Introduction to data science and network science

Liubov Tupikina ITMO, 2024-2025

Figure 7.11

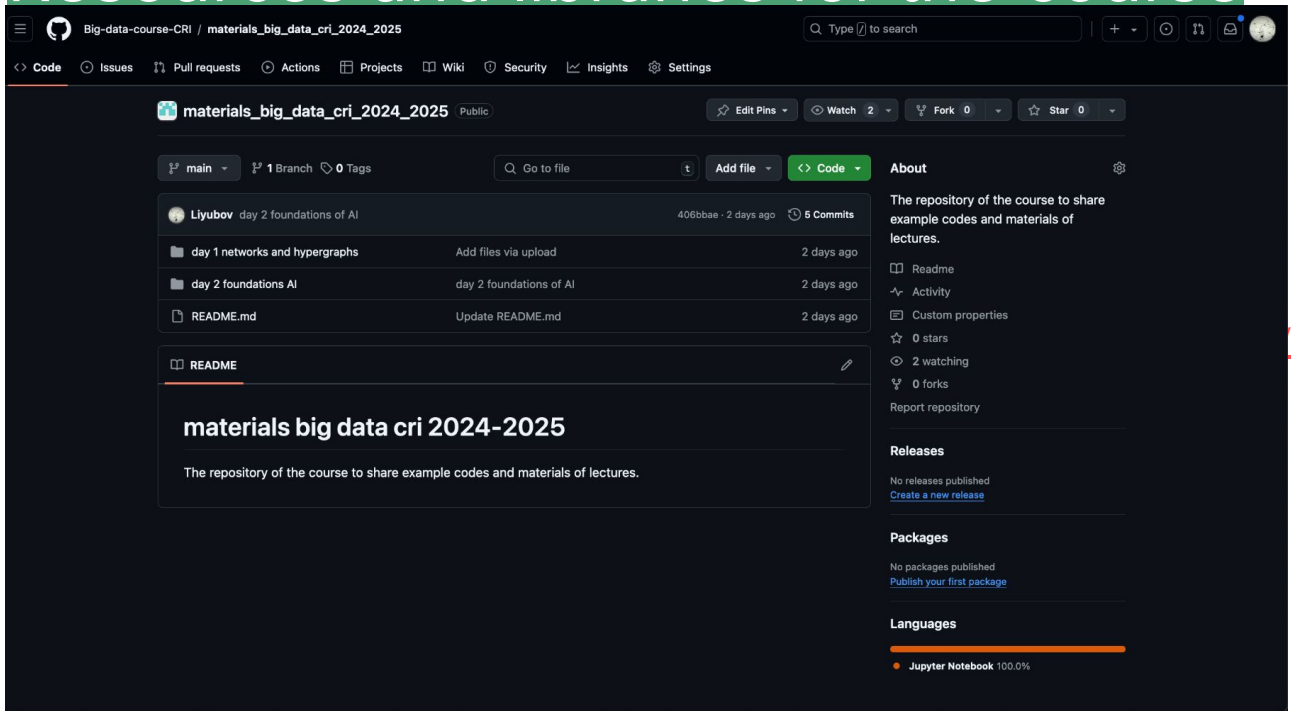
Resources and libraries for the course

Standard libraries (Python): numpy, matplotlib, scikit learn, seaborn
Network libraries: networkx, osmnx (open street data), PySal

Support materials

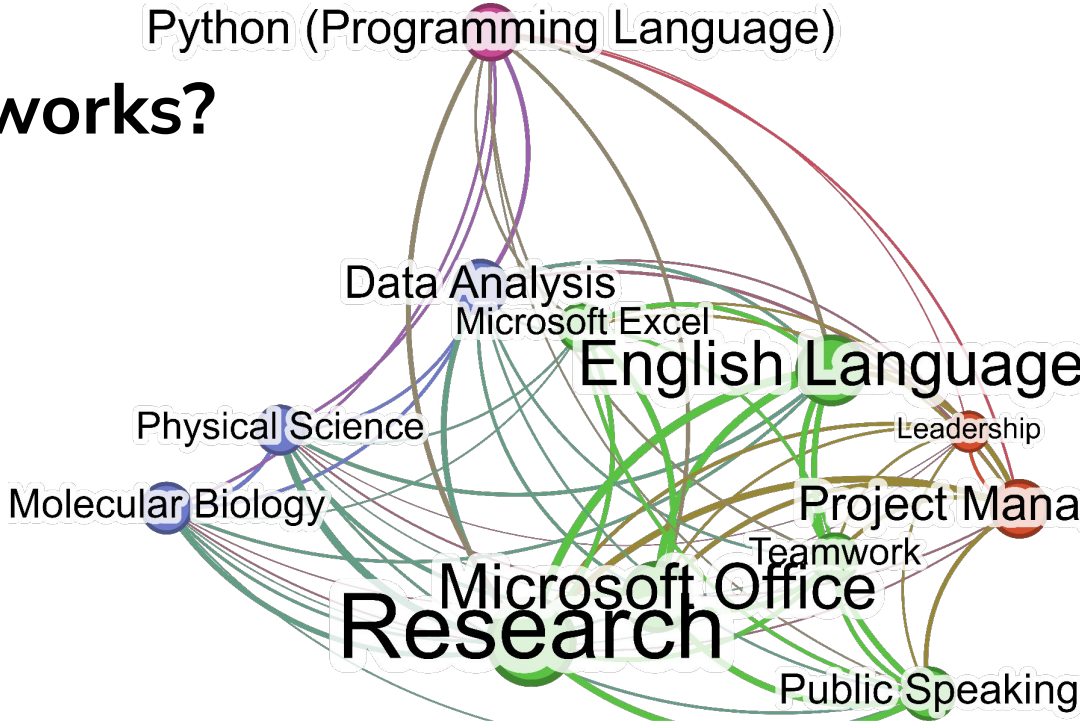
- Big data course Marc and Liubov https://github.com/Big-data-course-CRI/
- Correlaid, Complex system conference CSS 2023 and TidyTuesday https://github.com/rfordatascience/tidytuesday
- Network science book http://networksciencebook.com/
- Network repository https://networks.skewed.de/
- Visualisation tools https://gephi.org/users/download/
- Network datasets https://www.complex-networks.net/datasets.html#chap8

Resources and libraries for the course



https://github.com/Big-data-course-CRI/materials_big_data_cri_2024_2025

What are graphs / networks?

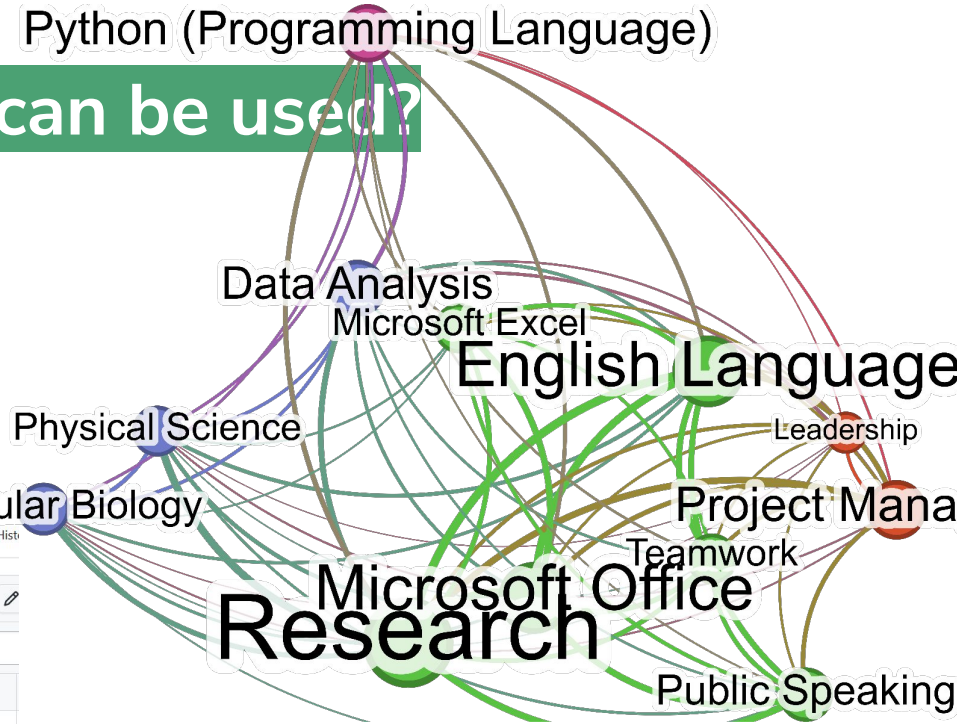


Where graphs / networks can be used?

Networks are good to represent data.

What are structures which we can process using networks?

See other examples of data in Github of the course



ziqingcheryl organized_version 688a7a3 · 9 months ago

Preview | Code | Blame | 8 lines (8 loc) · 56.3 KB | Raw

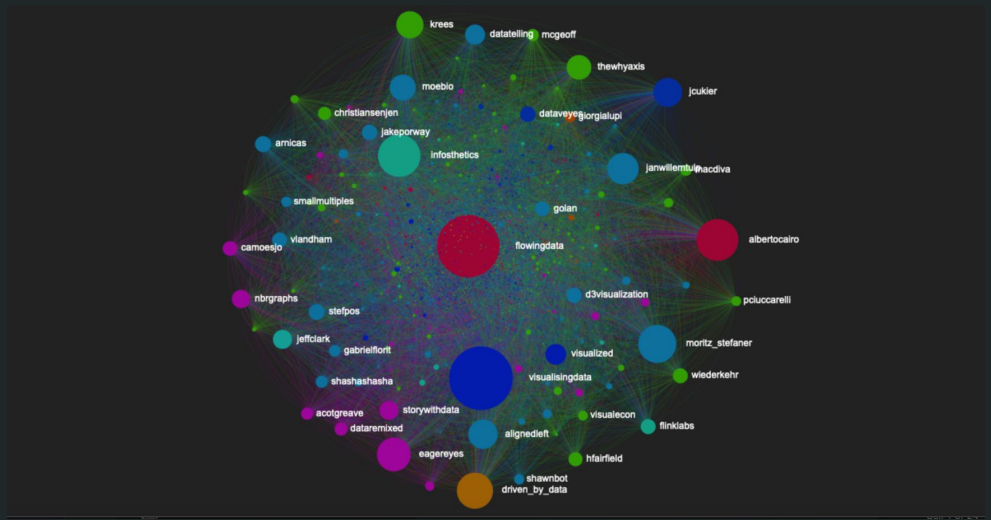
Search this file

		Name	Occupation	City	State	Country
2	1	7f87716741aeae4d227491b5a7d87d4e	Stagiaire at INSTITUT GUSTAVE ROUSSY			France
3	2	c10674167315089247ea5fa8c98256f6	Initiatrice de projet at Tous Tes Possibles	Paris	Île-de-France	France
4	3	bf83d3cbf7ebfeeecdedd9519df56af8	PhD Student at Medical University of Vienna	Österreich		Austria
5	4	94369f5f833008a9b93d6e9ae7a6a533	Chargée de communication grand public et jeunes at ADEME	Paris	Île-de-France	France
6	5	67af1831de65104374b77b9a597f4671	Stagiaire UX/ Product Owner at Tyllt	Talence	Nouvelle-Aquitaine	France
7	6	95dcb67fb5d78d0c099249d553343451	Research Associate at King's College London			United Kingdom

Which data can be represented using networks representation?

https://www.wired.com/2014/09/coupland-bell-labs/

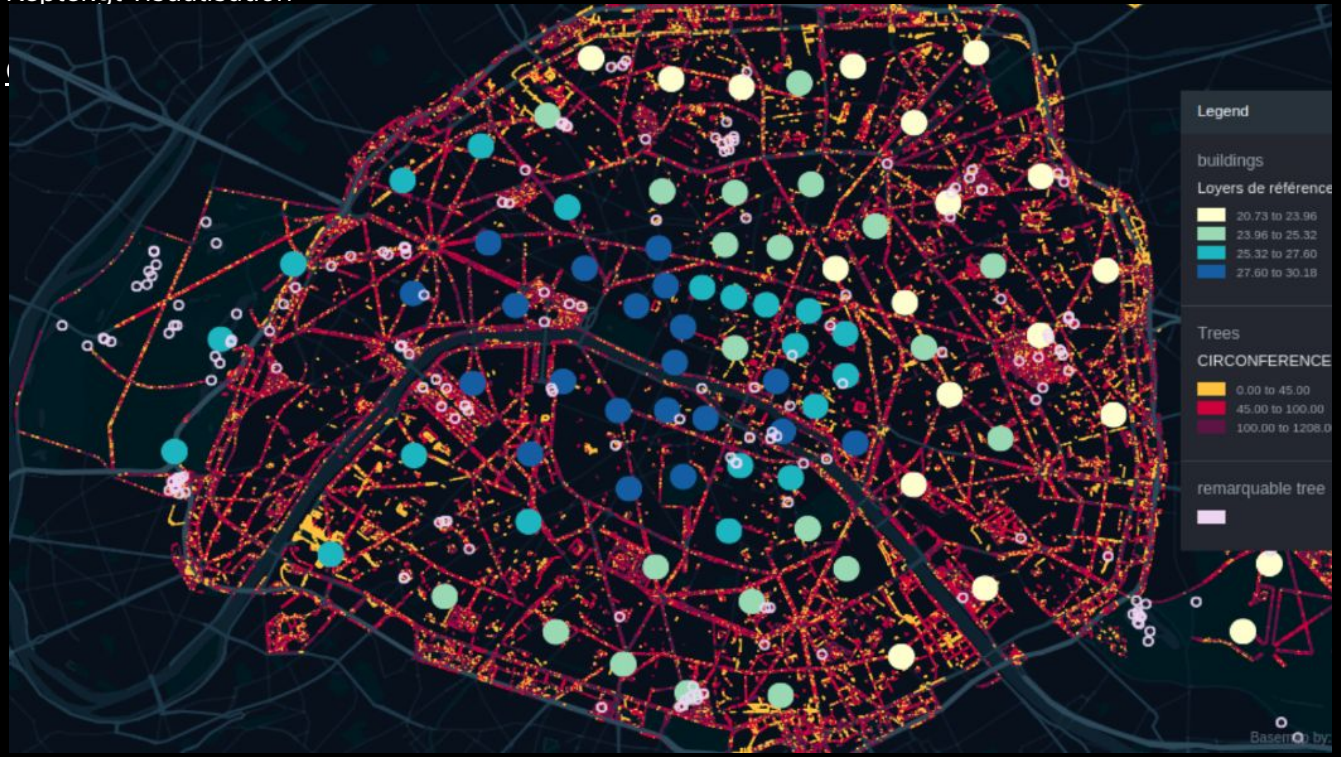
Which data can be represented using networks representation?



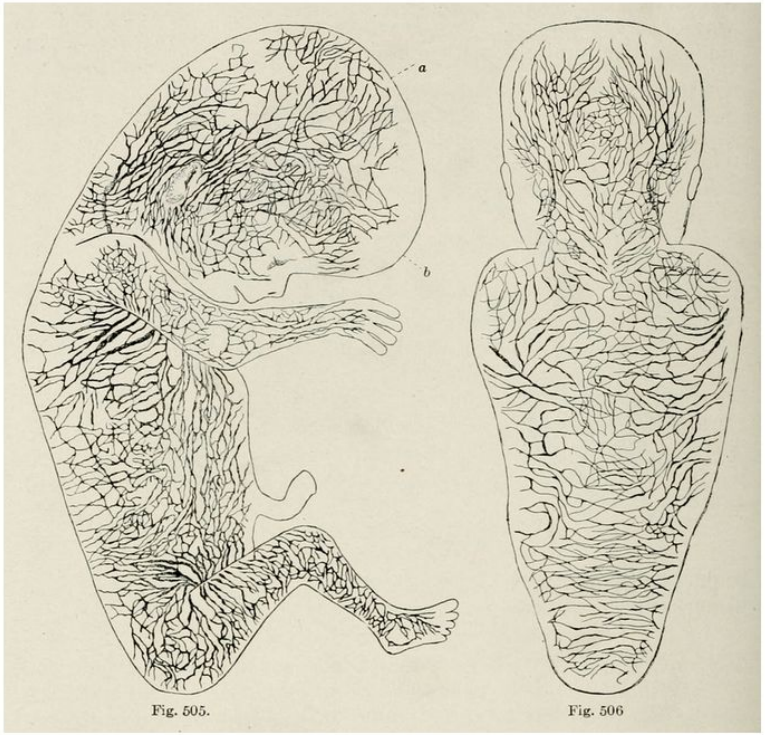
Social media: Twitter, Instagram, Facebook data analysis
Ask Marc Santolini on his project or check other projects
https://exploring-data.com/vis/visualisingdata-census-twitter-network/#infosthetics

Networks and hypergraphs in data

Network data: Olympics in Paris, statistics, the Guardian data stories, Humanitarian data HDX
Kepler.gl visualisation

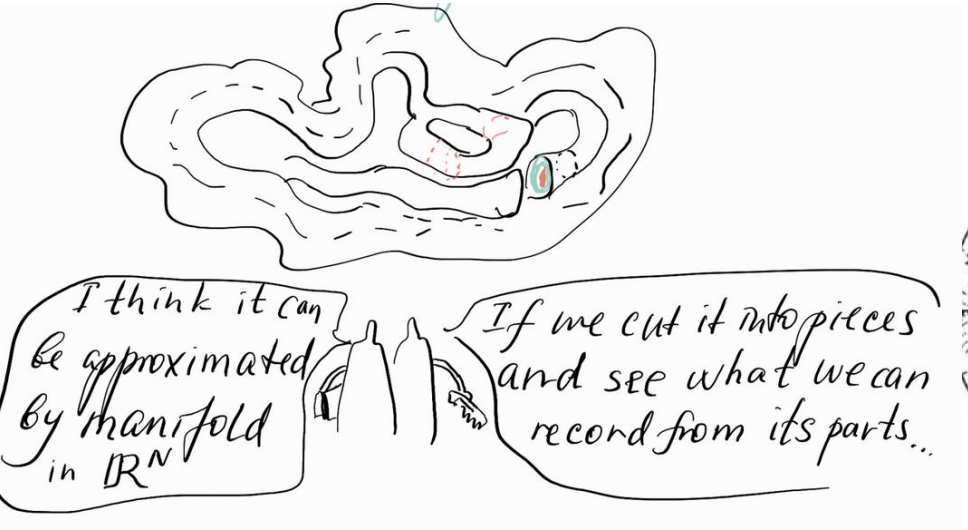


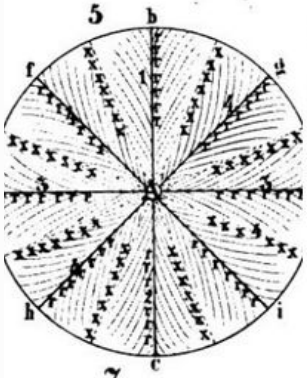
NETWORKS DESCRIBE HOW THINGS CONNECT AND INTERACT

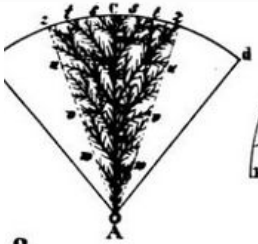


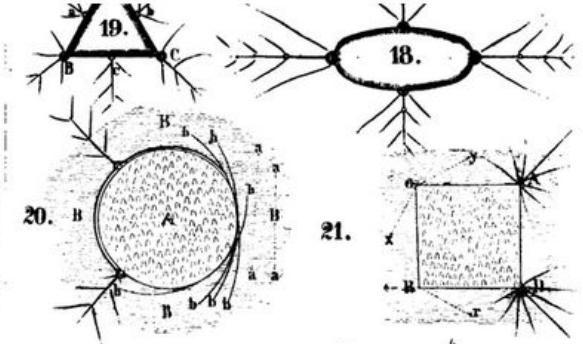
Distension of the lymphatic vessels in the human foetus, from Franz Kreibel, *Manual of human embryology*, 1910

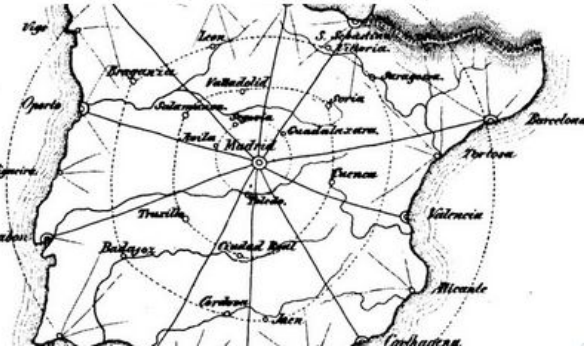
NETWORKS DESCRIBE HOW THINGS CONNECT AND INTERACT











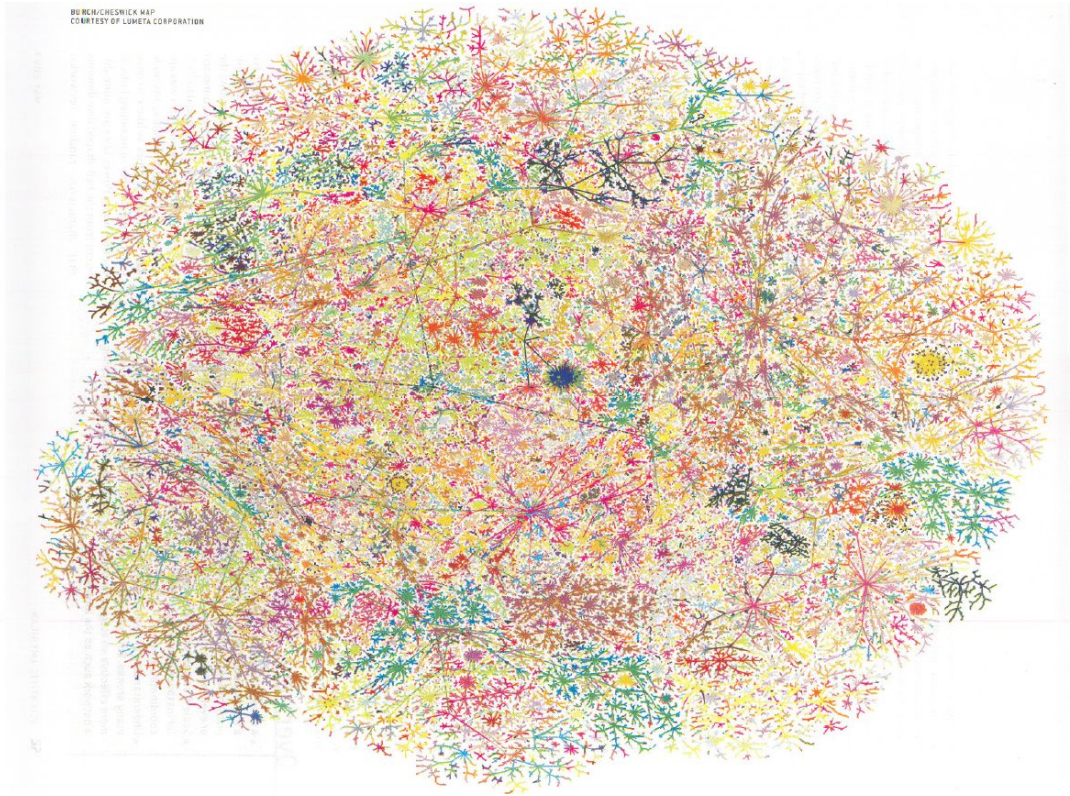
Network theory, graph theory

1. Implementation of algorithms using networks
2. Representing your data using networks
3. Ways of solving problems using networks representation

What is network science?

One 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.
(P. Holme) https://petterhol.me/



Network science

Network **measures and network types**

Networks in **time and space, dynamics on** networks

Networks from **data**

Figure 7.11

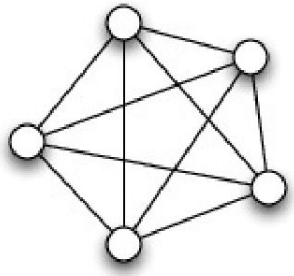
How did the network science start?

http://networksciencebook.com/chapter/2#bridges

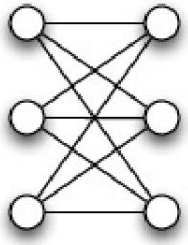
Network science, graph and topology theory

- Graph theory:
  Koenigsberg problem 1736
  Eurlerian paths algorithms 1873

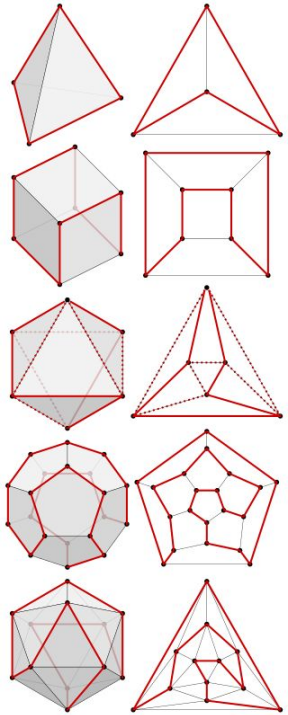
- Soft matter physics



K_5

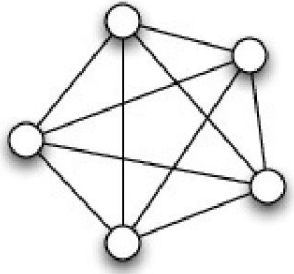


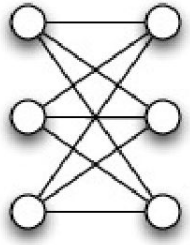
$K_{3,3}$



Network science and graph theory

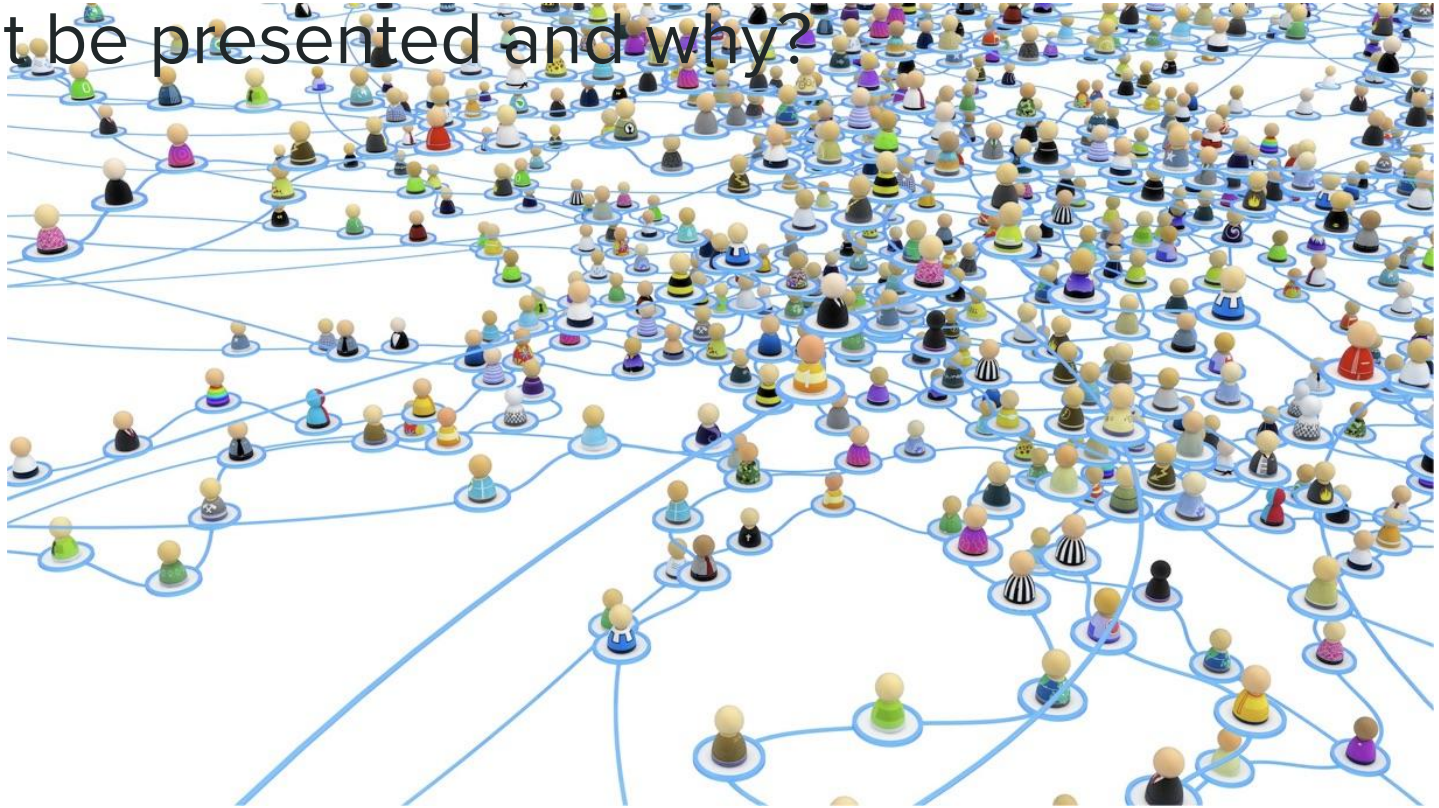
Graph (discrete mathematics),
a structure made of vertices V
and edges E (subset of two vertices).





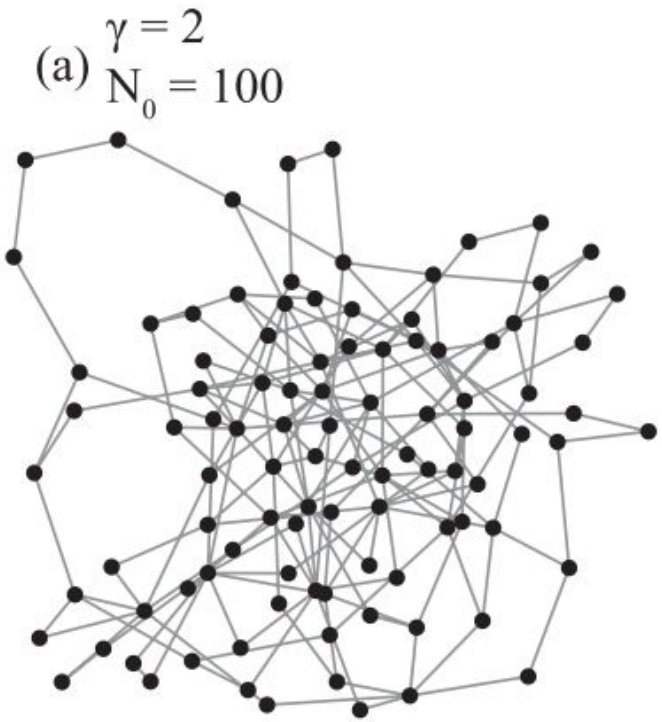
Networks (from real world)
can be represented/encoded as graphs.

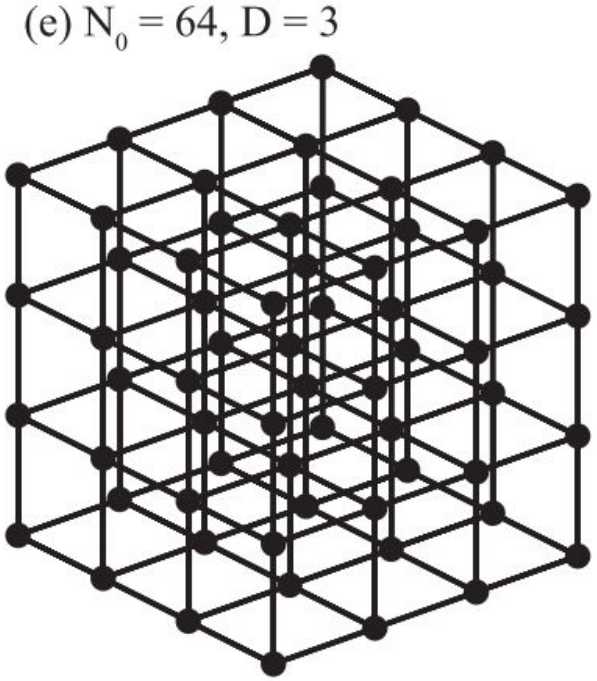
Can anything be presented as a network?
What cannot be presented and why?



Examples of network representations

M.Stella et al.





What we will look at in network science?

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

Defining a network to the computer

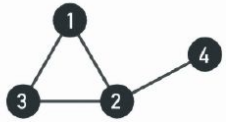
How to represent a network: **edgelist** and **adjacency matrix.**

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

a. **Adjacency matrix**

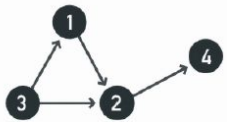
$$A_{ij} = \begin{matrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{matrix}$$

b. **Undirected network**



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

c. **Directed network**

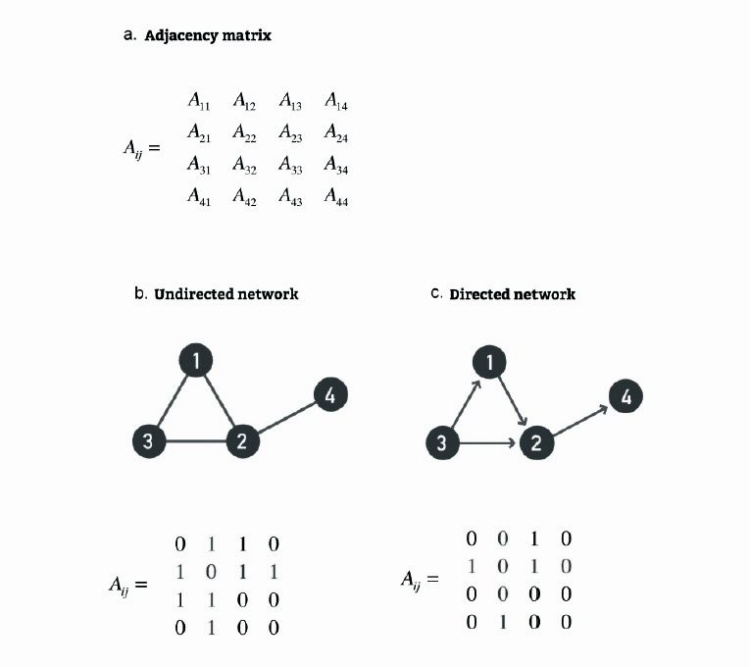


$$A_{ij} = \begin{matrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

Network definitions

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

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



Questions to check:

How to encode the network data to computer?

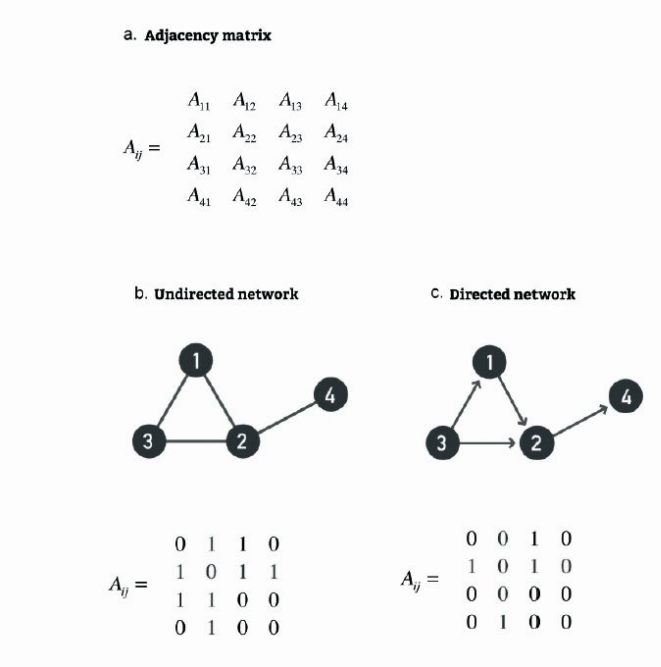
What is the most efficient representation of a network for computer?

Notebook to encode the network

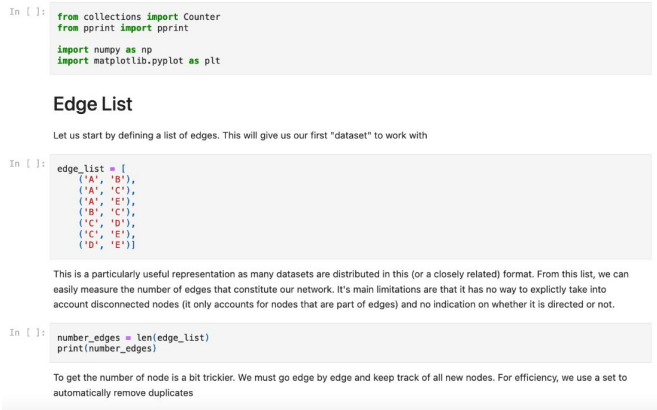
Network definitions

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

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



[Notebook](#) to encode the network



```
In [ ]:
from collections import Counter
from pprint import pprint

import numpy as np
import matplotlib.pyplot as plt
```

Edge List

Let us start by defining a list of edges. This will give us our first "dataset" to work with

```
In [ ]:
edge_list = [
    ('A', 'B'),
    ('A', 'C'),
    ('A', 'E'),
    ('B', 'C'),
    ('C', 'D'),
    ('C', 'E'),
    ('D', 'E')]
```

This is a particularly useful representation as many datasets are distributed in this (or a closely related) format. From this list, we can easily measure the number of edges that constitute our network. It's main limitations are that it has no way to explictly take into account disconnected nodes (it only accounts for nodes that are part of edges) and no indication on whether it is directed or not.

```
In [ ]:
number_edges = len(edge_list)
print(number_edges)
```

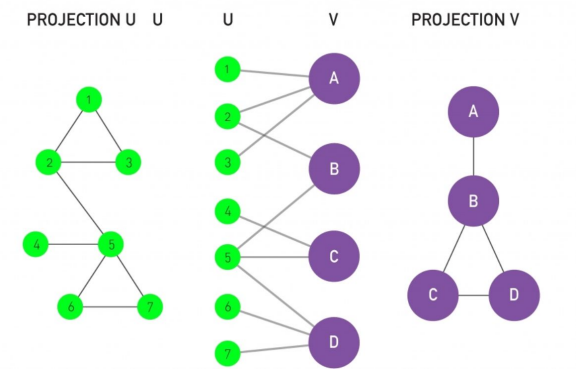
To get the number of node is a bit trickier. We must go edge by edge and keep track of all new nodes. For efficiency, we use a set to automatically remove duplicates

Network types based on links, nodes properties

How to encode some more information which we want to include into our network?

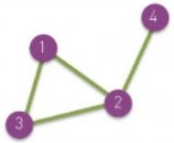
E.g. bipartite networks, or layed networks?

Try it yourself in the Notebook
Try it with networkX precoded library Notebook here



Network types based on links, nodes properties

a. Undirected

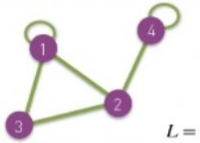


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

b. Self-loops

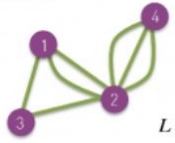


$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

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

c. Multigraph
(undirected)

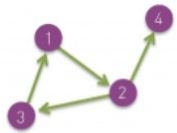


$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

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

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

d. Directed

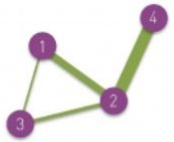


$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

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

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

e. Weighted
(undirected)

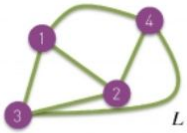


$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

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

$< k >= \frac{2L}{N}$

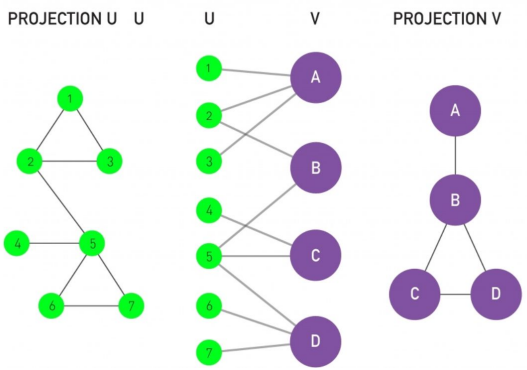
f. Complete Graph
(undirected)



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

$A_{ii} = 0 \qquad A_{i \neq j} = 1$

$L = L_{max} = \frac{N(N-1)}{2} \qquad < k >= N - 1$



Network types based on links, nodes properties

Contact

Mailing list
Issue tracker
Source

Releases

Stable (notes)
3.3 — April 2024
Documentation

Latest (notes)
3.4 development
Documentation

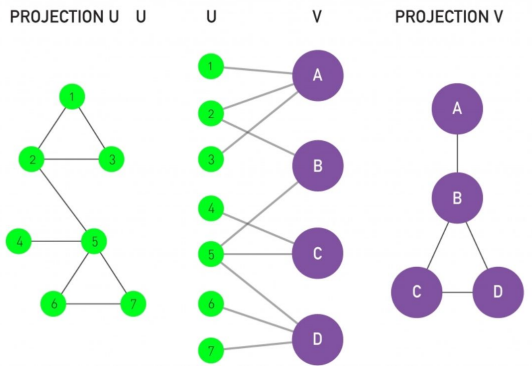
Archive

NetworkX
Network Analysis in Python

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Software for complex networks

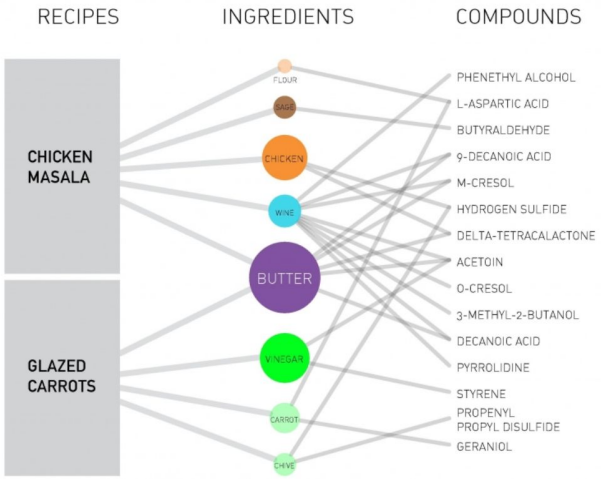
- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

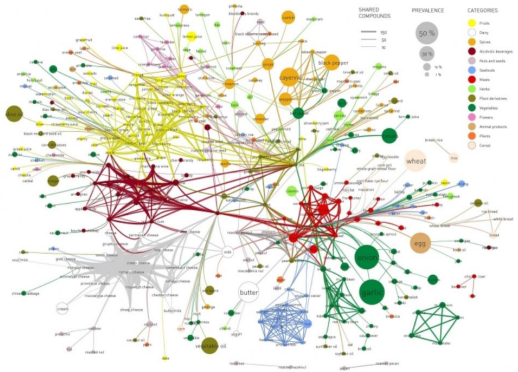


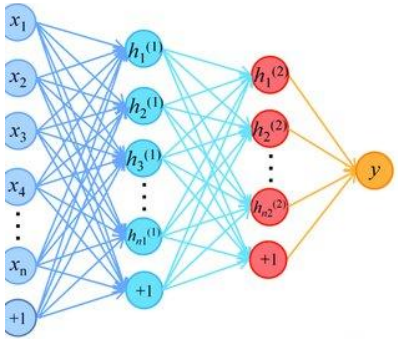
https://networkx.org/documentation/stable/tutorial.html

Network types based on links, nodes properties

bipartite networks





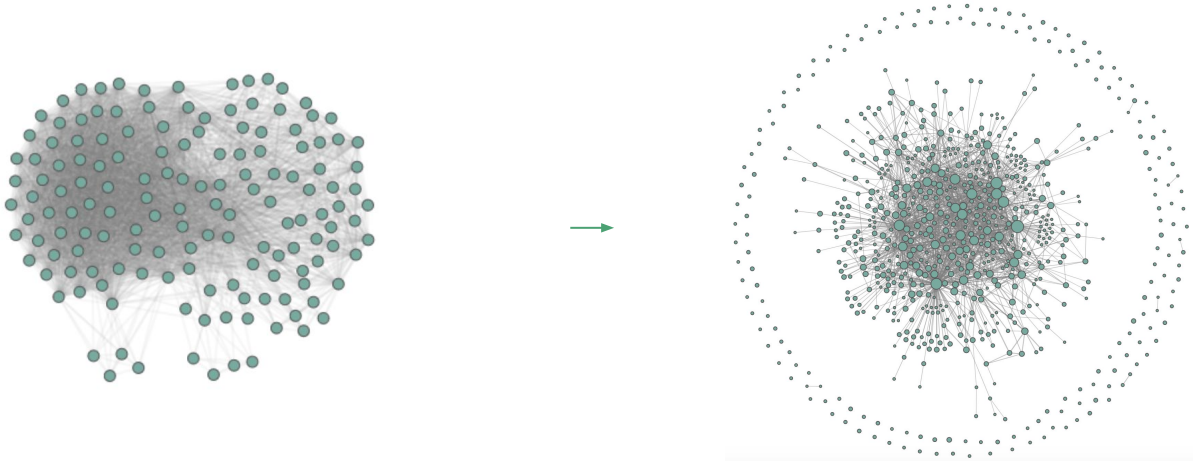


Network measures

Main idea is to characterise their properties.
Local measures for each node (degree)
Global measures for the whole network (**density** - number of links normalised by number of links in a **complete graph)**

Network measures and layout

Local measures for each node. **Global measures** for the whole network.
Layouts of the same network (left, right).



Network statistics

Network statistics

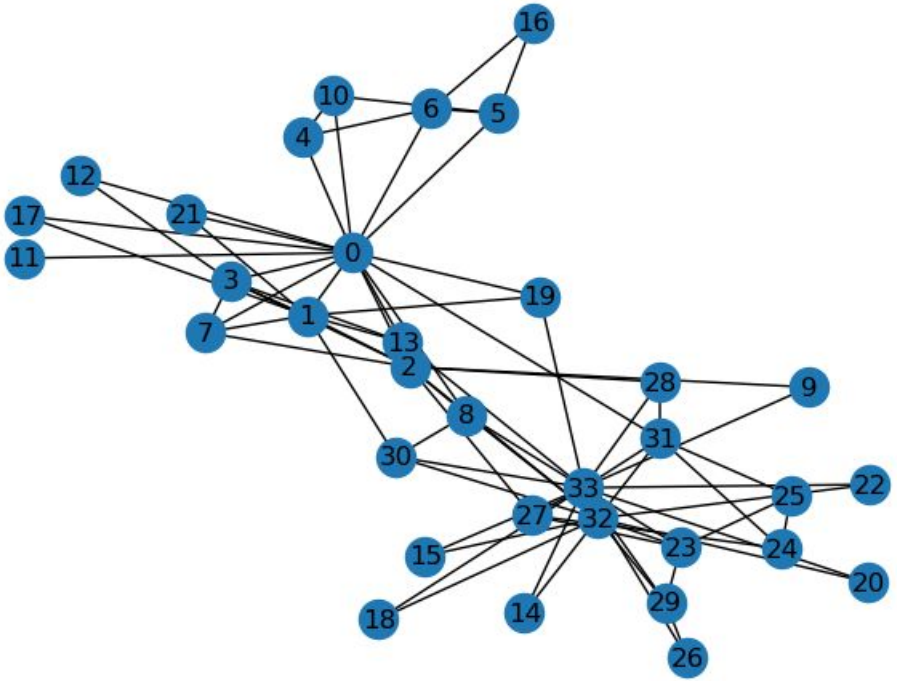
What are nodes with highest centrality?

Size n = 34
Volume m = 78
Loop count l = 0
Triangle count t = 45
Square count q = 154 4-
Maximum degree dmax = 17
Average degree d = 4.588 24
Size of Large Connected Component N = 34
Diameter δ = 5 50
Median distance δM = 2
Mean distance δm = 2.443 26
Gini coefficient G = 0.385 370
Power law exponent γ = 1.780 96

Quick check-in

What are network measures for this network?

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



Network statistics

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.