Digital epidemiology

Lesson 7

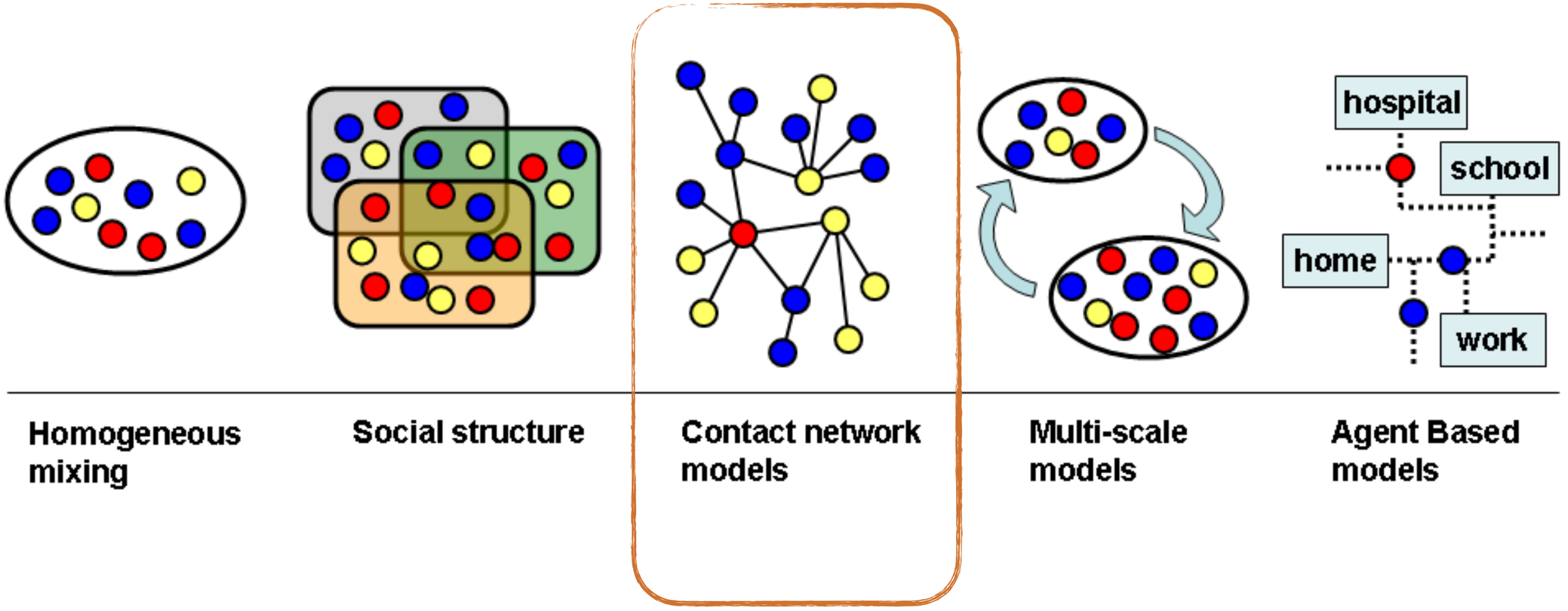
Michele Tizzoni
Dipartimento di Sociologia e Ricerca Sociale
Via Verdi 26, Trento
Ufficio 6, 3 piano

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

C2S2 Center for Computational Social Science and Human Dynamics

Models


hospital
school
home
work
Homogeneous mixing
Social structure
Contact network models
Multi-scale models
Agent Based models

Network theory: basics

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

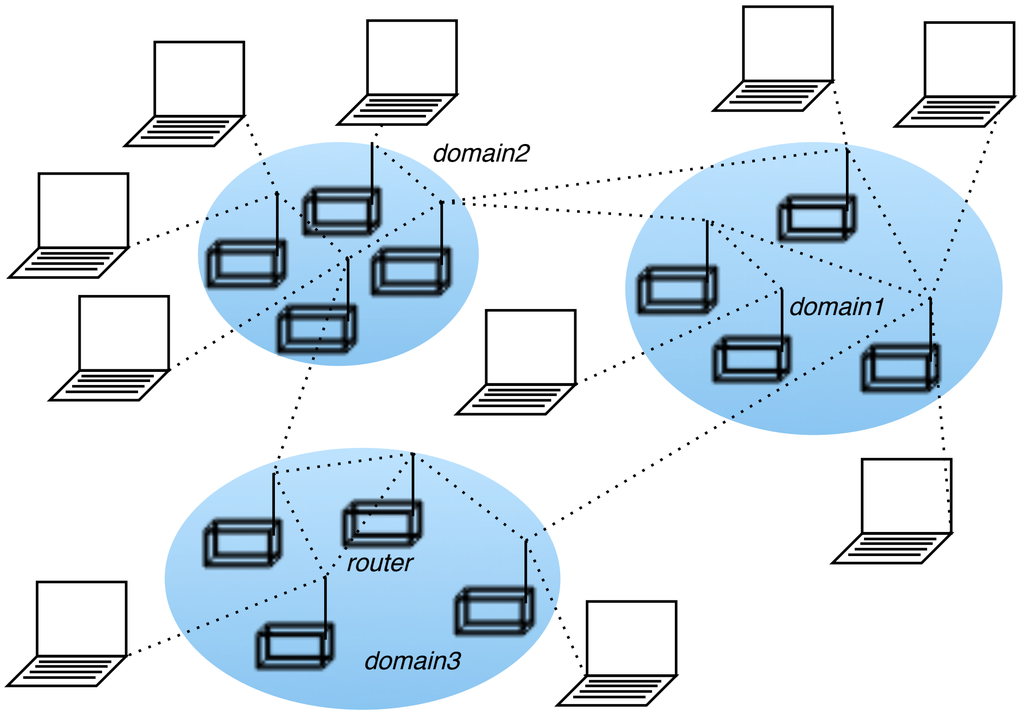
Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving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

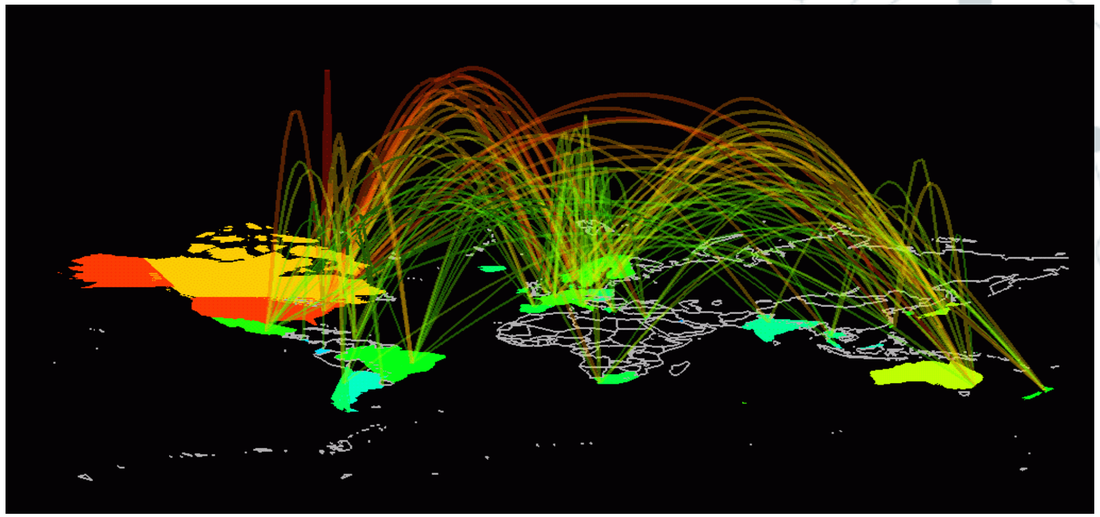
Source: John L. Casti, Encyclopædia Britannica

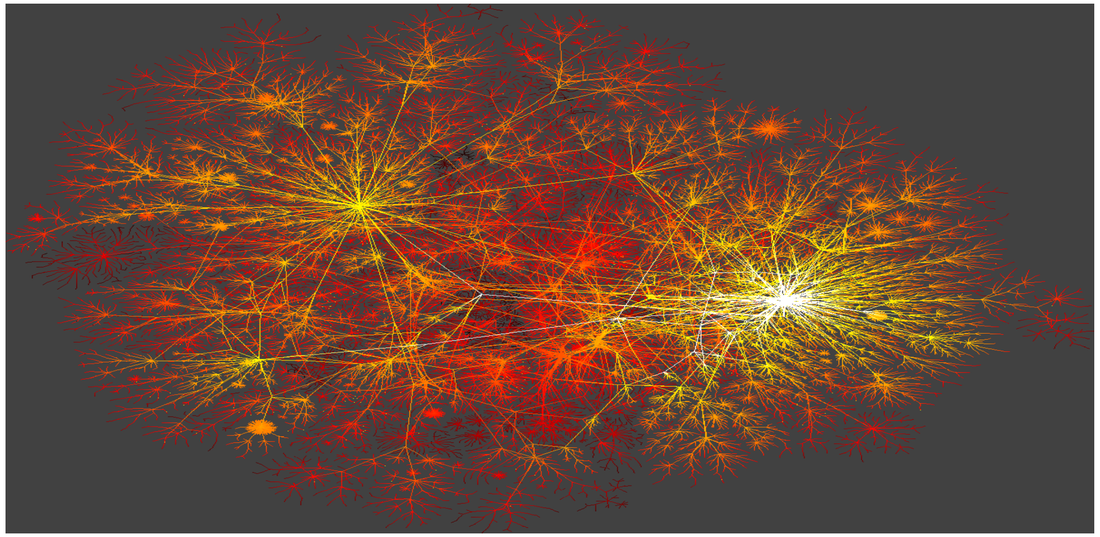
Complexity

Behind each complex system
there is a **network**
that defines the interactions
between the components

Internet

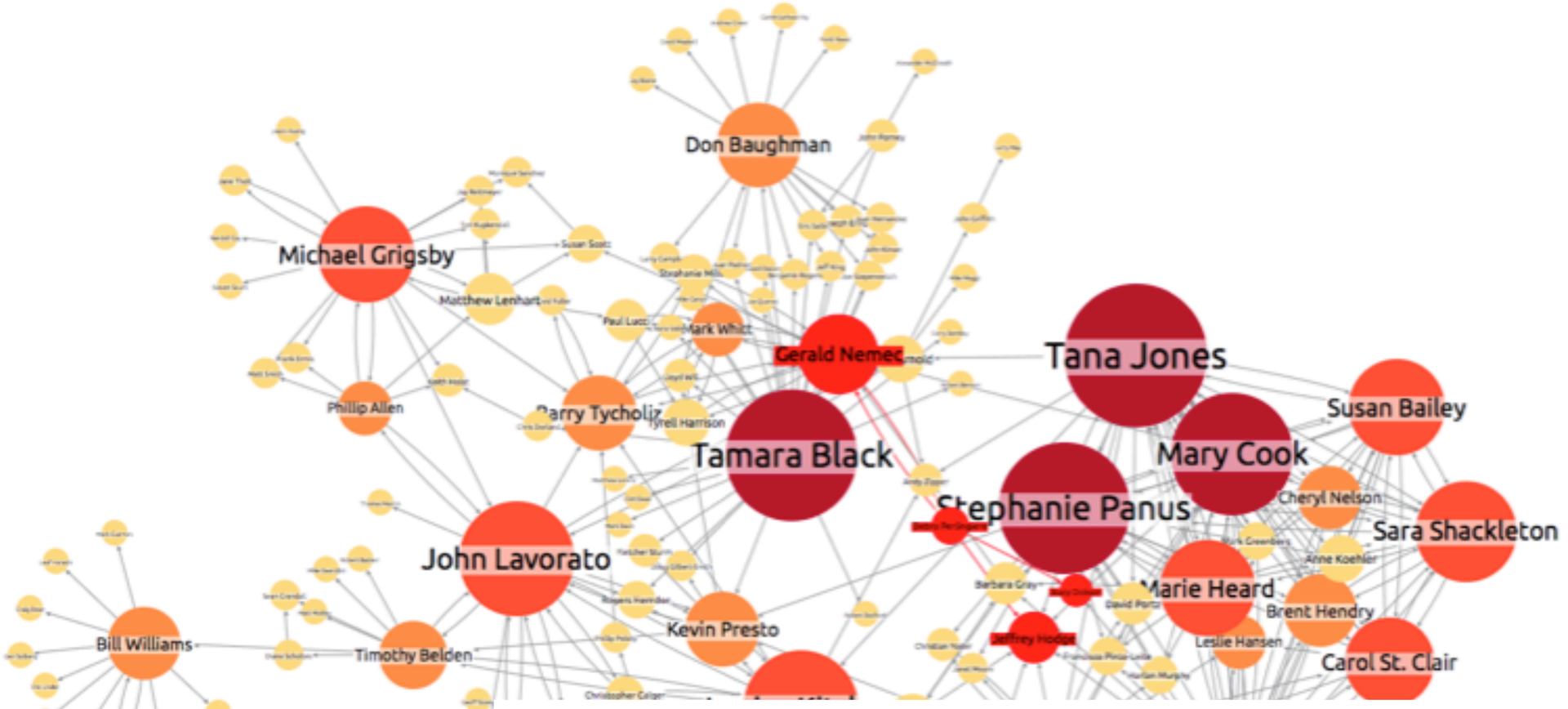






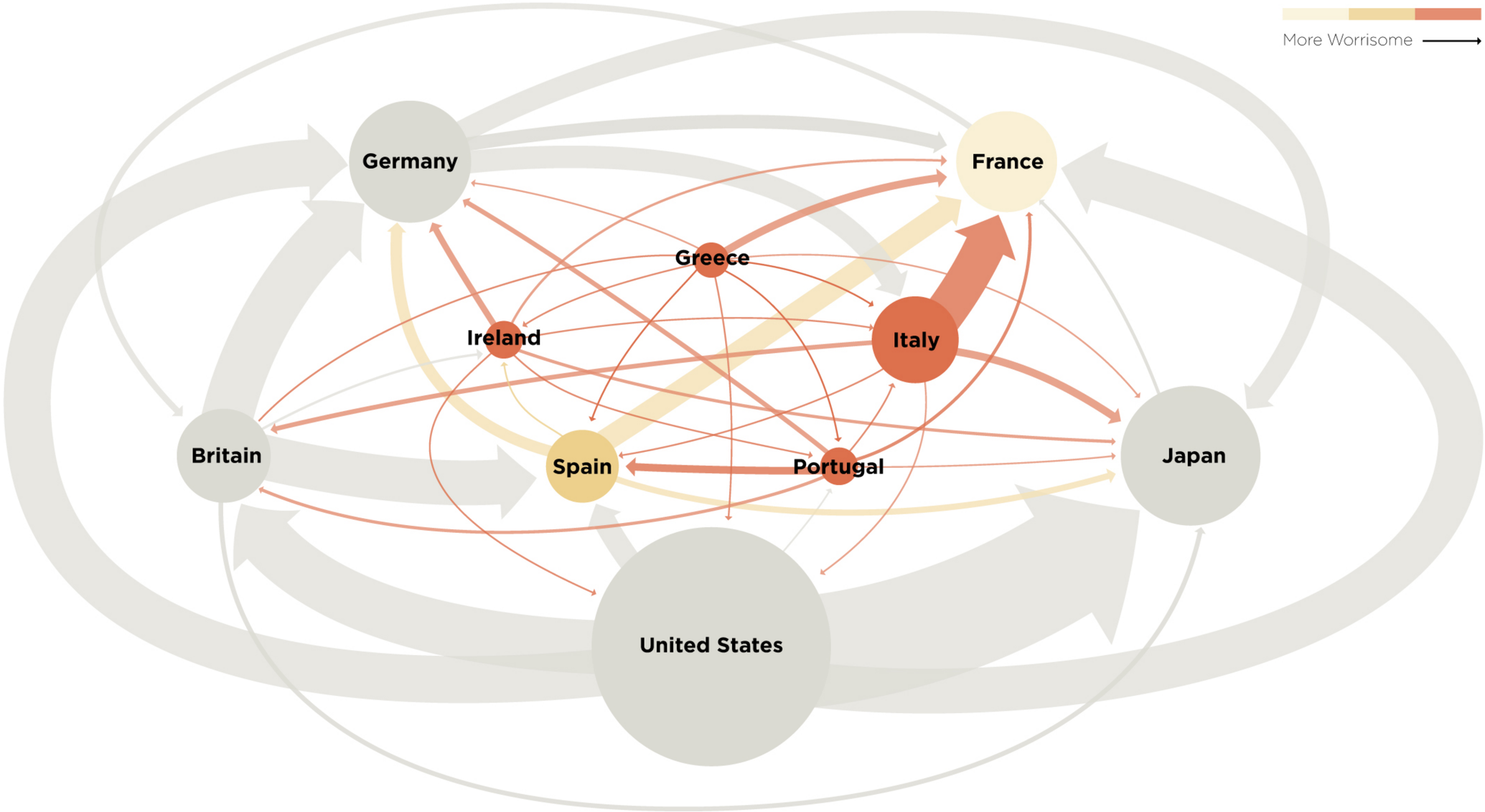
online interactions

Email communication



The **Enron email dataset** contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.

Financial networks



many more…

- **biological** (protein-protein, brain)
- **infrastructure/transport** (air travel, road networks)
- **mobility networks** (human movements)
- **word co-occurence**
- **off-line interactions** (proximity, friendship)
- **scientific collaborations** (co-authorship, citations)

Network datasets

http://snap.stanford.edu

By *Jure Leskovec*

STANFORD UNIVERSITY

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- BIOSNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : Talk, editing, voting, and article data from Wikipedia
- Temporal networks : networks where edges have timestamps
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also available from SuiteSparse Matrix Collection by Tim Davis.

Network datasets

http://networks.skewed.de

Netzschleuder network catalogue, repository and centrifuge

Networks | Stats | API | Git | Issues | Contribute | Health | About

Tip: click on the table header to sort the list. Hover your mouse over it to obtain a legend.

Multiple regexp terms separated by '&' Search

Name	Title	Nodes	Edges	$\langle k \rangle$	σ_k	λ_h	τ	r	c	$\oslash$	S	Kind	Mode	n	Tags
7th_graders	Vickers 7th Graders (1981)	29	740	25.52	20.34	17.73	1.71	-0.01	0.76	2	1.00	Directed	Unipartite	1	Social Offline Multilayer Unweighted Metadata
academia_edu	Academica.edu (2011)	200,169	1,398,063	6.98	46.24	109.99	78.34	-0.02	0.04	16	1.00	Directed	Unipartite	1	Social Online Unweighted
add_health	Adolescent health (ADD HEALTH) (1994)	2,587	12,969	5.01	5.65	11.92	29.03	0.29	0.17	10	0.98	Directed	Unipartite	84	Social Offline Weighted
adjnoun	Word adjacencies of David Copperfield	112	425	7.59	6.85	11.54	2.27	-0.13	0.16	5	1.00	Undirected	Unipartite	1	Informational Language Unweighted
advogato	Advogato trust network (2009)	6,541	51,127	7.82	34.13	68.61	20.71	-0.05	0.11	9	0.77	Directed	Unipartite	1	Social Online Weighted
amazon_copurchases	Amazon co-purchasing network (2003)	410,236	3,356,824	8.18	16.30	40.36	1805.09	-0.01	0.25	22	1.00	Directed	Unipartite	4	Economic Commerce Unweighted
amazon_ratings	Amazon customer ratings (2010)	3,376,972	5,838,041	3.46	19.33	83.61	610.18	-0.02	0.00	28	0.86	Undirected	Bipartite	1	Economic Preferences Weighted Timestamps
ambassador	Philippines Ambassador bombing (2000)	16	19	2.38	2.23	3.23	3.96	-0.21	0.59	4	0.69	Undirected	Unipartite	15	Social Offline Weighted Temporal
anybeat	Anybeat social network (2013)	12,645	67,053	5.30	89.97	92.23	41.02	-0.12	0.02	10	1.00	Directed	Unipartite	1	Social Online Unweighted
arxiv_authors	Arxiv authors (1993-2003)	133,280	396,160	5.94	27.24	92.56	158.60	0.21	0.32	14	0.13	Undirected	Unipartite	5	Social Collaboration Unweighted Projection
arxiv_citation	arXiv citation networks (1993-2003)	34,546	421,578	12.20	30.90	74.33	63.13	-0.01	0.15	14	1.00	Directed	Unipartite	2	Informational Citation Unweighted

Tools

- network visualization
  - Gephi
  - D3
  - igraph
- Python libraries
  - NetworkX
  - Graph-tool
  - SNAP

Approaches

- **Physics of complex systems**
  - microscopic modeling
  - statistical physics tools (mean-field)
  - universal features
- **Computer science**
  - machine learning
  - link prediction
  - classification
  - clustering

Ranking

Google

network science

Tutti | Immagini | Notizie | Libri | Video | Altro | Impostazioni | Strumenti

Circa 2.790.000.000 risultati (0,53 secondi)

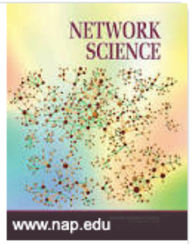
Suggerimento: Cerca risultati solo in **italiano**. Puoi specificare la lingua di ricerca in Preferenze.

Network Science | SpringerOpen Journal | SpringerOpen.com
Annuncio appliednetsci.springeropen.com/
Applied **Network Science** – open-access journal for researchers and practitioners

About the journal
Discover the advantages of publishing with SpringerOpen

Editorial board
Read the profiles of our expert international editorial board

Network science is an academic field which studies complex **networks** such as telecommunication **networks**, computer **networks**, biological **networks**, cognitive and semantic **networks**, and social **networks**, considering distinct elements or actors represented by nodes (or vertices) and the connections between the elements or ...



Network science - Wikipedia
https://en.wikipedia.org/wiki/Network_science

Informazioni su questo risultato | Feedback

Network science - Wikipedia
https://en.wikipedia.org/wiki/Network_science ▾ Traduci questa pagina
Network science is an academic field which studies complex **networks** such as telecommunication **networks**, computer **networks**, biological **networks**, cognitive and semantic **networks**, and social **networks**, considering distinct elements or actors represented by nodes (or vertices) and the connections between the elements or ...

Network science

Campo di studi

Tradotto dall'inglese - La scienza di rete è un campo accademico che studia reti complesse come reti di telecomunicazione, reti di computer, reti biologiche, reti cognitive e semantiche e reti sociali, considerando i diversi elementi o attori rappresentati dai nodi e le connessioni tra gli elementi o gli attori come collegamenti. Wikipedia (inglese)

Vedi la descrizione originale

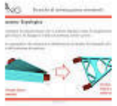
Ricerche correlate

Visualizza altri 10 elementi


Algoritmo


Apprendi… automatico

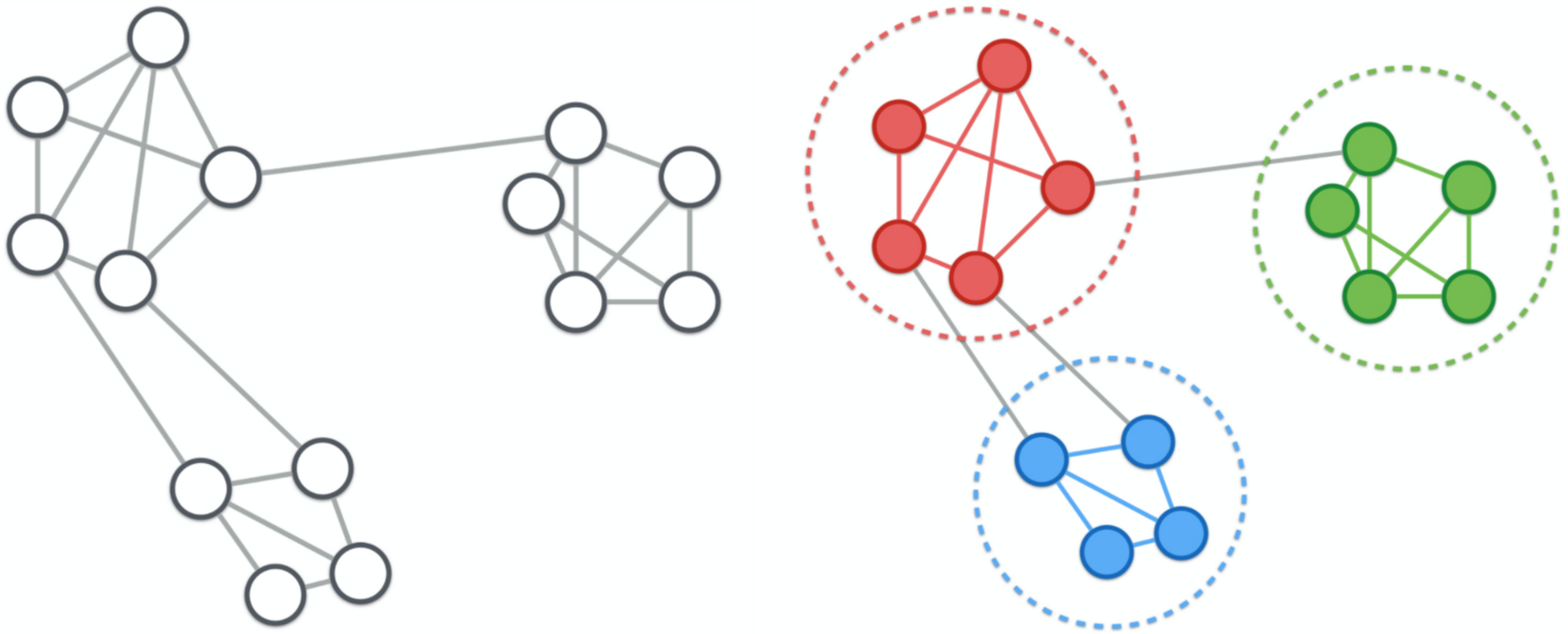

Rete di computer


Ottimizza…


Ricerca scientifica

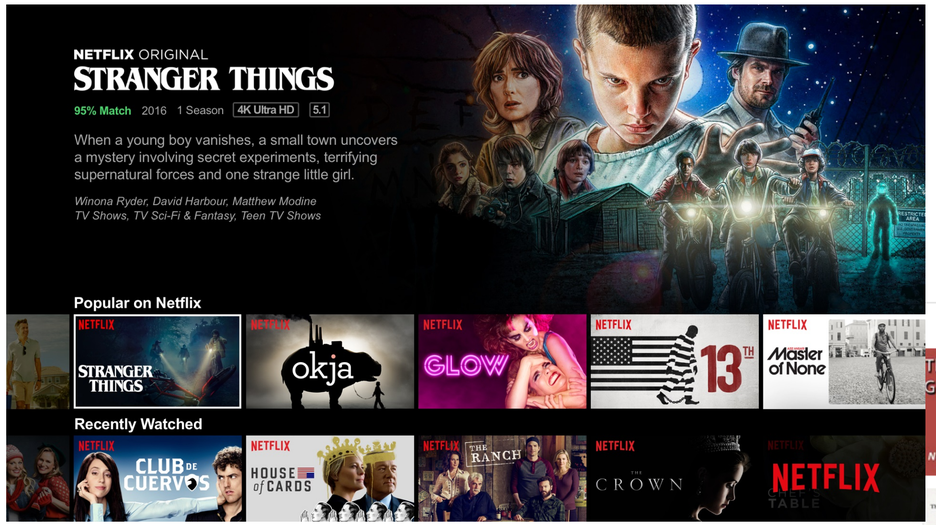
Feedback

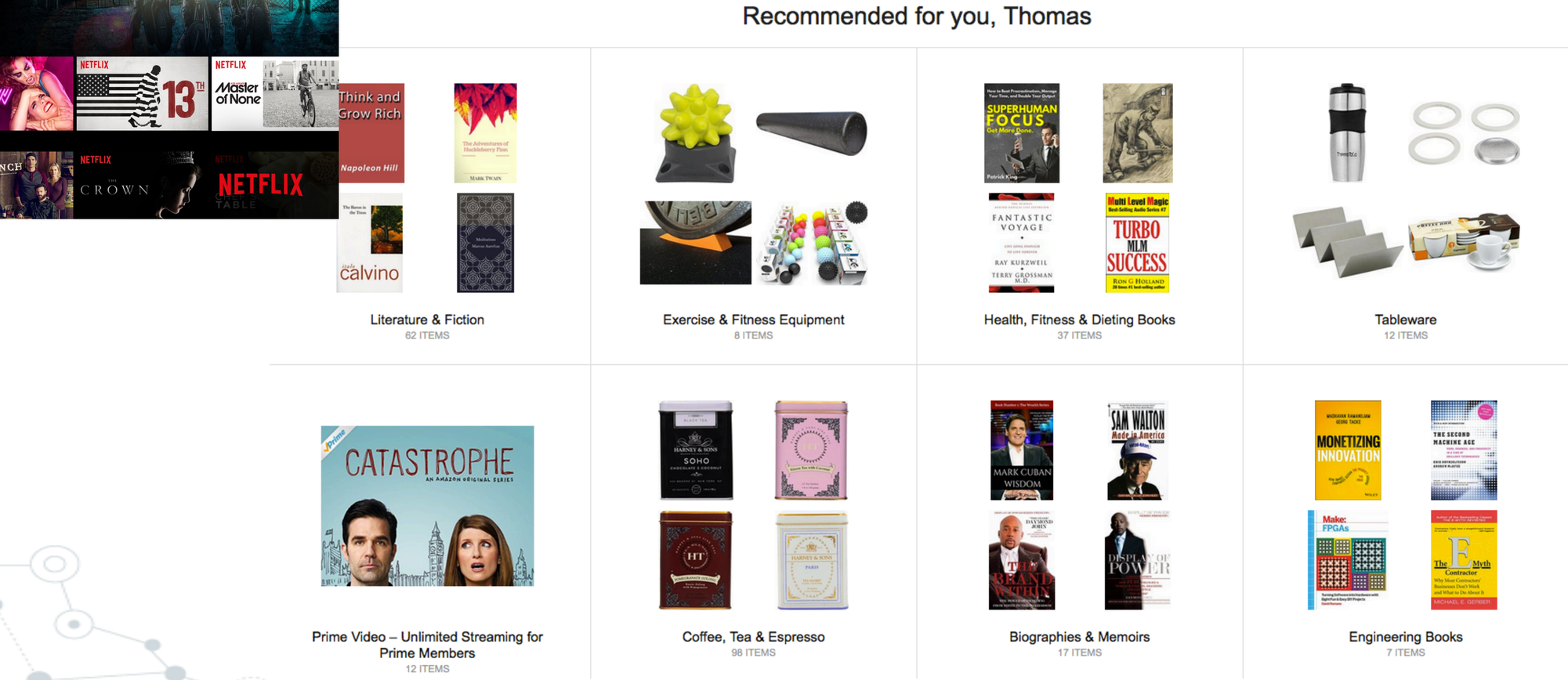
Community detection



application: identify similar customers based on their purchases

Recommendation





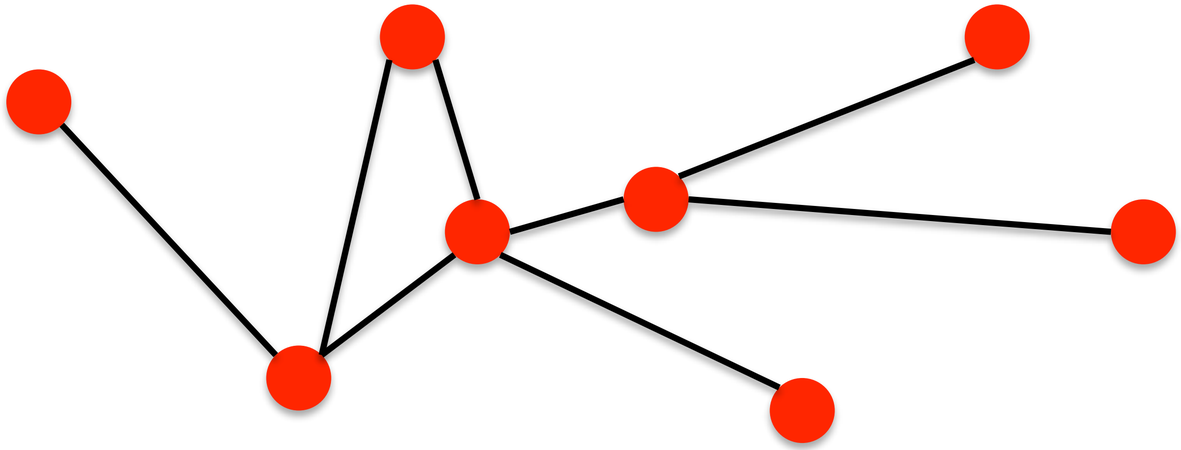
References

- M. E. J. Newman, “Networks: an introduction” Oxford University Press

- Albert-László Barabási, “Network Science” http://networksciencebook.com/

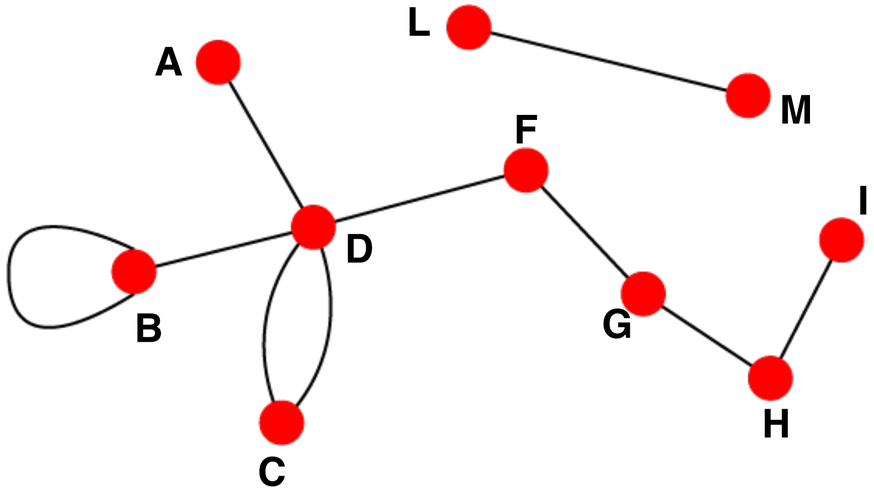
Some formalism

Components

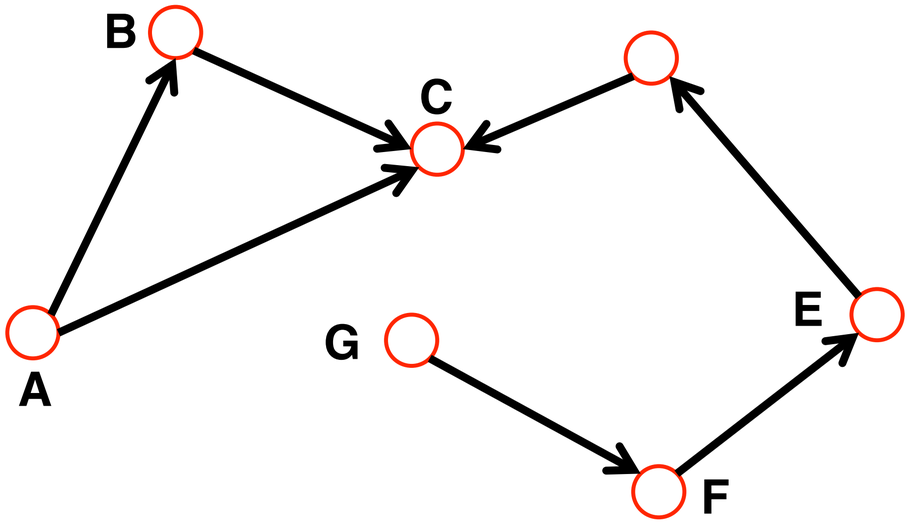


- **components**: nodes, vertices N
- **interactions**: links, edges L
- **system**: network, graph (N,L)

Undirected vs directed



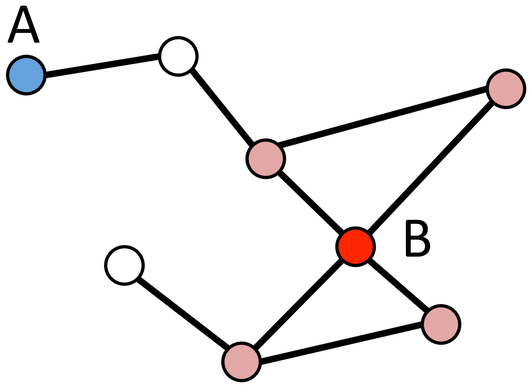
co-authorship
actor networks
co-occurrence



phone calls
hyperlinks
scientific citations

Degree and degree distribution

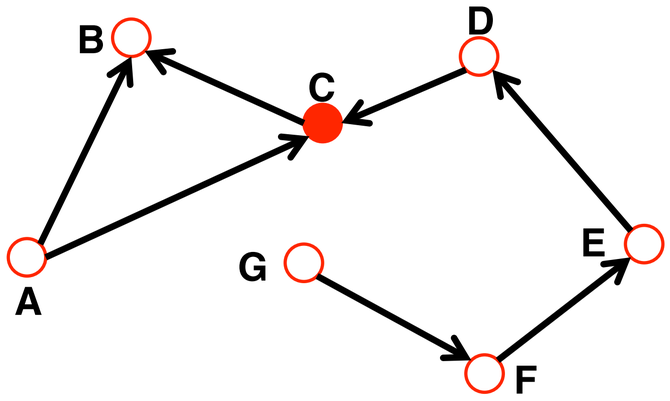
Undirected



$k_A = 1 \qquad k_B = 4$

***node degree**: the number of links connected to the node*

Directed



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

***Source:** degree in = 0*

***Sink:** degree out = 0*

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:

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

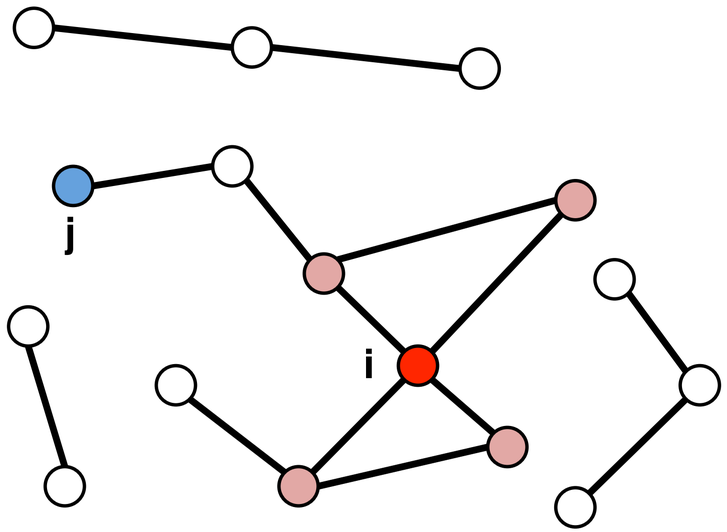
Distribution of x:

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

where p_x follows

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

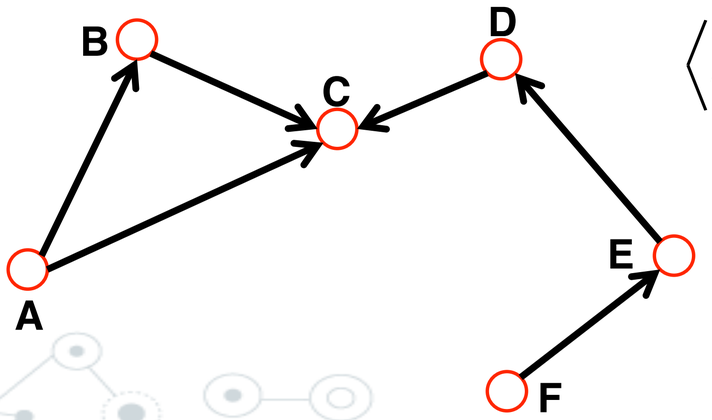
Undirected



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

N – the number of nodes in the graph

Directed



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

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

Degree distribution

$$P(k) = \frac{N_k}{N}$$

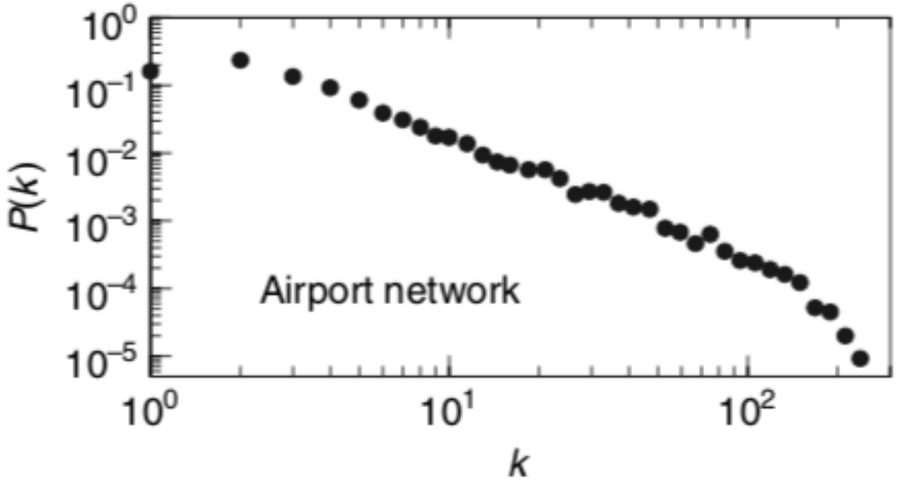
probability that a random chosen node has degree k

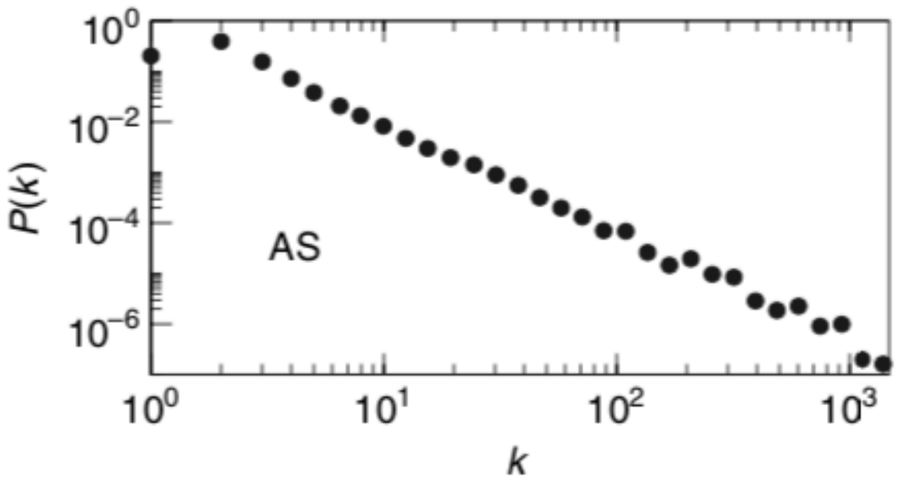
$$\langle k \rangle = \sum_k kP(k)$$

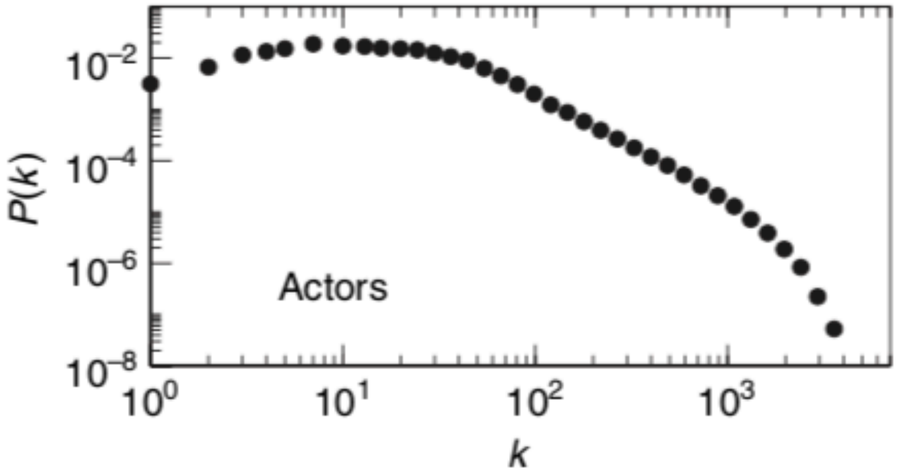
$$\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$$

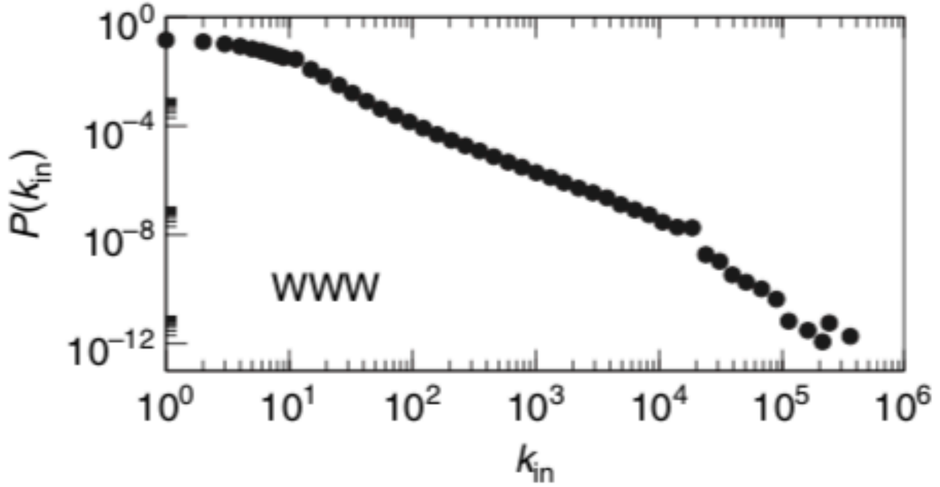
$$\langle k^2 \rangle = \sum_k k^2 P(k)$$

Real world networks

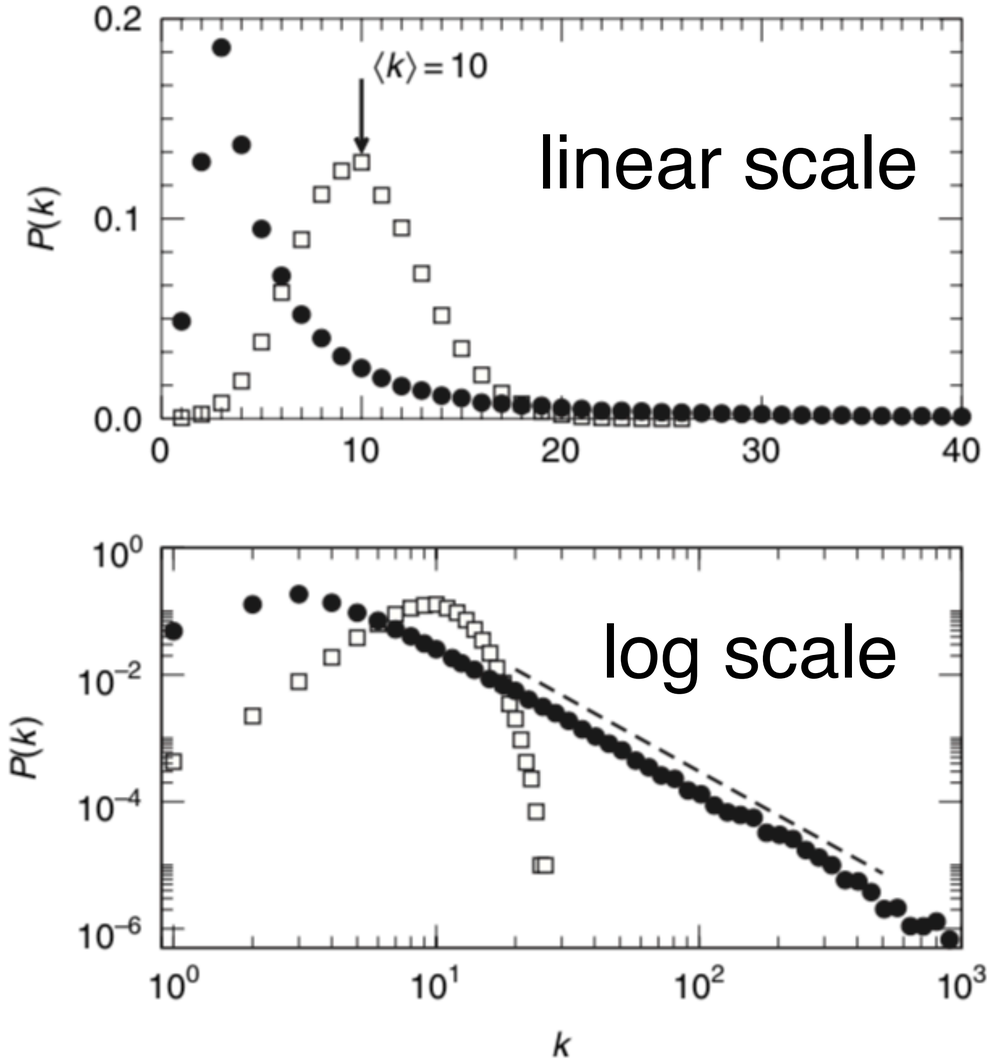








Real world networks



Poisson (homogeneous)

vs

power-law (heterogeneous)

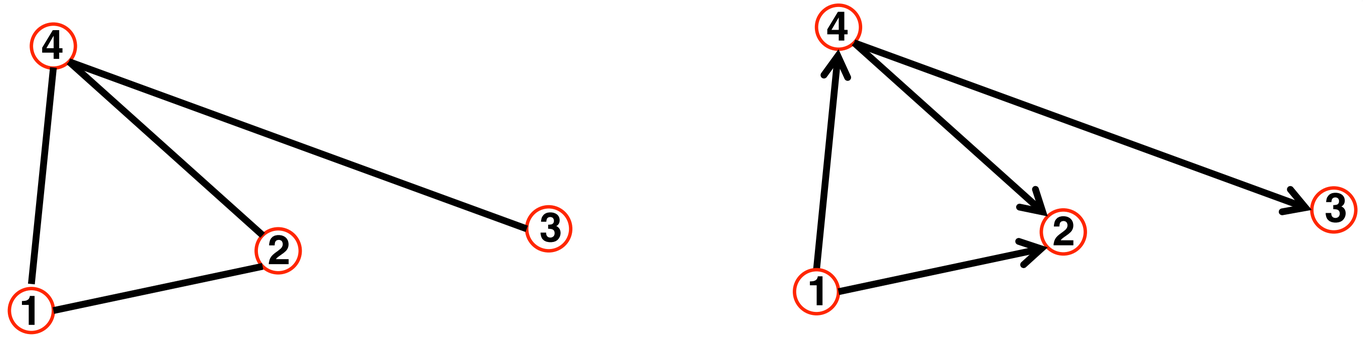
Broad degree distributions

Power-law tails

$$P(k) \sim k^{-\gamma}, 2 < \gamma < 3$$

No characteristic scale

Adjacency matrix



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

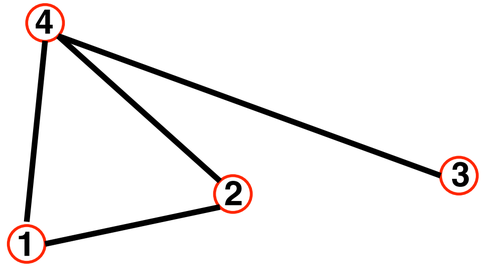
A_{ij}=0 if nodes *i* and *j* are not connected to each other.

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

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

Adjacency matrix

Undirected



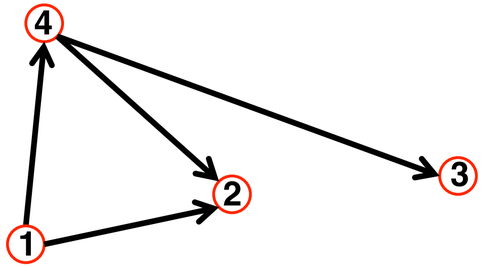
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

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

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

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

Directed



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

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

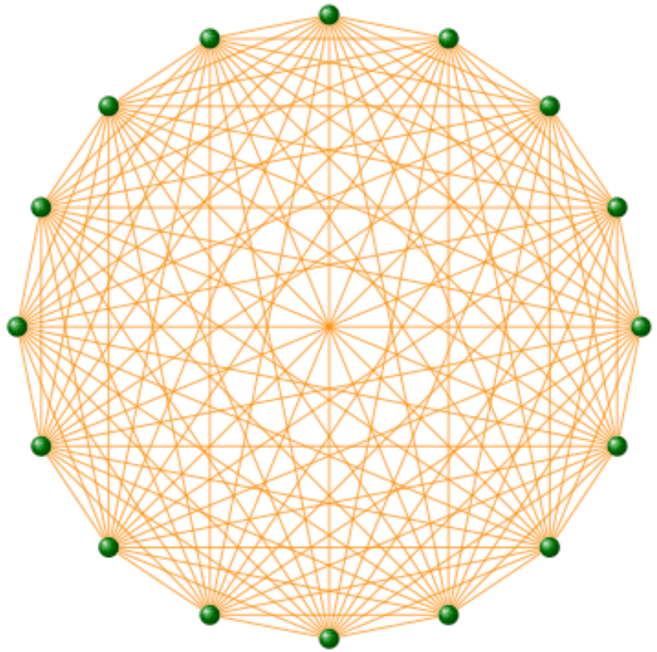
$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

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

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

Real networks are sparse!

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**

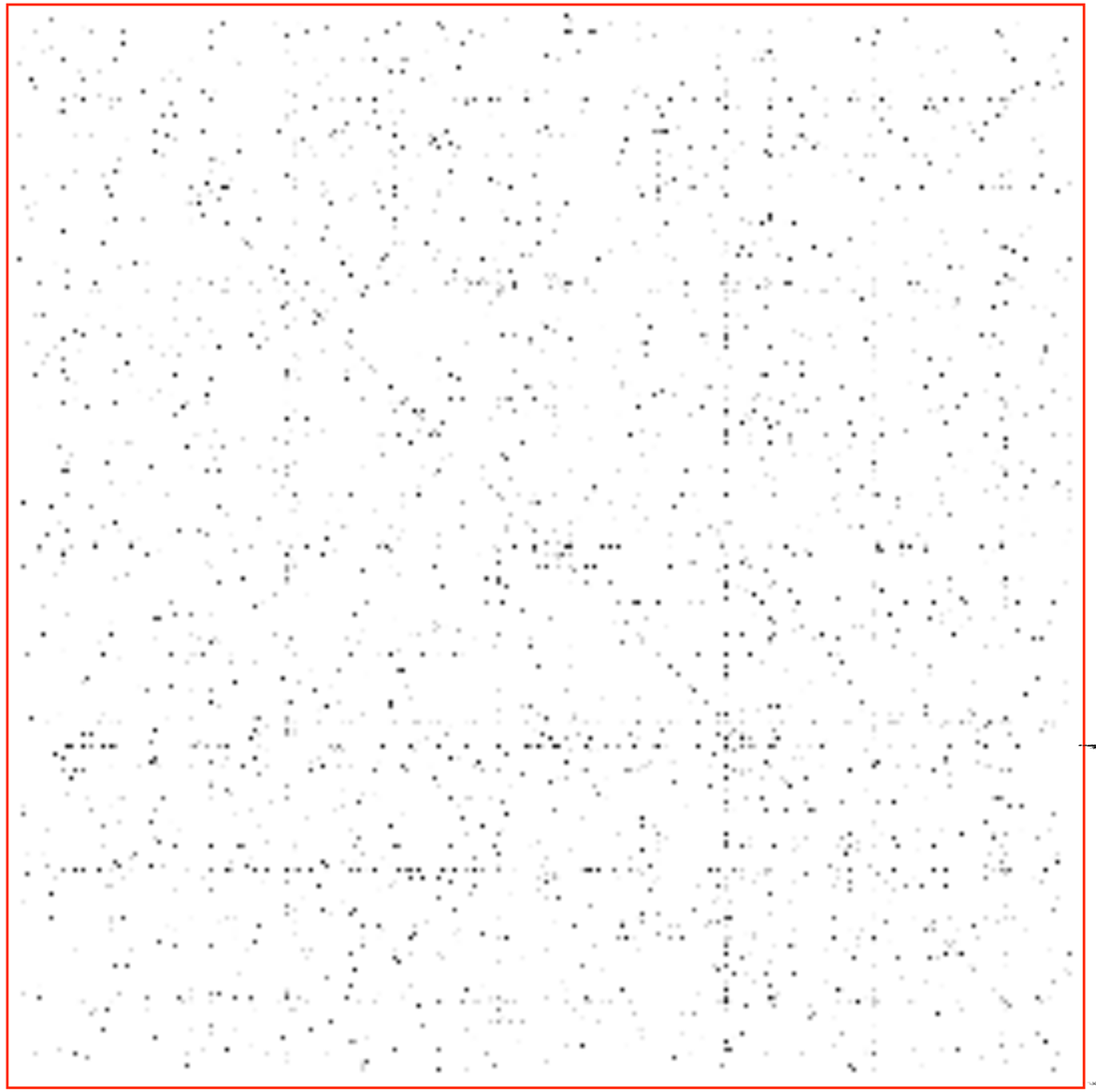
Most networks observed in real systems are **sparse**

$$L \ll L_{max} \qquad \langle k \rangle \ll N - 1$$

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

(Source: Albert, Barabasi, RMP2002)

The adjacency matrix of the yeast protein-protein interaction network, consisting of 2,018 nodes, each representing a yeast protein.

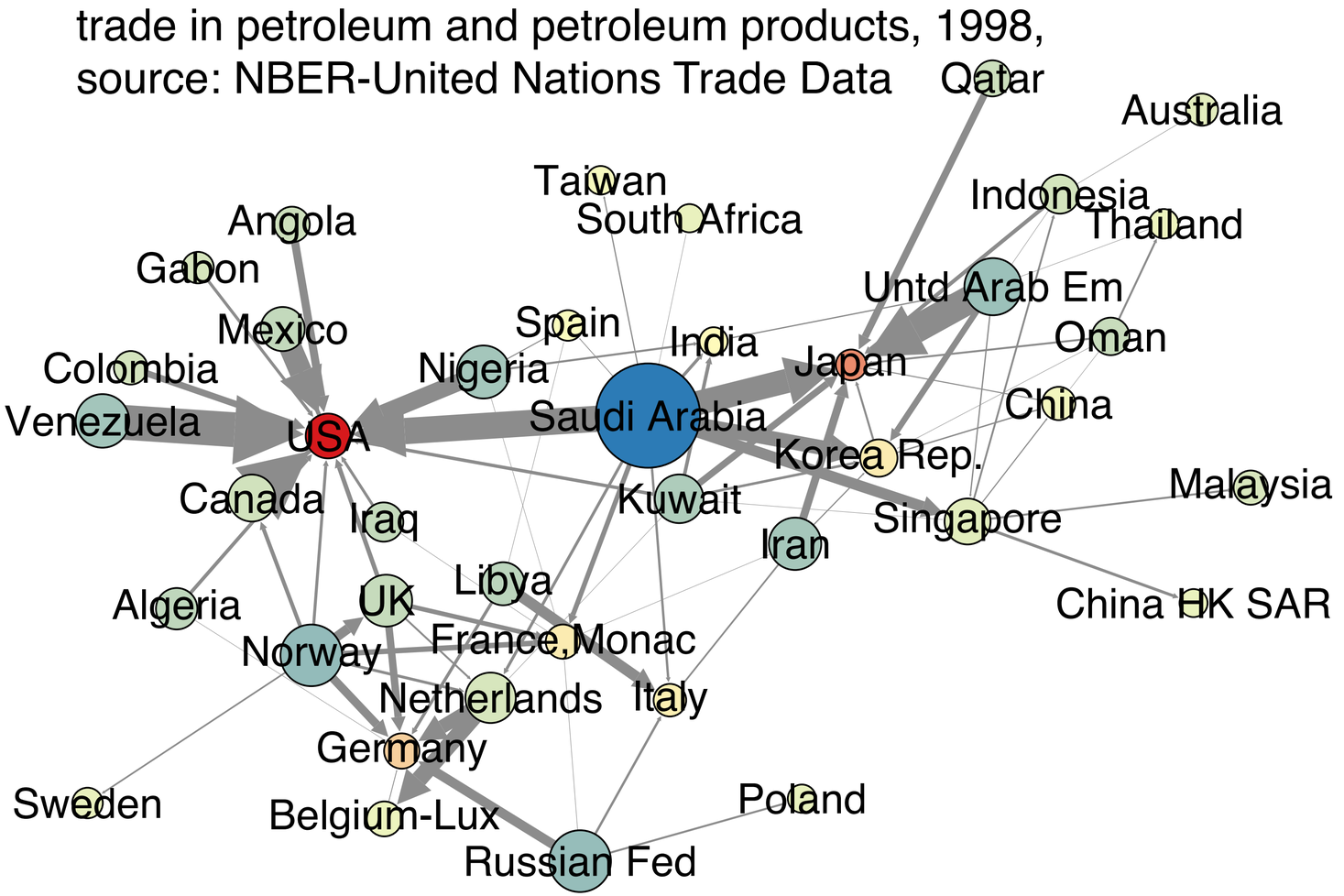


The adjacency matrix is not efficient to store the network

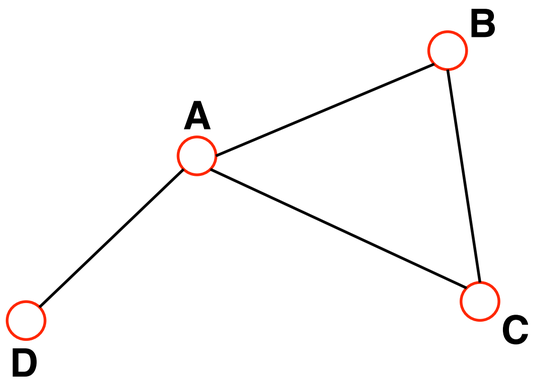
Weighted networks

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

$$s_i = \sum_j w_{ij}$$

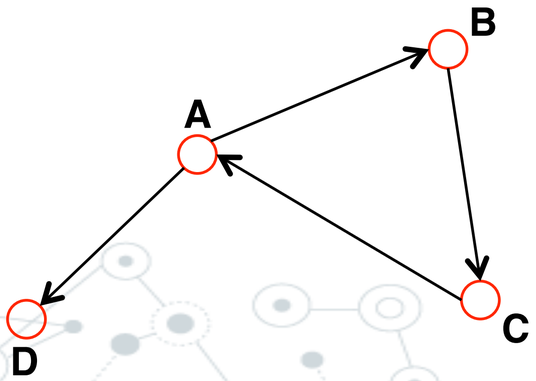


Distance



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them. The *diameter of a graph* is the length of the longest geodesic path between any pair of vertices in the network for which a path actually exists.

*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).

Paths

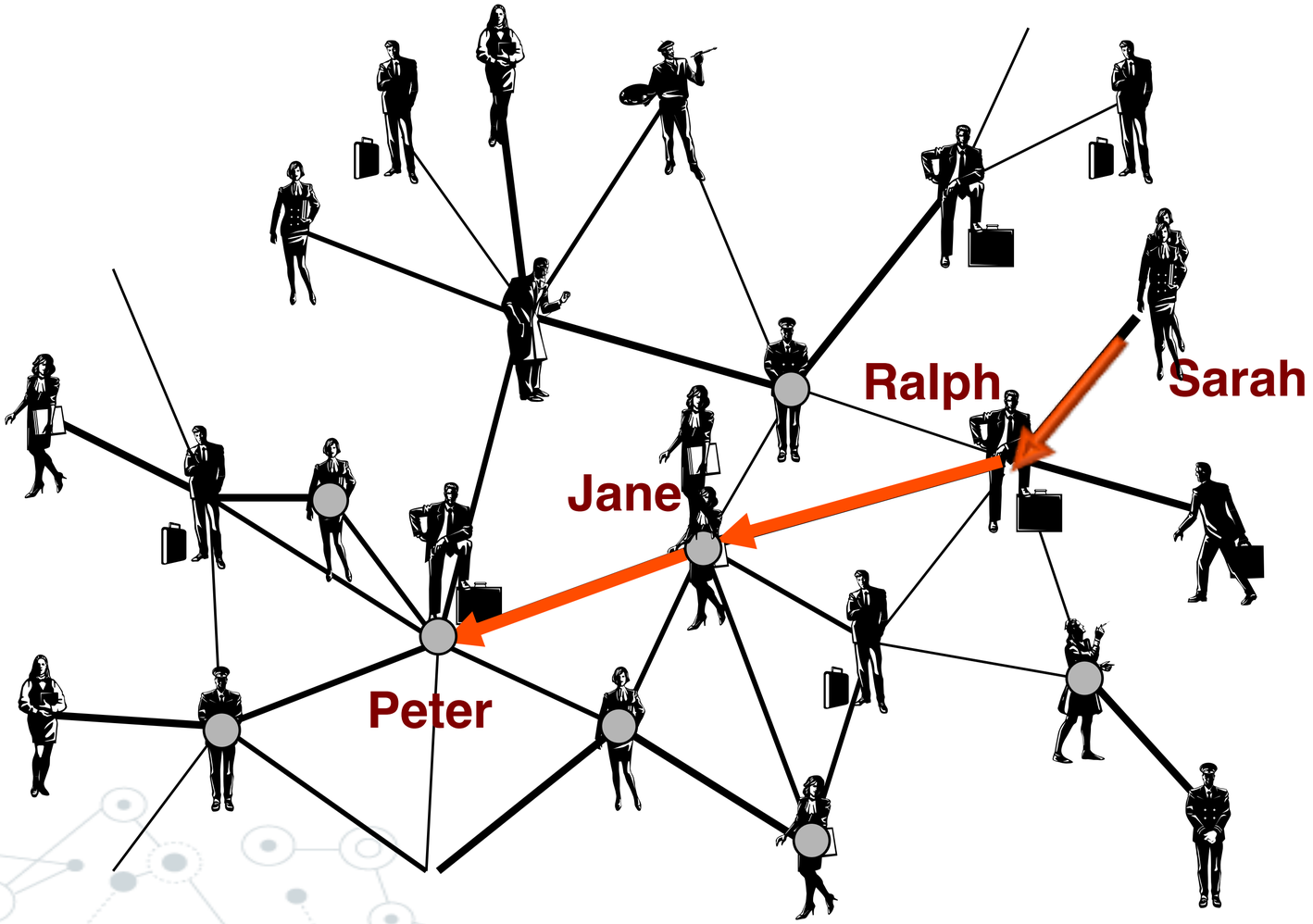
Diameter: $\boldsymbol{d_{max}}$ the maximum distance between any pair of nodes in the graph.

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

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

In an *undirected graph* $d_{ij} = d_{ji}$, *so* we only need to count them once: $\langle d \rangle = \frac{1}{L_{max}} \sum_{i,j>i} d_{ij}$

Real world networks



Small world effect

Frigyes Karinthy, 1929
Stanley Milgram, 1967

Six degrees of separation

Stanley Milgram (1967)

Two targets in Boston and
Sharon, MA.

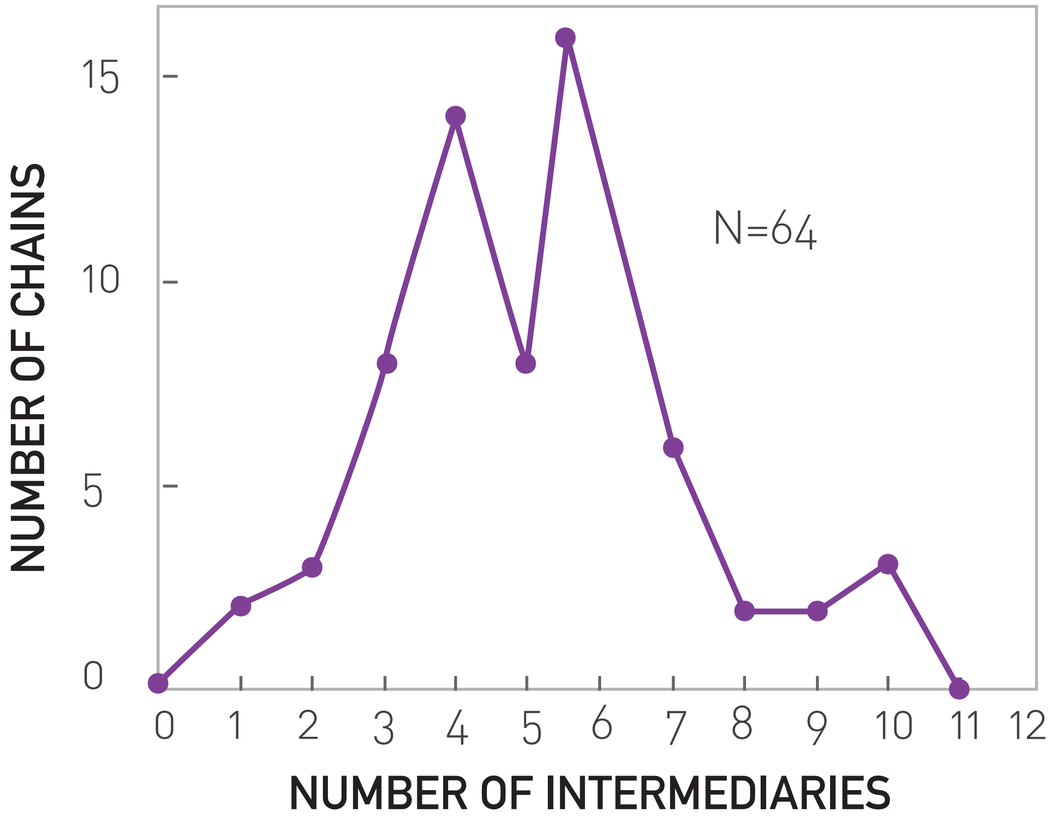
Randomly selected residents of
Wichita and Omaha were asked
to forward a letter to someone
who is most likely to know the
target person.

Six degrees of separation

Stanley Milgram (1967)

Two targets in Boston and Sharon, MA.

Randomly selected residents of Wichita and Omaha were asked to forward a letter to someone who is most likely to know the target person.



Six degrees of separation

Real networks are
small-world!

$$\langle d \rangle \simeq \log(N)$$

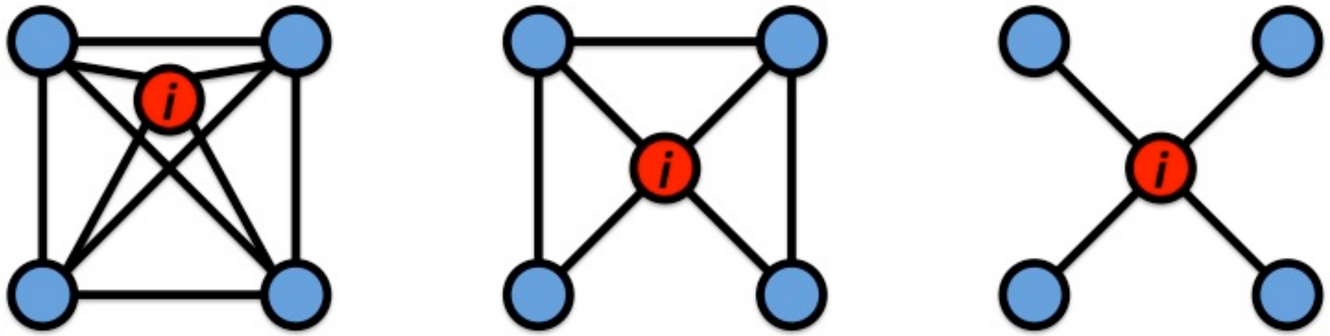
Clustering coefficient

The clustering coefficient of a node captures the degree to which the neighbours of a given node link to each other, i.e. **what fraction of your neighbors are connected?**

Clustering coefficient

The clustering coefficient of a node captures the degree to which the neighbours of a given node link to each other, i.e. **what fraction of your neighbors are connected?**

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

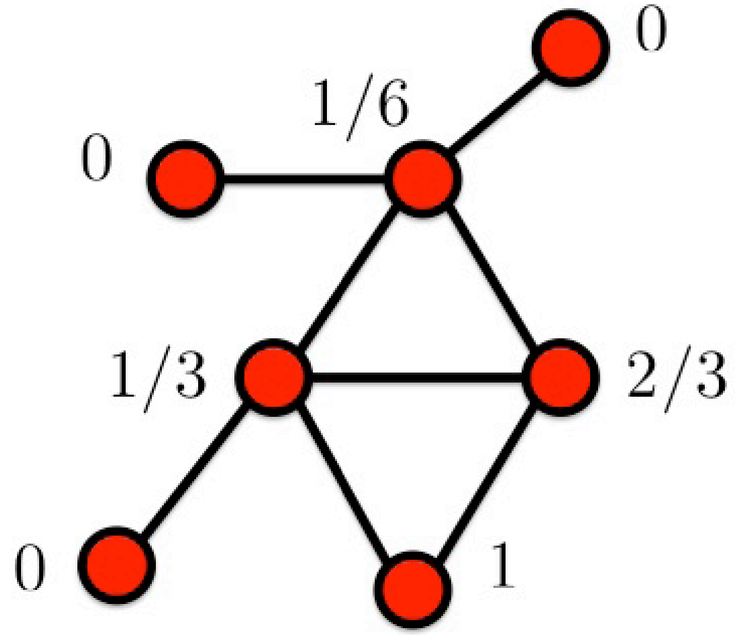


$C_i = 1$ $C_i = 1/2$ $C_i = 0$

Clustering coefficient

The degree of clustering of a whole network is captured by the **average clustering coefficient**, representing the average of C over all nodes i = 1,…,N

$$\langle C_i \rangle = \frac{1}{N} \sum_i C_i$$



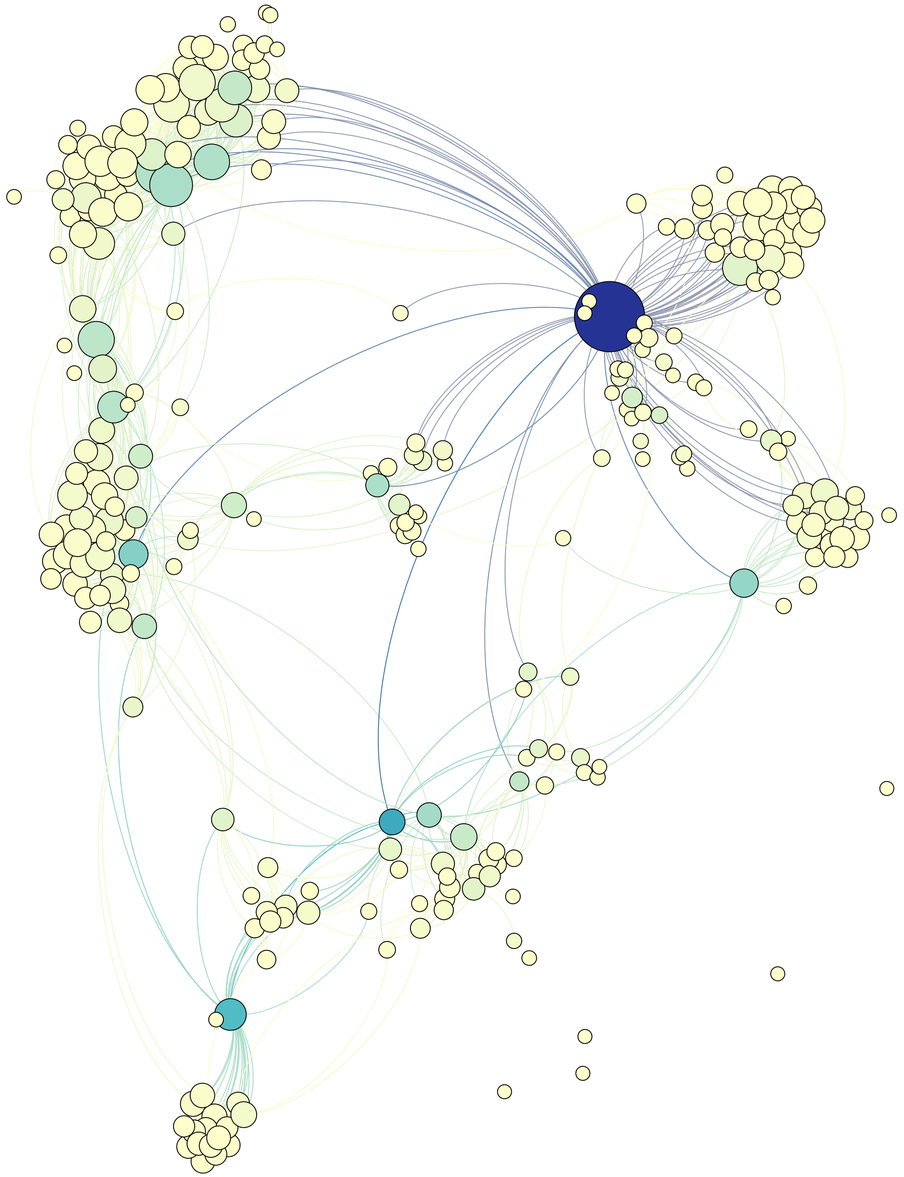
$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

Real world networks

- Real world networks are **highly clustered**

- Average clustering coefficient can have values >0.5

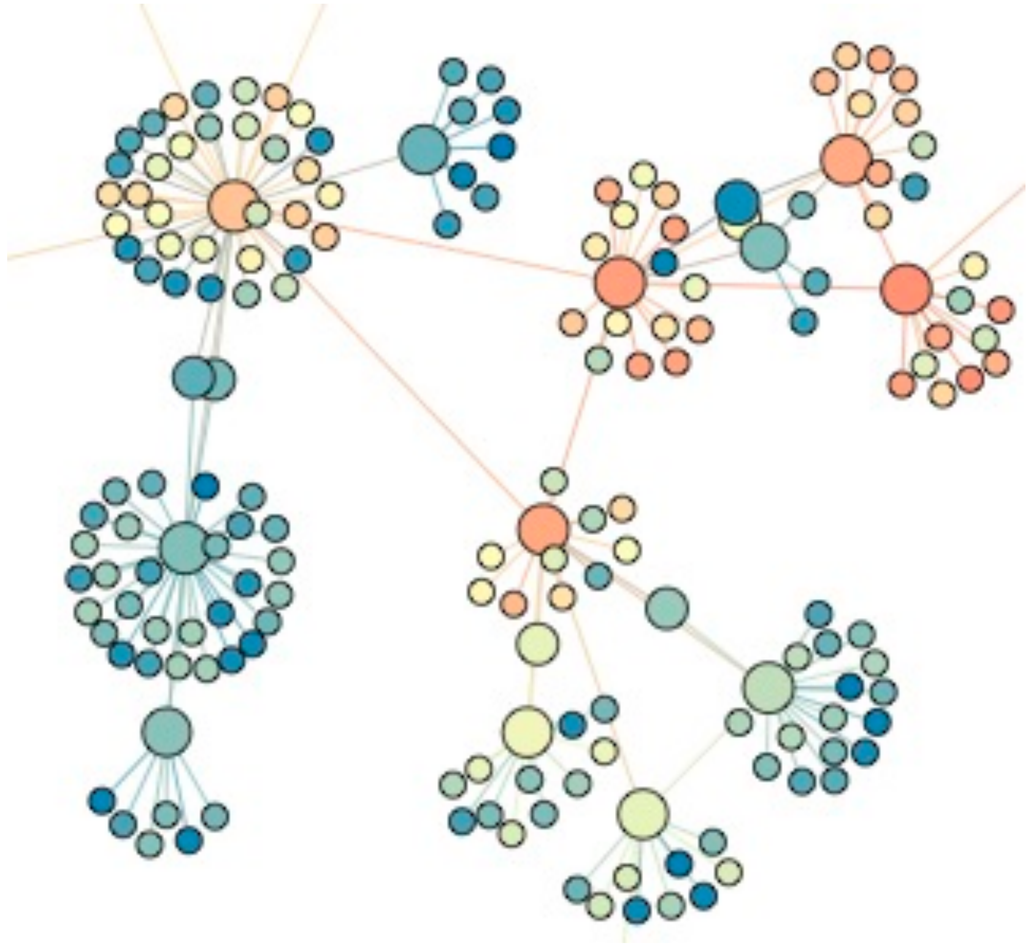
- **Triadic closure** in social networks is a common phenomenon



Centrality measures

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality
- Katz centrality
- Pagerank

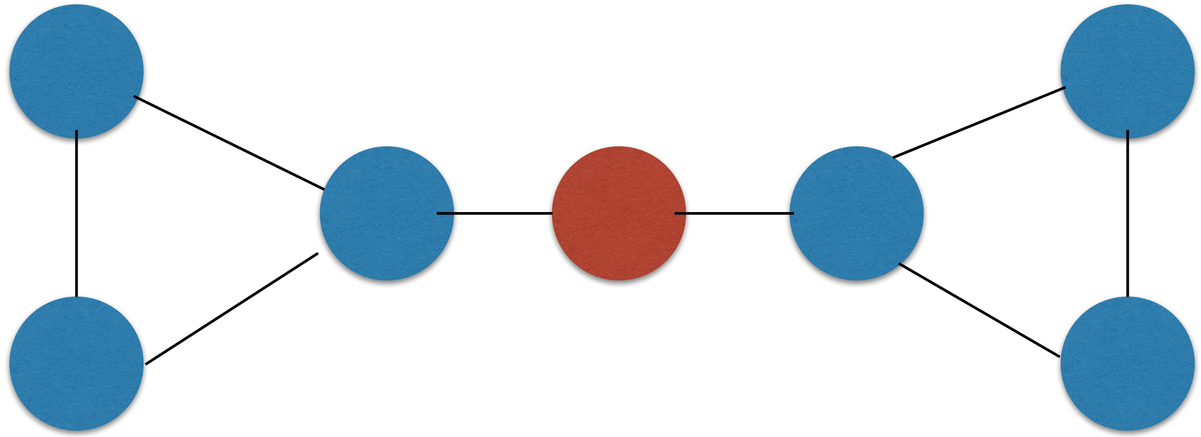
…



Betweenness centrality

Betweeness captures a node's brokerage

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

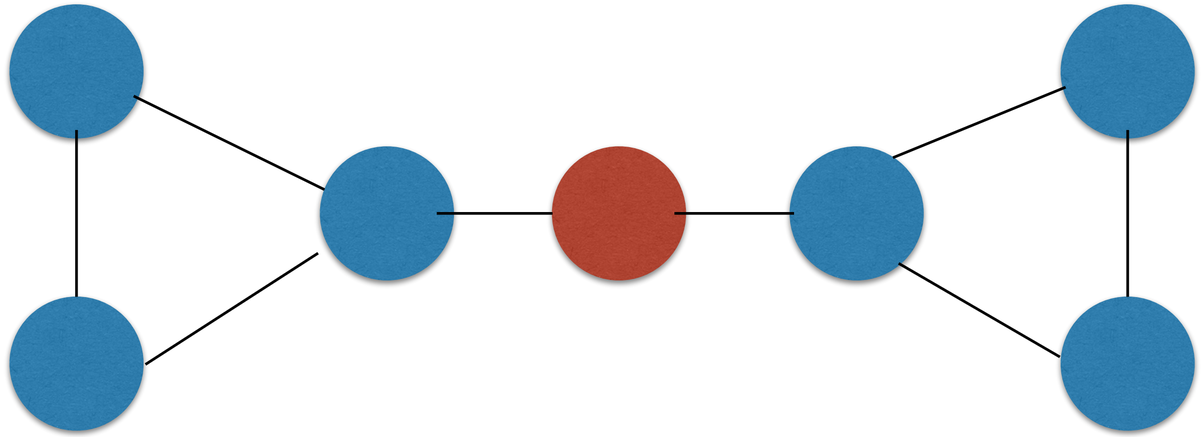


Betweenness centrality

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

g_{jk} = #shortest paths connecting j and k

$g_{jk}(i)$ = #shortest paths connecting j and k through i

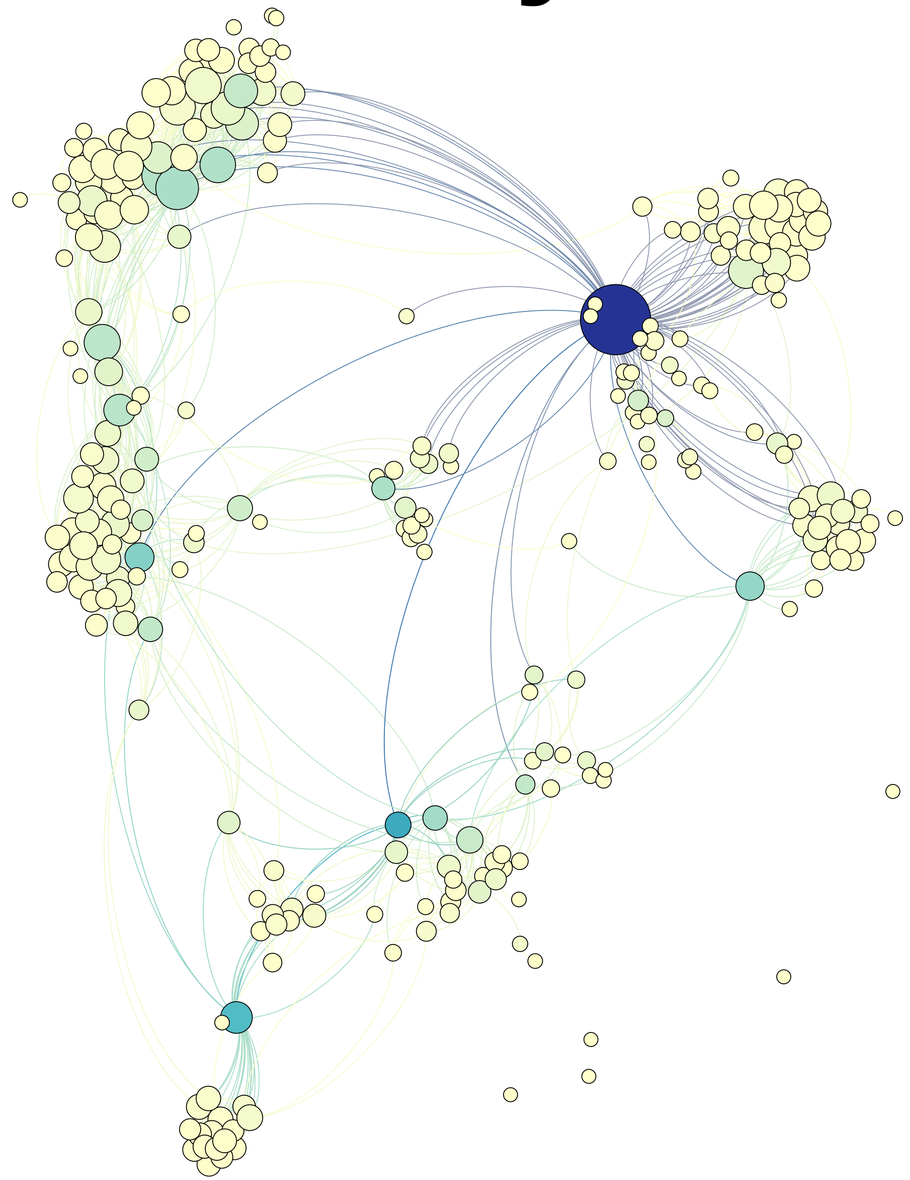


Betweenness centrality

My Facebook graph

Node size is proportional
to the degree

Node color is proportional
to the betweenness



Next.. Network theory: models