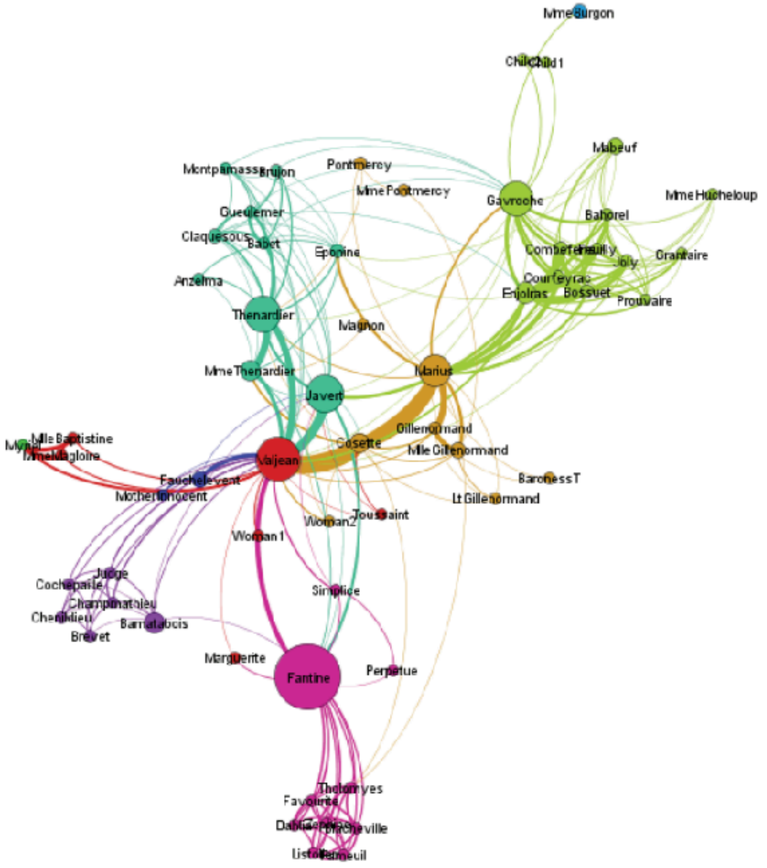
Class 25: Network analysis and visualization

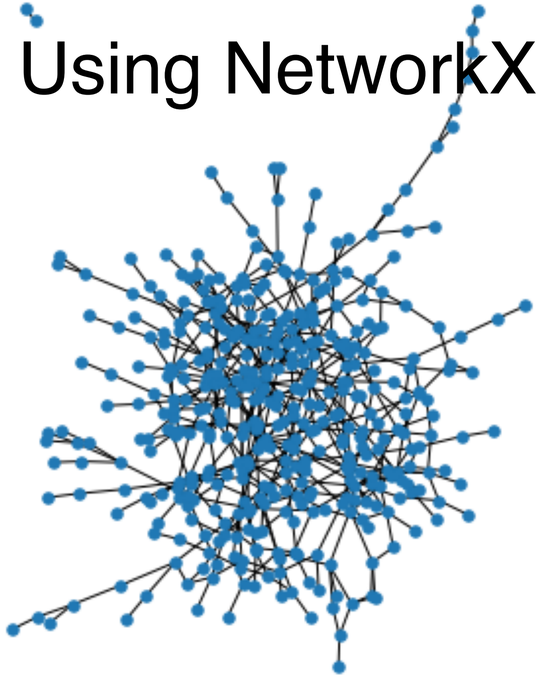
Instructor: Michael Szell
Nov 27, 2019



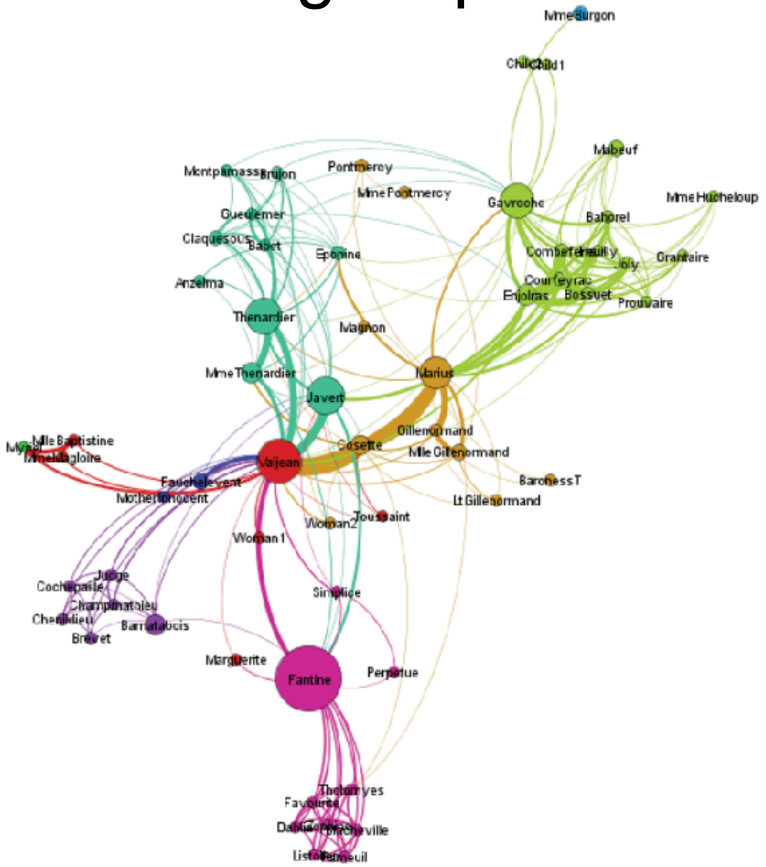
Today you will learn about analyzing and visualizing networks

Centrality and community: Who is important?

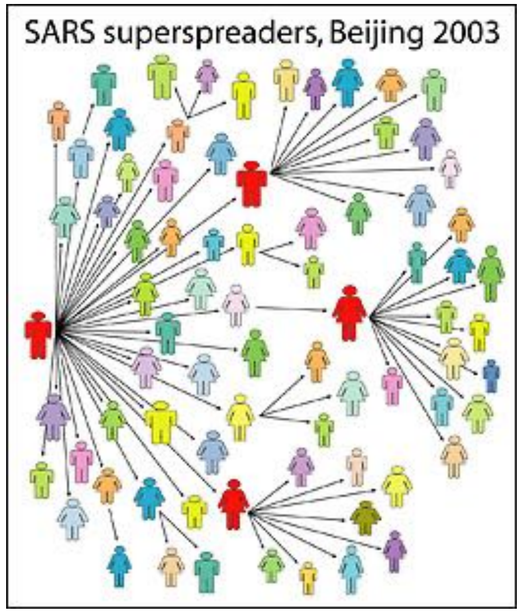
Using NetworkX



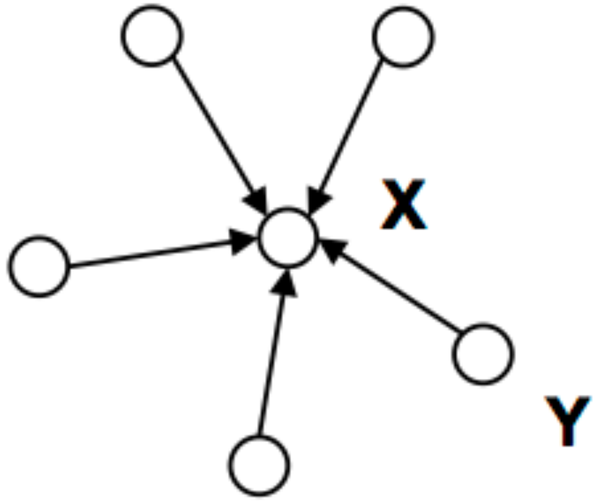
Using Gephi



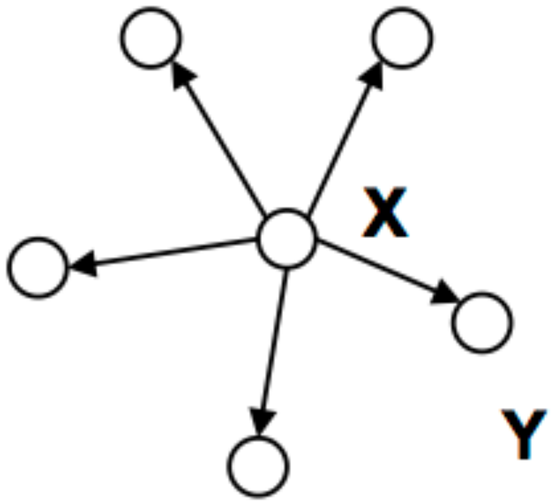
Centrality is a measure of a node's importance in the network



Degree can be a measure of importance

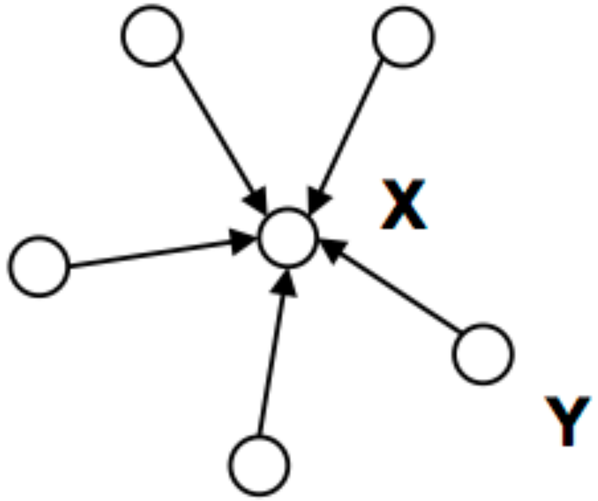


indegree



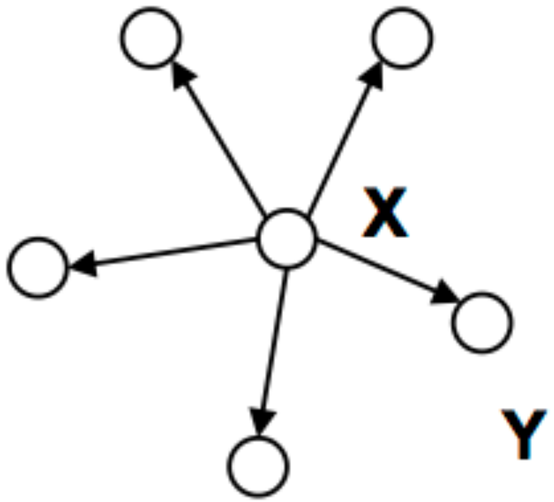
outdegree

Degree can be a measure of importance



indegree

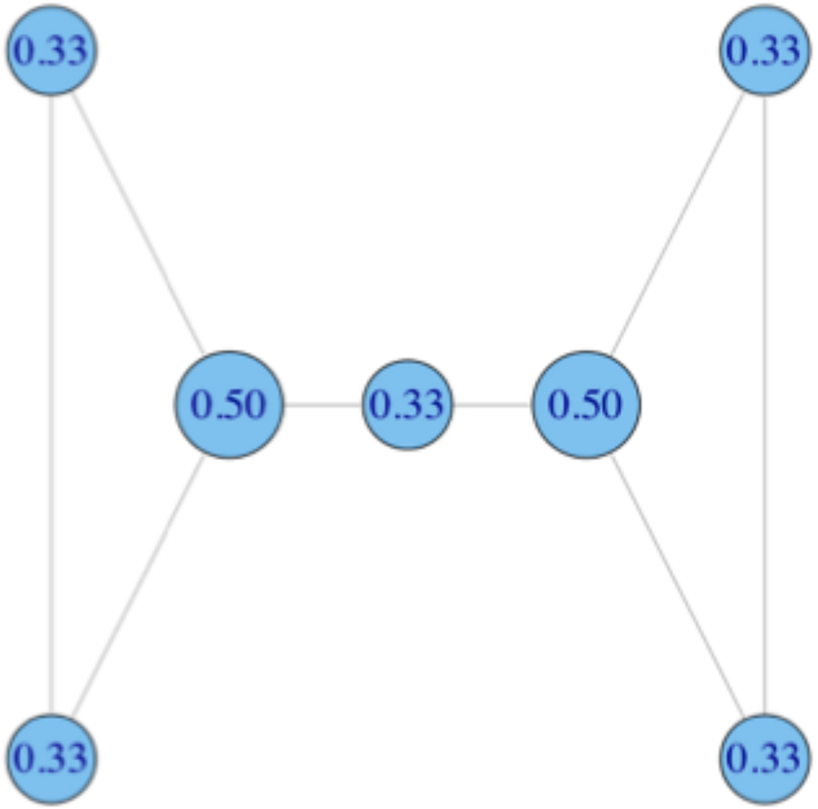
How popular you are

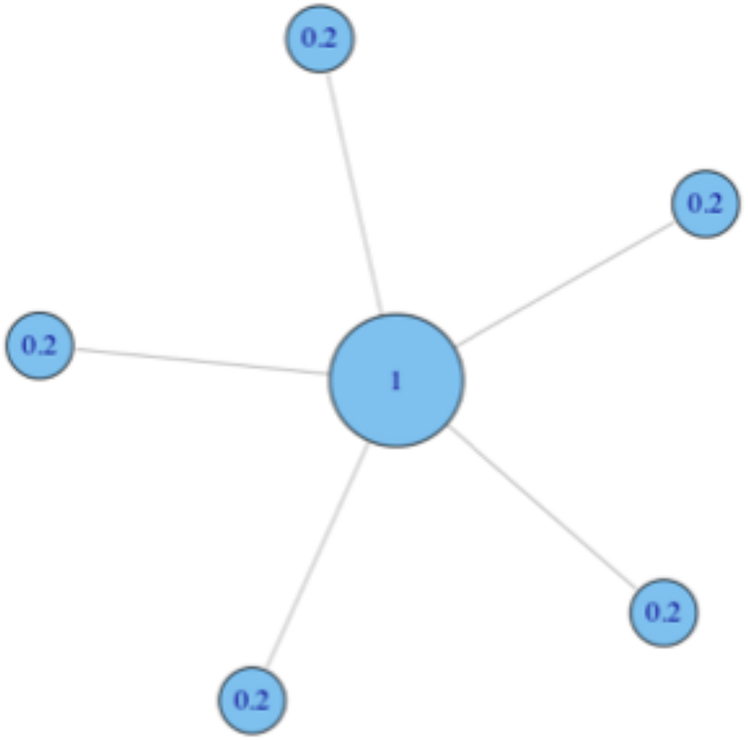


outdegree

How many people you know

Degree centrality measures the degree (normalized)

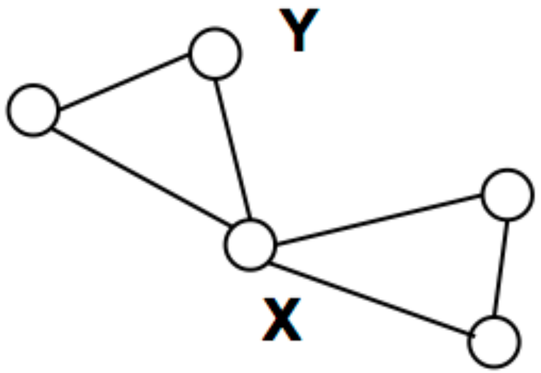


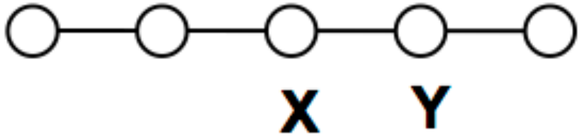


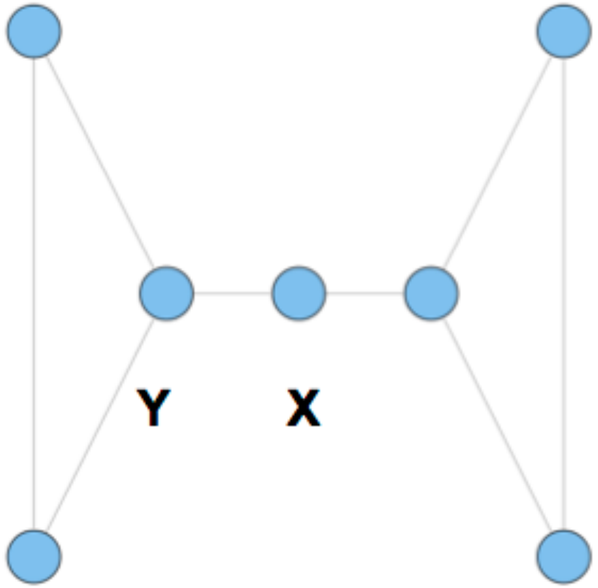
$$C_D(i) = \frac{k_i}{N-1}$$

Degree is not everything

Who is more important: X or Y?

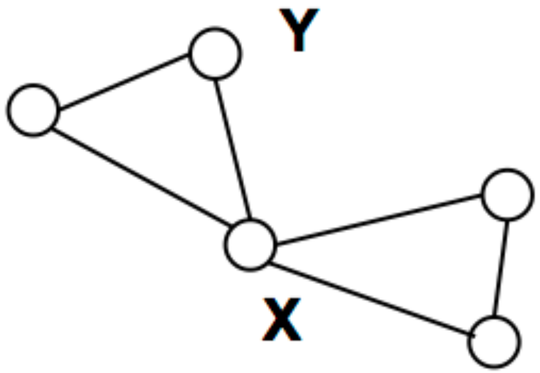


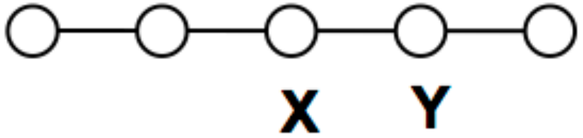


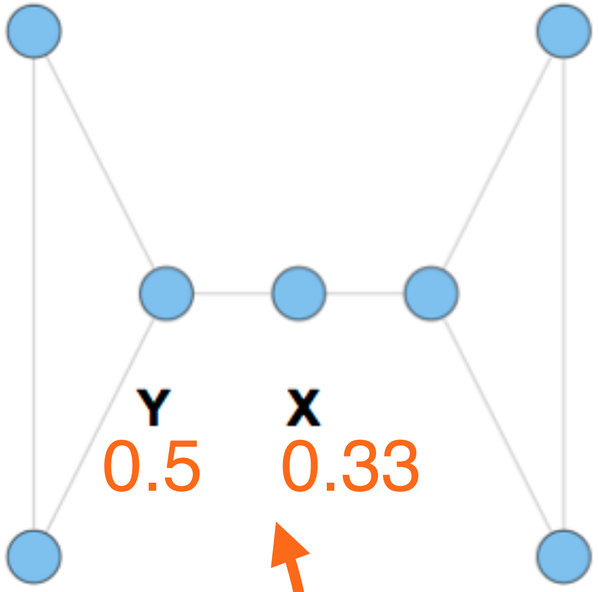


Degree is not everything

Who is more important: X or Y?







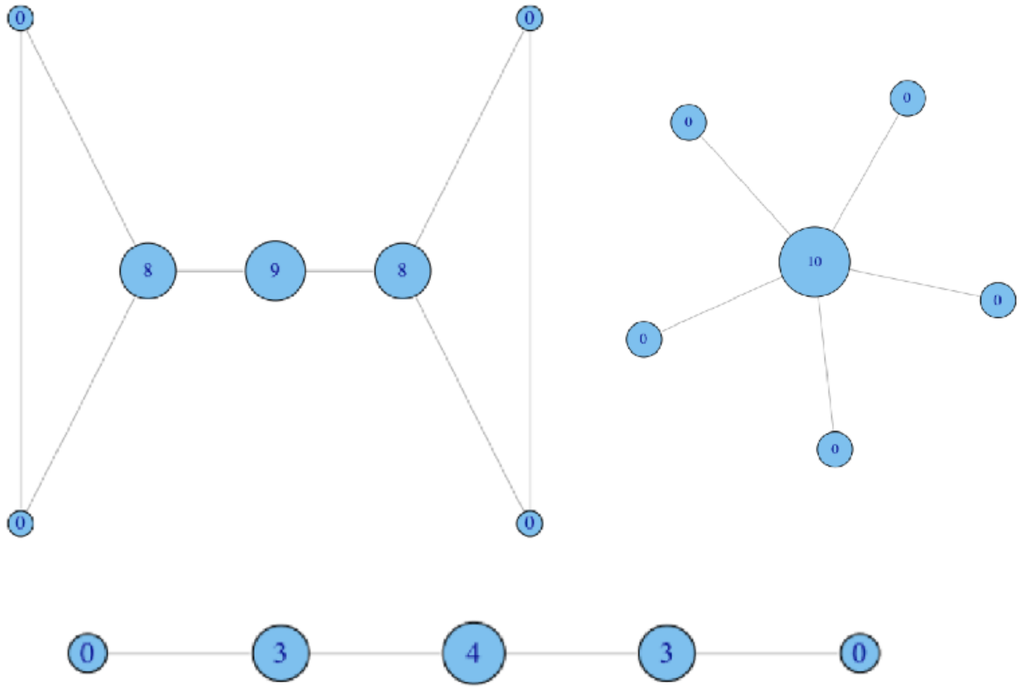
If we care about the ability to broker between groups, or to catch information originating anywhere, degree centrality can fail

Betweenness centrality captures being part of shortest paths

$$C_B(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

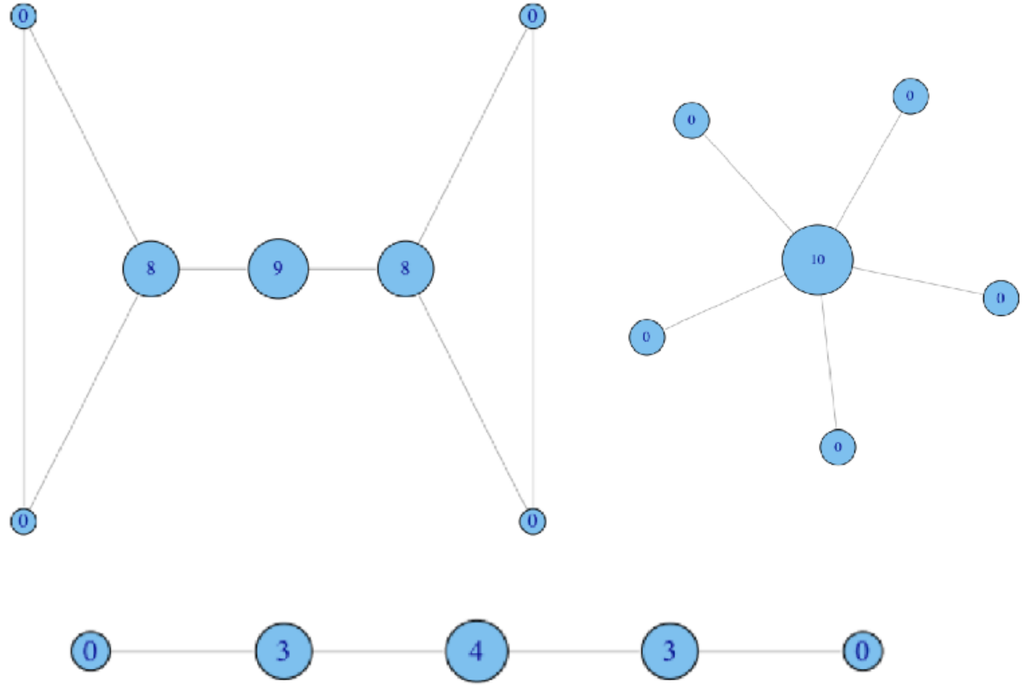
$\sigma_{jk}(i)$ ← Number of shortest paths between j and k that go through i

σ_{jk} ← Number of all shortest paths between j and k

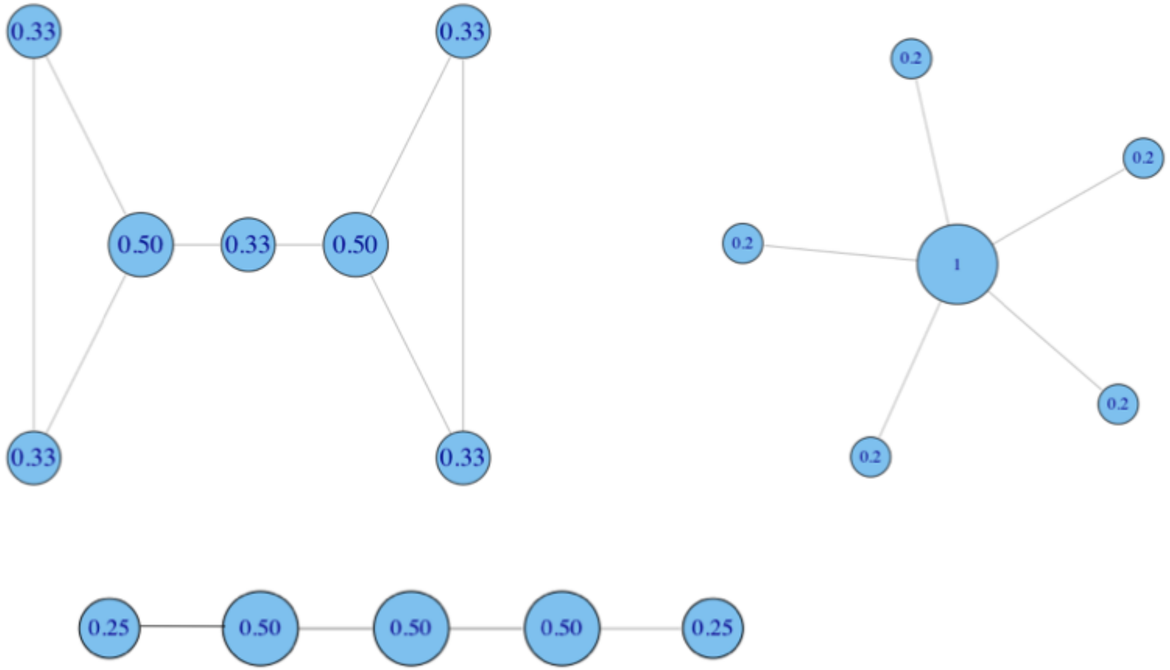


Betweenness centrality captures being part of shortest paths

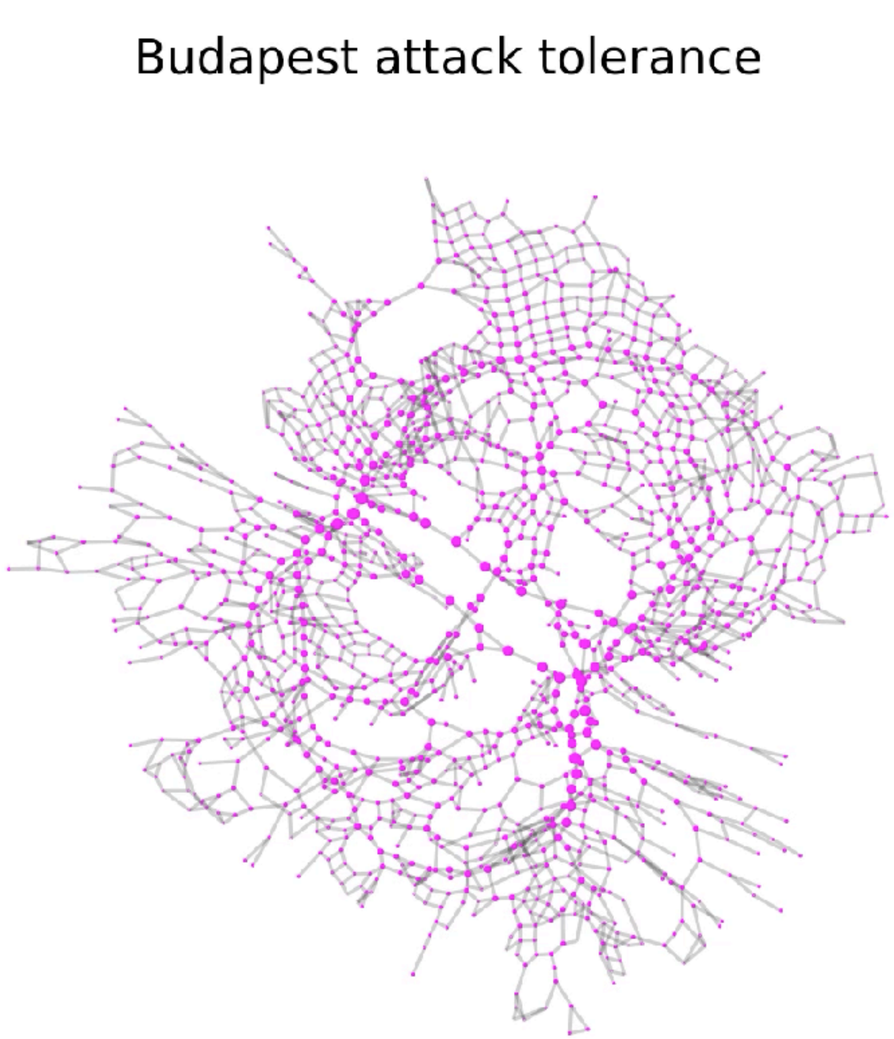
Betweenness centrality

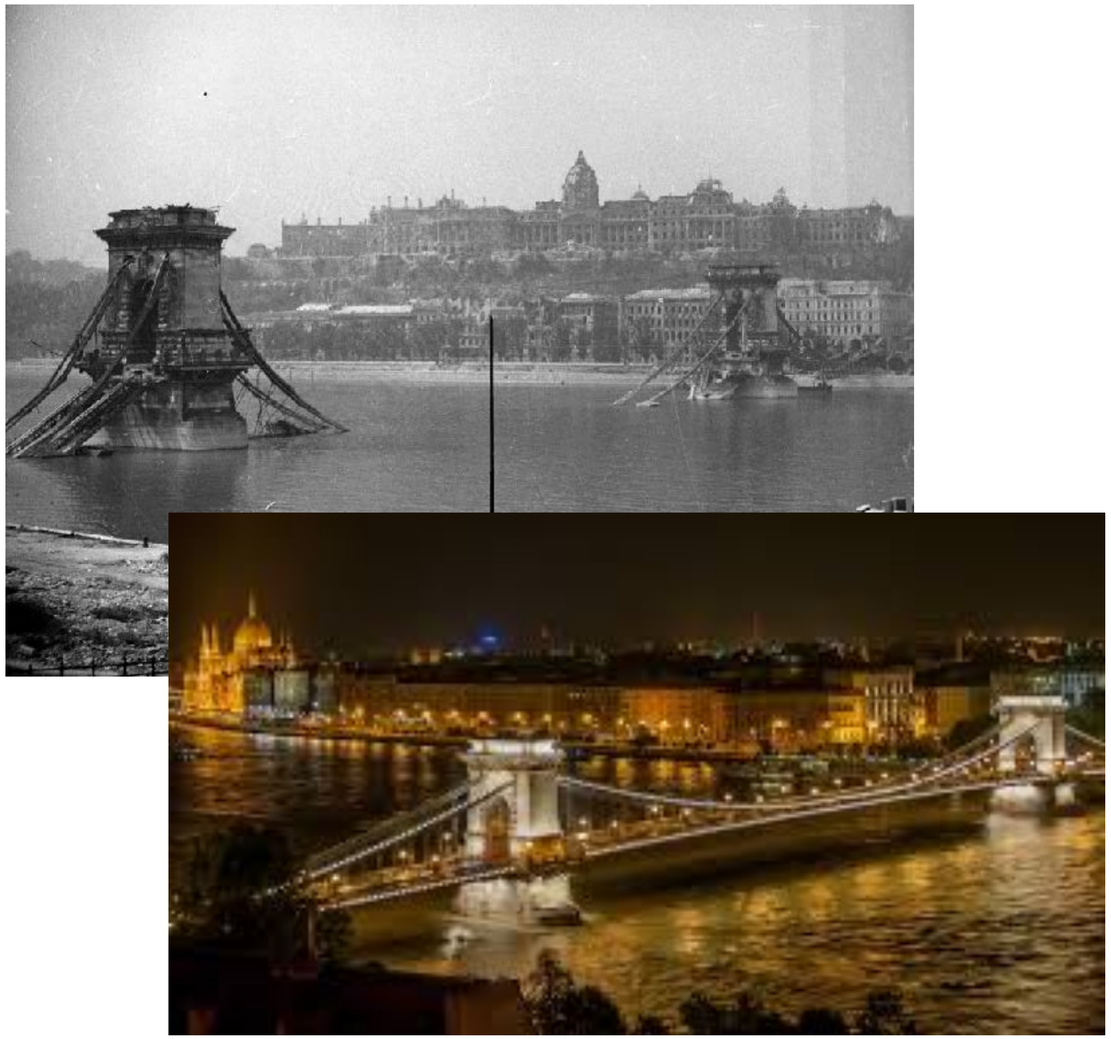


Degree centrality



Betweenness centrality identifies bottlenecks in transport networks

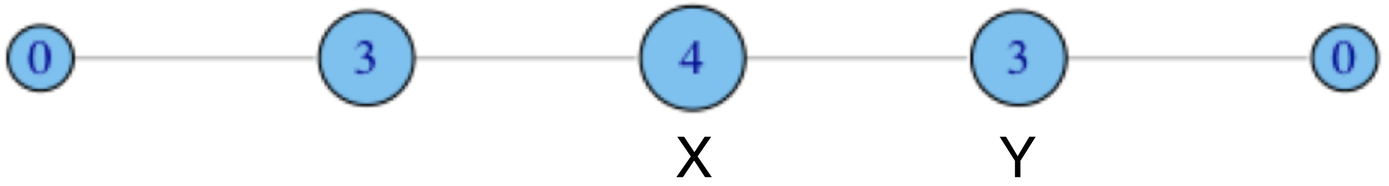




Video by Luis Guillermo Natera Orozco

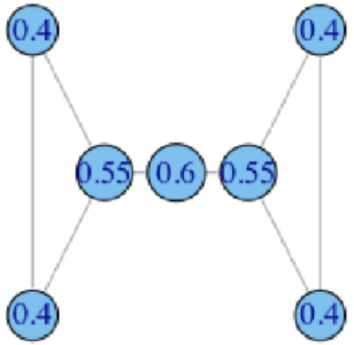
Betweenness is not everything

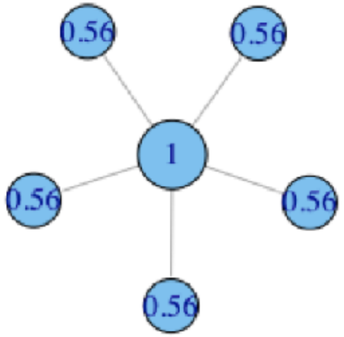
X and Y do not differ much for betweenness



but X is closer to other nodes than Y

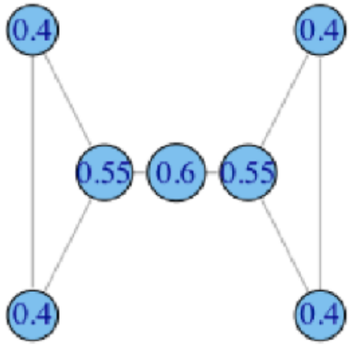
Closeness centrality captures being close to all other nodes

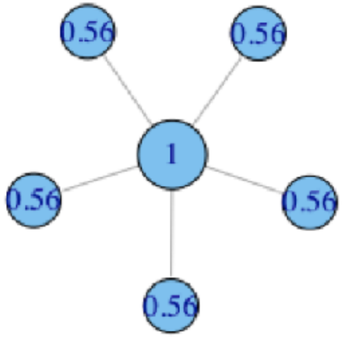




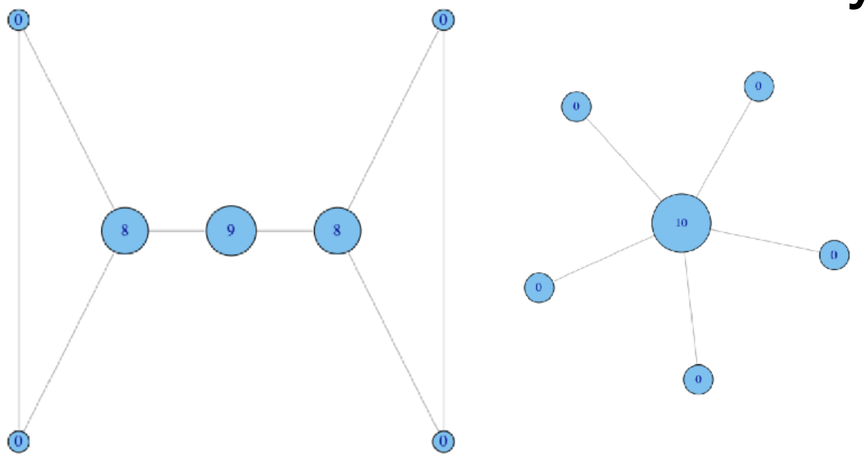
$$C_C(i) = \frac{N-1}{\sum_{j=1}^{N} d(i,j)}$$

Closeness centrality captures being close to all other nodes

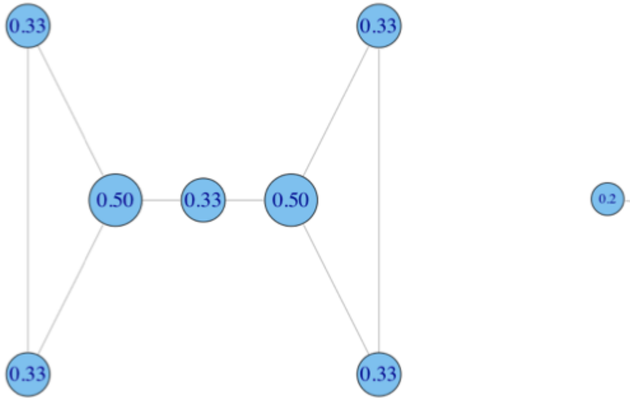


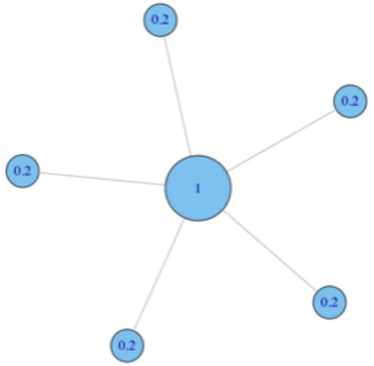


Betweenness centrality



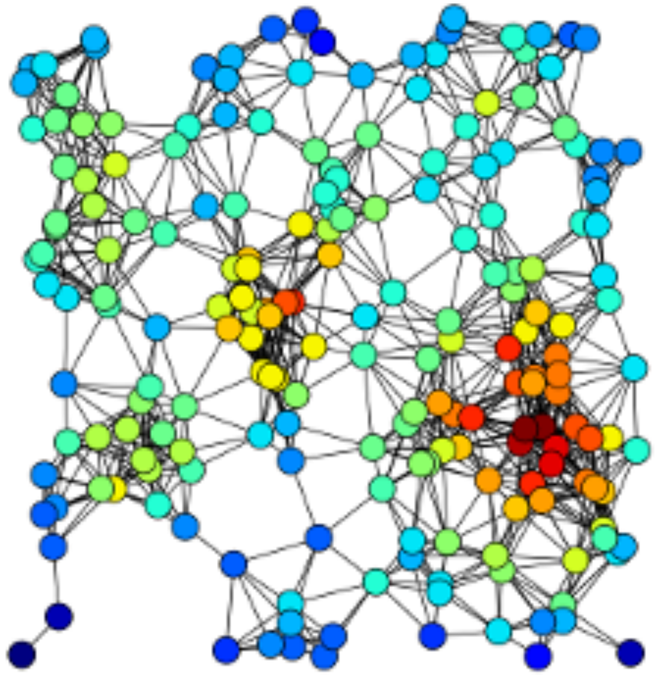
Degree centrality



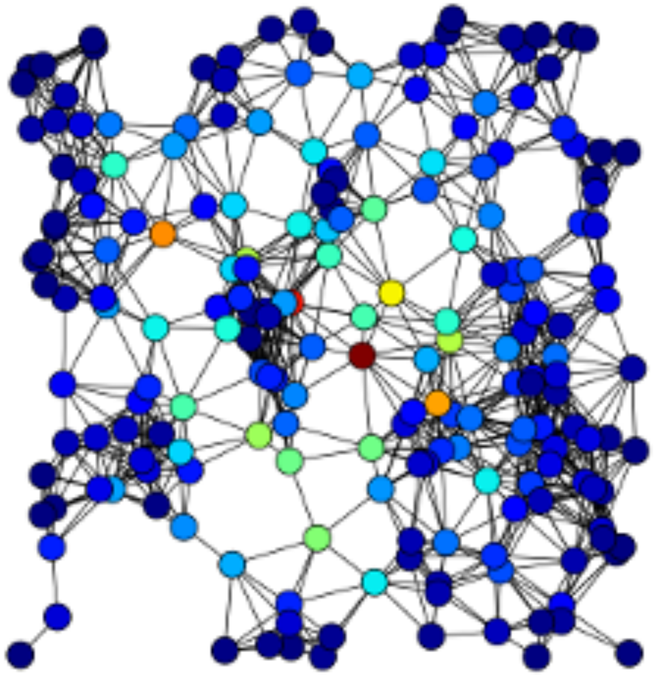


Which node is important? Depends on our definition of importance

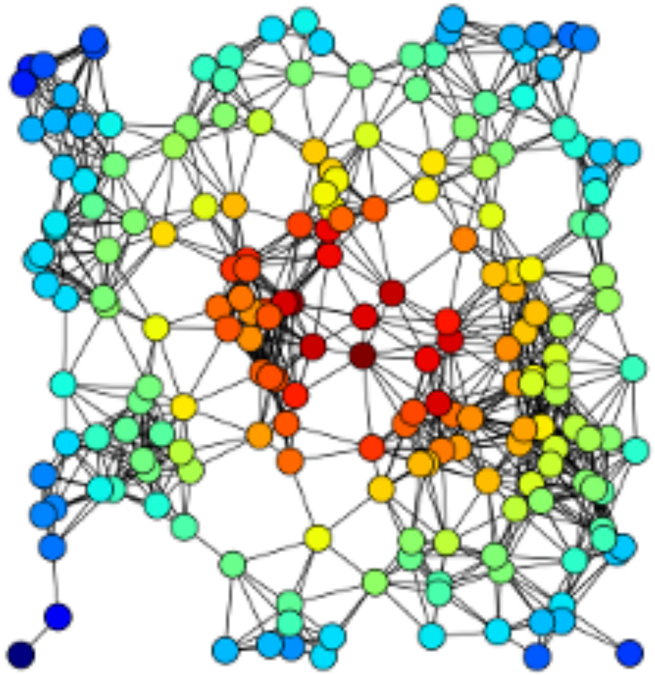
Degree



Betweenness

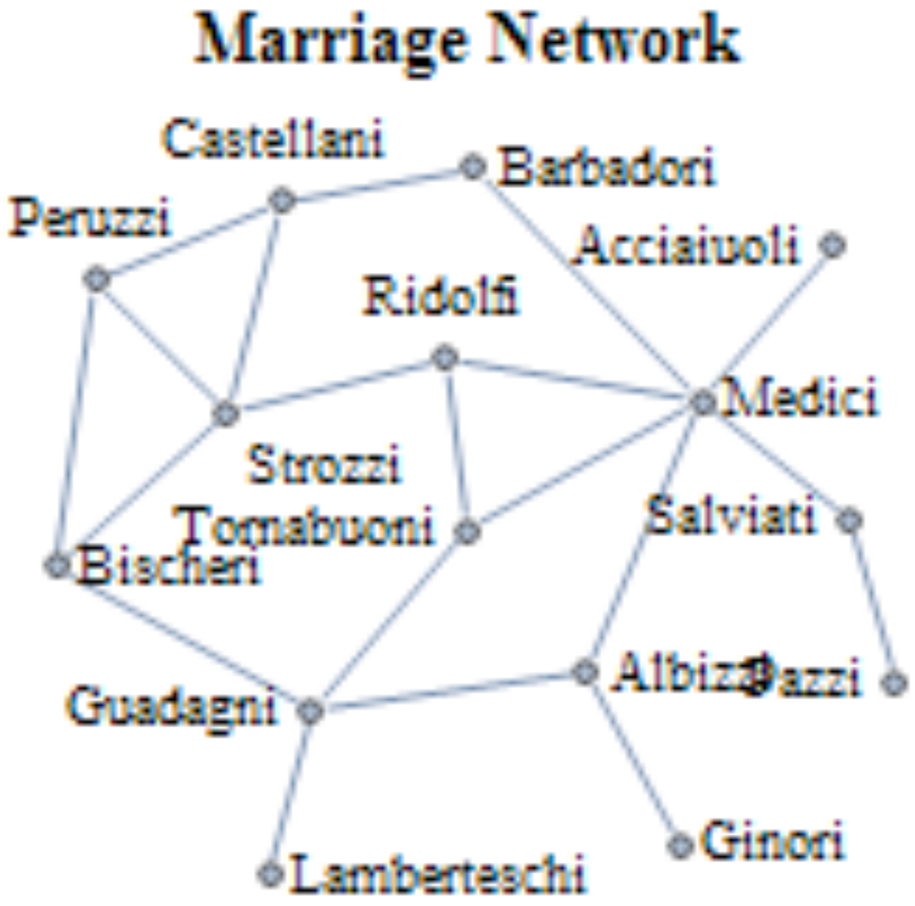


Closeness

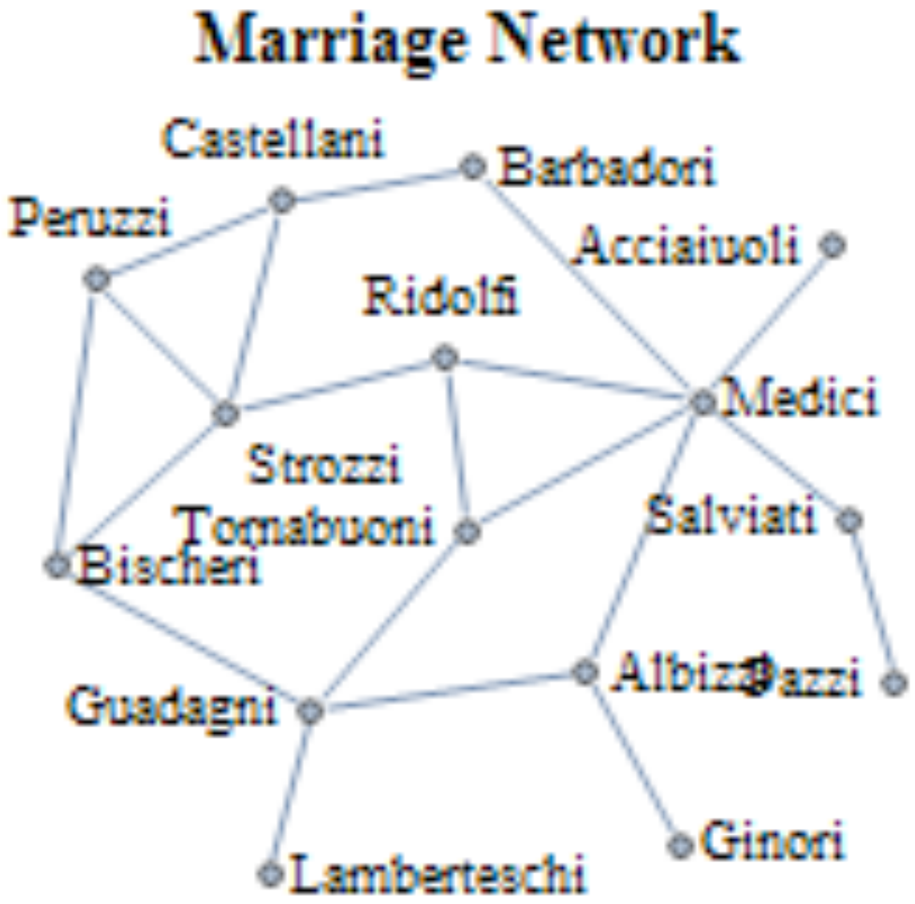


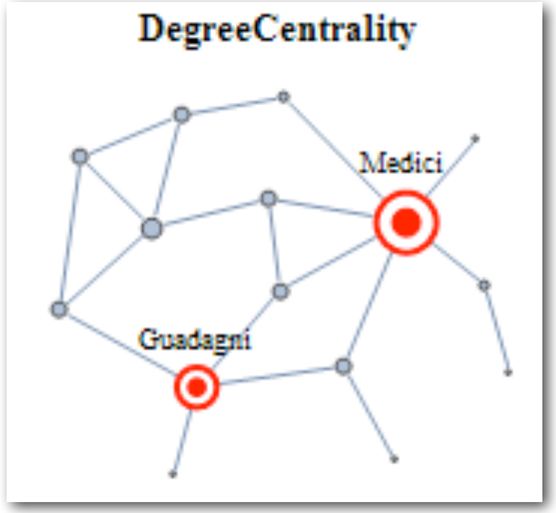
http://en.wikipedia.org/wiki/Centrality

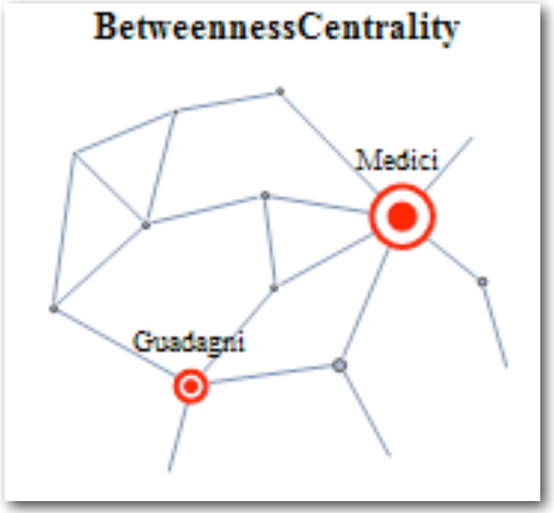
Influence of florentine families can be analyzed with centralities

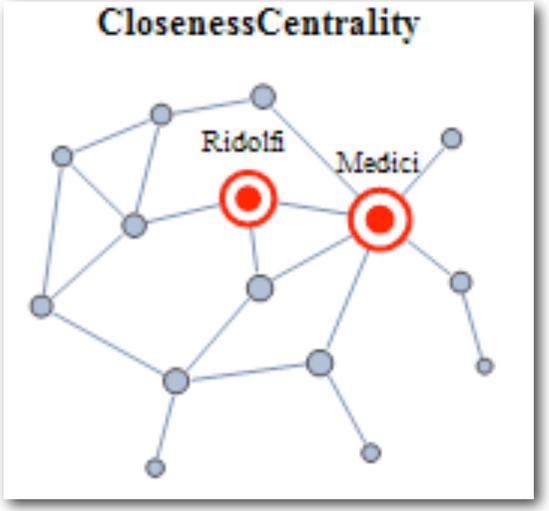


Influence of florentine families can be analyzed with centralities



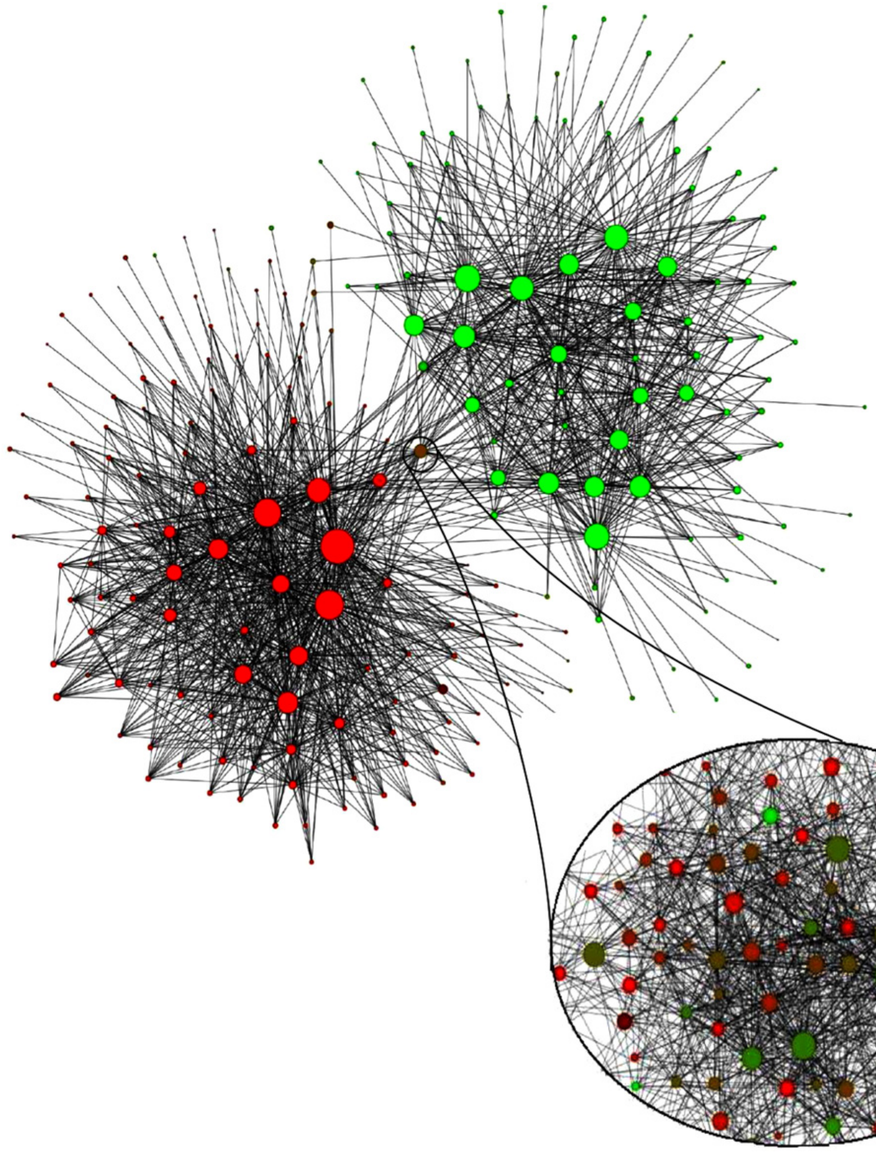


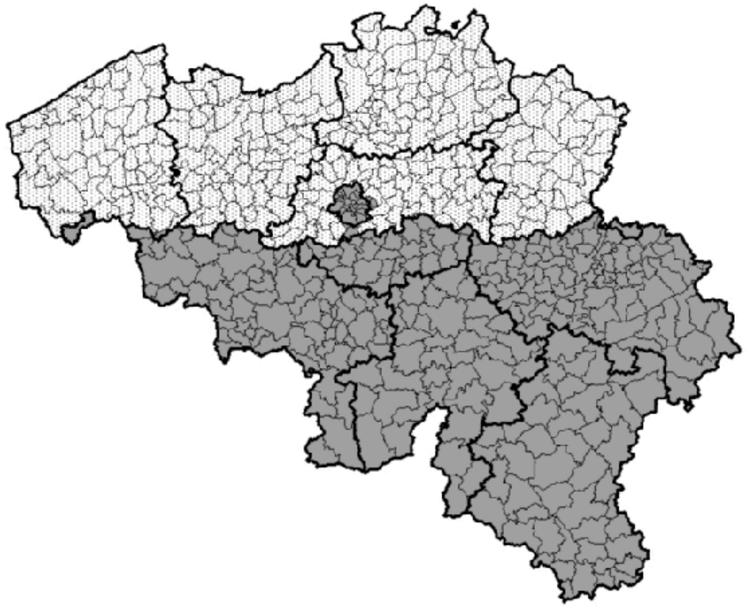


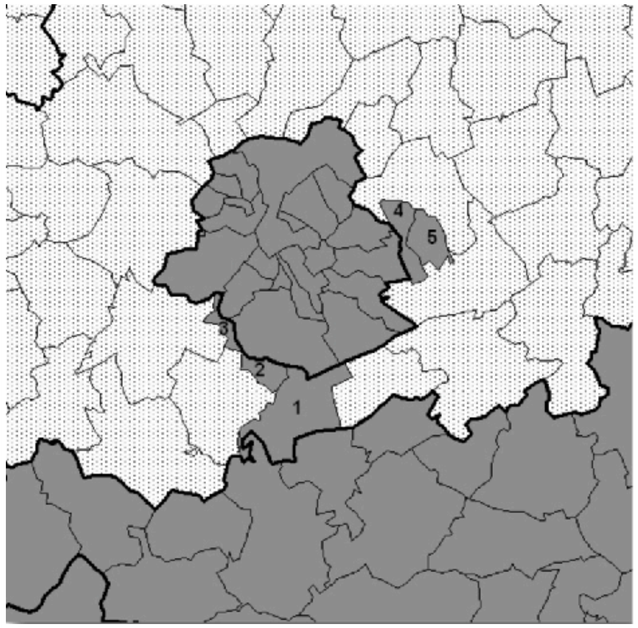


Community detection

Communities are sets of nodes with many links to each other

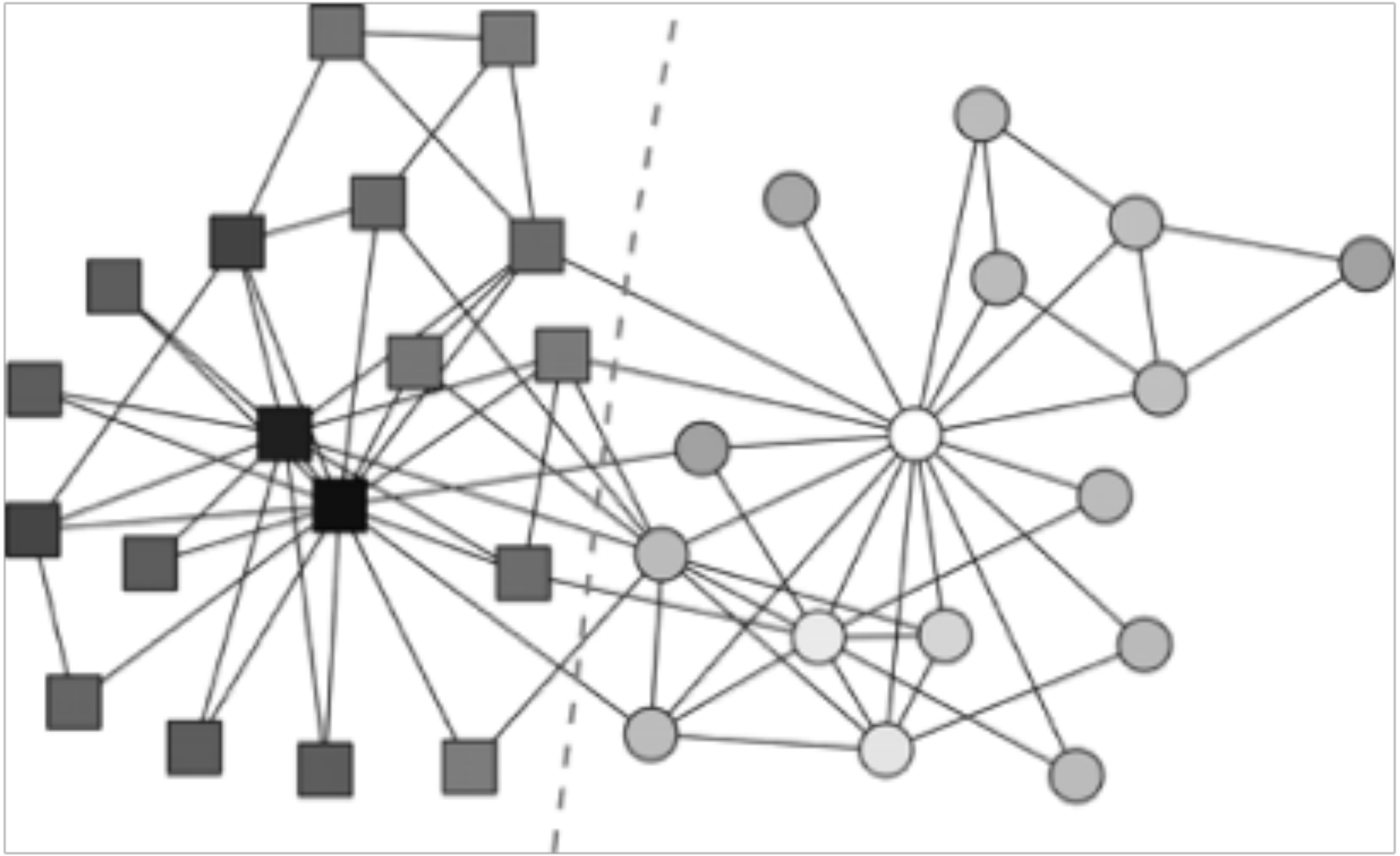






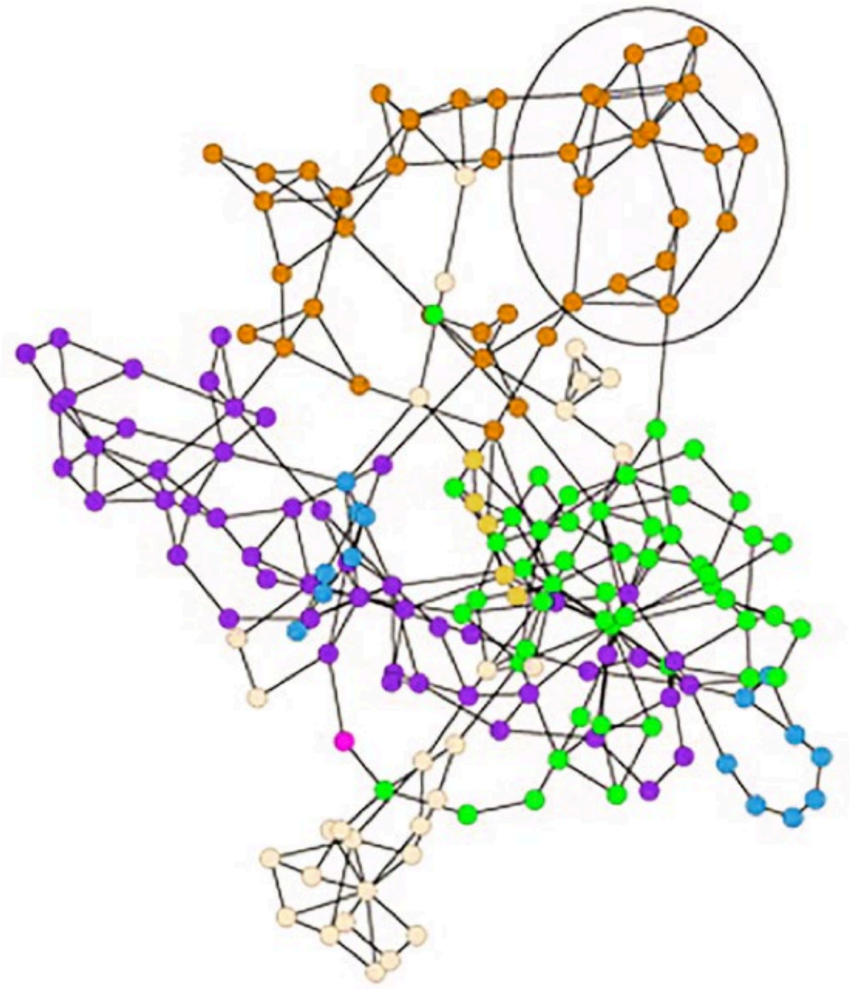
Blondel, Krings, Thomas. Brussels Studies 42 (2010)

The Zachary Karate Club is a benchmark for community detection



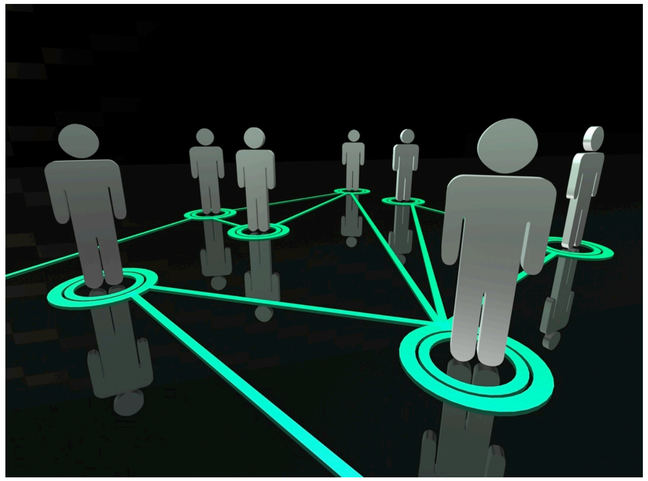
W.W. Zachary, J. Anthropol. Res. 33:452-473 (1977)

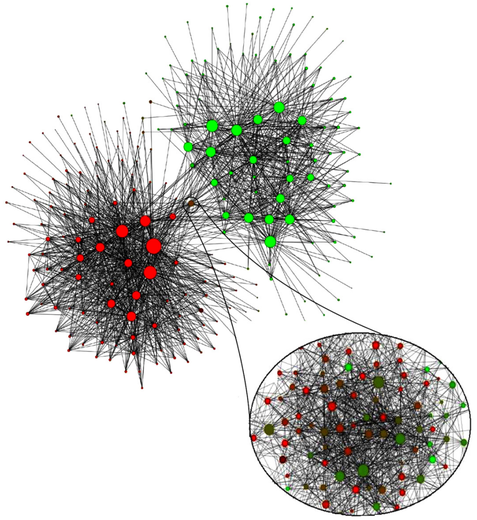
Here, communities are groups of molecules with more patterns of reactions with each other than with others



Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. S*cience*, *297*(5586), 1551-1555.

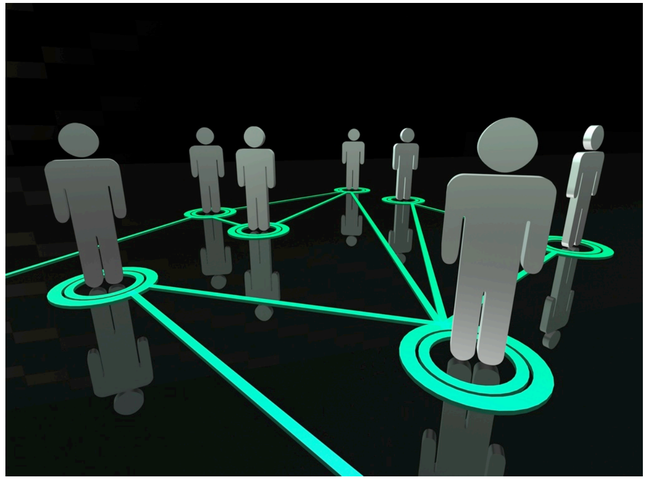
The study of communities focuses on the mesoscopic scale

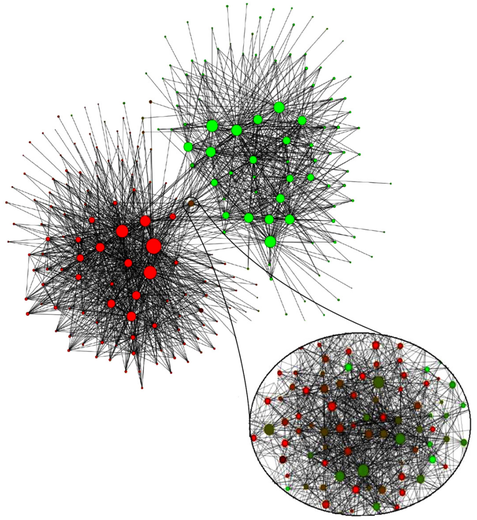




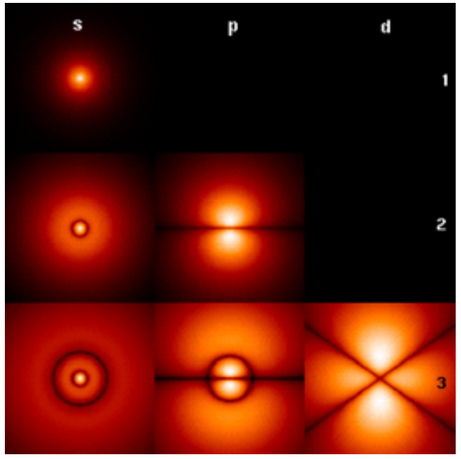
Microscopic Mesoscopic Macroscopic

The study of communities focuses on the mesoscopic scale

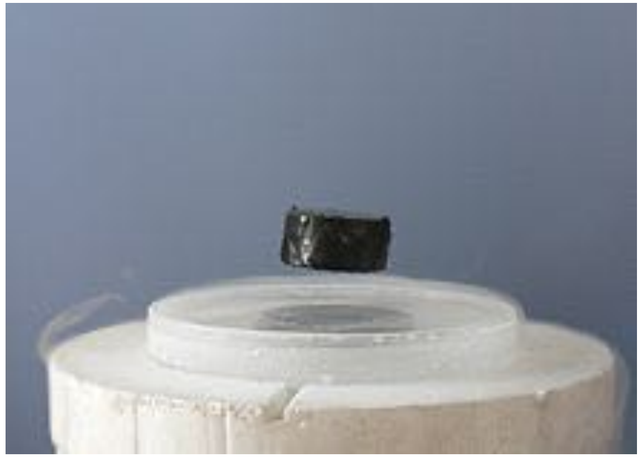




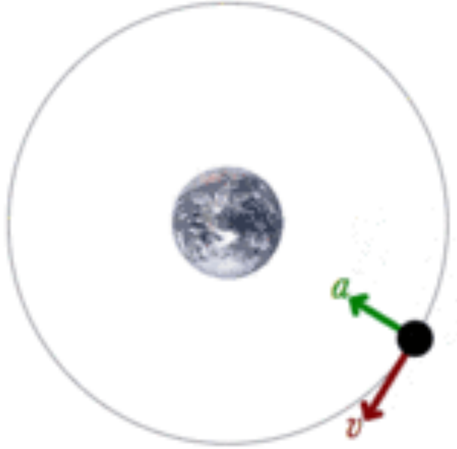
Microscopic



Mesoscopic



Macroscopic



Community detection has a long history

- **1927** The sociologists Stuart Rice used the similarity of voting patterns to identify communities in small political bodies [8].
- **1949:** Duncan R Luce and Albert D Perry used complete subgraphs to model cliques of individuals [9].
- **1950:** George Homans showed that social groups can be revealed by rearranging the rows and the columns of the adjacency matrix [3] (Figure 9.3).
- **1955:** Robert Weiss and Eugene Jacobson identified work groups within a government agency by removing the connectors, individuals that are linked to different groups [10].
- **1970s:** The need to minimize the communication time between processors on a chip has fueled an extensive literature on graph partitioning (Box 9.x) [11], leading to the Kernighan-Lin algorithm [12].
- **1973:** Mark Granovetter explored the role of communities on an individual's ability to find a job [13].

Community detection relies on some hypotheses

H1 (Fundamental Hypothesis)

A network's community structure is uniquely encoded in its wiring diagram

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

Community detection relies on some hypotheses

H1 (Fundamental Hypothesis)

A network's community structure is uniquely encoded in its wiring diagram

H2 (Connectedness Hypothesis)

A community corresponds to a connected subgraph

Community detection relies on some hypotheses

H1 (Fundamental Hypothesis)

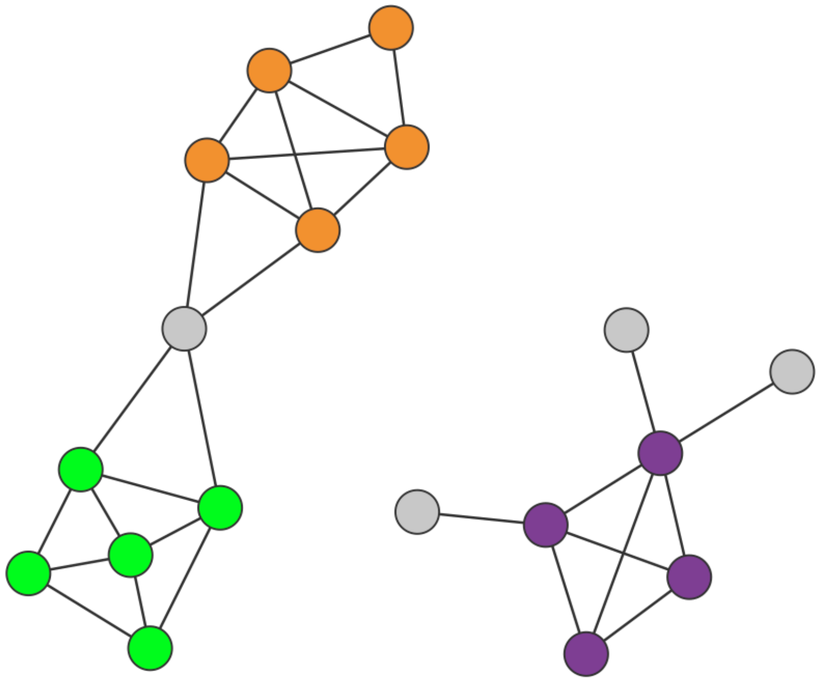
A network's community structure is uniquely encoded in its wiring diagram

H2 (Connectedness Hypothesis)

A community corresponds to a connected subgraph

H3 (Density Hypothesis)

Communities are locally dense subgraphs



Community detection relies on some hypotheses

H1 (Fundamental Hypothesis)

A network's community structure is uniquely encoded in its wiring diagram

H2 (Connectedness Hypothesis)

A community corresponds to a connected subgraph

H3 (Density Hypothesis)

Communities are locally dense subgraphs

H4 (Random Hypothesis)

Randomly wired networks are not expected to have a community structure

The random hypothesis leads to the measure of modularity M, which compares structure to a random model

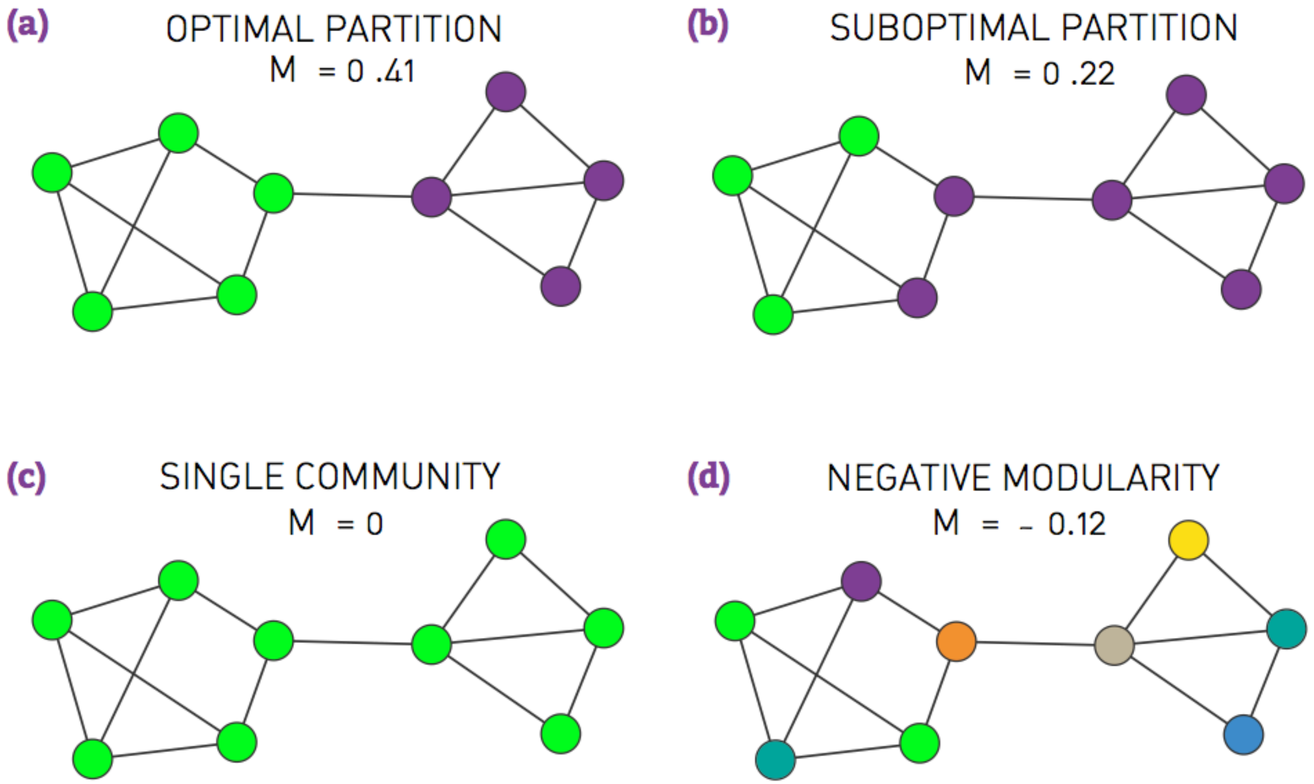
H5 (Maximal Modularity Hypothesis)

The partition with maximum modularity M for a given network offers the optimal community structure

The random hypothesis leads to the measure of modularity M, which compares structure to a random model

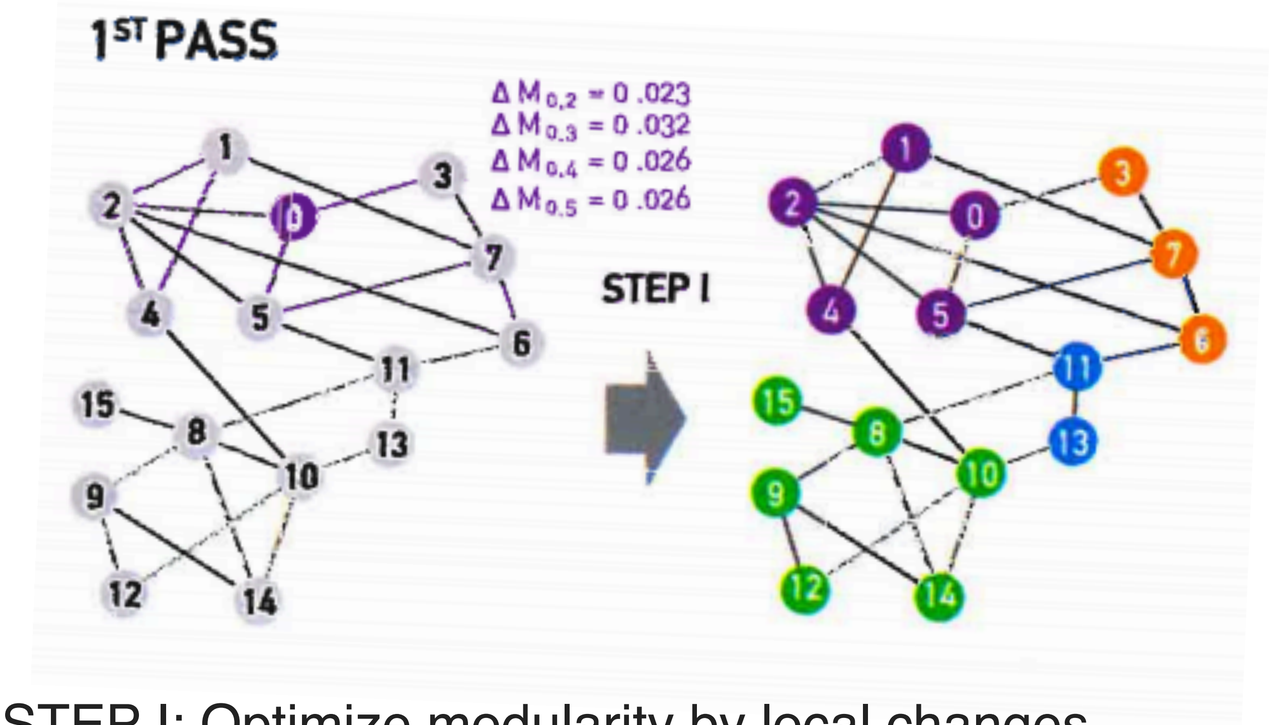
H5 (Maximal Modularity Hypothesis)

The partition with maximum modularity M for a given network offers the optimal community structure



Goal: Find the partition that maximizes M

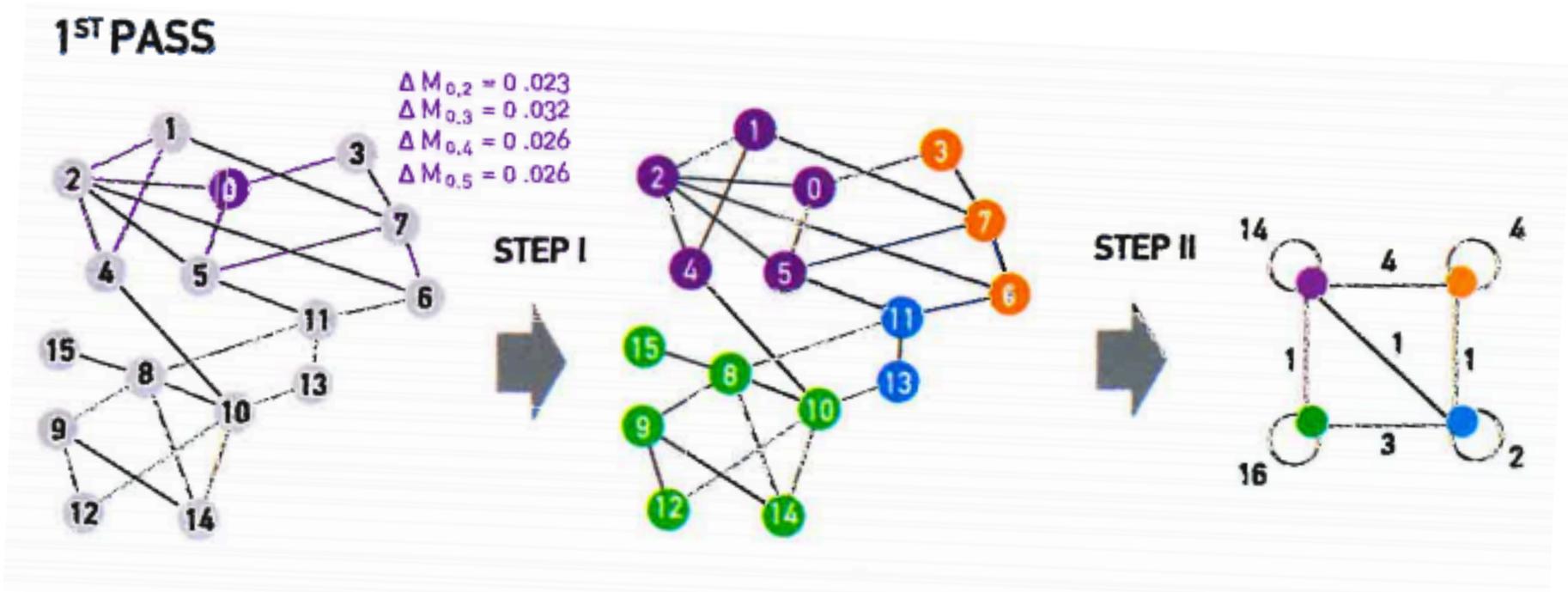
The Louvain method implements modularity maximization



STEP I: Optimize modularity by local changes.
Which of its neighbors should a node join?

Blondel, V. D. et al. (2008). *Journal of statistical mechanics 2008*(10), P10008.

The Louvain method implements modularity maximization

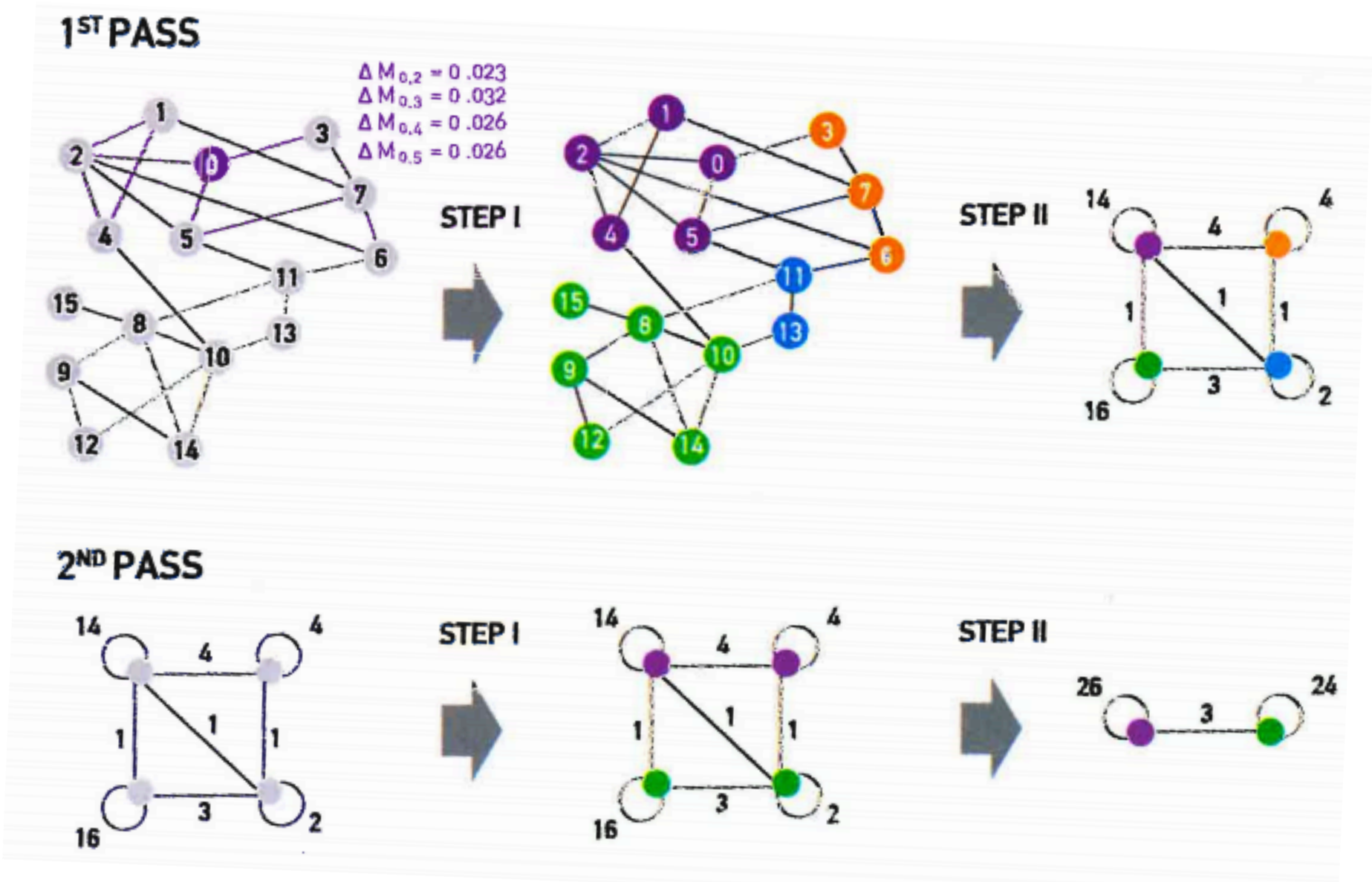


STEP I: Optimize modularity by local changes. Which of its neighbors should a node join?

STEP II: Aggregate the communities into single nodes. Weights are number of links.

Blondel, V. D. et al. (2008). *Journal of statistical mechanics 2008*(10), P10008.

The Louvain method implements modularity maximization



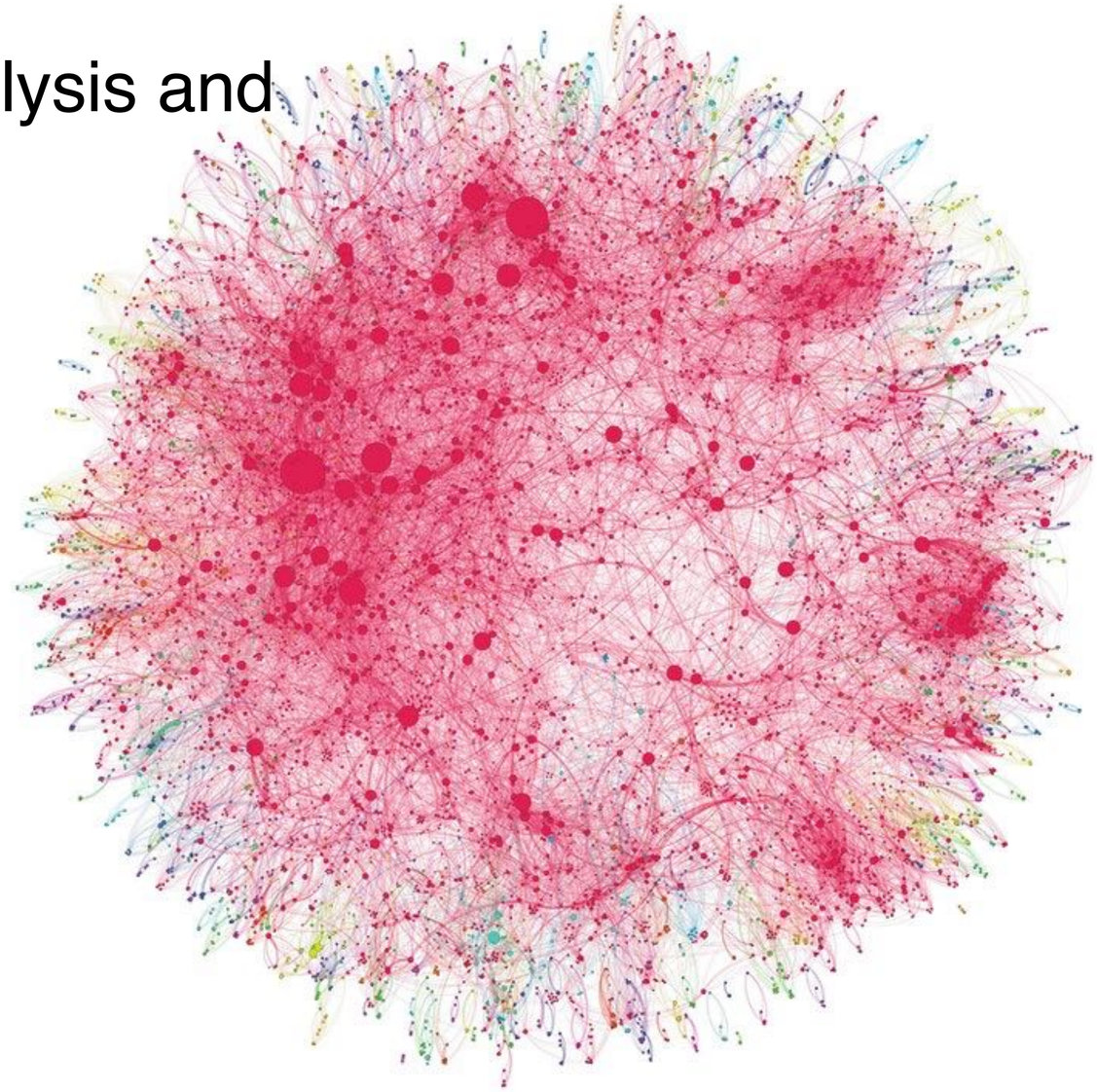
Blondel, V. D. et al. (2008). *Journal of statistical mechanics 2008*(10), P10008.

Next we are learning to use Gephi - the "Photoshop for networks"

Gephi is an open-source network analysis and visualization software written in Java

Gephi

gephi.org



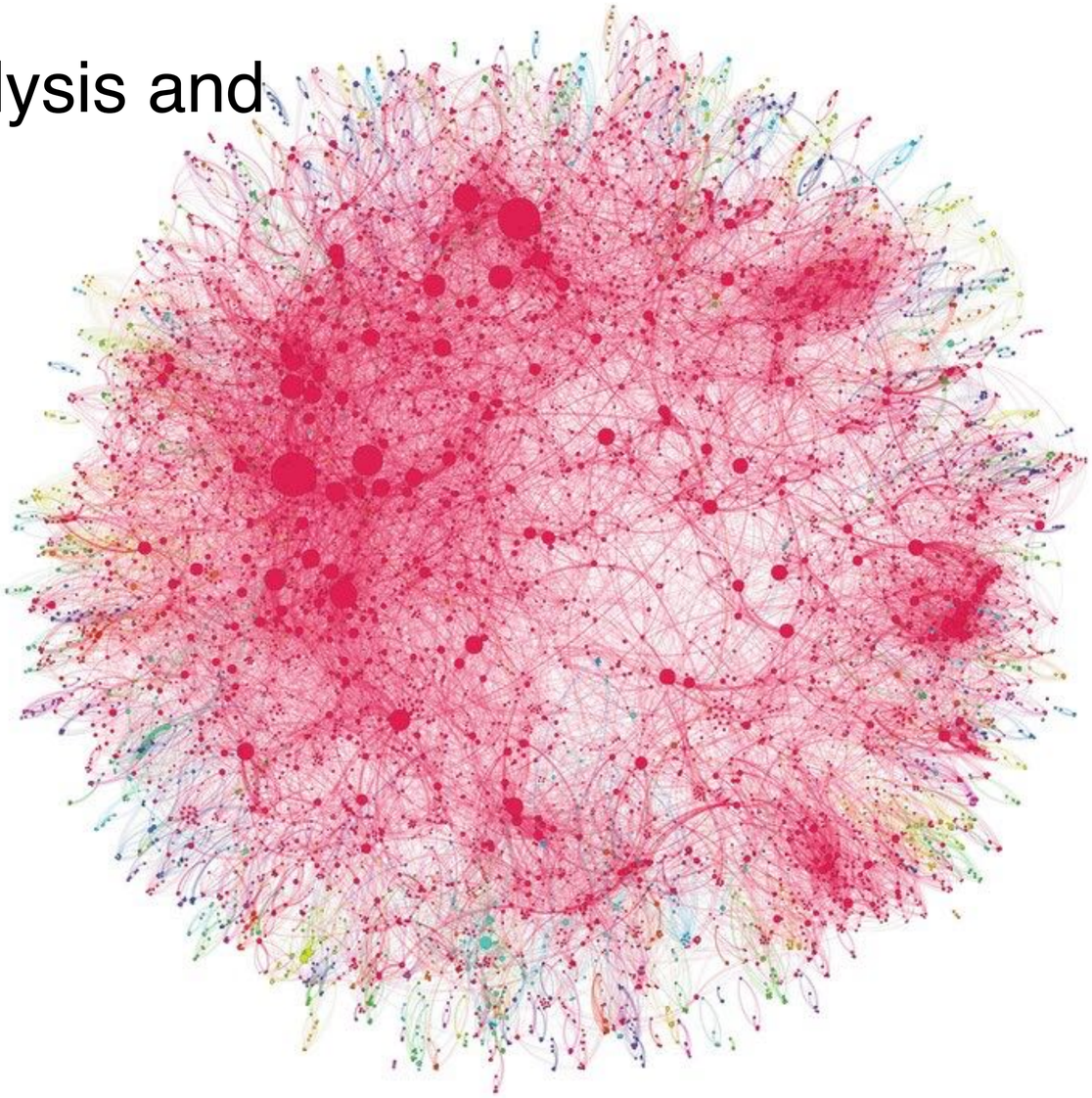
https://www.flickr.com/photos/speedoflife/8240814593/

Next we are learning to use Gephi - the "Photoshop for networks"

Gephi is an open-source network analysis and visualization software written in Java

Gephi

gephi.org

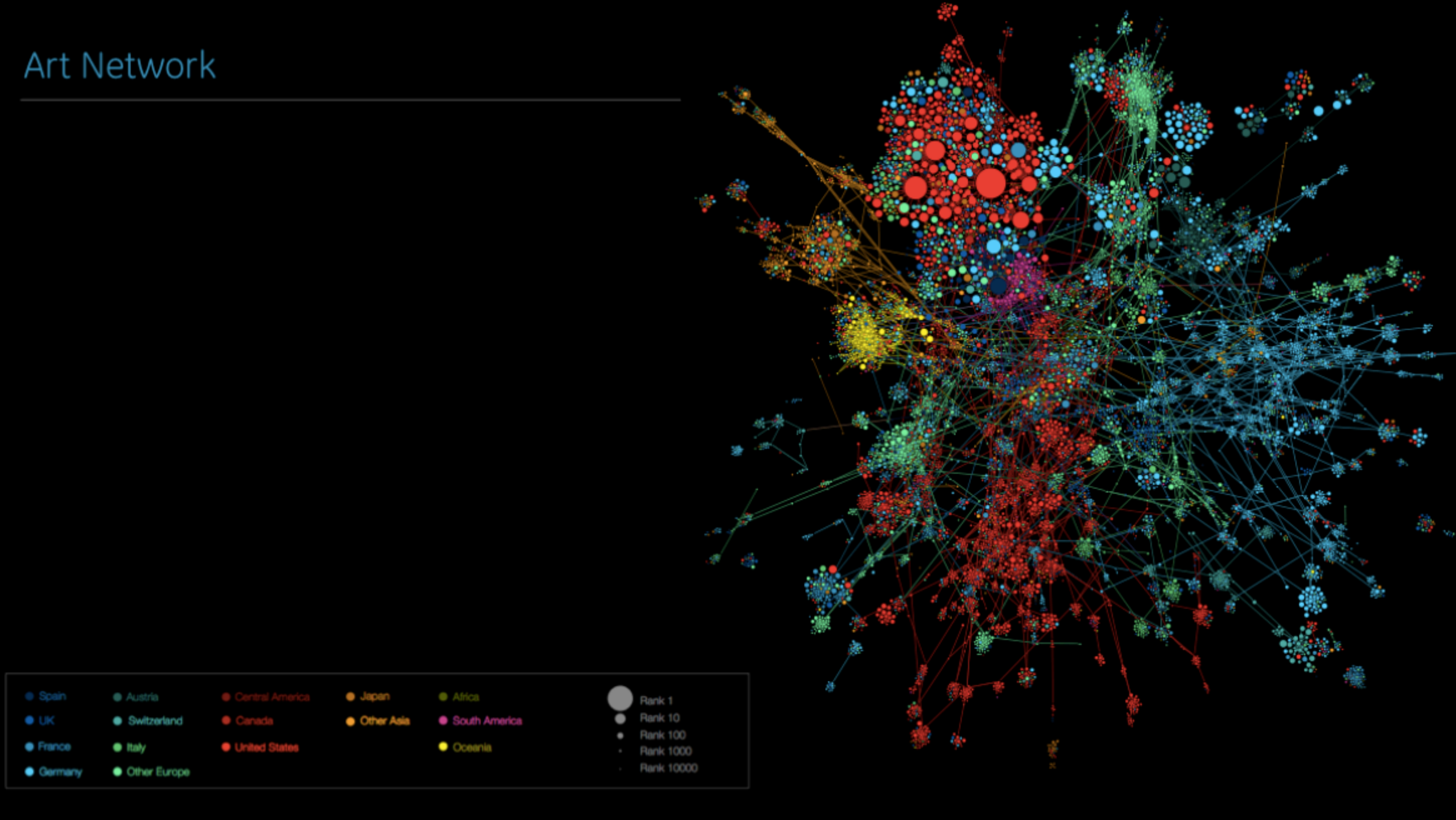


https://www.flickr.com/photos/speedoflife/8240814593/

Nice vector-image rendering
Many layouts
Quick black box network analysis

Not maintained anymore
No UNDO!

Art Network


Spain
UK
France
Germany
Austria
Switzerland
Italy
Other Europe
Central America
Canada
United States
Japan
Other Asia
Africa
South America
Oceania
Rank 1
Rank 10
Rank 100
Rank 1000
Rank 10000

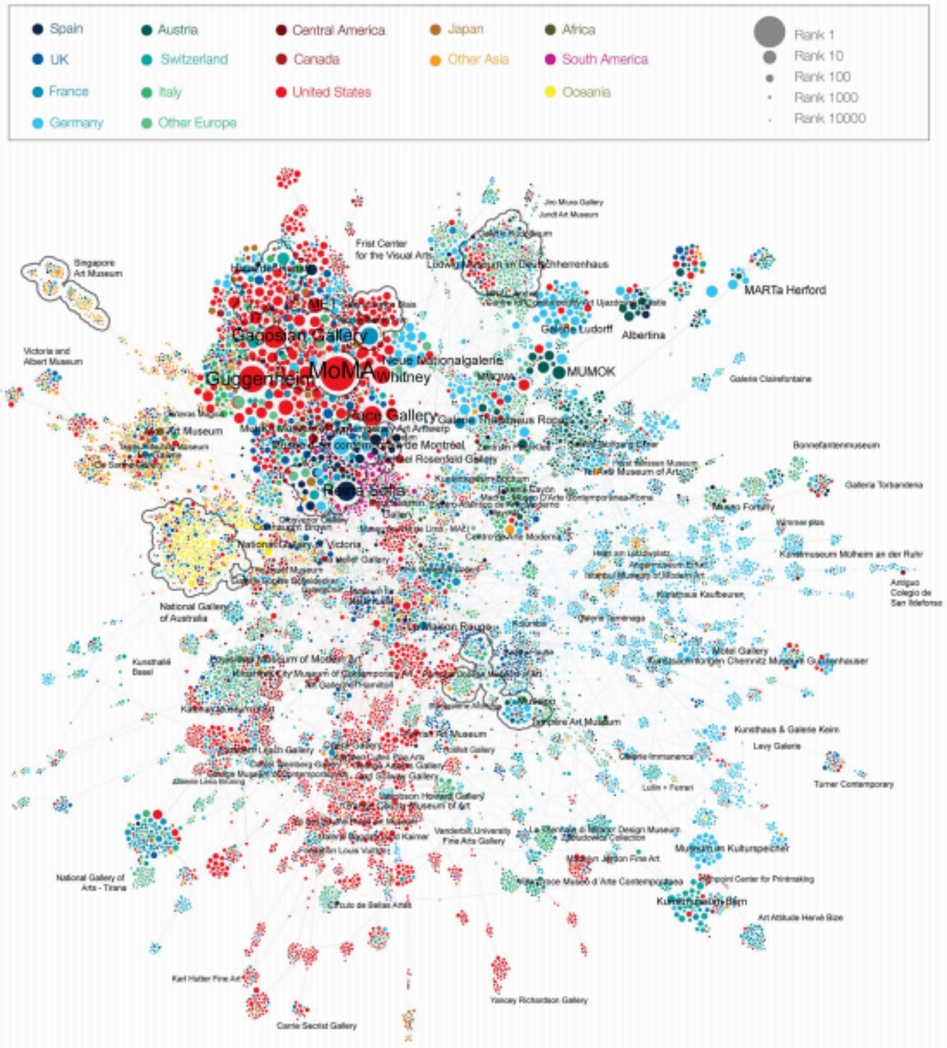


Fig. 1. Coexhibition network. Force-directed layout of the order $\tau = \infty$ coexhibition network, whose nodes are institutions (galleries, museums). Node size is proportional to each institution's eigenvector centrality. Nodes are connected if they both exhibited the same artist, with link weights being equal to the number of artists' coexhibitions. Node colors encode the region in which institutions are located. Links are of the same colors as their end nodes, or gray when end nodes have different colors. For visualization purposes, we only show the 12,238 nodes corresponding to institutions with more than 10 exhibits; we pruned the links by keeping the most statistically significant links (*20*) (supplementary text S2.2). We implemented community detection on the pruned network (*21*), identifying 122 communities (supplementary text S2.3). We highlighted five of them, the full community breakdown being shown in fig. S3. We also show the names of the most prestigious institution for each community.

http://science.sciencemag.org/content/362/6416/825/tab-pdf

Gephi

Let's now learn network analysis with NetworkX

NetworkX is the most used Python package for network analysis

Open source, easy to install
Many measures, algorithms, generators
Flexible: nodes can be anything, links can hold any attributes

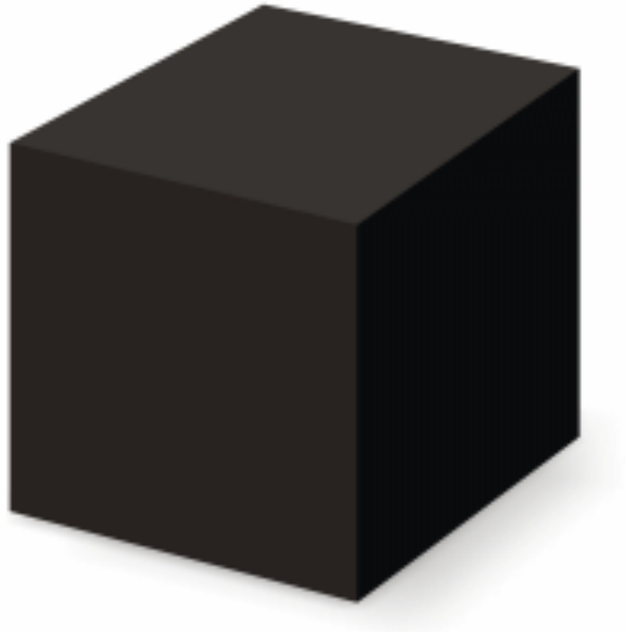
Data structures can be confusing
Relatively slow

https://networkx.github.io/

There are many other alternatives with pros and cons

Why did we go through all the trouble of coding Graph.py?

We cannot treat tools as black boxes but need to understand how algorithms work to:



- Make sense of results
- Pick the right algorithm for the problem at hand
- Fix things when they (invariably) go wrong

Jupyter