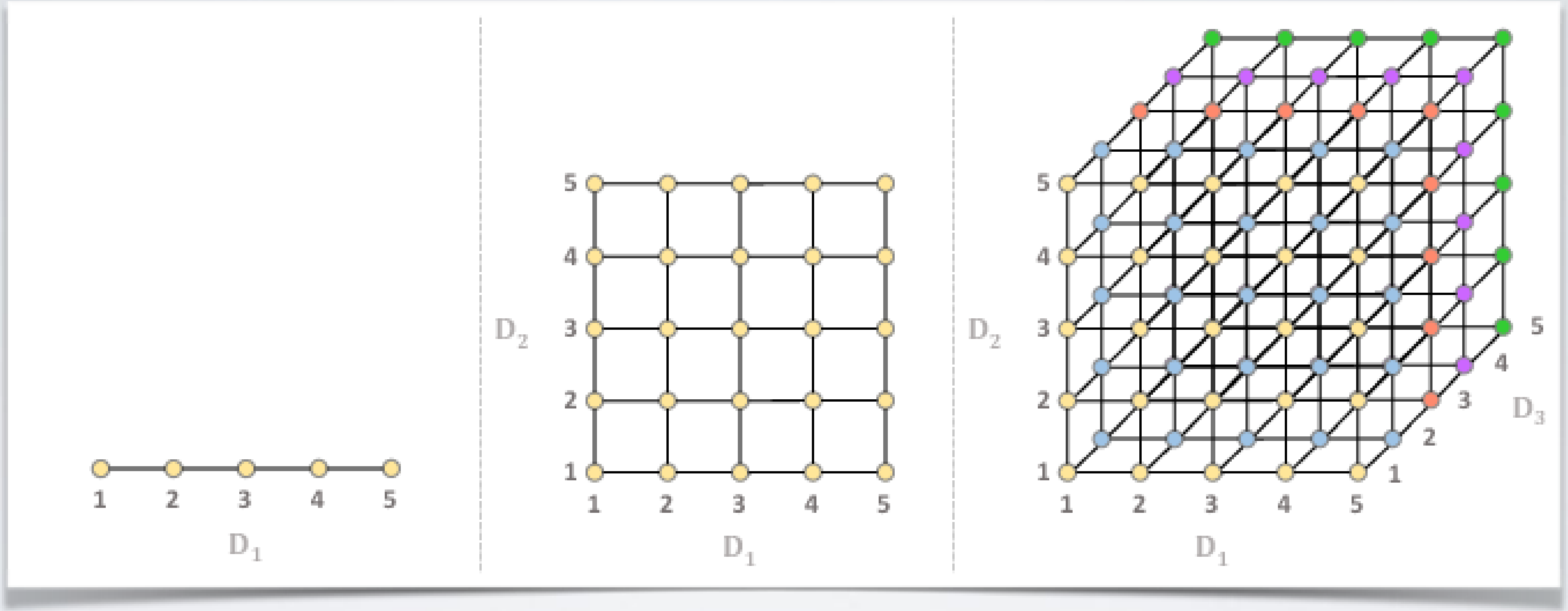
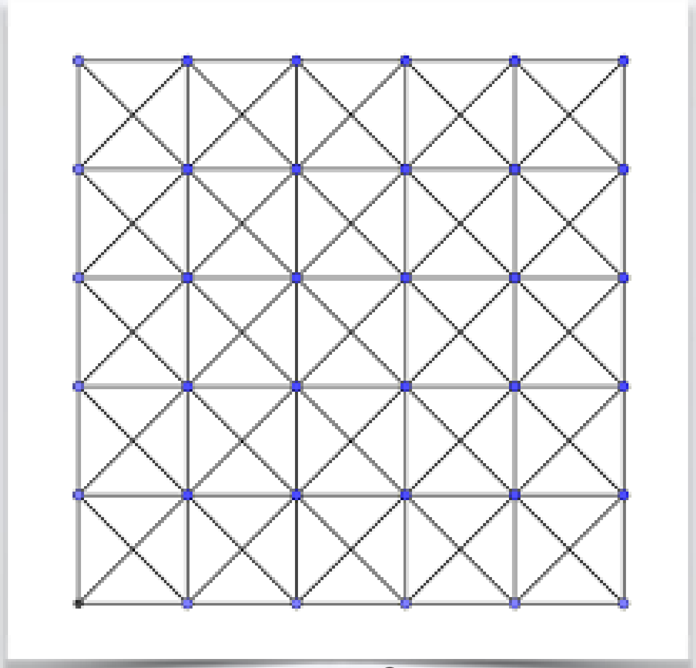
M2 TIW – M2 BIO-INFO

DATA ANALYSIS

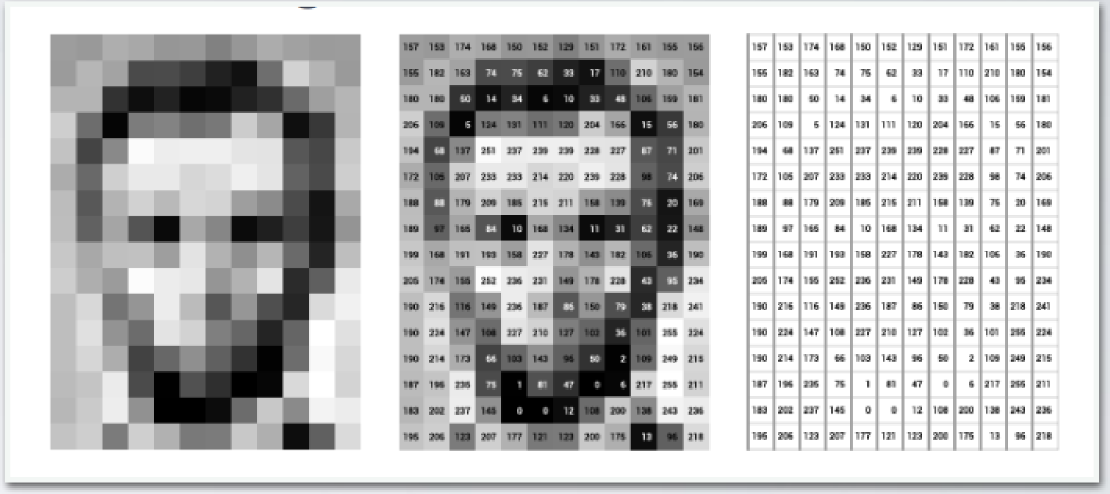
Network Data Mining

NETWORKS/GRAPHS



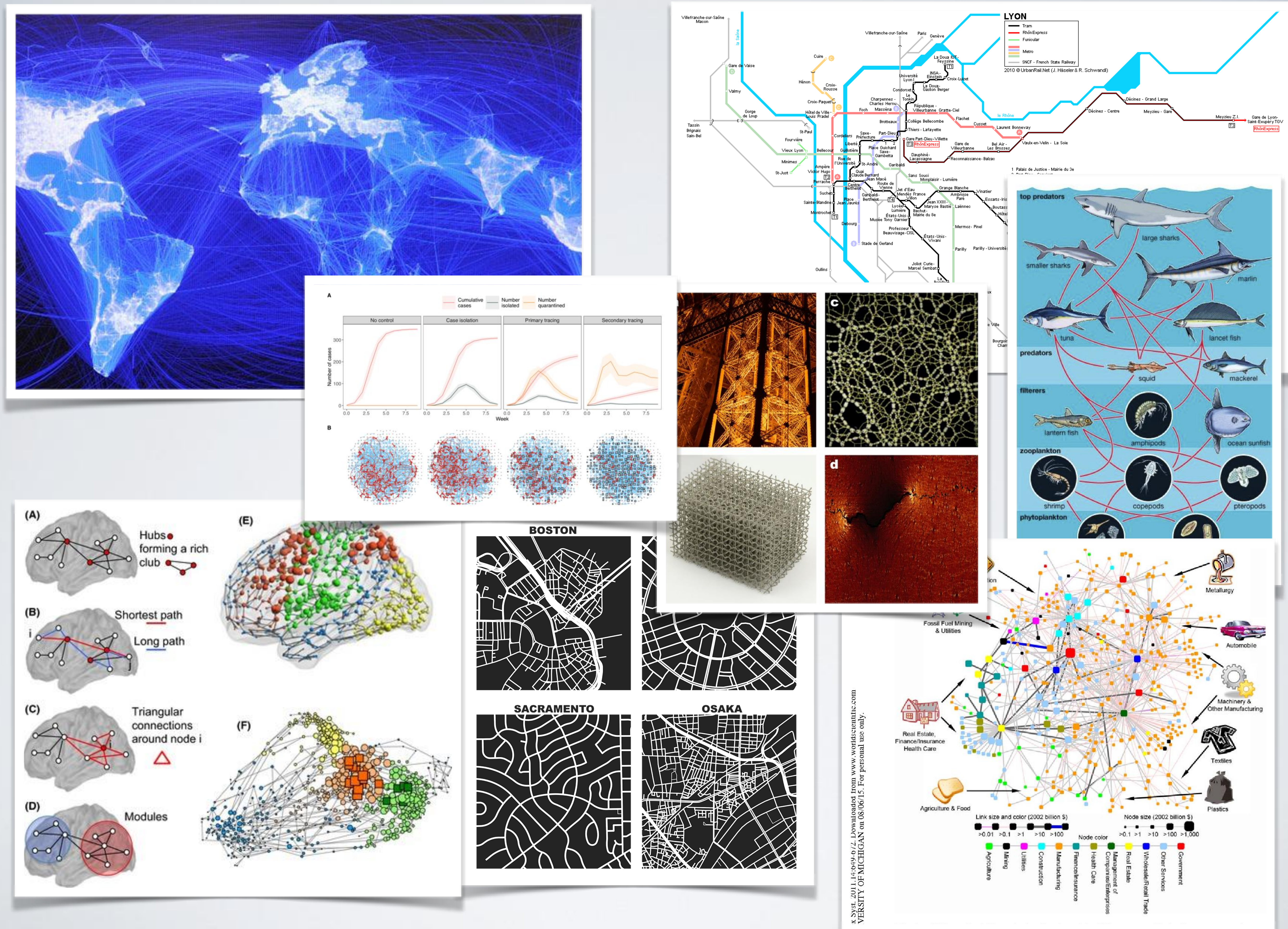


NETWORKS/GRAPHS

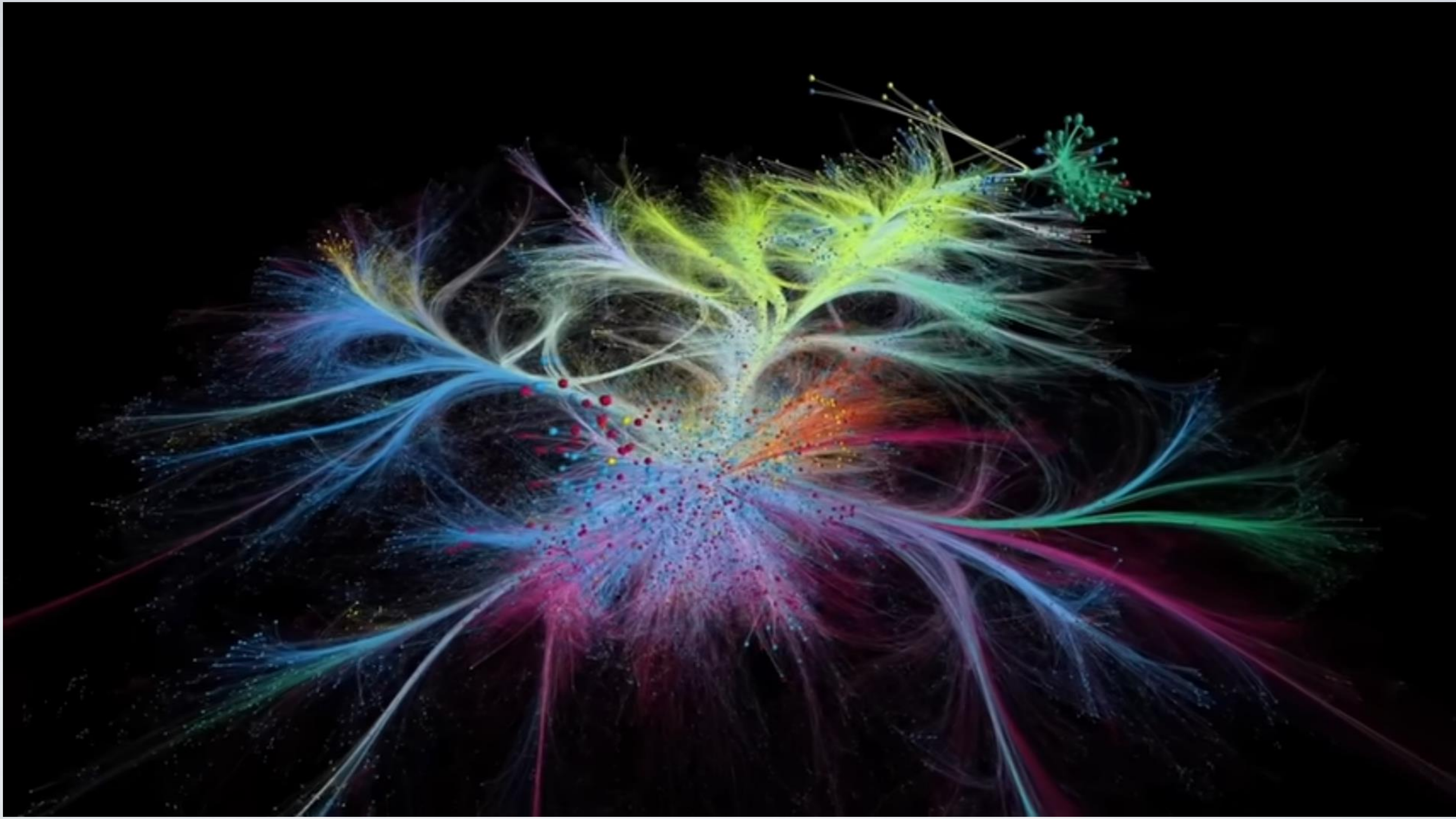


- Structured data
  - Text
    - Sequence. Each item is **before** or **after** the other ones. And it is important
    - 1D organisation
  - Images
    - Each pixel has a position in 2D grid, it is on the **left**, **right**, **top** or **bottom** compared with the other ones. And it is important
    - 2D organisation
  - Variants: Video (3D), time series (1D continuous), spatial (2D/3D continuous), etc.
  - **Networks**: Neighborhoods are not constrained. The graph is the structure
    - Generalization of discrete structures (text, images, videos)

NETWORKS ARE EVERYWHERE


LYON
Tram
RhônExpress
Funicular
Metro
SNCF - French State Railway
2010 © UrbanRail.Net (J. Häseler & R. Schwandl)
A
Cumulative cases
Number isolated
Number quarantined
No control
Case isolation
Primary tracing
Secondary tracing
Number of cases
Week
B
c
d
top predators
large sharks
smaller sharks
marlin
tuna
lancet fish
predators
squid
mackerel
filterers
lantern fish
amphipods
ocean sunfish
zooplankton
shrimp
copepods
pteropods
phytoplankton
(A)
Hubs forming a rich club
(B)
Shortest path
Long path
(C)
Triangular connections around node i
(D)
Modules
(E)
(F)
BOSTON
SACRAMENTO
OSAKA
Metallurgy
Fossil Fuel Mining & Utilities
Automobile
Machinery & Other Manufacturing
Real Estate, Finance/Insurance Health Care
Textiles
Agriculture & Food
Plastics
Link size and color (2002 billion $)
Node size (2002 billion $)
Node color
Agriculture
Mining
Utilities
Construction
Manufacturing
Finance/Insurance
Health Care
Management of Companies/Enterprises
Real Estate
Wholesale/Retail Trade
Other Services
Government

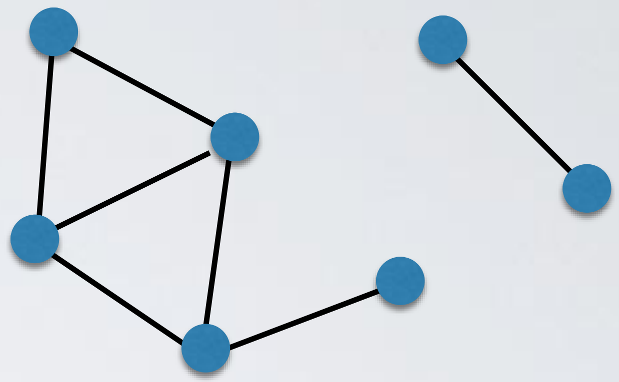
150 YEARS OF PUBLICATIONS



Sources:
[150 years of Nature: a data graphic charts our evolution]
[Nature 150 Interactive]
[Ce réseau décrypte 150 ans de découvertes scientifiques]

GRAPHS & NETWORKS

Networks often refers to real systems
- www,
- social network
- metabolic network.
- Language: (Network, node, link)



Vertex	Edge
person	friendship
neuron	synapse
Website	hyperlink
company	ownership
gene	regulation

Graph is the mathematical representation of a network
- Language: (Graph, vertex, edge)

In most cases we will use the two terms interchangeably.

NETWORK REPRESENTATIONS

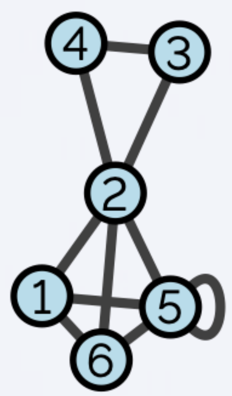
Networks: Graph notation

Graph notation : $G = (V, E)$

V	set of vertices/nodes.
E	set of edges/links.
$u \in V$	a node.
$(u, v) \in E$	an edge.

Network - Graph notation

Graph



Graph notation

$$G = (V, E)$$
$$V = \{1, 2, 3, 4, 5, 6\}$$
$$E = \{(1, 2), (1, 6), (1, 5), (2, 4), (2, 3), (2, 5), (2, 6), (6, 5), (5, 5), (4, 3)\}$$

GRAPH REPRESENTATION

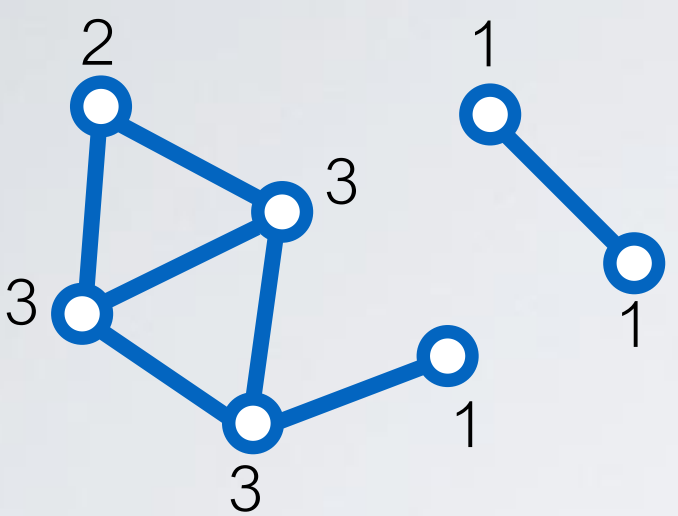
Node-Edge description

N_u	**Neighbourhood** of u, nodes sharing a link with u.
k_u	**Degree** of u, number of neighbors $\|N_u\|$.
N_u^{out}	**Successors** of u, nodes such as $(u, v) \in E$ in a directed graph
N_u^{in}	**Predecessors** of u, nodes such as $(v, u) \in E$ in a directed graph
k_u^{out}	**Out-degree** of u, number of outgoing edges $\|N_u^{out}\|$.
k_u^{in}	**In-degree** of u, number of incoming edges $\|N_u^{in}\|$
$w_{u,v}$	**Weight** of edge (u, v).
s_u	**Strength** of u, sum of weights of adjacent edges, $s_u = \sum_v w_{uv}$.

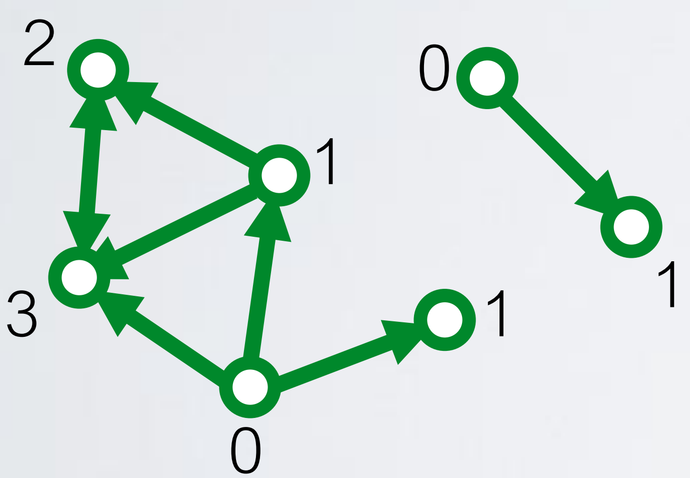
Node degree

Number of connections of a node

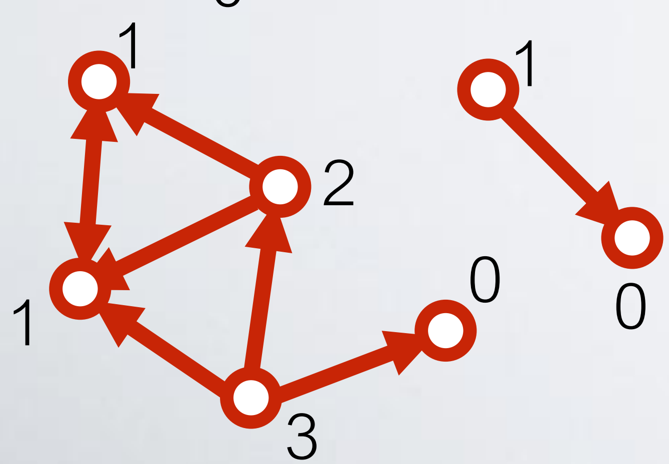
- Undirected network



- Directed network



In degree



Out degree

SIZE

Counting nodes and edges

N/n	**size**: number of nodes $\|V\|$.
L/m	number of edges $\|E\|$
L_{max}	Maximum number of links

Undirected network: $\binom{N}{2} = N(N-1)/2$

Directed network: $\binom{N}{2} = N(N-1)$

DENSITY

Network descriptors 1 - Nodes/Edges

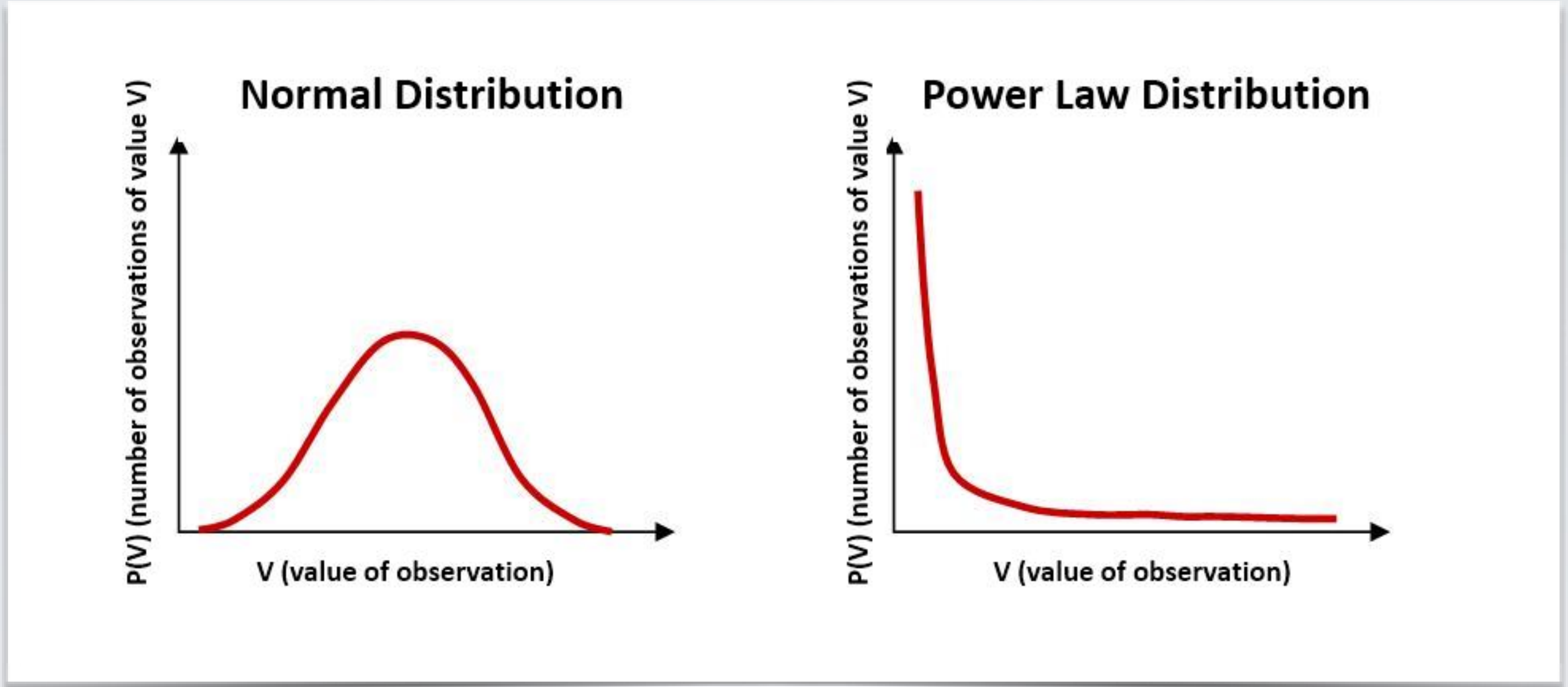
$\langle k \rangle$	**Average degree**: Real networks are sparse, i.e., typically $\langle k \rangle \ll n$. Increases slowly with network size, e.g., $d \sim \log(m)$ $$\langle k \rangle = \frac{2m}{n}$$
$d/d(G)$	**Density**: Fraction of pairs of nodes connected by an edge in G. $$d = L/L_{\max}$$

DENSITY

	#nodes	#edges	Densité	Deg. Moyen
Wikipedia HL	2M	30M	1.5×10^{-5}	30
Twitter 2015	288M	60B	1.4×10^{-6}	416
Facebook 2015	1.4B	400B	4×10^{-9}	570
Brain c. Elegans	280	6393	0,16	46
Roads Calif.	2M	2.7M	6×10^{-7}	2,7
Airport traffic	3k	31k	0,007	21

Attention: It's difficult to compare density of graphs with different sizes

DEGREE DISTRIBUTION



PDF (Probability Distribution Function)

DEGREE DISTRIBUTION

- In a fully random graph (Erdos-Renyi), degree distribution is (close to) a normal distribution centered on the average degree
- In real graphs, in general, it is not the case:
  - A high majority of small degree nodes
  - A small minority of nodes with very high degree (Hubs)
- Often modeled by a **power law**
  - More details later in the course

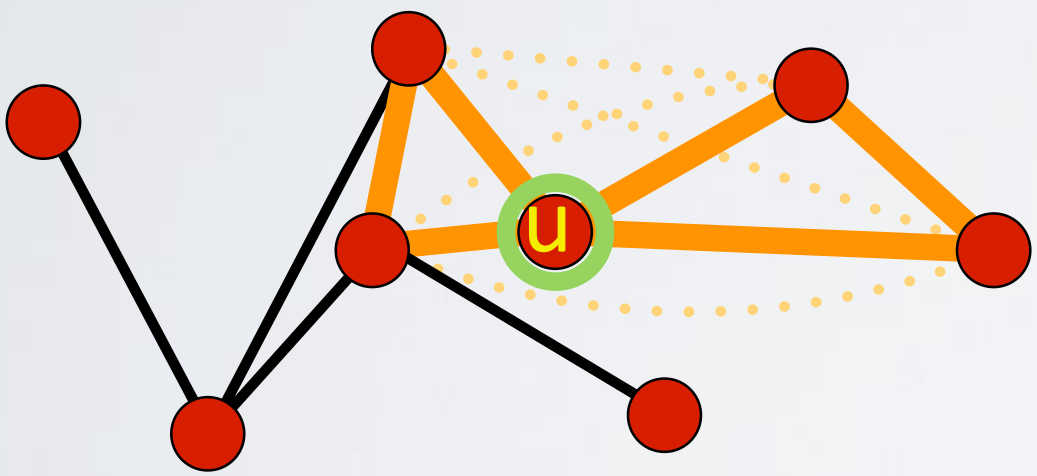
CLUSTERING COEFFICIENT

- **Clustering coefficient** or **triadic closure**

- Triangles are considered important in real networks
  - Think of social networks: ***friends of friends are my friends***
  - # triangles is a big difference between real and random networks

CLUSTERING COEFFICIENT

C_u - **Node clustering coefficient:** density of the subgraph induced by the neighborhood of u, $C_u = d(H(N_u))$. Also interpreted as the fraction of all possible triangles in N_u that exist, $\frac{\delta_u}{\delta_u^{\max}}$



Edges: 2
Max edges: 4*3/2=6
C_u=2/6=1/3

Triangles=2
Possible triangles=$\binom{4}{2}$=6
C_u=2/6=1/3

SUBGRAPHS

Subgraphs

Subgraph $H(W)$ (induced subgraph): subset of nodes W of a graph $G = (V, E)$ and edges connecting them in G, i.e., subgraph $H(W) = (W, E'), W \subset V, (u, v) \in E' \iff u, v \in W \wedge (u, v) \in E$
Clique: subgraph with $d = 1$
Triangle: clique of size 3
Connected component: a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph

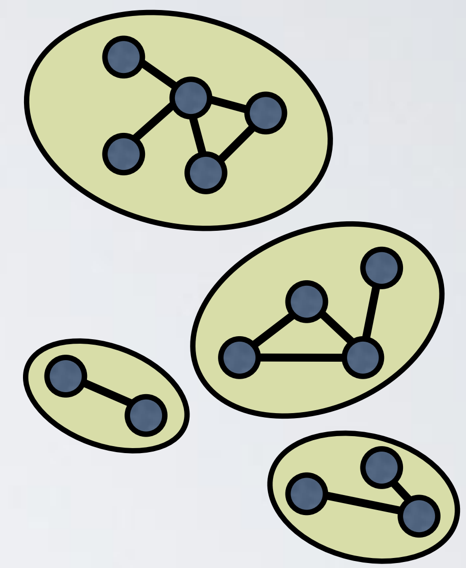
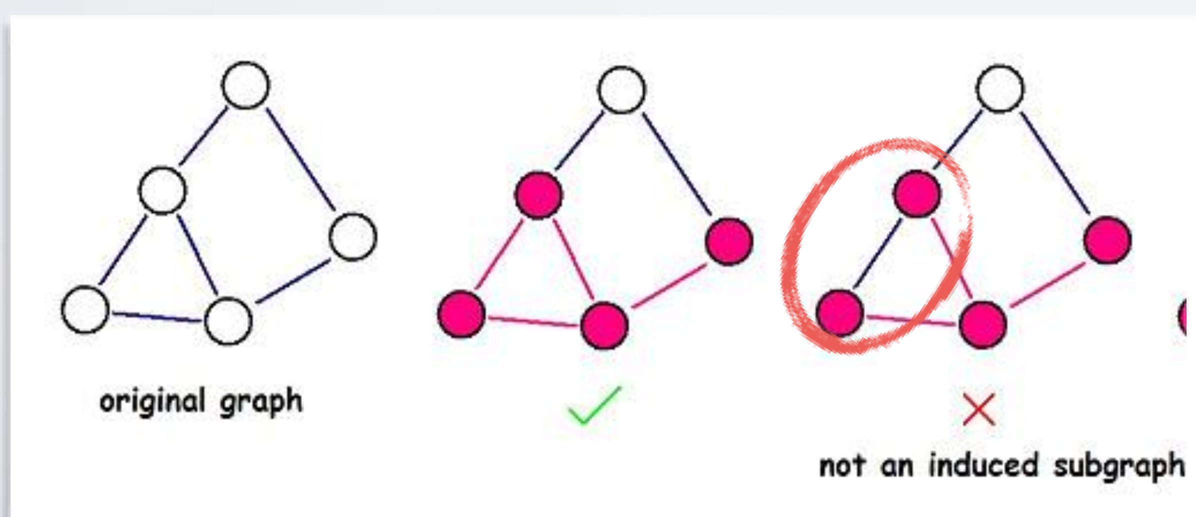
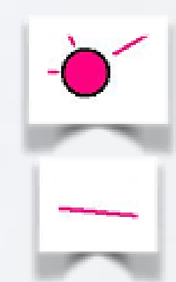


Figure after Newman, 2010



After “A. DZY Loves Physics”



Nodes/Edges in the subgraph

CLUSTERING COEFFICIENT

$\langle C \rangle$ - **Average clustering coefficient:** Average clustering coefficient of all nodes in the graph, $\bar{C} = \frac{1}{N}\sum_{u \in V} C_u$.

Be careful when interpreting this value, since all nodes contributes equally, irrespectively of their degree, and that low degree nodes tend to be much more frequent than hubs, and their C value is very sensitive, i.e., for a node u of degree 2, $C_u \in 0, 1$, while nodes of higher degrees tend to have more contrasted scores.

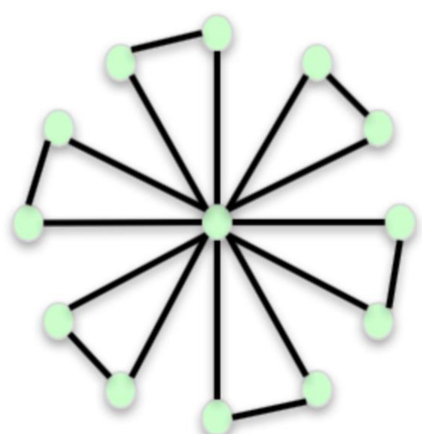
C^g - **Global clustering coefficient:** Fraction of all possible triangles in the graph that do exist, $C^g = \frac{3\Delta}{\Delta^{\max}}$

CLUSTERING COEFFICIENT

Global CC = Transitivity

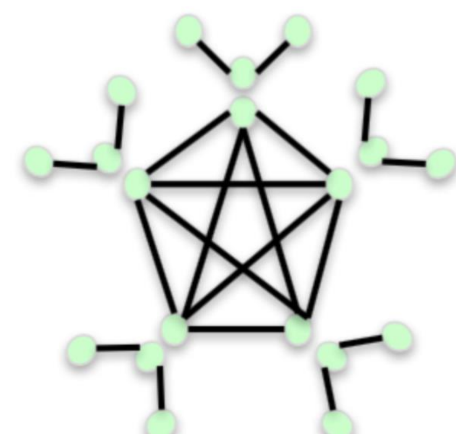
Transitivity vs. Average Clustering Coefficient

Both measure the tendency for edges to form triangles.
Transitivity weights nodes with large degree higher.



- Most nodes have high LCC
- The high degree node has low LCC

Ave. clustering coeff. = 0.93
Transitivity = 0.23



- Most nodes have low LCC
- High degree node have high LCC

Ave. clustering coeff. = 0.25
Transitivity = 0.86

https://pynetwork.readthedocs.io/en/latest/connectivity.html

CLUSTERING COEFFICIENT

- Global CC:
  - In random networks, GCC = density
    - =>very small for large graphs

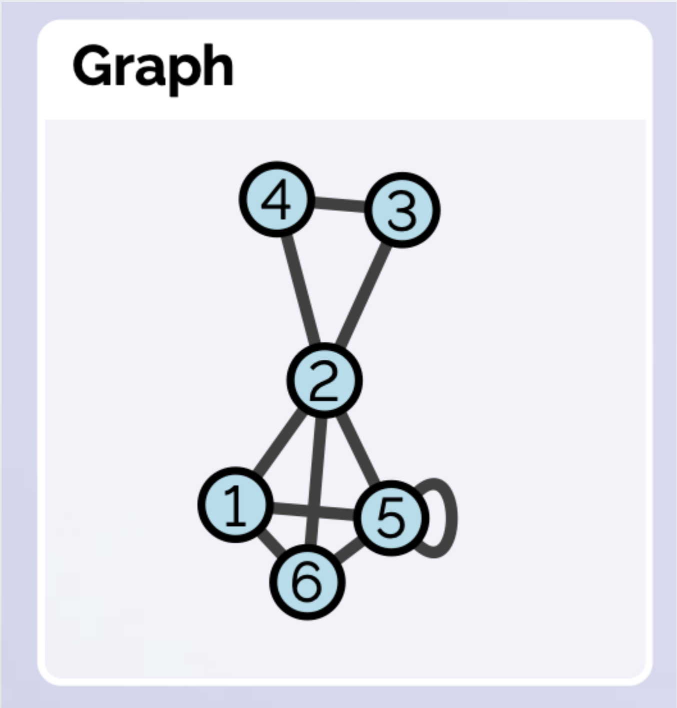
Network	Size	$\langle k \rangle$	C	C_{rand}	Reference
WWW, site level, undir.	153 127	35.21	0.1078	0.00023	Adamic, 1999
Internet, domain level	3015–6209	3.52–4.11	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001
Movie actors	225 226	61	0.79	0.00027	Watts and Strogatz, 1998
LANL co-authorship	52 909	9.7	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c
MEDLINE co-authorship	1 520 251	18.1	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c
SPIRES co-authorship	56 627	173	0.726	0.003	Newman, 2001a, 2001b, 2001c
NCSTRL co-authorship	11 994	3.59	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c
Math. co-authorship	70 975	3.9	0.59	5.4×10^{-5}	Barabási *et al.*, 2001
Neurosci. co-authorship	209 293	11.5	0.76	5.5×10^{-5}	Barabási *et al.*, 2001
E. coli, substrate graph	282	7.35	0.32	0.026	Wagner and Fell, 2000
E. coli, reaction graph	315	28.3	0.59	0.09	Wagner and Fell, 2000
Ythan estuary food web	134	8.7	0.22	0.06	Montoya and Solé, 2000
Silwood Park food web	154	4.75	0.15	0.03	Montoya and Solé, 2000
Words, co-occurrence	460.902	70.13	0.437	0.0001	Ferrer i Cancho and Solé, 2001
Words, synonyms	22 311	13.48	0.7	0.0006	Yook *et al.*, 2001b
Power grid	4941	2.67	0.08	0.005	Watts and Strogatz, 1998
C. Elegans	282	14	0.28	0.05	Watts and Strogatz, 1998

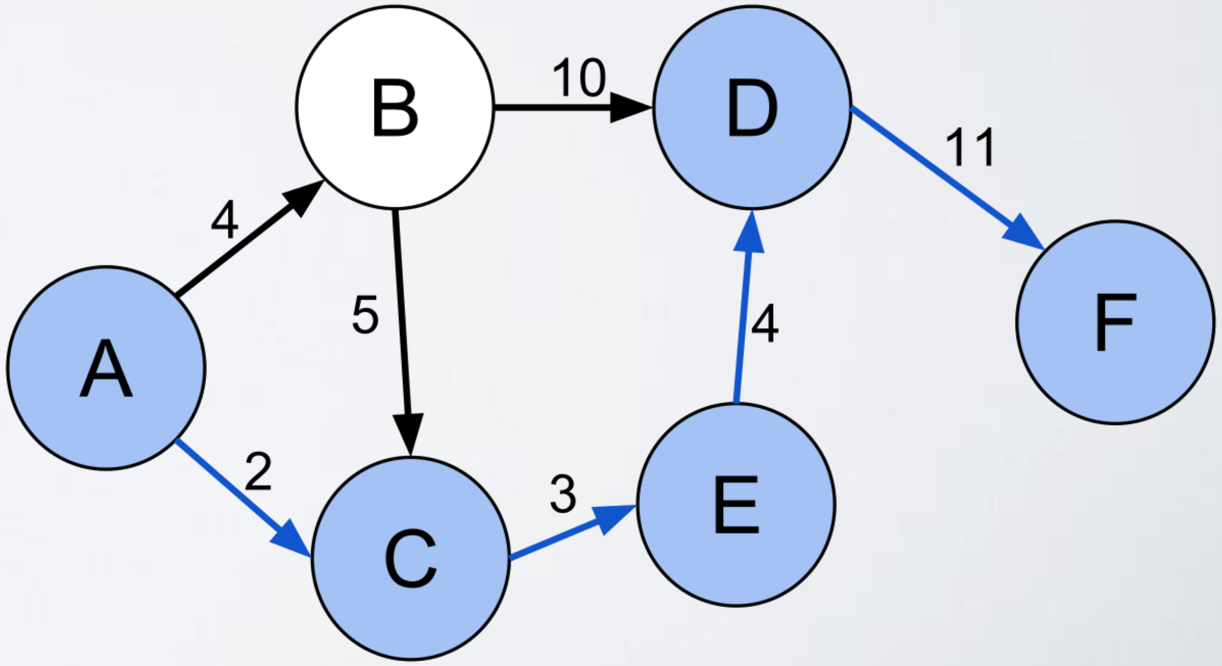
Albert, R. et.al. Rev. Mod. Phy. (2002)

PATH RELATED SCORES

Paths - Walks - Distance

Walk: Sequences of adjacent edges or nodes (e.g., **1.2.1.6.5** is a valid walk)
Path: a walk in which each node is distinct.
Path length: number of edges encountered in a path
Weighted Path length: Sum of the weights of edges on a path
Shortest path: The shortest path between nodes u, v is a path of minimal *path length*. Often it is not unique.
Weighted Shortest path: path of minimal *weighted path length*.
$\ell_{u,v}$: **Distance**: The distance between nodes u, v is the length of the shortest path





PATH RELATED SCORES

Network descriptors 2 - Paths

$\ell_{\max}$ | **Diameter**: maximum *distance* between any pair of nodes.

$\langle \ell \rangle$ | **Average distance**:

$$\langle \ell \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$$

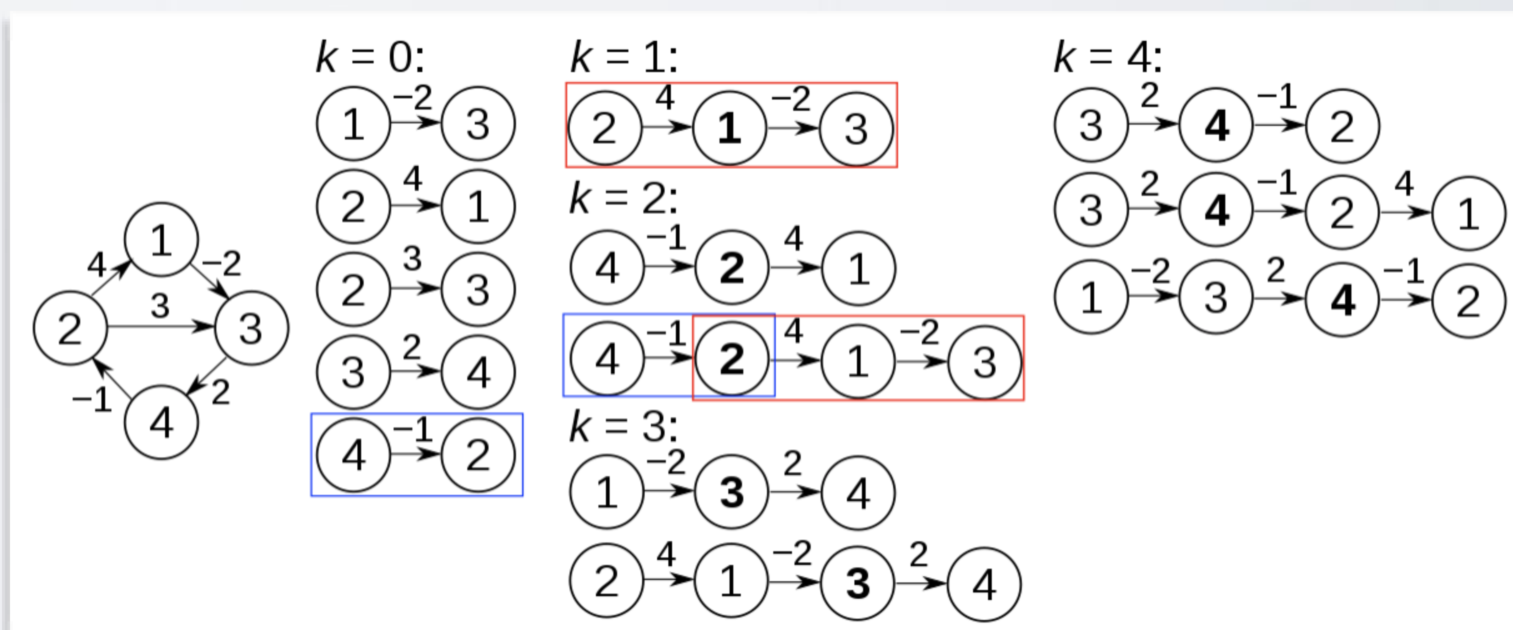
All shortest path algorithm

finding shortest paths in a **weighted graph** with **positive** or **negative edge weights** (but with no negative cycles)

```
proc FloydWarshall(G=(V,E,w))
// let dist be a |V| × |V| array of minimum distances initialized to ∞ (infinity)
for each edge (u,v)
   dist[u][v] ← w(u,v)  // the weight of the edge (u,v)
for each vertex v
   dist[v][v] ← 0
for k from 1 to |V|
   for i from 1 to |V|
      for j from 1 to |V|
         if dist[i][j] > dist[i][k] + dist[k][j]
             dist[i][j] ← dist[i][k] + dist[k][j]
         end if
```

Checking and updating all paths going through nodes k=1, 2, 3, … , N by assuming that:

shp(i,j,k)=
min(shp(i,j,k-1)), shp(i,k,k-1)+shp(k,j,k-1))



Complexity: $O(n^3)$

AVERAGE PATH LENGTH

- The famous 6 degrees of separation (Milgram experiment)
  - (More on that next slide)
- Not too sensible to noise
- Tells you if the network is “stretched” or “hairball” like

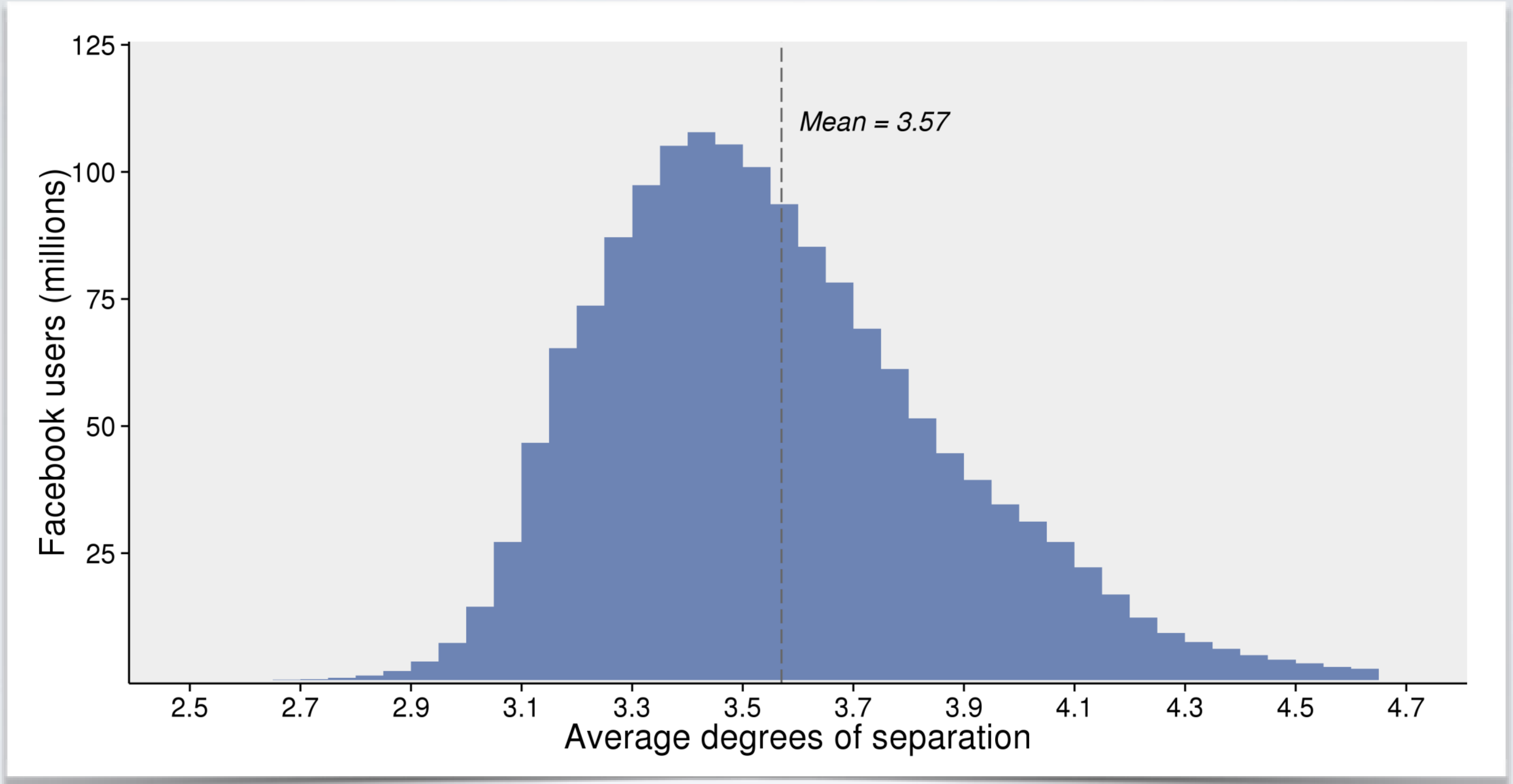
SIDE-STORY: MILGRAM EXPERIMENT

- Small world experiment (60's)
  - Give a (physical) mail to random people
  - Ask them to send to someone they don't know
    - They know his city, job
  - They send to their most relevant contact
- Results: In average, 6 hops to arrive

SIDE-STORY: MILGRAM EXPERIMENT

- Many criticism on the experiment itself:
  - Some mails did not arrive
  - Small sample
  - …
- Checked on “real” complete graphs (giant component):
  - MSN messenger
  - Facebook
  - The world wide web
  - …

SIDE-STORY: MILGRAM EXPERIMENT



Facebook

SMALL WORLD

Small World Network

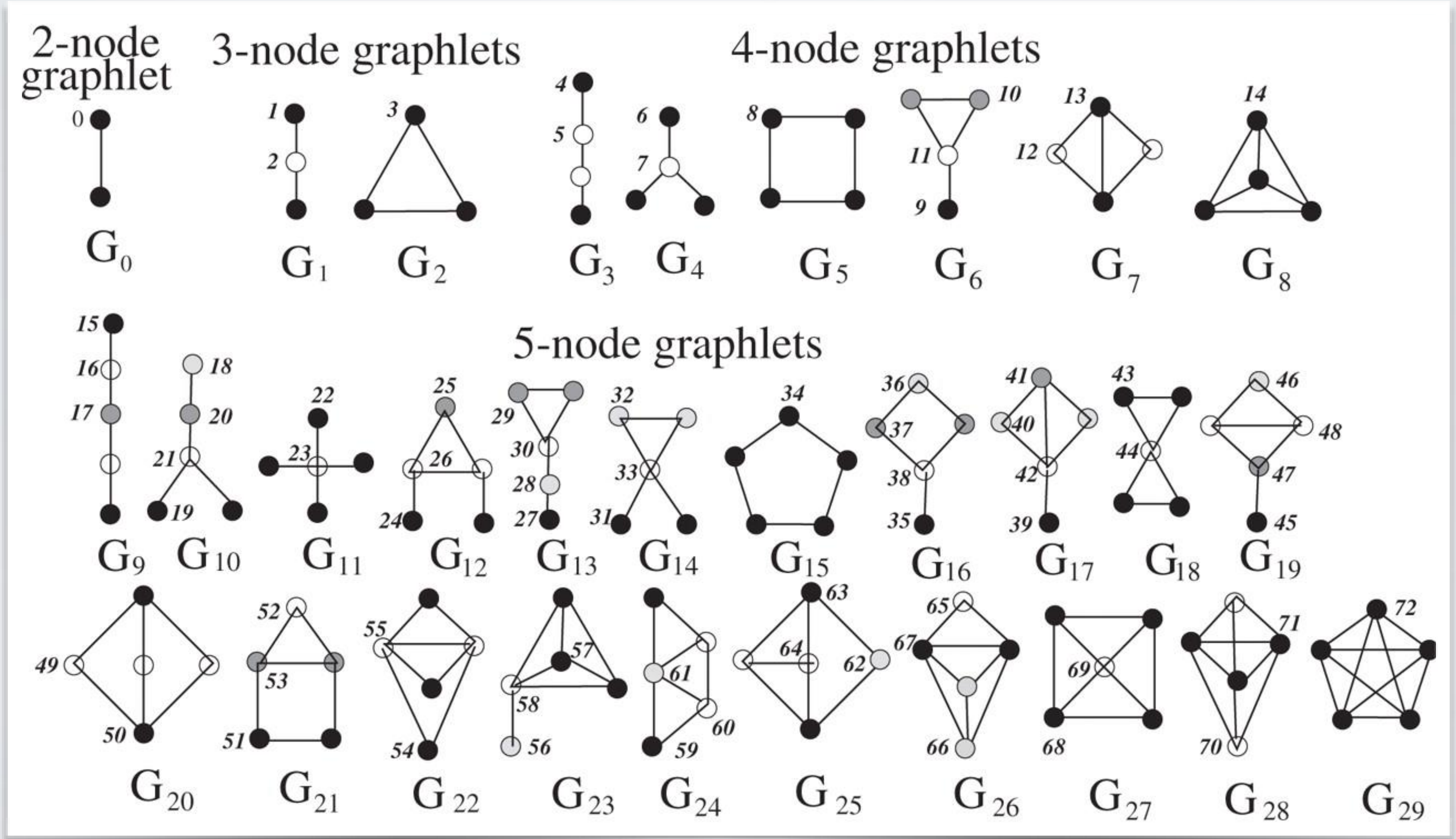
A network is said to have the **small world** property when it has some structural properties. The notion is not quantitatively defined, but two properties are required:

- Average distance must be short, i.e., $\langle \ell \rangle \approx \log(N)$
- Clustering coefficient must be high, i.e., much larger than in a random network , e.g., $C^g \gg d$, with d the network density

NETWORK DESCRIPTORS

- Many other network descriptors exist:
  - Modularity (later in community detection class)
  - Centralization (comparing the centrality scores between most central and less central, see later)
  - Rich-club coefficient: tendency of high-degrees to connected to high-degrees, cf random network class
  - Motif profiles (how often do specific subgraphs appear)
  - Network Resilience (see practicals)
  - etc.

GRAPHLETS

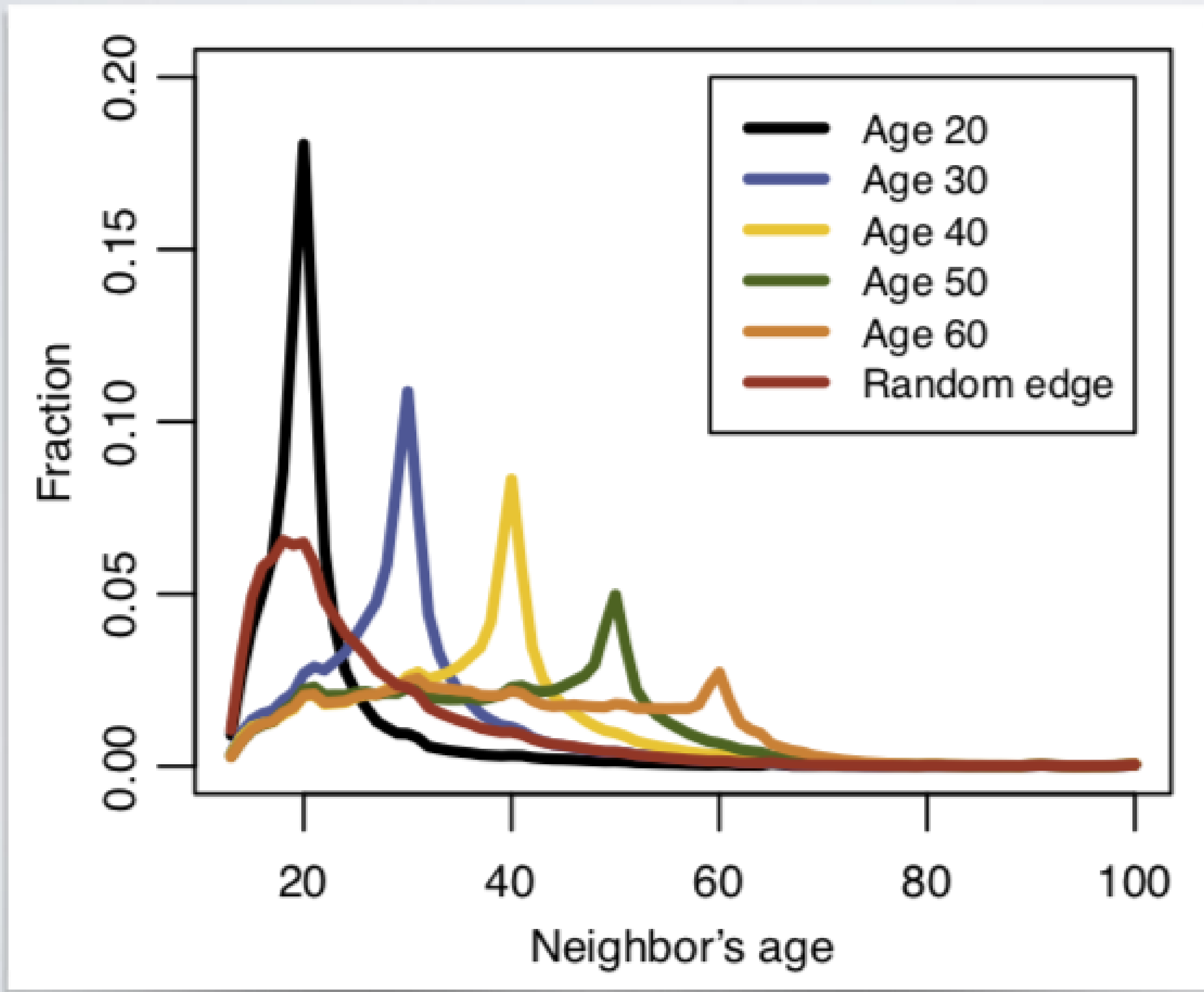


2-node graphlet
3-node graphlets
4-node graphlets
5-node graphlets
0
G0
1
2
3
G1
G2
4
5
6
7
8
9
10
11
12
13
14
G3
G4
G5
G6
G7
G8
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
G9
G10
G11
G12
G13
G14
G15
G16
G17
G18
G19
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
G20
G21
G22
G23
G24
G25
G26
G27
G28
G29

EXEMPLE OF GRAPH ANALYSIS

- 721M users (nodes) (active in the last 28 days)
- 68B edges
- Average degree: 190 (average # friends)
- Median degree: 99
- Connected component: 99.91%

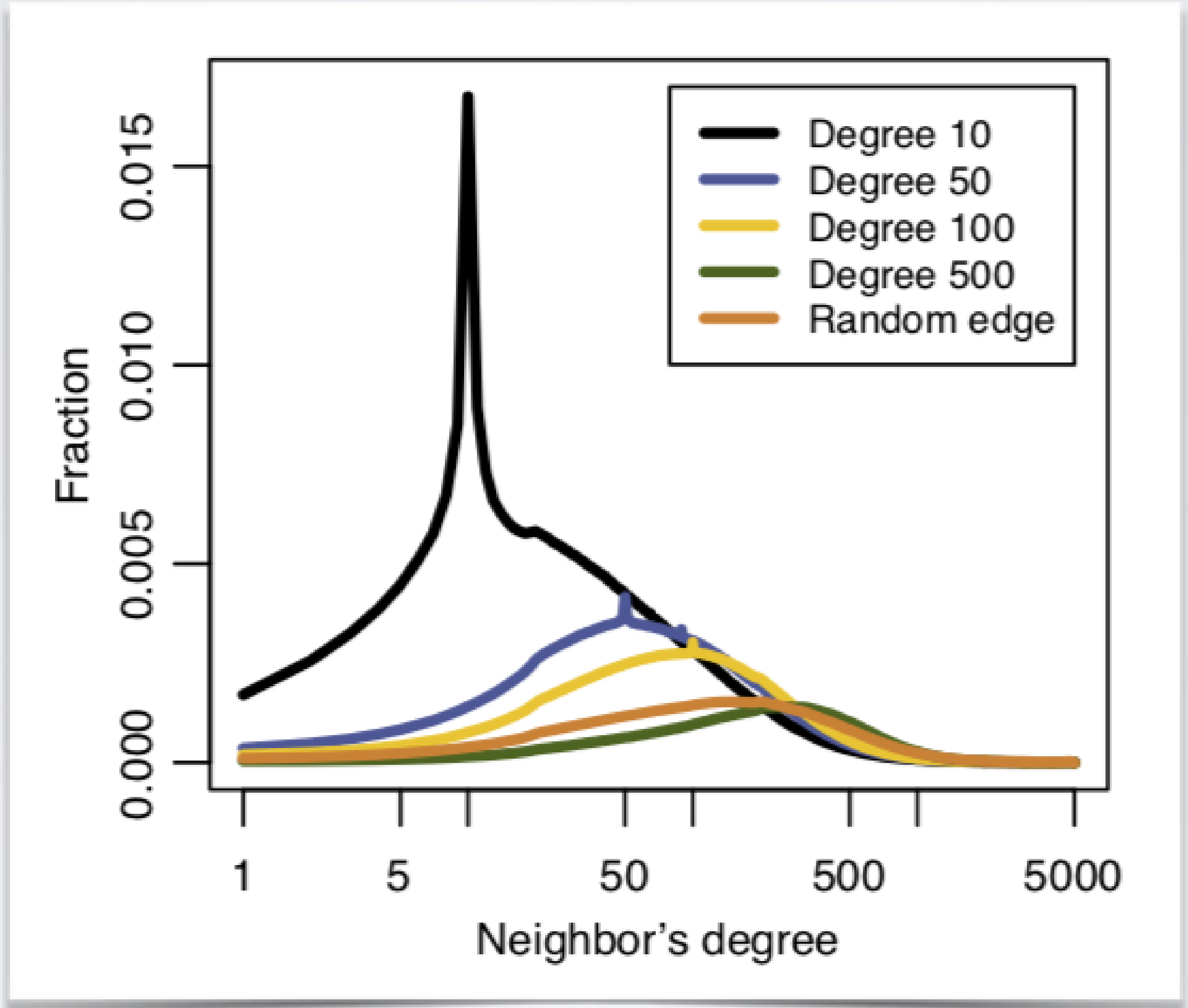
EXEMPLE OF GRAPH ANALYSIS



Age homophily

(More next class)

EXEMPLE OF GRAPH ANALYSIS



Many of my friends have the
Same # of friends than me!

CENTRALITIES

Characterizing/Discovering important nodes

CENTRALITY

- We can measure nodes importance using so-called **centrality**.
- Poor terminology: nothing to do with being central in general
- Usage:
  - Some centralities have straightforward interpretation
  - Centralities can be used as ***node features*** for machine learning on graph
    - (Classification, link prediction, …)

NODE DEGREE

- **Degree**: how many neighbors

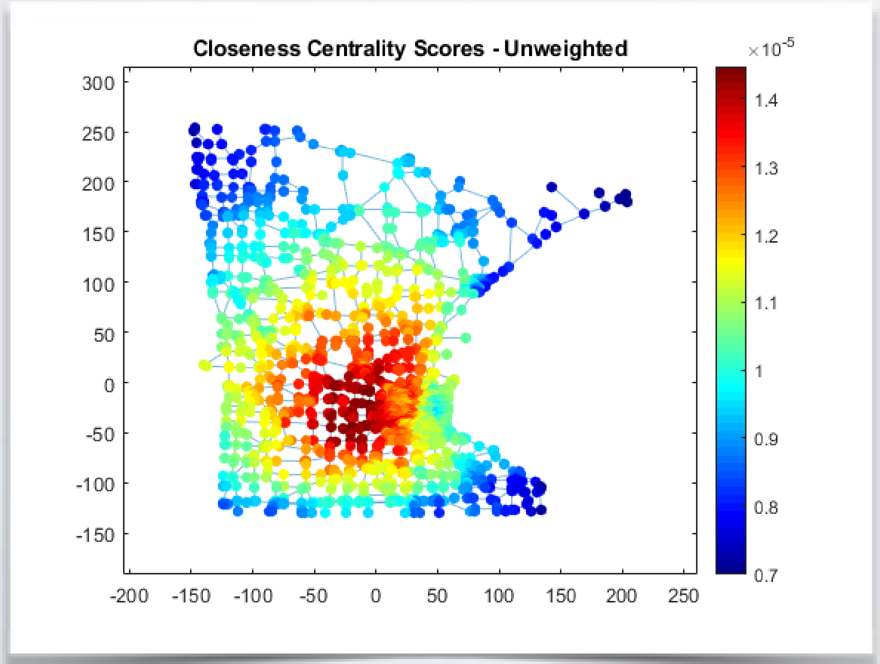
- Often enough to find important nodes
  - Main characters of a series talk with the more people
  - Largest airports have the most connections
  - …

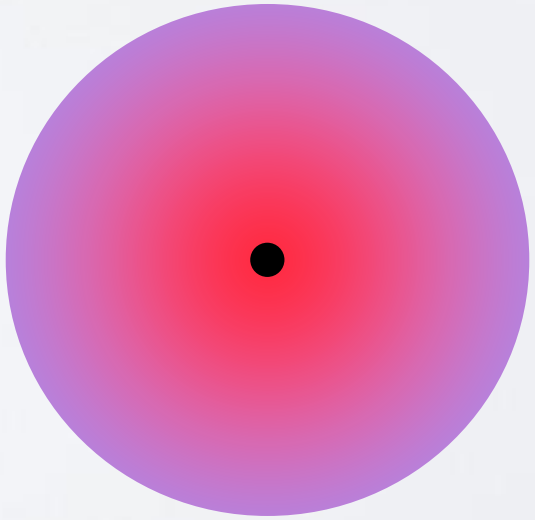
- But not always
  - Facebook users with the most friends are spam
  - Webpages/wikipedia pages with most links are simple lists of references
  - …

FARNESS, CLOSENESS HARMONIC CENTRALITY

FARNESS, CLOSENESS

- How close the node is to all other nodes
- Parallel with the center of a figure:
  - Center of a circle is the point of shorter average distance to any points in the circle





FARNESS, CLOSENESS

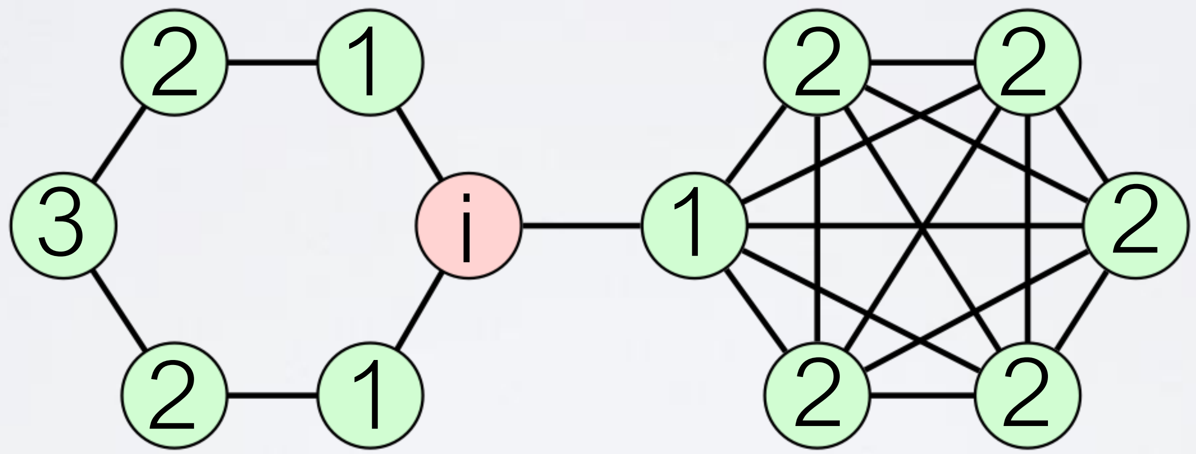
Farness: Average distance to all other nodes in the graph

$$\text{Farness}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \ell_{u,v}$$

CLOSENESS CENTRALITY

Closeness: Inverse of the farness, i.e., how close the node is to all other nodes in term of shortest paths.

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$

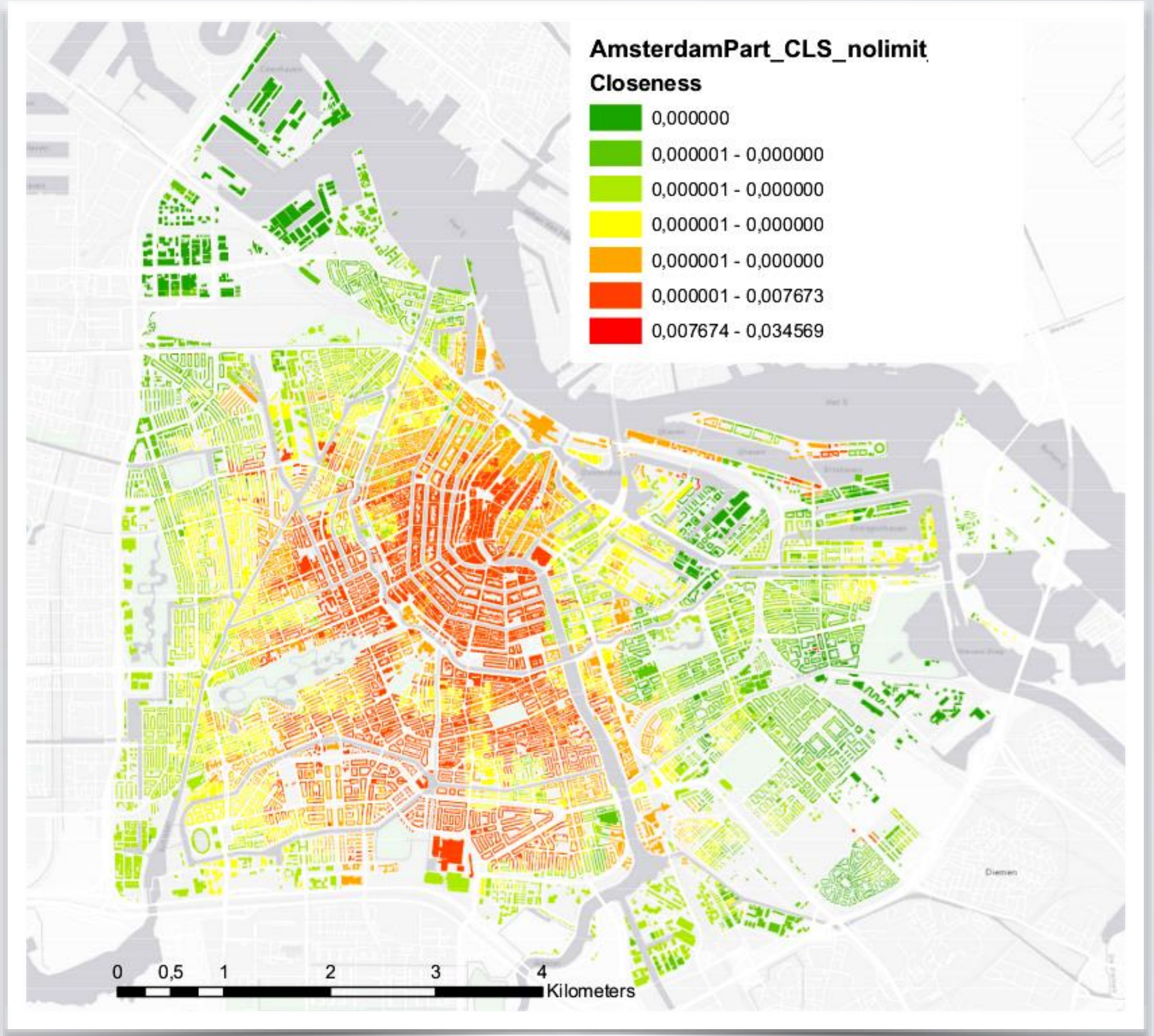


$$C_{cl}(i) = \frac{12-1}{(3 \times 1 + 7 \times 2 + 1 \times 3)} = \frac{11}{20} = 0.55$$

CLOSENESS CENTRALITY

Closeness: Inverse of the farness, i.e., how close the node is to all other nodes in term of shortest paths.

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$

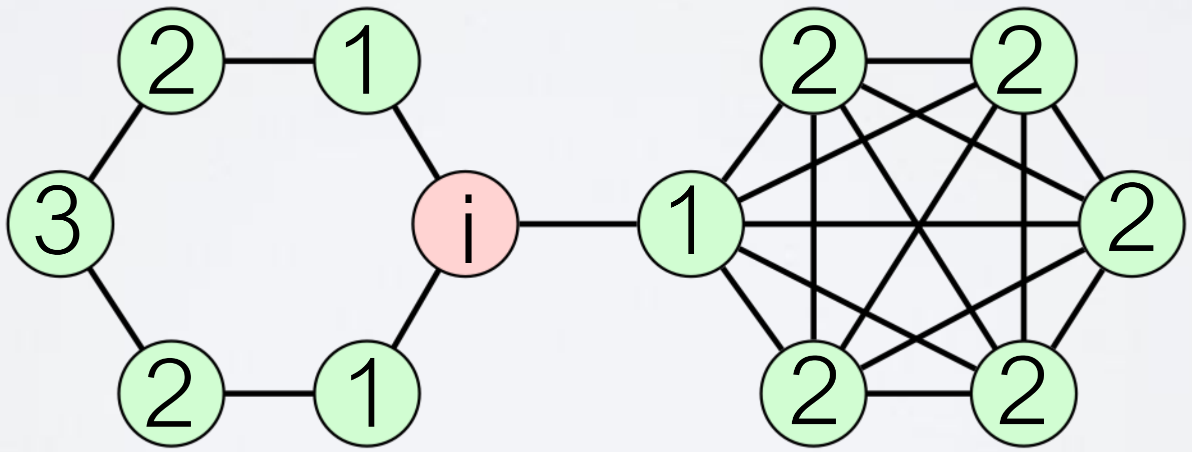


1=all nodes are at distance one

Harmonic Centrality

Harmonic centrality: A variant of the closeness defined as the average of the inverse of distance to all other nodes (Harmonic mean). Well defined on disconnected network with $\frac{1}{\infty} = 0$. Its interpretation is the same as the closeness.

$$\text{Harmonic}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \frac{1}{\ell_{u,v}}$$



$$C_h(i) = \frac{1}{12-1}\left(3 \times \frac{1}{1} + 7 \times \frac{1}{2} + 1 \times \frac{1}{3}\right) = \frac{41}{66} = 0.6212$$

BETWEENNESS CENTRALITY

- Measure how much the node plays the role of a bridge
- Betweenness of ***u:*** ***f***raction of all the shortest paths between all the pairs of nodes going through u.

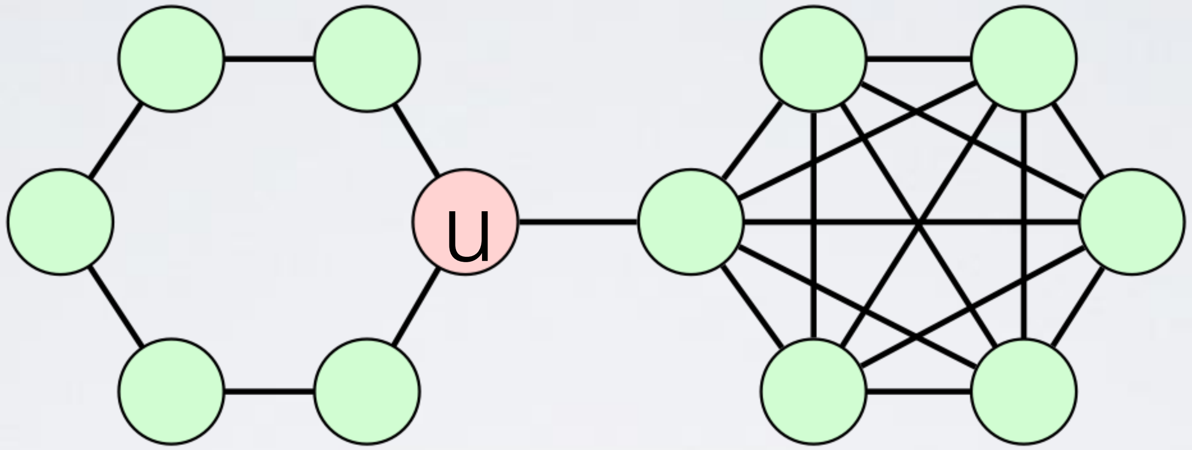
$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with σ_{st} the number of shortest paths between nodes s and t and $\sigma_{st}(v)$ the number of those paths passing through v.
The betweenness tends to grow with the network size. A normalized version can be obtained by dividing by the number of pairs of nodes, i.e., for a directed graph: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.

Betweenness Centrality

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

directed graph: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.



$$C_B(u) = 2\frac{5 * 6 + 1 + \frac{1}{2} + \frac{1}{2}}{11 * 10} = \frac{64}{110}$$

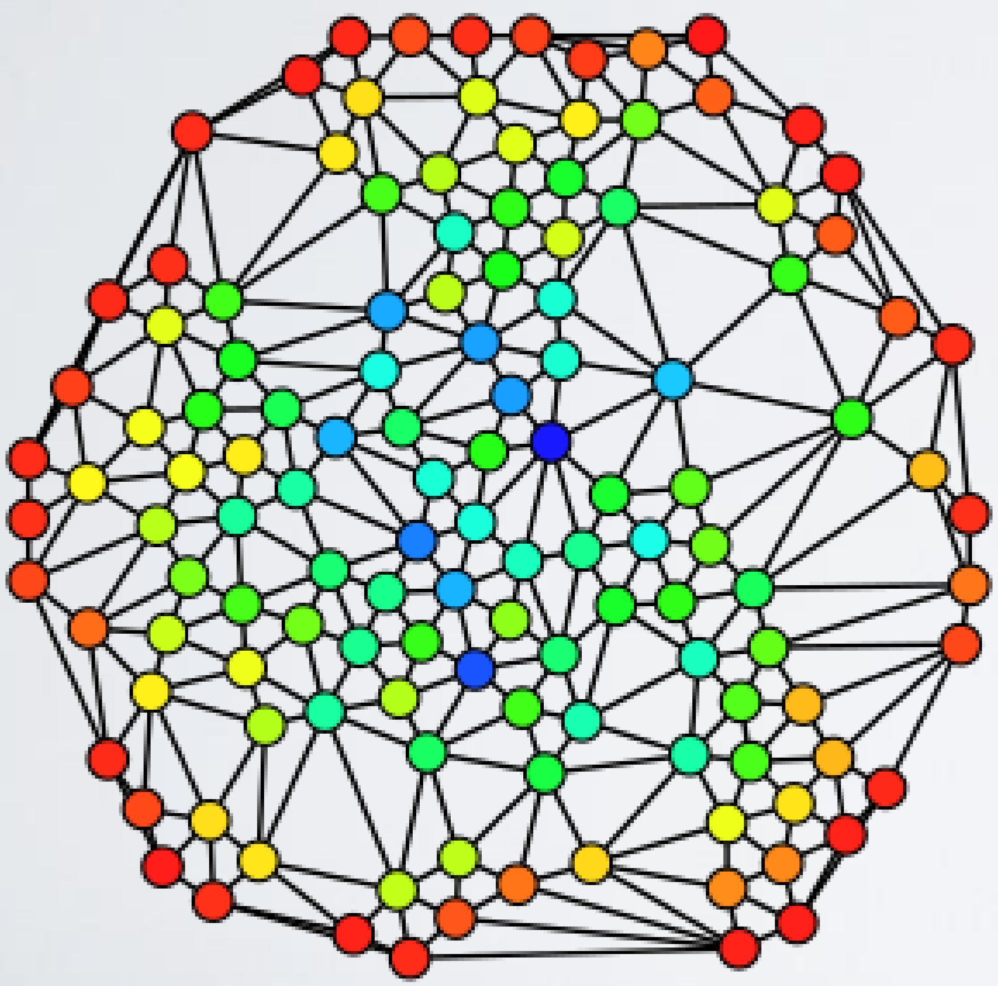
Exact computation:

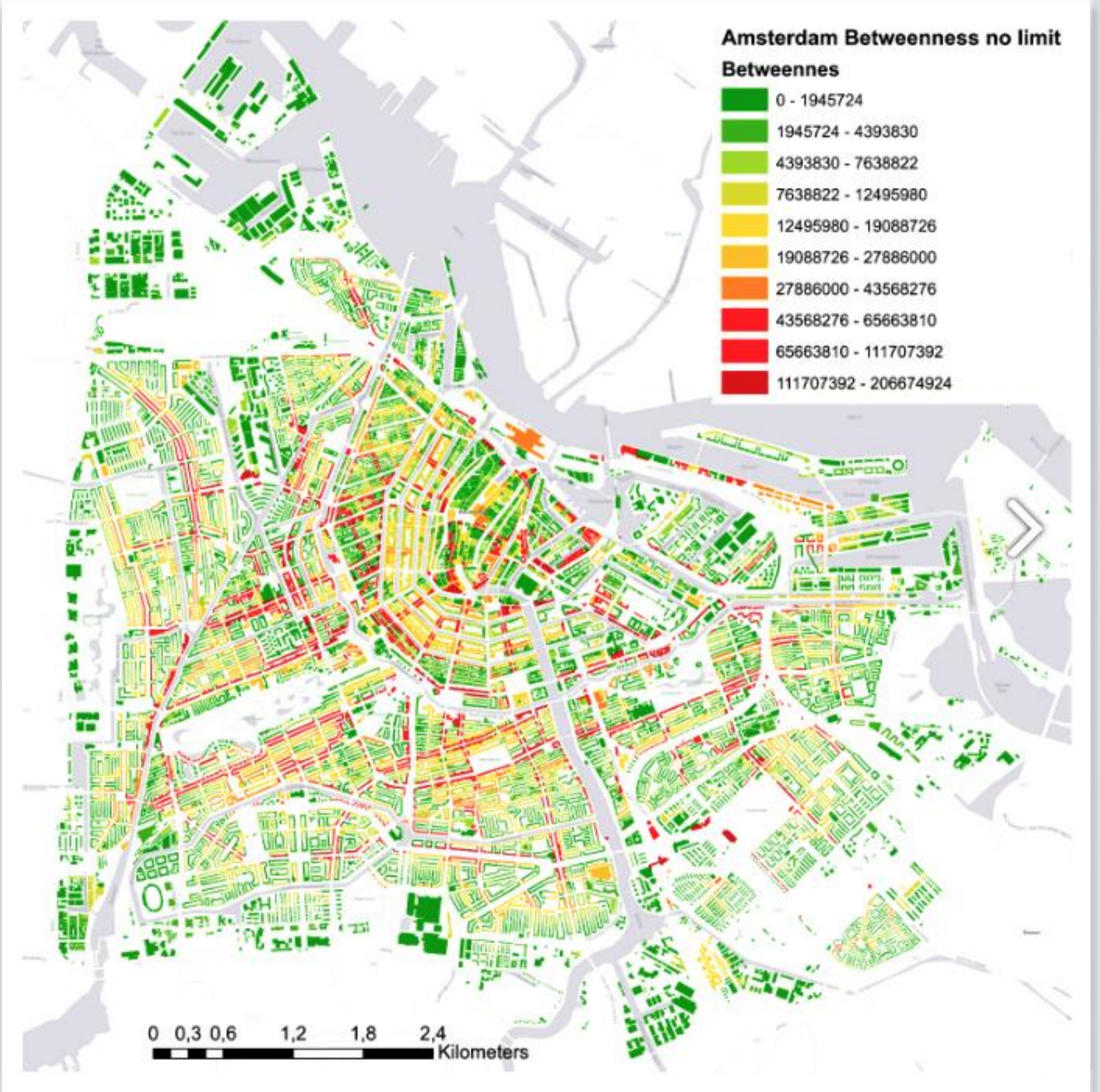
Floyd-Warshall: *$O(n^3)$ time complexity*
$O(n^2)$ space complexity

Approximate computation

Dijskstra: *O(n(m+n log n)) time complexity*

BETWEENNESS CENTRALITY





(blue higher)

(red higher)

EDGE - BETWEENNESS

Same definition as for nodes

Can you guess the edge of highest betweenness in the European rail network ?

RECURSIVE DEFINITIONS

RECURSIVE DEFINITIONS

- Recursive importance:
  - **Important nodes** are those connected ***to important nodes***
- Several centralities based on this idea:
  - Eigenvector centrality
  - PageRank
  - …

RECURSIVE DEFINITION

- We would like scores such as :
  - Each node has a score (centrality),
  - If every node "sends" its score to its neighbors, the sum of all scores received by each node will be equal to its original score

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad (1)$$

- With λ a normalisation constant

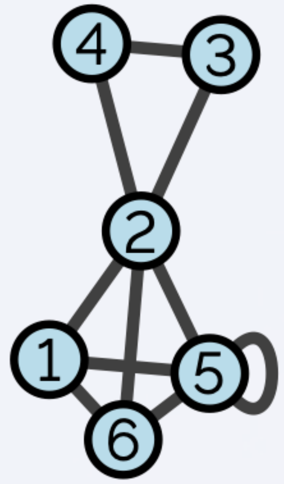
RECURSIVE DEFINITION

- This problem can be solved by what is called the ***power method:***
  - 1) We initialize all scores to random values
  - 2)Each score is updated according to the desired rule, until reaching a stable point (after normalization)

- Why does it converge?
  - Perron-Frobenius theorem (see next slide)
  - =>True for undirected graphs with a single connected component

ADJACENCY MATRIX

Graph



A **- Adjacency Mat.**

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

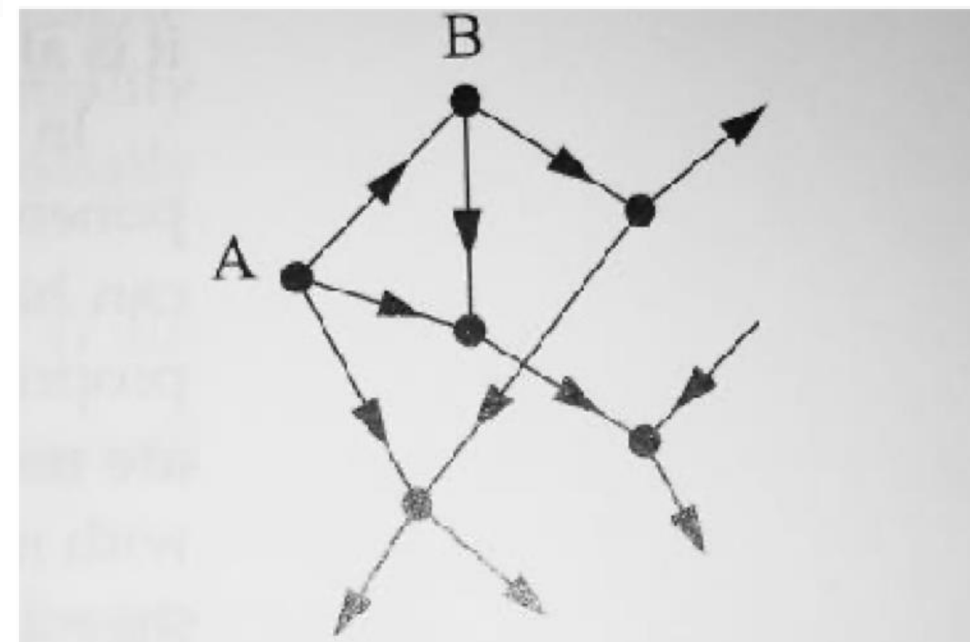
EIGENVECTOR CENTRALITY

- What we just described is called the Eigenvector centrality
- A couple eigenvector (x) and eigenvalue (λ) is defined by the following relation: $Ax = \lambda x$
  - x is a column vector of size ***n,*** which can be interpreted as the scores of nodes
- What Perron-Frobenius algorithm says is that the power method will always converge to the ***leading eigenvector***, i.e., the eigenvector associated with the highest eigenvalue

Eigenvector Centrality

Some problems in case of directed network:

- Adjacency matrix is asymmetric
- 2 sets of eigenvectors (Left & Right)
- 2 leading eigenvectors
    - Use right eigenvectors : consider nodes that are pointing towards you



-Vertex A is connected but has only outgoing link = Its centrality will be 0

-Vertex B has outgoing and an incoming link, but incoming link comes from A

= Its centrality will be 0

-etc.

Solution: Only in strongly connected component

Note: Acyclic networks (citation network) do not have strongly connected component

PageRank Centrality

- Eigenvector centrality generalised for directed networks

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

PageRank Centrality

- Eigenvector centrality generalised for directed networks

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

PAGERANK

- 2 main improvements over eigenvector centrality:
  - In directed networks, problem of source nodes
    - => Add a constant centrality gain for every node
  - Nodes with very high centralities give very high centralities to all their neighbors (even if that is their only in-coming link)
    - => What each node "is worth" is divided equally among its neighbors (normalization by the degree)

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad \Rightarrow \quad C_u^{t+1} = \alpha \sum_{v \in N_u^{in}} \frac{C_v^t}{k_v^{out}} + \beta$$

With by convention β=1 and α a parameter (usually 0.85) controlling the relative importance of β

PAGERANK

Matrix interpretation

Principal eigenvector of the “Google Matrix”:

First, define matrix S as:

- -Normalization by columns of A
- -Columns with only 0 receives 1/n (dead end)

-Finally, $G_{ij} = \alpha S_{ij} + (1-\alpha)/n$

Removing some trip probability from out-link

And distributing them at random among other nodes

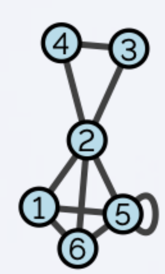
((1-0.85)/5=0.03)

(a) $$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

(c) $$S = \begin{pmatrix} 0 & 1/2 & 1/3 & 0 & 1/5 \\ 1 & 0 & 1/3 & 1/3 & 1/5 \\ 0 & 1/2 & 0 & 1/3 & 1/5 \\ 0 & 0 & 1/3 & 0 & 1/5 \\ 0 & 0 & 0 & 1/3 & 1/5 \end{pmatrix}$$

(e) $$G = \begin{pmatrix} 0.03 & 0.455 & 0.313 & 0.03 & 0.2 \\ 0.88 & 0.03 & 0.313 & 0.313 & 0.2 \\ 0.03 & 0.455 & 0.03 & 0.313 & 0.2 \\ 0.03 & 0.03 & 0.313 & 0.03 & 0.2 \\ 0.03 & 0.03 & 0.03 & 0.313 & 0.2 \end{pmatrix}$$

Graph



A - Adjacency Mat.

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

Random W. mat.

$$\begin{pmatrix} 0 & \frac{1}{5} & 0 & 0 & \frac{1}{4} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{3} \\ 0 & \frac{1}{5} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{5} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{5} & 0 & 0 & \frac{1}{4} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{5} & 0 & 0 & \frac{1}{4} & 0 \end{pmatrix}$$

PageRank - as Random Walk

Main idea: **The PageRank computation can be interpreted as a Random Walk process with restart**

Teleportation probability: the parameter α gives the probability that in the next step of the RW will follow a Markov process or with probability 1-α it will jump to a random node

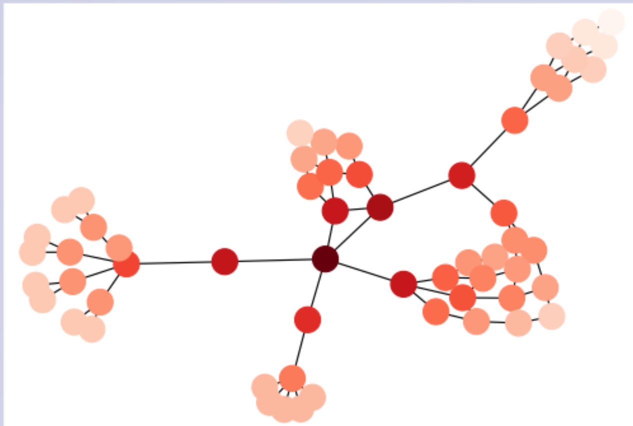
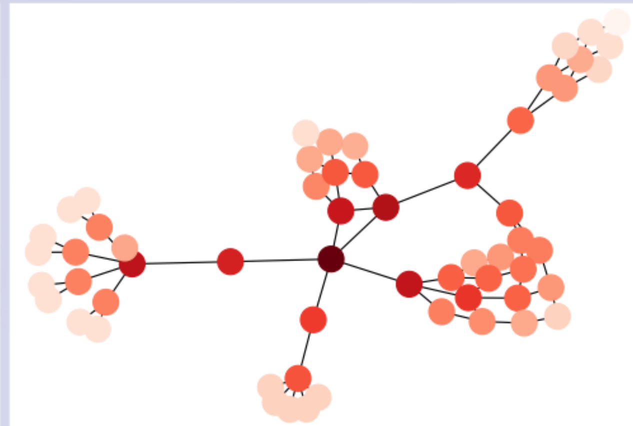
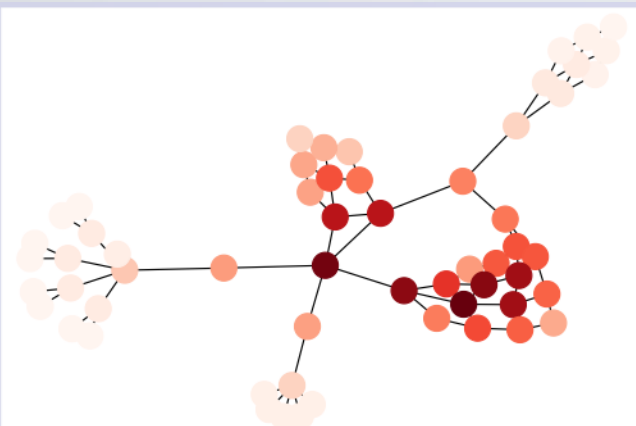
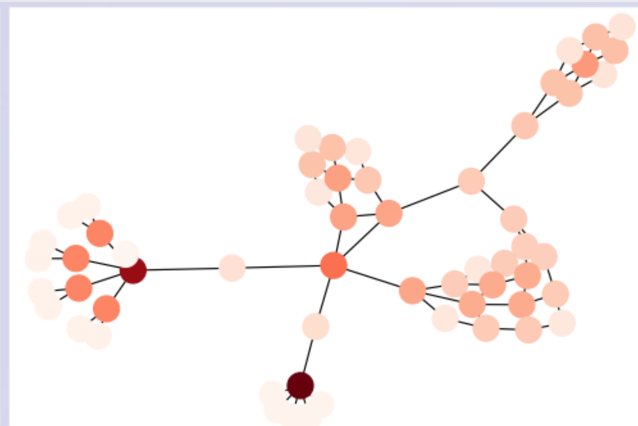
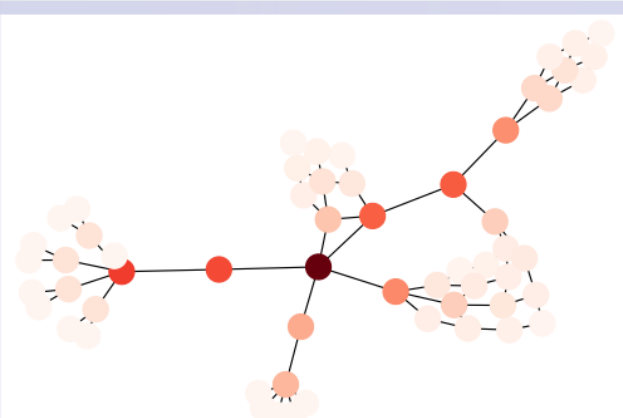
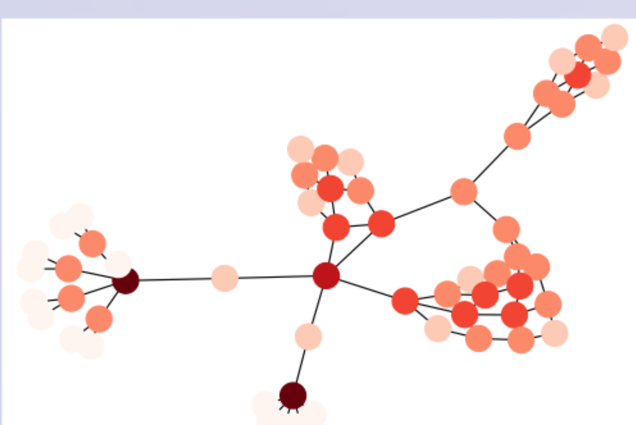
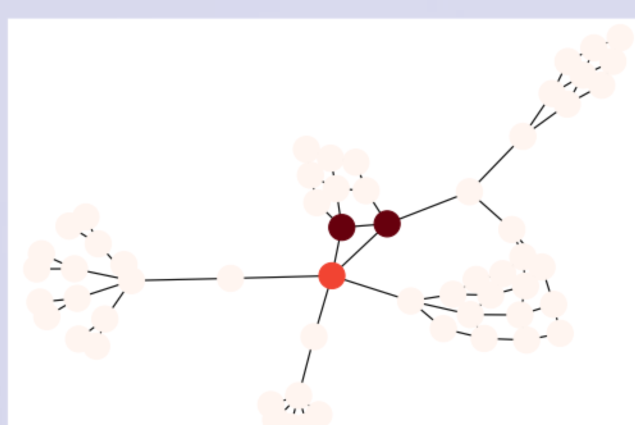
Pagerank score of a node thus corresponds to the probability of this random walker to be on this node after an infinite number of hops.

PAGERANK

- Then how do Google rank when we do a research?
- Compute pagerank (using the power method for scalability)
- Create a subgraph of documents related to our topic
- Of course now it is certainly much more complex, but we don't really know:
  "Most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art" [Page, Brin, 1997]

OTHERS

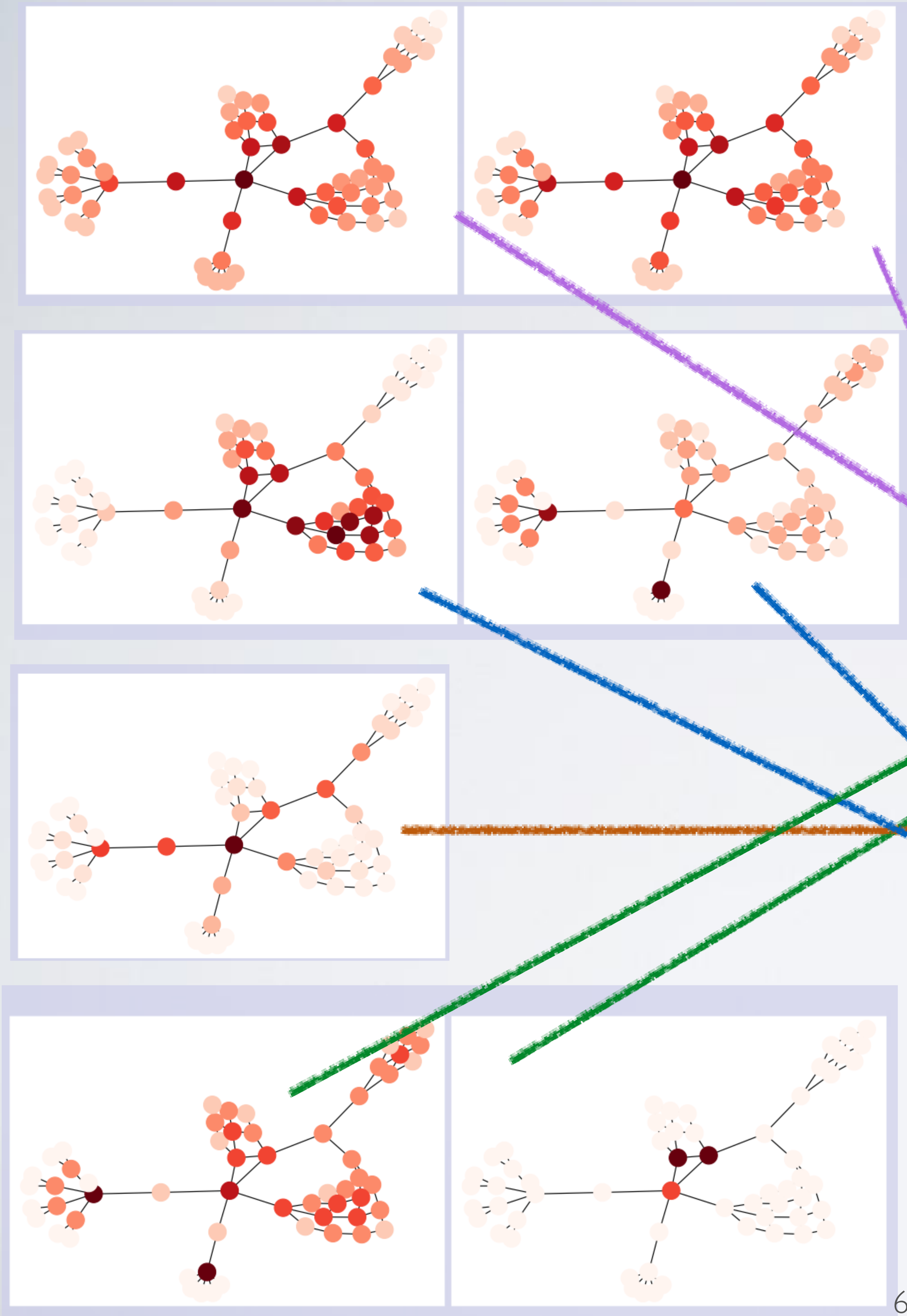
- Many other centralities have been proposed
  - 50+ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4646361/)
- The problem is how to interpret them ?
- Can be used as supervised tool:
  - Compute many centralities on all nodes
  - Learn how to combine them to find chosen nodes
  - Discover new similar nodes
  - (roles in social networks, key elements in an infrastructure, …)

Which is which ?

Degree
Clustering coefficient
Closeness
Harmonic Centrality
Betweenness
Eigenvector
PageRank

Which is which ?



Degree

Clustering coefficient

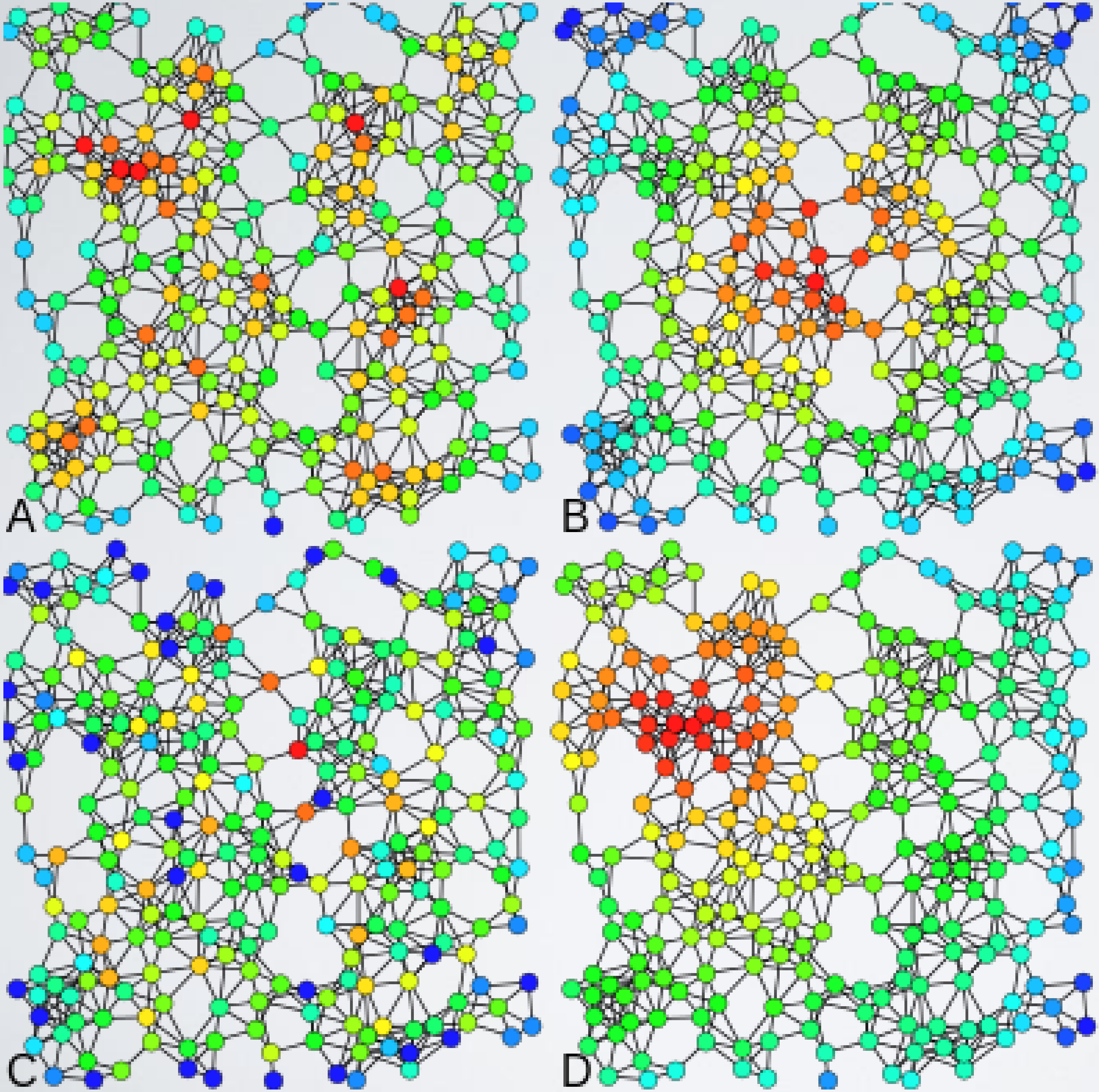
Closeness

Harmonic Centrality

Betweenness

Eigenvector

PageRank



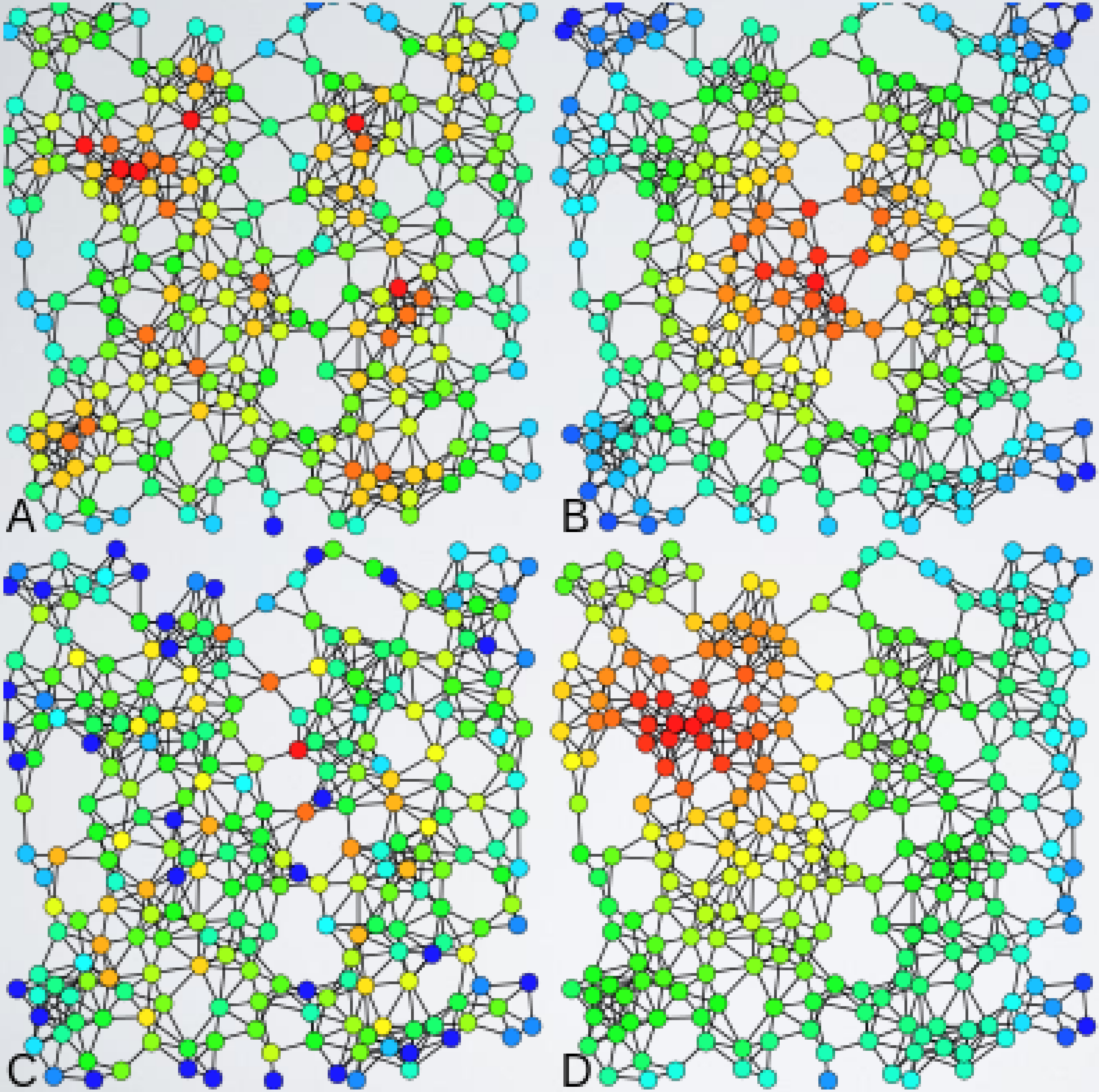
Try again :)

Degree

Betweenness

Closeness

Eigenvector



Try again :)

Degree: A
Betweenness: C
Closeness: B
Eigenvector: D