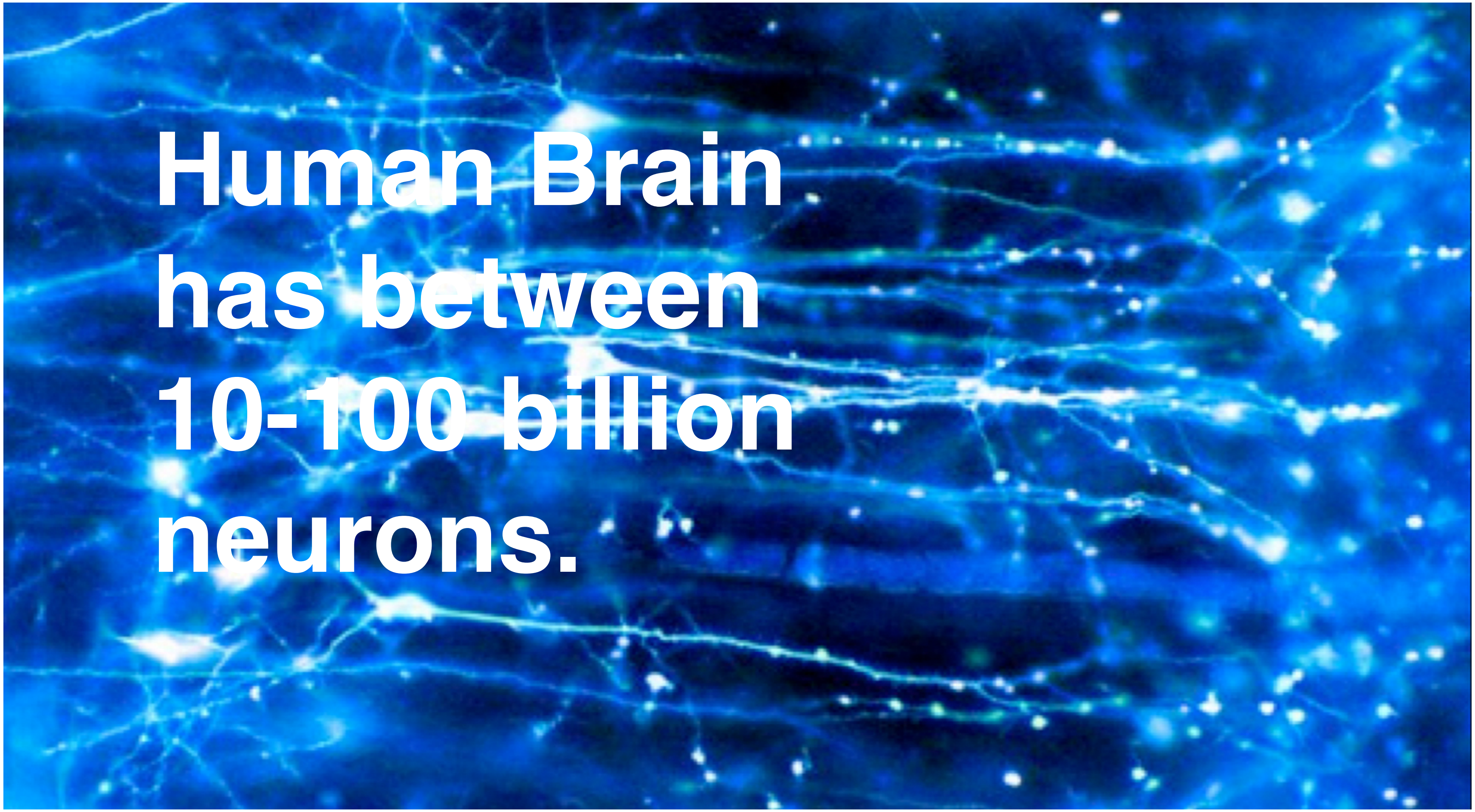
S. Rinzivillo – rinzivillo@isti.cnr.it

DATA VISUALIZATION AND VISUAL ANALYTICS

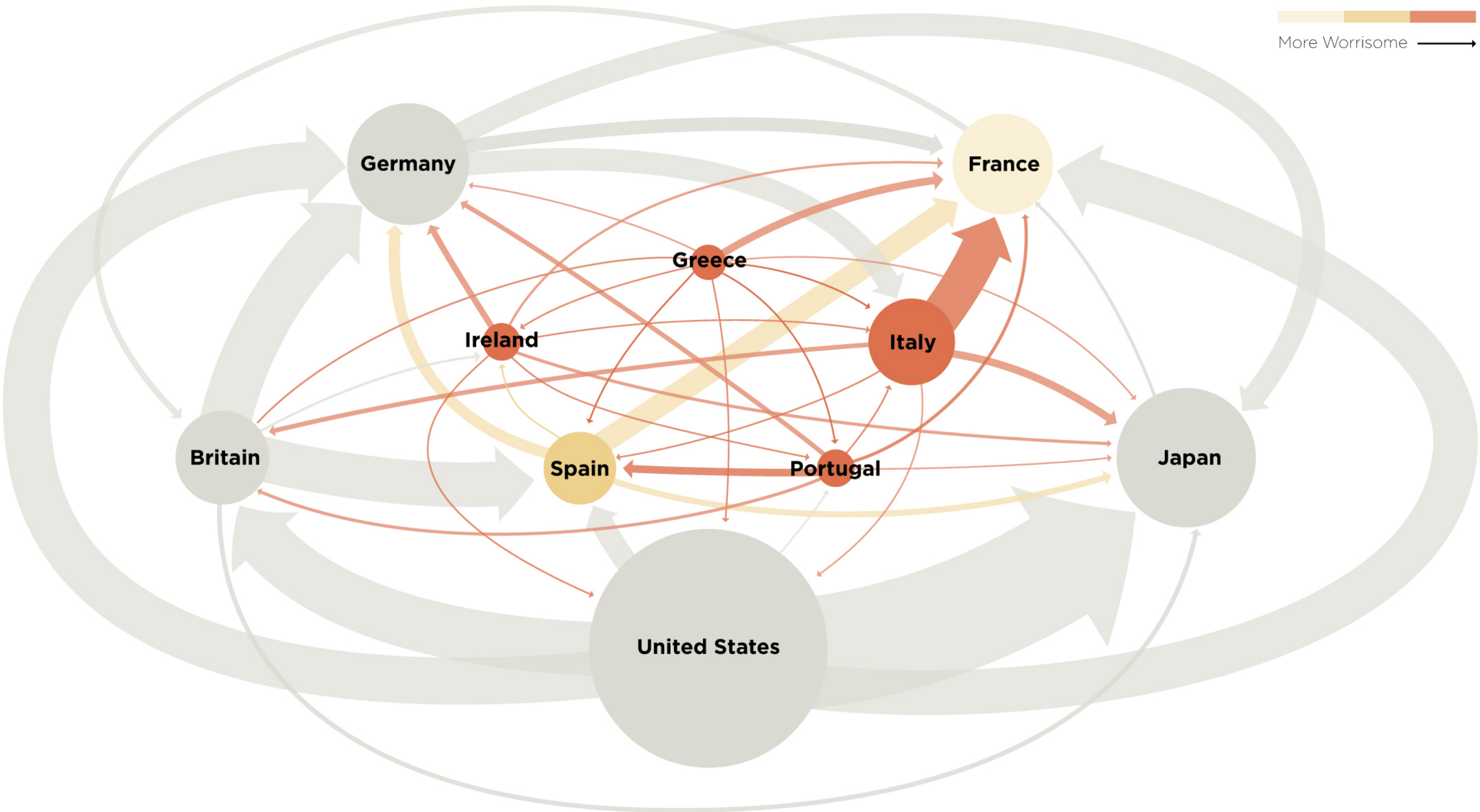
Relational Structures: Networks

Networks, networks everywhere

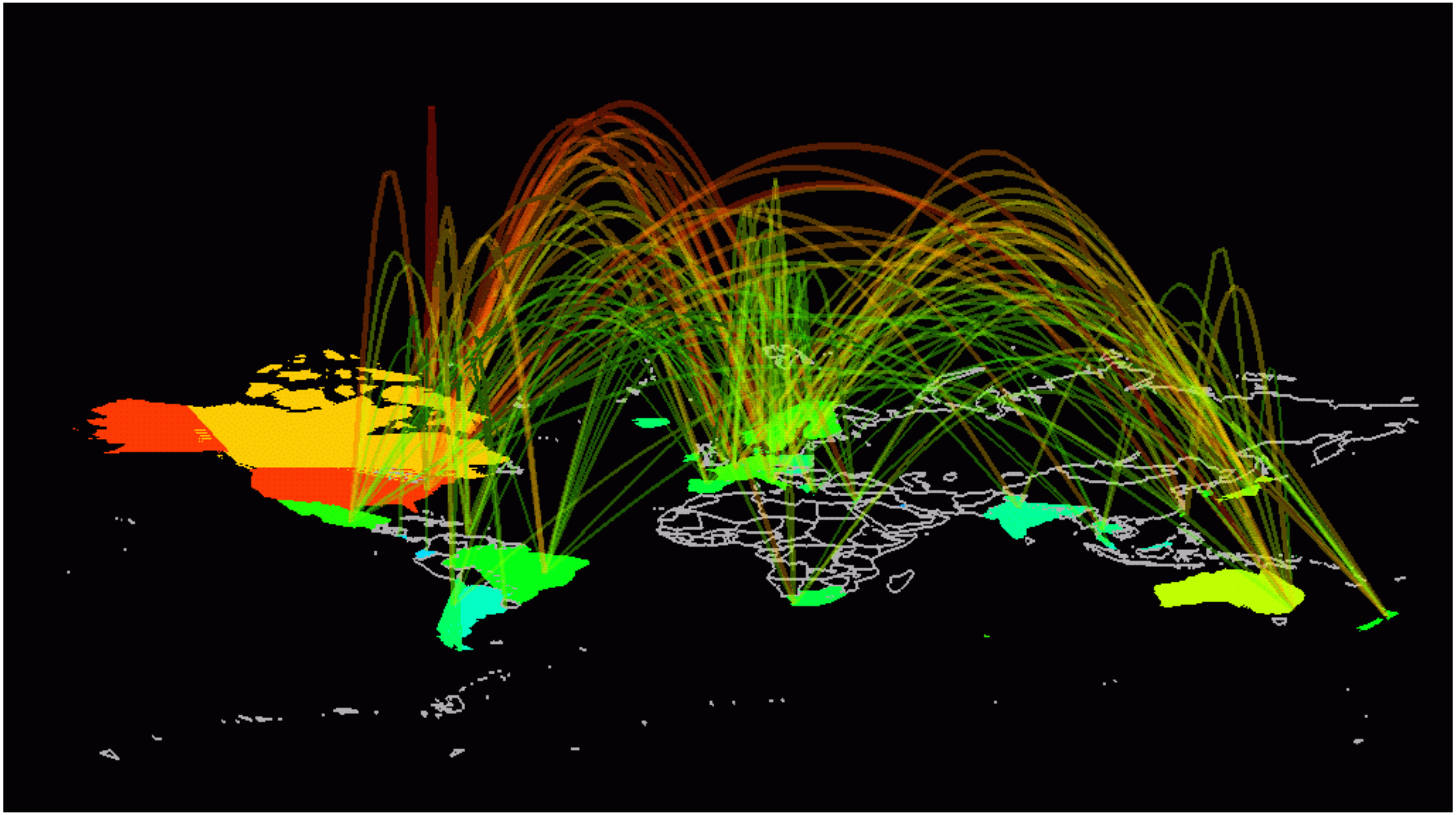
Networks, networks everywhere



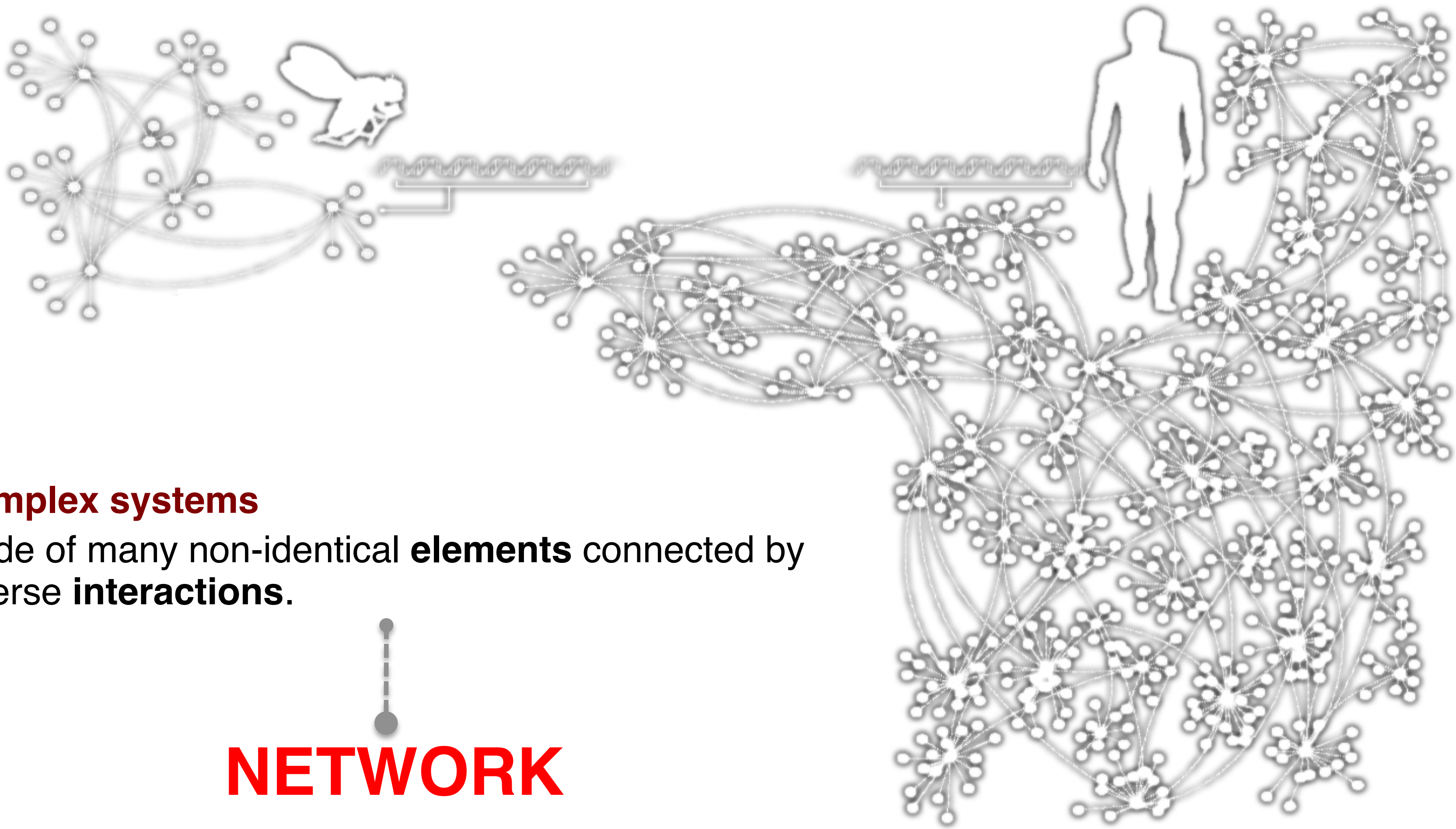
Financial network



Internet



Human Gene



Complex systems

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

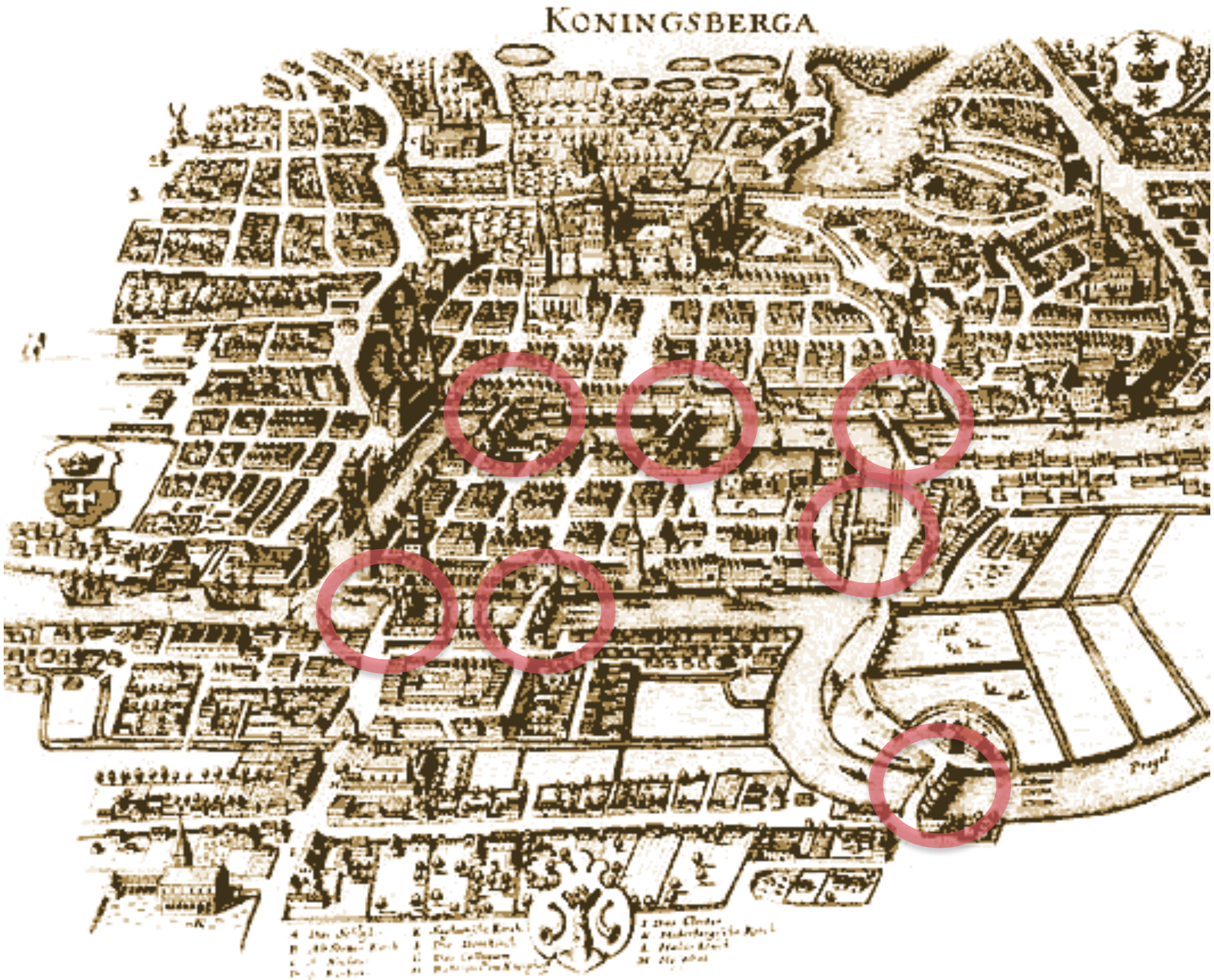
NETWORK

Networks

- Data main focus is relationship
- Study the patterns of connection among different parts of a complex system
- Visualization has a key roles to add insights to numerical analysis

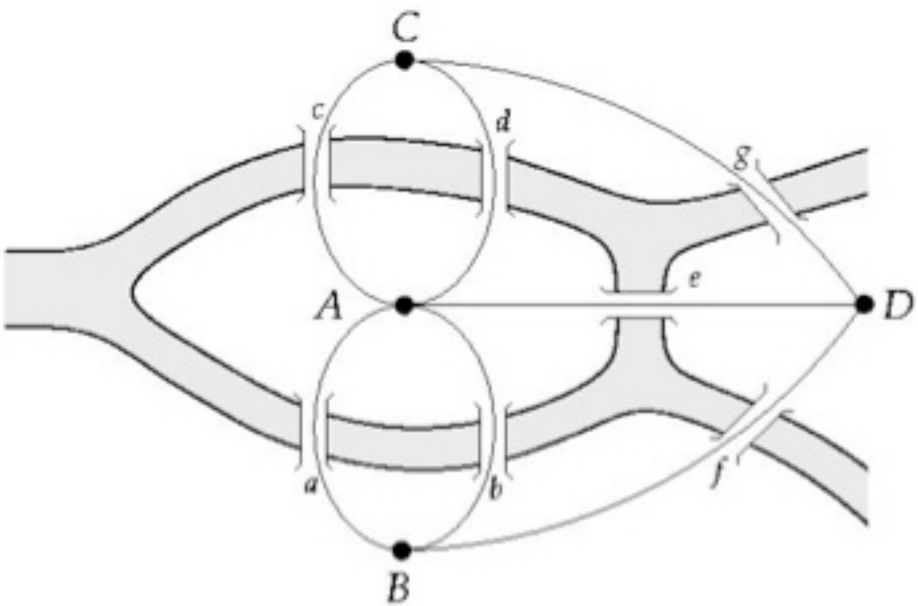
GRAPH THEORY

The bridges of Konisberg



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

The bridges of Konisberg

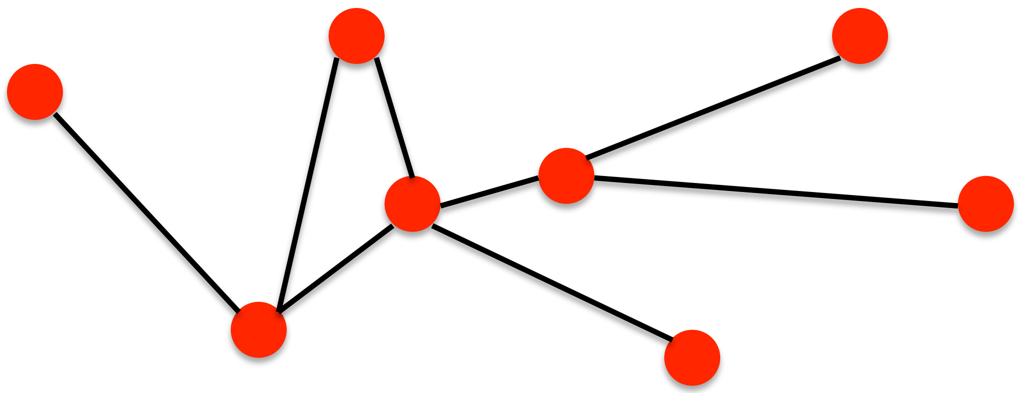


Can one walk across the seven bridges and never cross the same bridge twice?

- Solved by Leonard Euler in 1736
- Euler's Theorem: eulerian paths
  - If a graph has more than two nodes of odd degree, there in no path
  - If a graph is connected and has no odd degree nodes, it has at least one path

NETWORKS AND GRAPHS

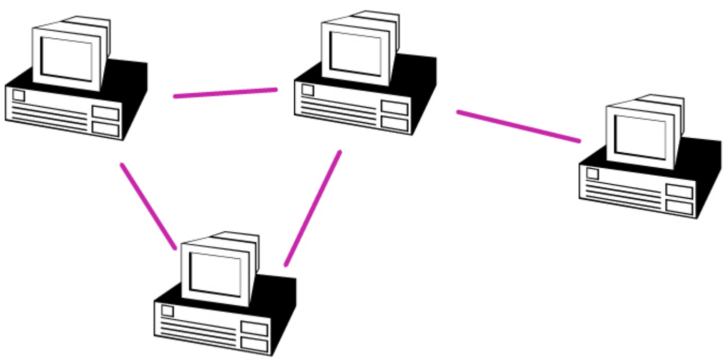
Basic Elements

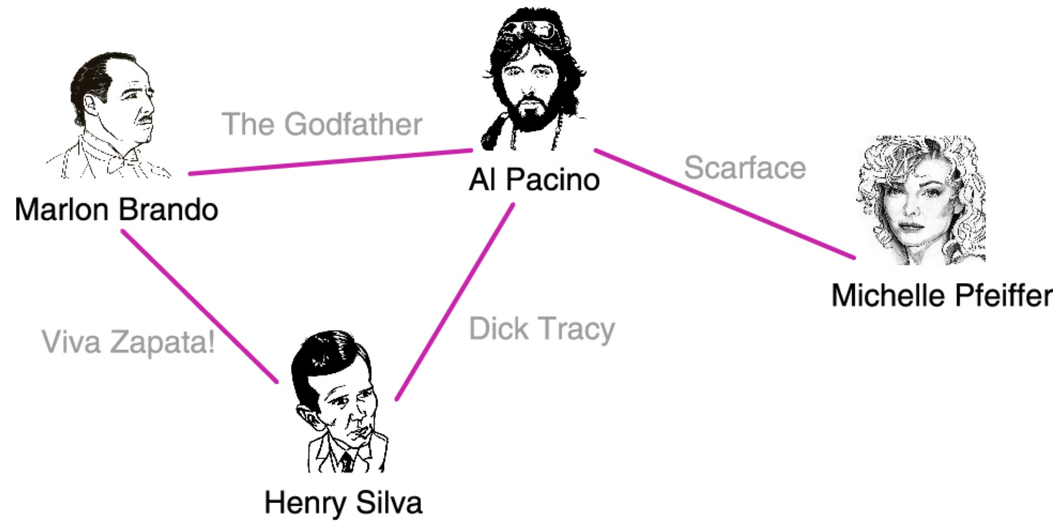


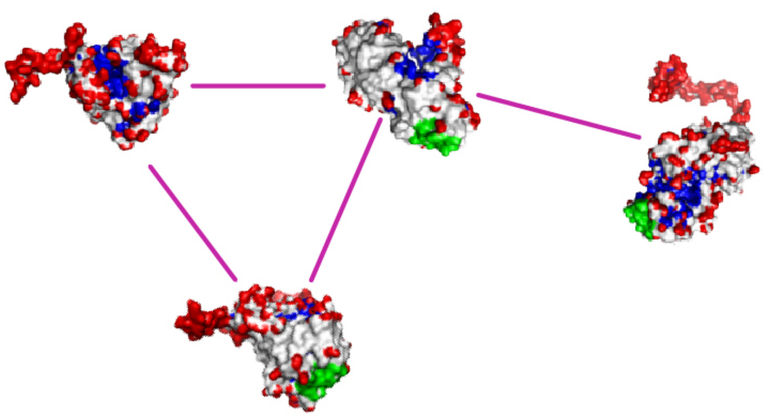
- **components**: nodes, vertices N
- **interactions**: links, edges L
- **system**: network, graph (N,L)

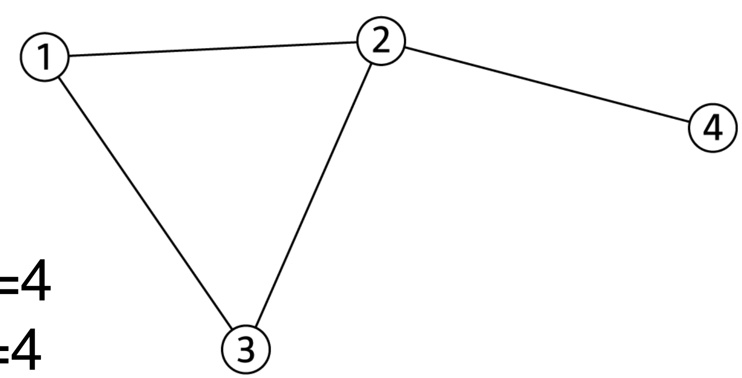
Networks or graphs? A common language

Network refer to a real system
Graph refers to mathematical representation of a network









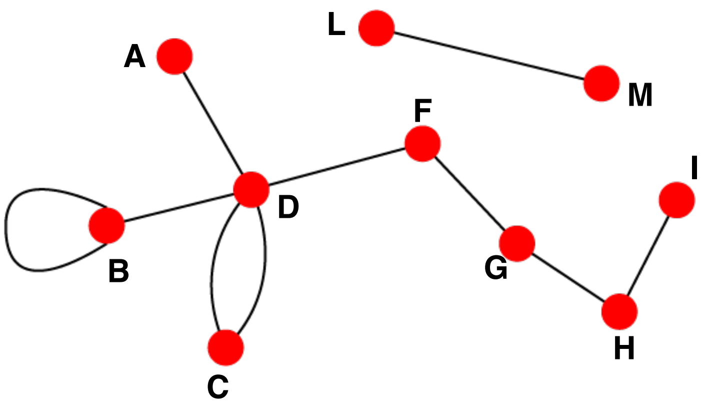
N=4
L=4

Undirected vs Directed Networks

Undirected

Links: undirected (*symmetrical*)

Graph:

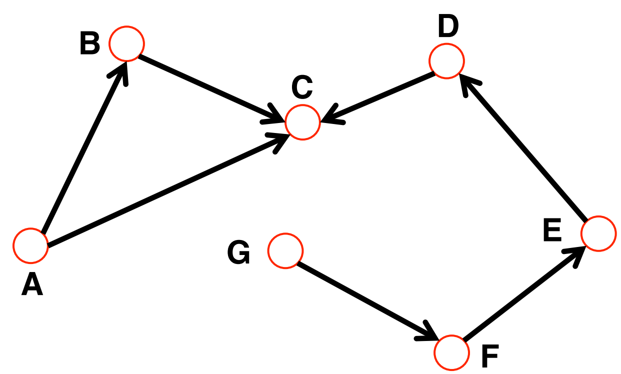


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

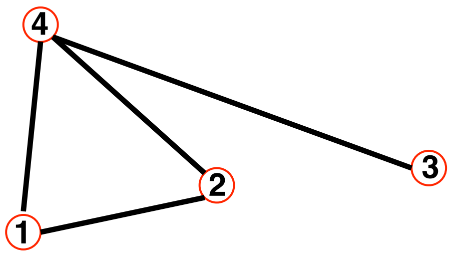
Digraph = directed graph:

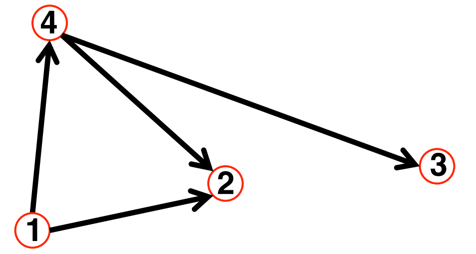


An undirected link is the superposition of two opposite directed links.

Directed links :
URLs on the www
phone calls
metabolic reactions

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.

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

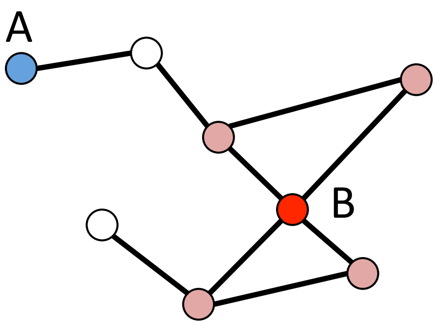
Note that for a directed graph (right) the matrix is not symmetric.

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

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

Node degree

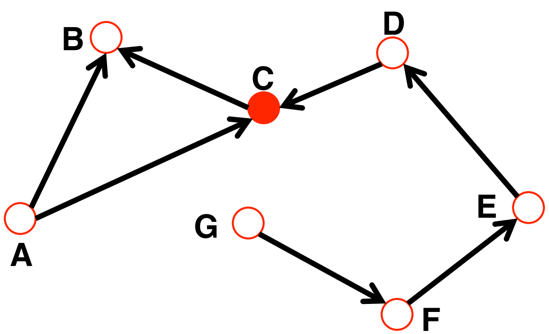
Undirected



Node degree: the number of links connected to the node.

$$k_A = 1 \qquad k_B = 4$$

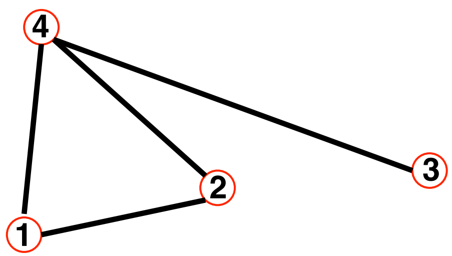
Directed



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

Adjacency Matrix and Node Degree

Undirected



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

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

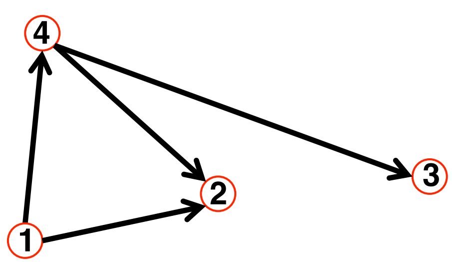
$$A_{ii} = 0$$

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

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

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

Directed



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

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

$$A_{ii} = 0$$

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

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

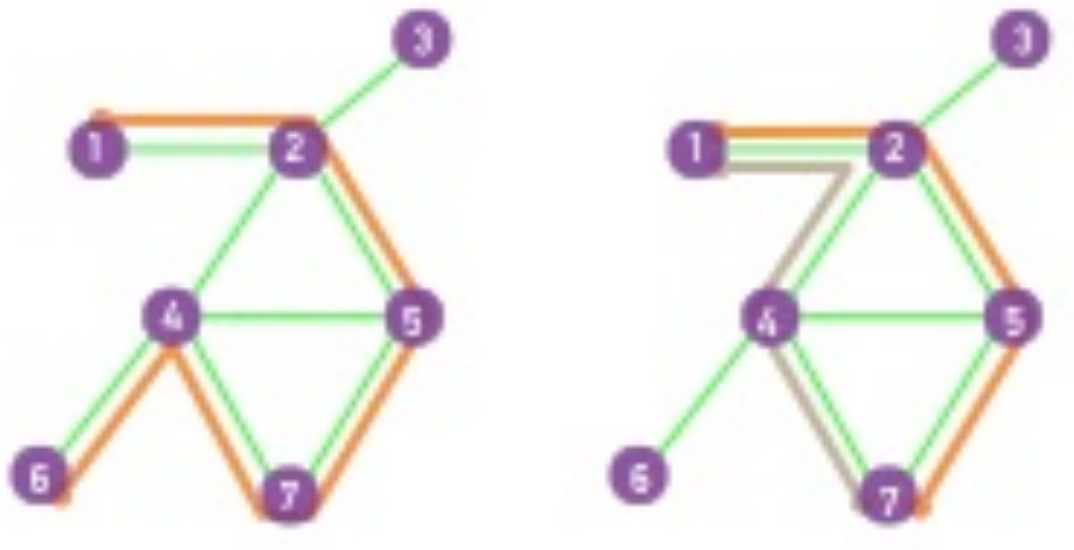
$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

Paths

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

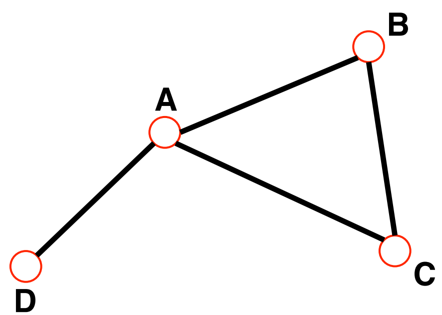
$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

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



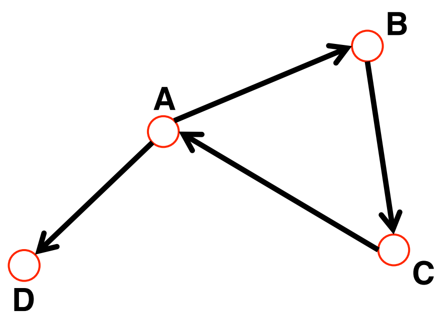
- In a directed network, the path can follow only the direction of an arrow.

Distance in a Graph



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.

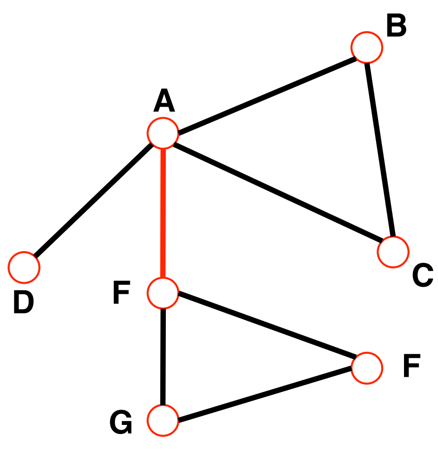


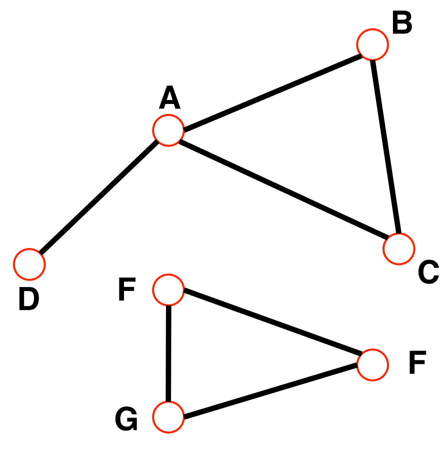
In directed graphs each path needs to follow the direction of the arrows.

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).

Connectedness

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

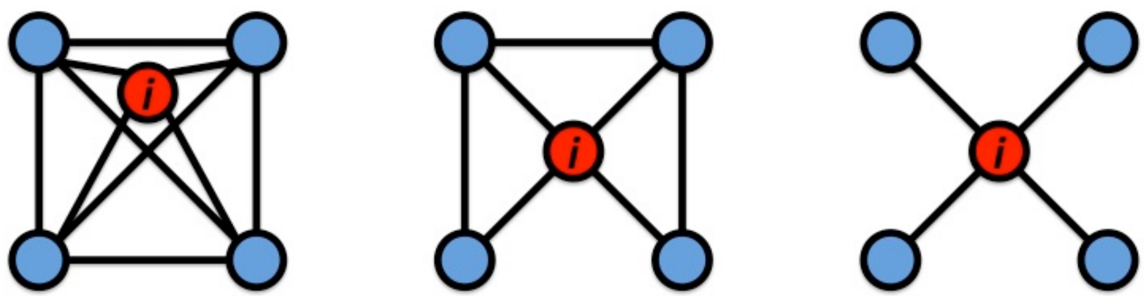
The rest: **Isolates**

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

Clustering Coefficient

- Clustering coefficient:
  - what fraction of your neighbors are connected?
- Node i with degree k_i
- C_i in [0,1]

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



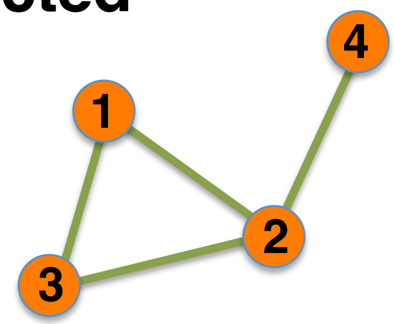
$C_i = 1$ $\quad$ $C_i = 1/2$ $\quad$ $C_i = 0$

Watts & Strogatz, Nature 1998.

SUMMARY

Undirected vs Directed

Undirected



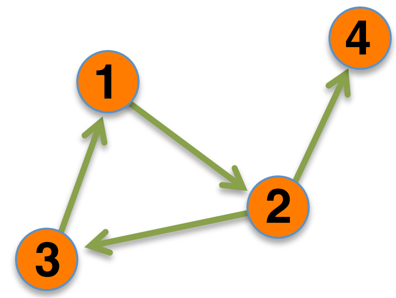
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions

Directed



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

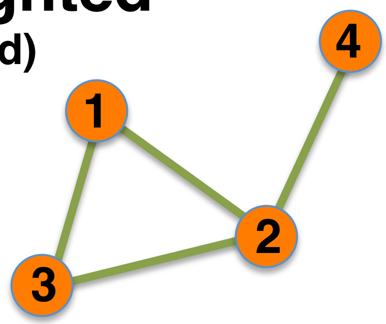
$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

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

WWW, citation networks

Unweighted vs Weighted

Unweighted
(undirected)



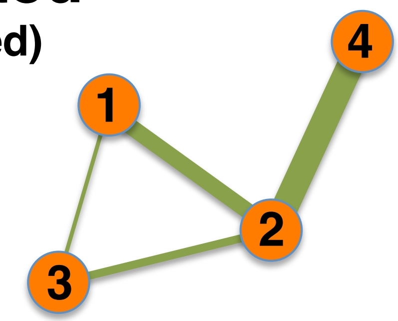
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

protein-protein interactions, www

Weighted
(undirected)



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

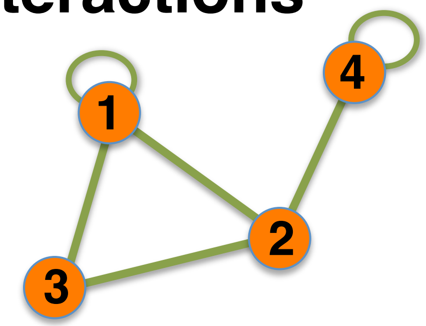
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

Call Graph, metabolic networks

Link properties

Self-interactions



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

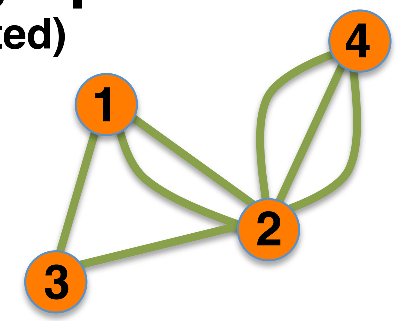
$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

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

Protein interaction network, www

Multigraph

(undirected)



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

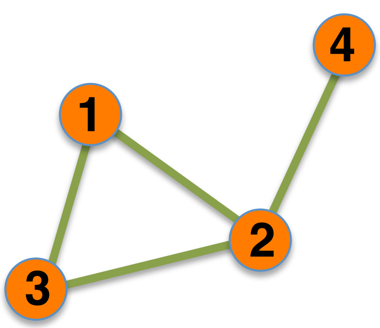
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

Social networks, collaboration networks

Network Visual Representation

- Three main methods
  a) Adjacency Lists
  b) Matrices
  c) Node-link diagrams



1: 2,3
2: 3,1,4
3: 1,2
4: 2

a)

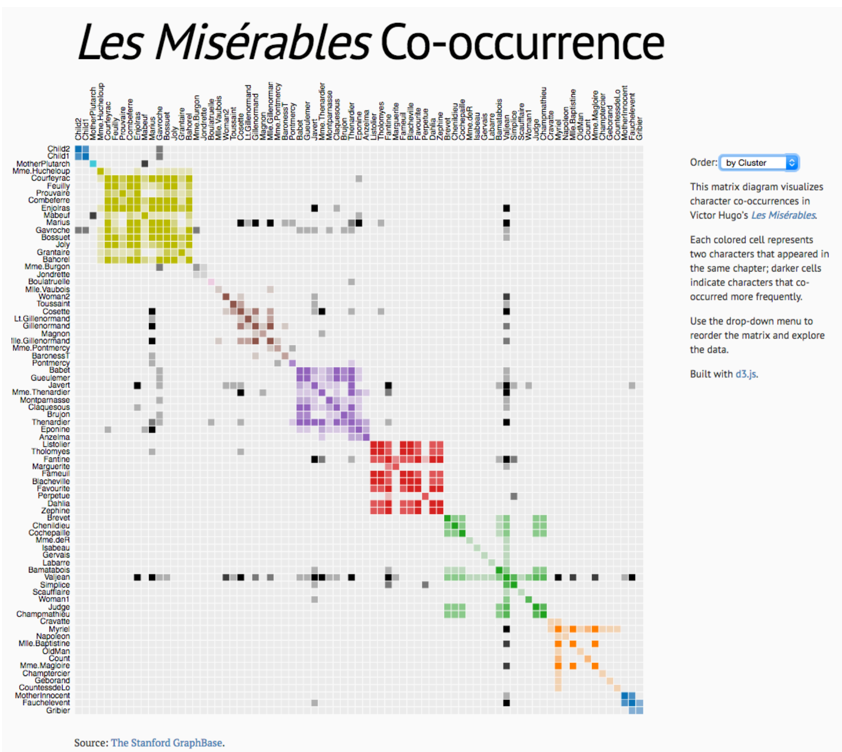
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

b)

1,2
1,3
2,3
2,4

c)

Adjacency Matrix

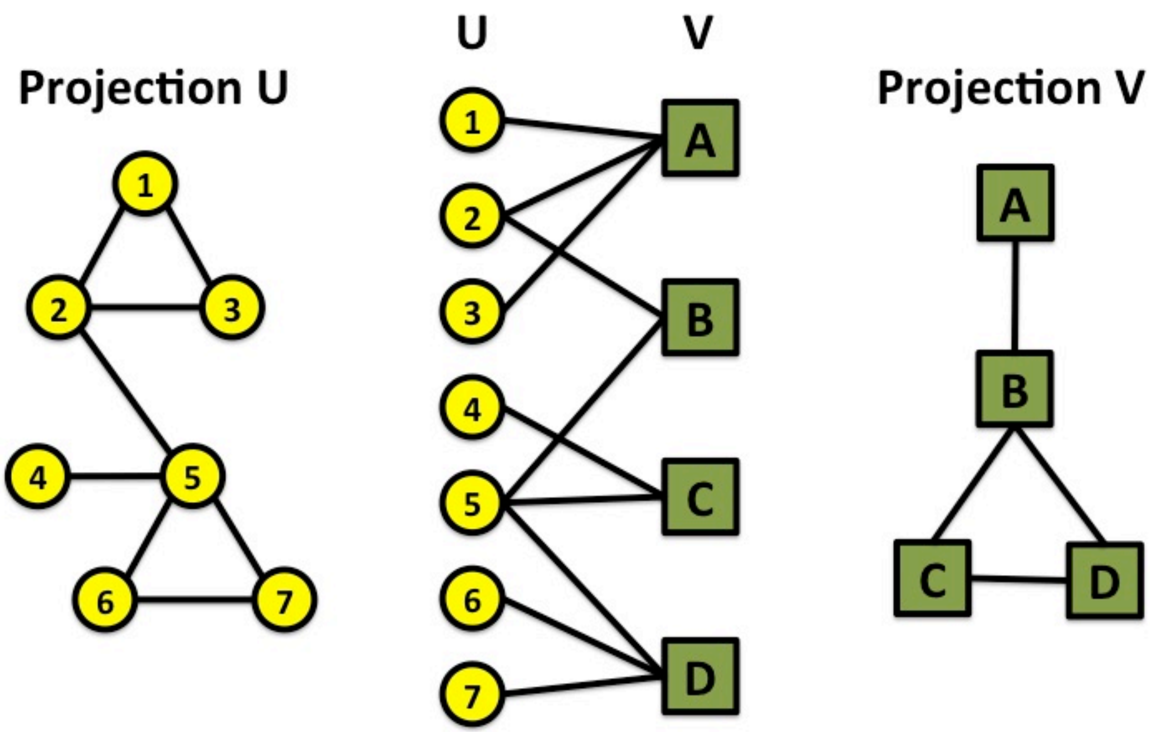


- Each cell ij represents an edge from vertex i to vertex j
- Effectiveness of visualization depens on rows/columns ordering
- First example by Jacques bertin (with paper strips rearranged by hand)
- Effective also for highly connected graphs

https://bost.ocks.org/mike/miserables/

Bipartite Graphs

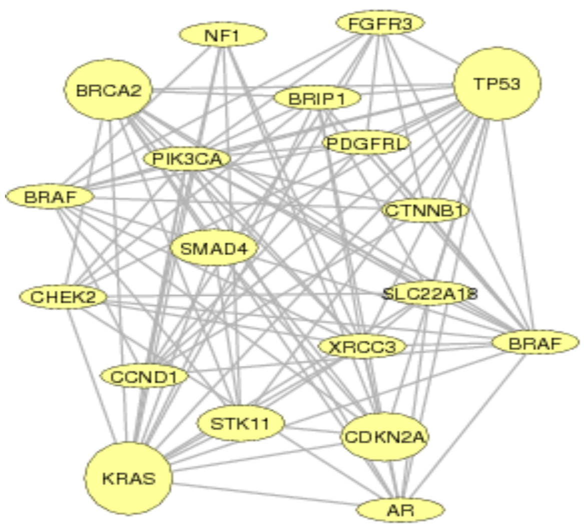
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



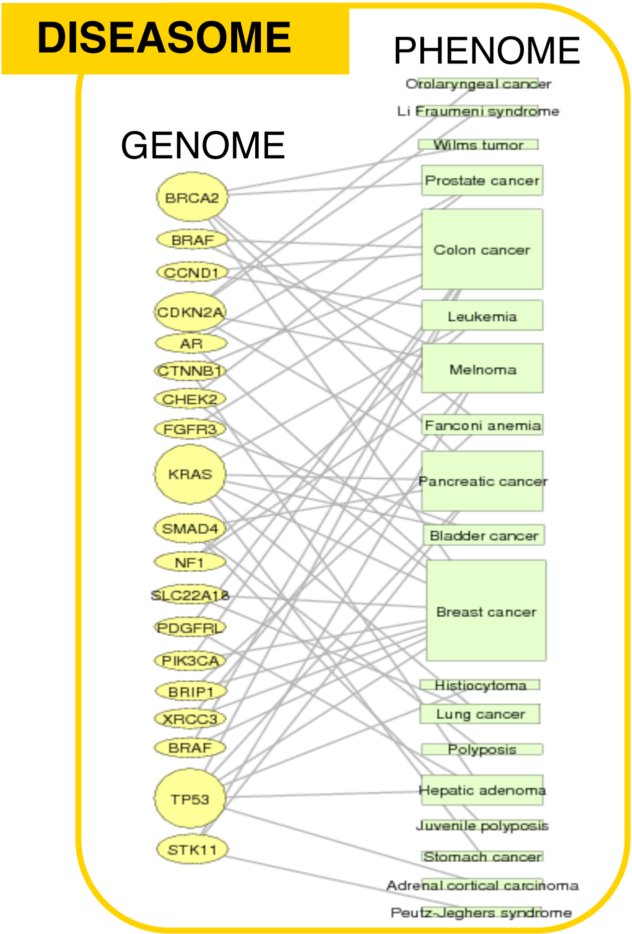
Examples:

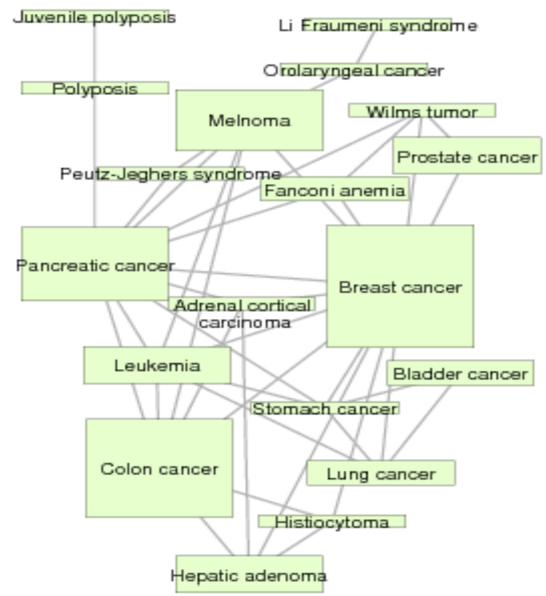
Hollywood actor network
Collaboration networks
Disease network (diseasome)

Gene Disease Network



Gene network





Disease network

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

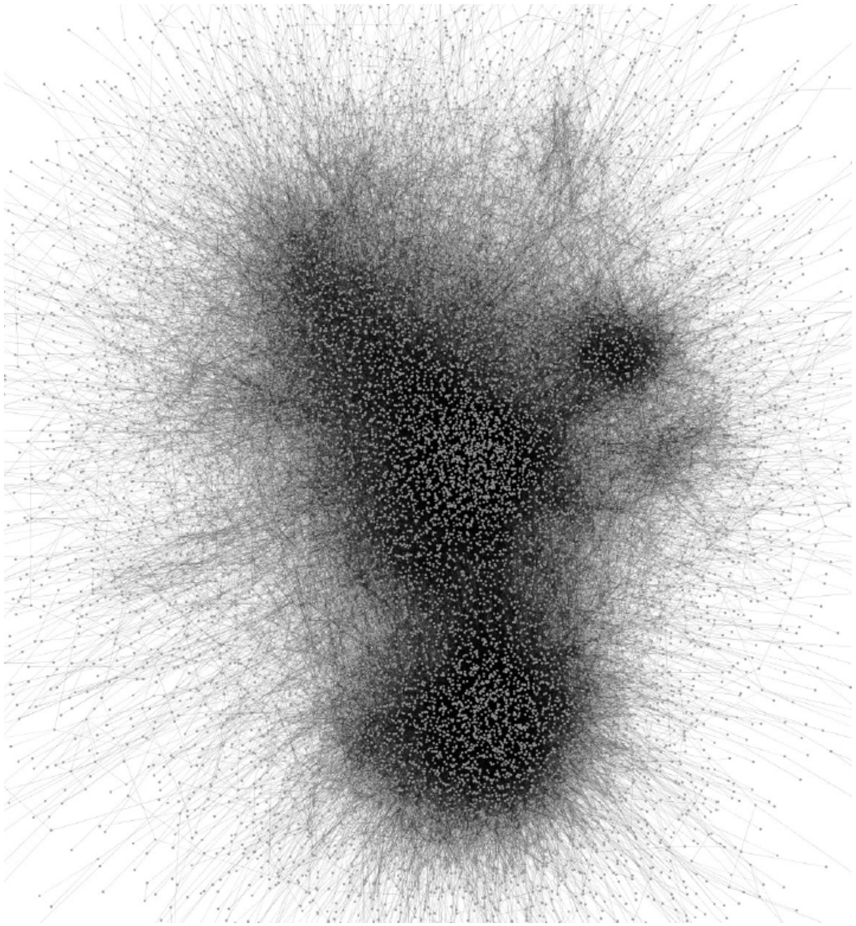
Node-link representation

http://bl.ocks.org/mbostock/4062045

- Symbolic elements for nodes
- Lines for connection among nodes
- Physical networks (roads, power grids) have a natural spatial encoding
- Abstract networks need layouts to infer a spatial position for nodes

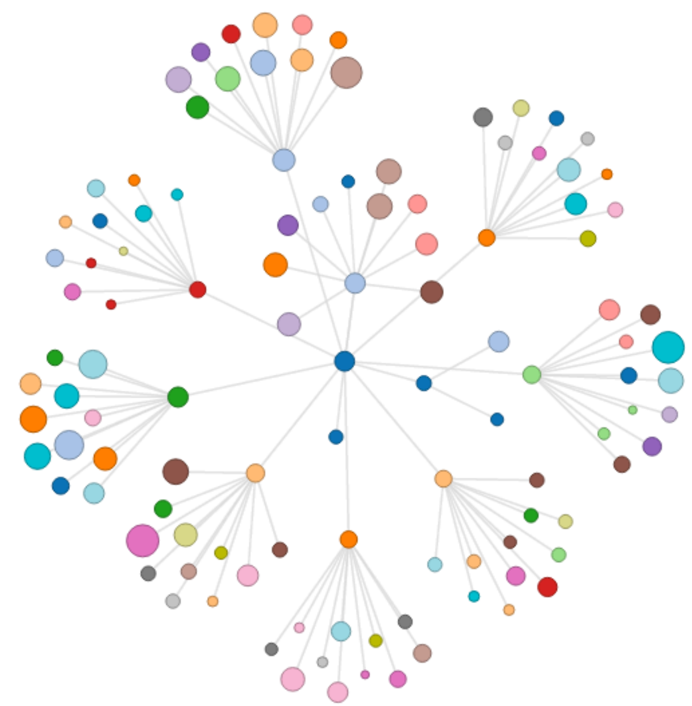
Problems of Node-Link diagrams

- Occlusion of node and link crossings
- Large networks may produce hairball like networks
- Many algorithms to produce effective layouts to reduce cluttering



Cluttering Reduction

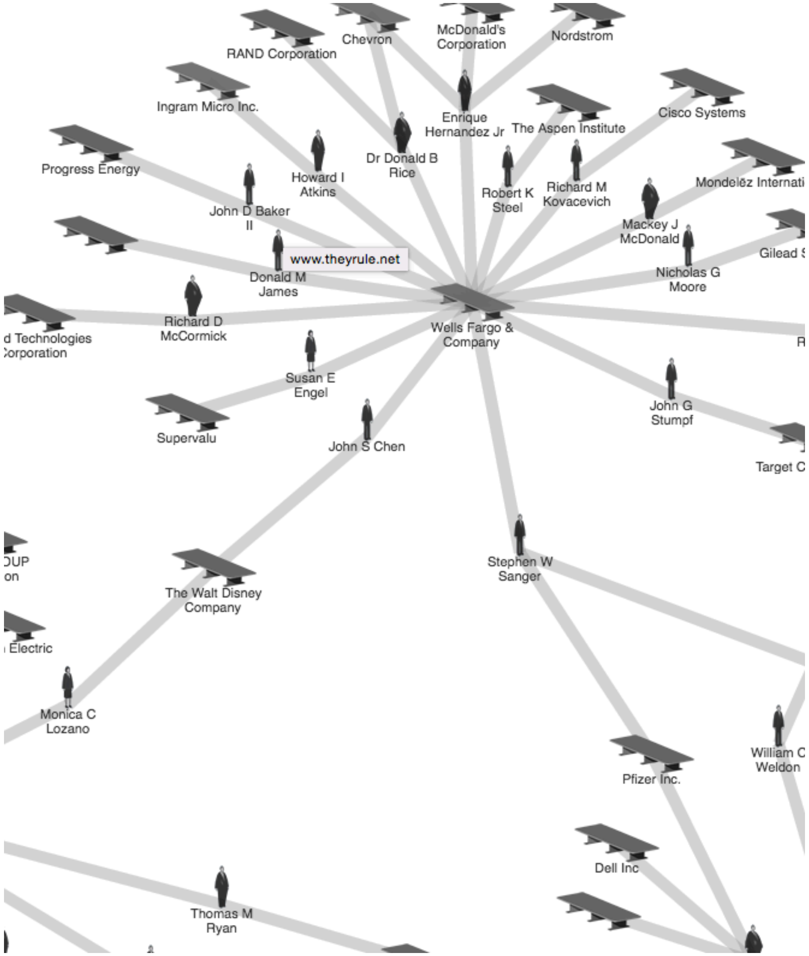
- Interaction to switch between different layouts
- Effective positioning of labels
  - Centered on nodes
  - Visualization based on interaction and mouse hover



http://projects.flowingdata.com/tut/interactive_network_demo/

Cluttering Reduction

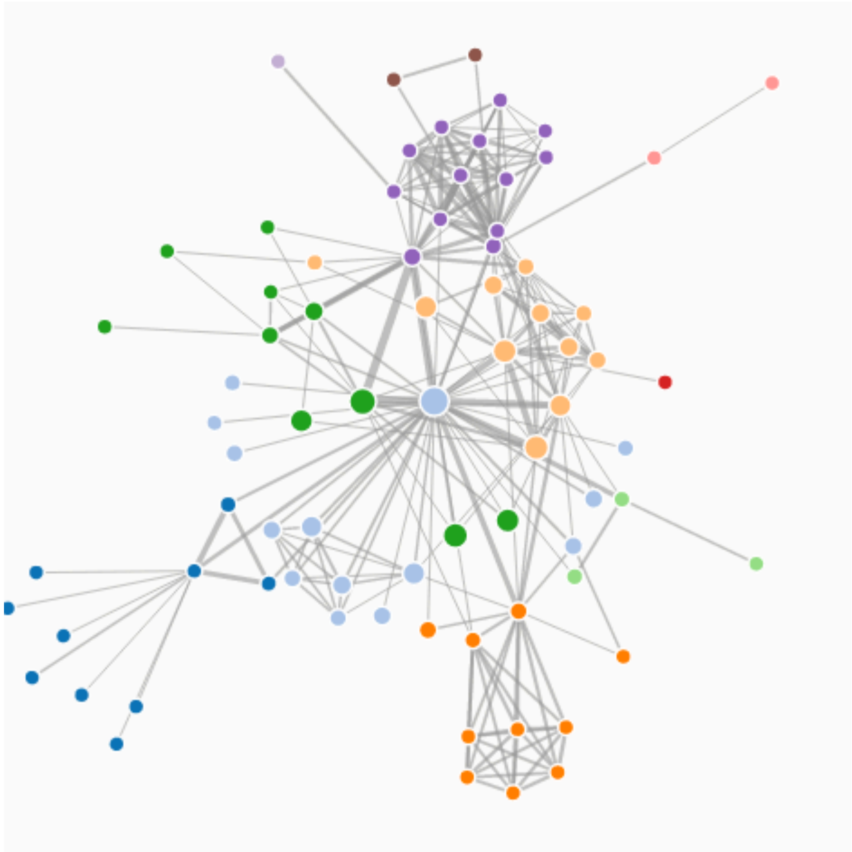
- Collapsing nodes into clusters



http://www.theyrule.net/

Cluttering Reduction

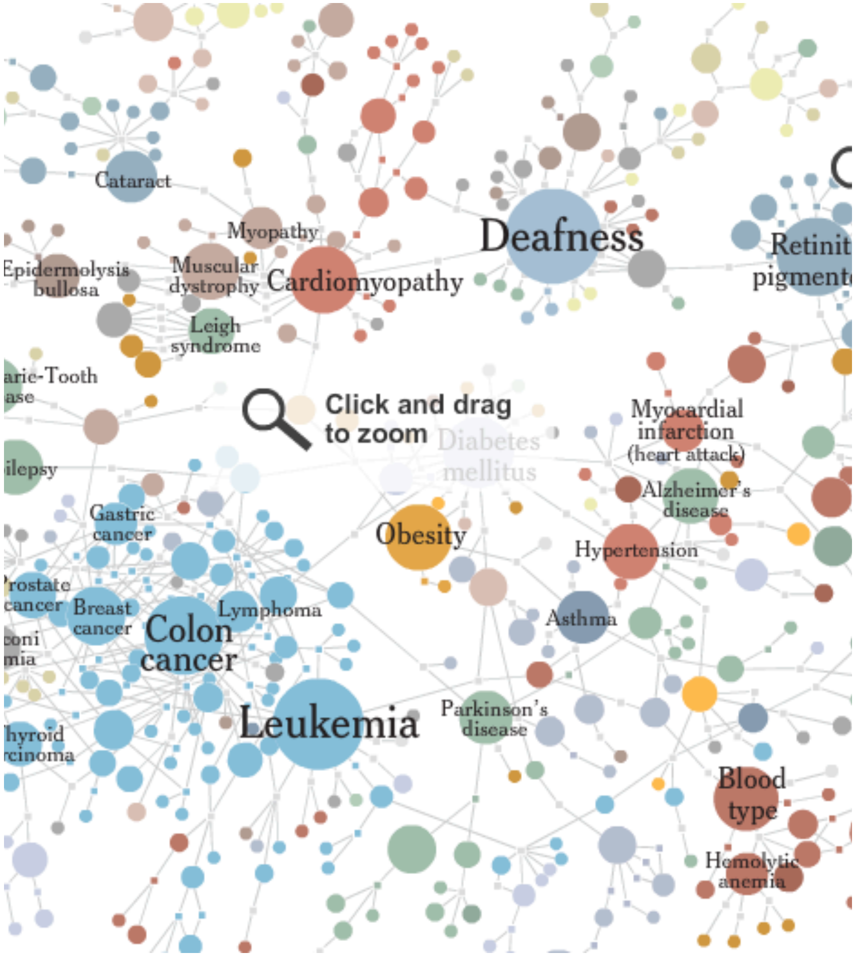
- Zooming and context distortion



https://bost.ocks.org/mike/fisheye/

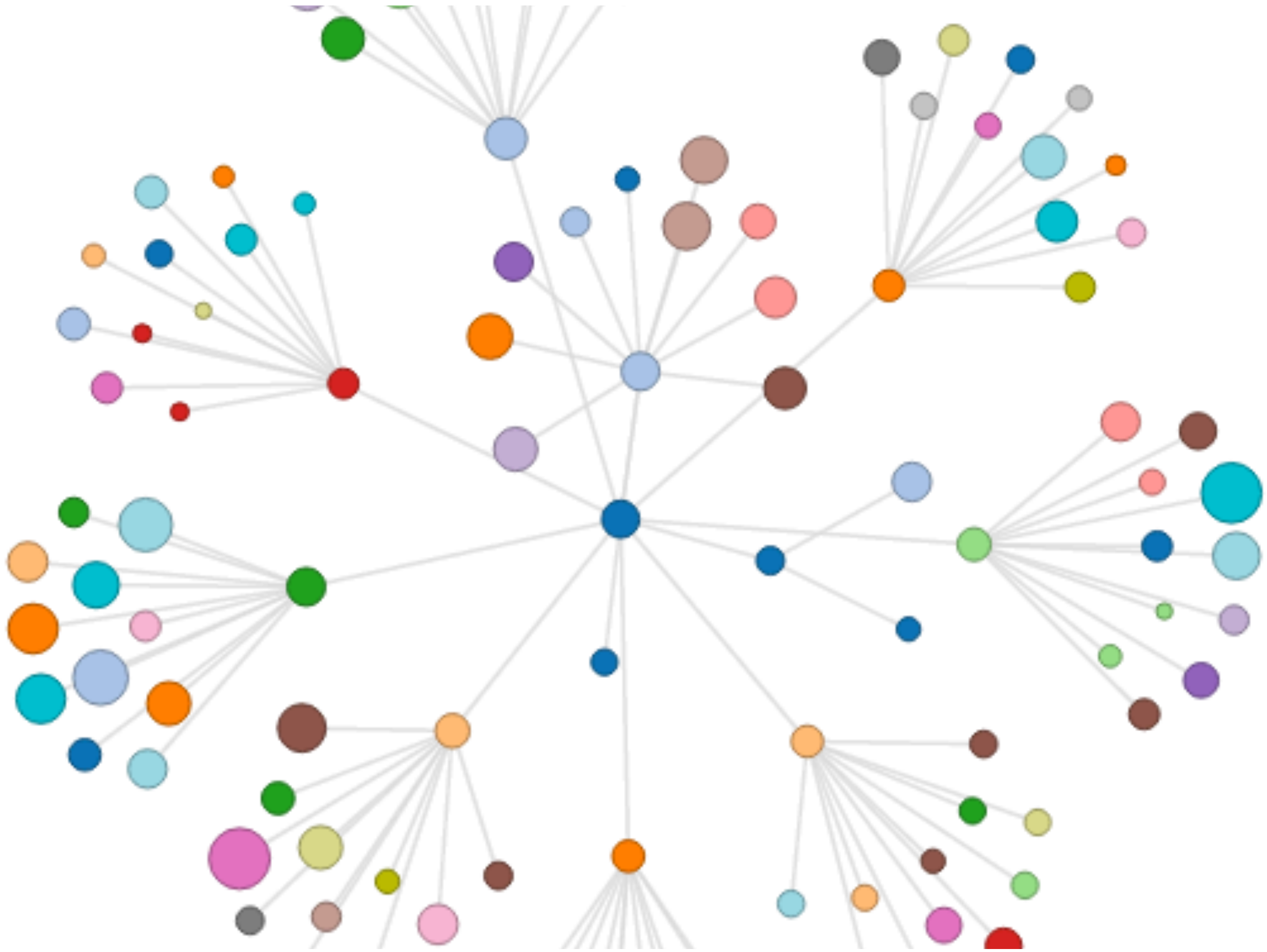
Cluttering Reduction

- Zooming and context distortion



http://www.nytimes.com/interactive/2008/05/05/science/20080506_DISEASE.html?_r=0

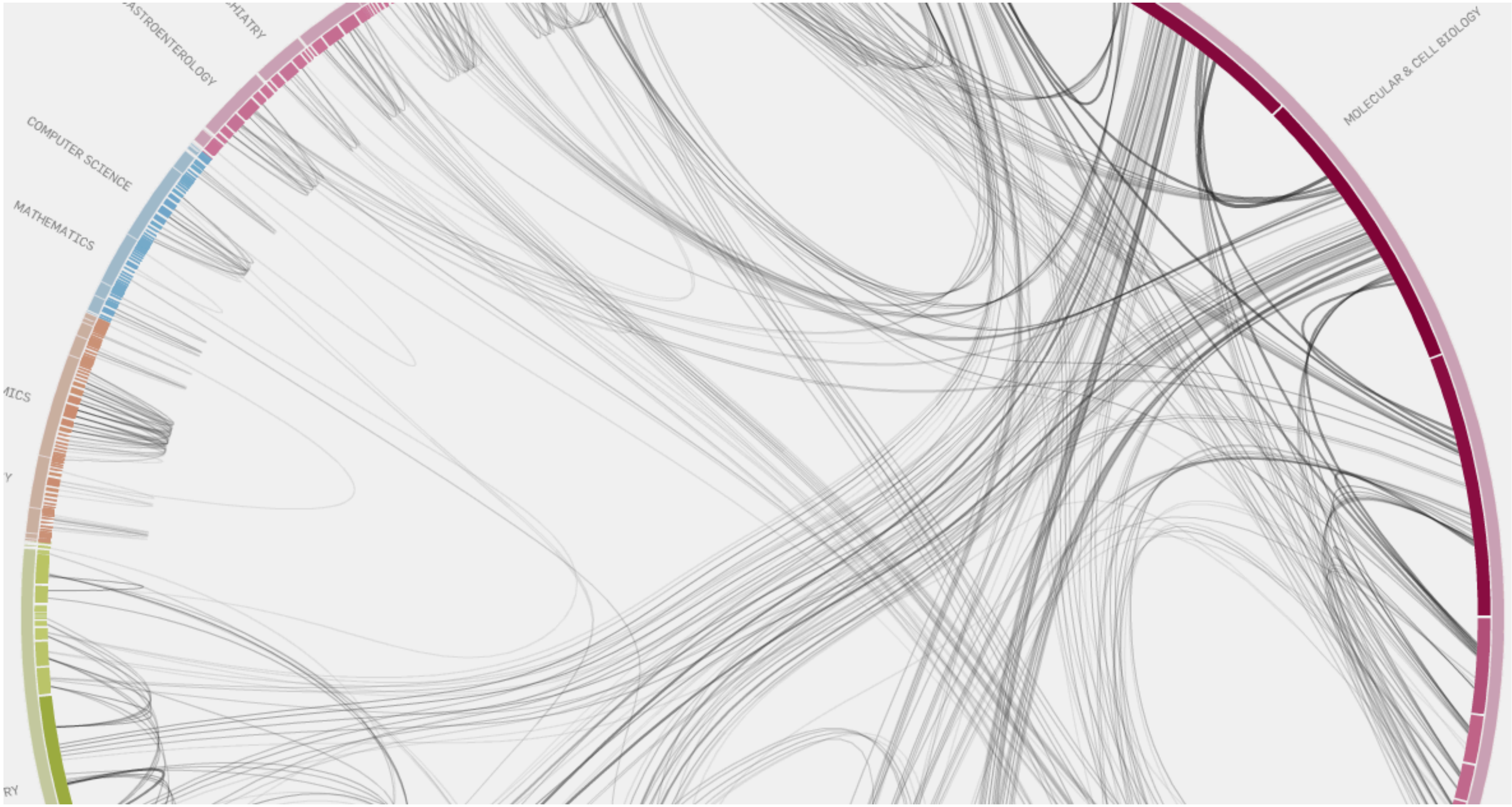
Case Study: Force Directed



http://projects.flowingdata.com/tut/interactive_network_demo/

Case Study: Information Flow

Circular Layout



http://www.eigenfactor.org/projects/well-formed/

Case Study: Sankey type diagrams