Introduction to networks

Big Data course CRI 2021-2022

Marc Santolini
Liubov Tupikina

Outline of the day

1. Introduction to networks
2. Practical part: notebooks
3. Hands-on session

What will we need for the practical part

Standard libraries (Python): numpy, matplotlib, scikit learn, seaborn

Network libraries: networkx, osmnx (open street data)

Inspiration from

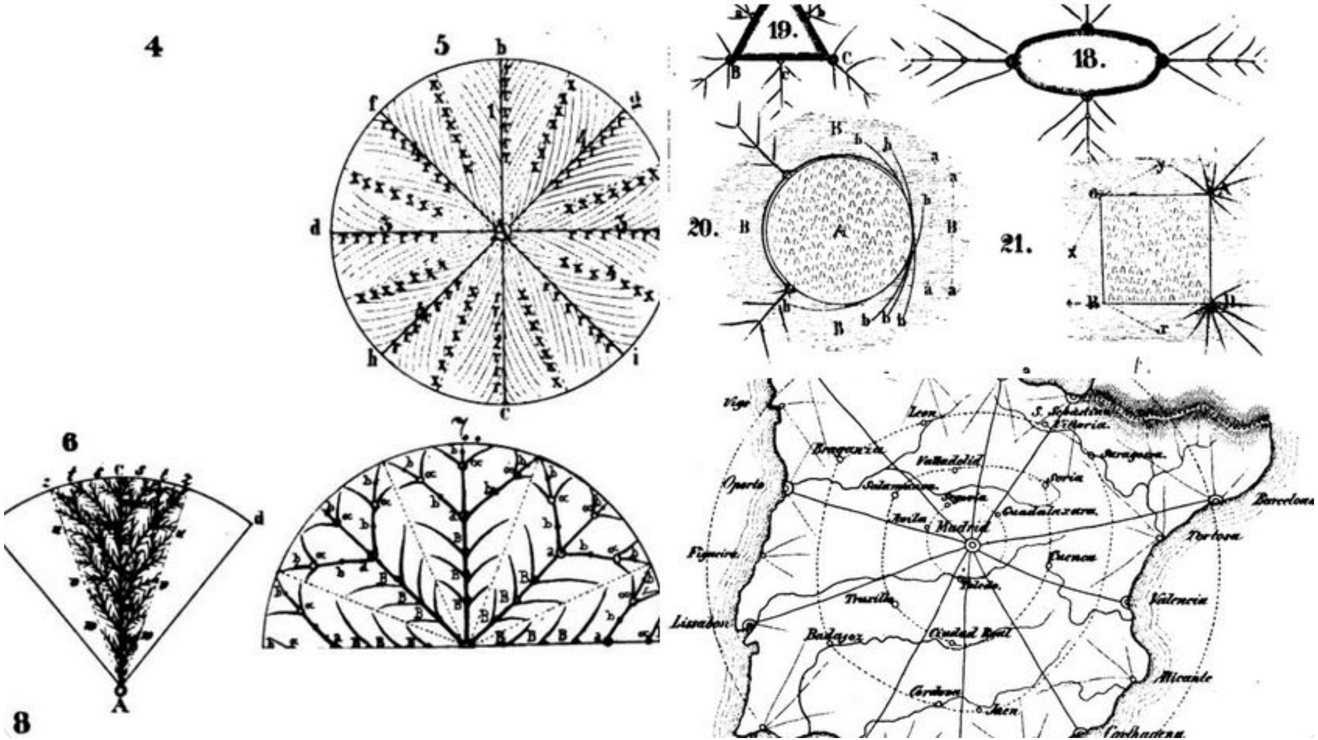
- Big data course Marc and Liubov https://github.com/Big-data-course-CRI/materials_big_data_cri_2019
- Correlaid and TidyTuesday https://github.com/rfordatascience/tidytuesday

Figure 7.11

Aaron Koblin's *Flight Patterns* (2005): visualization of the flight paths of aircraft crossing North America

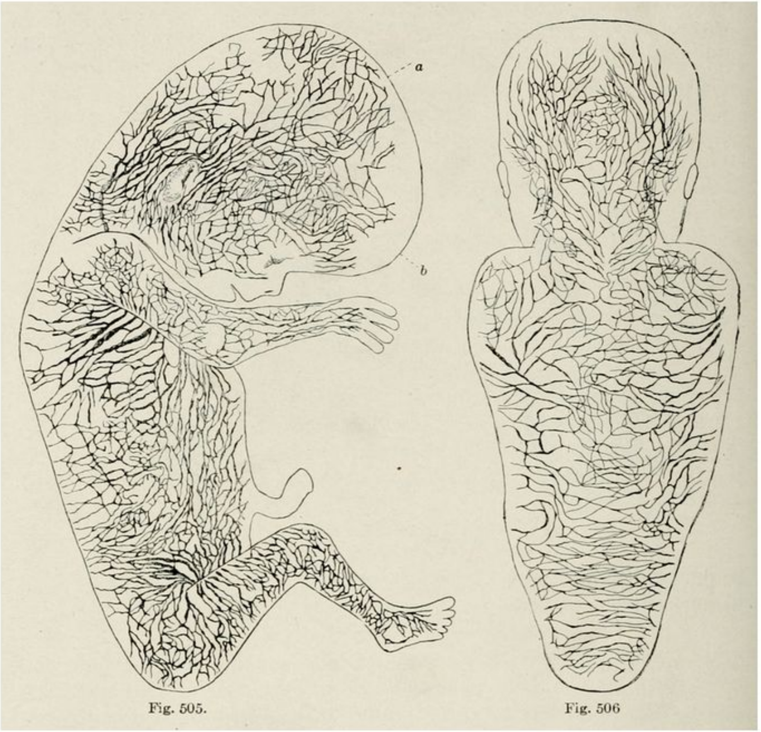
Some resources we will use: network book

http://networksciencebook.com/



Why networks?

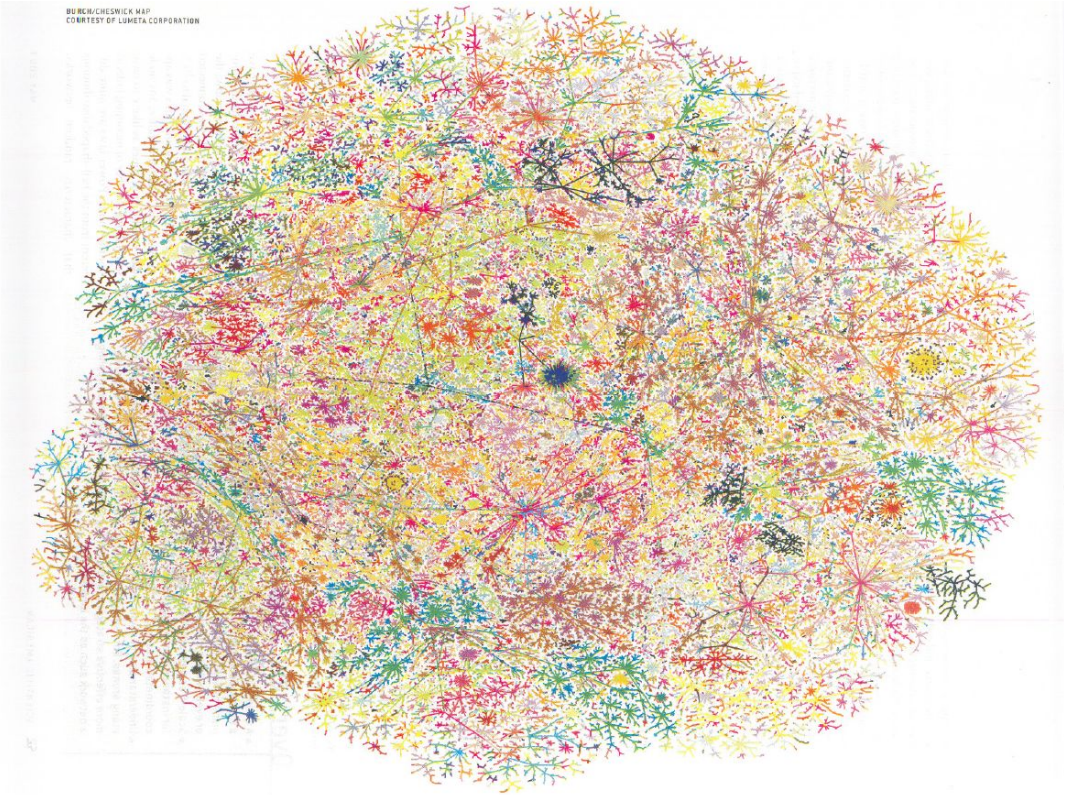
Where you can find them?



Distension of the lymphatic vessels in the human foetus, from Franz Kreibel, *Manual of human embryology*, 1910

What is network science?

The central idea in network science is that any node can influence other nodes, not only their direct connections. Such indirect influence happens through some external phenomenon—travel in a transportation network, information transfer in the Internet, vibrations in a spiderweb, etc.—and depends on how the network is connected. Thus, we can use the graph describing the interaction structure to understand how the system works, the roles of individual nodes, etc.
(Holme)



Network science

1. Network measures and network types
2. Networks in time and space
3. Networks from data

Ways to represent data in the form of a network

Figure 7.11

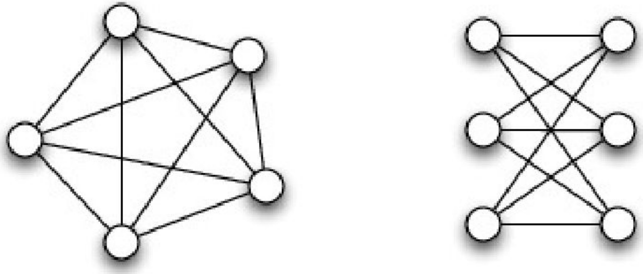
Aaron Koblin's *Flight Patterns* (2005): visualization of the flight paths of aircraft crossing North America

How did the network science start?

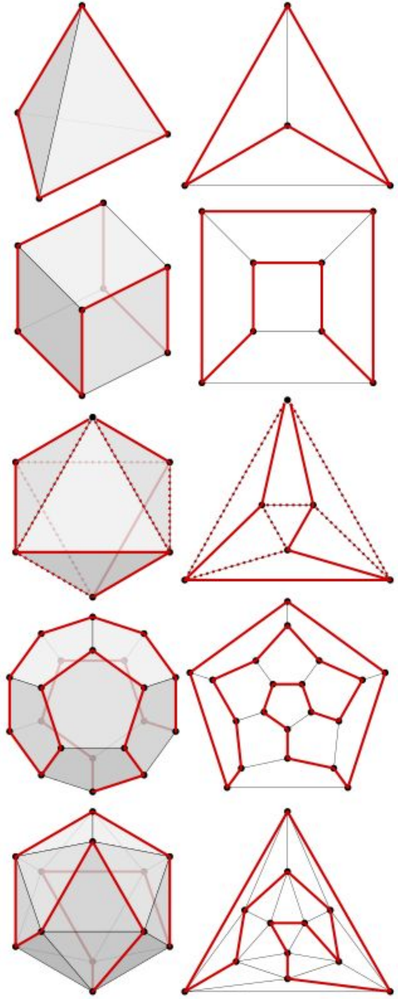
http://networksciencebook.com/

Network science how it started?

- Graph theory:
  Koenigsberg problem 1736
  Eurlerian paths algorithms 1873
- Soft matter physics
  https://arxiv.org/list/cond-mat.soft/recent



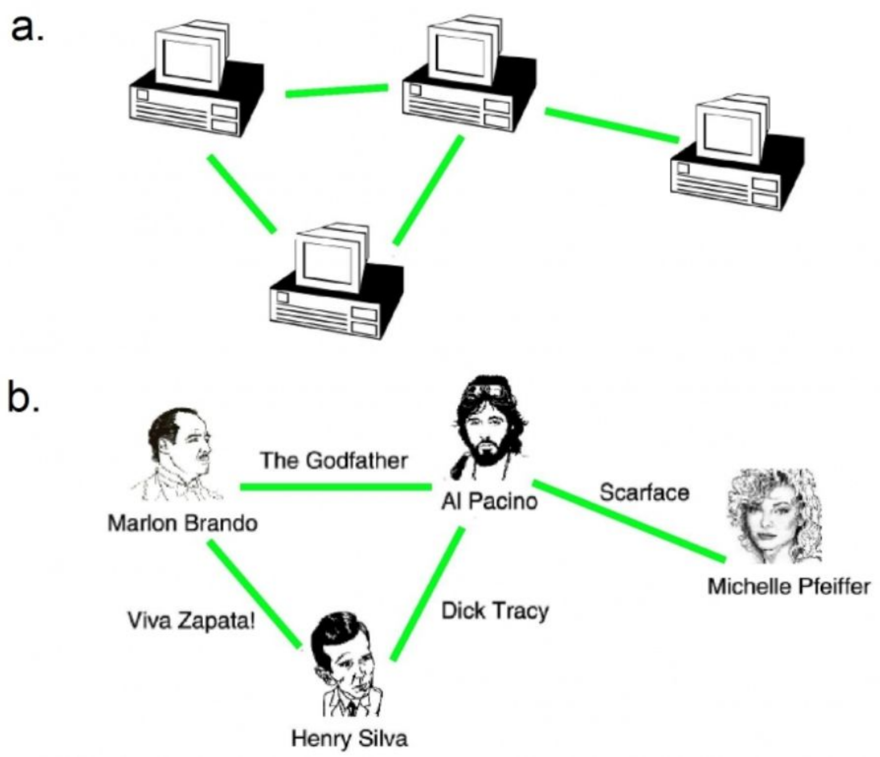
K5 K3,3



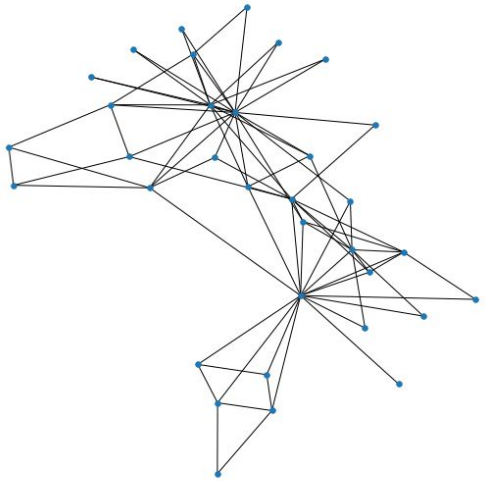
Networks vs. graph

Graph (discrete mathematics),
a structure made of vertices and edges.

Networks (from real world)
can be represented as graphs.

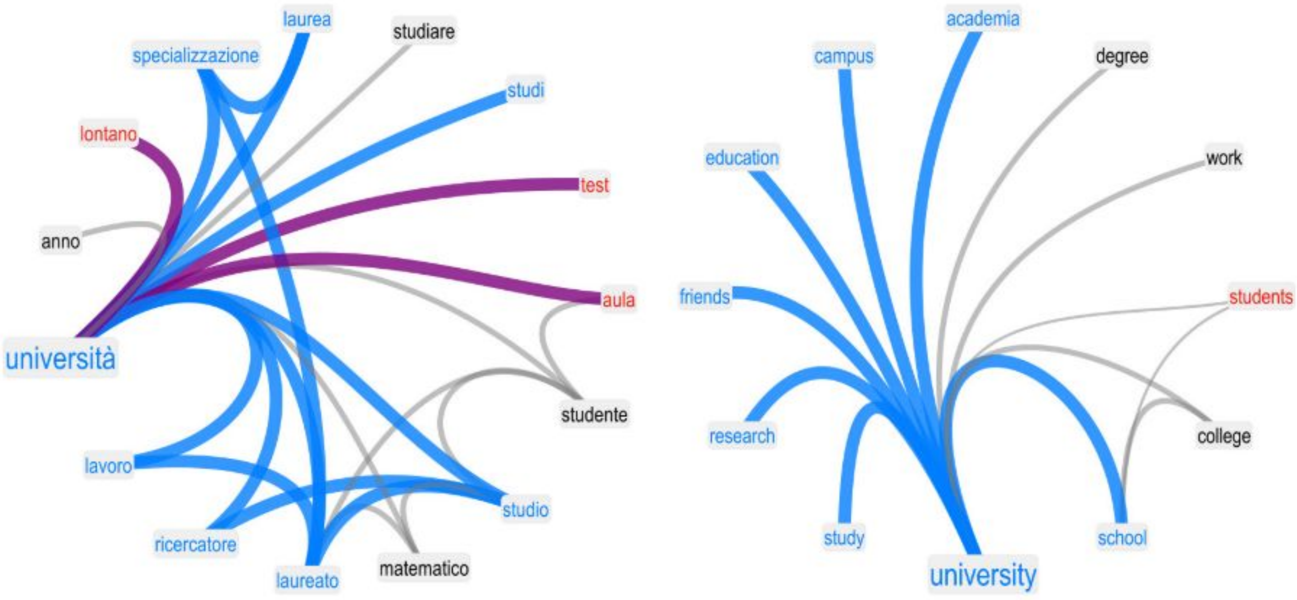


Can anything be presented as a network? What cannot be presented and why?



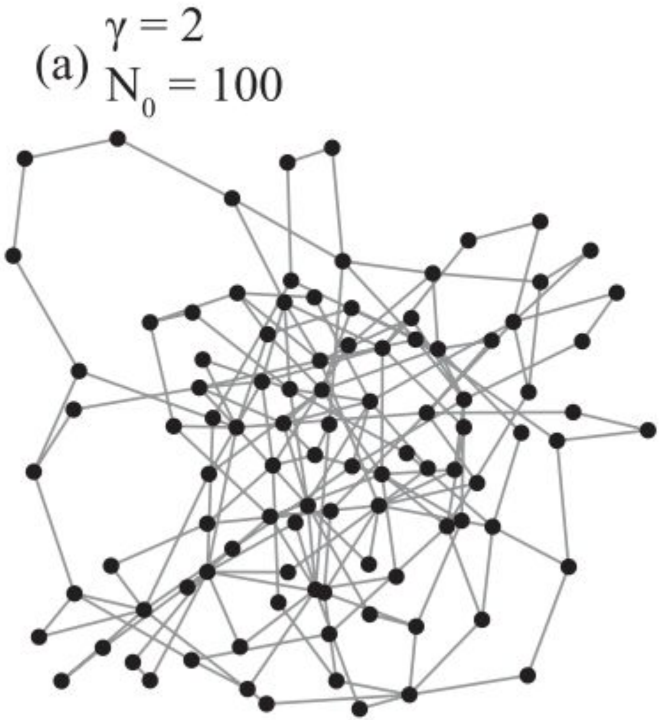
Examples of network representations

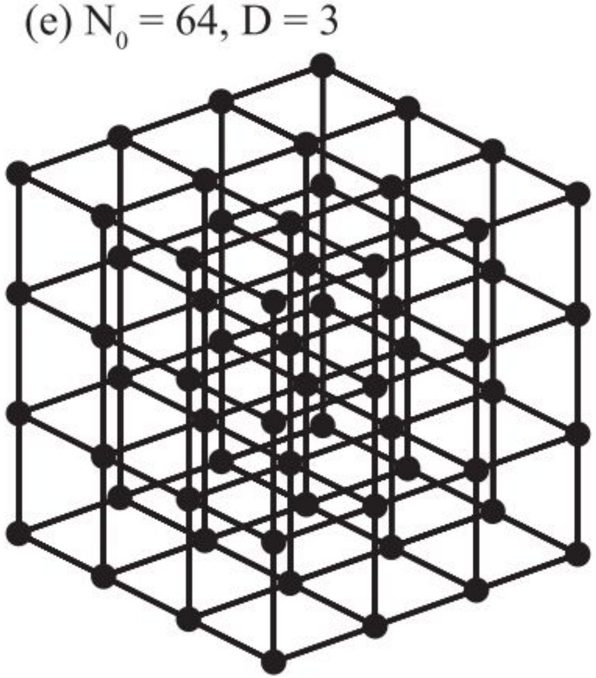
M.Stella et al.



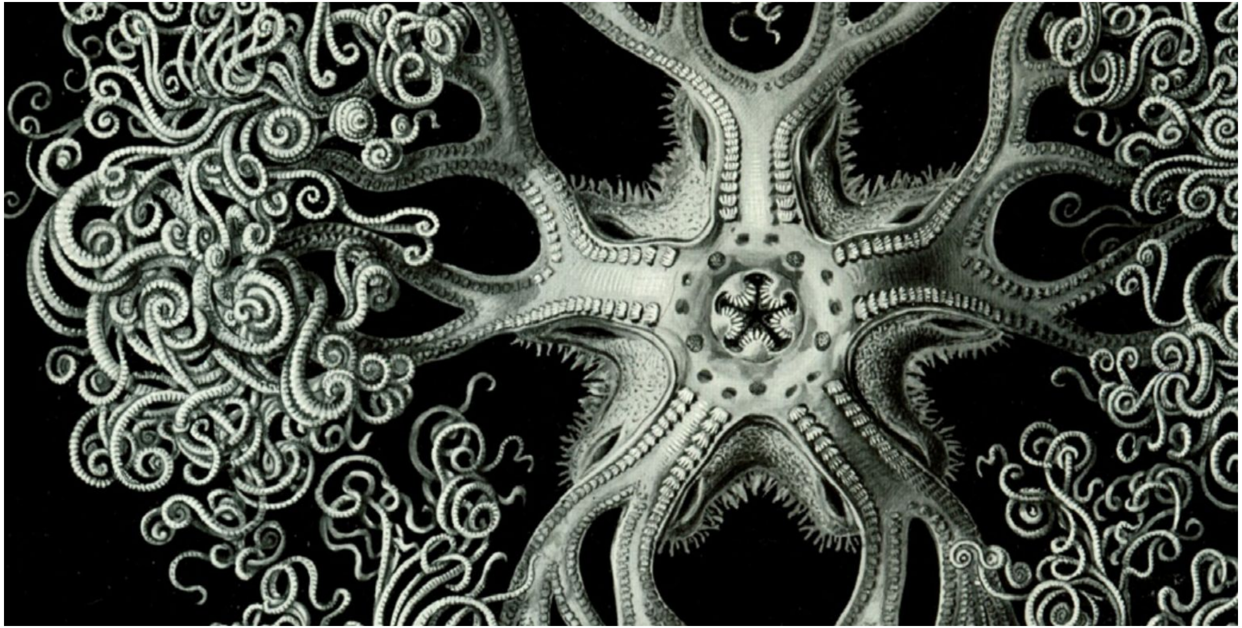
Examples of network representations

M.Stella et al.

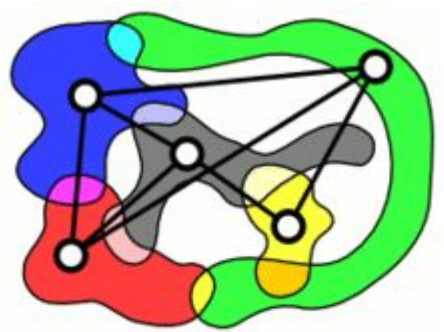


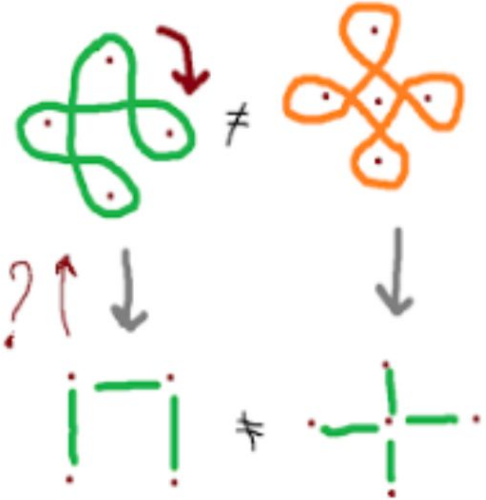


Can anything be presented as a network? What cannot be presented and why?



Can anything be presented as a network? What cannot be presented and why?





What we will look at in network science?

1. Network measures and network types
2. Networks in time and space
3. Networks from data

Wednesday class on modeling:

Networks and processes

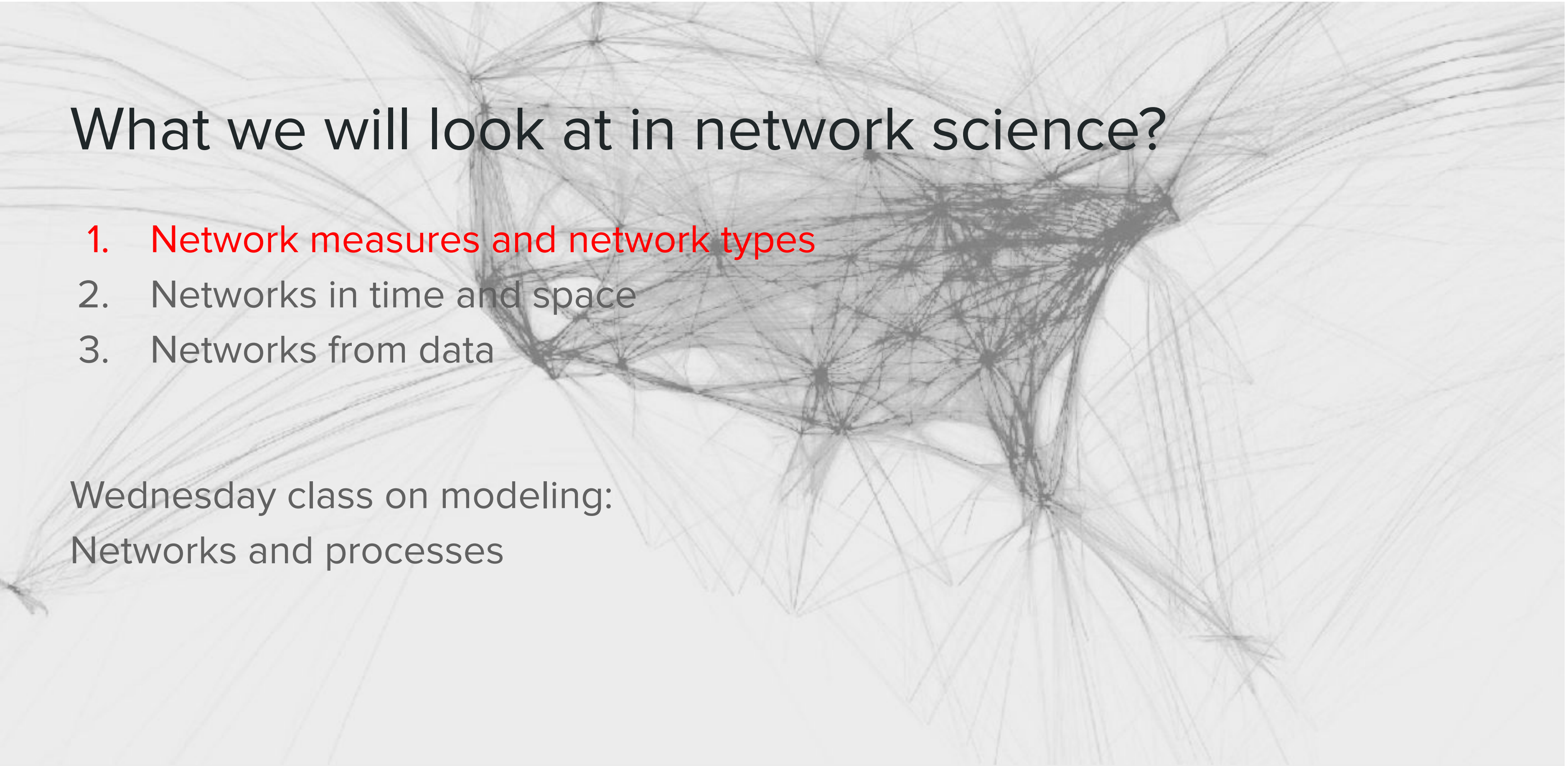


Figure 7.11

Aaron Koblin's *Flight Patterns* (2005): visualization of the flight paths of aircraft crossing North America

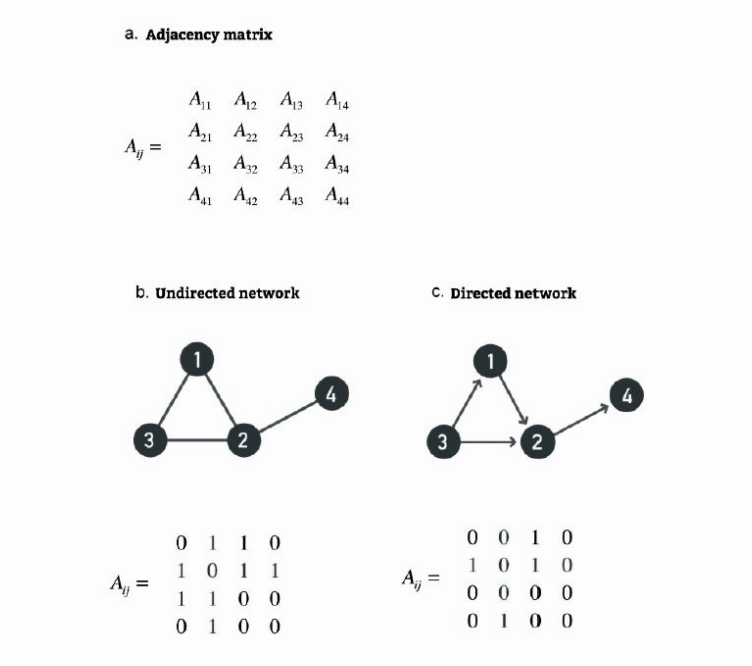
What we will look at in network science?

1. **Network definition and measures**
2. Networks in time and space
3. Networks from data

1. Network definitions

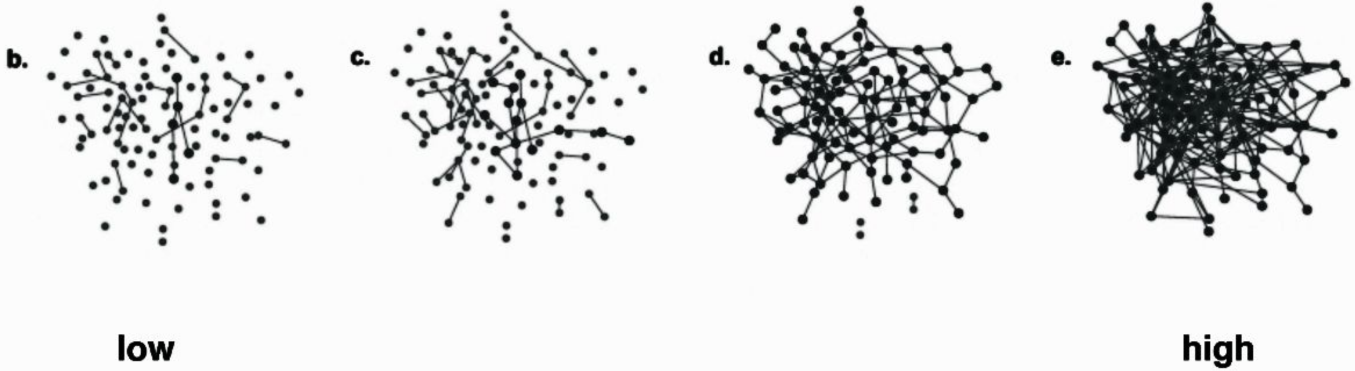
How to represent a network: **edgelist** and **adjacency matrix.**

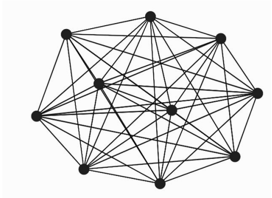
Adjacency matrix encodes the same information about the network as edgelists.



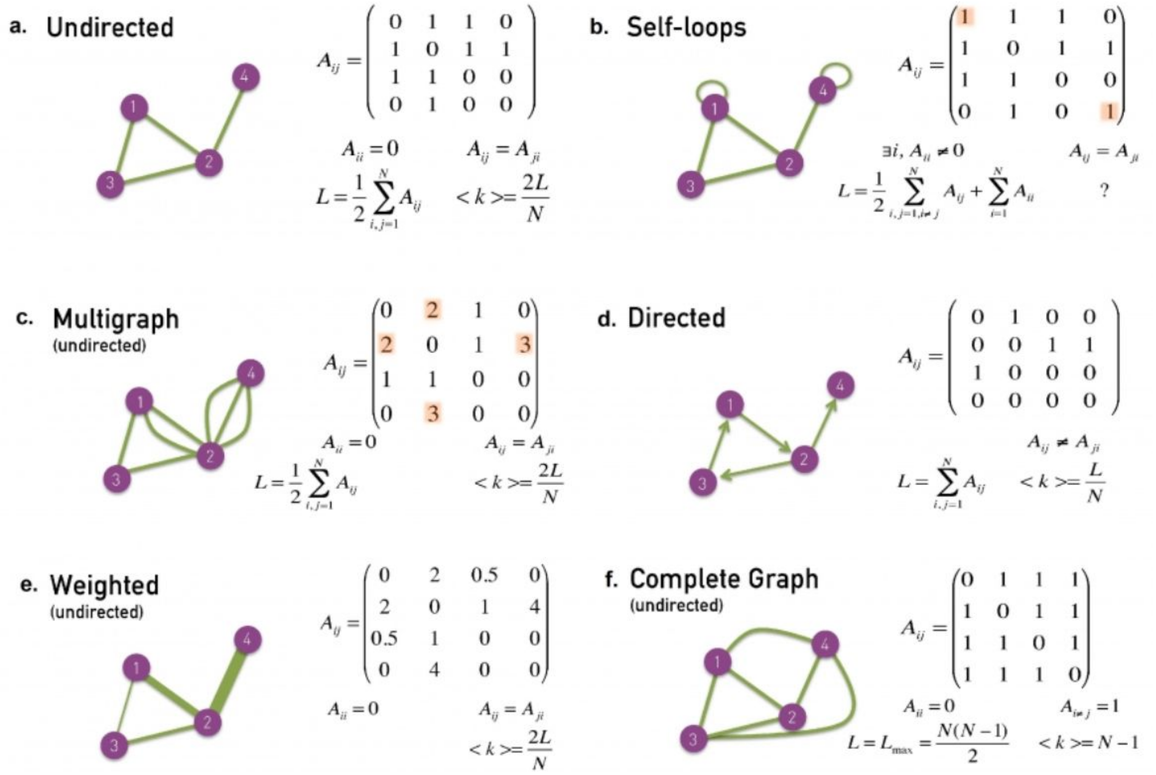
1. Network measures and definitions:

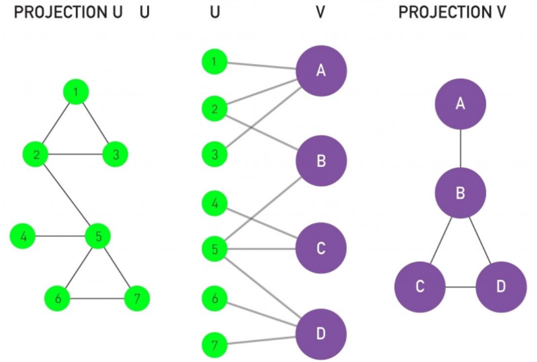
Density - number of links normalised by number of links in a **complete graph.**



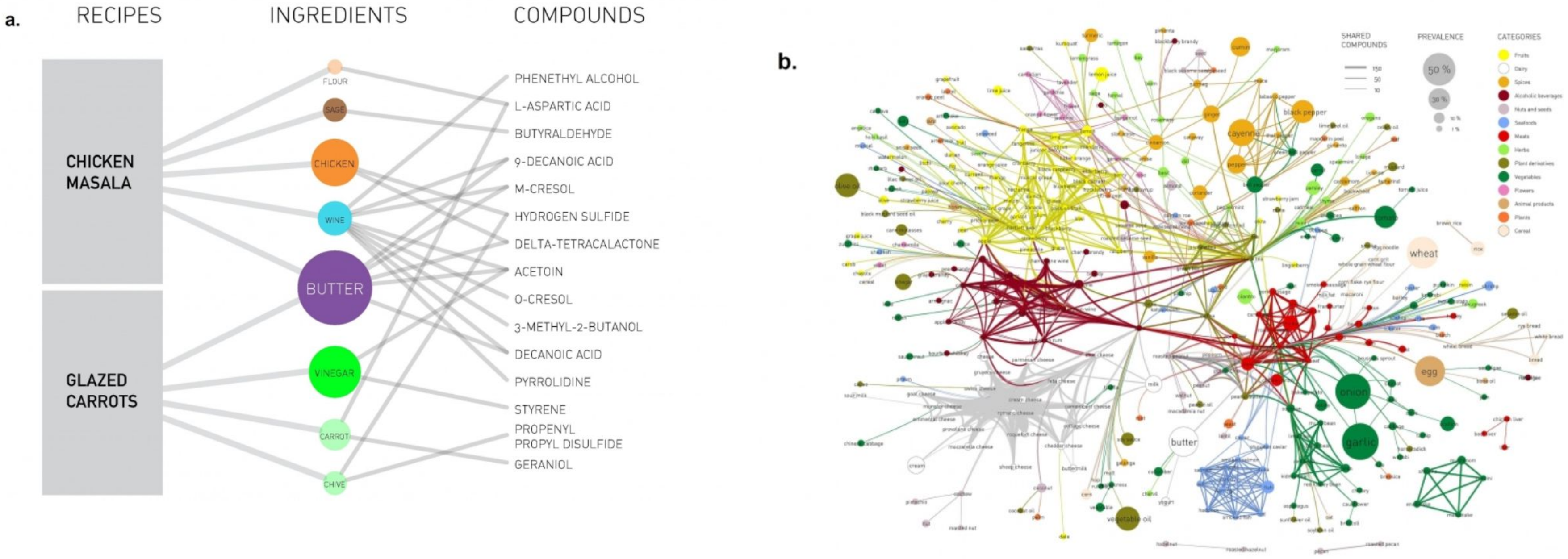


Networks based on their links/nodes properties



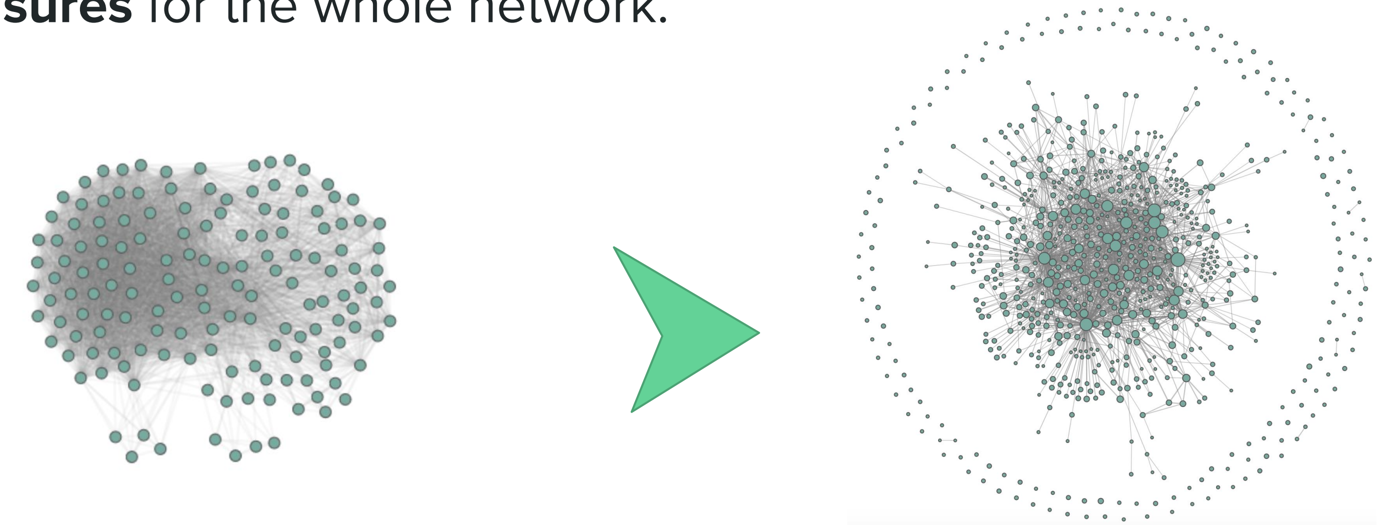


Network types: bipartite networks



Network measures main idea is to characterise their properties.

Local measures for each node.
Global measures for the whole network.

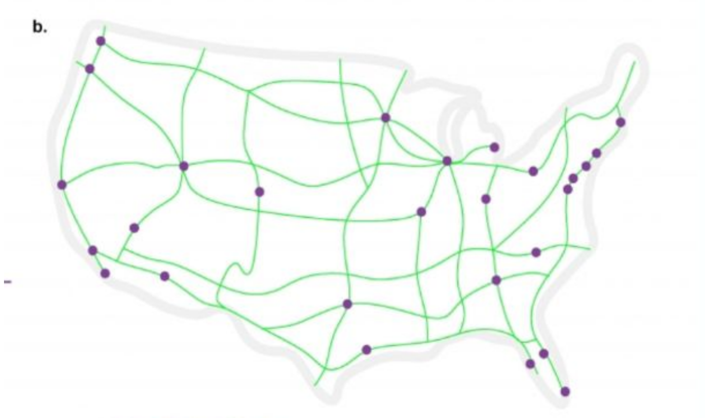


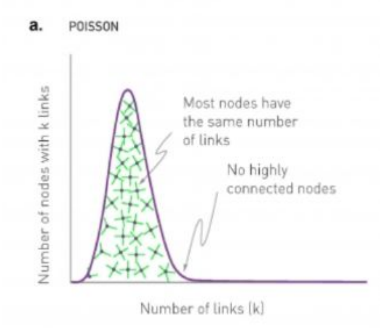
Network measures

Degree measure - is a local measure to characterise how many nodes each node is connected to.

Network measures

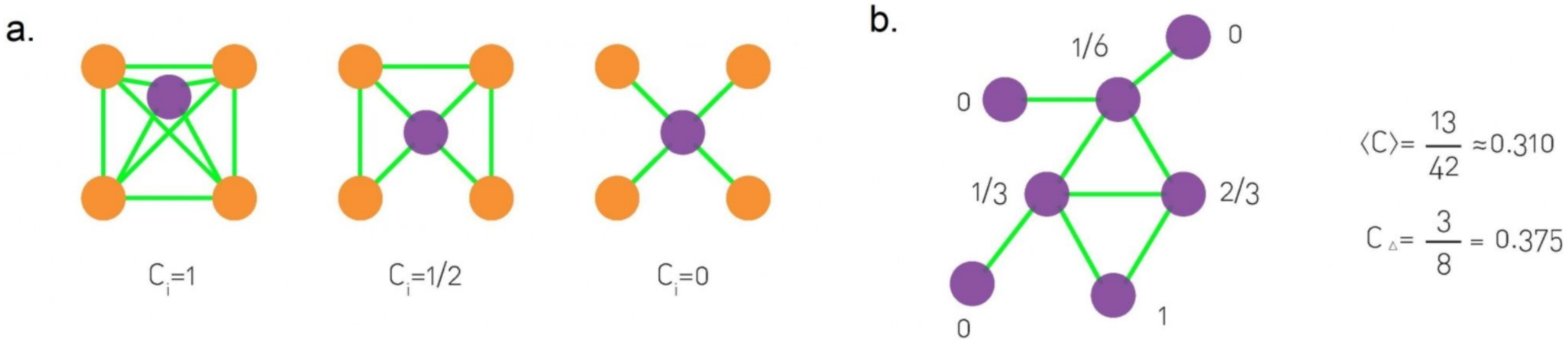
Degree measure - is a local measure to characterise how many nodes each node is connected to.
How to look into degree for N nodes?
Looking into the degree distribution: plotting how many nodes have degree=k.





Network measures

Clustering coefficient (do not confuse with clustering for the whole graph).
Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.



Network measures

Shortest path finding between nodes are used in other algorithms for networks.
Example path between node 1 and node 6 in a graph is then encoded as a sequence of nodes, e.g. (1,2,5,7,4,6)

One of the most known shortest path algorithm is Dijkstra’s algorithm (1956).

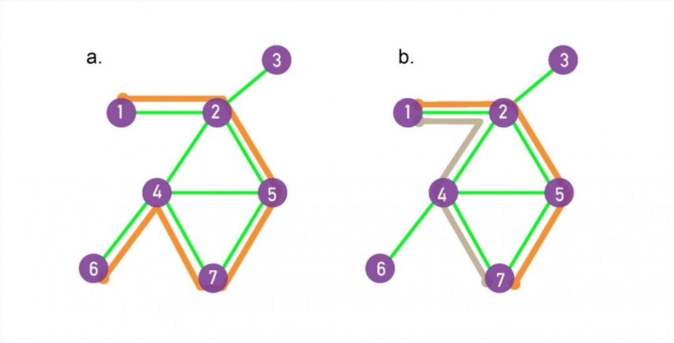


Image 2.12

Paths

- A path between nodes i_0 and i_n is an ordered list of n links P = {(i_0, i_1), (i_1, i_2), (i_2, i_3), ... ,(i_{n-1}, i_n)}. The length of this path is n. The path shown in orange in (a) follows the route 1→2→5→7→4→6, hence its length is n = 5.

Network measures

Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s, t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s, t$, $\sigma(s,t|v) = 0$ [2].

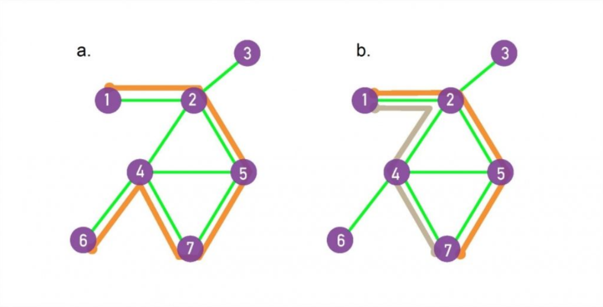


Image 2.12

Paths

- A path between nodes i_0 and i_n is an ordered list of n links P = {(i_0, i_1), (i_1, i_2), (i_2, i_3), ... ,(i_{n-1}, i_n)}. The length of this path is n. The path shown in orange in (a) follows the route 1→2→5→7→4→6, hence its length is n = 5.

Network measures

Closeness centrality of a node v is the reciprocal of the average shortest path distance to u over all n-1 reachable nodes.

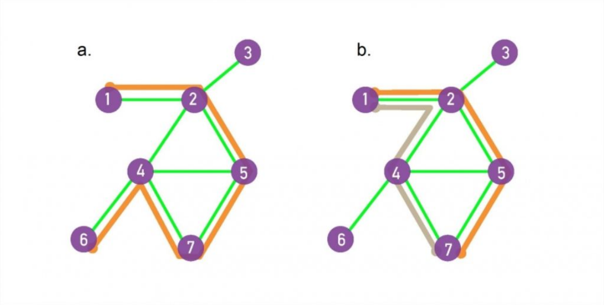
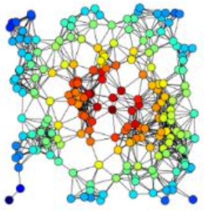


Image 2.12

Paths

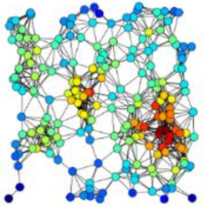
- A path between nodes i_0 and i_n is an ordered list of n links P = $\{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots ,(i_{n-1}, i_n)\}$. The length of this path is n. The path shown in orange in (a) follows the route 1→2→5→7→4→6, hence its length is n = 5.

Network measures
Networkx



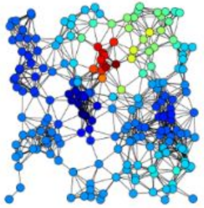
B

Closeness centrality



D

Degree centrality



F

Katz centrality

Table 2: Definitions of network science terms and variables.

Term/variable	Definition
N	number of nodes, N, in graph
E	number of edges, E, in graph
network density	ratio of the number of edges to the maximum number of possible edges $\frac{2E}{N(N-1)}$
$d(n_i, n_j)$	shortest path between node i and node j $d(n_i, n_j)$ where $n_i, n_j \in N$
hortest path length, L	average length of shortest path between pairs of nodes $L = \frac{1}{N(N-1)} \cdot \sum_{i \neq j} d(n_i, n_j)$
, D	largest shortest path between nodes $D = \max_{n_i \in N, n_j \in N} d(n_i, n_j)$
centrality	inverse of the sum of the length of the shortest paths between node i and all other nodes in the graph $C_i = \frac{1}{\sum_j d(n_i, n_j)}$
i	number of edges attached to node i
legree, $\langle k \rangle$	average number of edges per node in network $\langle k \rangle = \frac{1}{N} \sum_{n=i}^{N} k_i$
tering coefficient, c_i	number of edges between the neighbors of node i divided by the maximum number of edges between those neighbors $c_i = \frac{2\lvert e_{jk} \rvert}{k_i(k_i - 1)}$ where $n_j, n_k \in N_i$, $e_{jk} \in E$
lustering coefficient, $\langle C \rangle$	average clustering coefficient of nodes in the network $\langle C \rangle = \frac{1}{N} \sum_{n=i}^{N} c_i$
ity, Q	proportion of edges that fall within subgroups of nodes minus the expected proportion if edges were randomly distributed, range $[-1, 1]$
fficiency, E_G	measure of how efficiently information is exchanged in the network $E_G = \frac{1}{n(n-1)} \sum_{i \neq j \in N} \frac{1}{d(n_i, n_j)}$
nnected component	largest group of nodes in the network that are connected to each other in a single component
stribution, $P(k)$	probability distribution of node degrees in the network
	power-law exponent for the degree distribution
rld structure	network with short average path lengths and relatively high clustering coefficient (relative to a random graph with similar density)
network	network with a degree distribution that is power-law distributed

Network measures

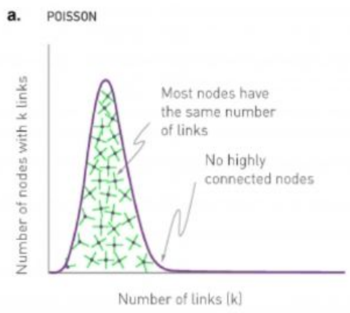
Most of the measures can be estimated directly using networkx python library.

TABLE 2: Definitions of network science terms and variables.

Term/variable	Definition
N	number of nodes, N, in graph
E	number of edges, E, in graph
network density	ratio of the number of edges to the maximum number of possible edges $\frac{2E}{N(N-1)}$
distance, $d(n_i, n_j)$	shortest path between node i and node j $d(n_i, n_j)$ where $n_i, n_j \in N$
average shortest path length, L	average length of shortest path between pairs of nodes $L = \frac{1}{N(N-1)} \cdot \sum_{i \neq j} d(n_i, n_j)$
diameter, D	largest shortest path between nodes $D = \max_{n_i \in N, n_j \in N} d(n_i, n_j)$
closeness centrality	inverse of the sum of the length of the shortest paths between node i and all other nodes in the graph $C_i = \frac{1}{\sum_j d(n_i, n_j)}$
degree, k_i	number of edges attached to node i
average degree, $\langle k \rangle$	average number of edges per node in network $\langle k \rangle = \frac{1}{N} \sum_{n=1}^{N} k_i$
local clustering coefficient, c_i	number of edges between the neighbors of node i divided by the maximum number of edges between those neighbors $c_i = \frac{2\lvert e_{jk} \rvert}{k_i(k_i-1)}$ where $n_j, n_k \in N_i$, $e_{jk} \in E$
average clustering coefficient, $\langle C \rangle$	average clustering coefficient of nodes in the network $\langle C \rangle = \frac{1}{N} \sum_{n=i}^{N} c_i$
modularity, Q	proportion of edges that fall within subgroups of nodes minus the expected proportion if edges were randomly distributed, range $[-1, 1]$
average efficiency, E_G	measure of how efficiently information is exchanged in the network $E_G = \frac{1}{n(n-1)} \sum_{i \neq j \in N} \frac{1}{d(n_i, n_j)}$
largest connected component	largest group of nodes in the network that are connected to each other in a single component
degree distribution, $P(k)$	probability distribution of node degrees in the network
γ	power-law exponent for the degree distribution
Small world structure	network with short average path lengths and relatively high clustering coefficient (relative to a random graph with similar density)
scale-free network	network with a degree distribution that is power-law distributed

Network measures

Global networks measures: given any network centrality we can estimate global (or average network measure).
The main idea is to compare each value to the average.
Yet mind that it gives only “average temperature over hospital”.



Nodes	Edges	$\langle k \rangle$	σ_k	λ_h
70	366	5.23	4.53	8.74

Examples:
https://networks.skewed.de/net/highschool

Quick check-in

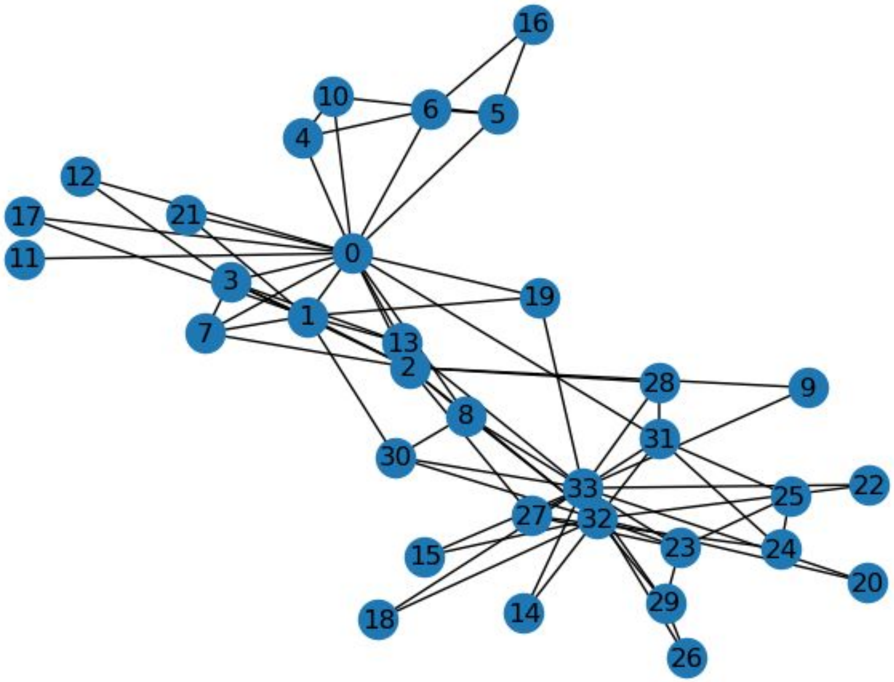
Karate club network

Zachary W. (1977). An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33, 452-473.

Quick check-in

What are network measures for this network?

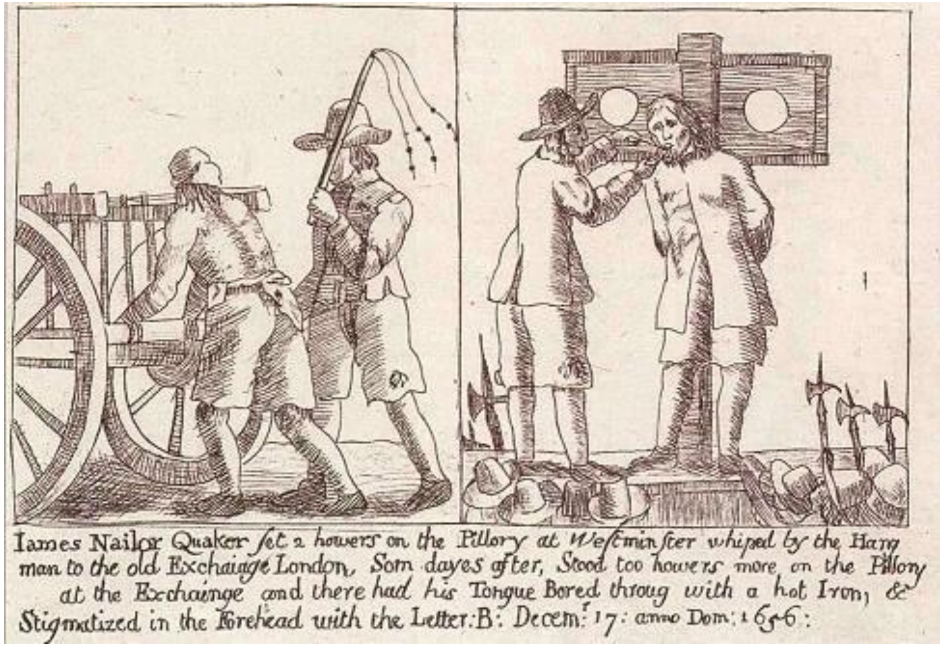
What node would have the highest betweenness centrality? What would be the best spreader?



Quick check-in

Can networks tell a new story about your data?

Typical examples:
Quakers, people who belong to a historically Protestant Christian set of denominations known formally as the Religious Society of Friends.



Random networks

Random networks

Network G on N nodes.
Model by Erdős (1913-1996) and Rényi (1921-1970) creates a random network with algorithm:
1. Create N nodes
2. Connect each pair of N labeled nodes with probability p.
You can do it yourself by tossing a coin each time.

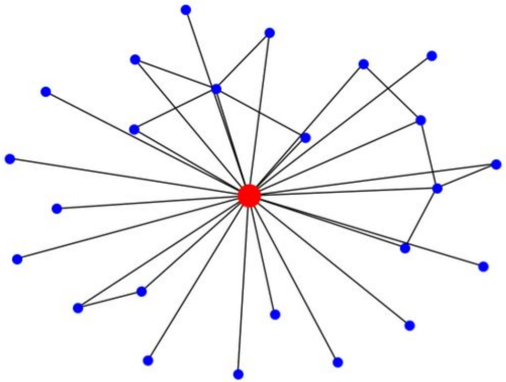
Corresponding class in networkx:

```
G_er = nx.erdos_renyi_graph(n, p2)
```

Random networks

Network G on N nodes and L links (N, L).
What is the degree distribution of Erdos-Renyi network?

```
G_er = nx.erdos_renyi_graph(n, p2)
```



Random networks

Network G on N nodes and m edges preferential attachment Model by Barabasi and Albert creates a random network with algorithm:

1. Create starting nodes
2. Connecta new node with m edges to existing nodes
3. Repeat (2.) x times for all non existing nodes

Corresponding class in networkx:

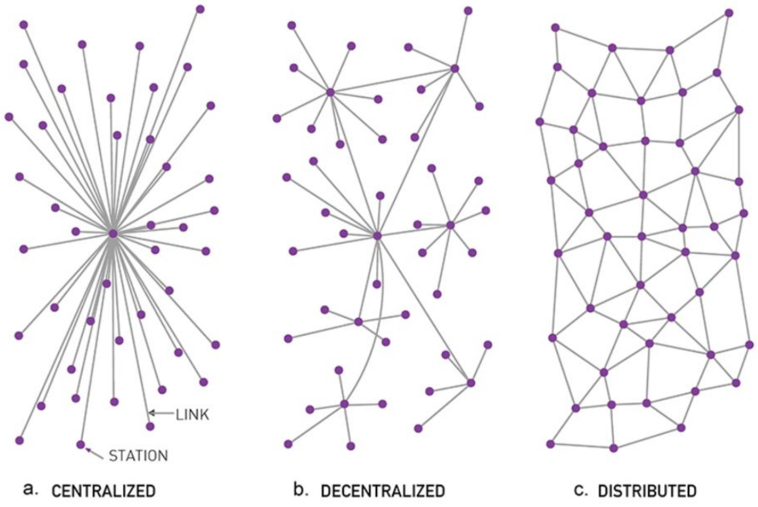
```
G_er = nx.barabasi_albert(n, m)
```

Network as a feeling

Network science requires intuition, e.g. how to construct a network, such that it would have a specific property, e.g robustness, or particular distribution?

Are random networks robust?

Fig. credits P. Barran



Practical part

Classroom notebooks

Network science

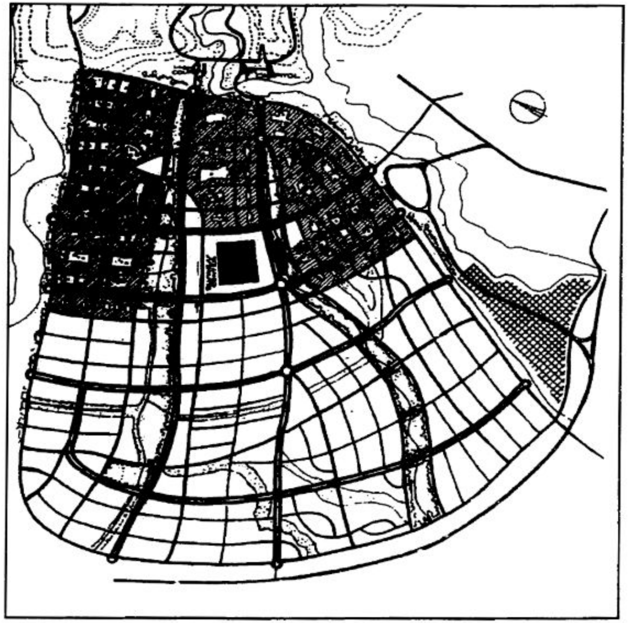
1. Network measures and network types
2. **Networks in time and space**
3. Networks from data

Figure 7.11

Aaron Koblin's *Flight Patterns* (2005): visualization of the flight paths of aircraft crossing North America

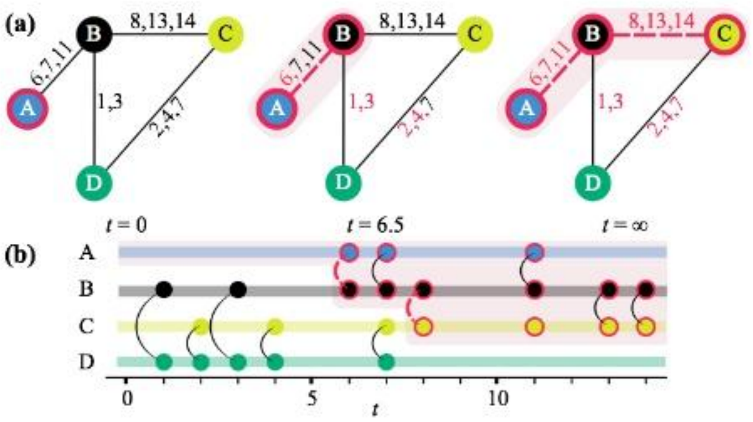
2. Networks in time and space

Master Plan for Chandigarh by Albert Mayer RAIC Journal, 1955 (Evenson Norma, Chandigarh, 1966)



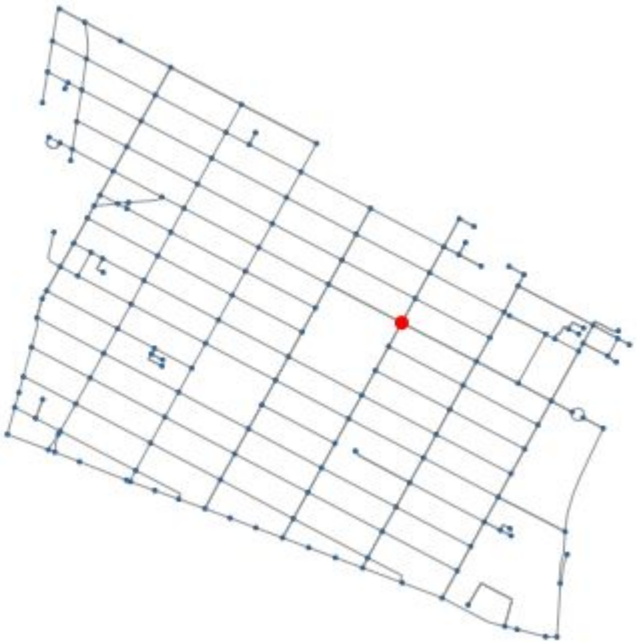
https://arxiv.org/pdf/1108.1780.pdf

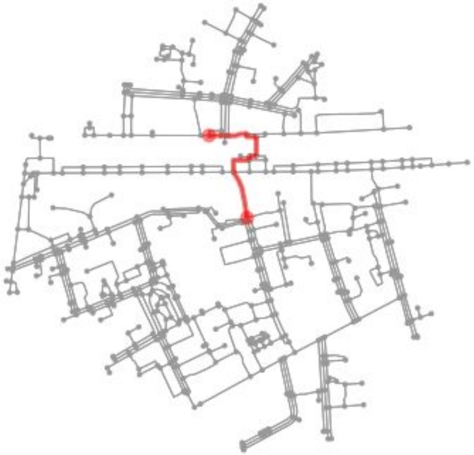
Temporality matters: reachability issue



Review Holme-Saramaki, Phys. Rep. (2012), arXiv:1108.1780

Networks in time and space

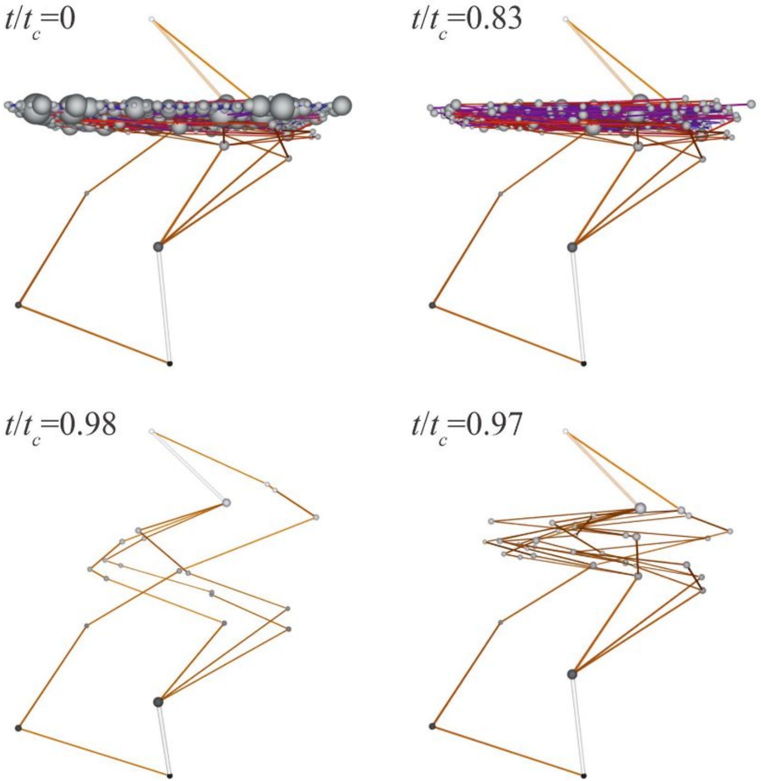




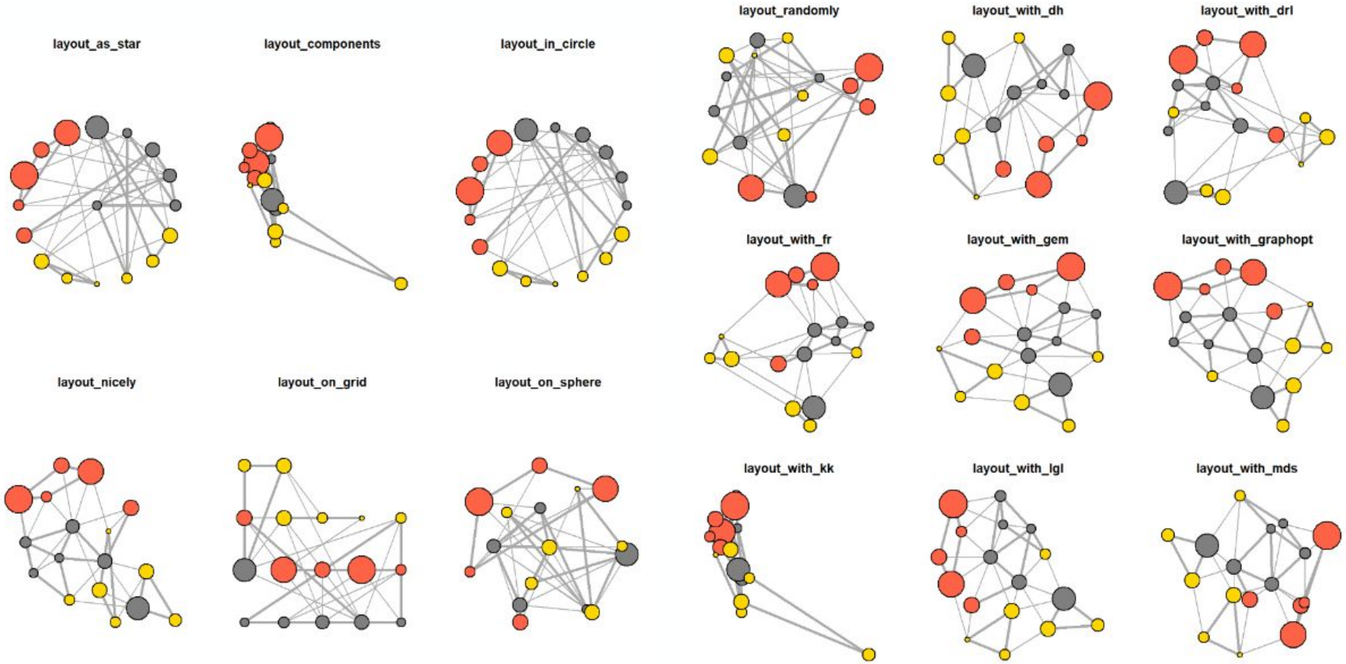
Osmnx for spatial networks analysis
https://arxiv.org/abs/1010.0302

Networks in time and space

Percolation of networks in time
[Nat.Comm. 2020]

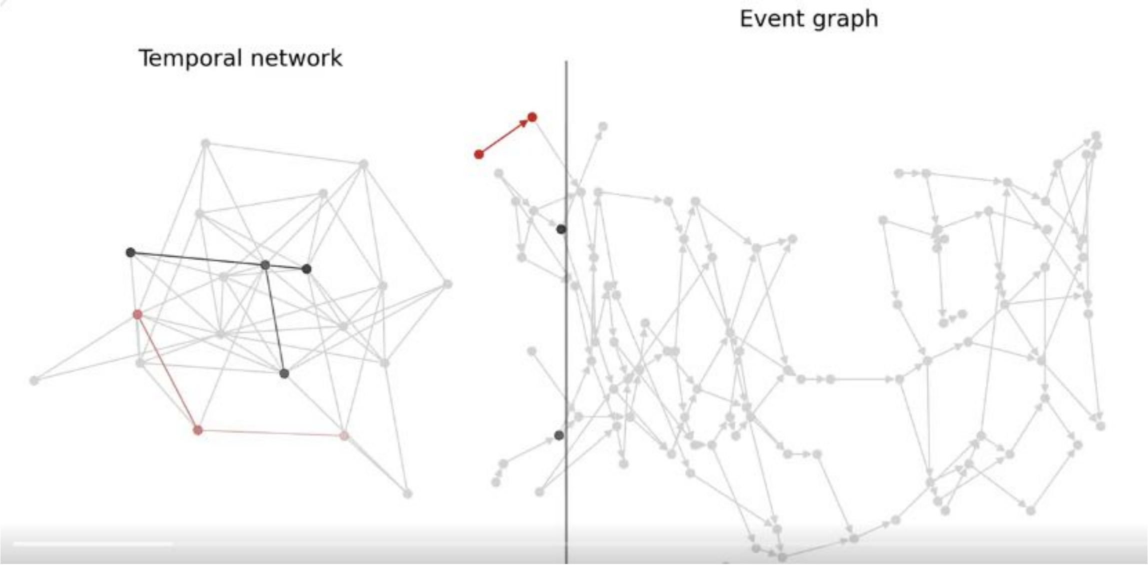


Networks layout



Networks sonification

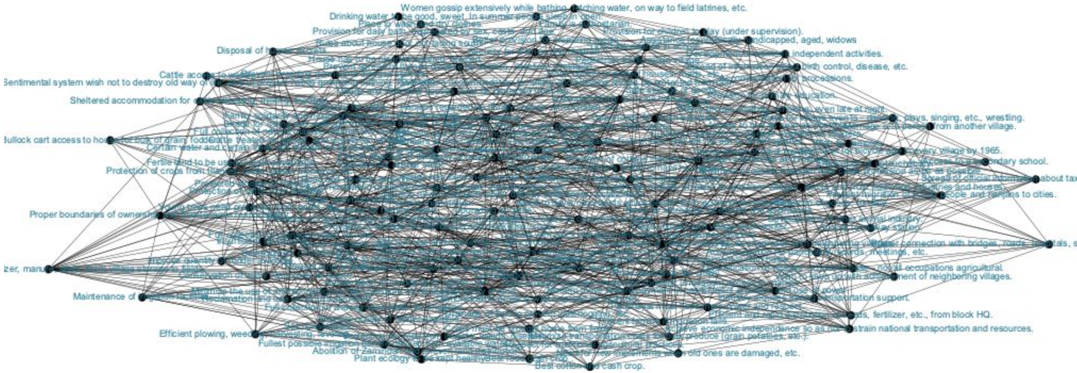
Directed percolation in temporal networks PRR 2022



Networks in time and space

Good resource on spatial networks

M.Barthelemy “Spatial networks”



Good resource on temporal networks

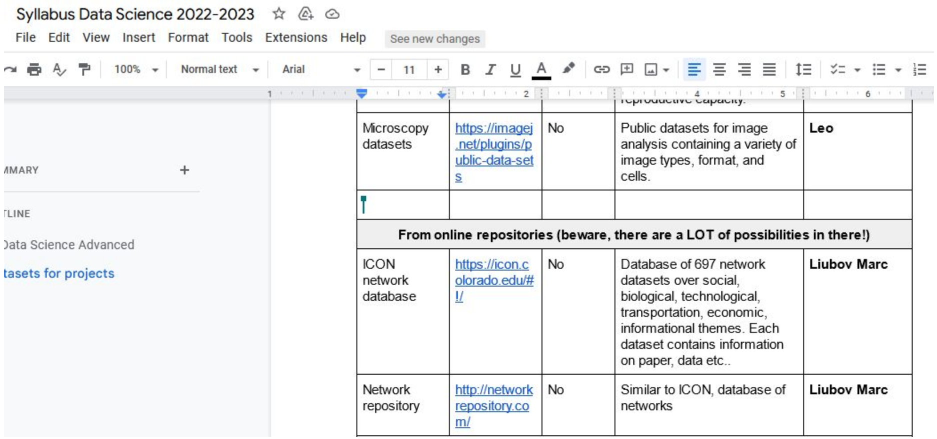
P.Holme, J.Saramaki “Temporal networks”

Holme blog https://petterhol.me/

Where can I get network data?

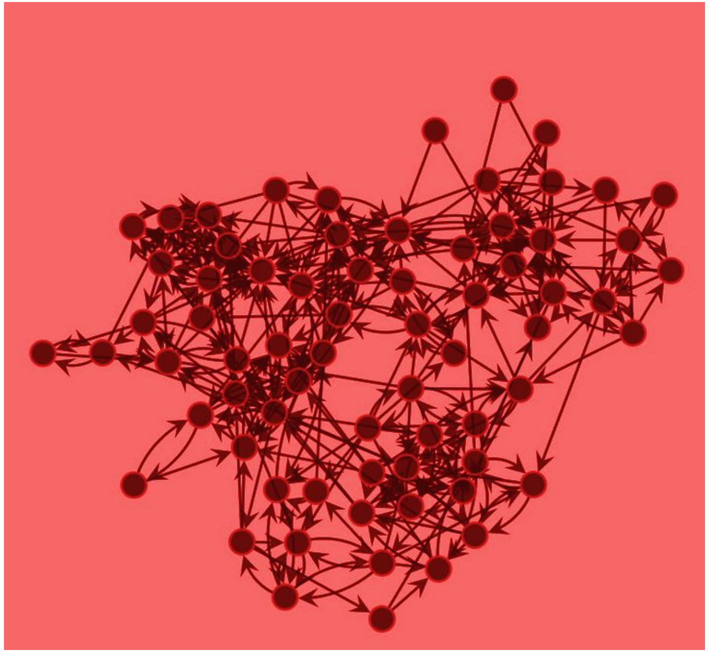
Example:
Highschool: Illinois high school students (1958). A network of friendships among male students in a small high school in Illinois from 1958. 70 nodes, 366 edges.
 https://networks.skewed.de/net/highschool

Example:
Facebook or wikipedia data
https://snap.stanford.edu/data/wiki-meta.html


Syllabus Data Science 2022-2023

File Edit View Insert Format Tools Extensions Help See new changes

Microscopy datasets	https://imagej.net/plugins/public-data-sets	No	Public datasets for image analysis containing a variety of image types, format, and cells.	**Leo**
From online repositories (beware, there are a LOT of possibilities in there!)				
ICON network database	https://icon.colorado.edu/#!/	No	Database of 697 network datasets over social, biological, technological, transportation, economic, informational themes. Each dataset contains information on paper, data etc..	**Liubov Marc**
Network repository	http://networkrepository.com/	No	Similar to ICON, database of networks	**Liubov Marc**



What we will look at in network science?

1. Network measures
2. Networks in time and space
3. **Networks from data**

Figure 7.11

Aaron Koblin's *Flight Patterns* (2005): visualization of the flight paths of aircraft crossing North America

Social networks analysis

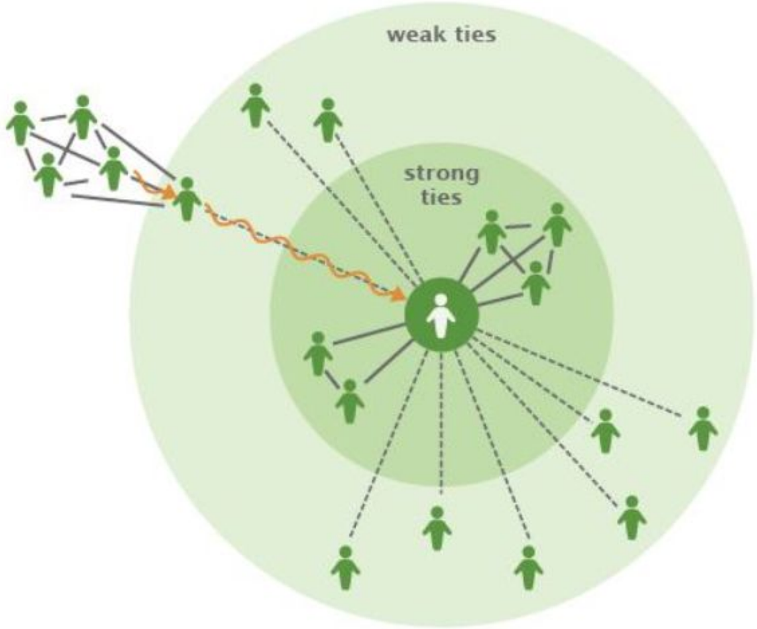
The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases.

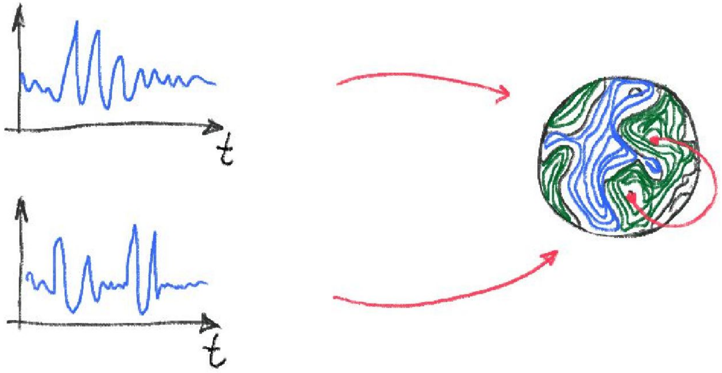
I will argue, in this paper, that the analysis of processes in interpersonal networks provides the most fruitful micro-macro bridge. In one way or another, it is through these networks that small-scale interaction becomes



How to construct networks from data?

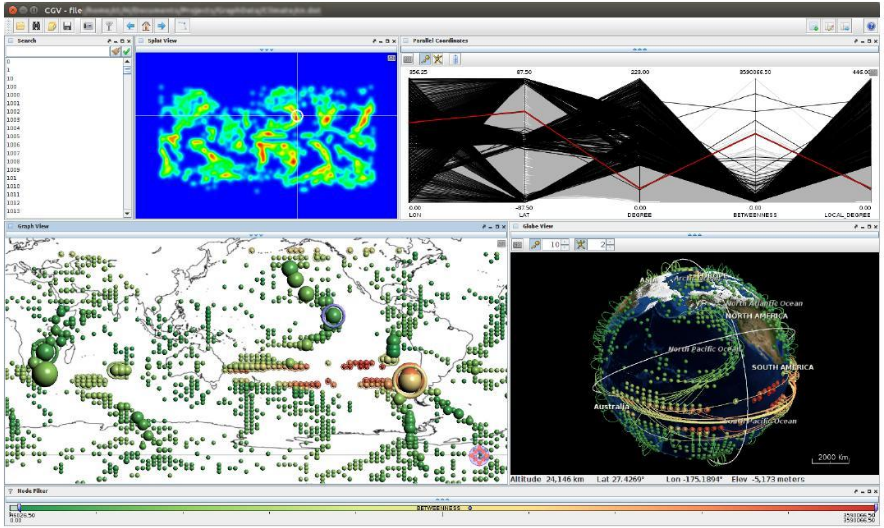
1. Directly build correspondence between links and edges with data (social networks, flights data)
2. Preprocess data (first build correlation from data)

(example of project from www.pik-potsdam.de Climate networks)



How to construct networks from data?

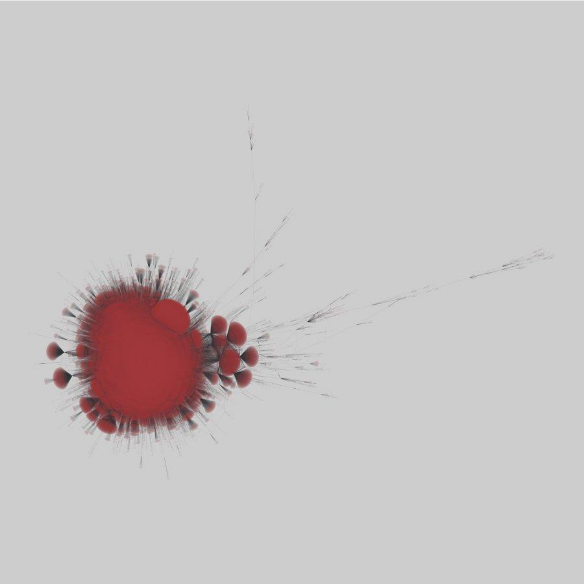
1. Directly build correspondence between links and edges with data (social networks, flights data)
2. **Preprocess data (first build correlation from data)**
   Working with data from big systems



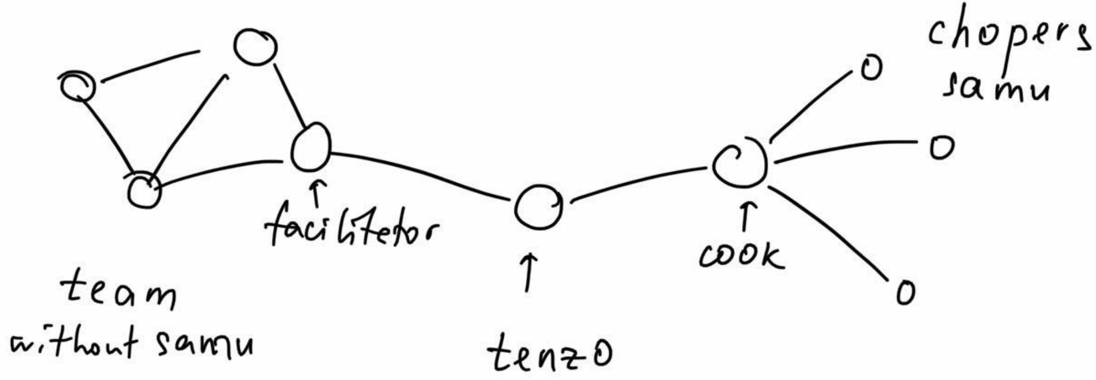
Network visualisaion

Can hairy ball be a good visualisation?
Discussions for class on visualisations

Linux kernel mailing list. A bipartite network of contributions by users to threads on the Linux kernel mailing list. 379554 nodes, 1565683 edges. https://networks.skewed.de/net/lkml_thread

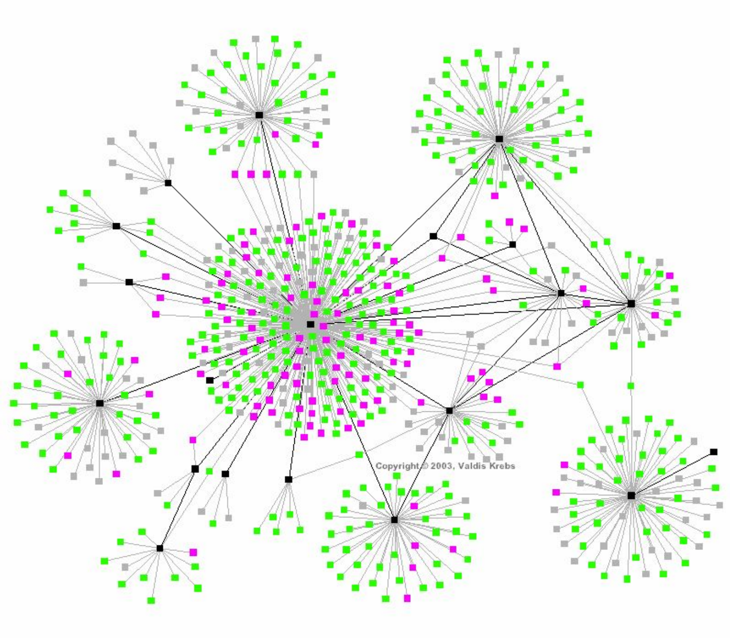


Analysis of communities



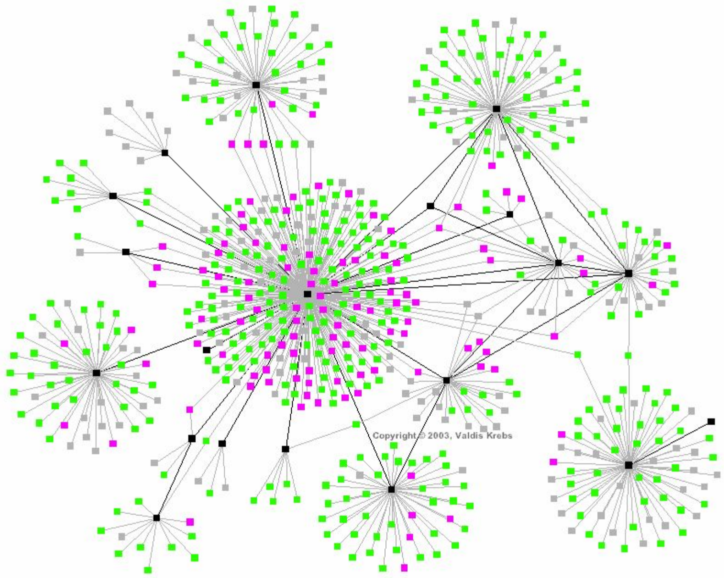
What we will look at in network science?

1. Network measures
2. Networks in time and space
3. Networks from data



What are spreading processes on networks?

Any examples of spreading processes?



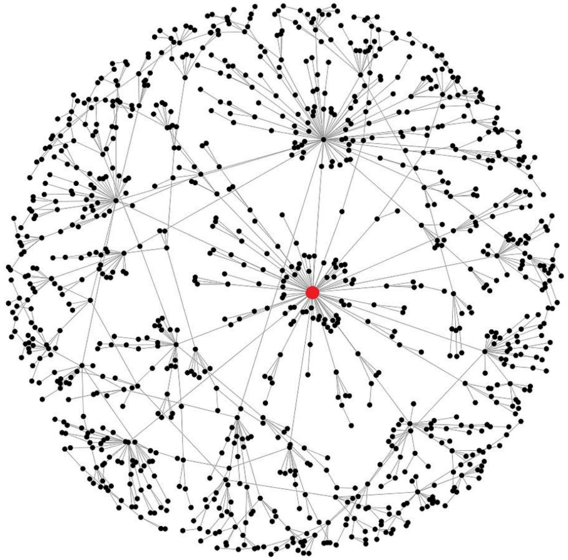
How to construct networks from data?

1. Decide what are links and nodes in your data?
2. Preprocess data (first build correlation from data)

Examples: social networks, flights data in the practical part

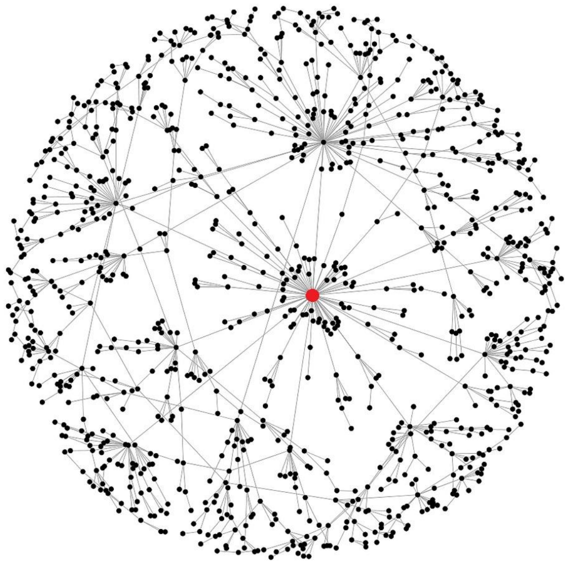
How to construct networks from data?

1. Directly build correspondence between links and edges with data (social networks, flights data)
2. Preprocess data (first build correlation from data)



How to construct networks from data?

1. Directly build correspondence between links and edges with data (social networks, flights data)
2. Preprocess data (first build correlation from data)



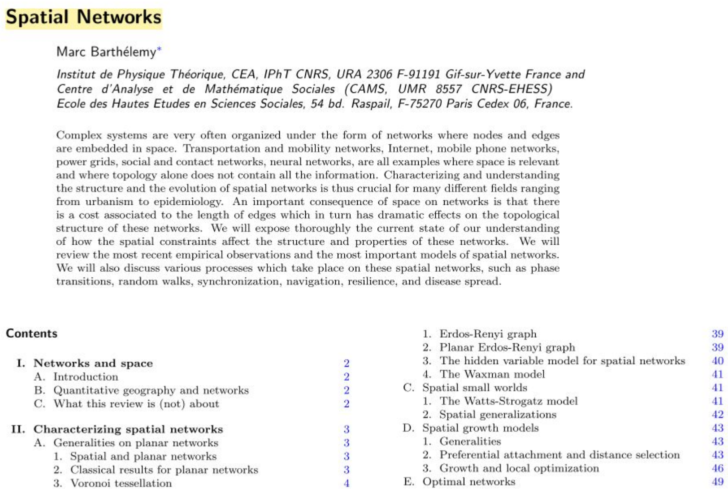
Remember that networks do not give one-to-one
Correspondence of your data.
Hence do not generalize

Network resources

http://networkrepository.com/networks.php
http://networksciencebook.com/chapter/3#advanced-b

Network resources

http://networkrepository.com/networks.php
http://networksciencebook.com/chapter/3#advanced-b



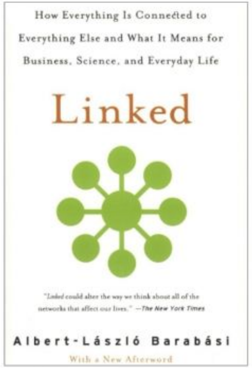
Spatial Networks

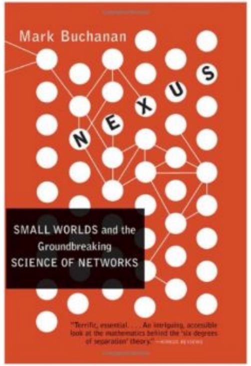
Marc Barthélemy*

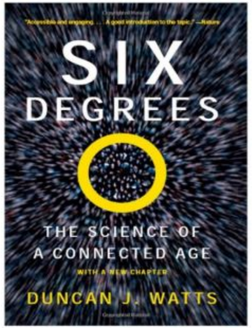
Institut de Physique Théorique, CEA, IPhT CNRS, URA 2306 F-91191 Gif-sur-Yvette France and Centre d'Analyse et de Mathématique Sociales (CAMS, UMR 8557 CNRS-EHESS) Ecole des Hautes Etudes en Sciences Sociales, 54 bd. Raspail, F-75270 Paris Cedex 06, France.

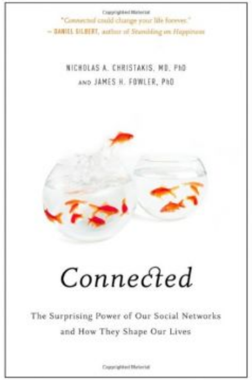
Complex systems are very often organized under the form of networks where nodes and edges are embedded in space. Transportation and mobility networks, Internet, mobile phone networks, power grids, social and contact networks, neural networks, are all examples where space is relevant and where topology alone does not contain all the information. Characterizing and understanding the structure and the evolution of spatial networks is thus crucial for many different fields ranging from urbanism to epidemiology. An important consequence of space on networks is that there is a cost associated to the length of edges which in turn has dramatic effects on the topological structure of these networks. We will expose thoroughly the current state of our understanding of how the spatial constraints affect the structure and properties of these networks. We will review the most recent empirical observations and the most important models of spatial networks. We will also discuss various processes which take place on these spatial networks, such as phase transitions, random walks, synchronization, navigation, resilience, and disease spread.

Contents









Jupyter notebooks and resources

https://classroom.google.com/u/0/w/NTQ0NTIwOTczMTQ2/t/all