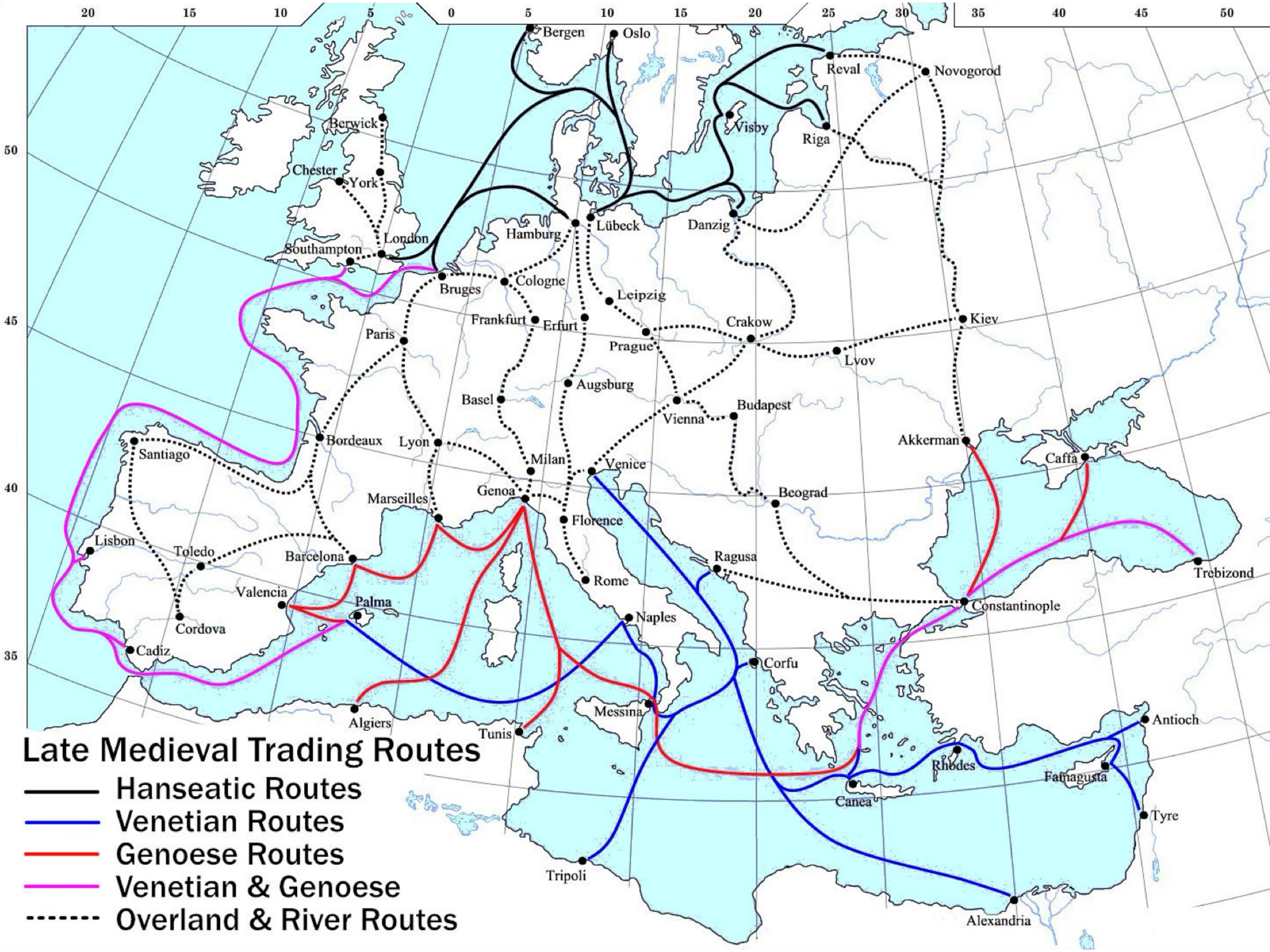
NETWORK SCIENCE FOR THE DIGITAL HUMANITIES

Shakti N. Menon

The Institute of Mathematical Sciences, Chennai

Mar 13th-24th 2025

Bits & Scripts II : Workshop on Computational Epigraphy



Late Medieval Trading Routes
Hanseatic Routes
Venetian Routes
Genoese Routes
Venetian & Genoese
Overland & River Routes
20
15
10
5
0
5
10
15
20
25
30
35
40
45
50
50
45
40
35
Bergen
Oslo
Reval
Novogorod
Visby
Riga
Berwick
Chester
York
London
Southampton
Hamburg
Lübeck
Danzig
Bruges
Cologne
Leipzig
Frankfurt
Erfurt
Paris
Prague
Crakow
Kiev
Lvov
Augsburg
Basel
Vienna
Budapest
Bordeaux
Lyon
Santiago
Milan
Venice
Akkerman
Caffa
Genoa
Marseilles
Florence
Beograd
Lisbon
Toledo
Barcelona
Ragusa
Trebizond
Rome
Valencia
Palma
Constantinople
Naples
Cordova
Cadiz
Corfu
Messina
Algiers
Tunis
Antioch
Rhodes
Famagusta
Canea
Tyre
Tripoli
Alexandria

WHAT DO MAPS AND OTHER SUCH ARTEFACTS ACTUALLY TELL US?

René Magritte, *La Trahison des images* ("*The Treachery of Images*", 1929)

"…What do you consider the largest map that would be really useful?"

"About six inches to the mile."

"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"

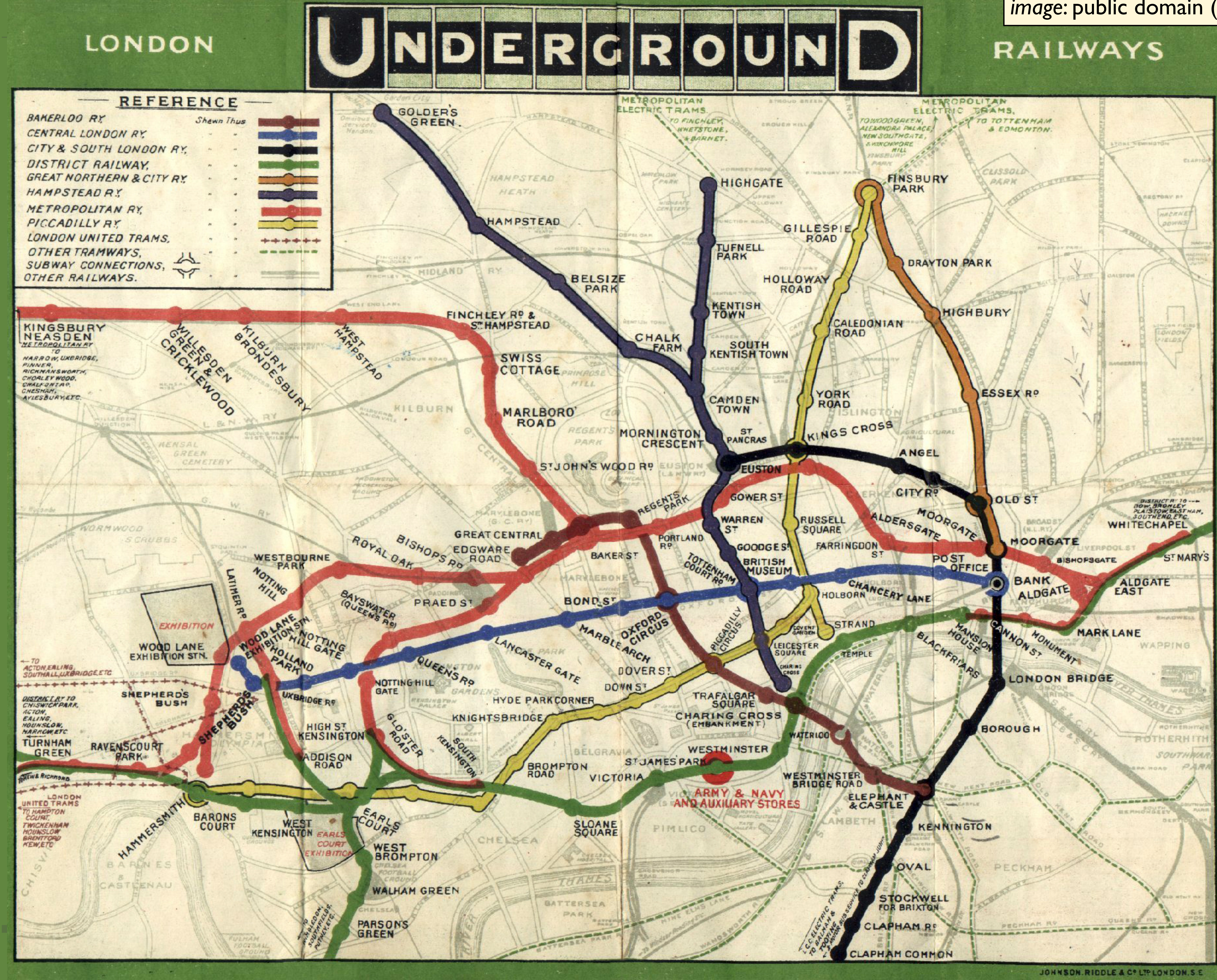
"Have you used it much?" I enquired.

"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."

Lewis Carroll, *"Sylvie and Bruno Concluded" (1893)*

The type/amount of information that is encoded within an artefact impacts the inferences that can be made from analysing it.

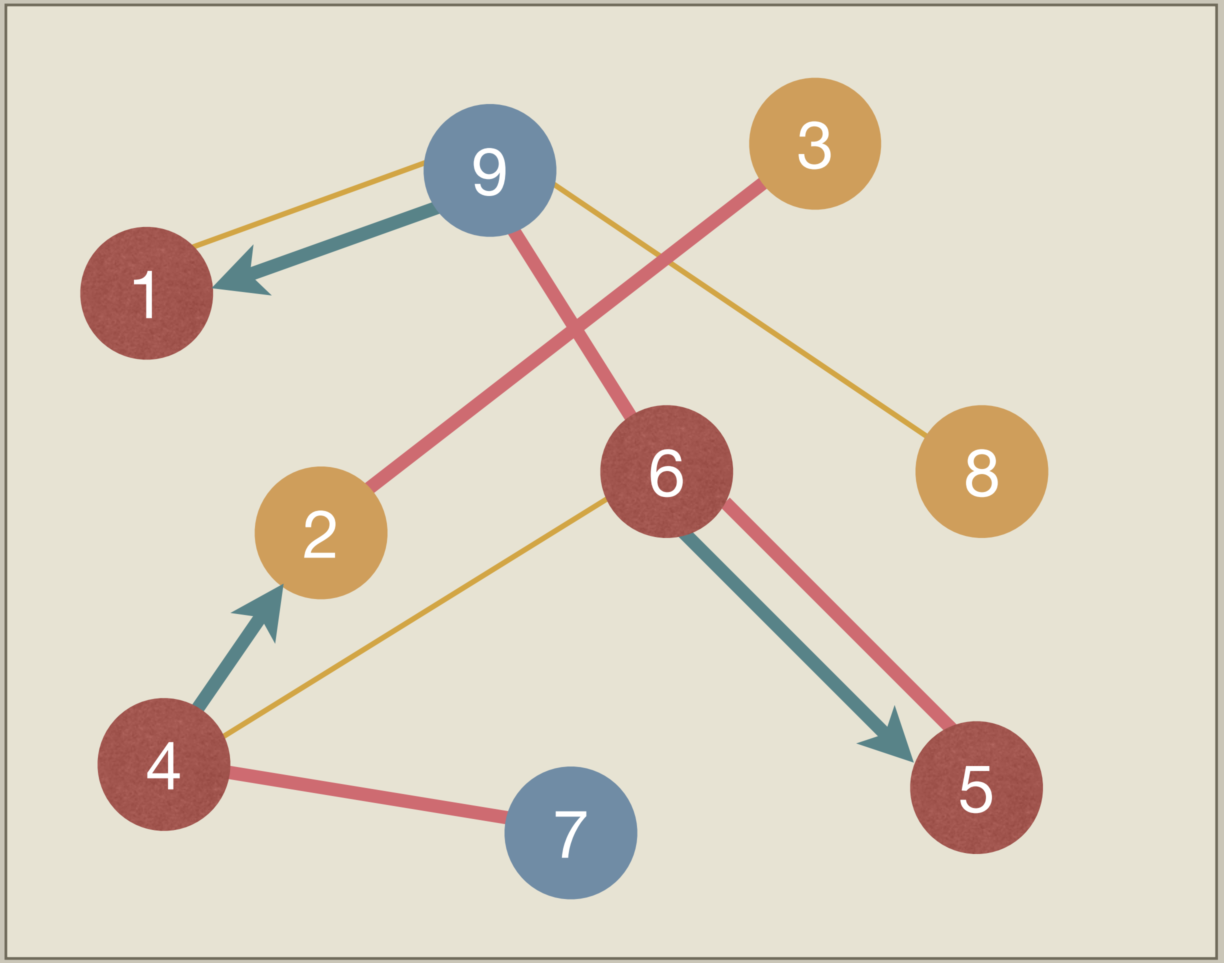
image: public domain (1908)

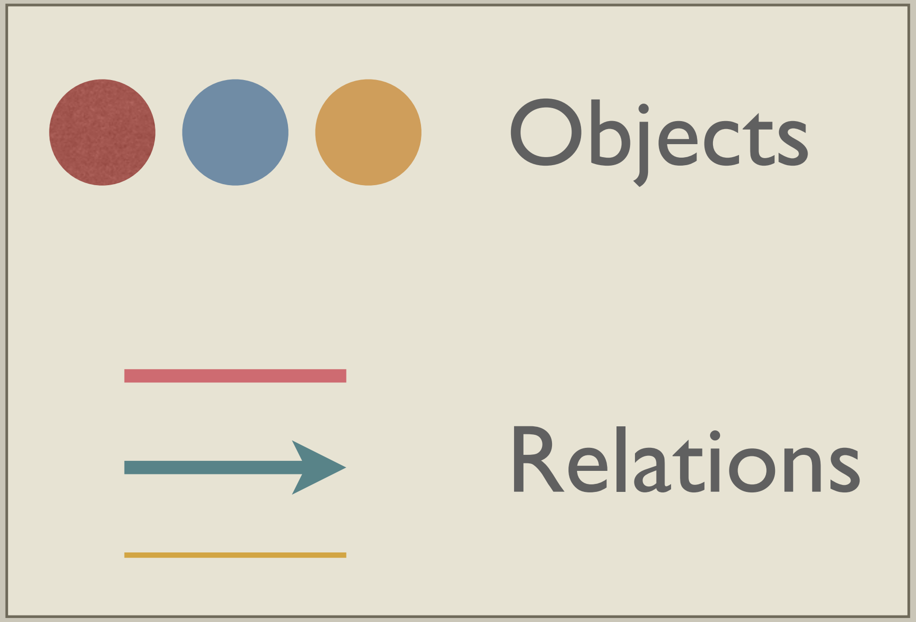


Harry Beck's Tube map (1933)

image: Rex Features

NETWORKS AS AN ABSTRACTION OF REALITY



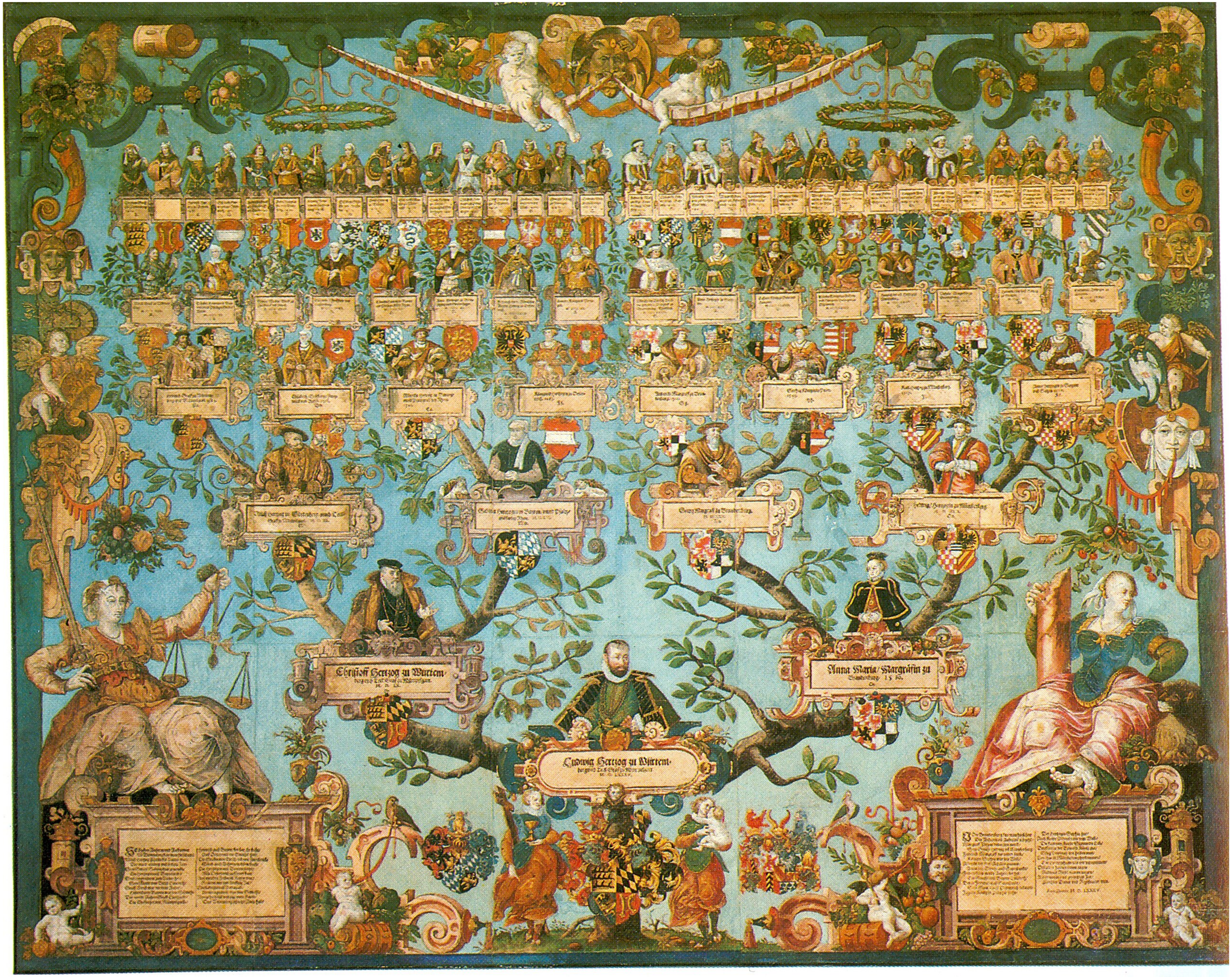


In general, this scenario can change over time

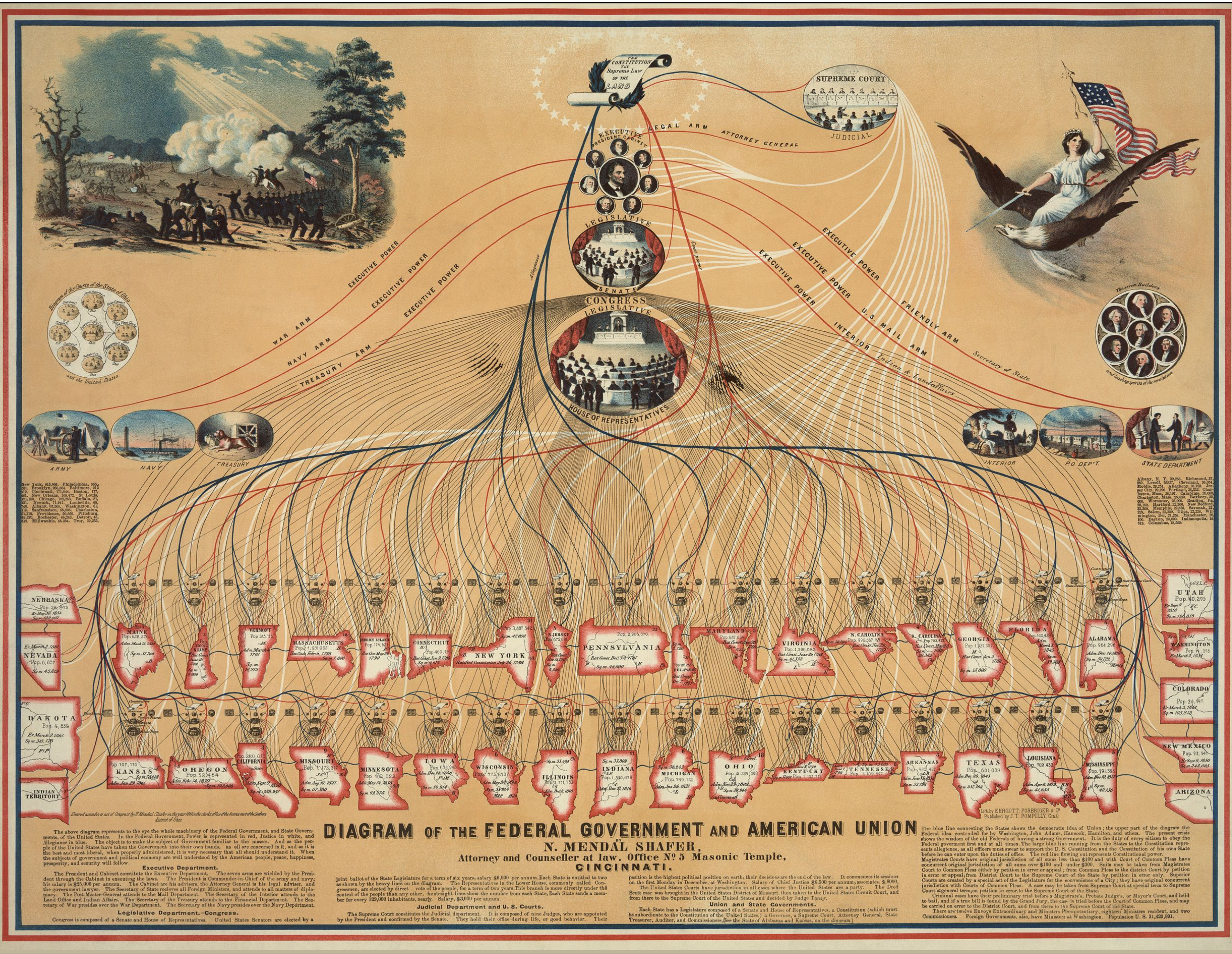
Every object has an associated set of attributes. Objects can also be classified based on certain common attributes.

Relations are classified based on type, direction, intensity, etc.

Family tree of Louis III, Duke of Württemberg (1585)



“Diagram of the Federal Government and American Union” by N. Mendal Shafer (1862)



Social network in ancient Athens (from Plutarch's *Parallel Lives*)

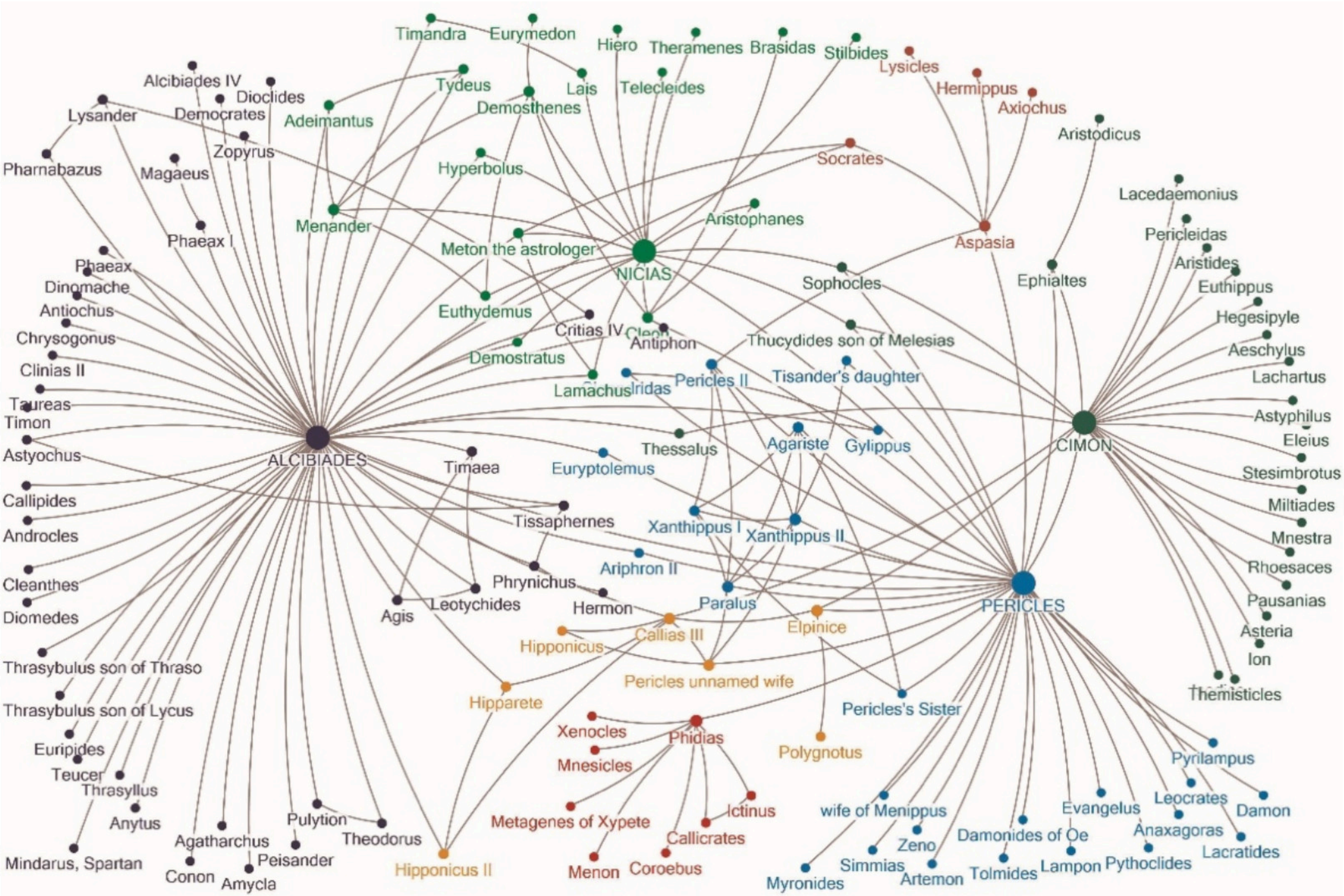
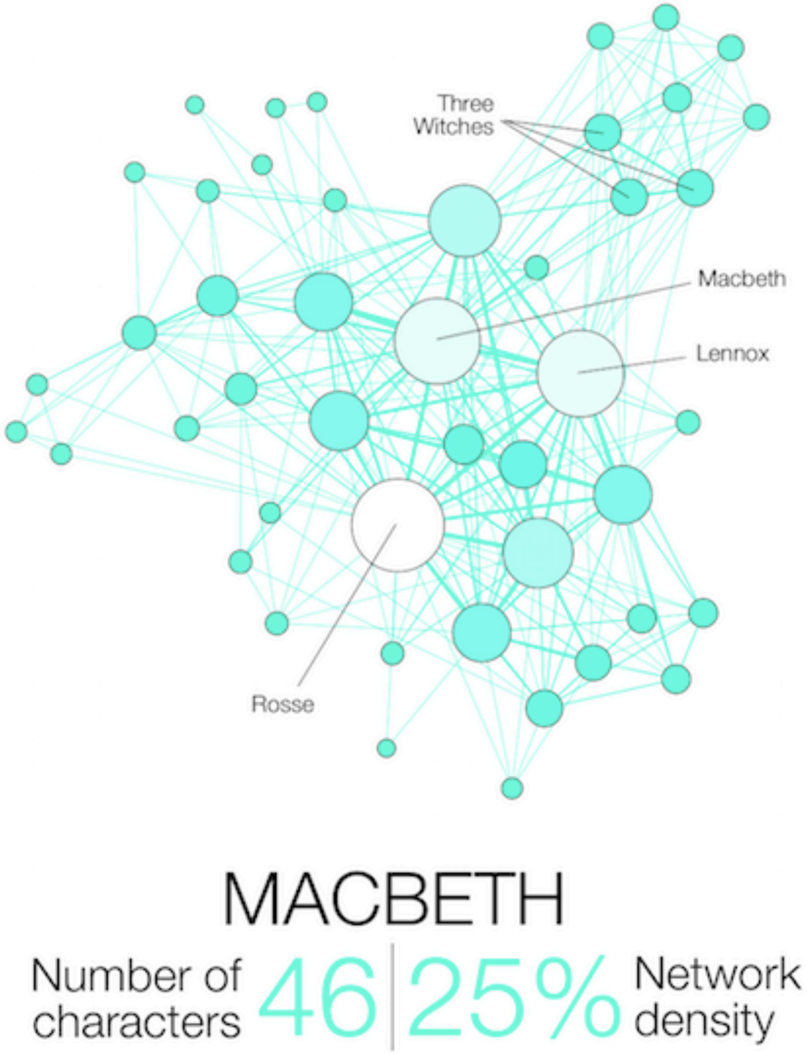
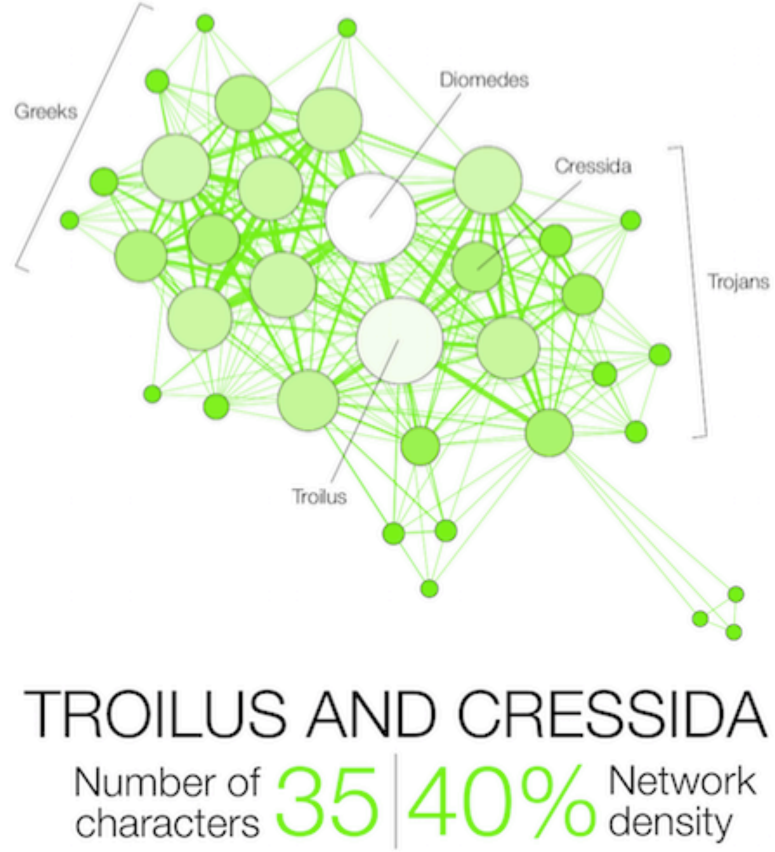
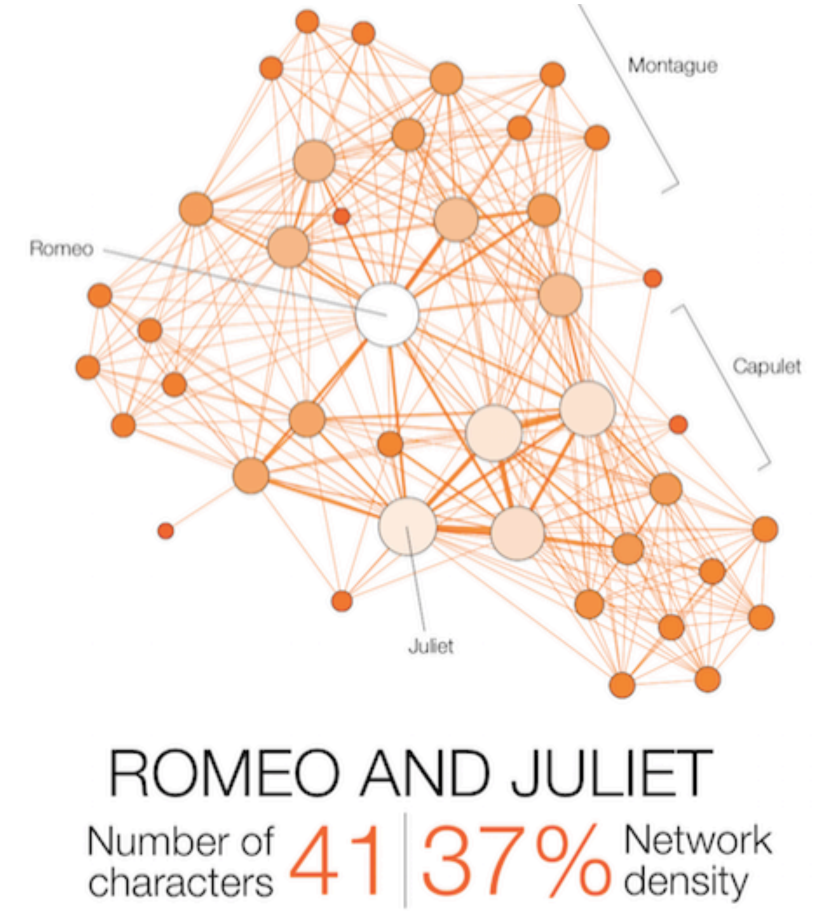


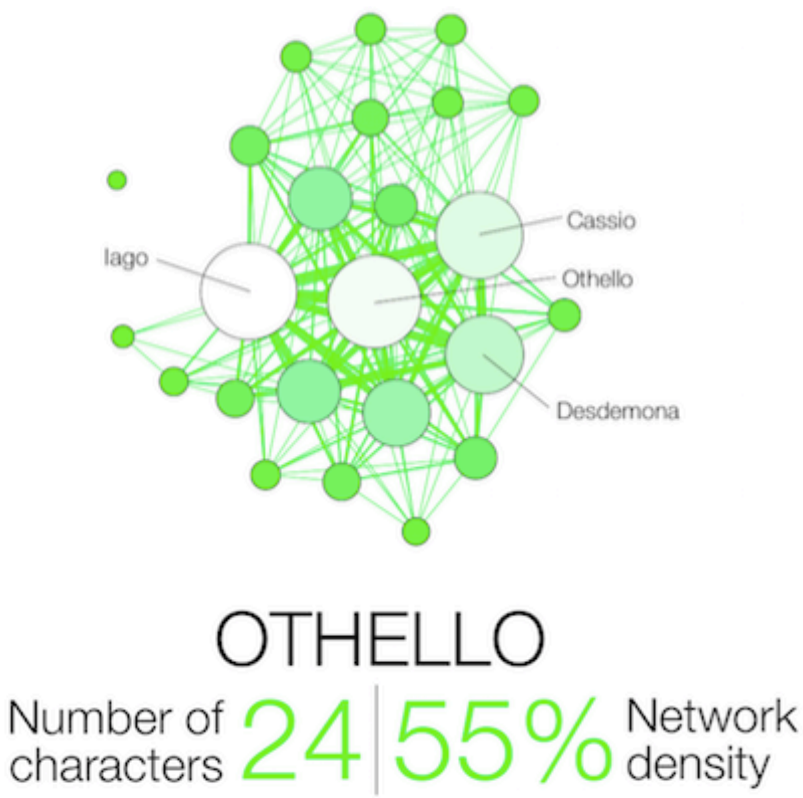
image: Cline, DH, "Athens as a Small World", Journal of Historical Network Research 4, 36-56 (2020).

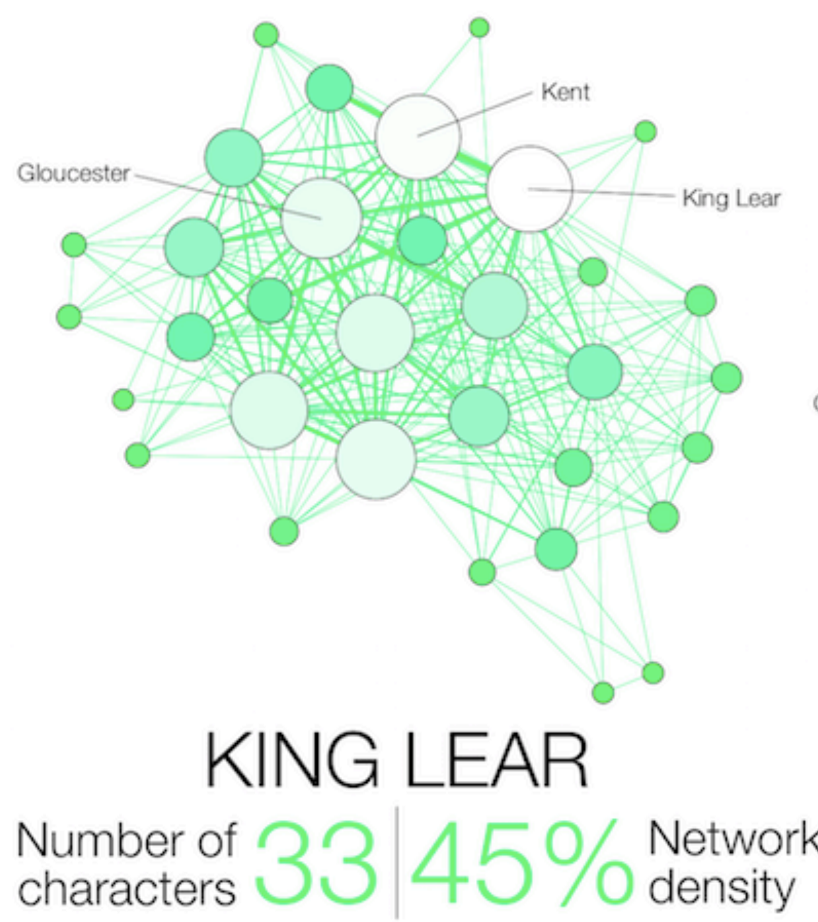
Shakespeare's tragedies











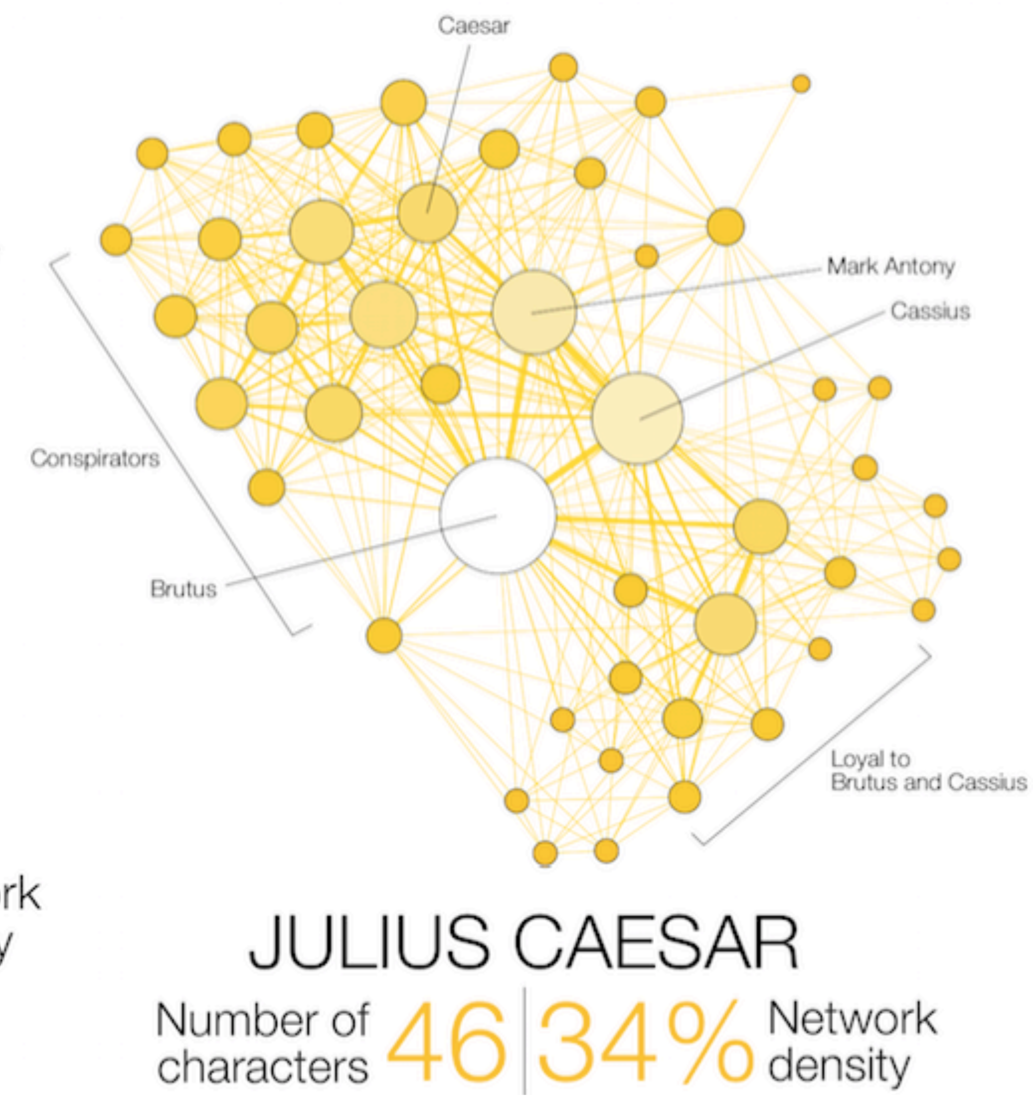
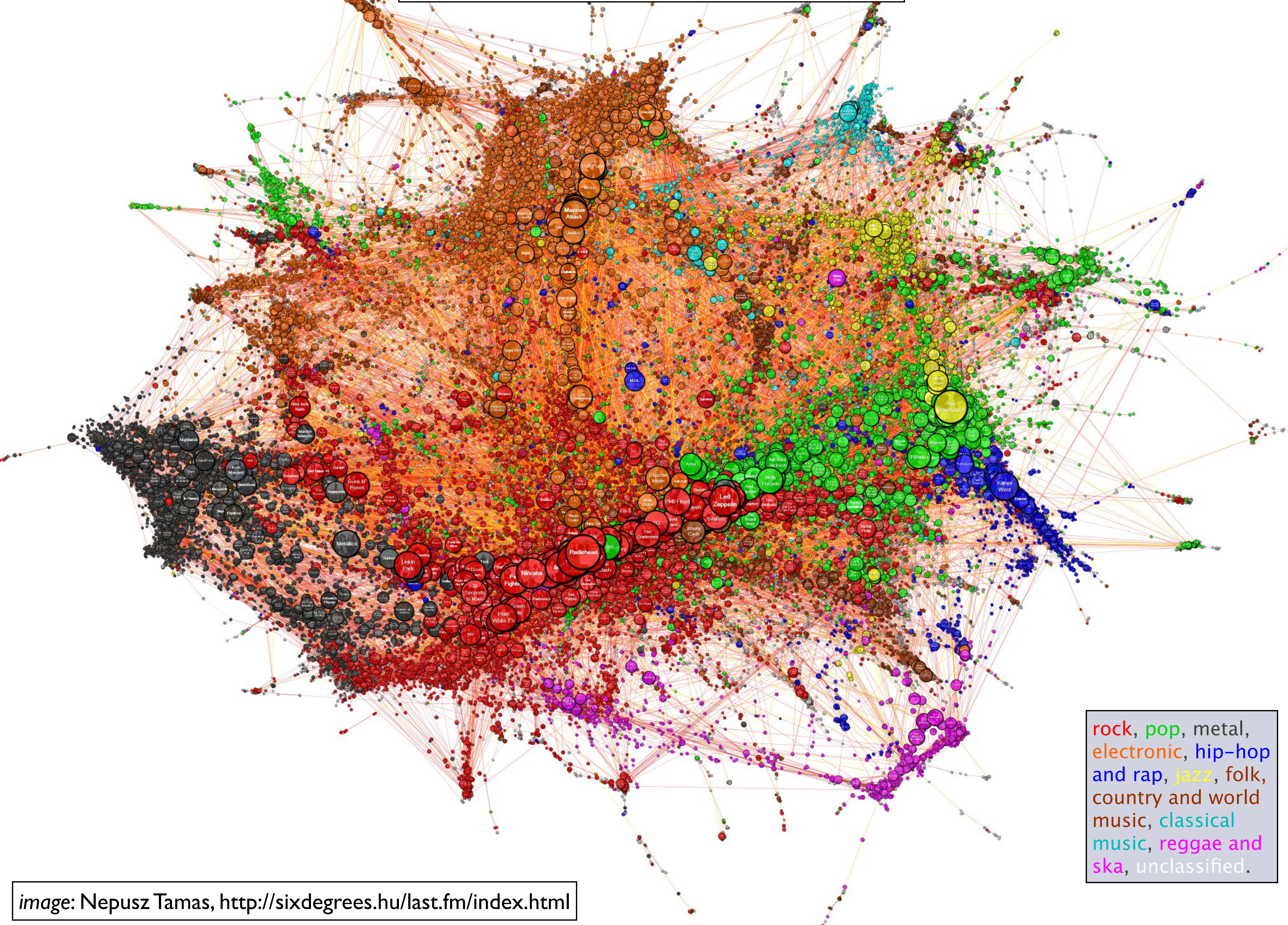


image: https://www.martingrandjean.ch/network-visualization-shakespeare/

Similarity Network of Musicians



rock, pop, metal, electronic, hip-hop and rap, jazz, folk, country and world music, classical music, reggae and ska, unclassified.

image: Nepusz Tamas, http://sixdegrees.hu/last.fm/index.html

The Political Blogosphere and the 2004 U.S. Election

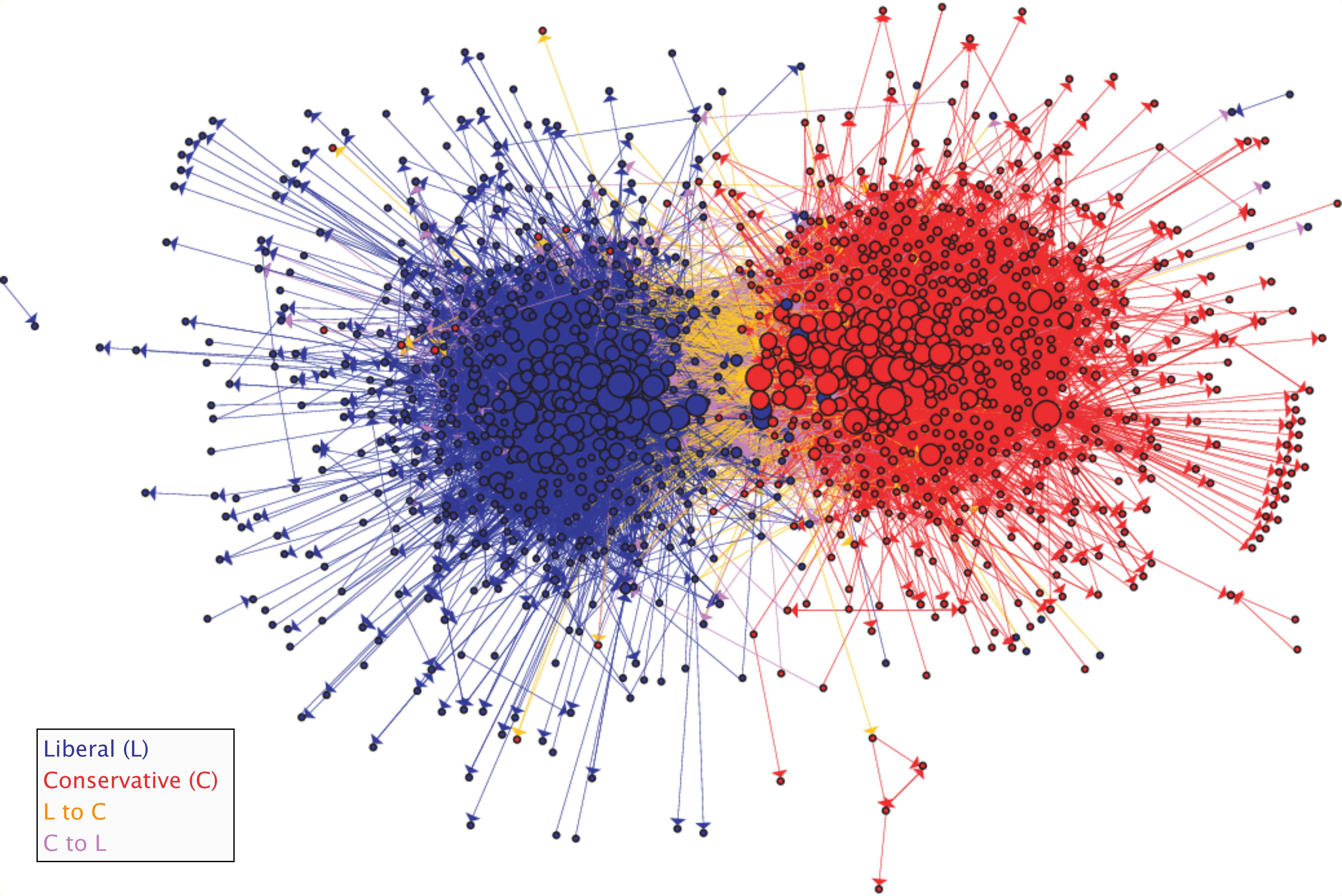
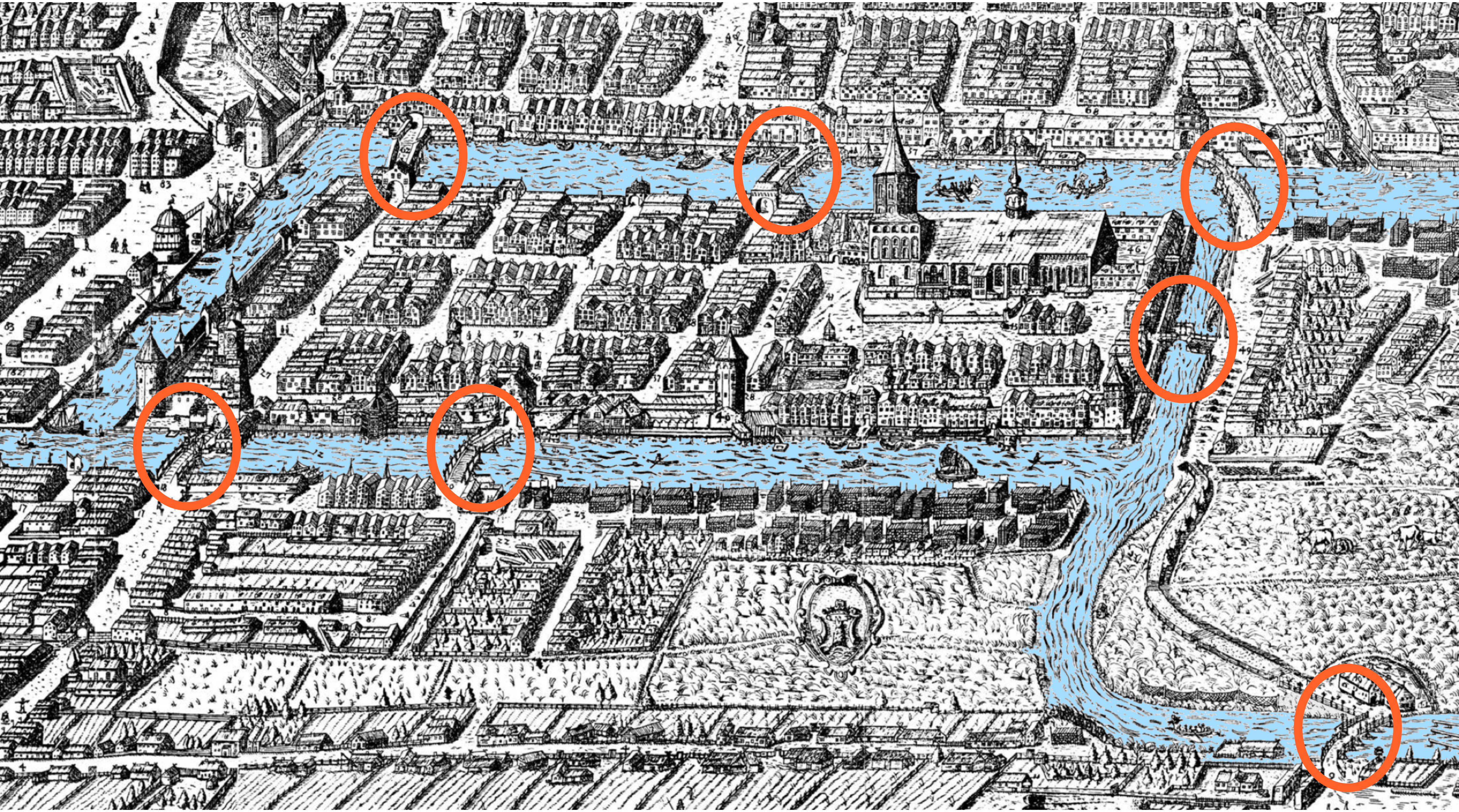
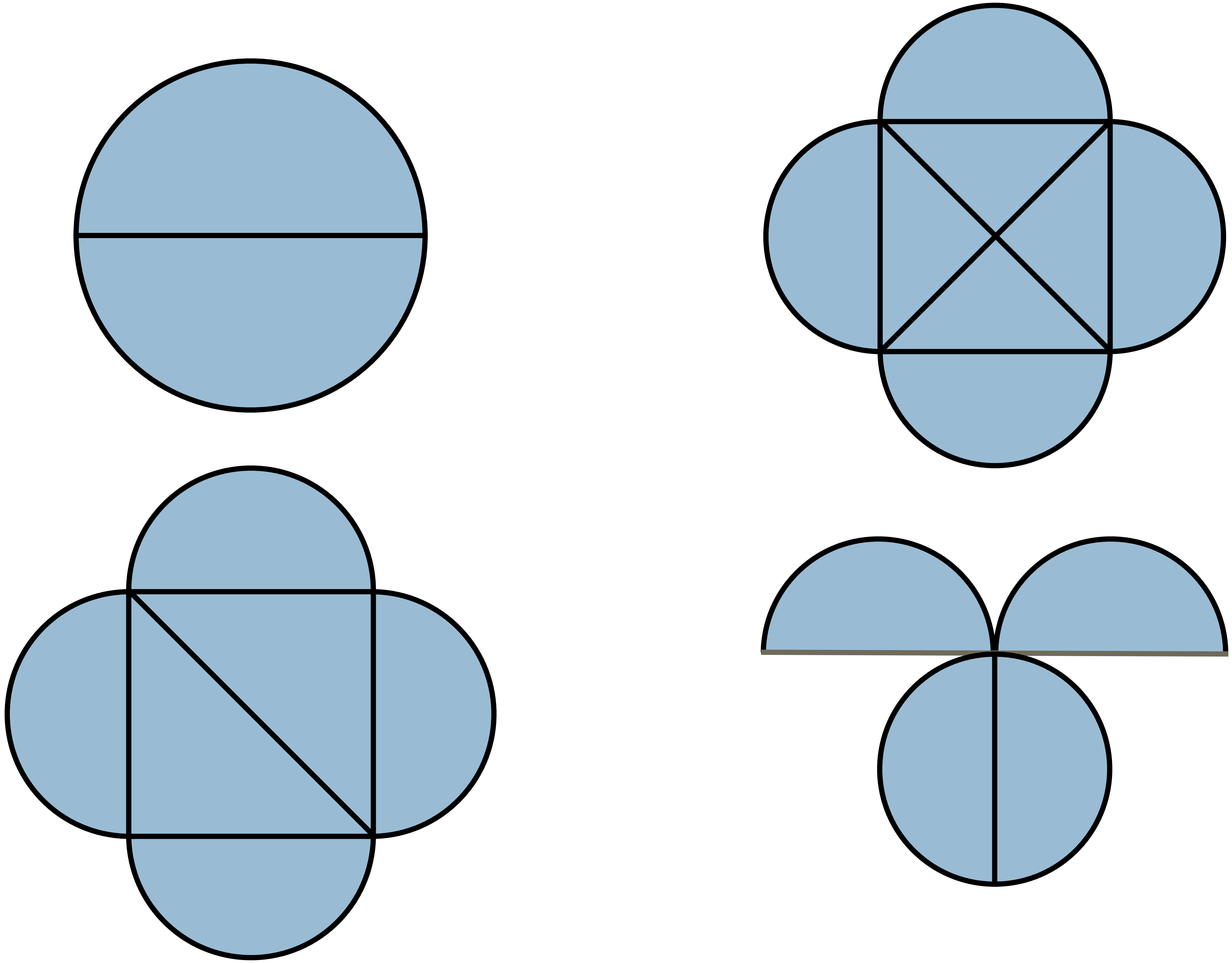


image: Lada Adamic & Natalie Glance, Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005).

THE SEVEN BRIDGES OF KÖNIGSBERG

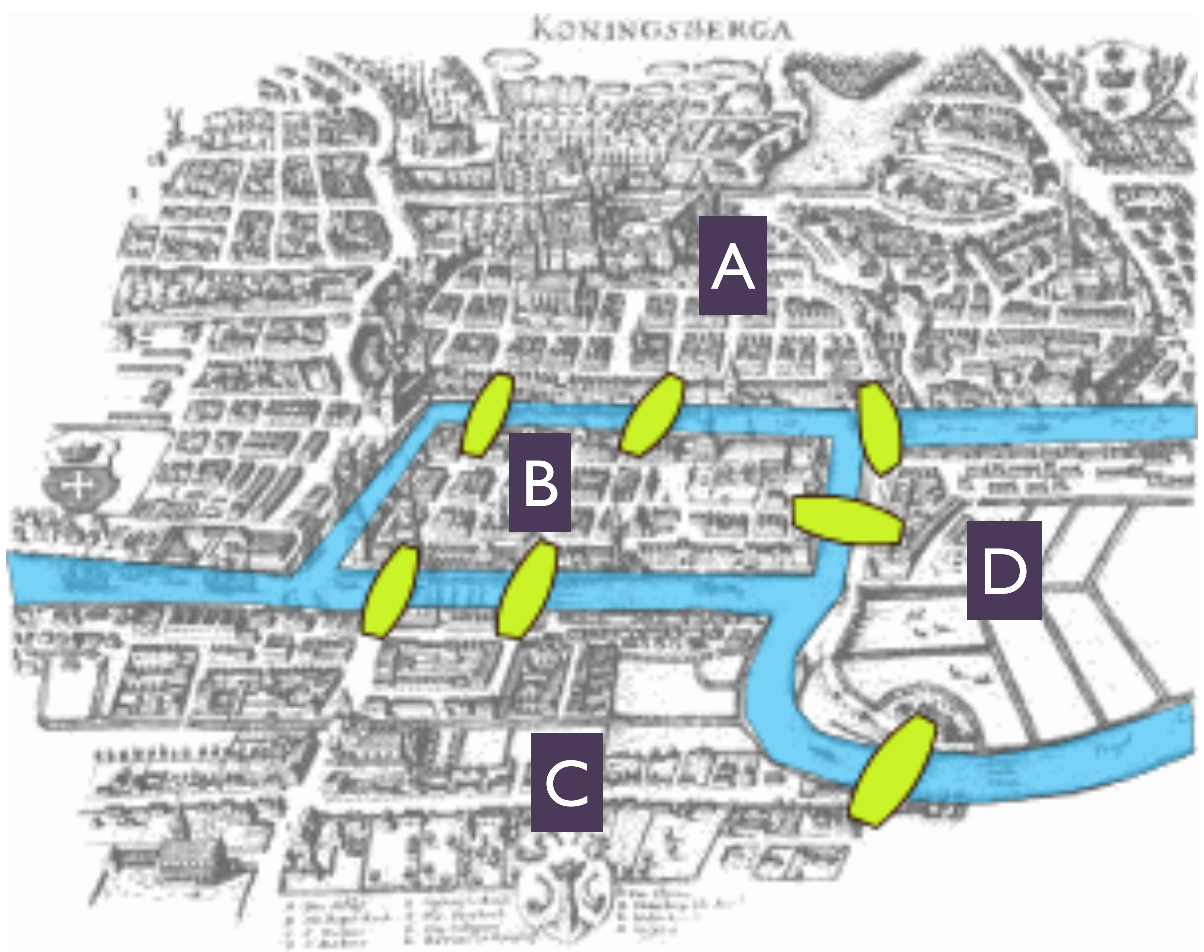


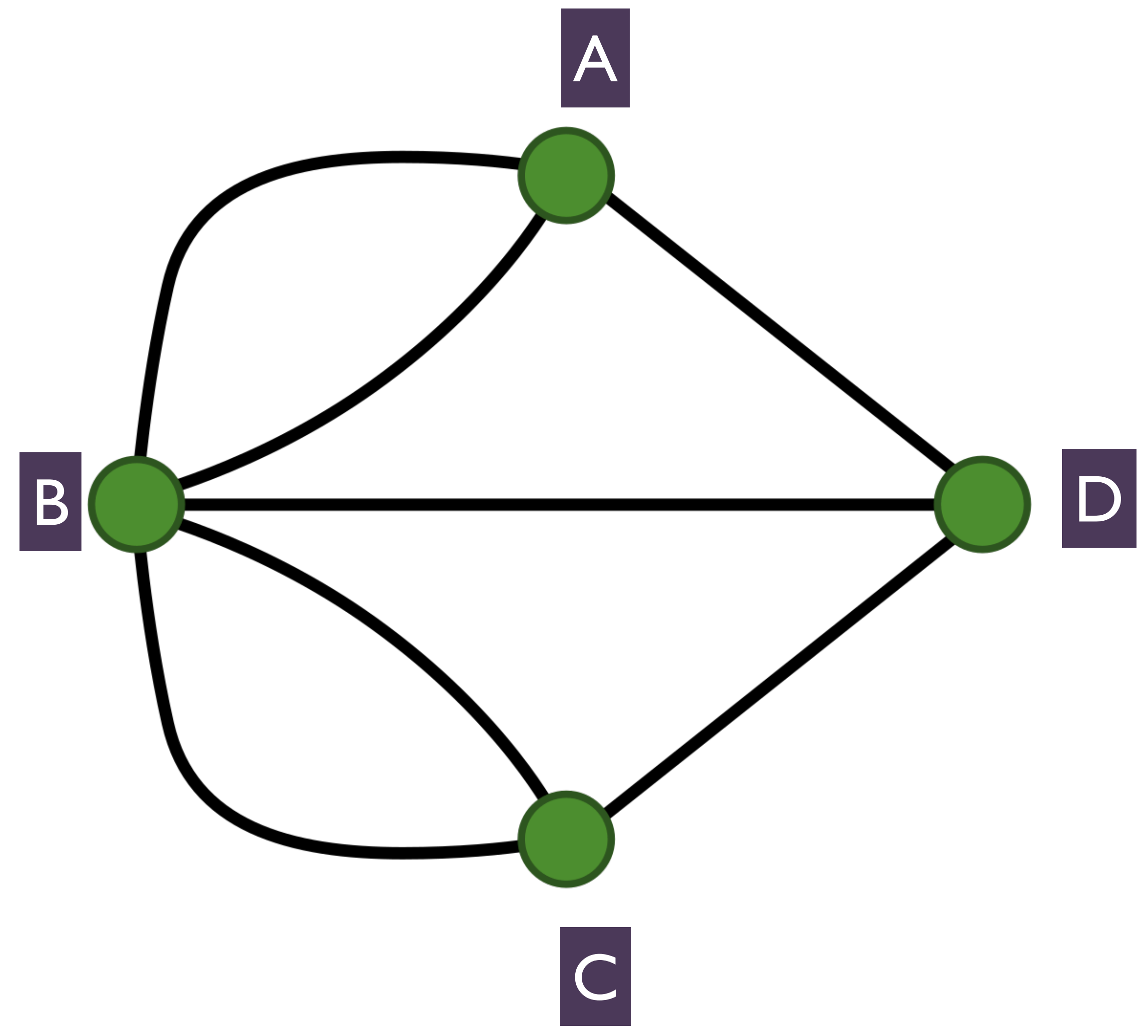


Can you draw these patterns

- without taking your pen off the paper, and
- without crossing any path twice?

EULER'S SOLUTION IN 1736





Each land mass can be viewed as a "vertex" and each bridge as a "link".

Only <u>terminal</u> vertices can have an odd number of links.

GRAPHS AND NETWORKS

Euler's work laid the foundation for the field of **graph theory**.

Any network of connections between entities can be analysed by viewing it as a graph that describes the manner in which a set of objects are connected.

Conversely, a network can simply be thought of as a graph where the objects and relations can be mapped to some real world setting.

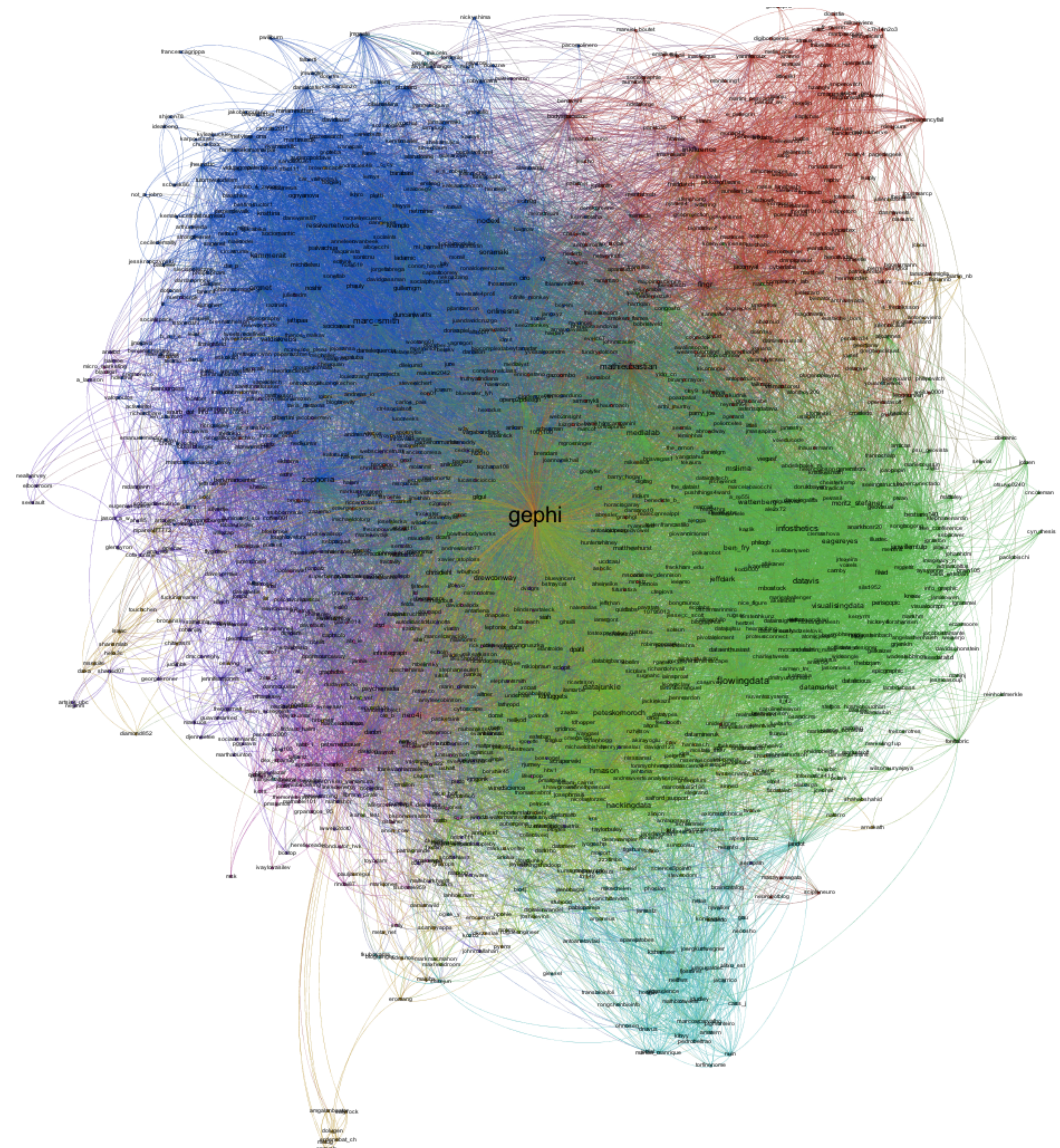
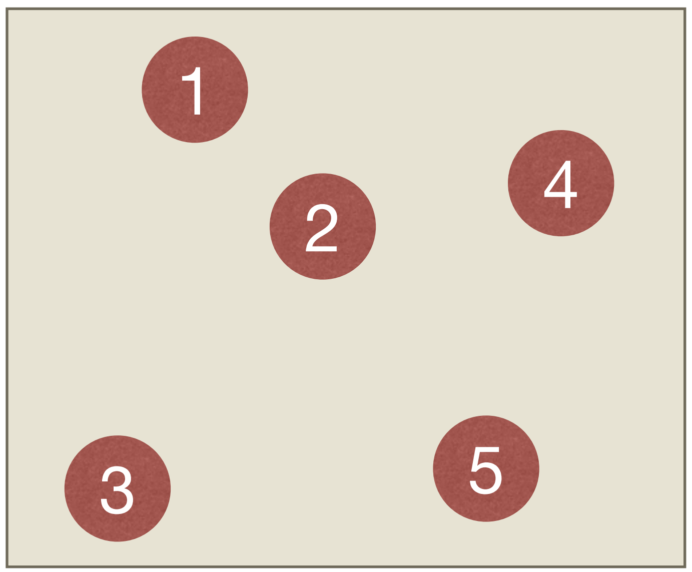
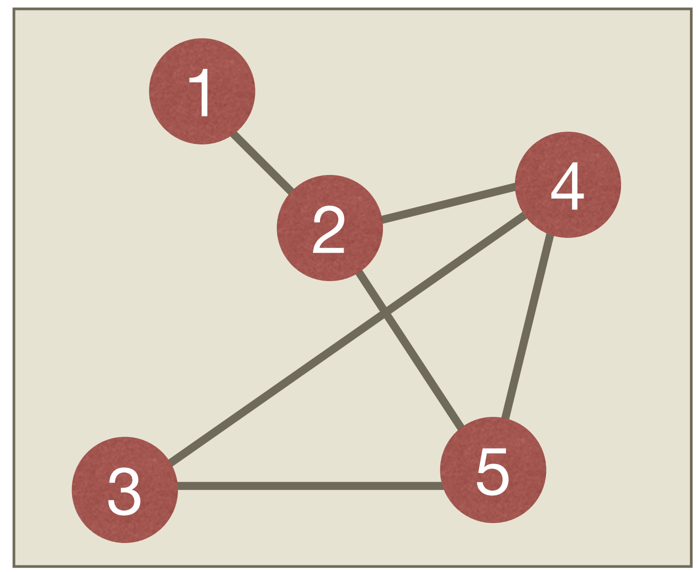


image: http://social-dynamics.org/a-gephi-visualization-of-gephi-on-twitter/

FUNDAMENTAL CONCEPTS: NODES AND LINKS



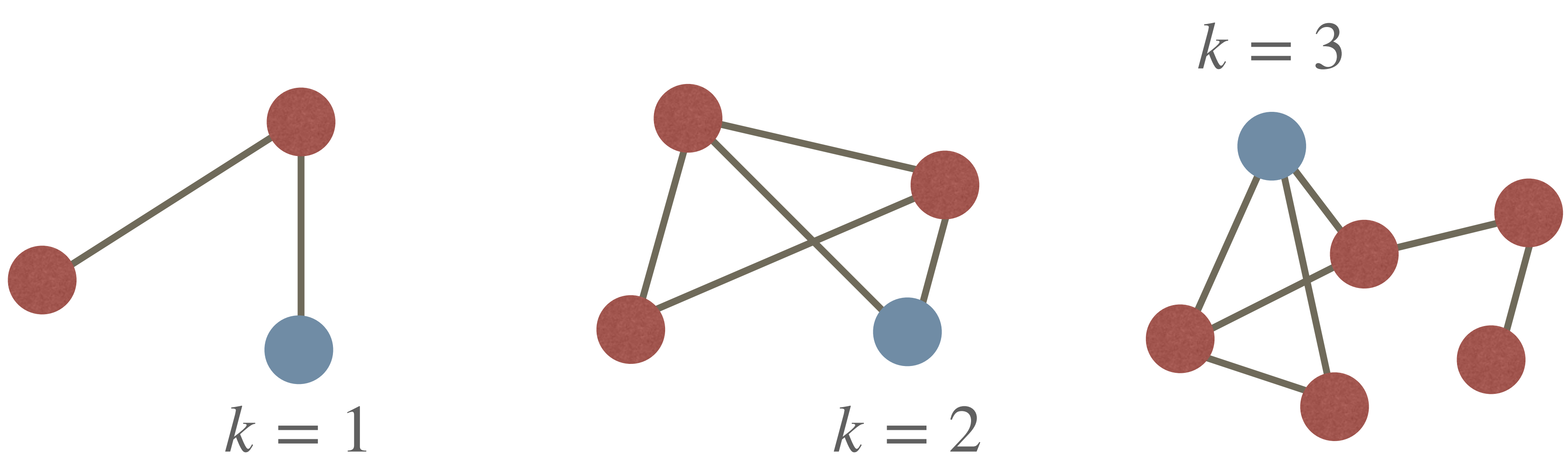
Nodes (or "Vertices")



Links (or "Edges")

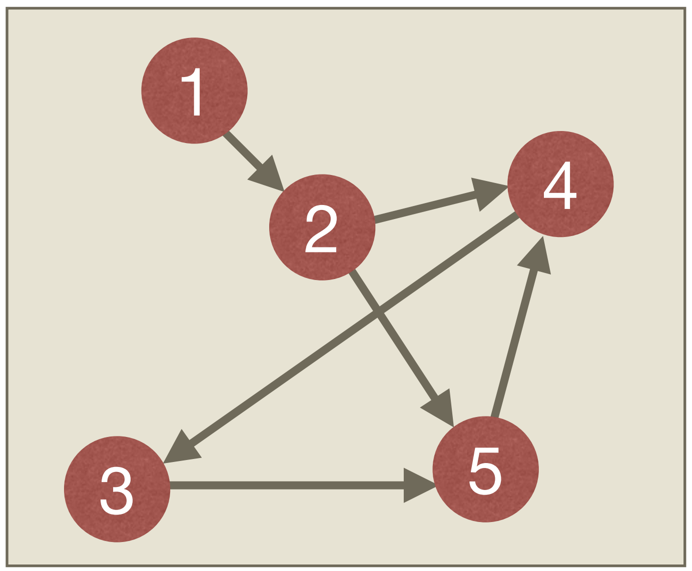
NODE	DEGREE
1	1
2	3
3	2
4	3
5	3

The total number of links associated with a node is its **degree** (k).

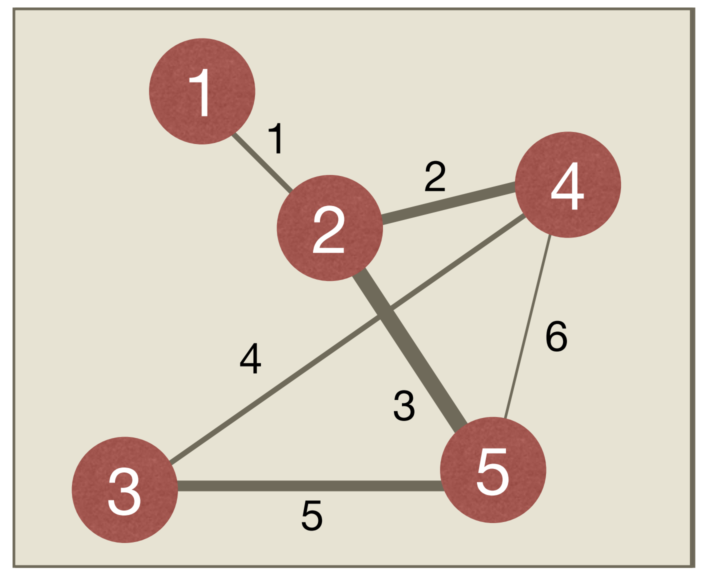


FUNDAMENTAL CONCEPTS: DIRECTED AND WEIGHTED GRAPHS

Directed graph



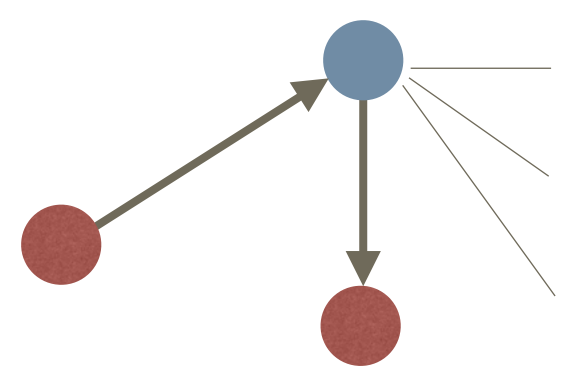
Weighted graph



EDGE	WEIGHT
1	1
2	2
3	3
4	1
5	2
6	0.5

NODE	IN-DEGREE	OUT-DEGREE	TOTAL DEGREE
1	0	1	1
2	1	2	3
3	1	1	2
4	2	1	3
5	2	1	3

In a <u>directed</u> graph a node's in-degree can be different to its out-degree



$k^{in} = 1$

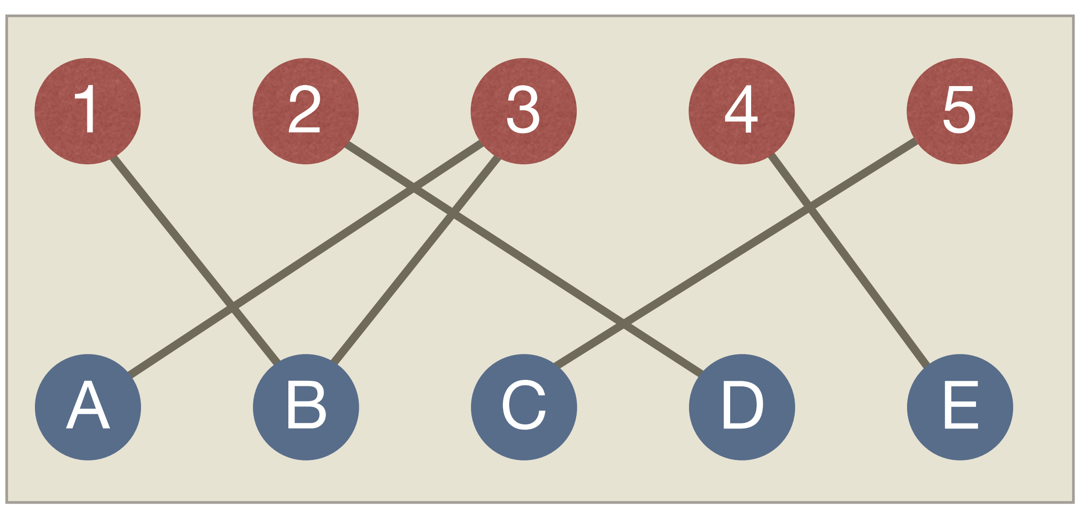
$k^{out} = 1$

$k^{total} = 2$

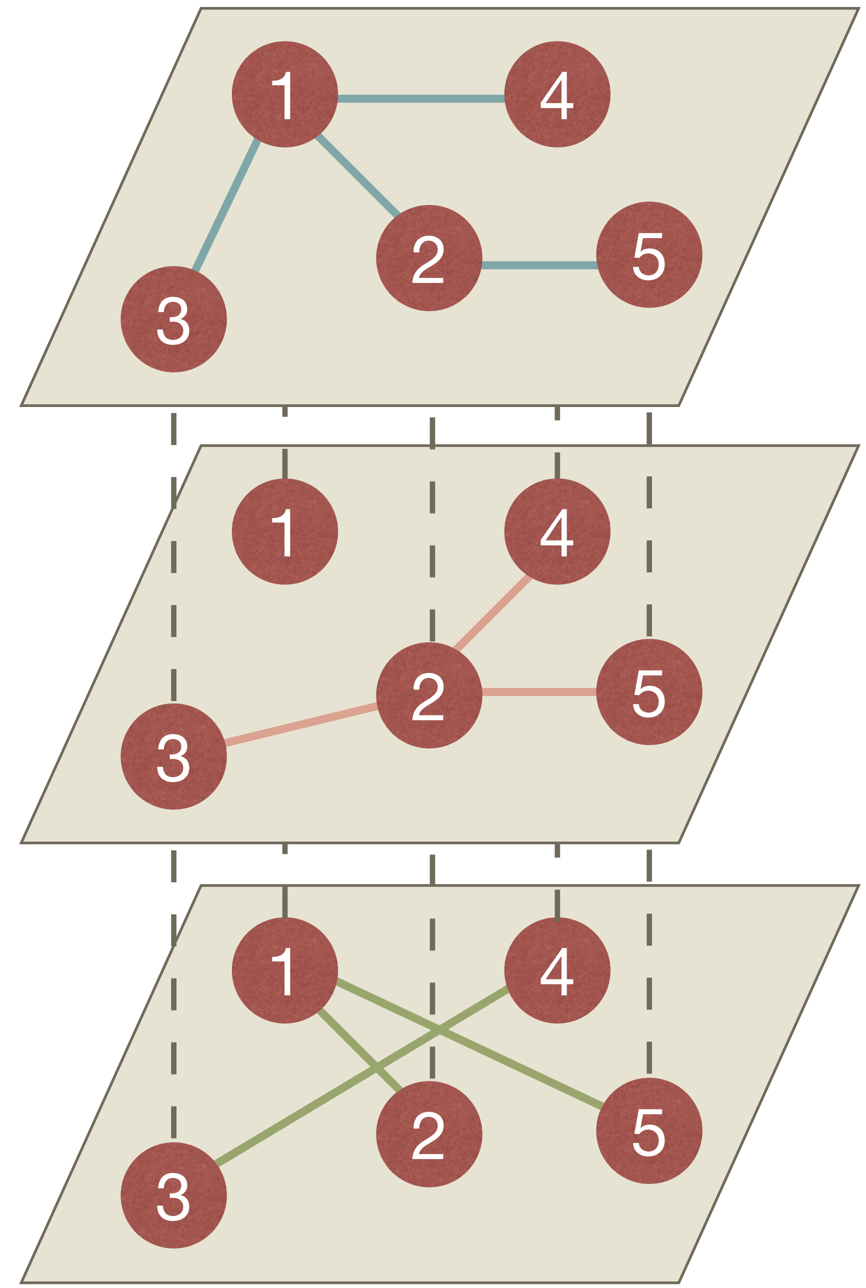
SOME OTHER TYPES OF GRAPHS

Graphs that describe relations between two different classes of objects are known as Bipartite graphs.

Graphs in which there may be different types of links between nodes are known as Multiplex graphs.

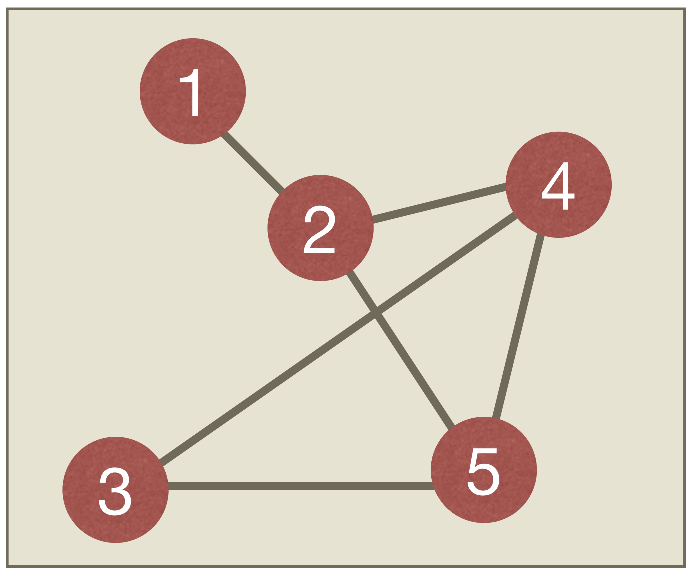


Bipartite graphs



Multiplex graph

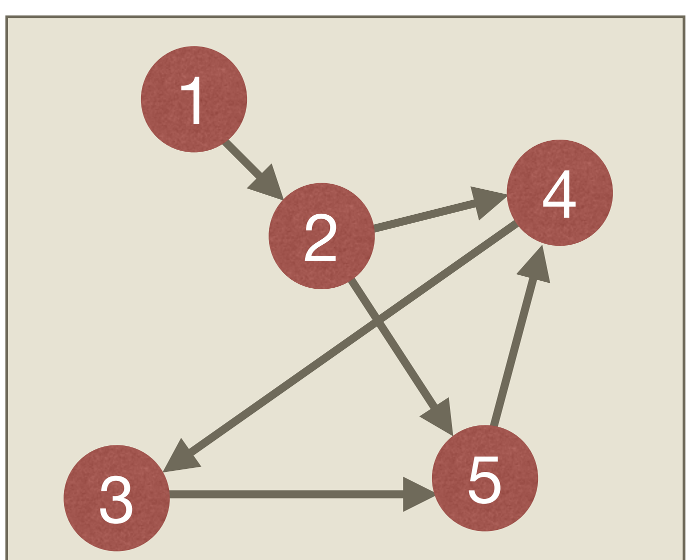
FUNDAMENTAL CONCEPTS: ADJACENCY MATRIX



Adjacency matrix

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	0	1	1
3	0	0	0	1	1
4	0	1	1	0	1
5	0	1	1	1	0

The Adjacency matrix $\mathbf{A}$ specifies all connections in the graph. If nodes i and j are connected then $A_{ij} = 1$ else $A_{ij} = 0$.



target →

source ↓

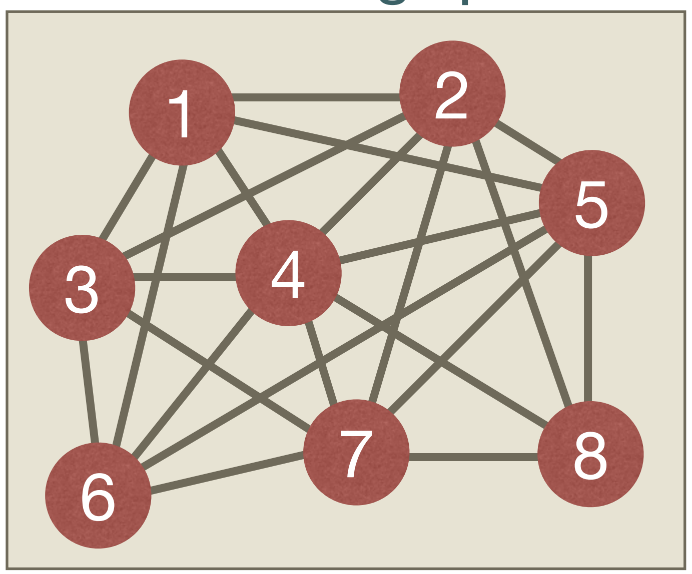
	1	2	3	4	5
1	0	1	0	0	0
2	0	0	0	1	1
3	0	0	0	0	1
4	0	0	1	0	0
5	0	0	0	1	0

In an undirected graph, the degree k_i of a node i can be obtained via:

$$k_i = \sum_i A_{ij} = \sum_j A_{ij}$$

FUNDAMENTAL CONCEPTS: DENSITY AND SPARSITY

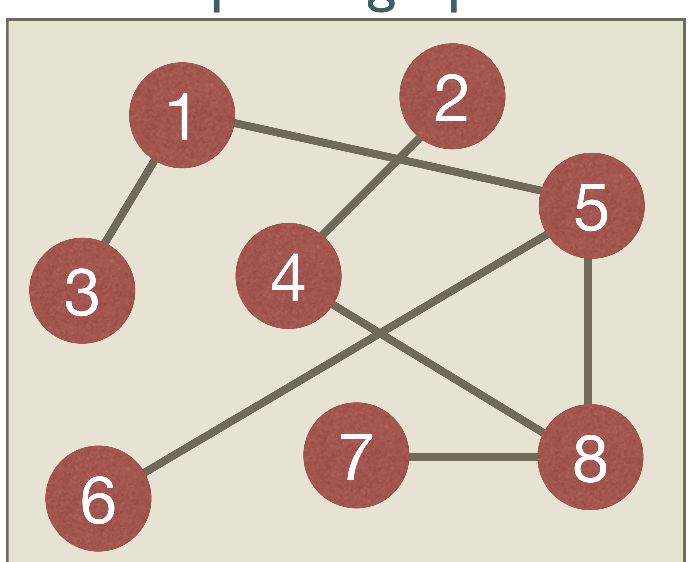
Dense graph



	1	2	3	4	5	6	7	8
1	0	1	1	1	1	1	0	0
2	1	0	1	1	1	0	1	1
3	1	1	0	1	0	1	1	0
4	1	1	1	0	1	1	1	1
5	1	1	0	1	0	1	1	1
6	1	0	1	1	1	0	1	0
7	0	1	1	1	1	1	0	1
8	0	1	0	1	1	0	1	0

The density ρ is the fraction of connected node pairs that exist in the graph.

Sparse graph

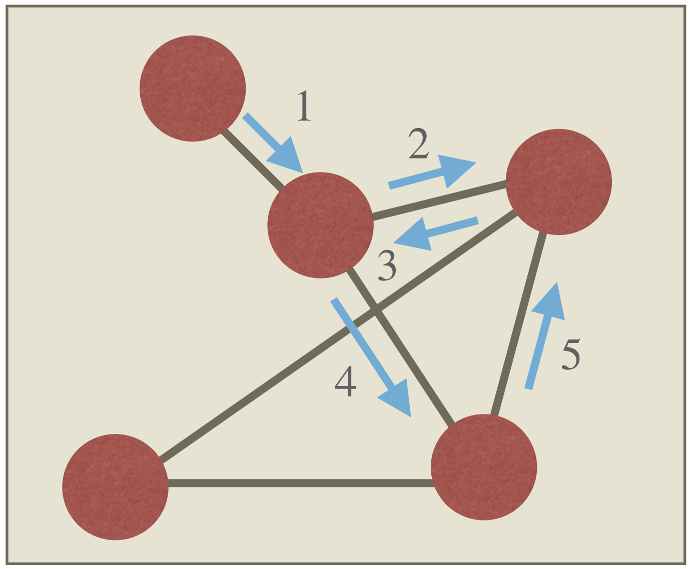


	1	2	3	4	5	6	7	8
1	0	0	1	0	1	0	0	0
2	0	0	0	1	0	0	0	0
3	1	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	1
5	1	0	0	0	0	1	0	1
6	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	0	1
8	0	0	0	1	1	0	1	0

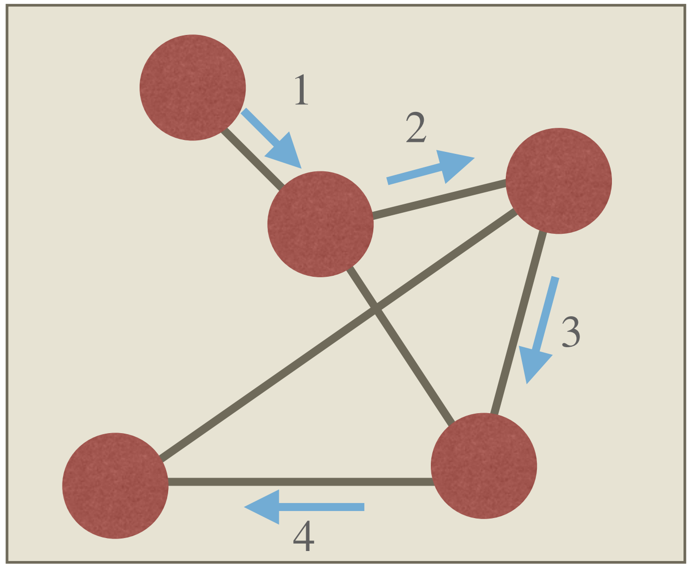
A graph is said to be dense if "most" of the possible links are present, and sparse if "most" are absent.

FUNDAMENTAL CONCEPTS: WALKS AND PATHS

Walk



Path

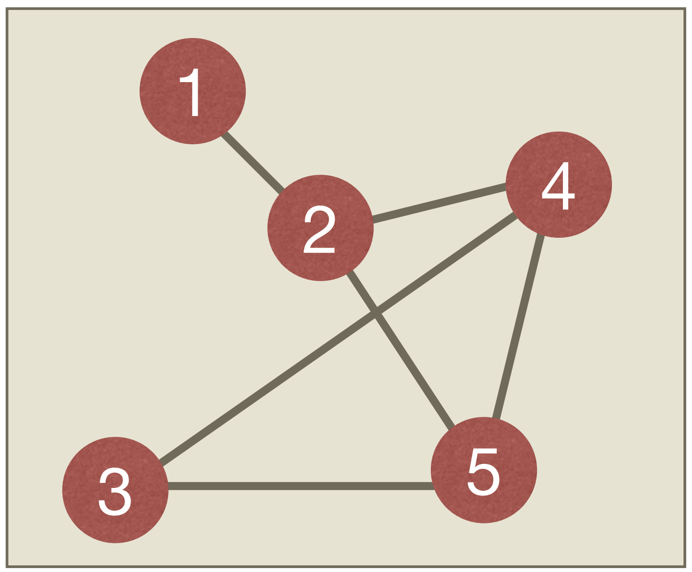


A walk is a route along the edges of a graph. In an undirected graph, an edge can be crossed in either direction.

The length of a walk is the number of hops taken along the route.

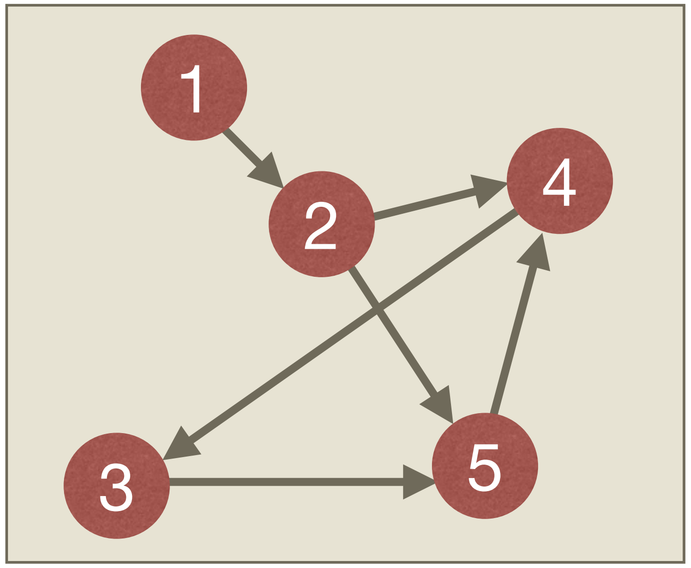
A path is a self-avoiding walk, i.e. one in which no edge is traversed twice.

FUNDAMENTAL CONCEPTS: SHORTEST PATH LENGTH



i-j	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
d_{ij}	1	3	2	2	2	1	1	1	1	1

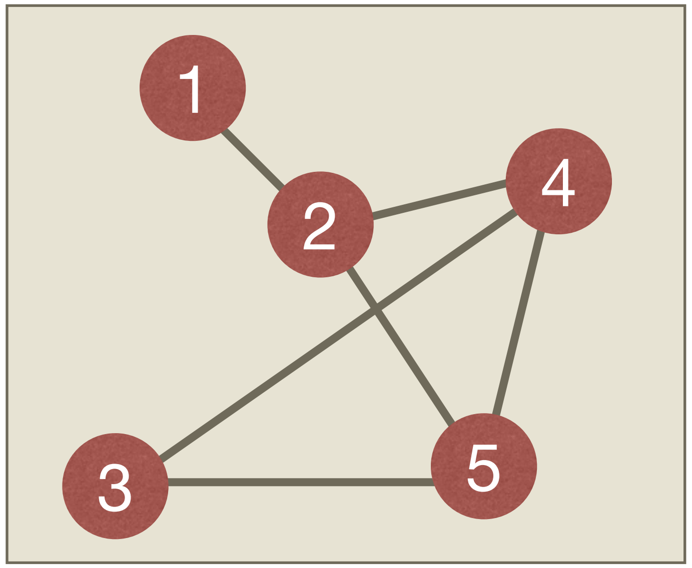
The shortest path length d_{ij} between two nodes i and j is the minimum number of links one has to cross to travel between them.



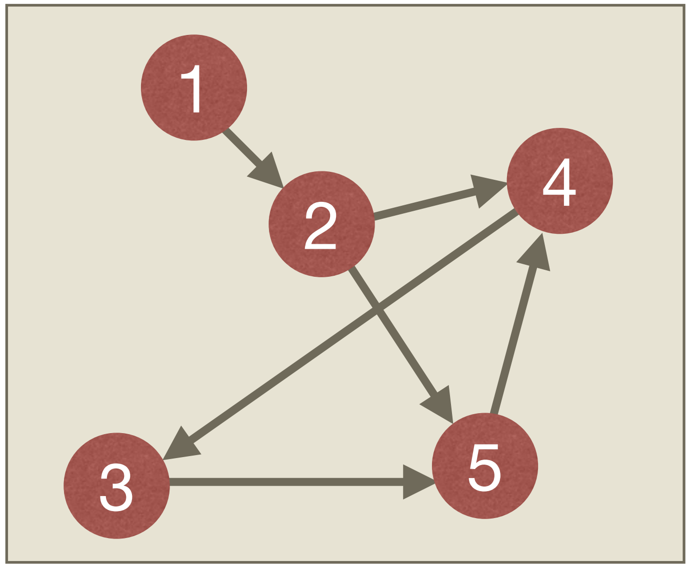
Can you work out what the shortest path length is between every pair of nodes of this directed graph?

FUNDAMENTAL CONCEPTS: DIAMETER

$d_{\max} = 3$



$d_{\max} = ?$



The diameter $d_{\max}$ of a network is the "longest shortest path" between all pairs of nodes i and j in the graph : $\max_{(i,j)}(d_{ij})$.

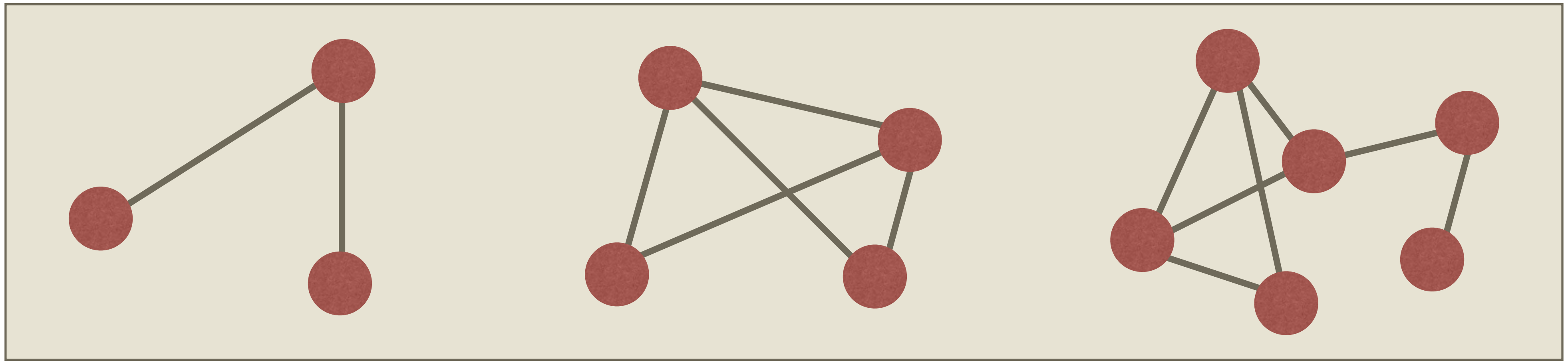
Can you work out the diameter of this directed graph?

FUNDAMENTAL CONCEPTS: AVERAGE PATH LENGTH

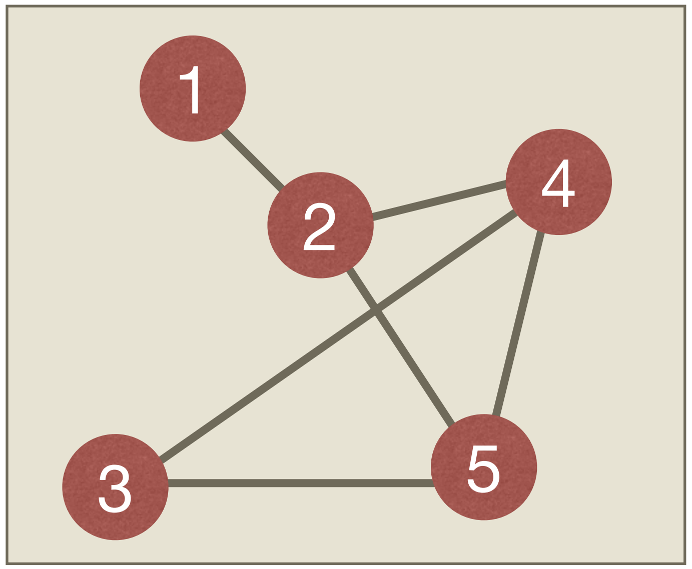
- The average path length is the average of the shortest path lengths between every pair of nodes in the graph.
- For a graph comprising N nodes, if $d(i, j)$ is the shortest number of steps between nodes i and j, then the average path length is:

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

What is the average path length of these graphs?



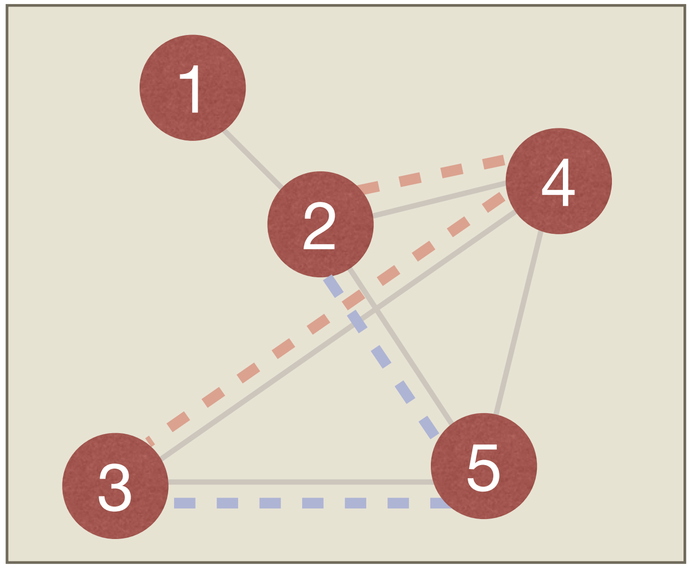
MORE ON PATH LENGTHS: TOTAL NUMBER OF WALKS OF A GIVEN LENGTH



#walks of length 1

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	0	1	1
3	0	0	0	1	1
4	0	1	1	0	1
5	0	1	1	1	0

The total number of walks $N_{ij}^{(1)}$ of length 1 between a pair of nodes (i, j) is just A_{ij}



#walks of length 2

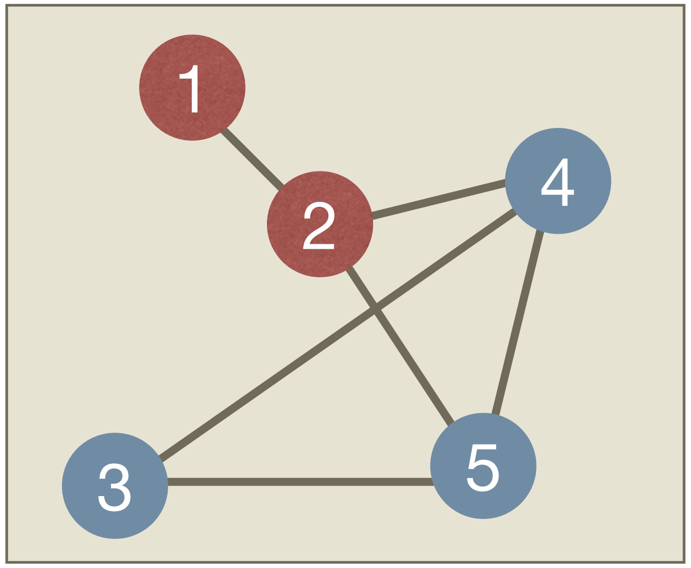
	1	2	3	4	5
1	1	0	0	1	1
2	0	3	2	1	1
3	0	2	2	1	1
4	1	1	1	3	2
5	1	1	1	2	3

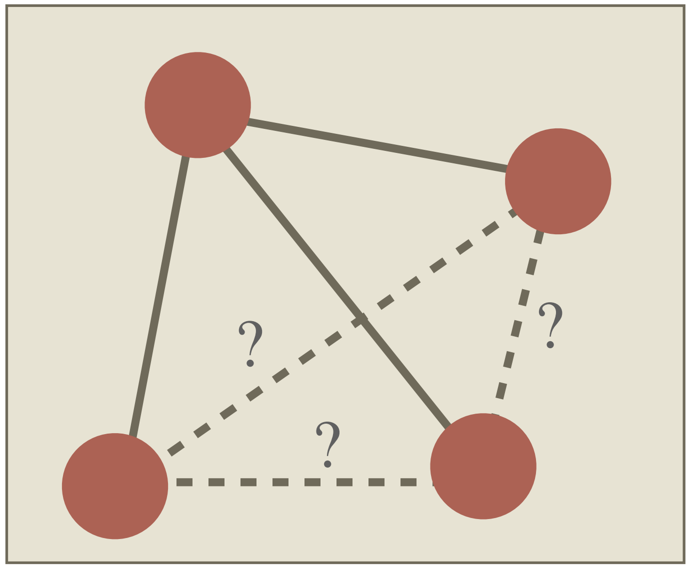
One can have multiple walks $N_{ij}^{(2)}$ of length 2 between a pair of nodes (i, j).

FUNDAMENTAL CONCEPTS: LOCAL CLUSTERING COEFFICIENT

- In real networks, one often finds that nodes that form links with one another also form links with those that the neighbour link to.
- The (local) clustering coefficient of a node measures the extent of connectivity of its local neighbourhood, i.e. how close they are to being a "clique" or a complete subgraph.
- If a node i in an undirected graph has k_i neighbours, there can be a maximum of $k_i(k_i - 1)/2$ links between them.
- The local clustering coefficient C_i of node i is the fraction of these links that exist.

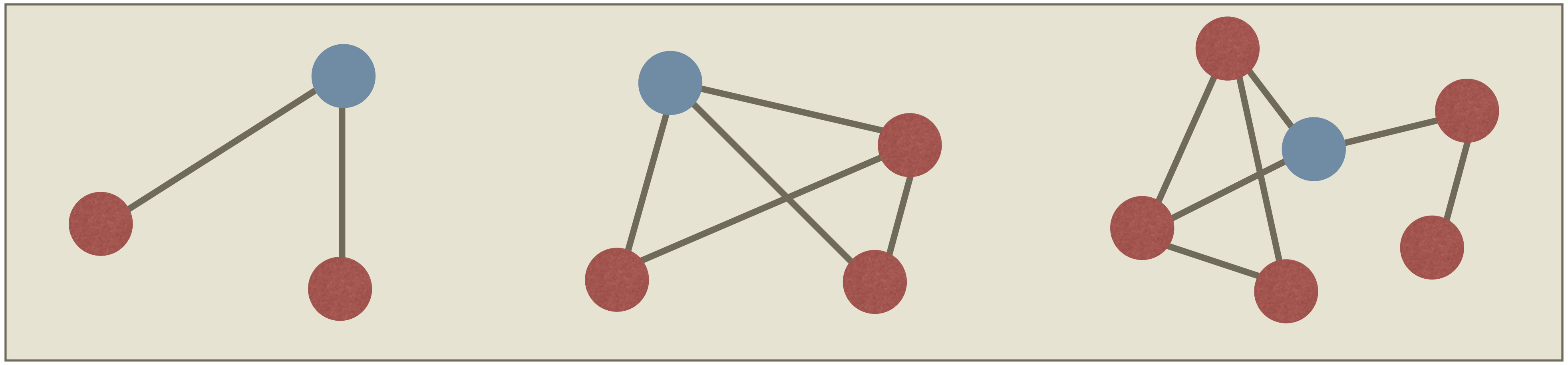
blue nodes form a clique



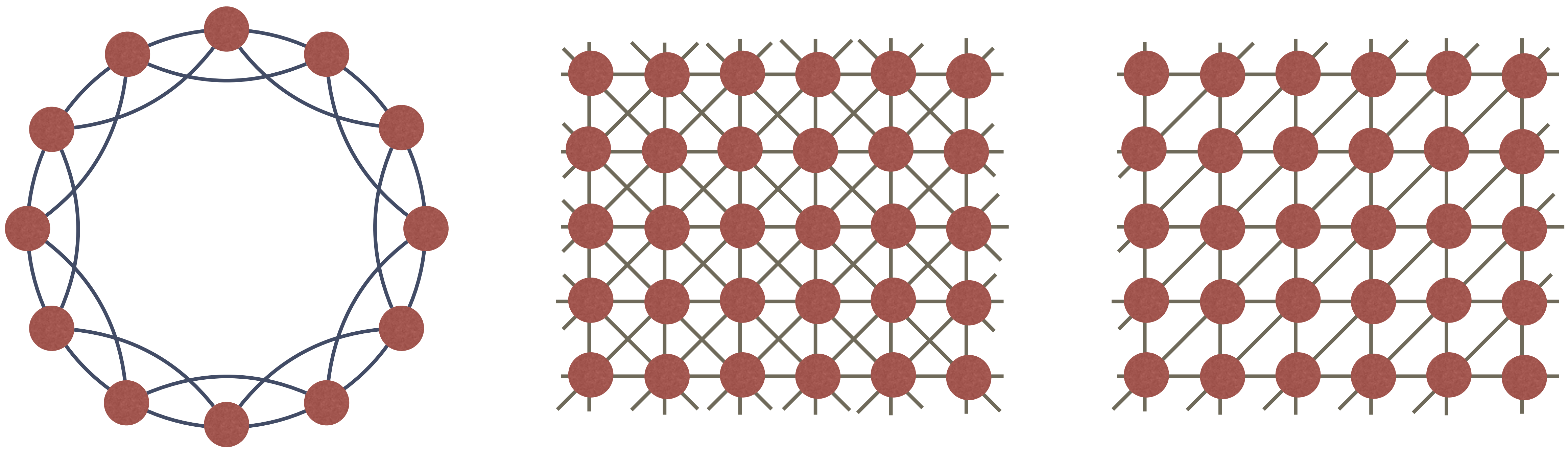


FUNDAMENTAL CONCEPTS: CLUSTERING COEFFICIENT

What is the clustering coefficient of the blue nodes?

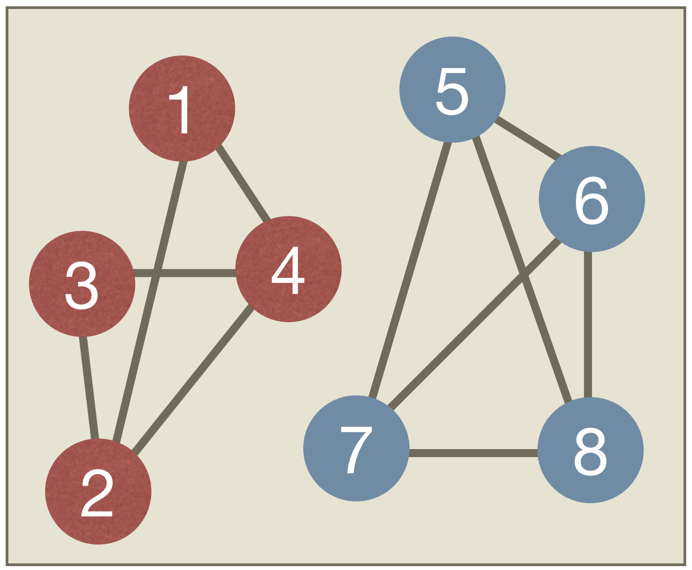


Calculate the clustering coefficients for a node in the following graphs:



$k_i = 4$ $k_i = 8$ $k_i = 6$

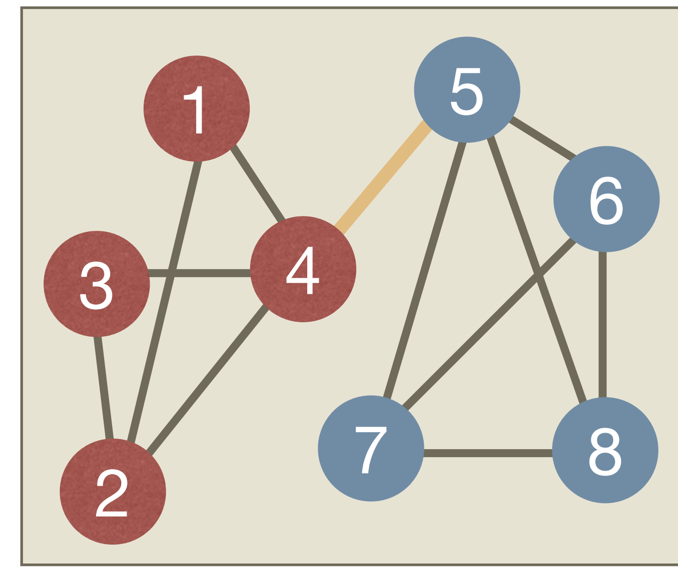
FUNDAMENTAL CONCEPTS: COMPONENTS



In an undirected network, a pair of nodes (i, j) are connected if there exists a path (of any length) between them.

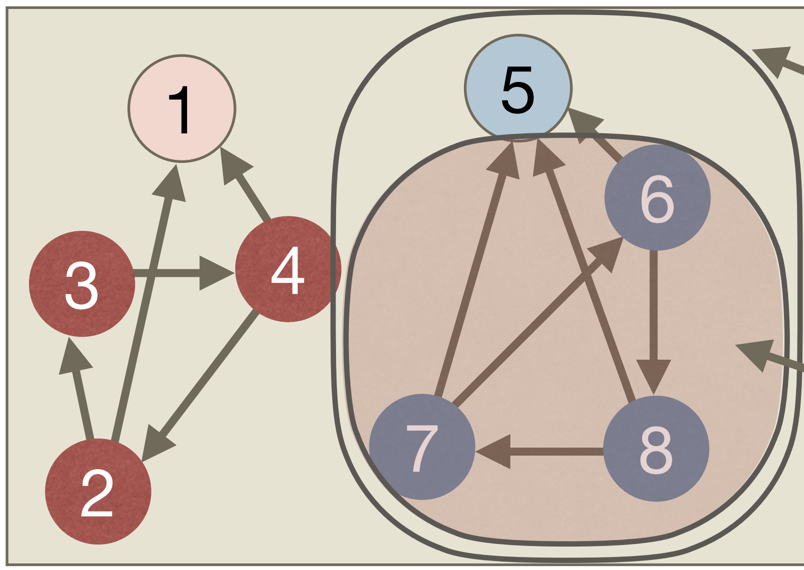
	1	2	3	4	5	6	7	8
1	0	1	0	1	0	0	0	0
2	1	0	1	1	0	0	0	0
3	0	1	0	1	0	0	0	0
4	1	1	1	0	0	0	0	0
5	0	0	0	0	0	1	1	1
6	0	0	0	0	1	0	1	1
7	0	0	0	0	1	1	0	1
8	0	0	0	0	1	1	1	0

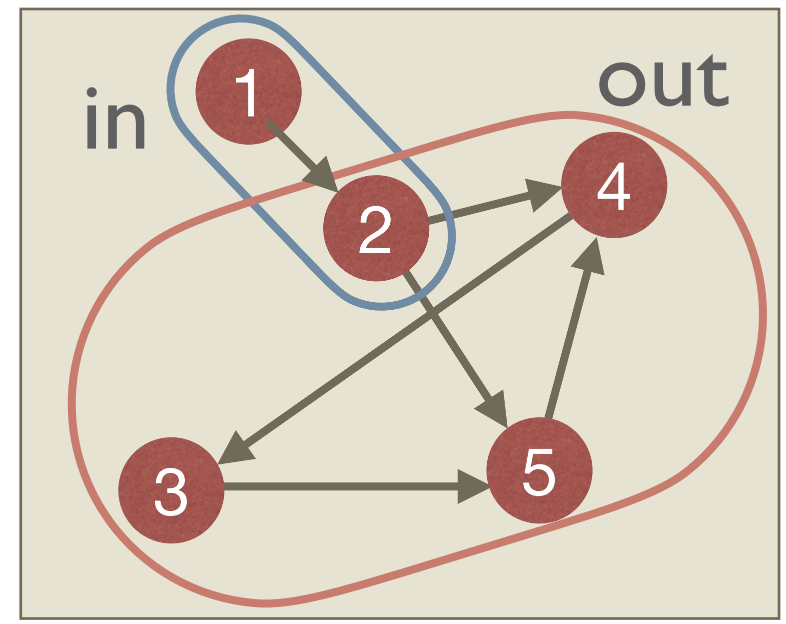
A component is a subset of the network in which all nodes are connected.



A bridge is a link that, when cut, causes the network to be disconnected.

FUNDAMENTAL CONCEPTS: COMPONENTS



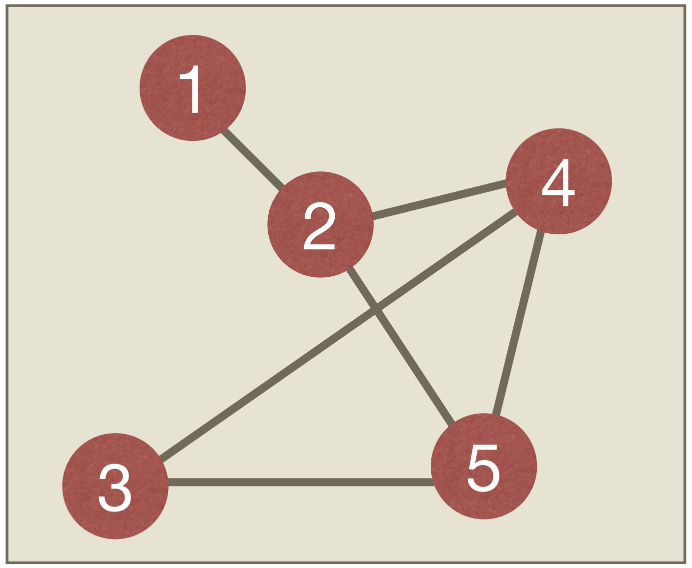


In an directed network, a strongly connected component is one where exists a path between all constituent nodes.

A weakly connected component is a connected component that exists if one were to ignore the directed nature of the edges.

The in-component of a node in a directed network is the set that can reach it, and its out-component is the set that can be reached from it.

FUNDAMENTAL CONCEPTS: BETWEENNESS CENTRALITY



On the right are the list of all possible shortest paths between every pair of nodes in the above network.

SHORTEST PATHS	
1-2	{1,2}
1-3	{1,2,5,3}, {1,2,4,3}
1-4	{1,2,4}
1-5	{1,2,5}
2-3	{2,4,3}, {2,5,3}
2-4	{2,4}
2-5	{2,5}
3-4	{3,4}
3-5	{3,5}
4-5	{4,5}

To find the betweenness centrality of a node, we count the fraction of times it appears in the shortest paths between other nodes.

FUNDAMENTAL CONCEPTS: BETWEENNESS CENTRALITY

	SHORTEST PATHS
1-2	{1,2}
1-3	{1,2,5,3}, {1,2,4,3}
1-4	{1,2,4}
1-5	{1,2,5}
2-3	{2,4,3}, {2,5,3}
2-4	{2,4}
2-5	{2,5}
3-4	{3,4}
3-5	{3,5}
4-5	{4,5}

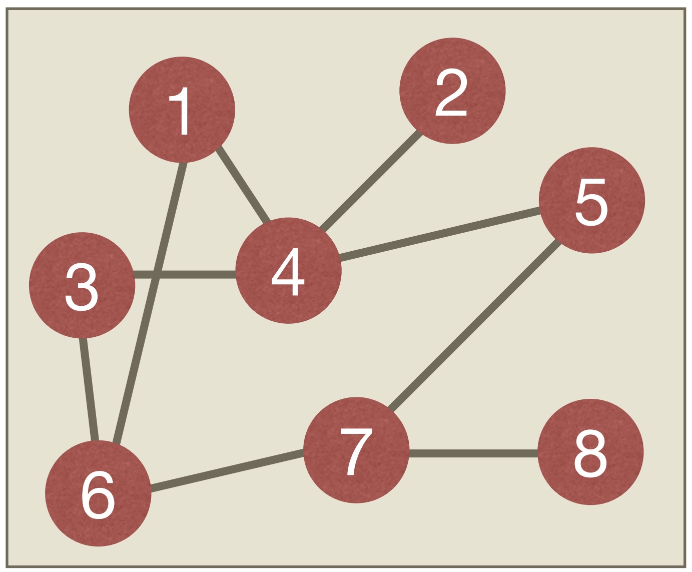
	OCCURRENCES	C_B
1	0	0
2	2/2 + 1 + 1	3
3	0	0
4	1/2 + 1/2	1
5	1/2 + 1/2	1

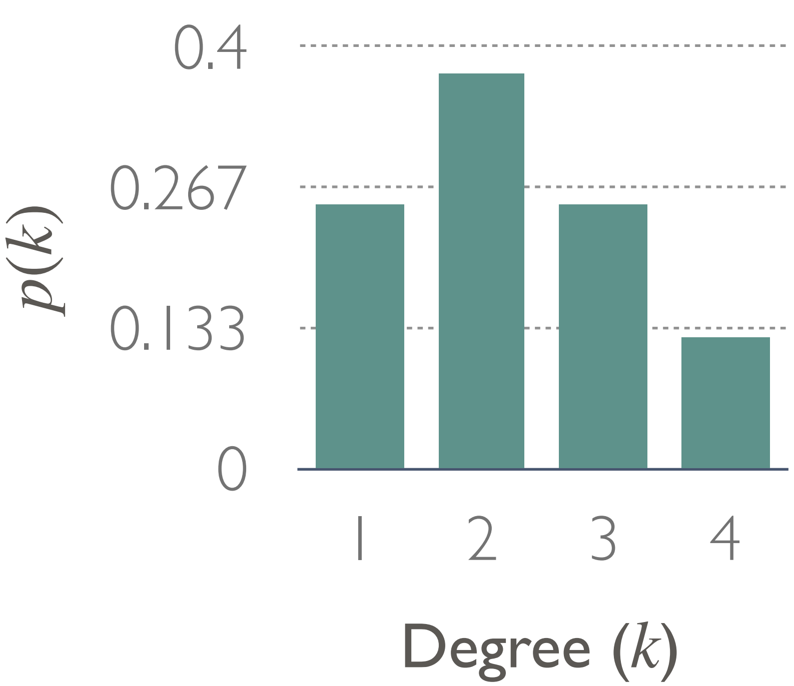
If σ_{st} is the no. of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of these containing node v, then:

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

Is the betweenness centrality.

FUNDAMENTAL CONCEPTS: DEGREE DISTRIBUTION





- The degree distribution $p(k)$ is the normalised histogram of the number of nodes of a given degree (N_k):

$$p(k) = N_k / N$$

- It can be interpreted as the probability that a randomly selected node in the network has degree k

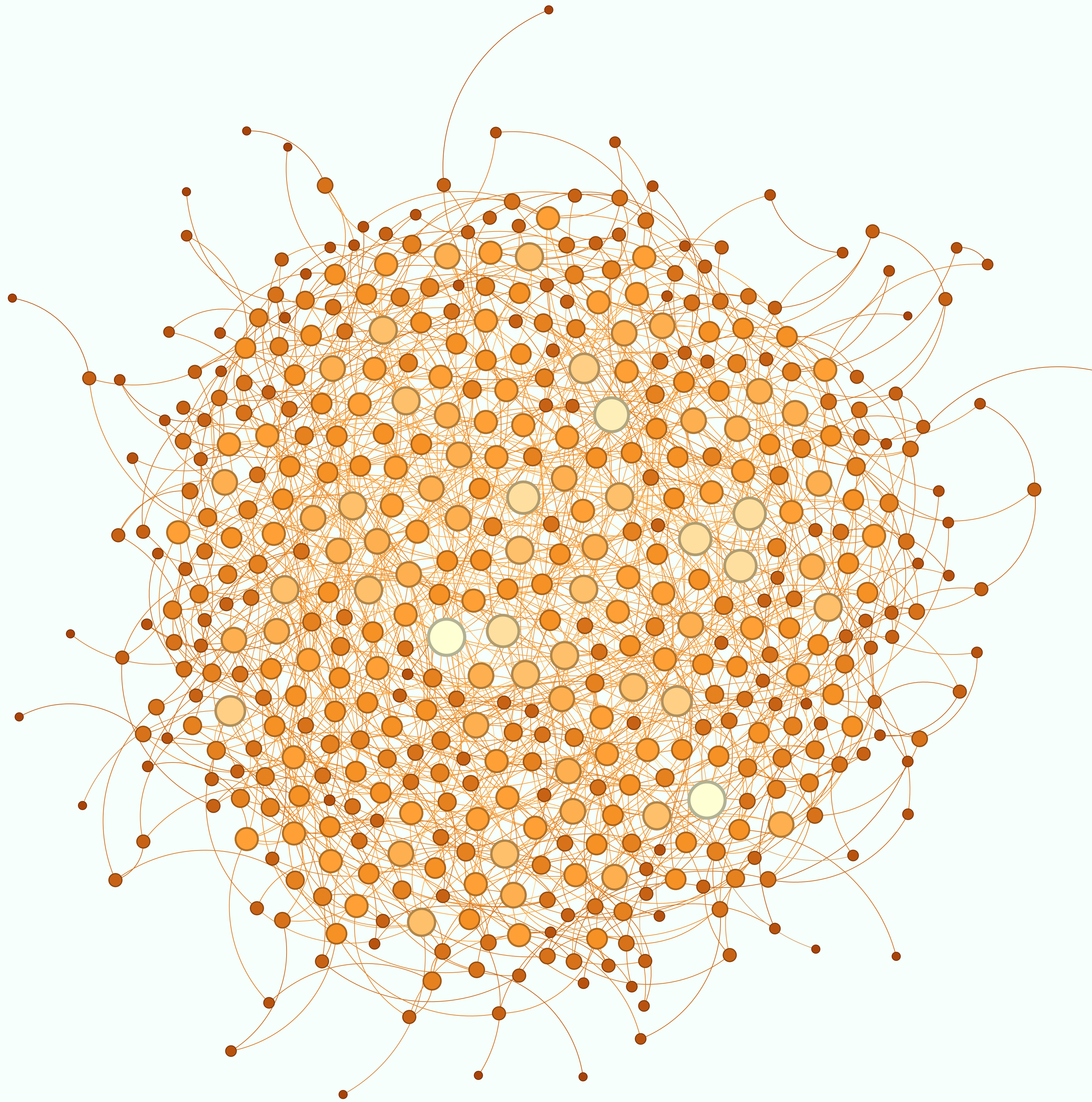
Average degree

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \sum_{k=0}^{\infty} k\, p(k)$$

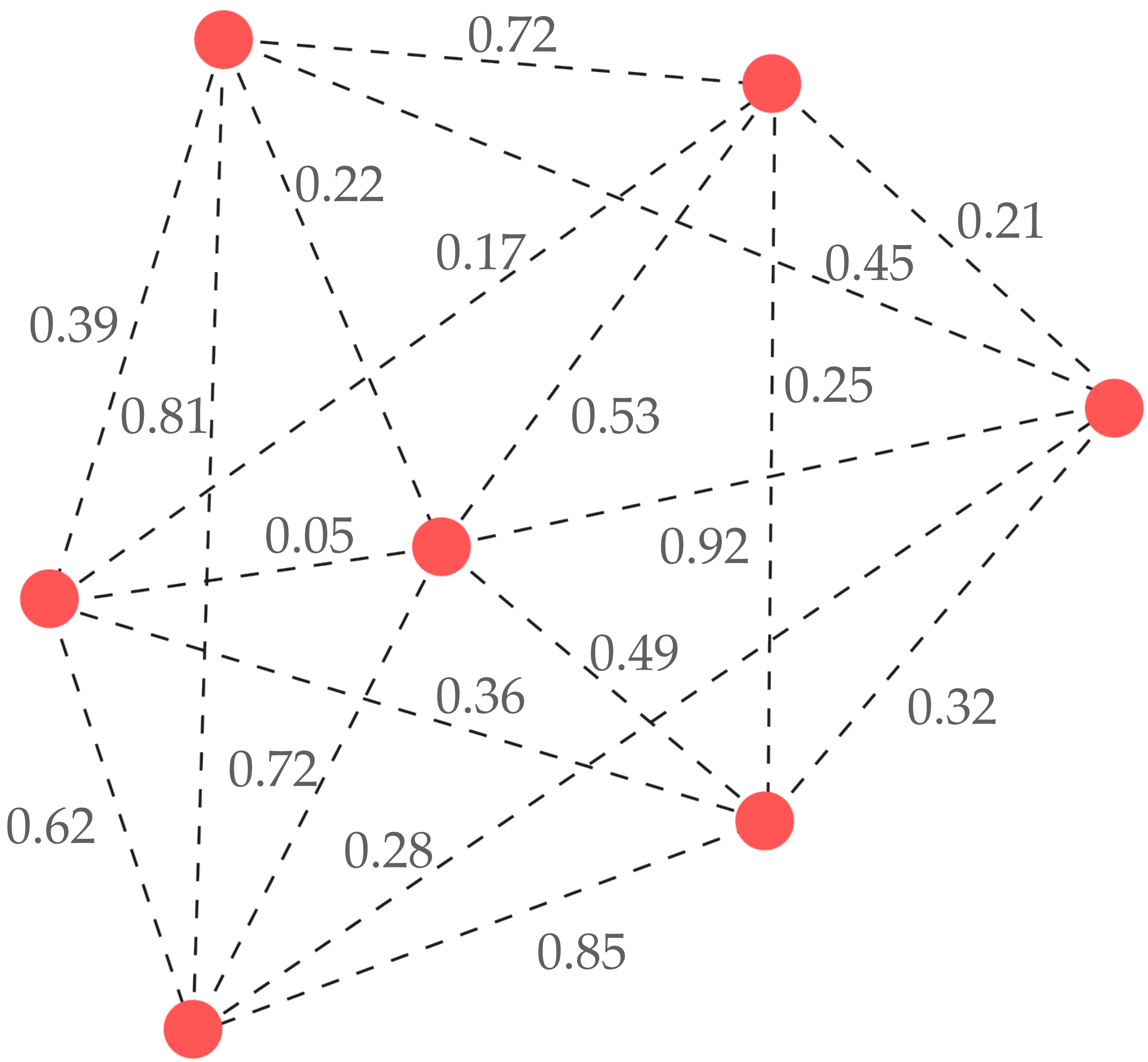
WHY STUDY RANDOM GRAPHS?

Random graphs provide *null models* against which we can test certain hypotheses.

Once the key attributes of the network responsible for certain properties have been identified it is then possible to generate numerous *surrogate* networks that can be used in place of the empirical network for further study.



GENERATING RANDOM GRAPHS



In 1959 a model was proposed for generating a random graph comprising n nodes.

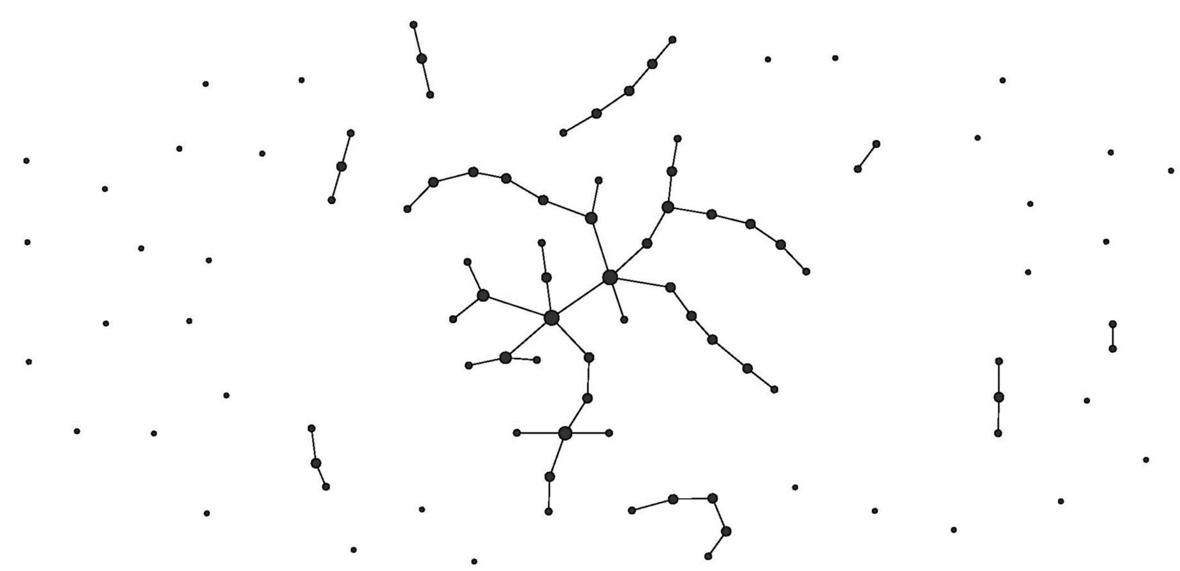
Assign a random number between [0,1] to every potential link.

Keep only those links whose values are less than a specified threshold $p \in [0,1]$.

ERDŐS-RÉNYI RANDOM GRAPHS

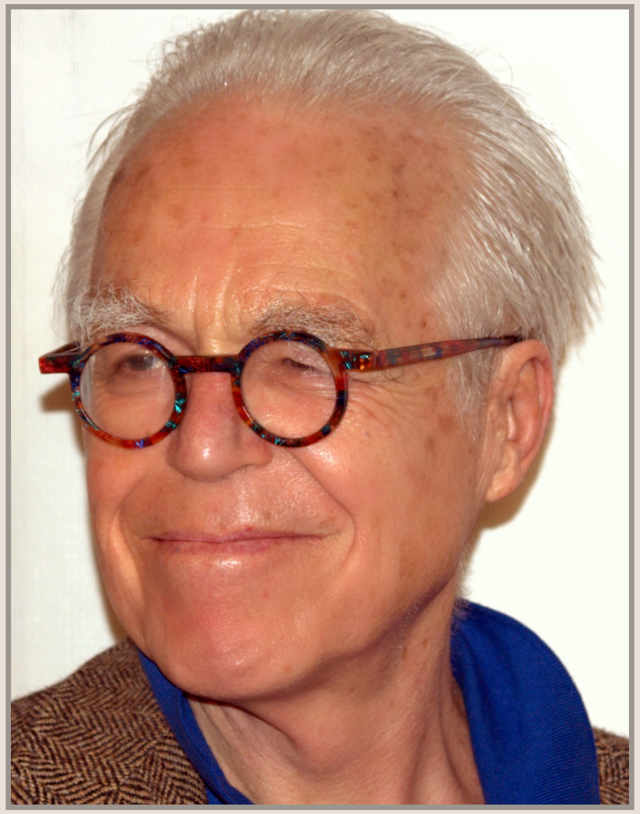
- This is known as the $G(n, p)$ model, and the resulting graphs are commonly referred to as Erdős-Rényi (ER) random graphs
- For certain choices of (n, p), the resulting graph may have multiple connected components.
- For large n, the number of links in the graph is approximately equal to $\dfrac{n(n-1)}{2}p$.

Pál Erdős Alfréd Rényi Edgar Gilbert



“The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Árpád Pásztor, someone I not only know, but is to the best of my knowledge a good friend of mine - so I could easily ask him to send a telegram via the general director telling Ford that he should talk to the manager and have the worker in the shop quickly hammer together a car for me, as I happen to need one.”

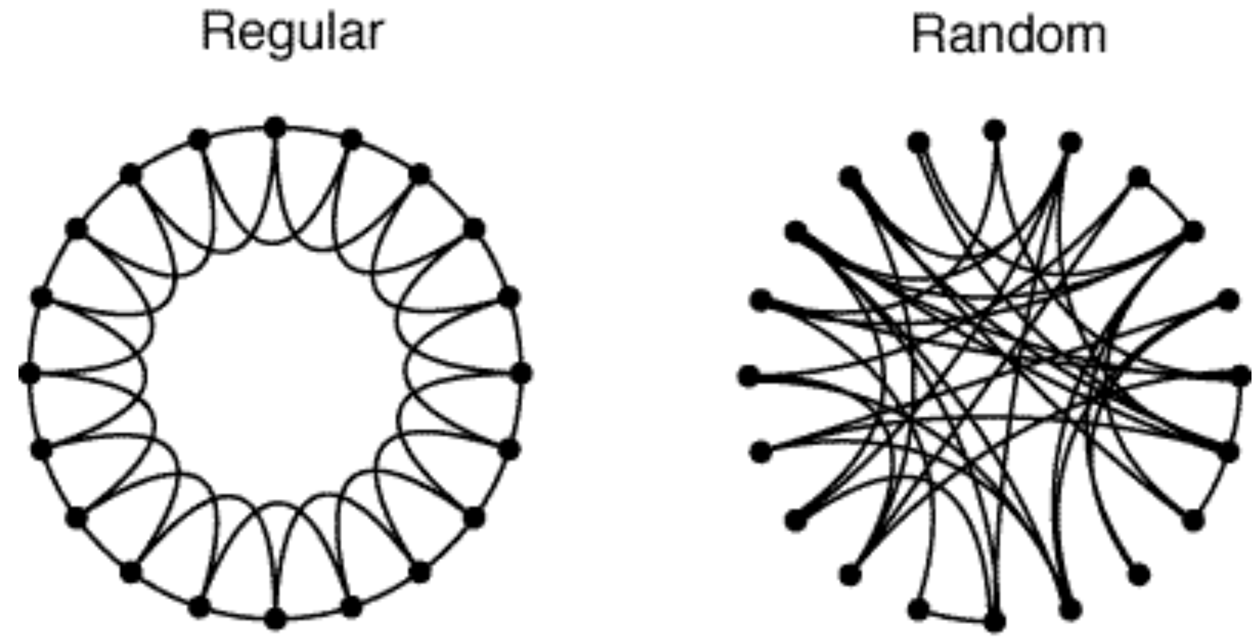
Frigyes Karinthy, “*Láncszemek (Chains)*” (1929).



“Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. Fill in the names.”

John Guare, “*Six Degrees of Separation*” (1990).

SMALL WORLD NETWORKS



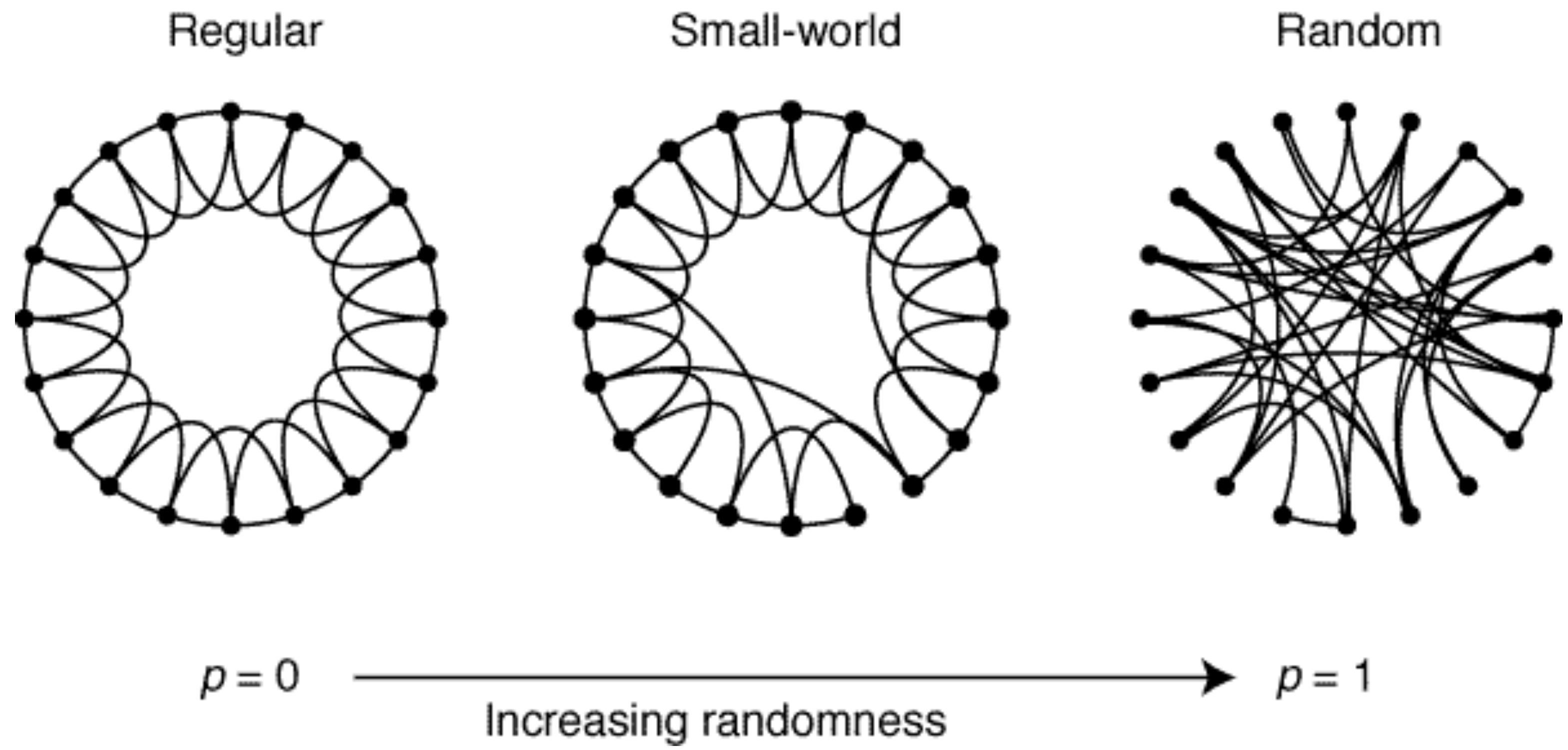
A regular graph has high clustering coefficient C and high average path length L. A random graph has the opposite properties.

Watts and Strogatz (1998) suggested a procedure for obtaining graphs with properties of both regular and random graphs:

- Start with a regular graph where each node has K neighbours.
- Cycle through each node, and consider the $K/2$ rightward links.
- Randomly rewire each of these links with probability p, avoiding self-loops and duplicate links.

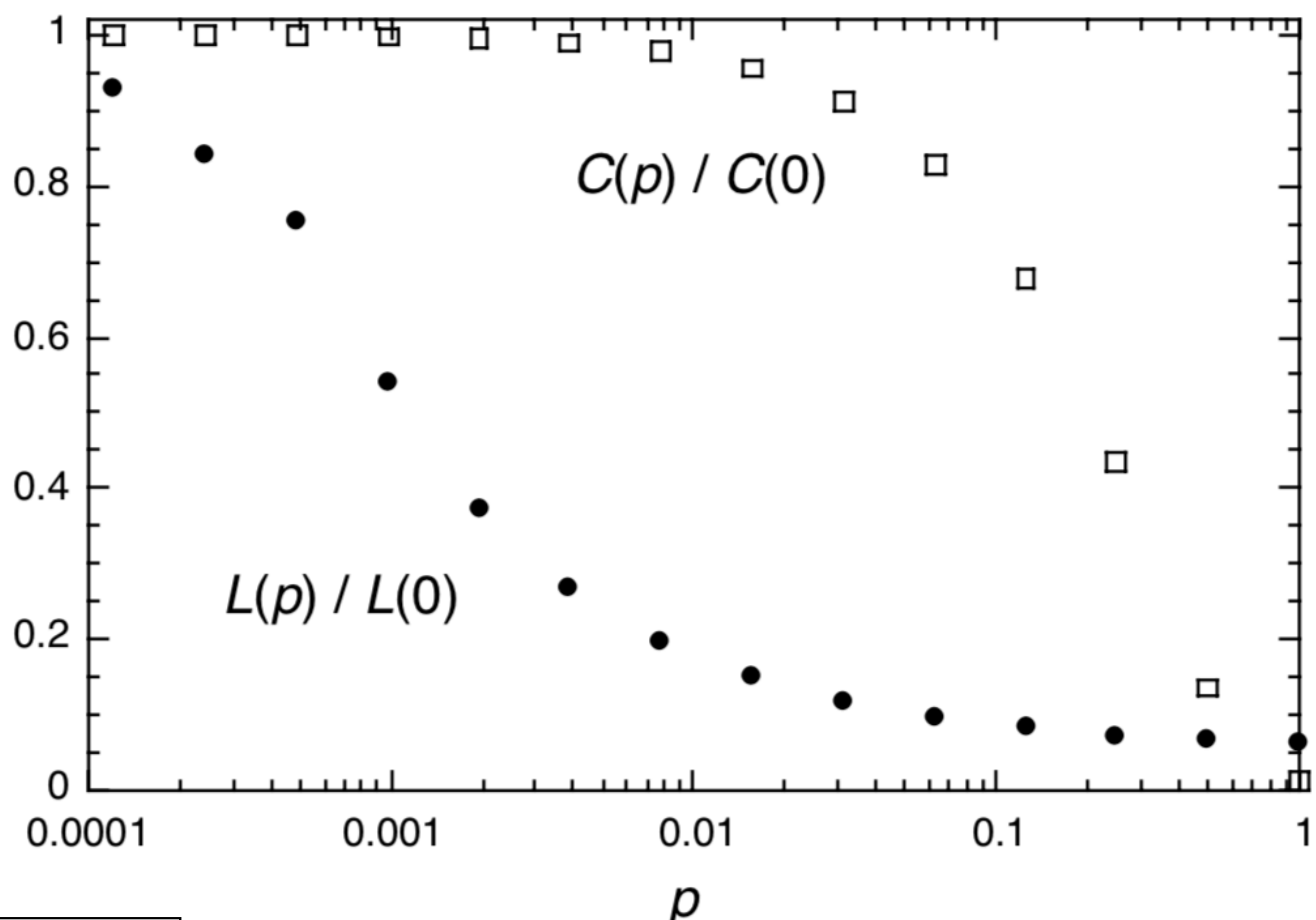
image: Watts, D. & Strogatz, S., *Nature* **393**, 440-442 (1998).

SMALL WORLD NETWORKS



For an intermediate p the graphs have a low average path length L and a high clustering coefficient C.

These are referred to as "small-world" networks.



images: Watts, D. & Strogatz, S., *Nature* **393**, 440-442 (1998).

Six degrees of Kevin Bacon

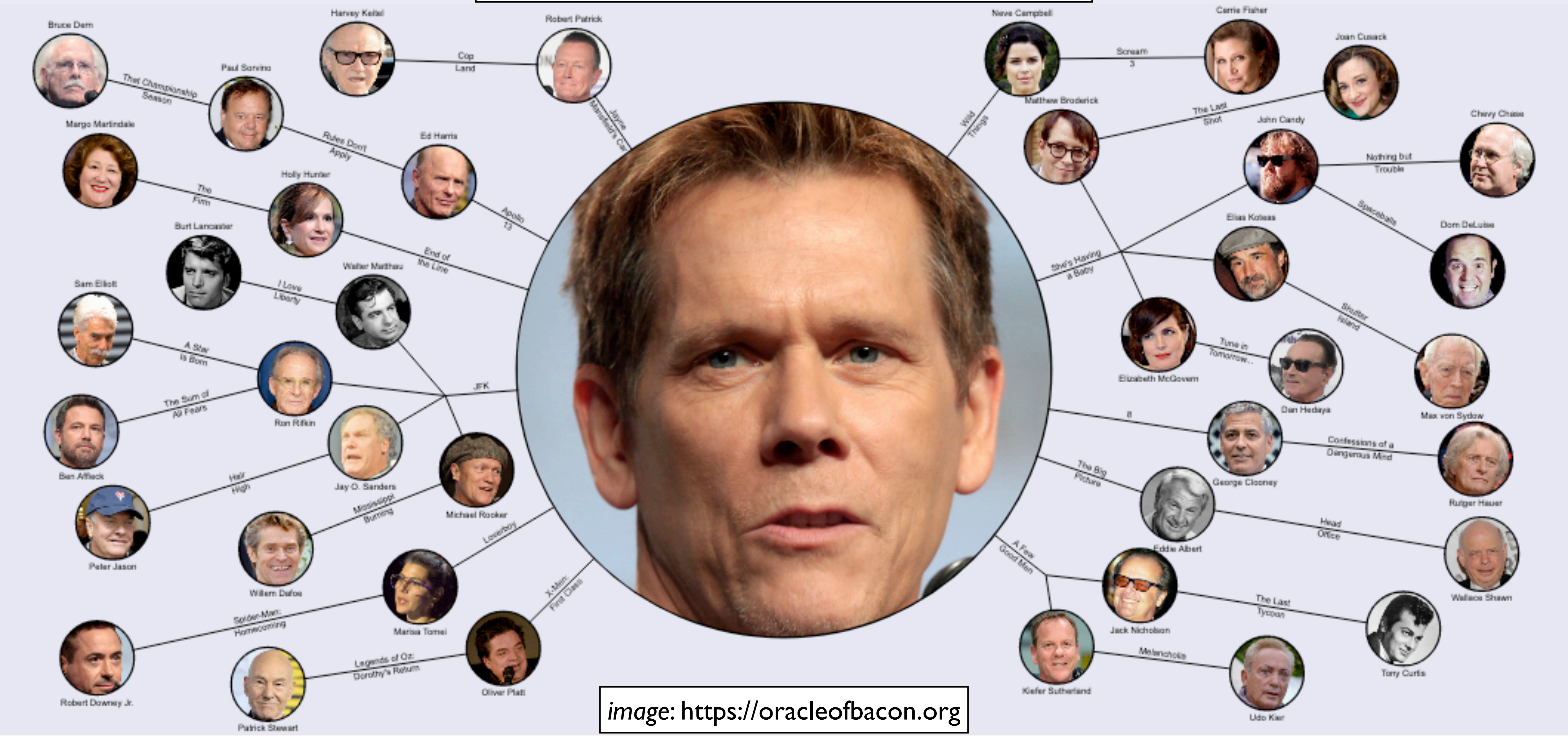
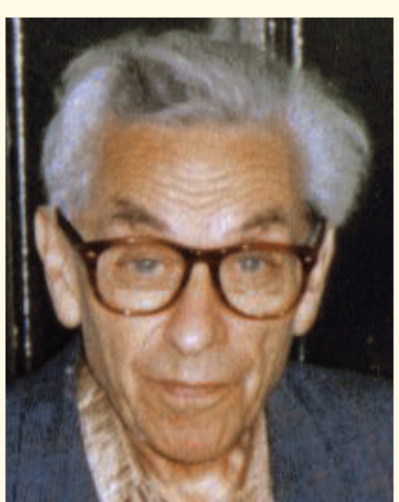
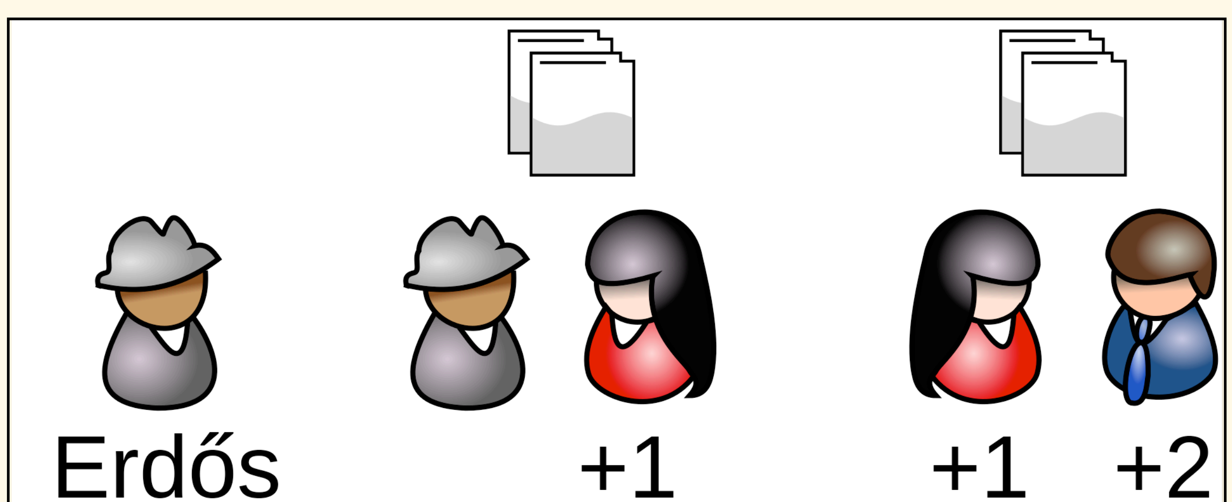


image: https://oracleofbacon.org





Erdős number

Collaborative "distance" between an author and Pál Erdős.

"RICH GET RICHER"

The Barabási-Albert (BA) model of random graphs employs a mechanism of growth and preferential attachment:

- Starting with m_0 nodes, at each step add a new node and connect it to $m(\leq m_0)$ existing nodes.
- The probability of connecting to an existing node i is

$$p(k_i) = \frac{k_i}{\Sigma_j k_j}$$

Albert-László Barabási

Réka Albert

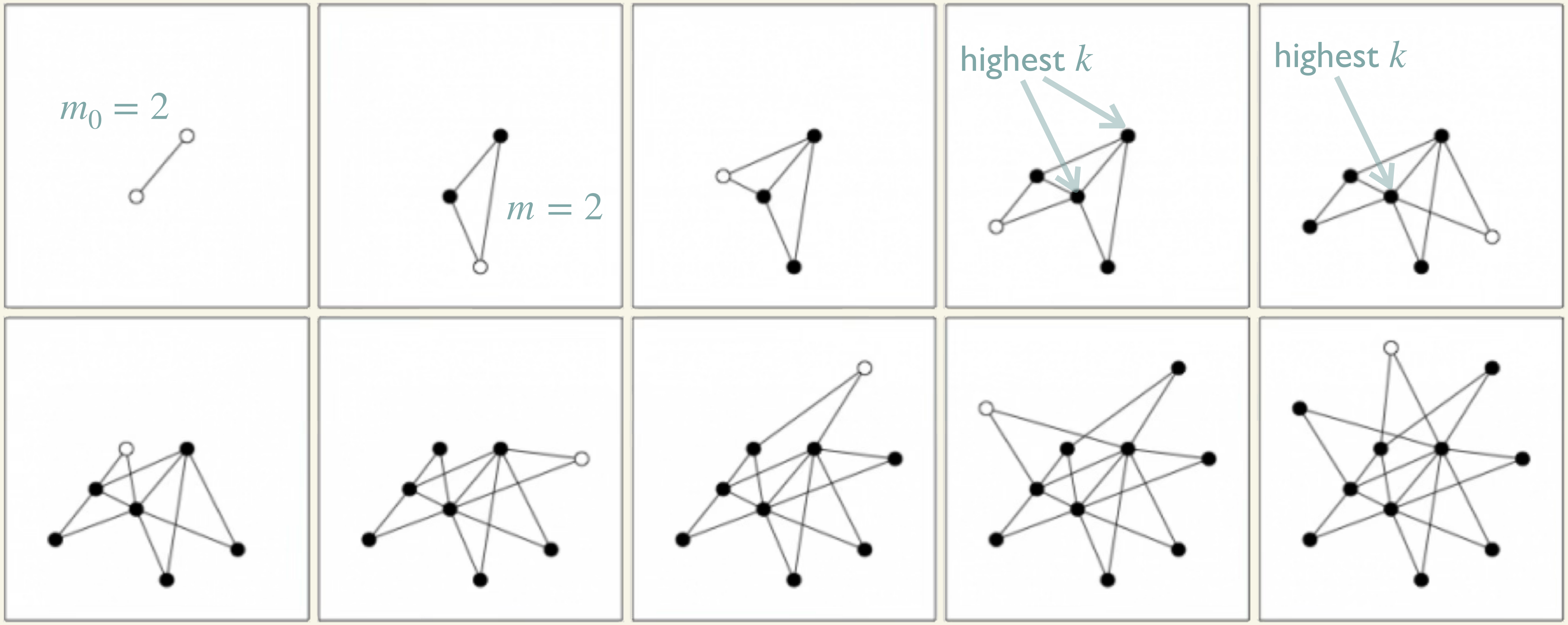
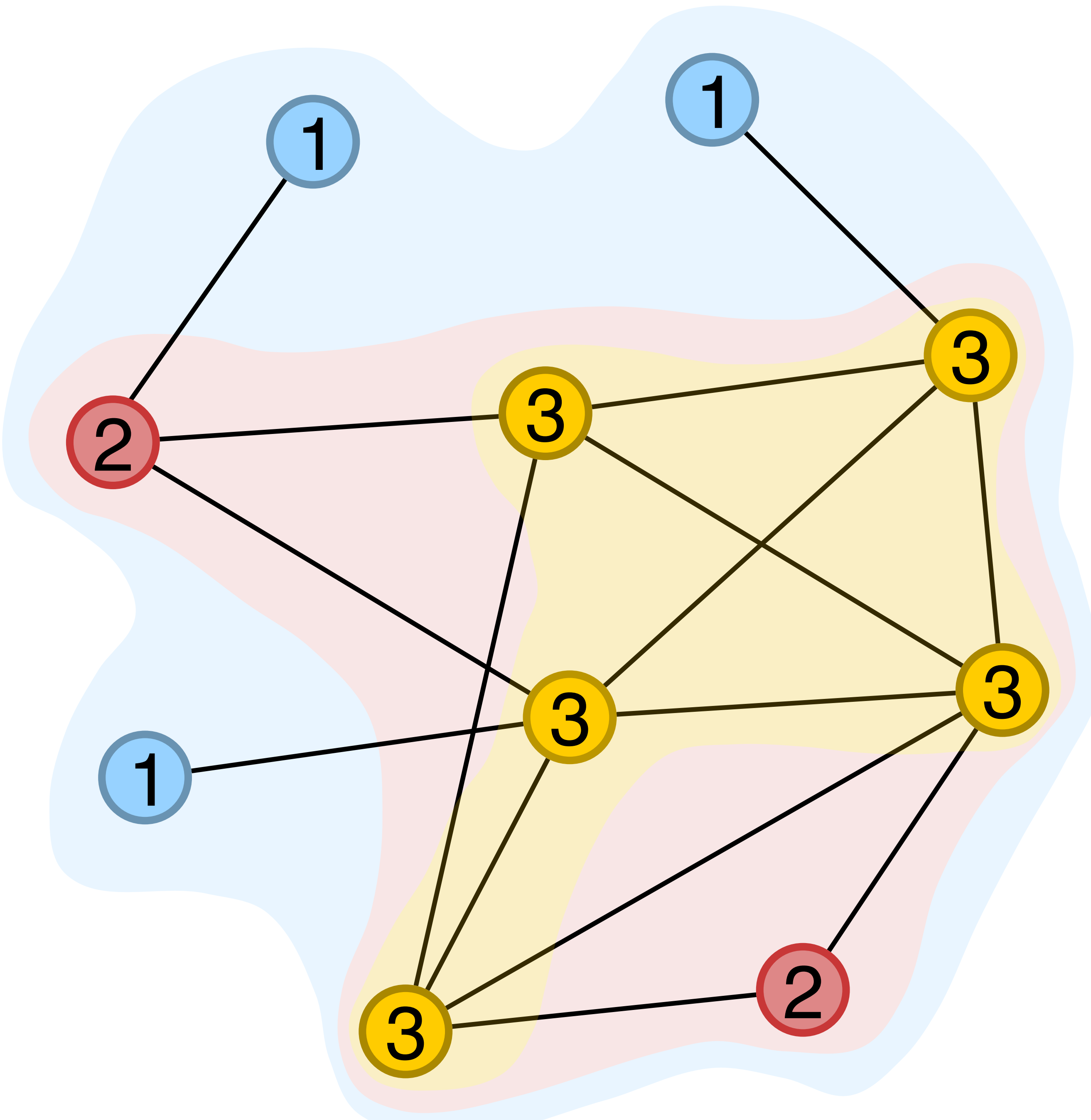


image: Barabási, A.-L., "Network Science" (http://networksciencebook.com/)

EXTRACTING STRUCTURAL FEATURES: CORE-PERIPHERY ORGANIZATION



One can deconstruct a network in terms of its core-periphery structure.

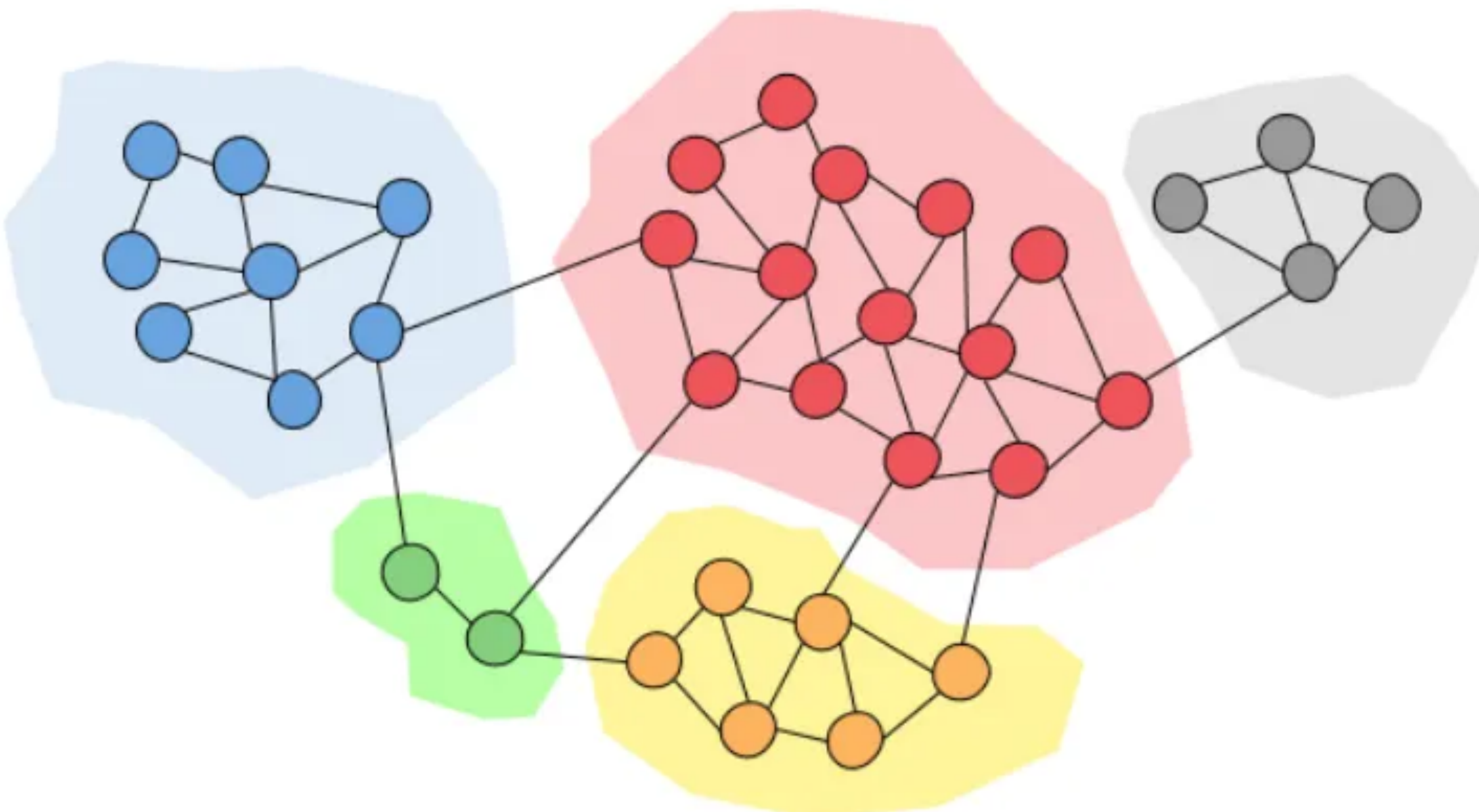
The k-core is the set of nodes of the network, within which each node has k links with each other.

A core need not be a single connected component.

The nodes of the highest k-core are referred to as core nodes and others are peripheral nodes.

EXTRACTING STRUCTURAL FEATURES: MODULARITY

A network is said to have a modular structure if there exist groups (or “communities”) of nodes that have a higher density of connections than that between groups.



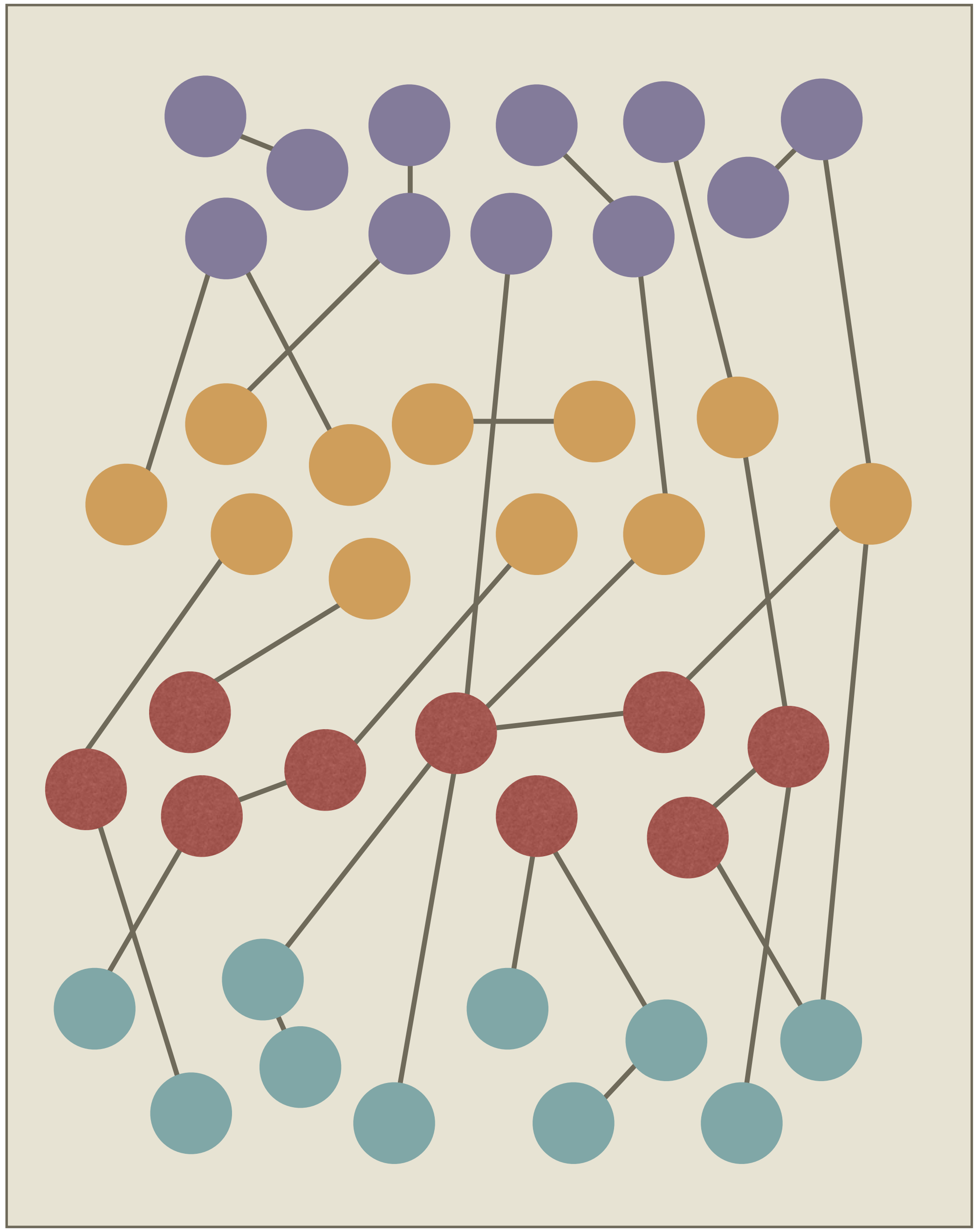
In practice, one has to first specify the modules/communities and then check if the density of intra-connections is more than that of the inter-connections.

image: https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae

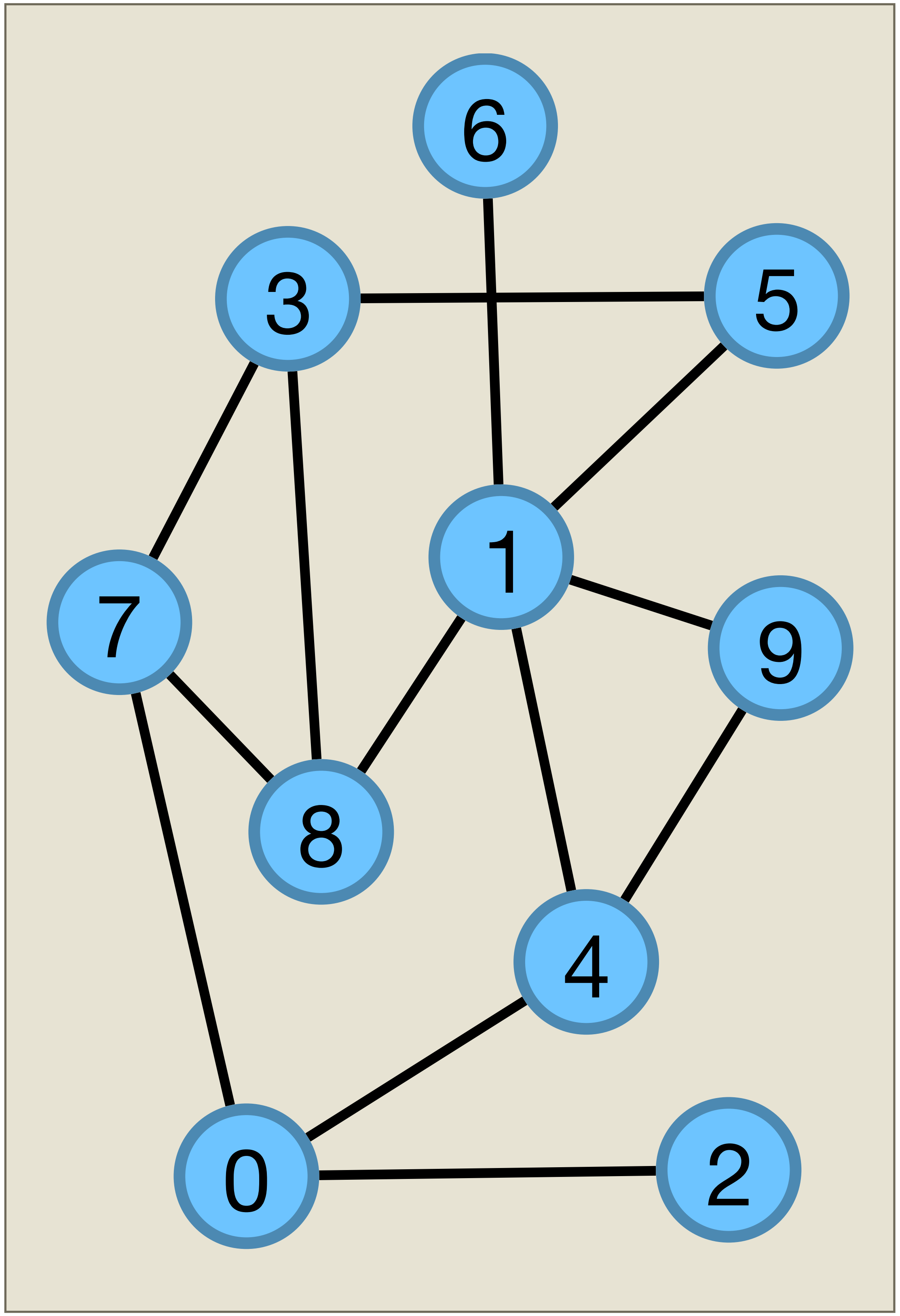
EXTRACTING STRUCTURAL FEATURES: HIERARCHY

A network is said to have a hierarchical structure if there exists "layers" of nodes, such that the density of connections between consecutive layers is higher than that within layers, or between non-consecutive layers.

As in the case of modularity, determining the optimal hierarchical decomposition of the network is a non-trivial problem.



HOMEWORK



1. Find the average degree of the network.
2. Find the shortest path length between nodes 3 & 4.
3. Find the diameter of the network
4. Find the number of walks of length 5 between nodes 2 & 3.
5. Find the clustering coefficient of nodes 1 & 8.
6. Find the betweenness centrality of nodes 3 & 4.