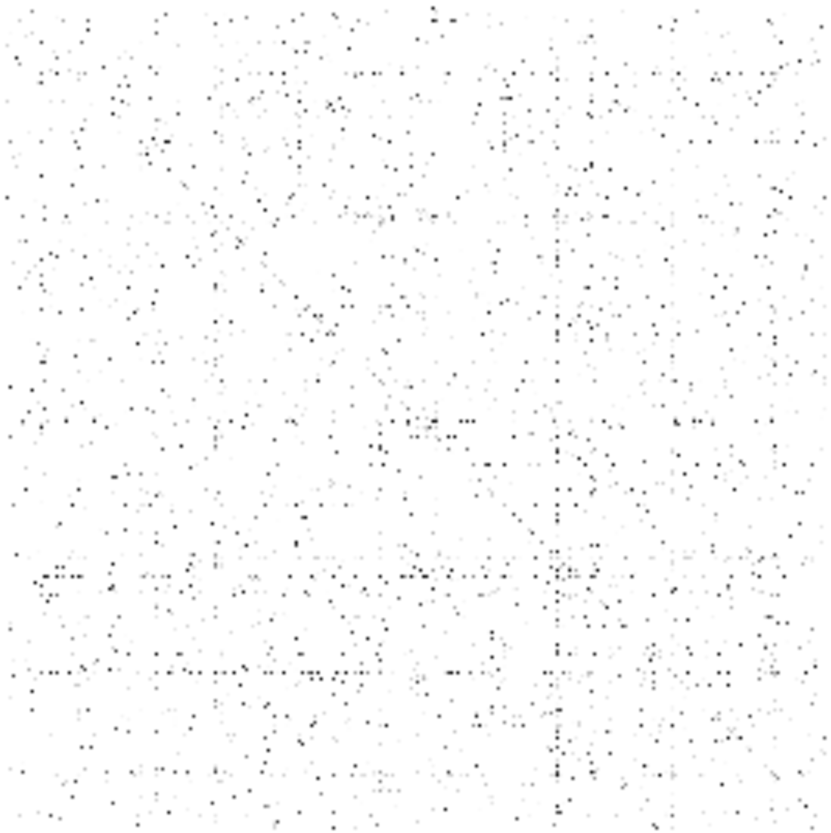
or

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

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

(Source: Albert, Barabasi, RMP2002)

ADJACENCY MATRICES ARE SPARSE



WEIGHTED AND UNWEIGHTED NETWORKS

So far we discussed only networks for which all links have the same weight, i.e. $A_{ij} = 1$. In many applications we need to study *weighted networks*, where each link (i, j) has a unique weight w_{ij}.
For example, in mobile call networks the weight can represent the total number of minutes two individuals talk with each other on the phone; on the power grid the weight is the amount of current flowing through a transmission line.

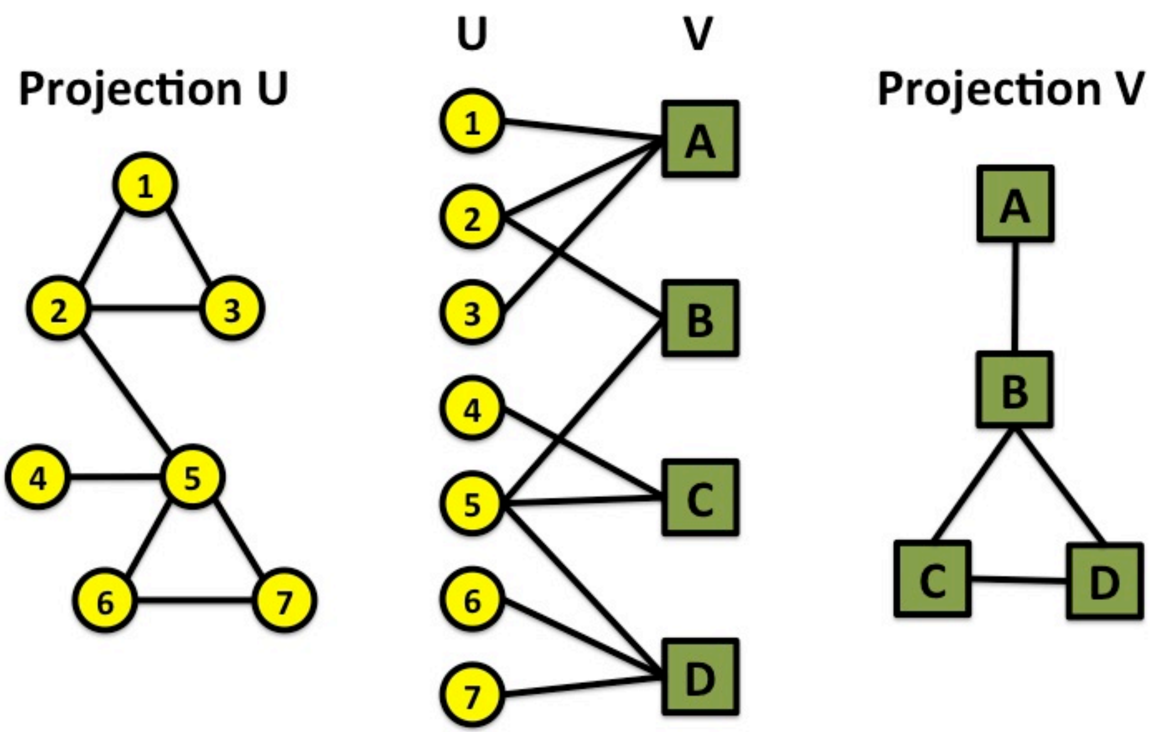
For *weighted networks* the elements of the adjacency matrix carry the weight of the link as

$$A_{ij} = w_{ij}$$

BIPARTITE NETWORKS

BIPARTITE GRAPHS

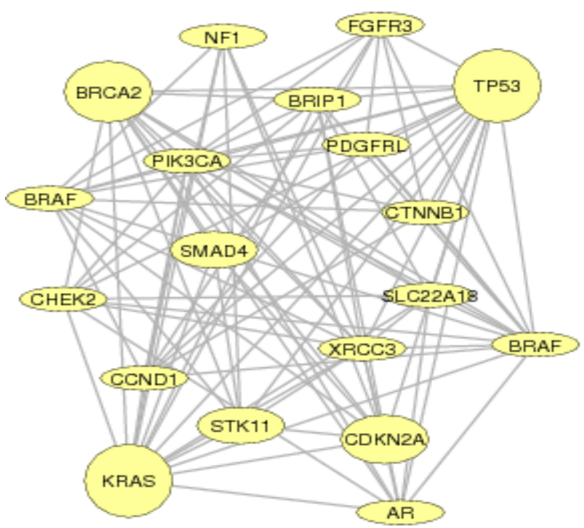
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are independent sets.



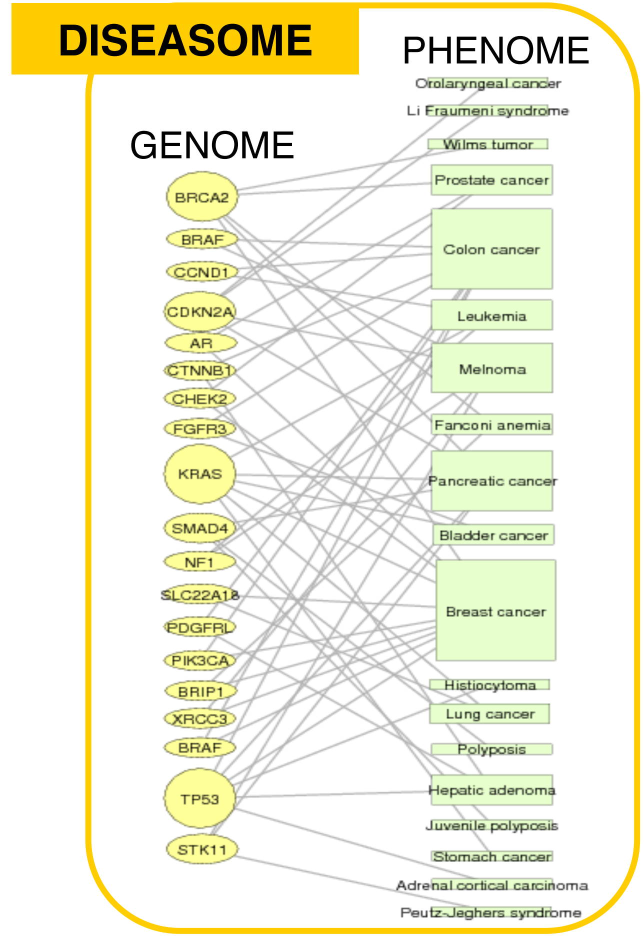
Examples:

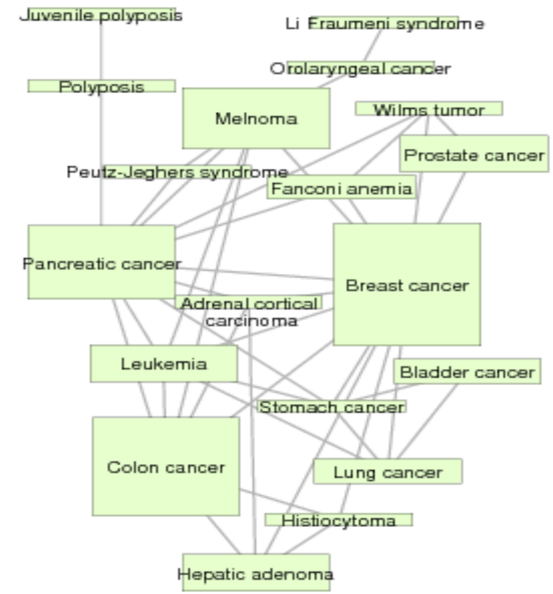
Hollywood actor network
Collaboration networks
Disease network (diseasome)

GENE NETWORK – DISEASE NETWORK



Gene network

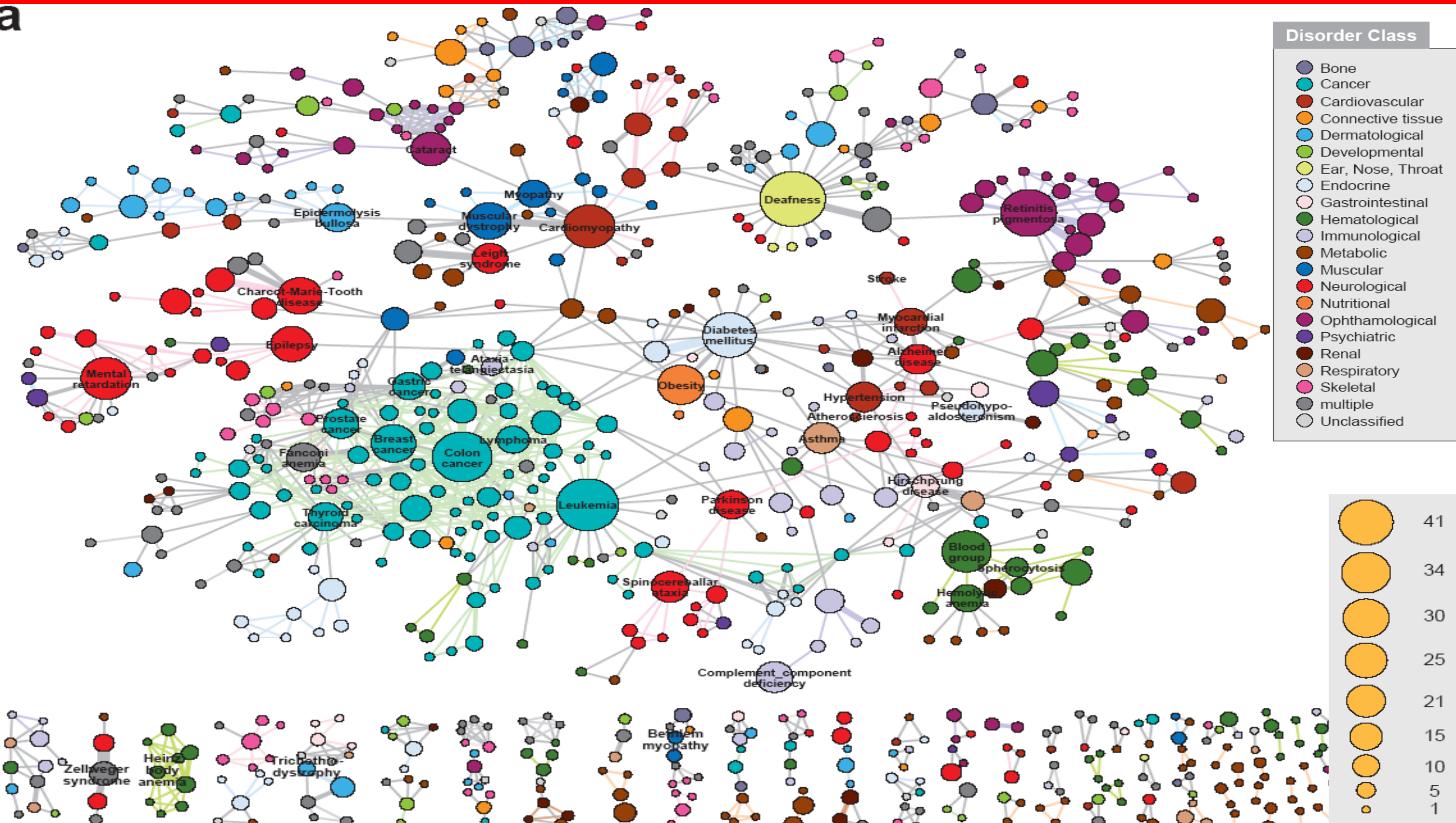




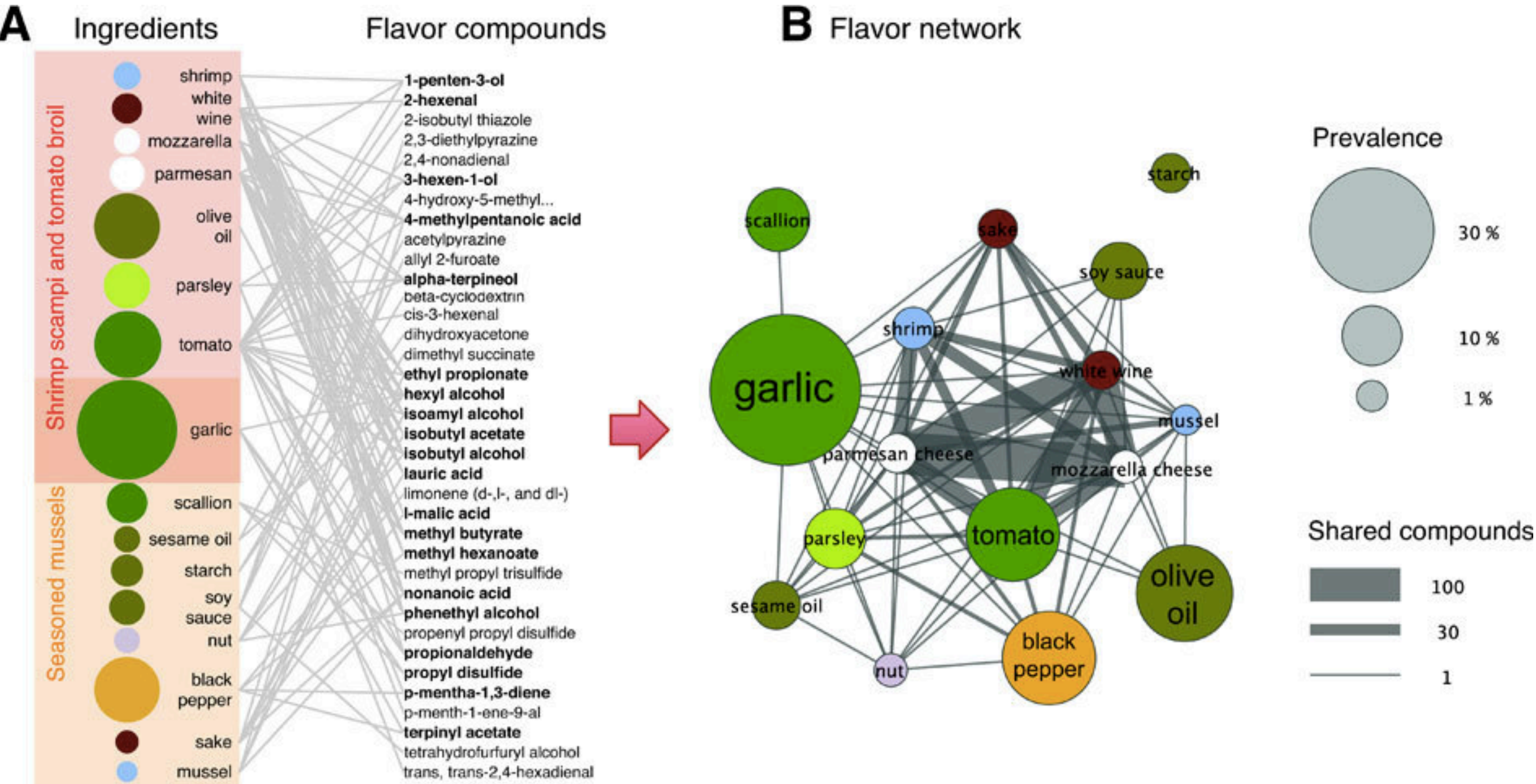
Disease network

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

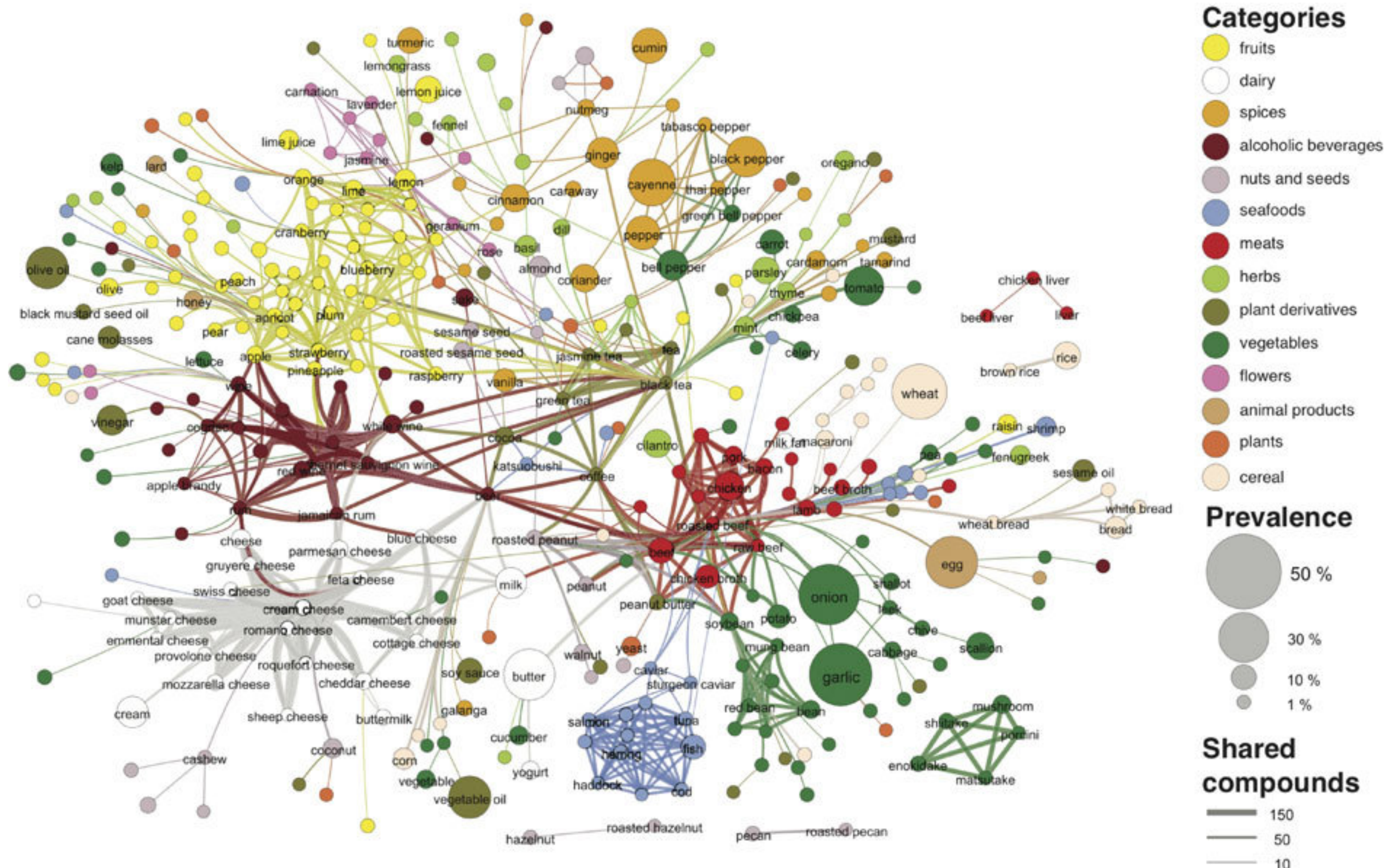
HUMAN DISEASE NETWORK



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

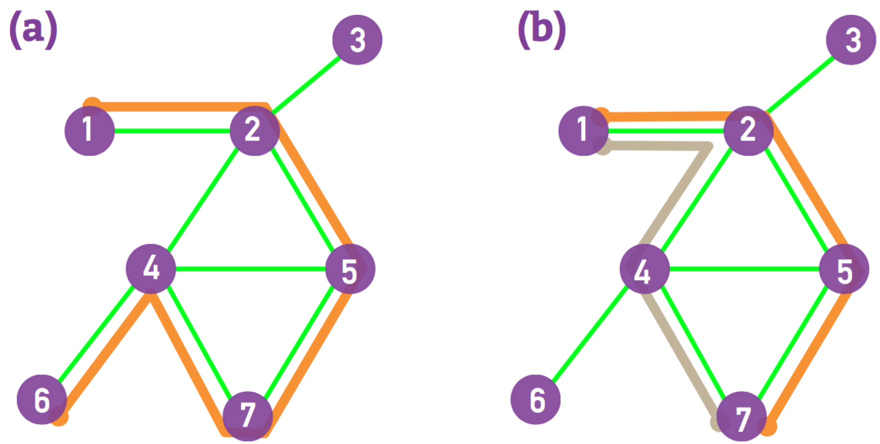
PATHOLOGY

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

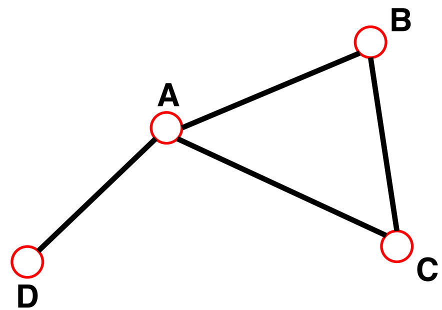
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



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

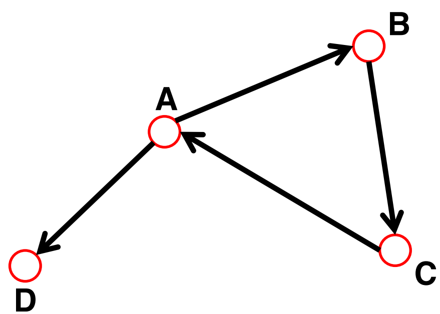
DISTANCE IN A GRAPH

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*

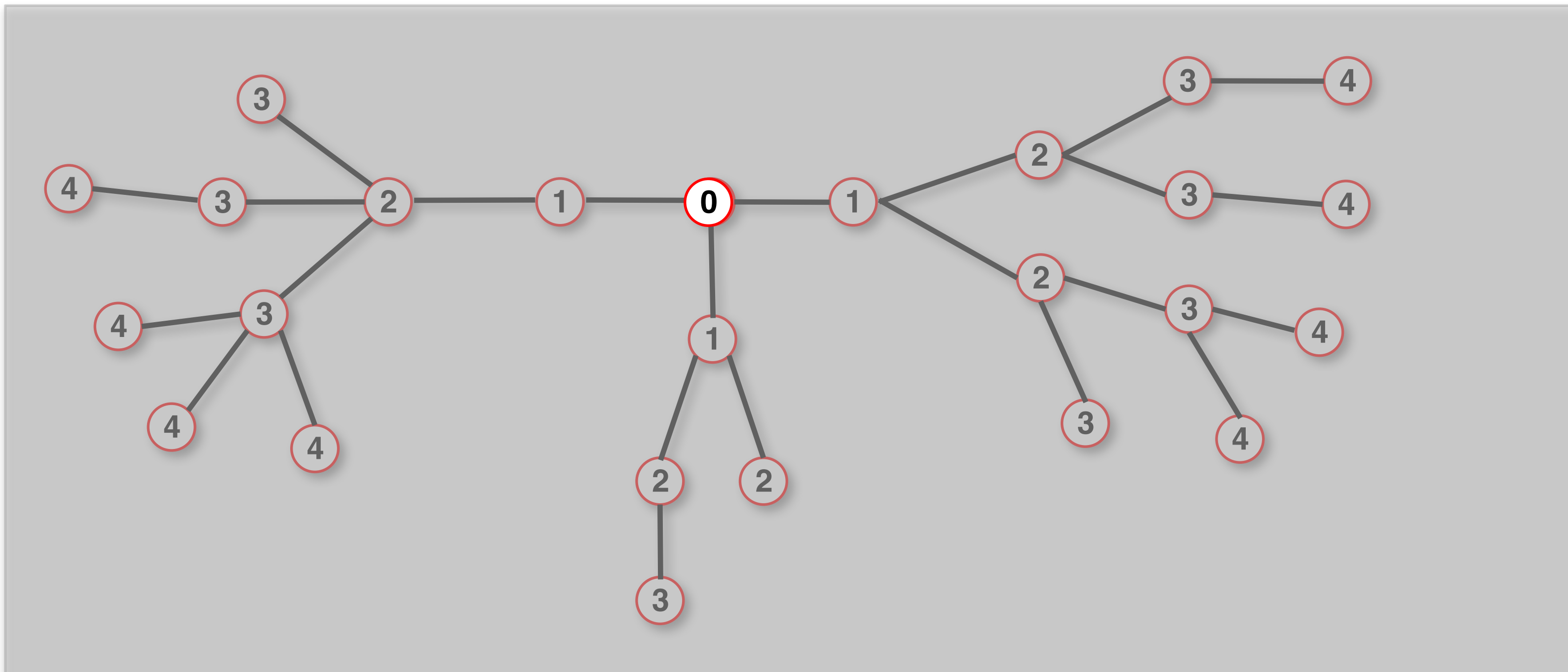
$$N_{ij}^{(n)} = [A^n]_{ij}$$

**holds for both directed and undirected networks.*

FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

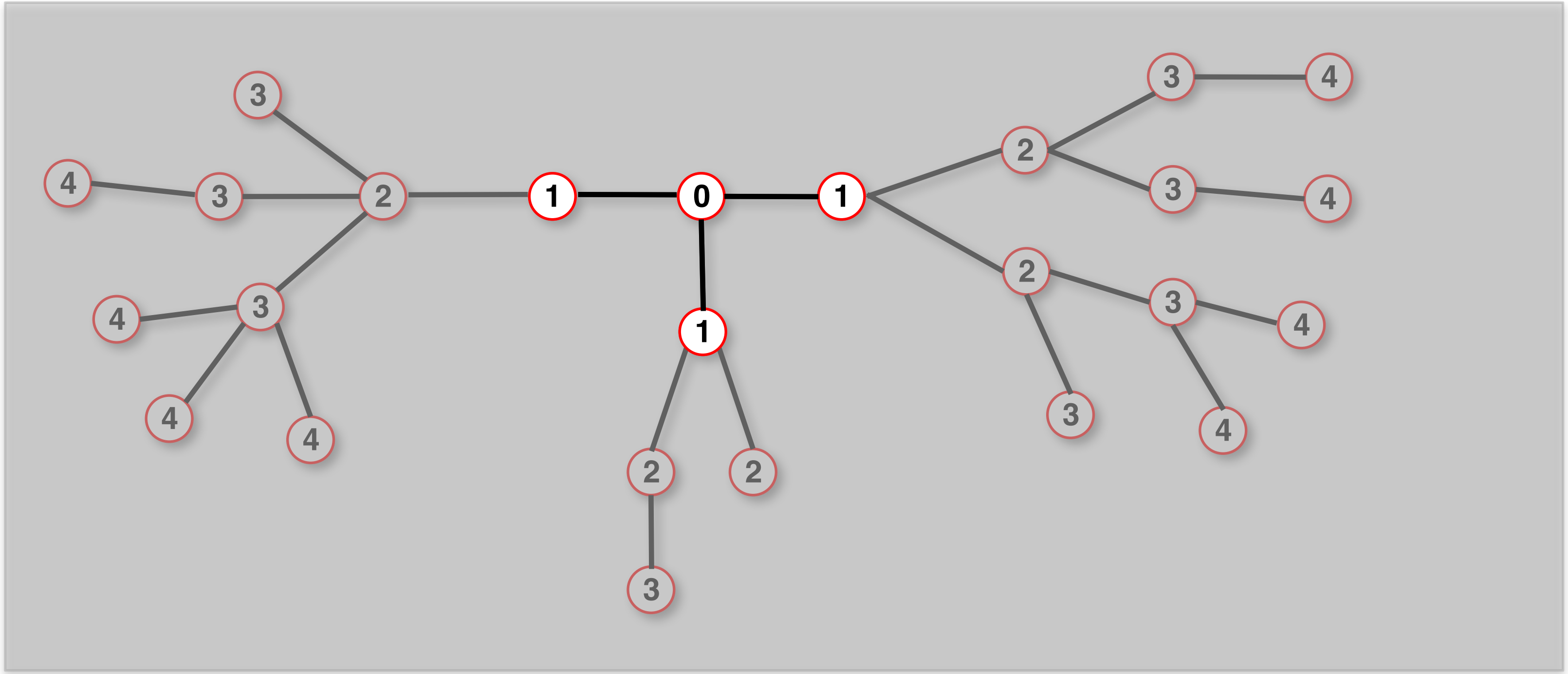
1.Start at 0.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

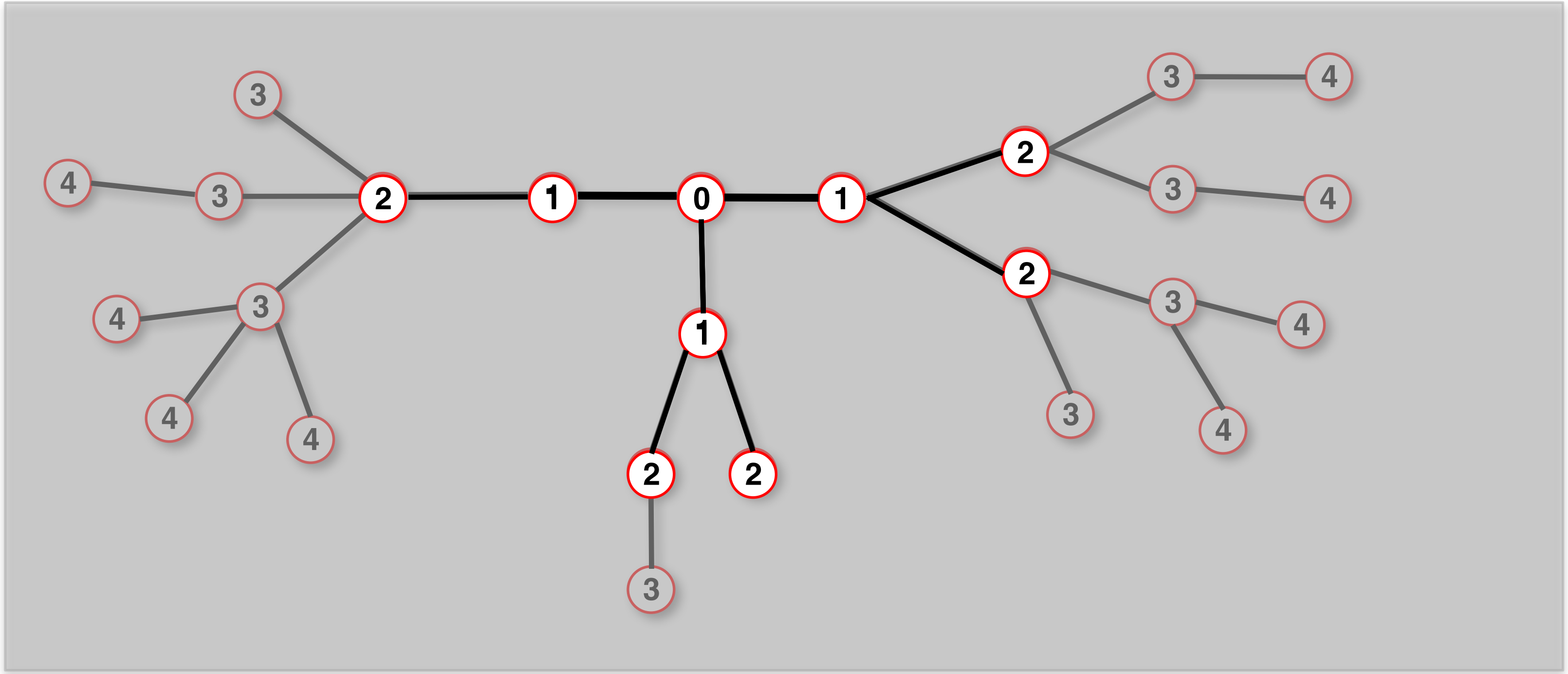
1.Start at 0.
2.Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

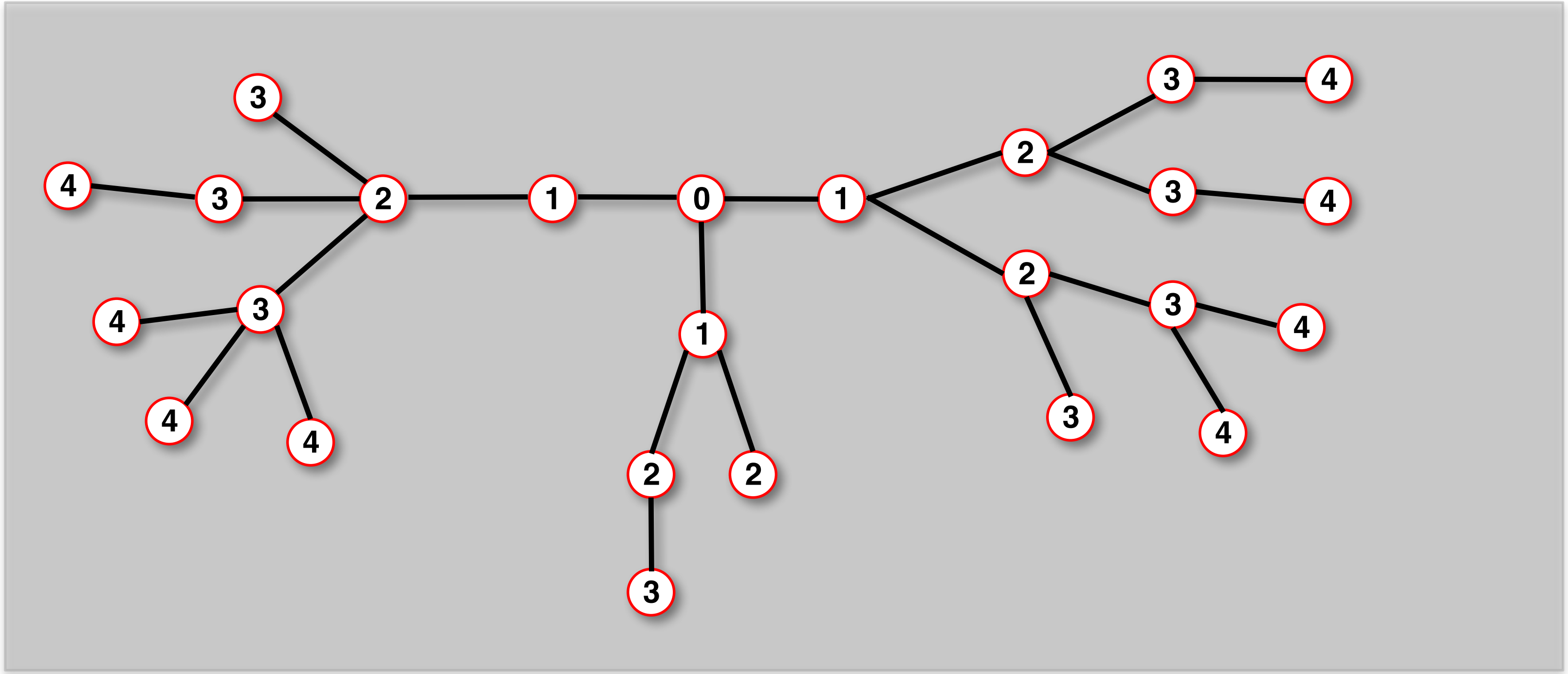
1. Start at 0.
2. Find the nodes adjacent to 0. Mark them as at distance 1. Put them in a queue.
3. Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.



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: average distance between all pairs of nodes in the network, where d_{ij} is the distance from node *i* to node j

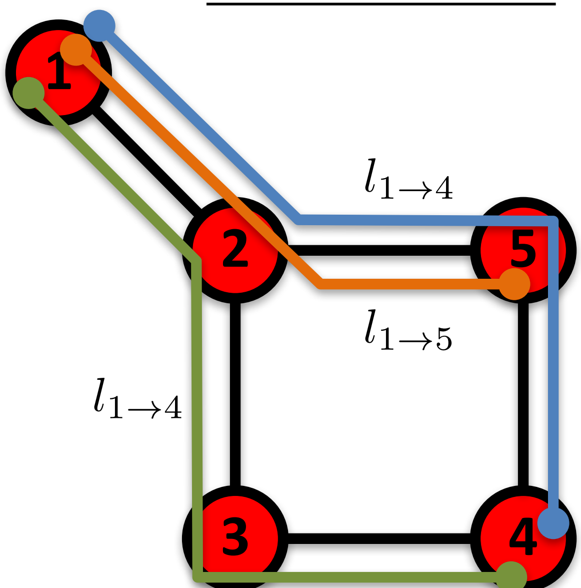
$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

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

Shortest Path



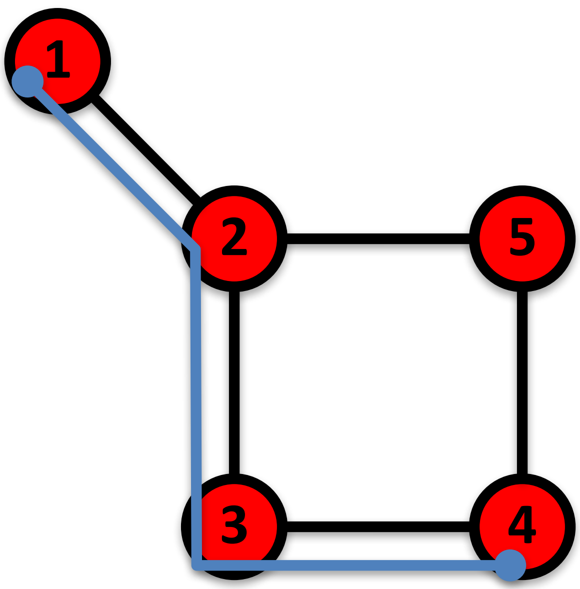
$$l_{1\to4} = 3$$

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

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

PATHOLOGY: SUMMARY

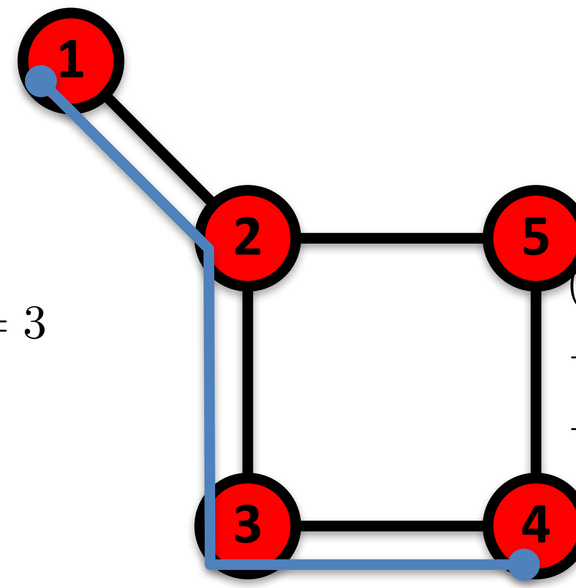
Diameter



$$l_{1\to 4} = 3$$

The longest shortest path in a graph

Average Path Length

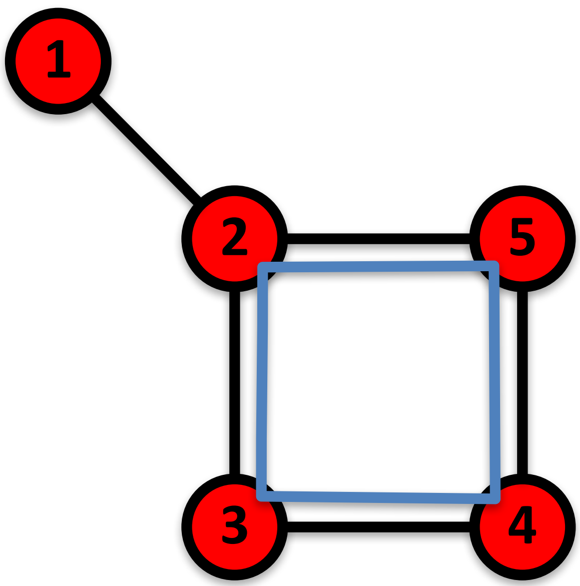


$$\left(l_{1\to 2} + l_{1\to 3} + l_{1\to 4} + l_{1\to 5} + l_{2\to 3} + l_{2\to 4} + l_{2\to 5} + l_{3\to 4} + l_{3\to 5} + l_{4\to 5}\right)/10 = 1.6$$

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

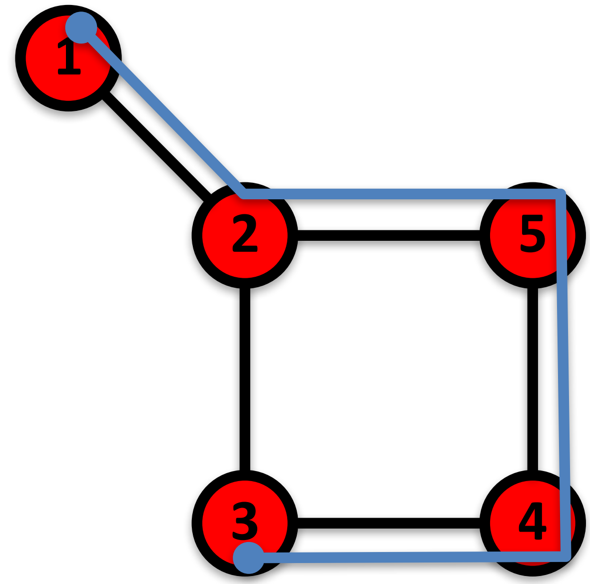
PATHOLOGY: SUMMARY

Cycle



A path with the same start
and end node.

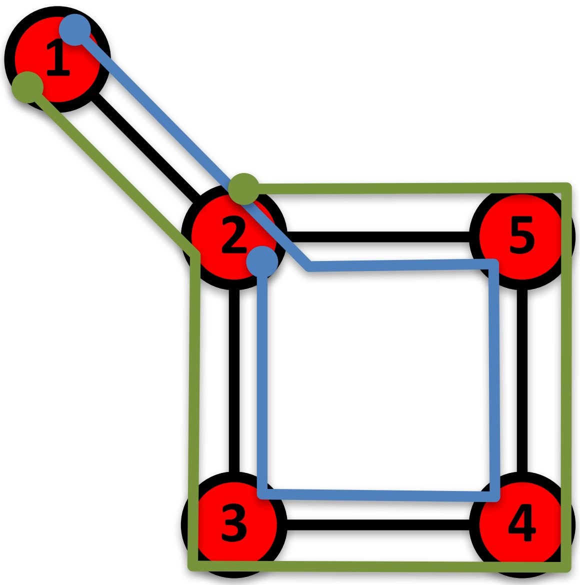
Self-avoiding Path



A path that does not intersect
itself.

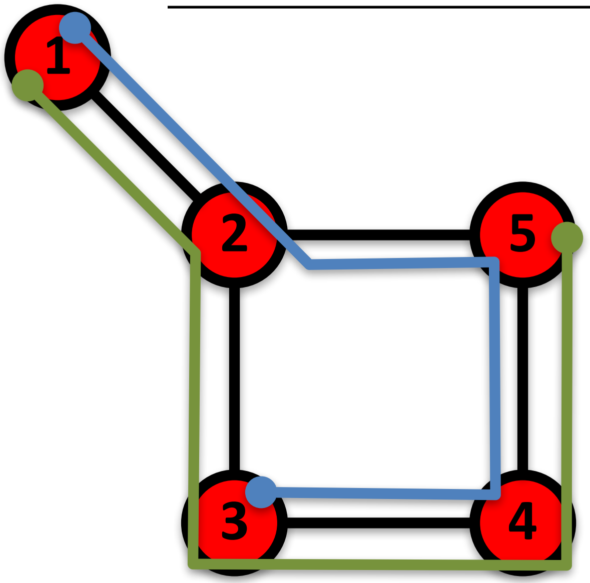
PATHOLOGY: SUMMARY

Eulerian Path



A path that traverses each link exactly once.

Hamiltonian Path

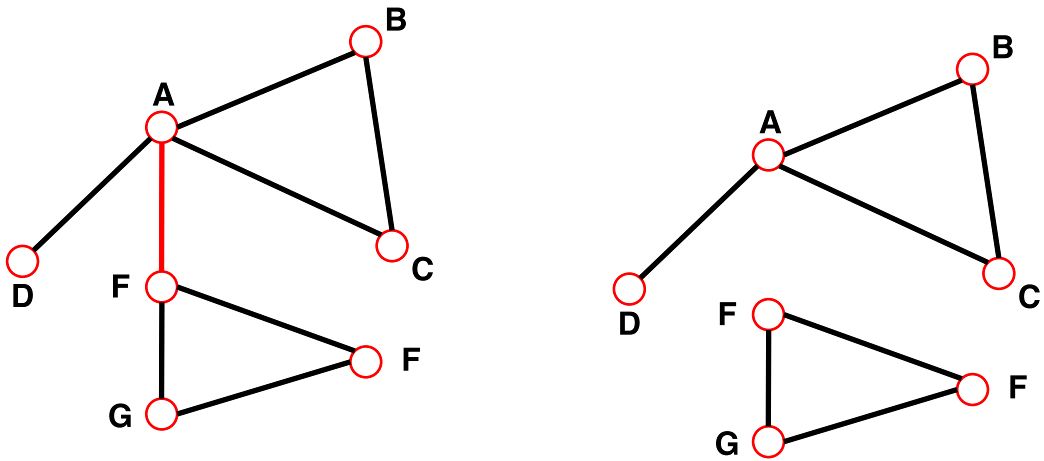


A path that visits each node exactly once.

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**

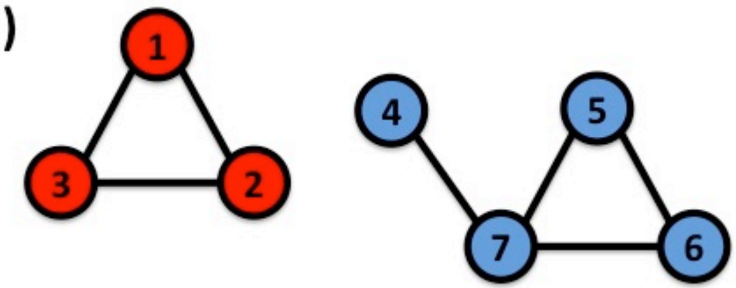
Bridge: if we erase it, the graph becomes disconnected.

CONNECTIVITY OF UNDIRECTED GRAPHS

Adjacency Matrix

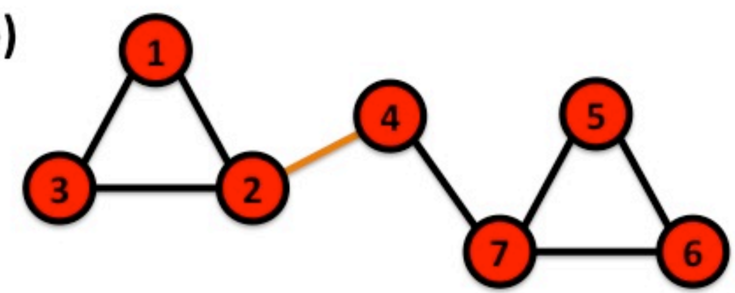
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

(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)



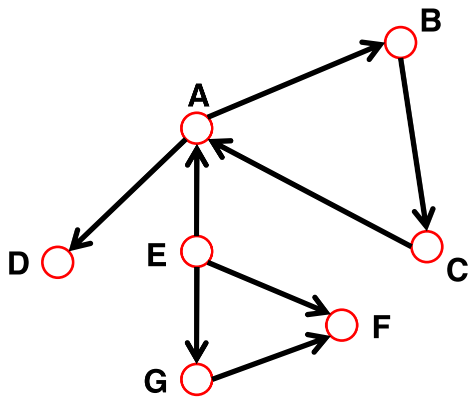
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

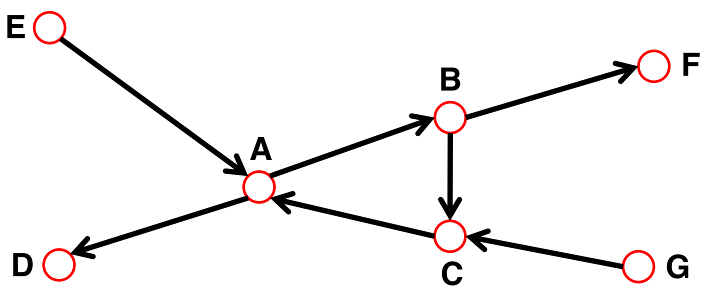
CONNECTIVITY OF DIRECTED GRAPHS

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,

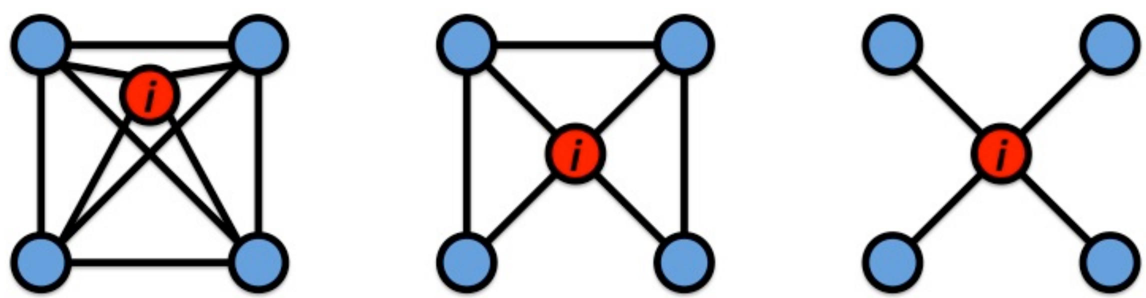
Out-component: nodes that can be reached from the scc.

Clustering coefficient

CLUSTERING COEFFICIENT

- **Clustering coefficient:** The clustering coefficient captures the degree to which the neighbors of a given node link to each other, i.e. what fraction of your neighbors are connected?
  - Node i with degree k_i
  - C_i in [0,1]
  - L_i number of links between the k_i of node i

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



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

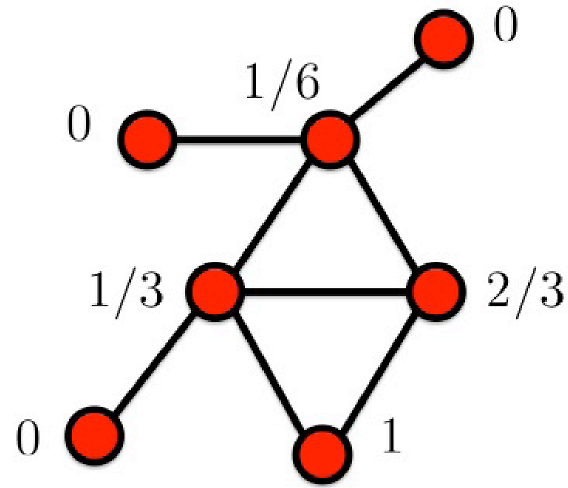
Watts & Strogatz, Nature 1998.

Clustering coefficient:

what fraction of your neighbors are connected?

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

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



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

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

The degree of clustering of a whole network is captured by the *average clustering coefficient*, $\langle C \rangle$, representing the average of *Ci* over all nodes *i* = 1,...,N

Watts & Strogatz, Nature 1998.

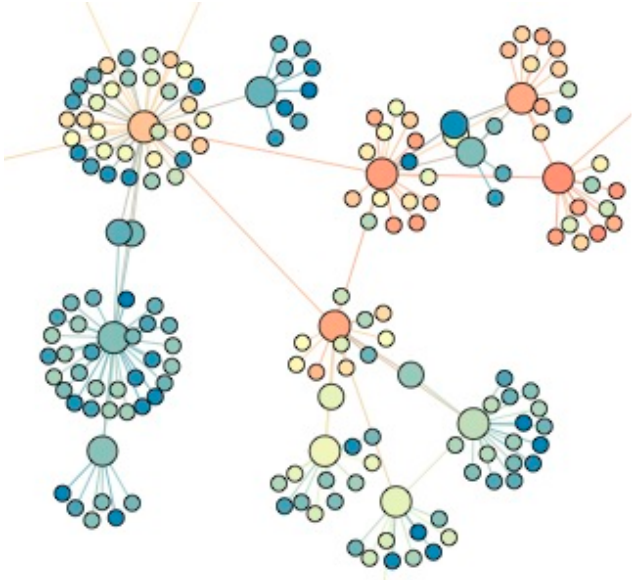
GLOBAL CLUSTERING COEFFICIENT

In the network literature we ocasionally encounter the *global clustering coefficient*, which measures the total number of closed triangles in a network. Indeed, L_i in the previous equation is the number of triangles that node *i* participates in, as each link between two neighbors of node *i* closes a triangle. Hence the degree of a network's global clustering can be also captured by the *global clustering coefficient*, defined as

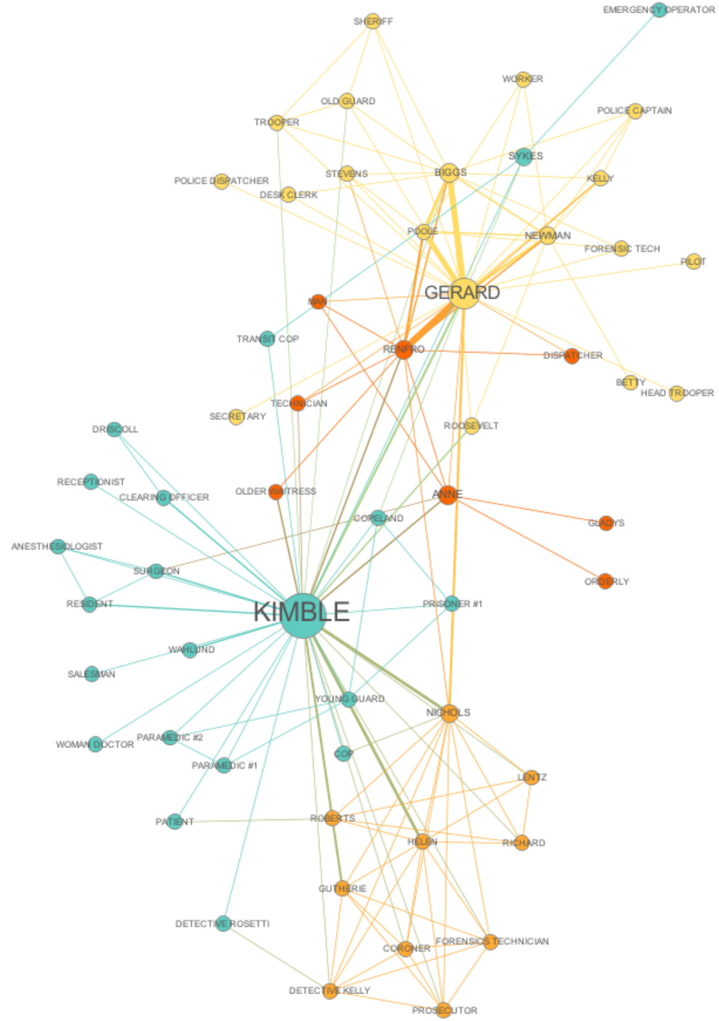
$$C_\Delta = \frac{3 \times Number\ Of\ Triangles}{Number\ Of\ Connected\ Triples}$$

Note that the average clustering coefficient <*C*> defined in and the global clustering coefficient are not equivalent.

CENTRALITY

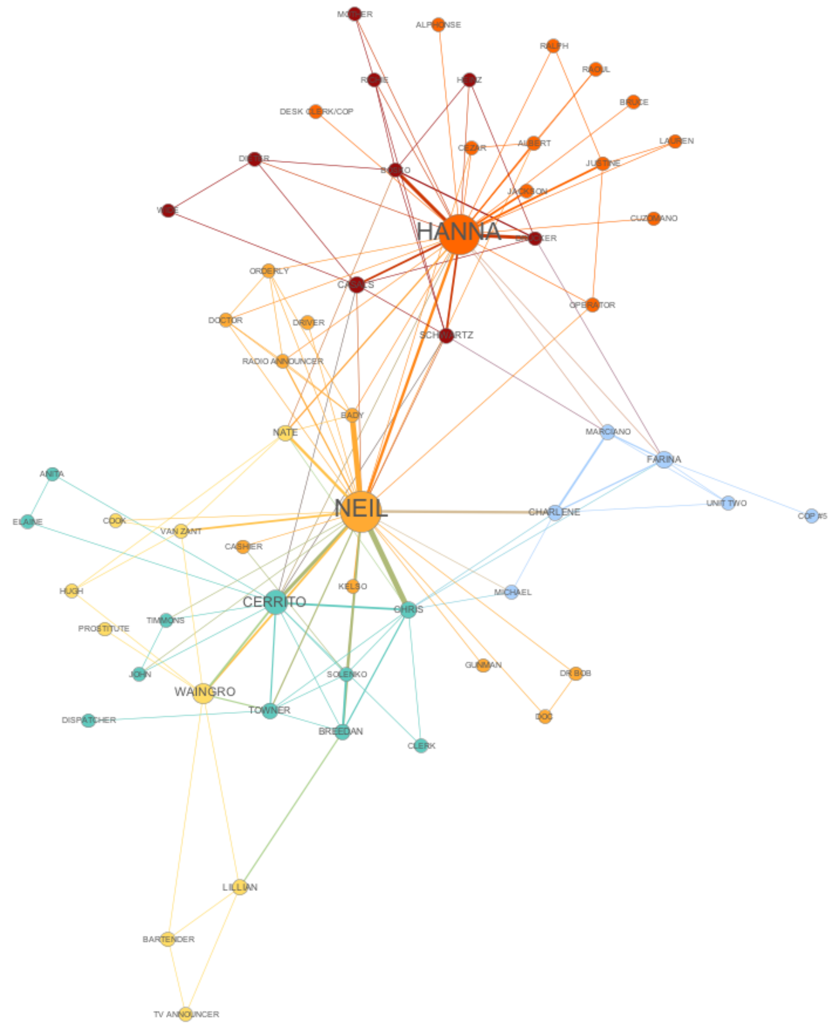


The Fugitive (1993)



moviegalaxies.com

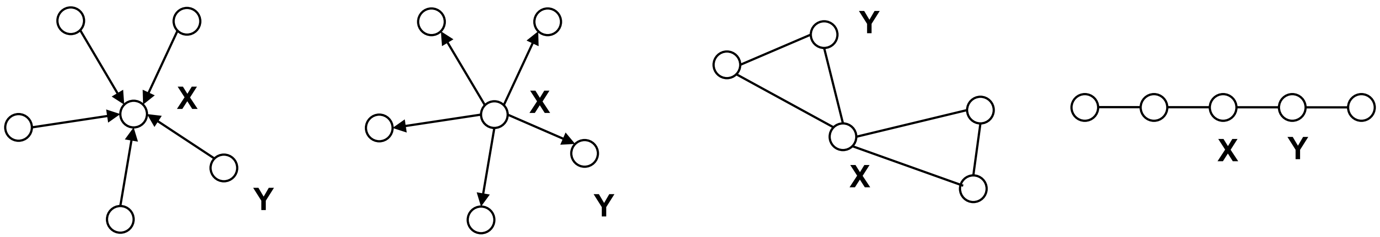
Heat (1995)



moviegalaxies.com

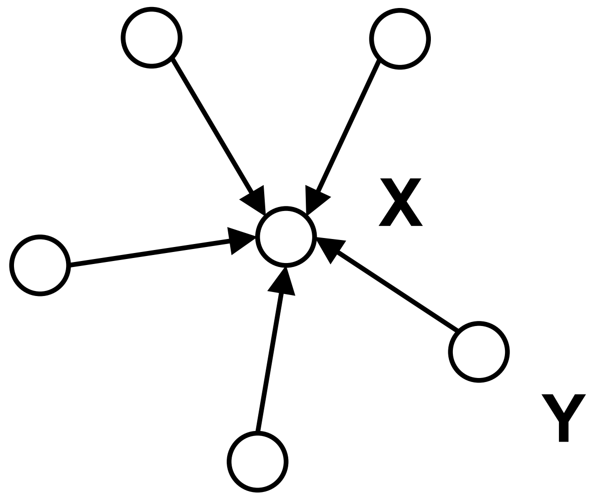
VARIOUS NOTIONS OF CENTRALITY

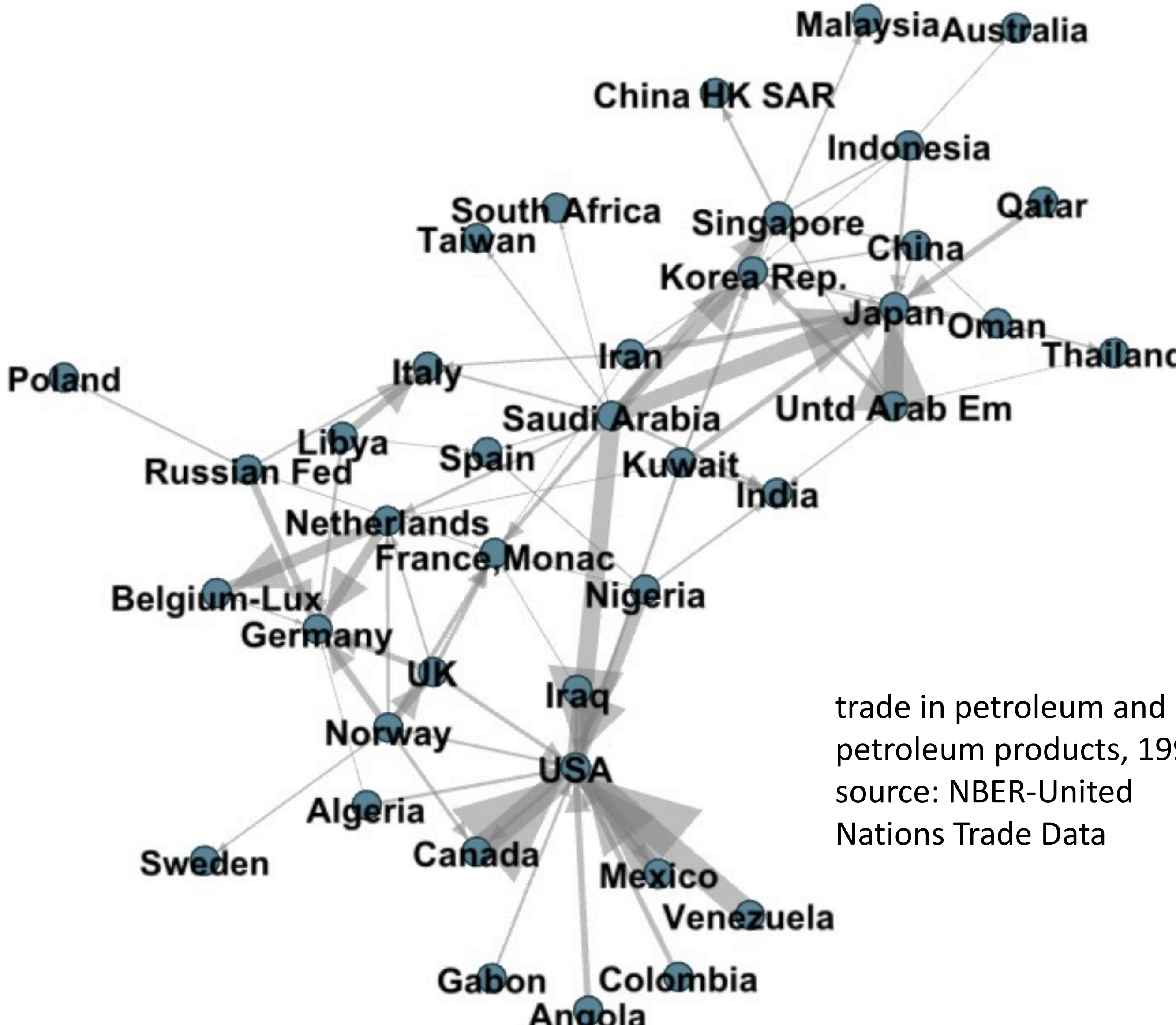
In each of the following networks, X has higher centrality than Y according to a particular measure



indegree outdegree betweenness closeness

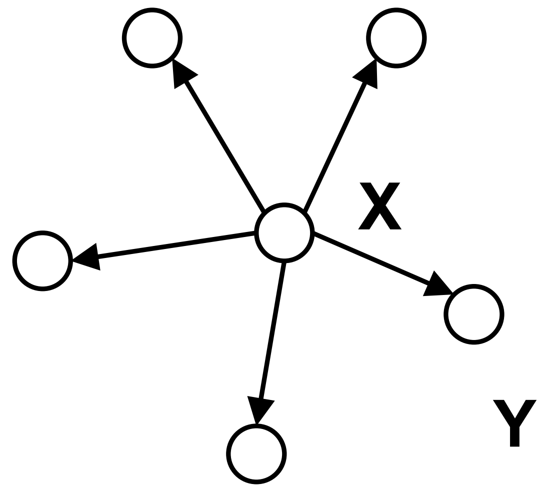
IN-DEGREE CENTRALITY

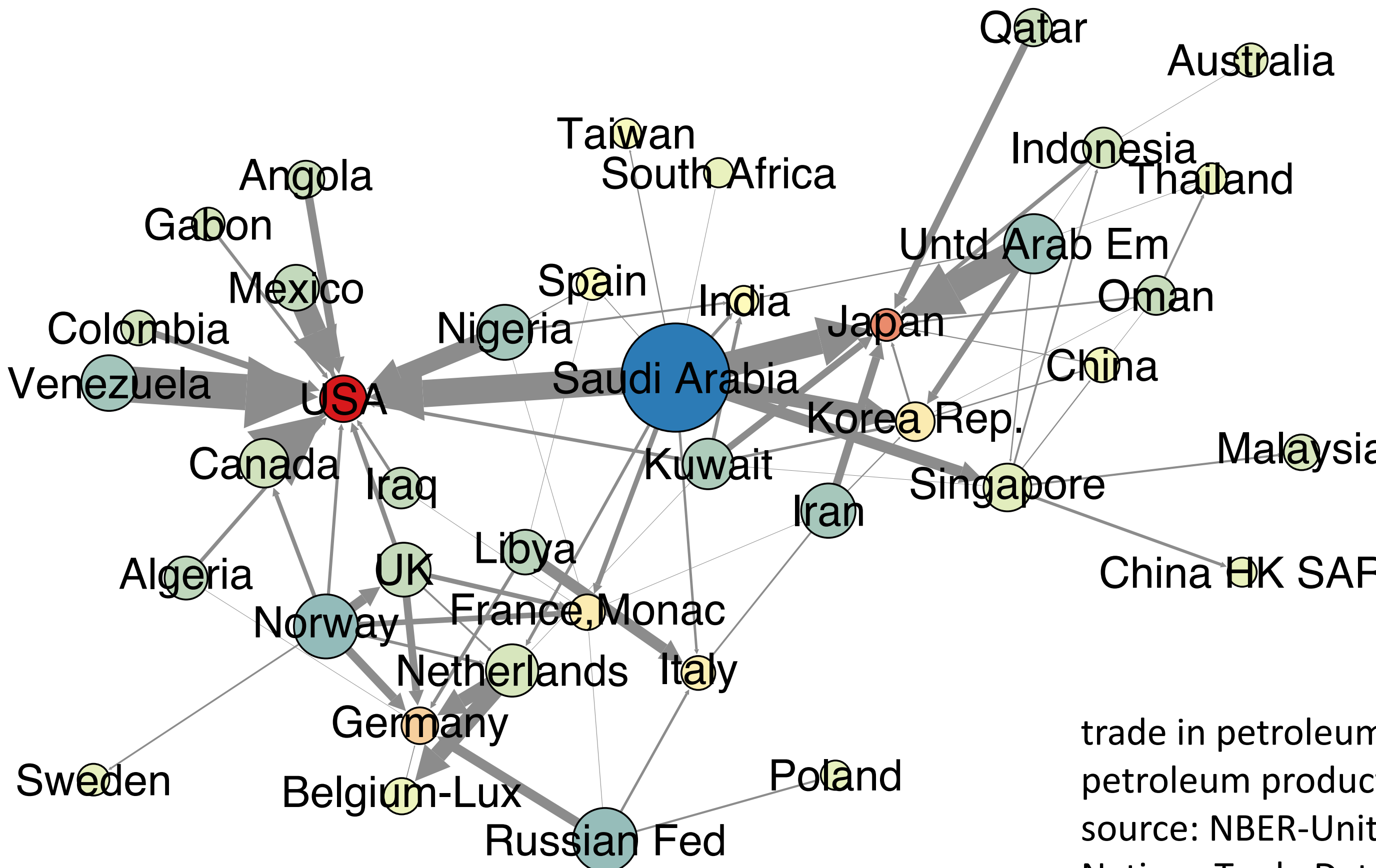




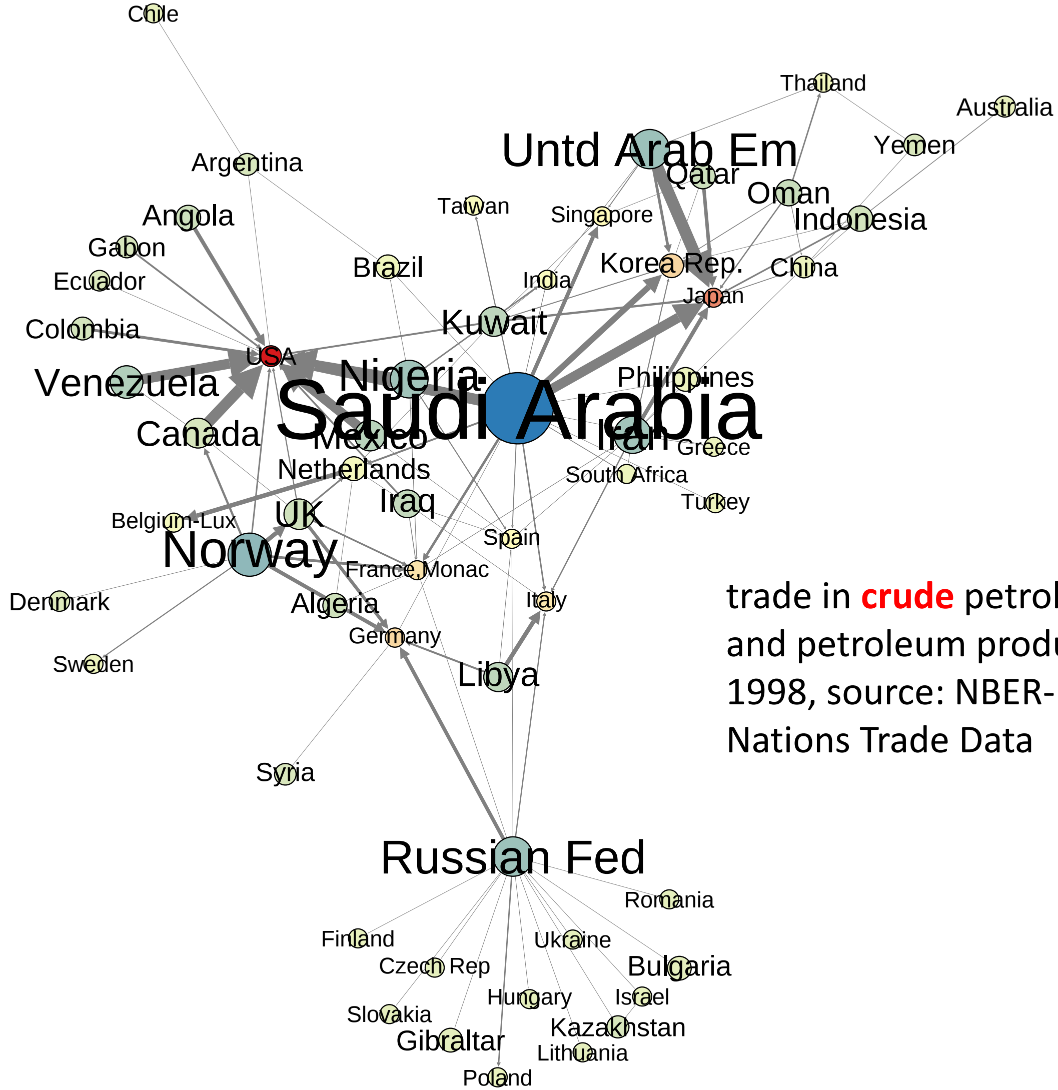
trade in petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data

OUT-DEGREE CENTRALITY





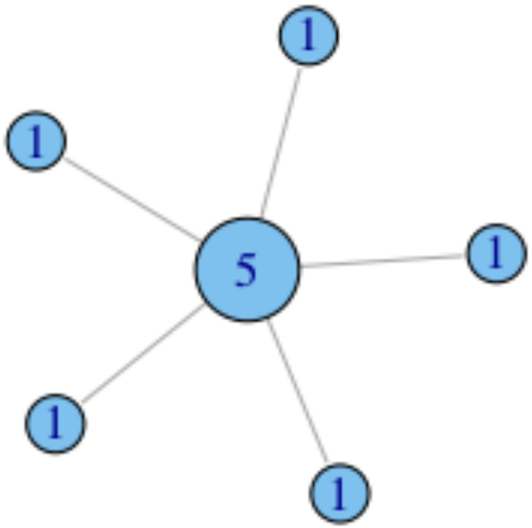
trade in petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data



trade in **crude** petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data

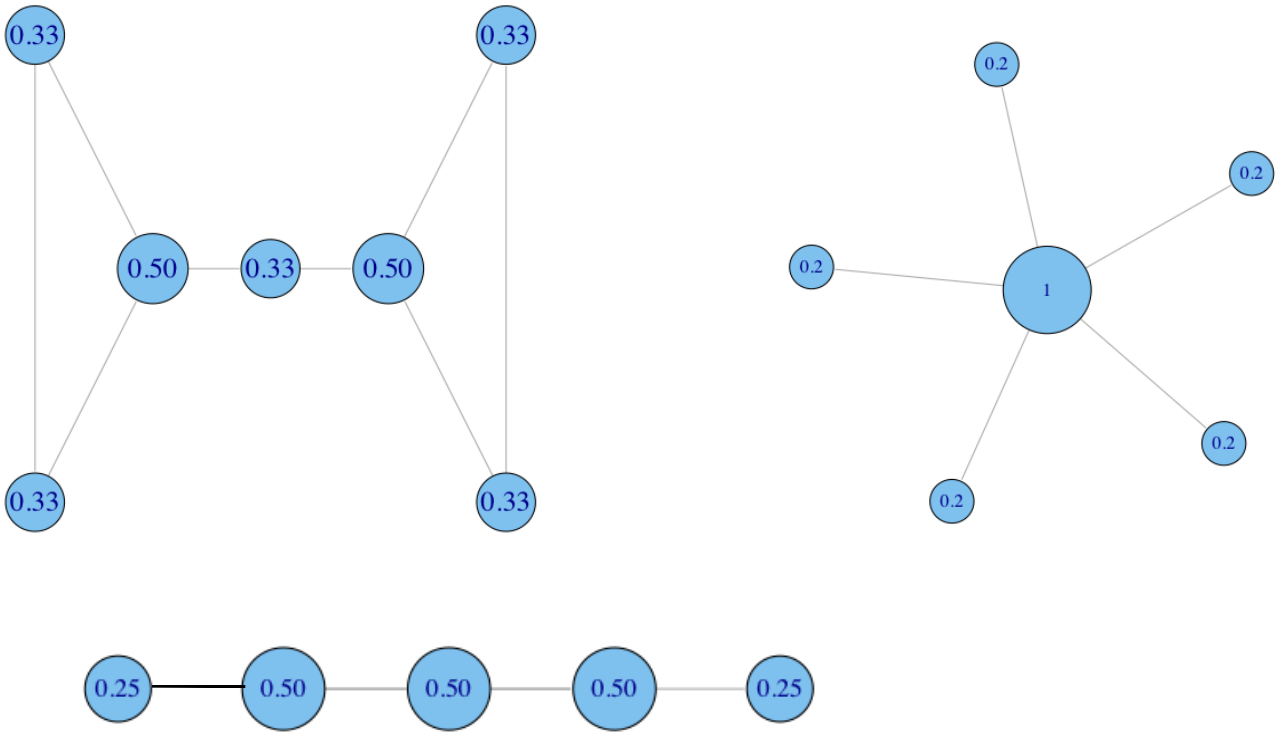
DEGREE CENTRALITY

Undirected degree, e.g. nodes with more friends are more central.



DEGREE CENTRALITY: NORMALIZATION

divide degree by the max. possible, i.e. (N-1)



DEGREE CENTRALITY

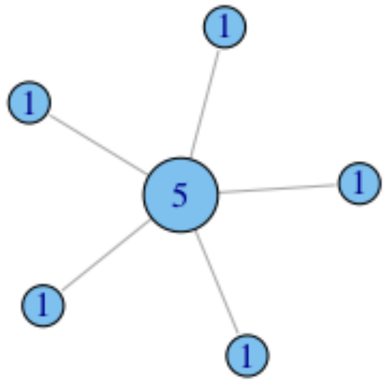
How much variation is there in the centrality scores among the nodes?

Freeman's general formula for centralization (can use other metrics, e.g. gini coefficient or standard deviation):

maximum value in the network

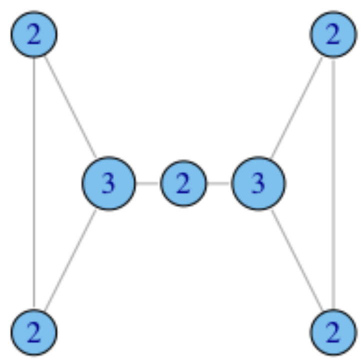
$$C_D = \frac{\sum_{i=1}^{g}\left[\boxed{C_D(n^*)} - C_D(i)\right]}{[(N-1)(N-2)]}$$

DEGREE CENTRALITY: TOY NETWORKS EXAMPLES



$C_D = 1.0$

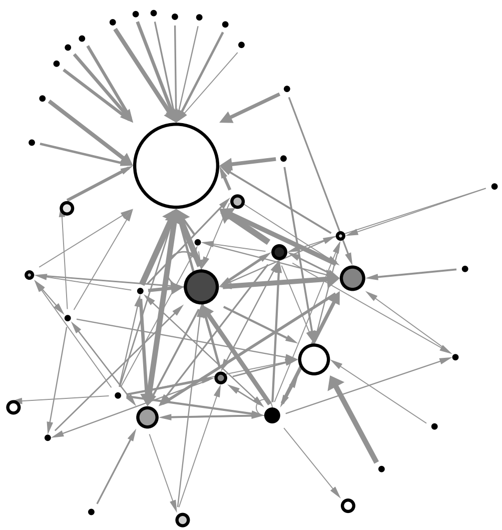
$C_D = 0.167$

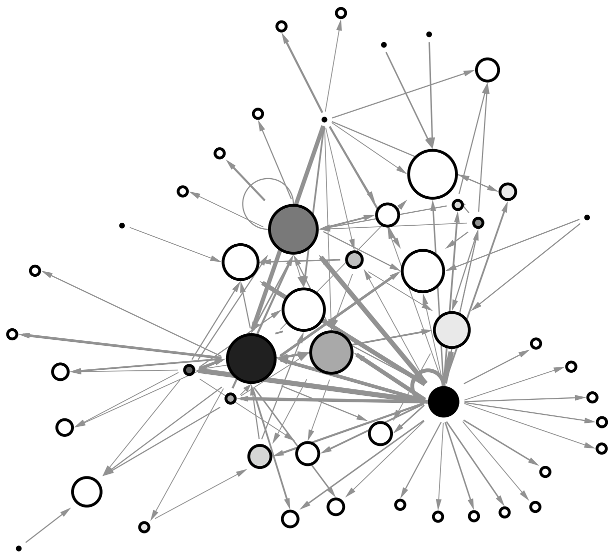


$C_D = 0.167$

DEGREE CENTRALITY: REAL-WORLD NETWORKS EXAMPLES

example financial trading networks



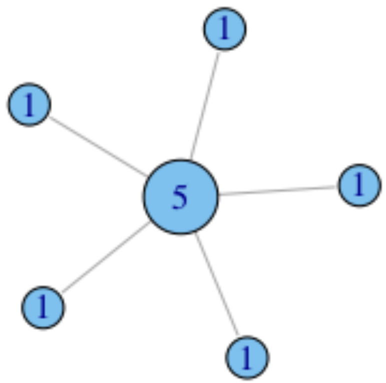


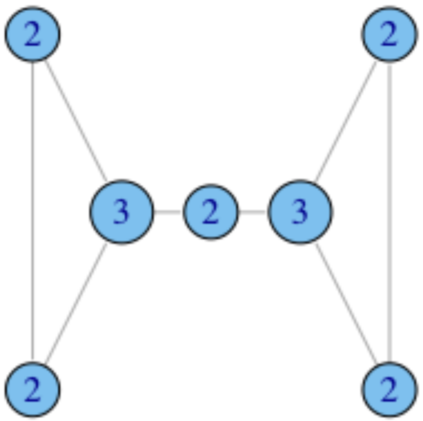
high in-centralization: one node buying from many others

low in-centralization: buying is more evenly distributed

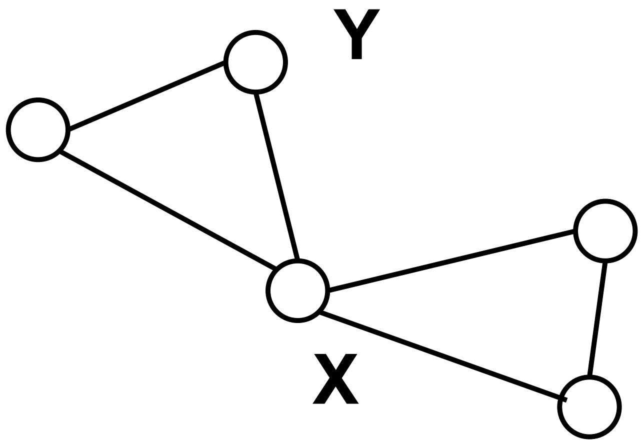
DEGREE CENTRALITY: WHAT IT DOES NOT CAPTURE?

In what ways does degree fail to capture centrality in the following graphs?



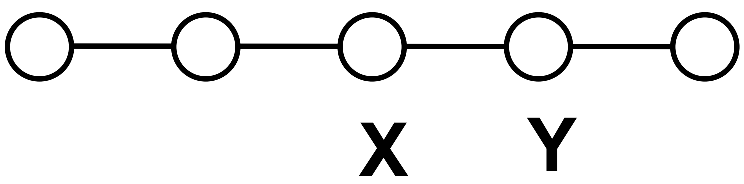


DEGREE CENTRALITY: BROKERAGE NOT CAPTURED



BETWEENNES CENTRALITY

- intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?



BETWEENNES CENTRALITY: DEFINITION

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$

Where g_{jk} = the total number of shortest paths connecting *jk*
$g_{jk}(i)$ = the number of those path that pass through *i* is on.

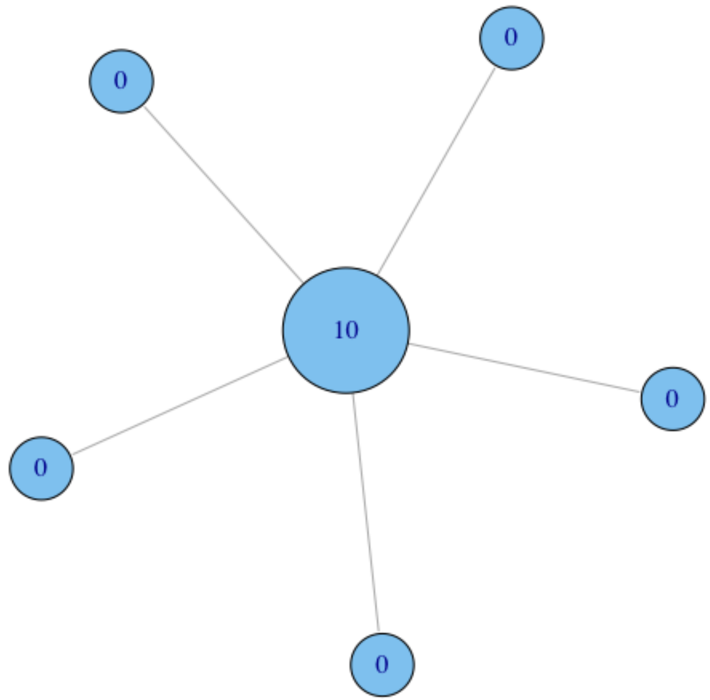
Note that the betweenness centrality of a node scales with the number of pairs of nodes not including j.
The betweenness value for each node *n* is normalized by dividing by the number of node pairs excluding *n*: (N-1)(N-2)/2, where N is the total number of nodes in the connected component that *n* belongs to. Thus, the betweenness centrality of each node is a number between 0 and 1.

$$C_B^{'}(i) = C_B(i) / \boxed{[(n-1)(n-2)/2]}$$

number of pairs of vertices excluding the vertex itself

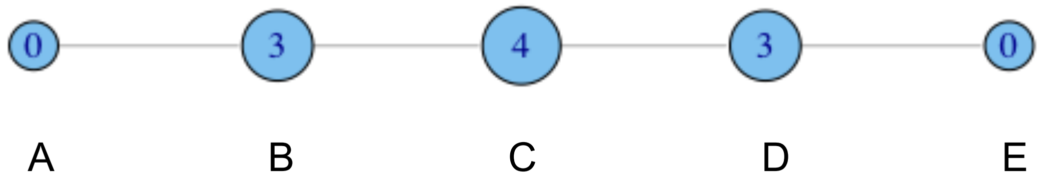
BETWEENNES CENTRALITY ON TOY NETWORKS

- non-normalized version:



BETWEENNES CENTRALITY ON TOY NETWORKS

- non-normalized version:

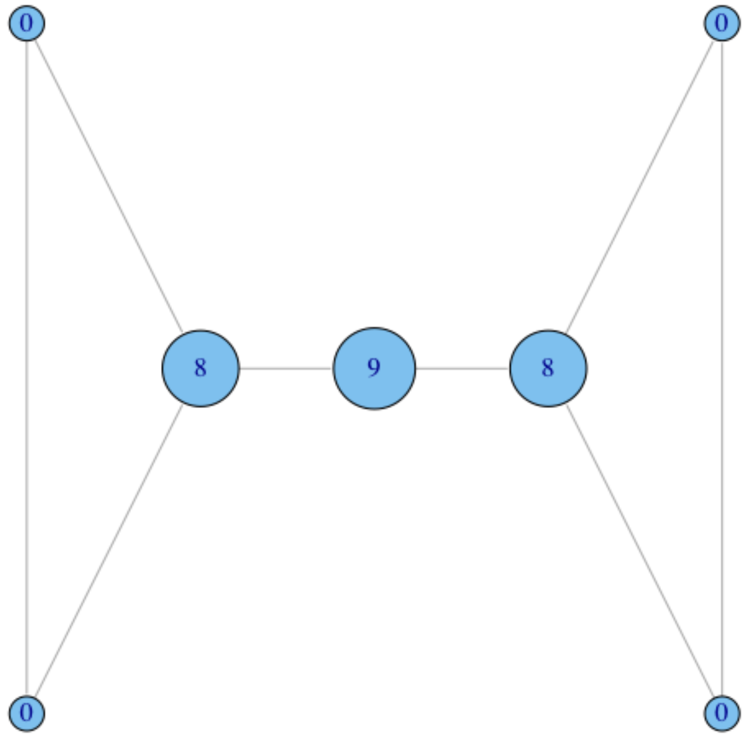


- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

- note that there are no alternate paths for these pairs to take, so C gets full credit

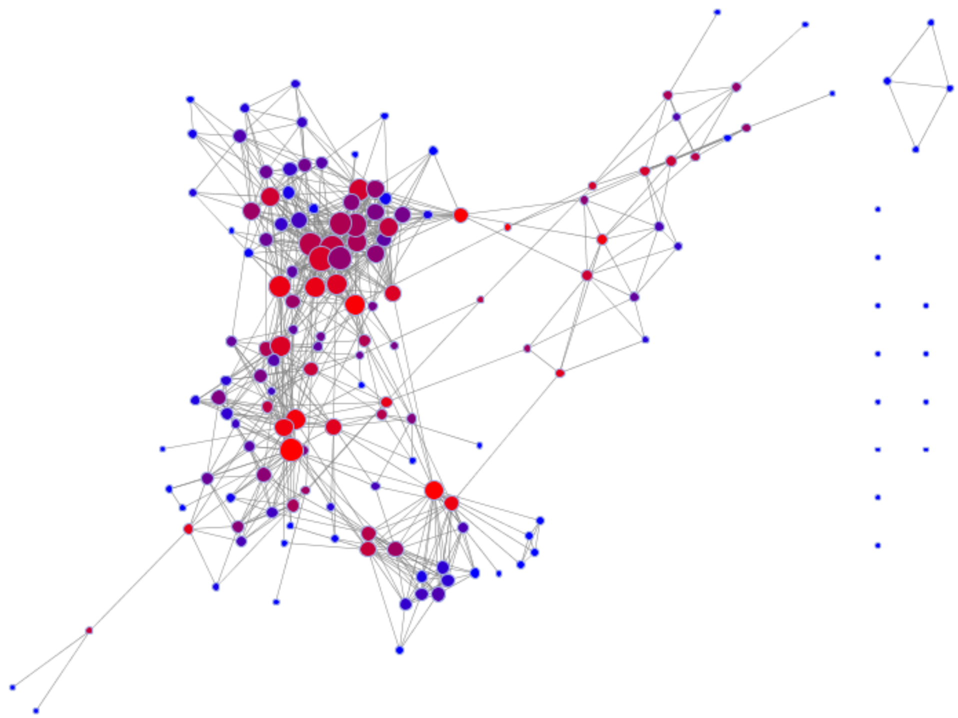
BETWEENNES CENTRALITY ON TOY NETWORKS

- non-normalized version:



BETWEENNES CENTRALITY: AN EXAMPLE

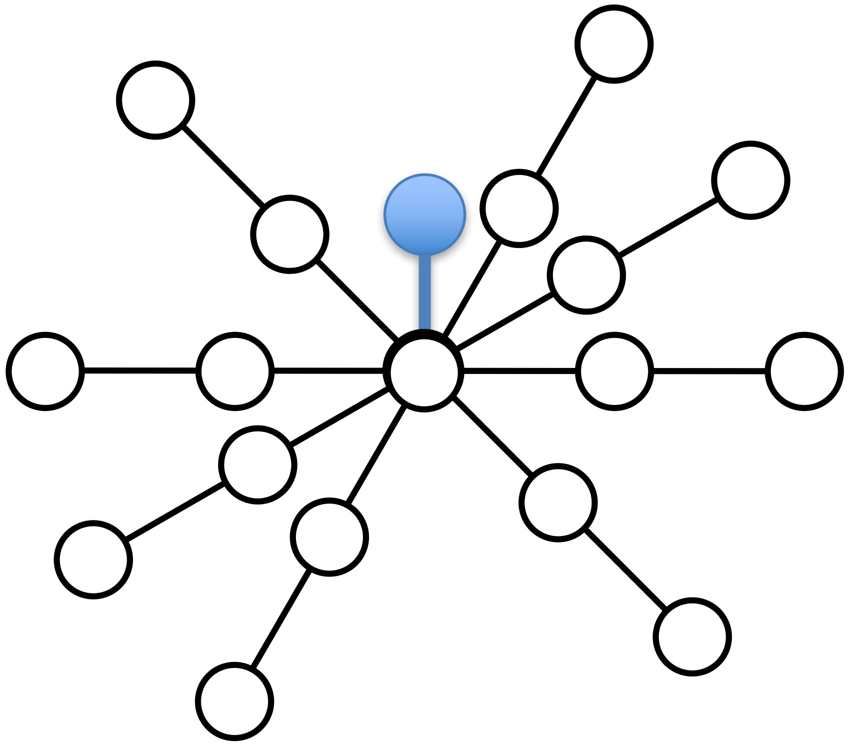
my Facebook network: nodes are sized by degree, and colored by betweenness.



CLOSENESS CENTRALITY

- What if it’s not so important to have many direct friends?
- Or be “between” others
- But one still wants to be in the “middle” of things, not too far from the center

CLOSENESS CENTRALITY: YOU DO NOT NEED TO BE IN A BROKERAGE POSITION



CLOSENESS CENTRALITY

Closeness is based on the length of the average shortest path between a node and all other nodes in the network

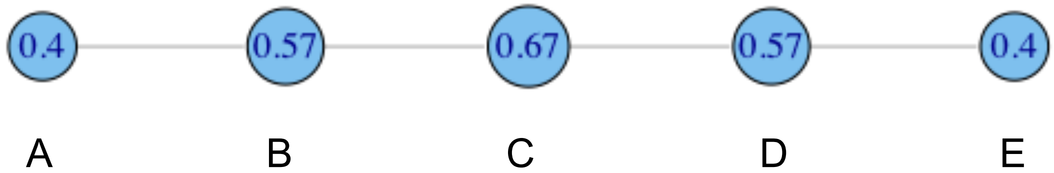
Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Normalized Closeness Centrality

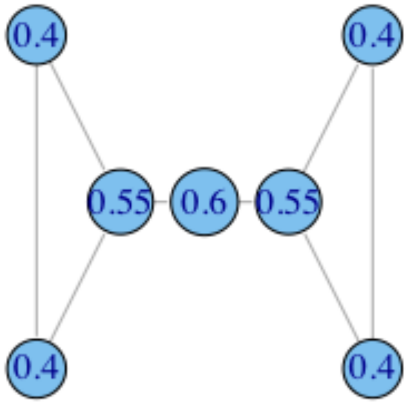
$$C_C^{'}(i) = (C_C(i))(N-1)$$

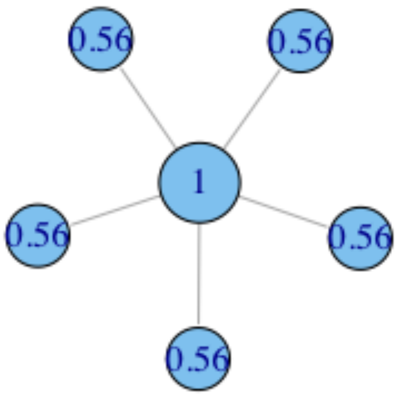
CLOSENESS CENTRALITY: TOY EXAMPLES



$$C_c^{'}(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1} \right]^{-1} = \left[\frac{1+2+3+4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

CLOSENESS CENTRALITY: TOY EXAMPLES





SUMMARY

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

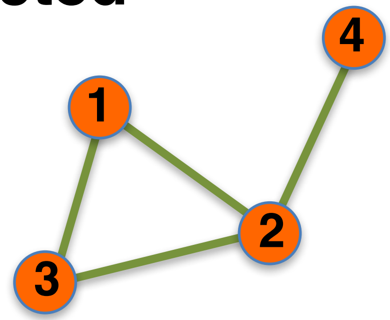
Degree distribution: **P(k)**

Path length: $\langle d \rangle$

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

Measures of centrality

Undirected



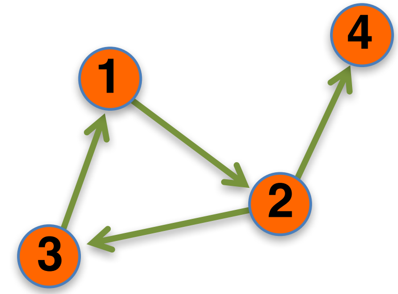
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions

Directed



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

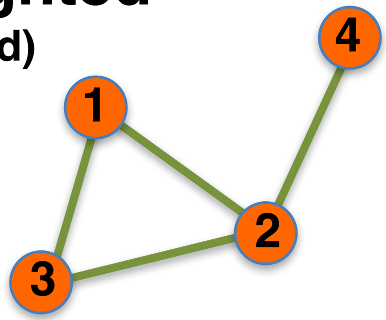
$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

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

WWW, citation networks

Unweighted

(undirected)



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

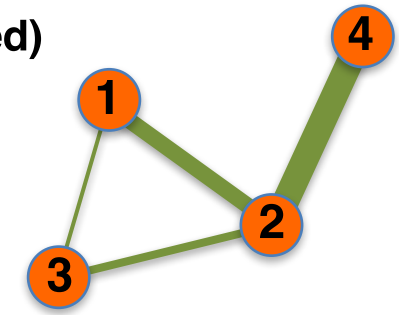
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

protein-protein interactions, www

Weighted

(undirected)



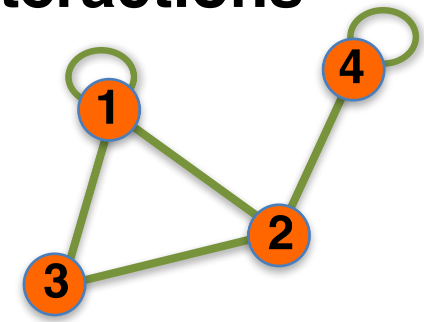
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

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

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

Call Graph, metabolic networks

Self-interactions



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

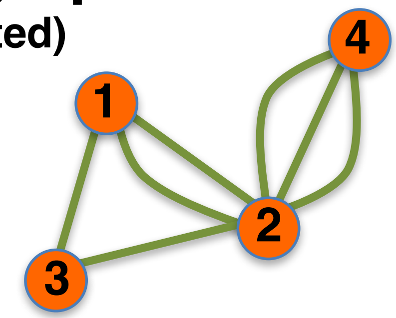
$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

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

Protein interaction network, www

Multigraph

(undirected)



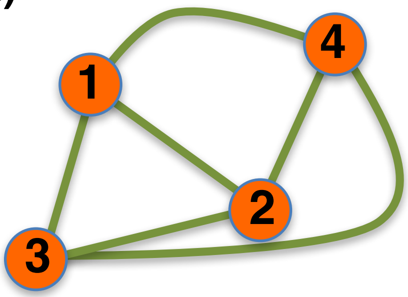
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

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

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

Social networks, collaboration networks

Complete Graph
(undirected)

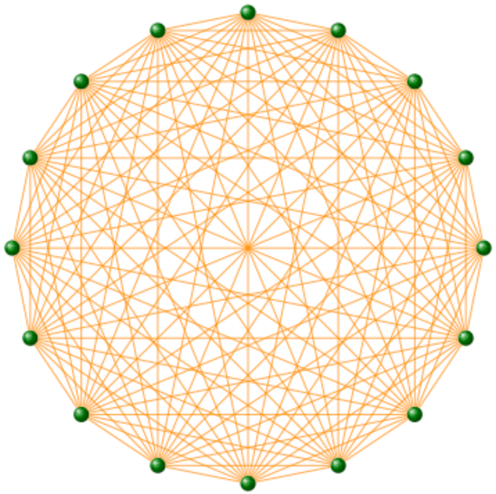


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions



GRAPHOLOGY: Real networks can have multiple characteristics

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

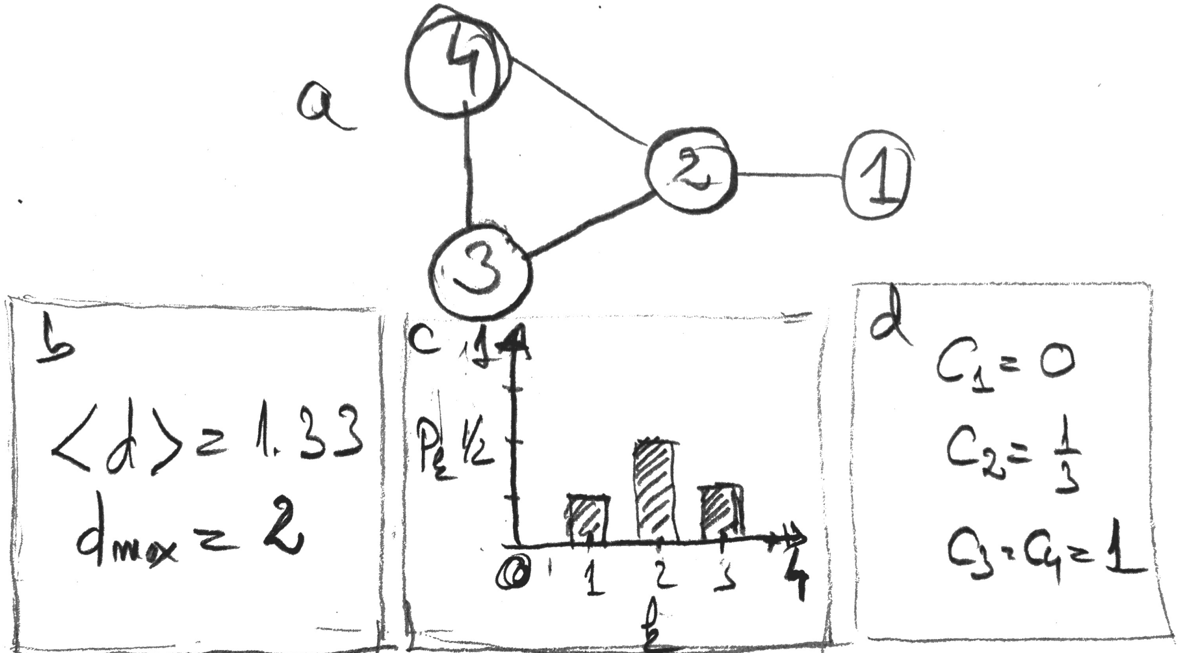
Protein Interactions > **u**ndirected unweighted with self-interactions

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

Mobile phone calls > directed, weighted.

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

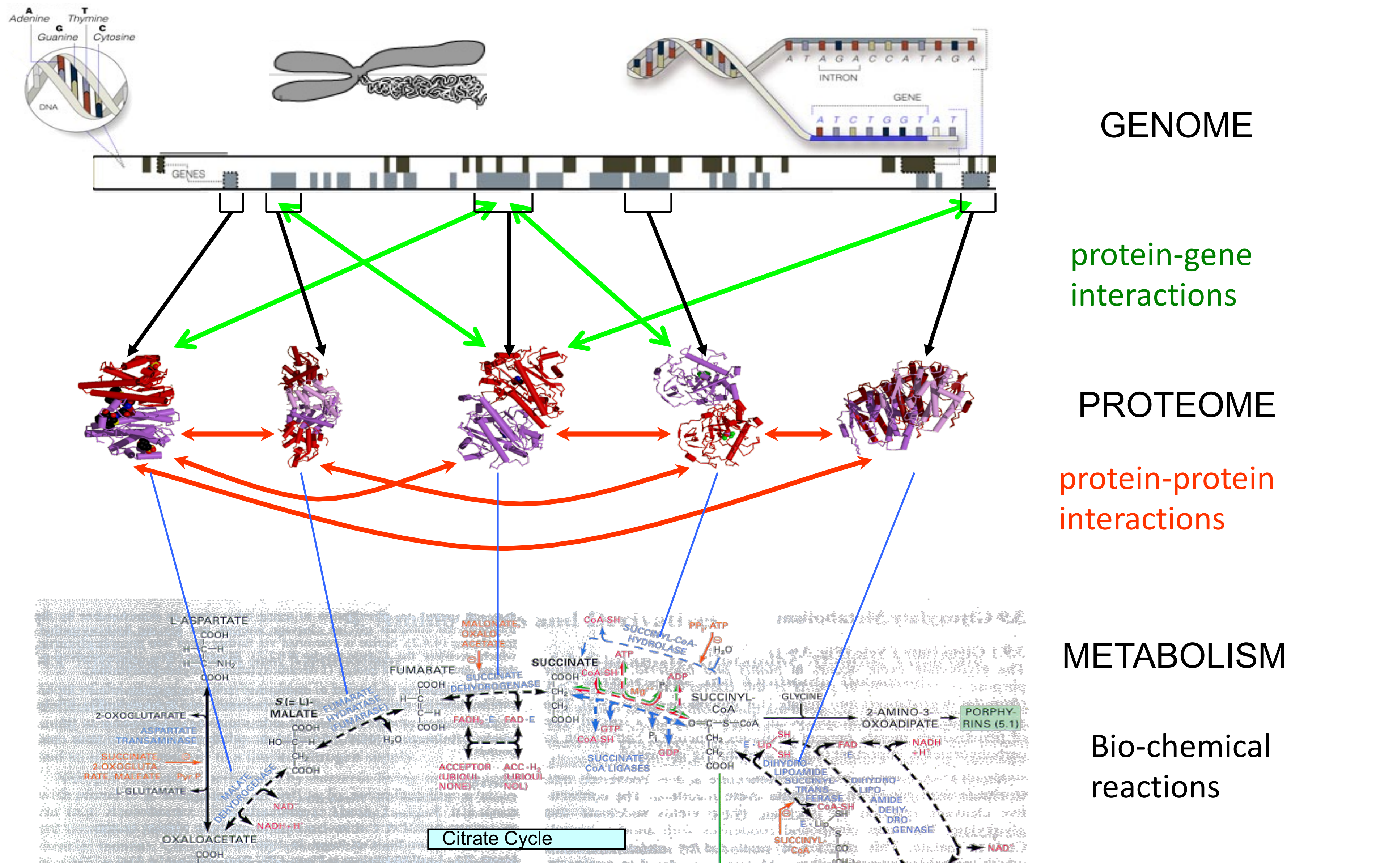
THREE CENTRAL QUANTITIES IN NETWORK SCIENCE



A. Degree distribution: p_k

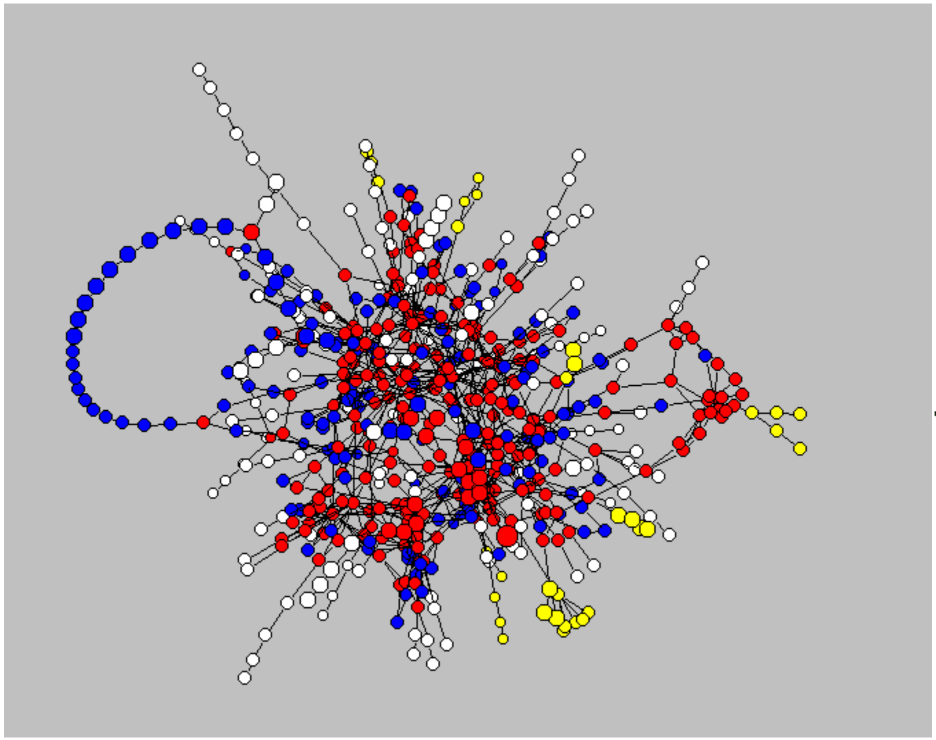
B. Path length: $\langle d \rangle$

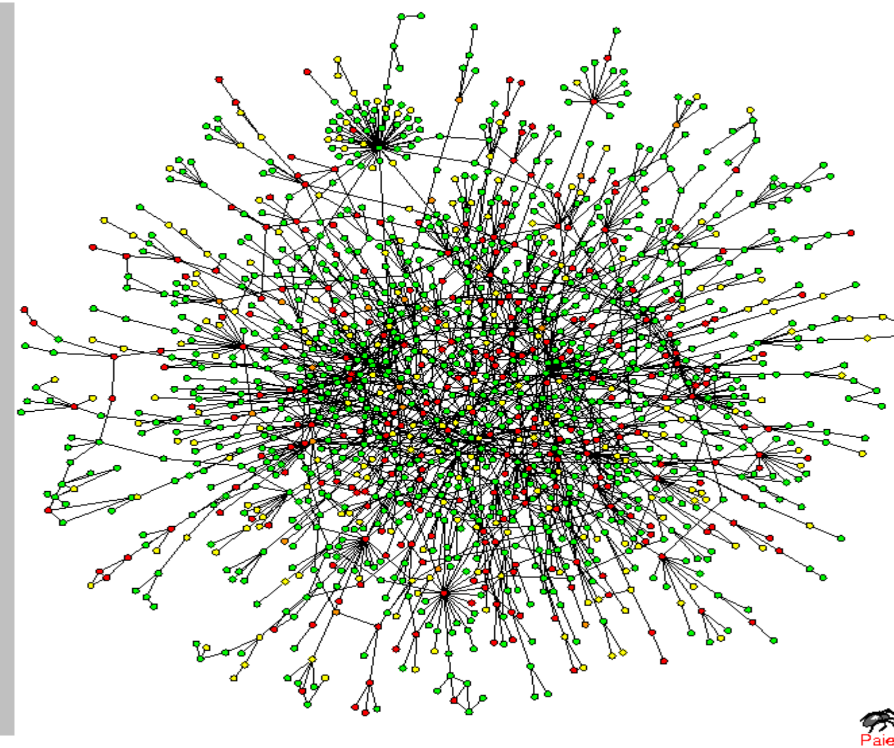
C. Clustering coefficient: $$C_i = \frac{2e_i}{k_i(k_i - 1)}$$


Adenine
Thymine
Guanine
Cytosine
DNA
INTRON
GENE
GENES
GENOME
protein-gene interactions
PROTEOME
protein-protein interactions
METABOLISM
Bio-chemical reactions
L-ASPARTATE
2-OXOGLUTARATE
L-GLUTAMATE
OXALOACETATE
MALATE
FUMARATE
SUCCINATE
SUCCINYL-CoA
GLYCINE
2-AMINO-3-OXOADIPATE
PORPHYRINS (5.1)
Citrate Cycle

Metabolic Network

Protein Interactions



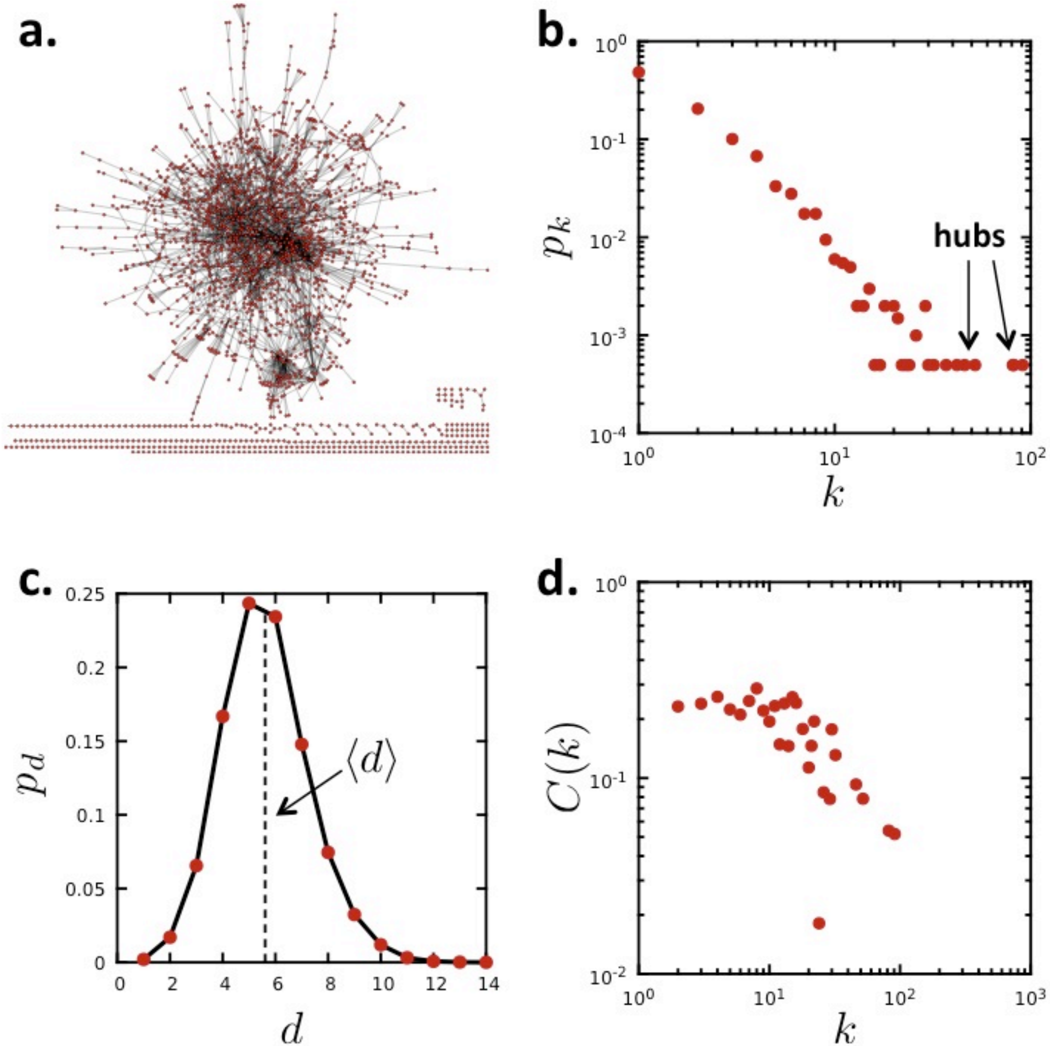


A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

Undirected network
N=2,018 proteins as nodes
L=2,930 binding interactions as links.
Average degree $\langle k \rangle = 2.90$.

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

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



<u>Undirected network</u>

N=2,018 proteins as nodes

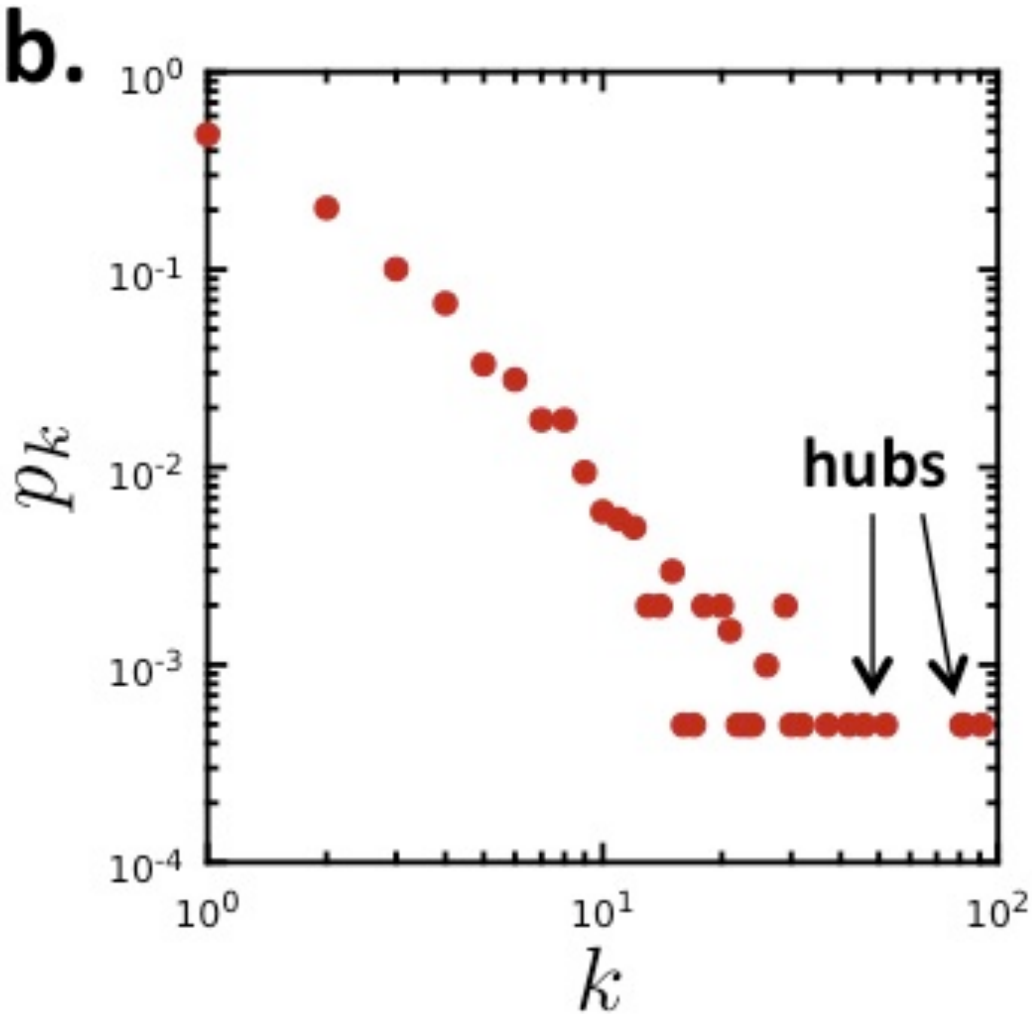
L=2,930 binding interactions as links.

Average degree <k>=2.90.

<u>Not connected: </u>185 components

the largest (giant component) 1,647 nodes

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

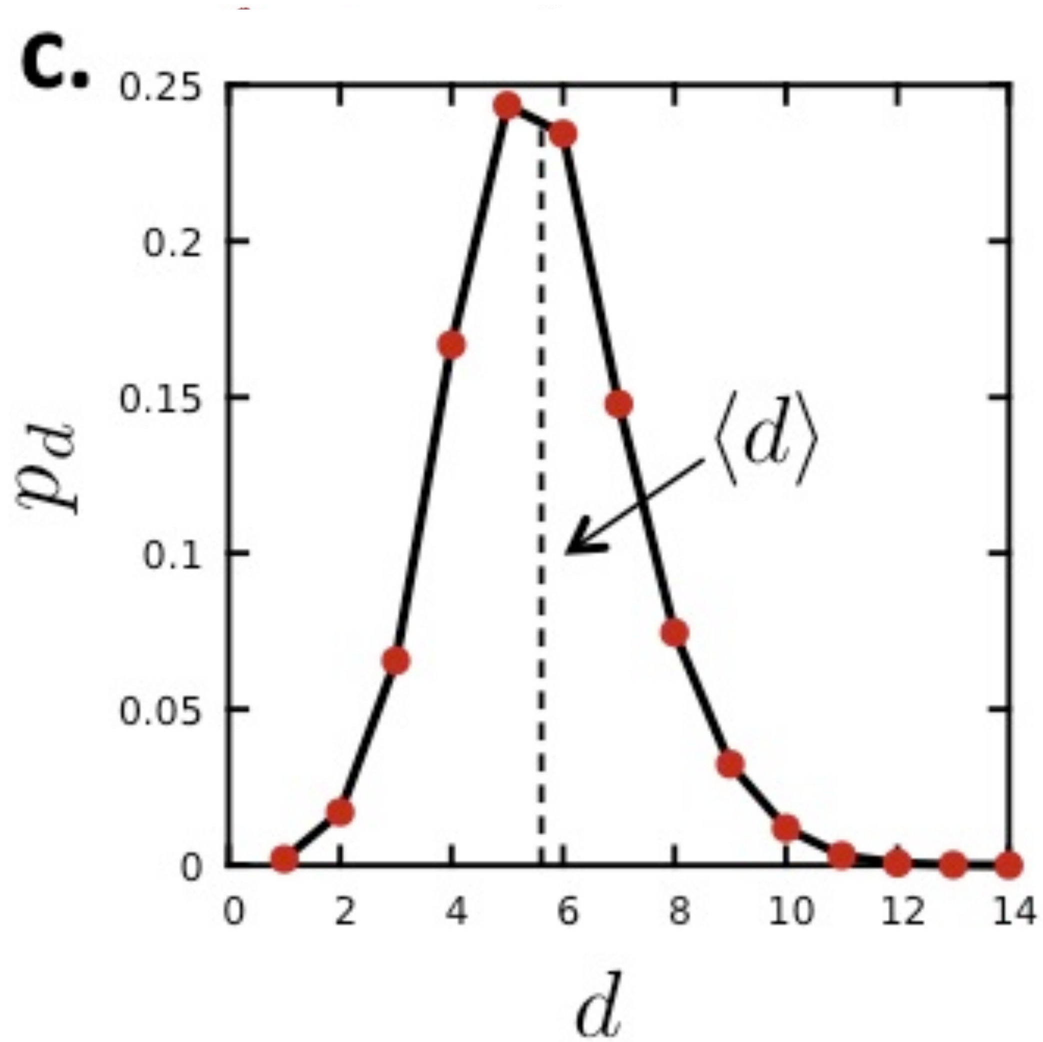


p_k is the probability that a node has degree **k**.

N_k = # nodes with degree k

$p_k = N_k / N$

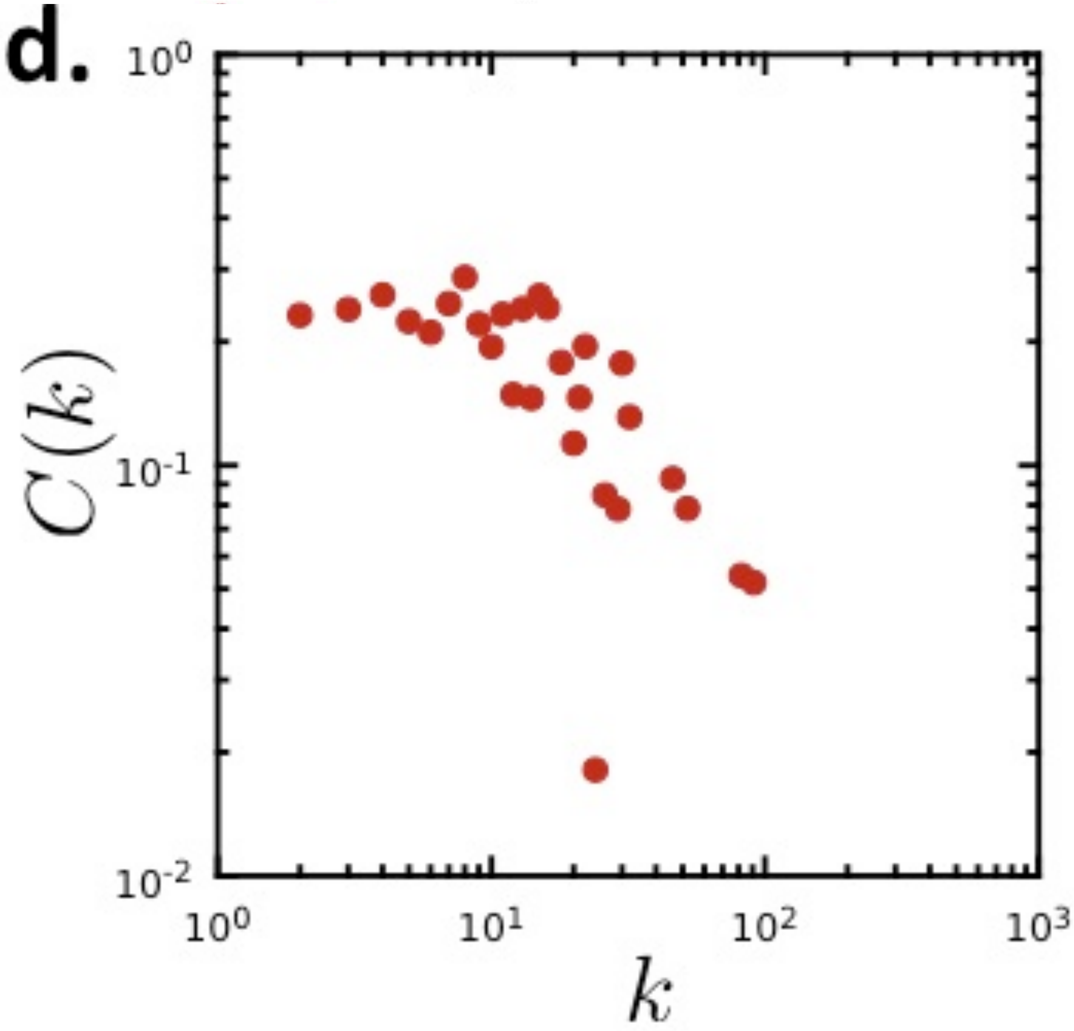
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

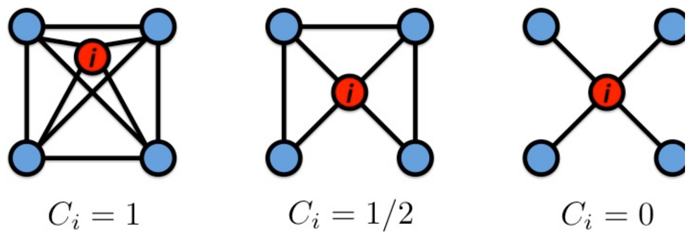


d_{max}=14

<d>=5.61

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK





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

<C>=0.12