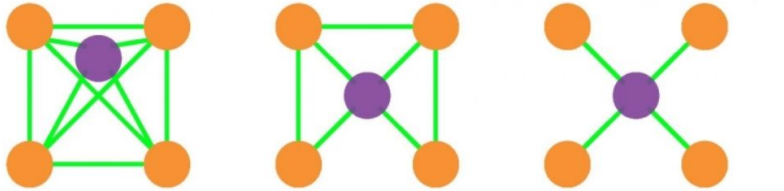
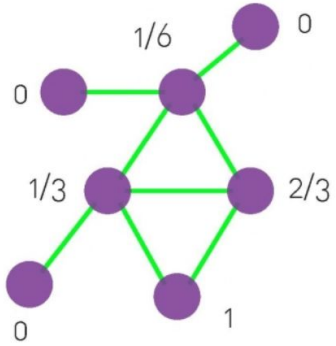
Network statistics

Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.

a.



b.

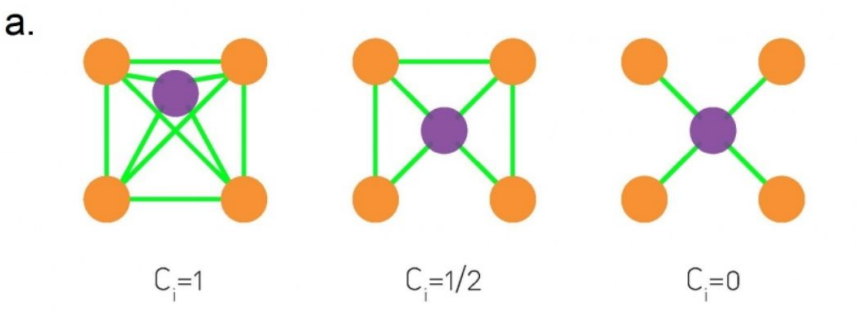


$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C_{\triangle} = \frac{3}{8} = 0.375$$

Network statistics

Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.



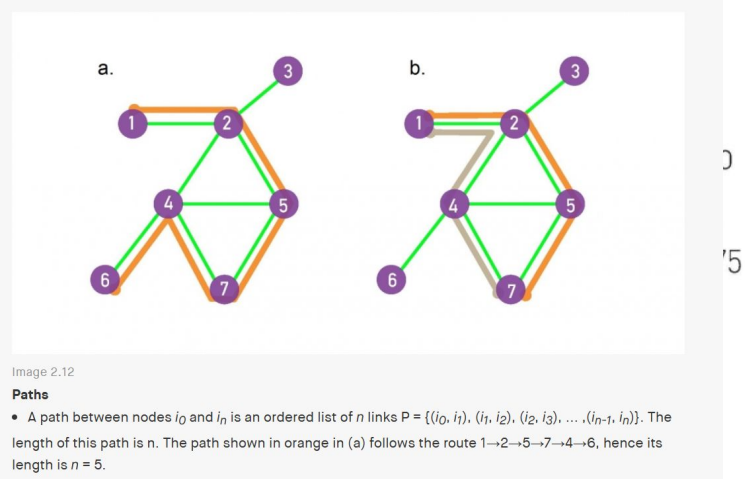


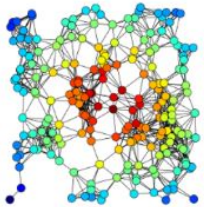
Image 2.12

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.

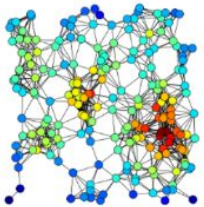
Shortest path finding between nodes are used in many algorithms for networks.
Example path between node 1 and node 6 in a graph is then encoded as a sequence of nodes, e.g. (1,2,5,7,4,6). One of the most known shortest path algorithm is Dijkstra's algorithm (1956).

Network measures



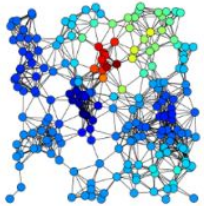
B

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 $C_i = \frac{1}{\sum_j d(n_i, n_j)}$
	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 $\langle C \rangle = \frac{1}{N} \sum_{n=i}^{N} c_i$
ty, 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

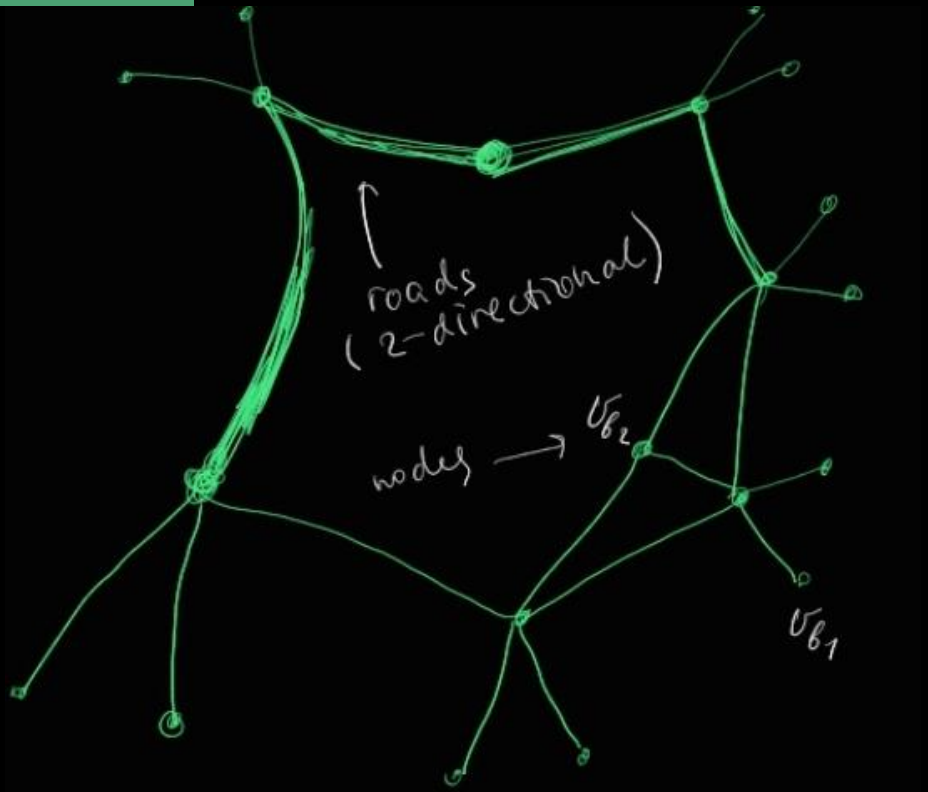
Network measures

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

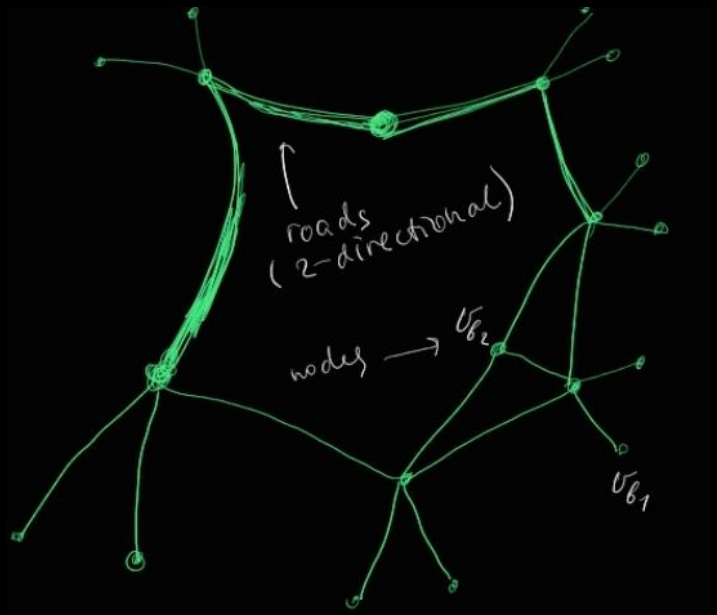
Table 2: Definitions of network science terms and variables.

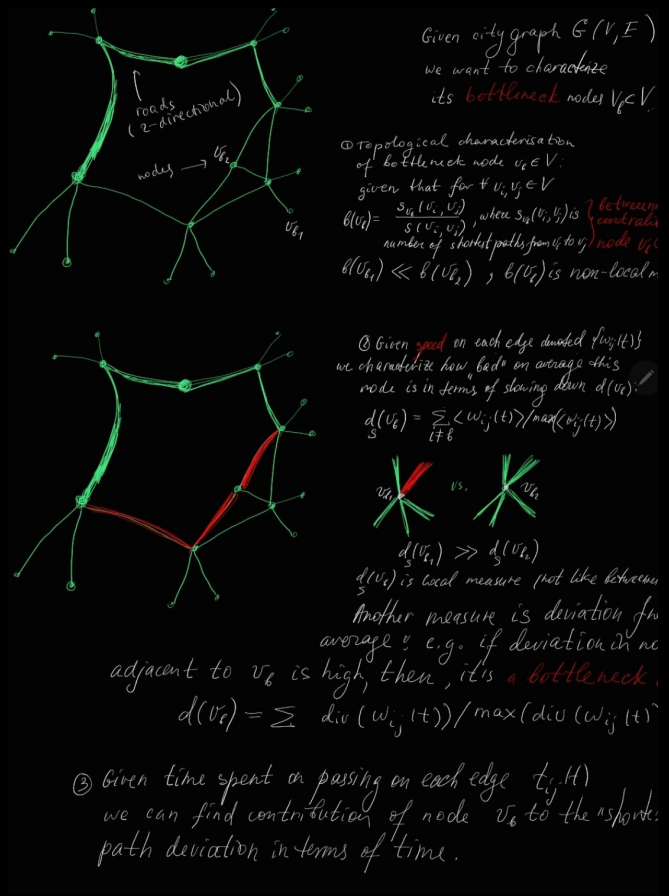
Term/variable	Definition
N	number of nodes, N, in graph
E	number of edges, E, in graph
network density	ratio of the number of edges to the maximum number of possible edges $\frac{2E}{N(N-1)}$
distance, $d(n_i, n_j)$	shortest path between node i and node j $d(n_i, n_j)$ where $n_i, n_j \in N$
average 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 $\langle k \rangle = \frac{1}{N} \sum_{n=i}^{N} k_i$
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

Getting intuition to work with networks



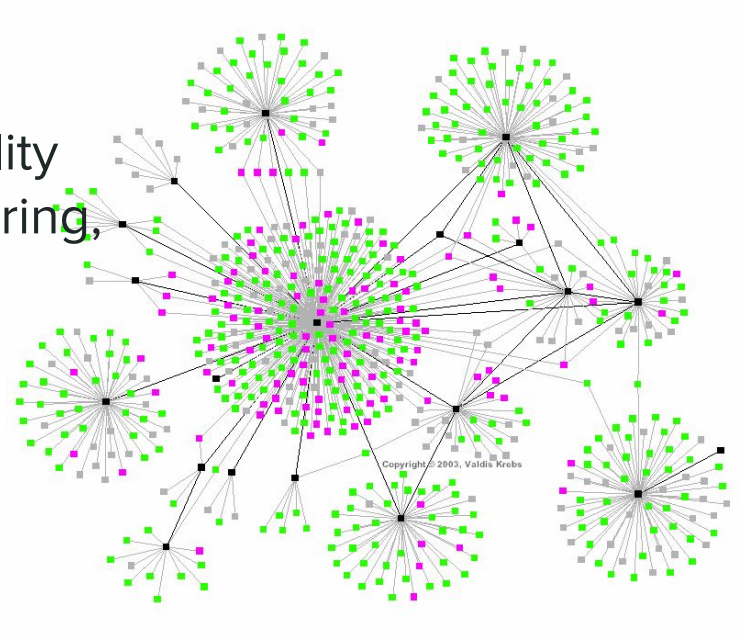
Getting intuition to work with networks





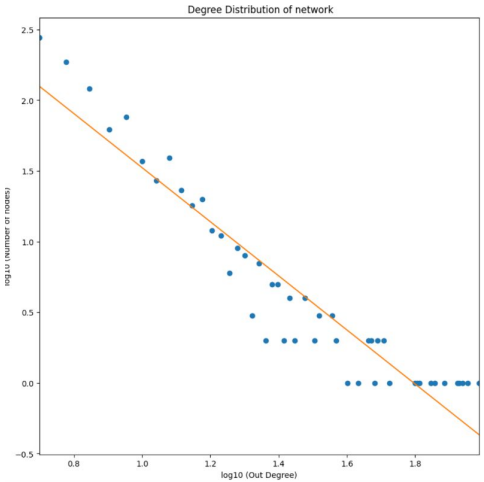
Getting intuition to work with networks

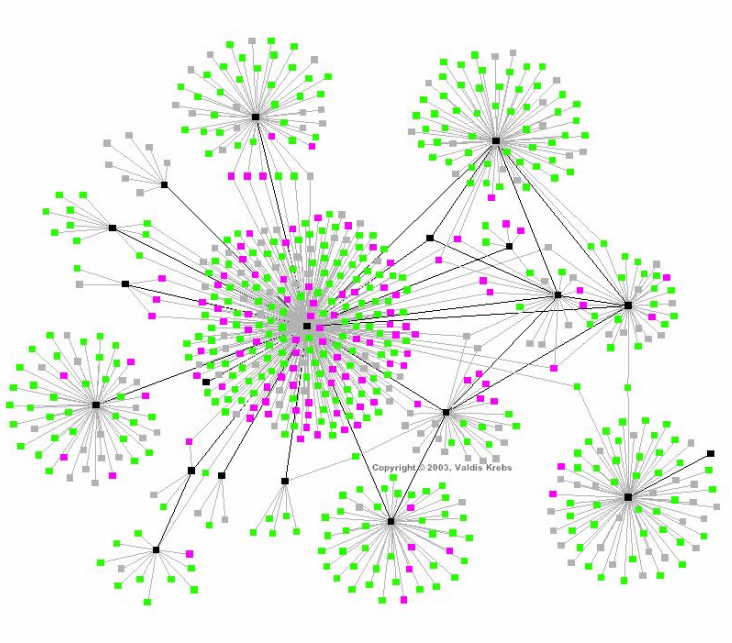
Bottlenecks - high betweenness centrality
Outliers - distributions of degree, clustering,
Betweenness
Hubs - nodes with high degree



Getting intuition to work with networks

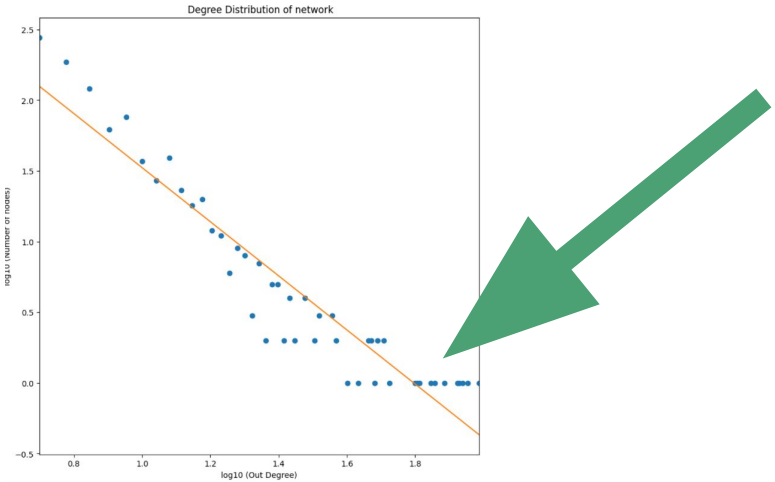
Bottlenecks - high betweenness centrality
Outliers - distributions of degree, clustering, betweenness

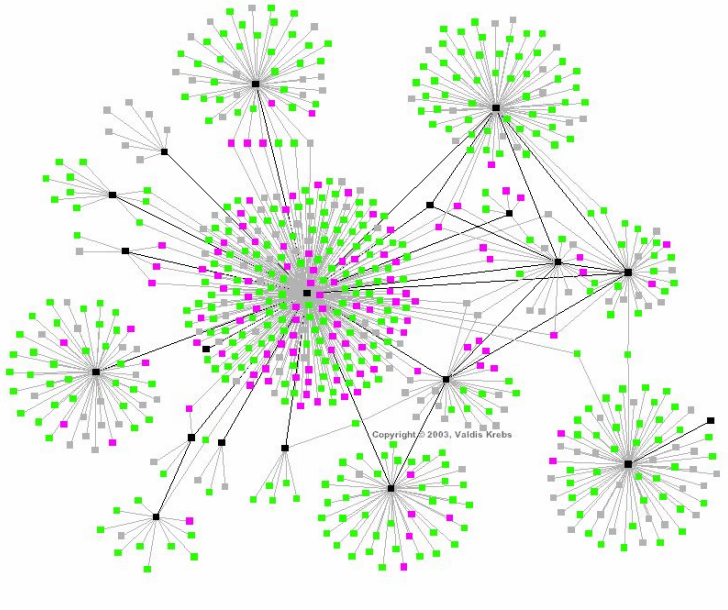




Getting intuition to work with networks

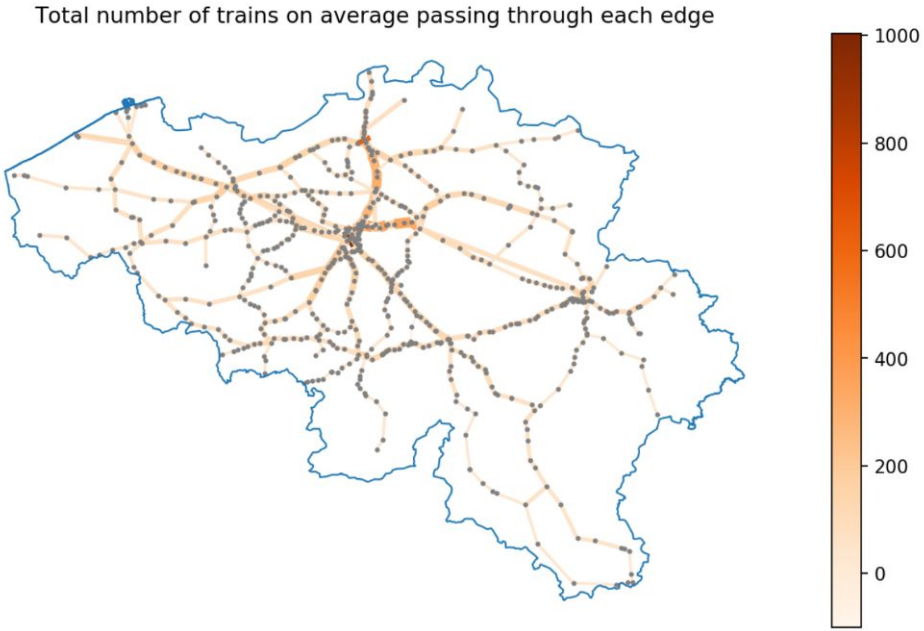
Outliers - from distributions of degree, clustering, betweenness
Notebooks from github





Getting intuition to work with networks

What are relevant questions to ask??



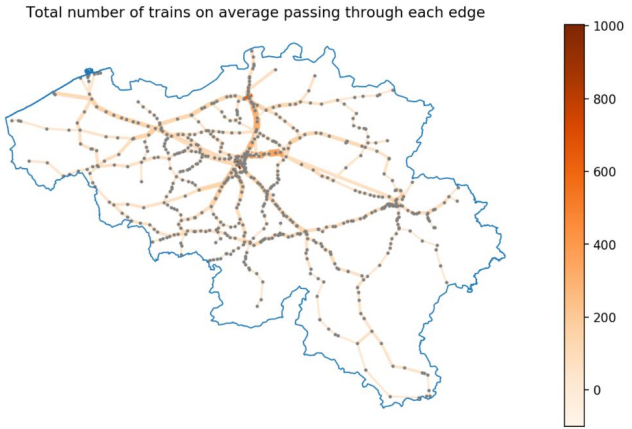
Dataset of train network

Getting intuition to work with networks

What are global characteristics to characterize bottleneck nodes in the transport network?

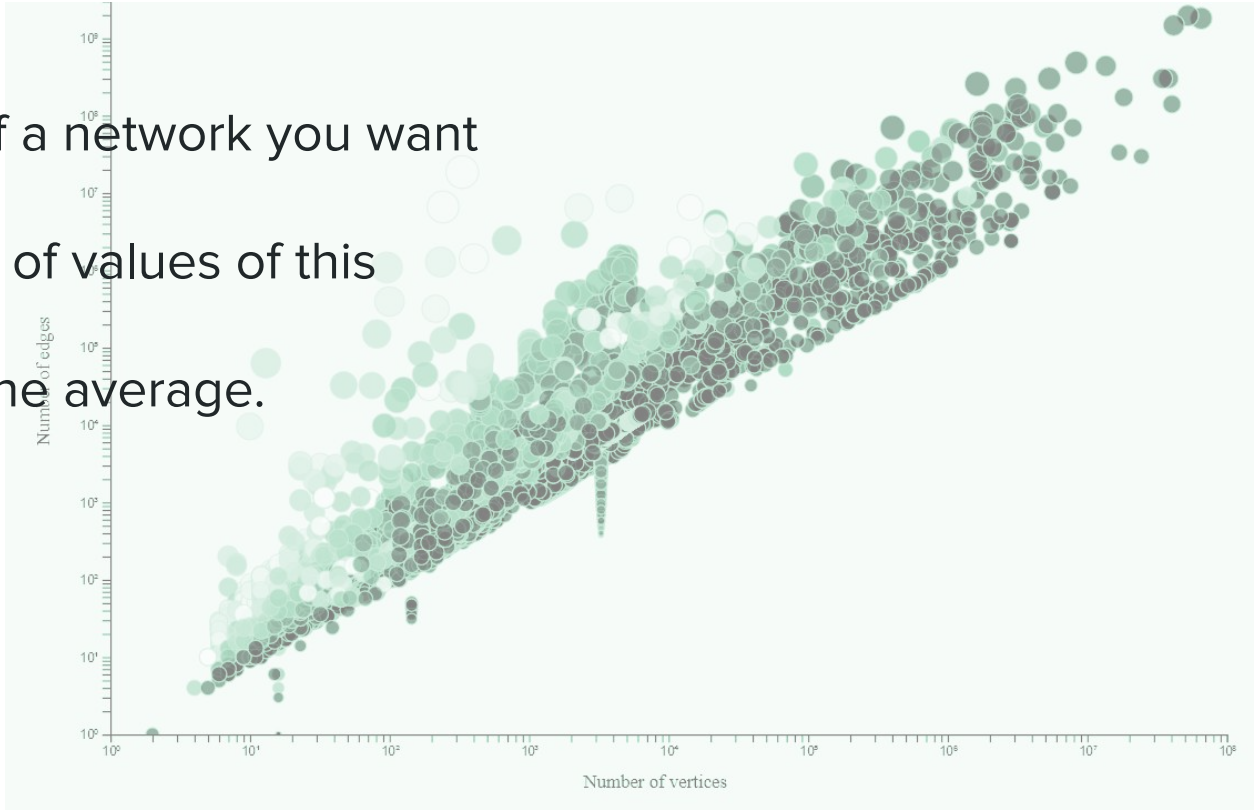
What are nodes with highest traffic?

Dataset of train network



How to make plots for your network?

Choose characteristics of a network you want to look at.
Calculate the distribution of values of this characteristics.
Compare each value to the average.

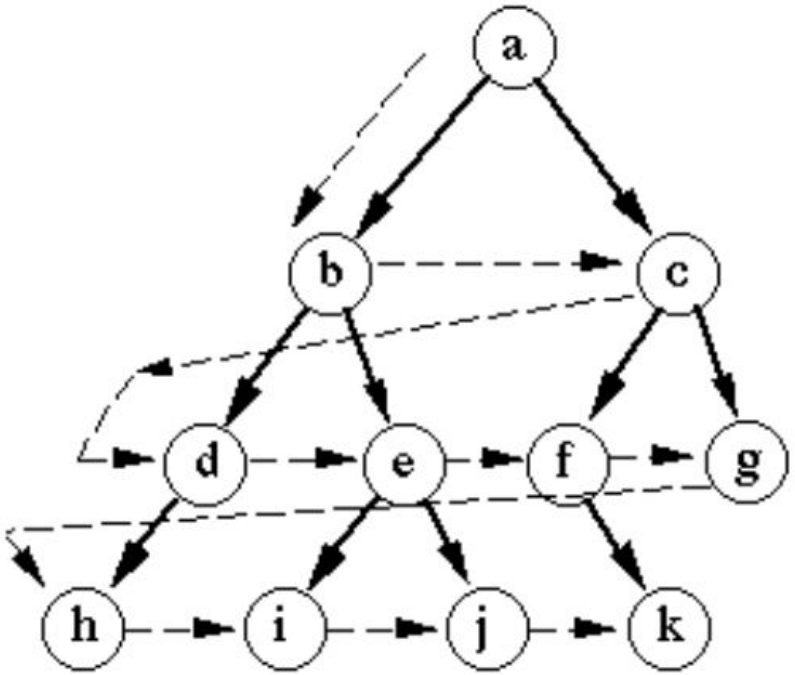


[Notebook]
[Datasets]

Algorithms on networks

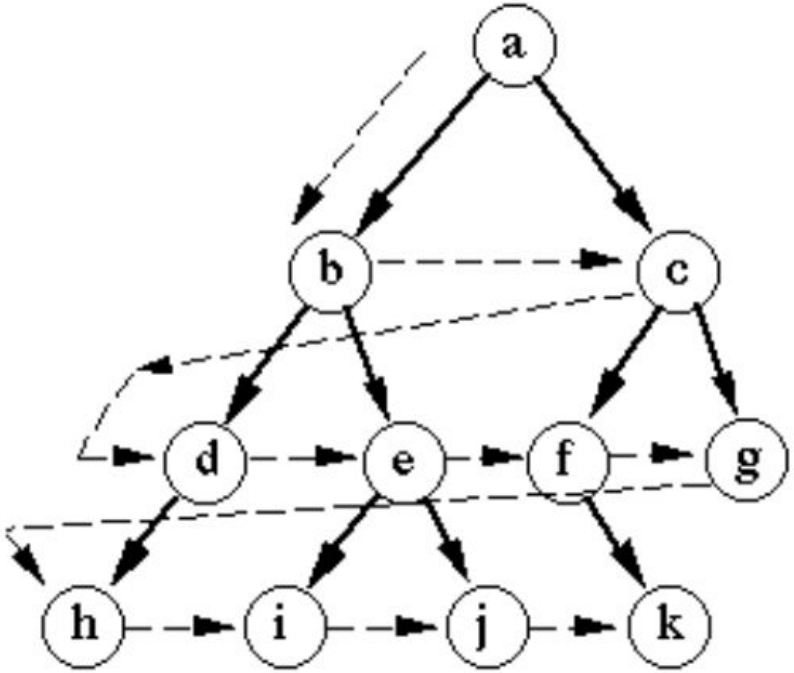
BFS algorithm
Pseudocode idea

BFS or “why your code on large networks is so slow”?

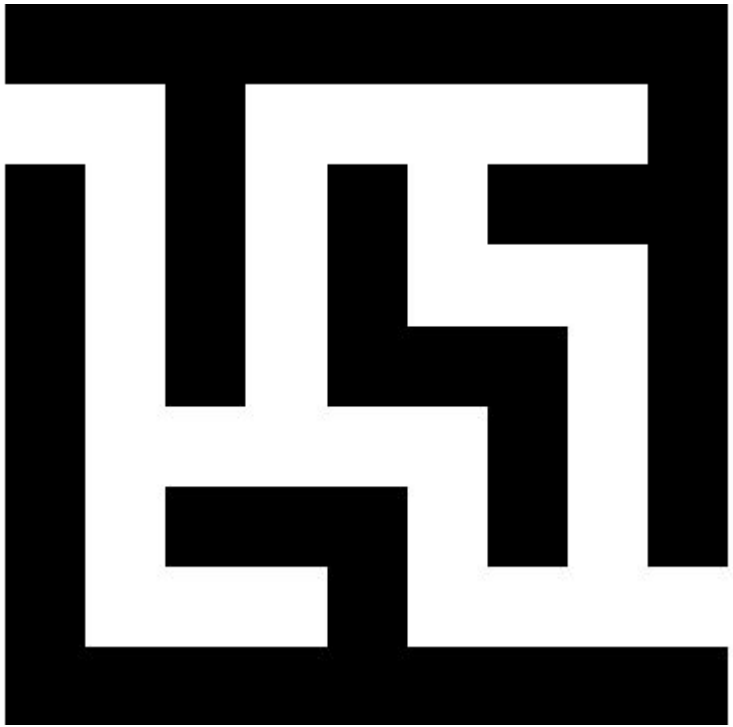


Breadth-first search

Algorithms on networks



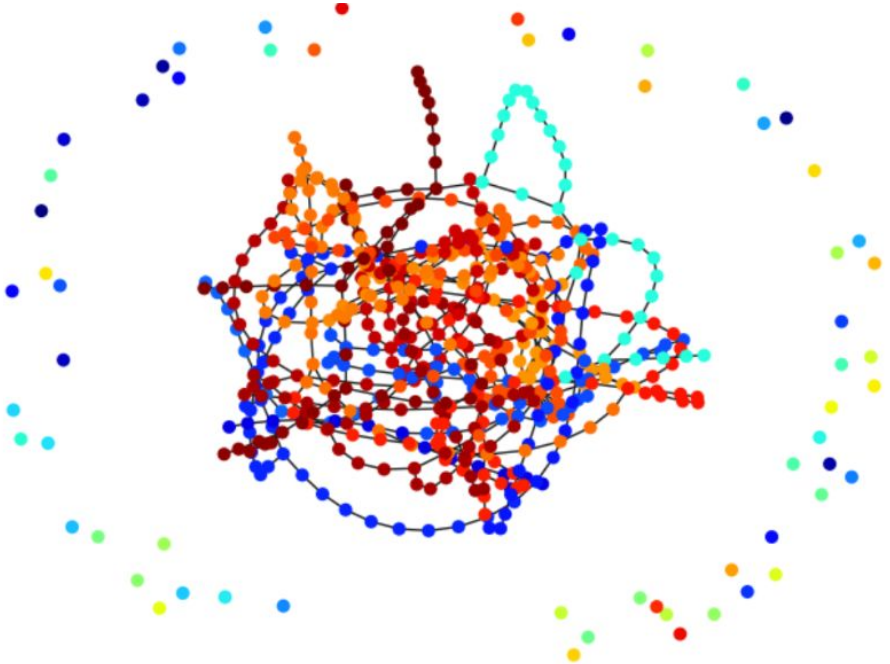
Breadth-first search



Algorithms on networks

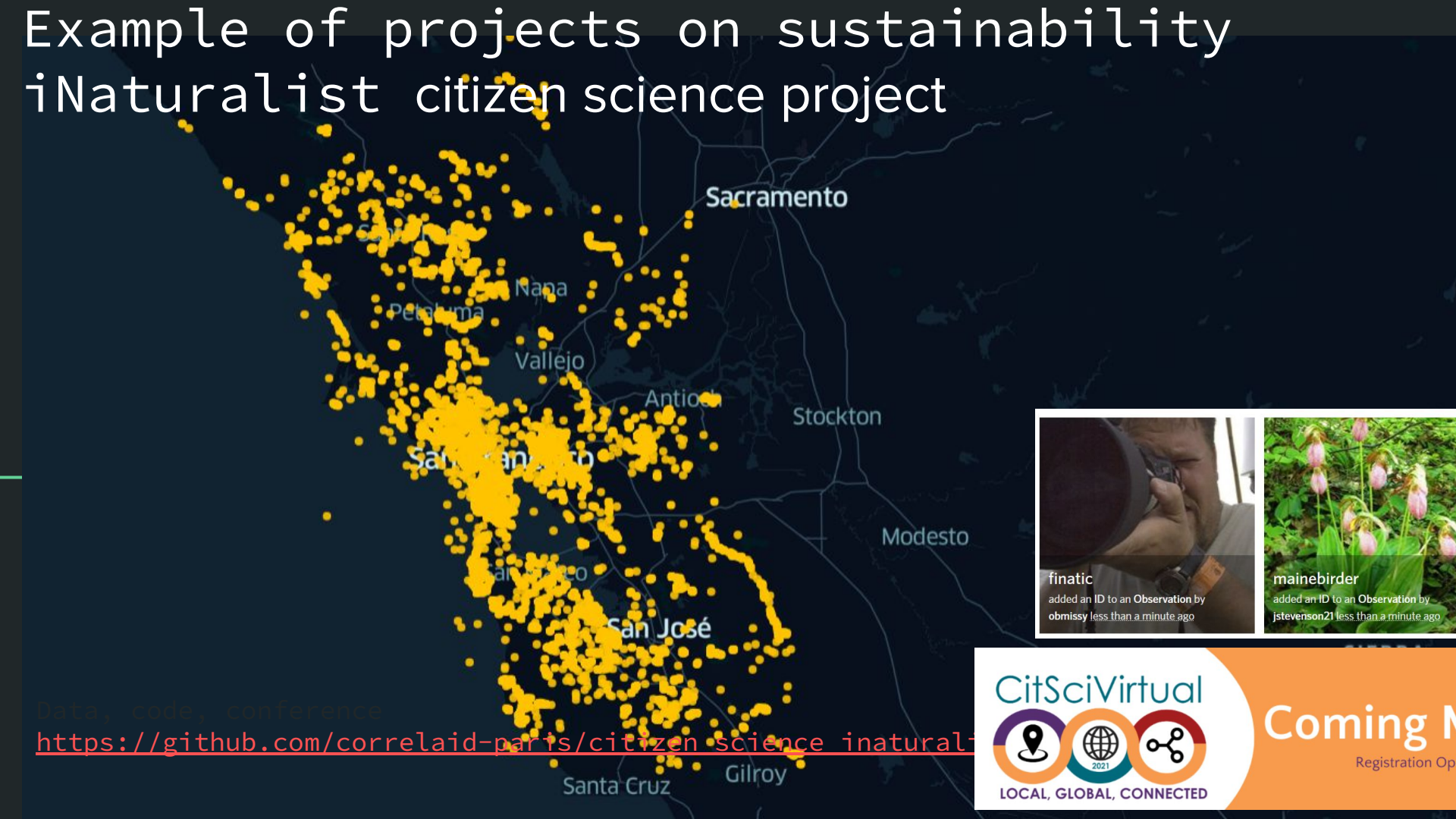
Louvain algorithm
Pseudocode idea

Communities in networks
See code in classroom



Example of projects on sustainability

iNaturalist citizen science project



Data, code, conference

https://github.com/correlaid-paris/citizen_science_inaturalist

Example of sustainability related projects

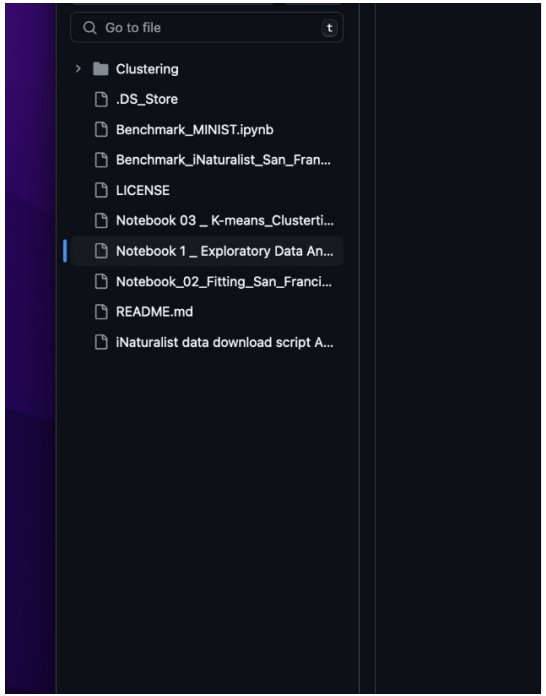
Github notebook and here

Data journalism examples

Data volunteering correlaid.org

Example of sustainability related projects

Github notebook and here



Loading: Francisco_Bay 2022
Loading: Francisco_Bay 2023

Out[4]:

	id	observed_on_string	observed_on	time_observed_at	created_time_zone	created_at	
0	20069	1:15 pm.	2016-07-14	2016-07-14T13:15:00-07:00	America/ Los_Angeles	2011-06-03T14:51:45-07:00	2020-0
1	20070	1:00 pm.	2016-03-25	2016-03-25T13:00:00-07:00	America/ Los_Angeles	2011-06-03T14:53:13-07:00	2020-0
2	68373	6:30	2016-02-12	2016-02-12T06:30:00-08:00	America/ Los_Angeles	2012-04-20T20:36:48-07:00	2020-0
3	158736	2:19	2016-10-14	2016-10-14T14:19:00-07:00	America/ Los_Angeles	2012-12-06T20:23:52-08:00	2016-1
4	538018	2016-04-10 2:20:00 PM PDT	2016-04-10	2016-04-10T14:20:00-07:00	America/ Los_Angeles	2014-02-20T15:40:40-08:00	2016-0

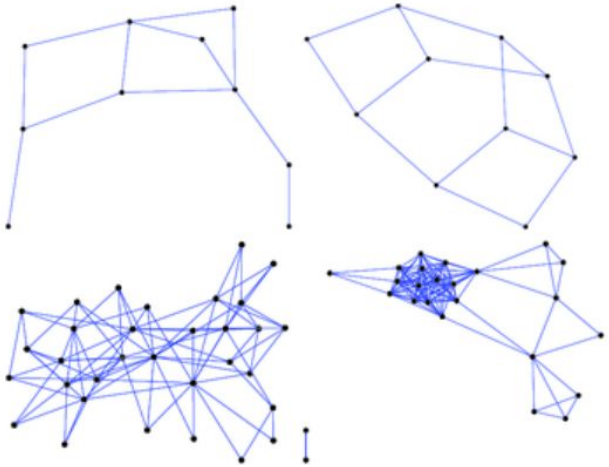
5 rows × 38 columns

Dataset comprises of 4214727 observations and 38 characteristics.

In [10]:
```
dfall.shape
```

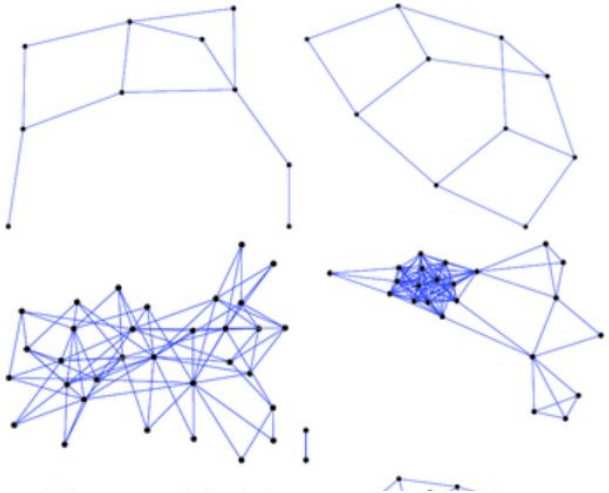
Out[10]: (4214727, 38)

Random networks: building null hypothesis



Degree distribution in my networks is really heterogeneous

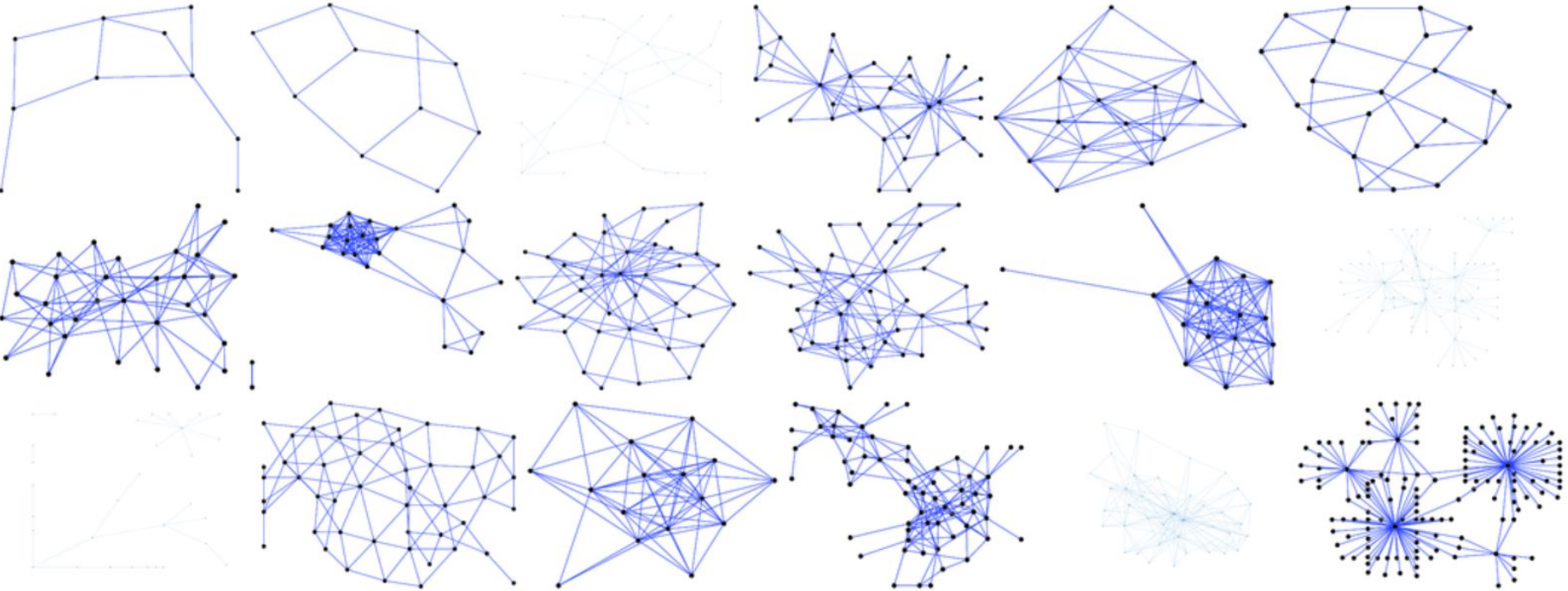
Random networks: building null hypothesis



Degree distribution in my networks is really heterogeneous

Well, did you test the null hypothesis?

Random networks: building null hypothesis



Random networks: model by Erdős (1913-1996) and Rényi (1921-1970)

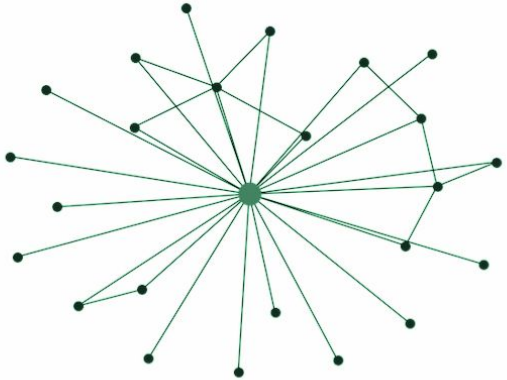
Pseudocode

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: model by Erdős (1913-1996) and Rényi (1921-1970)

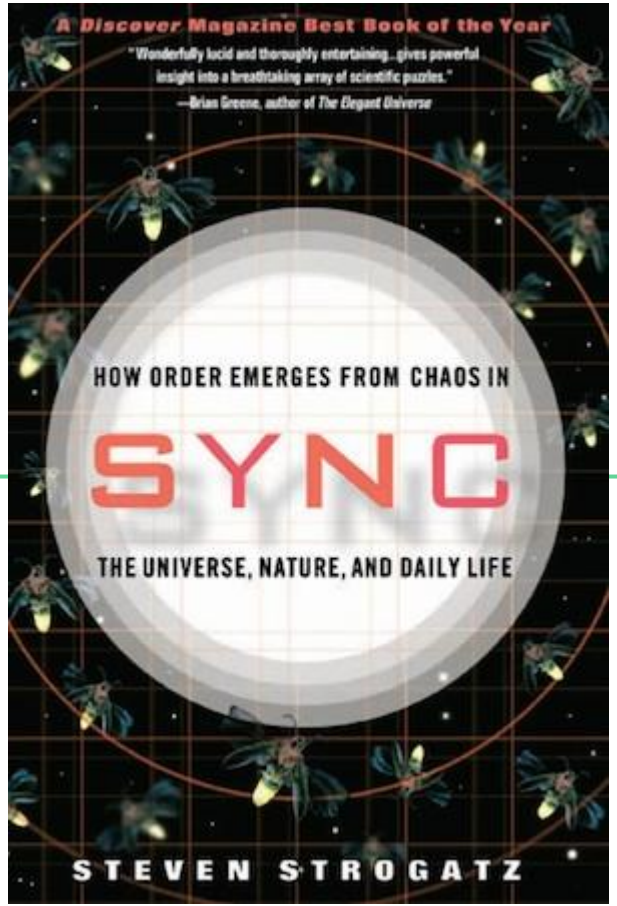


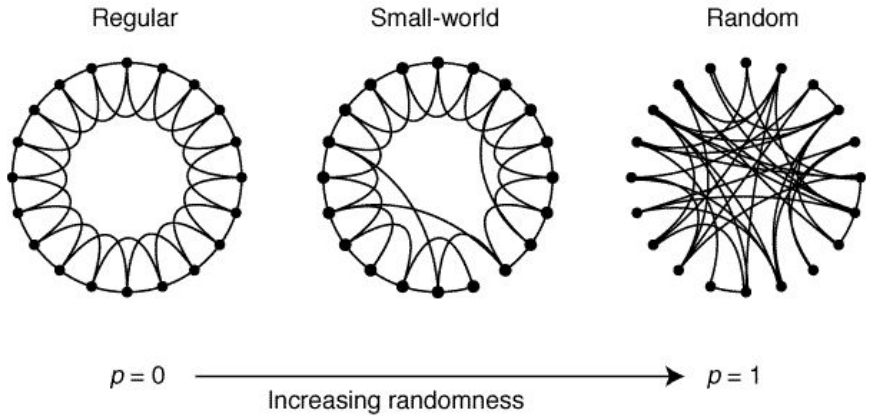
1. Create N nodes
2. Connect each pair of N labeled nodes with probability p. You can do it yourself by tossing a coin each time.

Corresponding class in networkx:

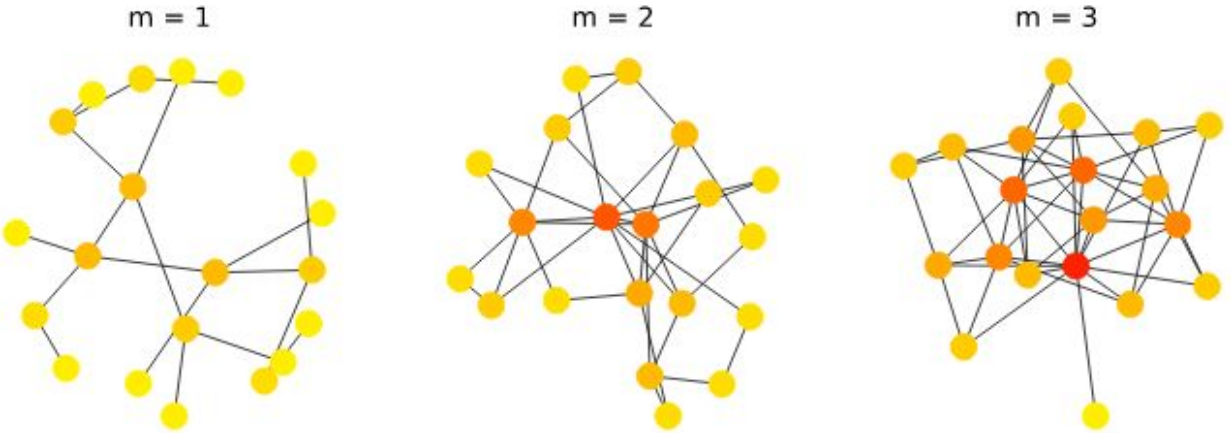
```
G_er = nx.erdos_renyi_graph(n, p2)
```

Random networks: Watts-Strogatz network





Random networks: model by Barabasi Albert



A graph of
nodes is grown by attaching new nodes each with m edges that are preferentially attached to existing nodes with high degree.

A. L. Barabási and R. Albert “Emergence of scaling in random networks”, Science, 1999.

Random networks: model Barabasi Albert

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

Network and robustness

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?

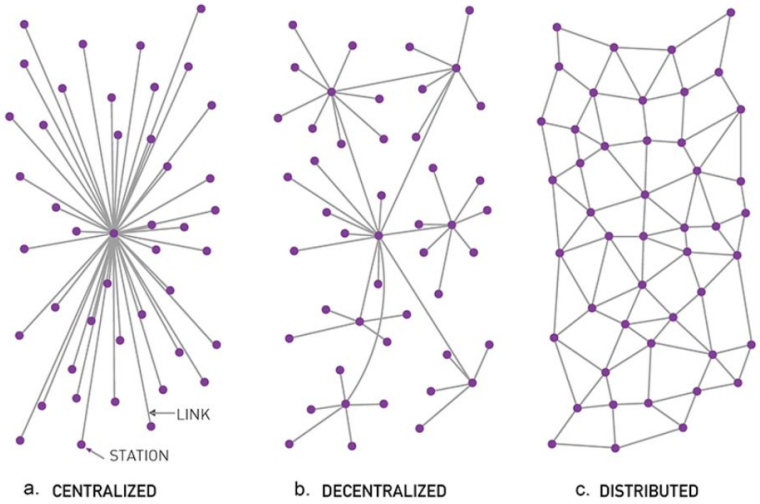
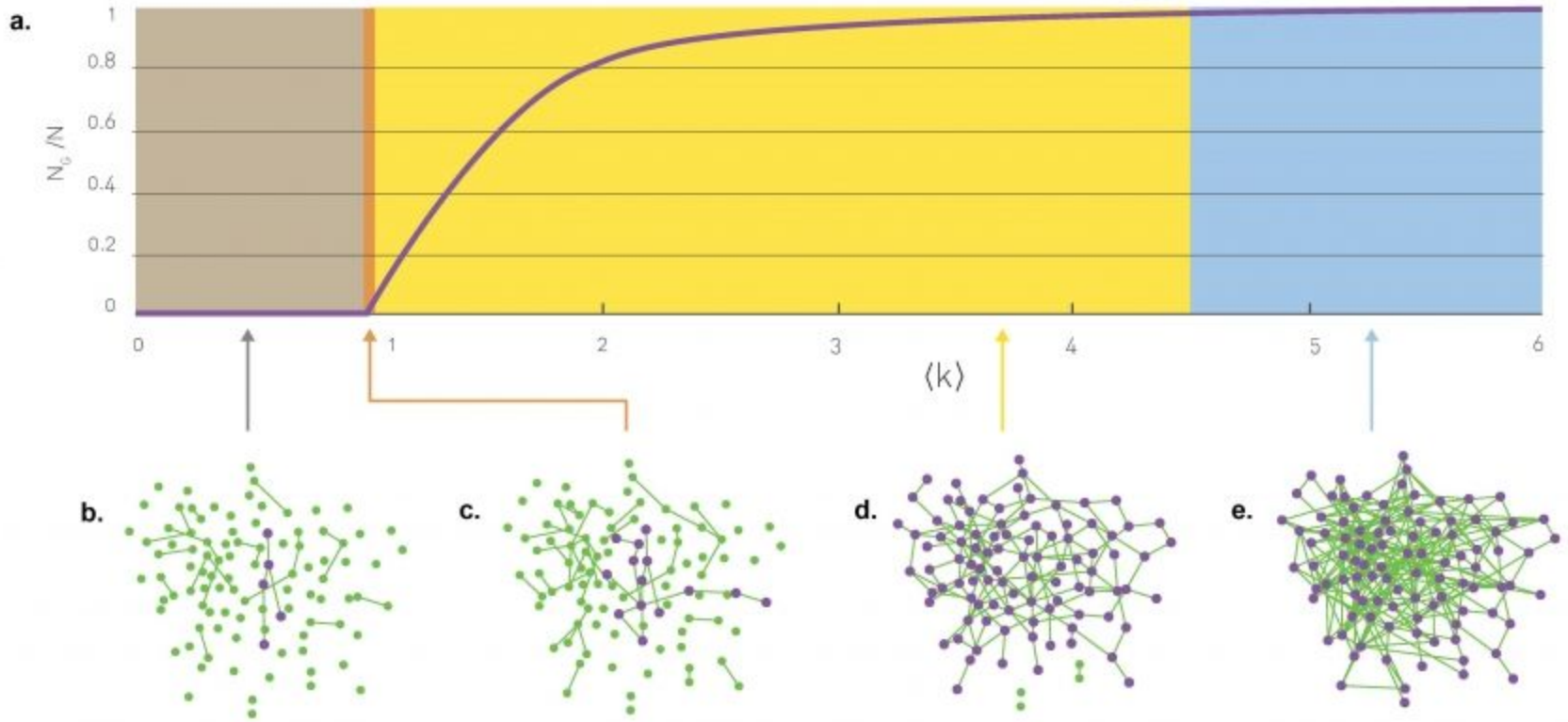


Fig. credits P. Barran

Example of a universal law at the collective scale

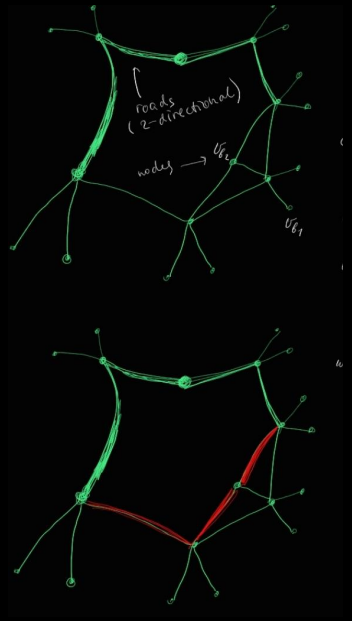
Emergence of a giant component in a network above a threshold of number of links



Hands-on notebooks

Github
https://github.com/Big-data-course-CRI/materials_big_data_cri_2024_2025/tree/main/day%201%20networks%20and%20hypergraphs

Notebook on networks generation and basic network measures
https://colab.research.google.com/drive/1WmwG30LMmkSoOP5Uc7YIXWlFtM68TIm6?usp=sharing



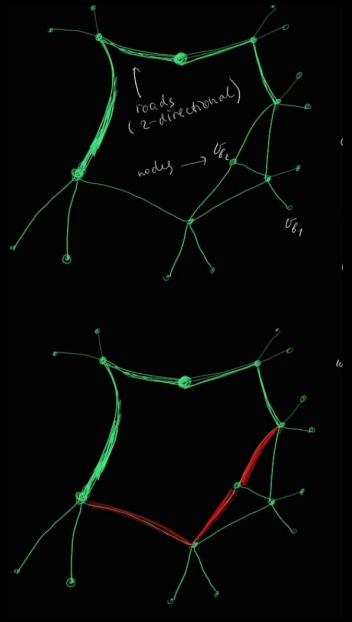
Hands-on notebooks

Github
https://github.com/Big-data-course-CRI/materials_big_data_cri_2024_2025/tree/main/day%201%20networks%20and%20hypergraphs

Extra notebook on geopandas and networks
https://drive.google.com/file/d/1PVFUTS0CuFj4whjUgt7SDROqXb6QQpRK/view?usp=sharing

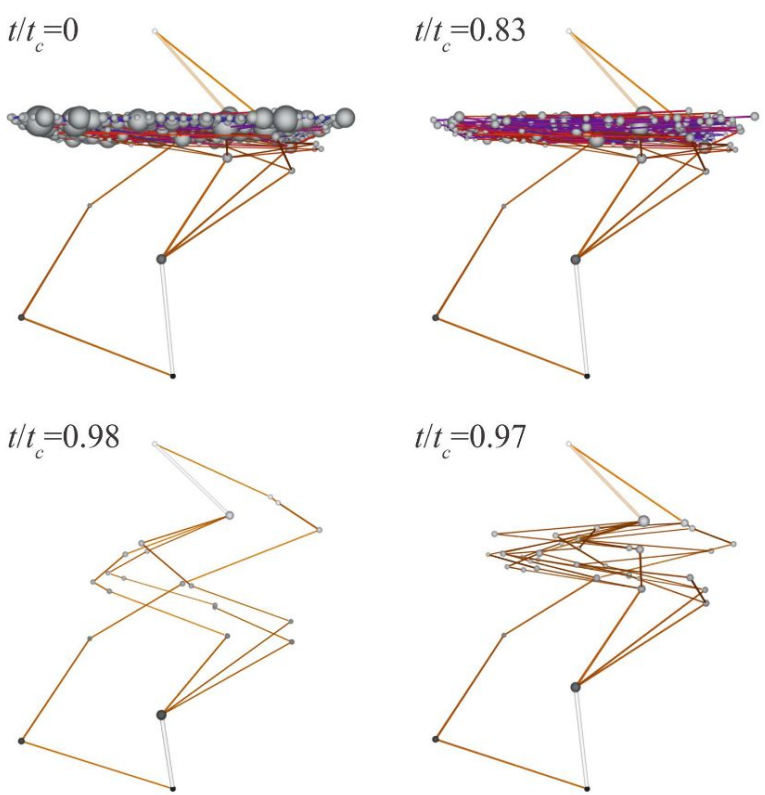
Extra notebook on hypergraphs
https://colab.research.google.com/drive/1bc633d1b5tBtIetFJ57nWFYPV_JrGahO?usp=sharing

Extra notebook on neural networks

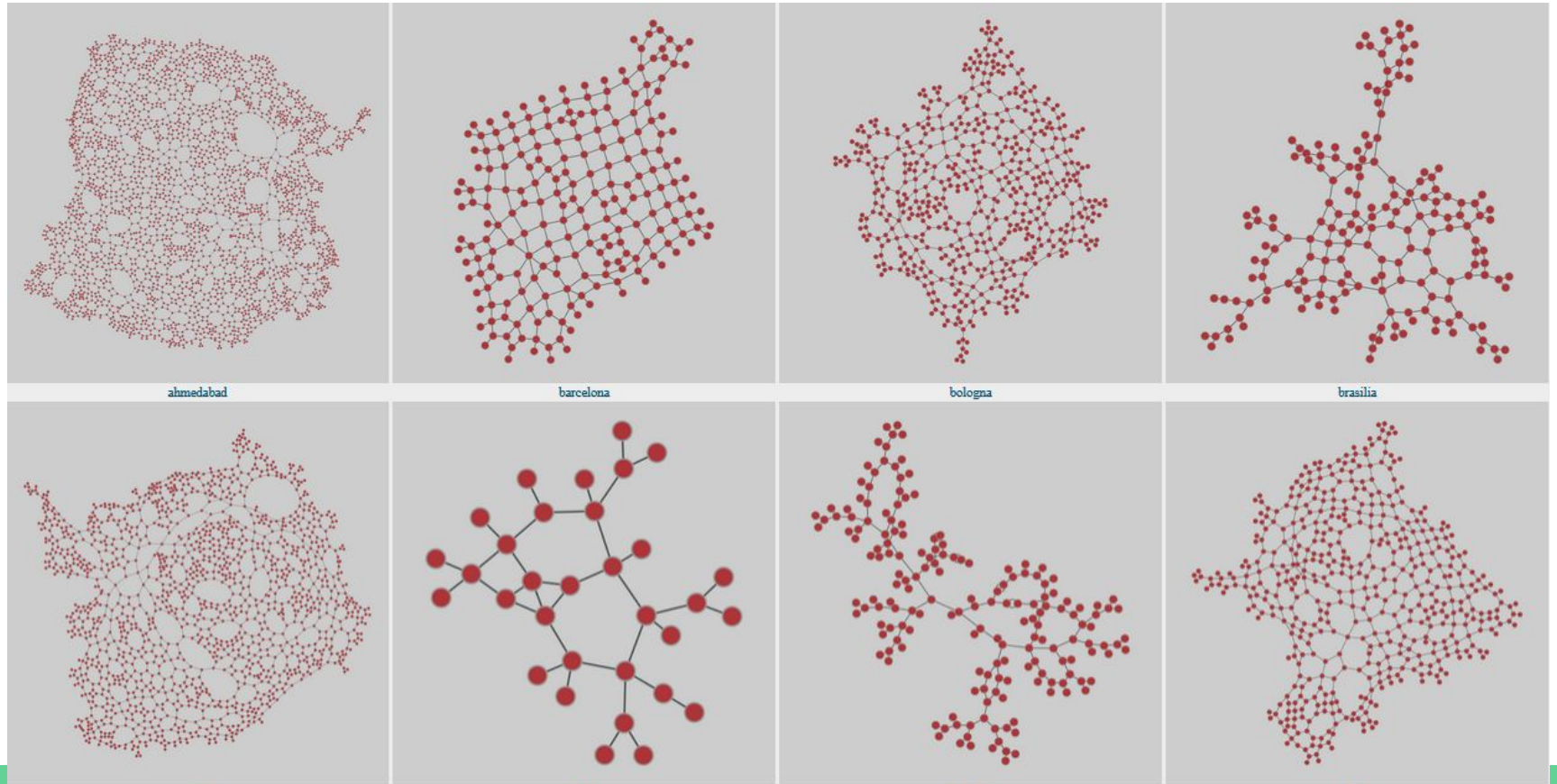


Networks in time and space

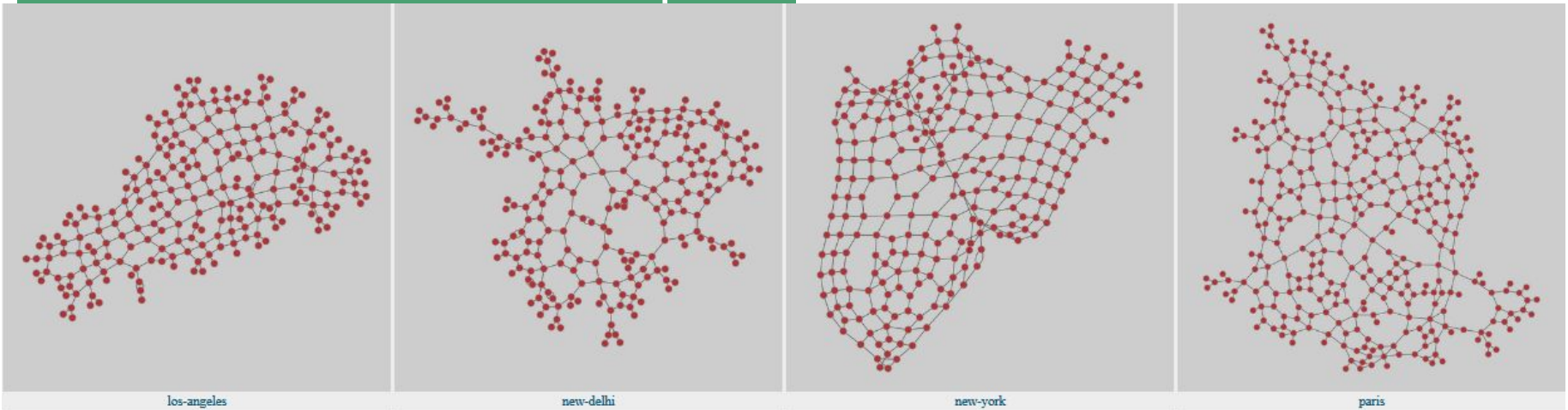
Percolation of networks in time
Nat.Comm. 2020

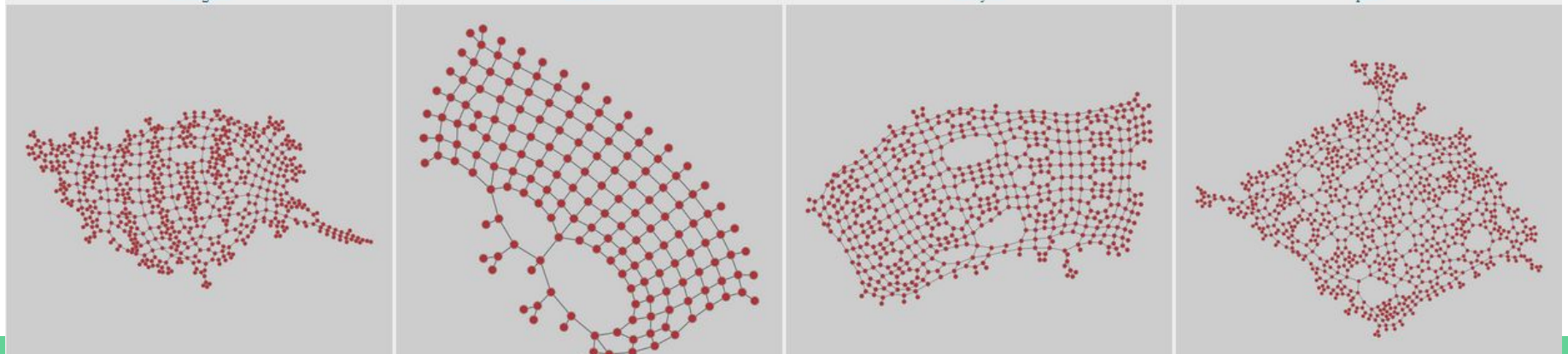


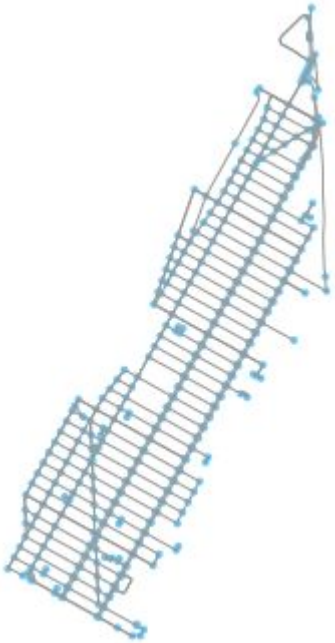
Networks in time and space

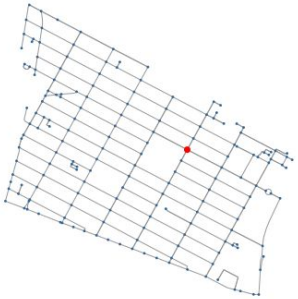


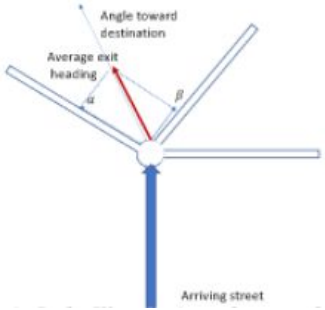
Networks in time and space







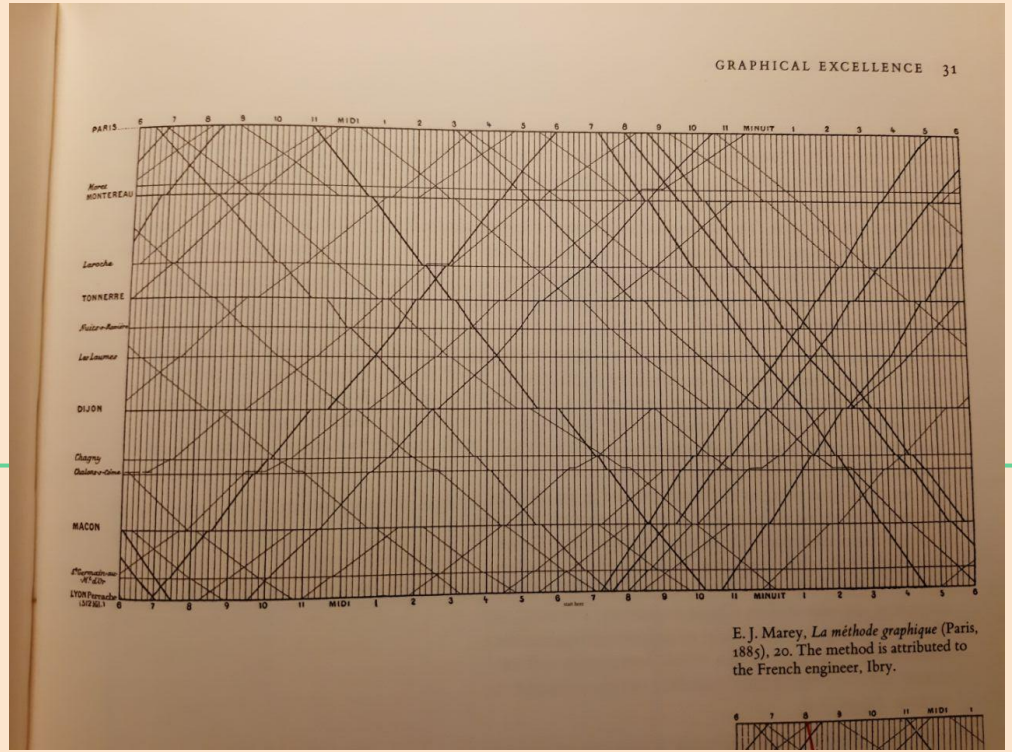



Angle toward destination
Average exit heading
α
β
Arriving street

Openstreetmaps analysis

Links to github
https://github.com/cityinteractionlab/openstreetmaps_osmnx_workshop
https://github.com/gboeing/osmnx-examples
https://github.com/cityinteractionlab/openstreetmaps_osmnx_workshop
Google colab
(geo packages required)

Data analysis from 19th century



Tufte book: spatial data visualisation

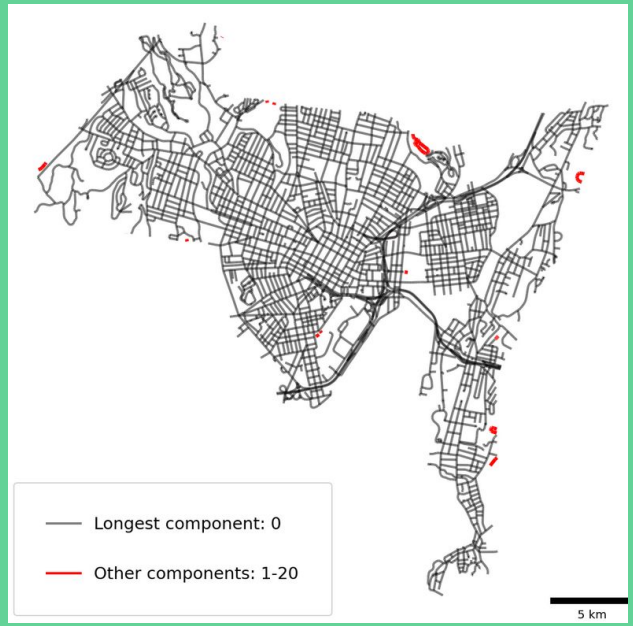
Resources pages

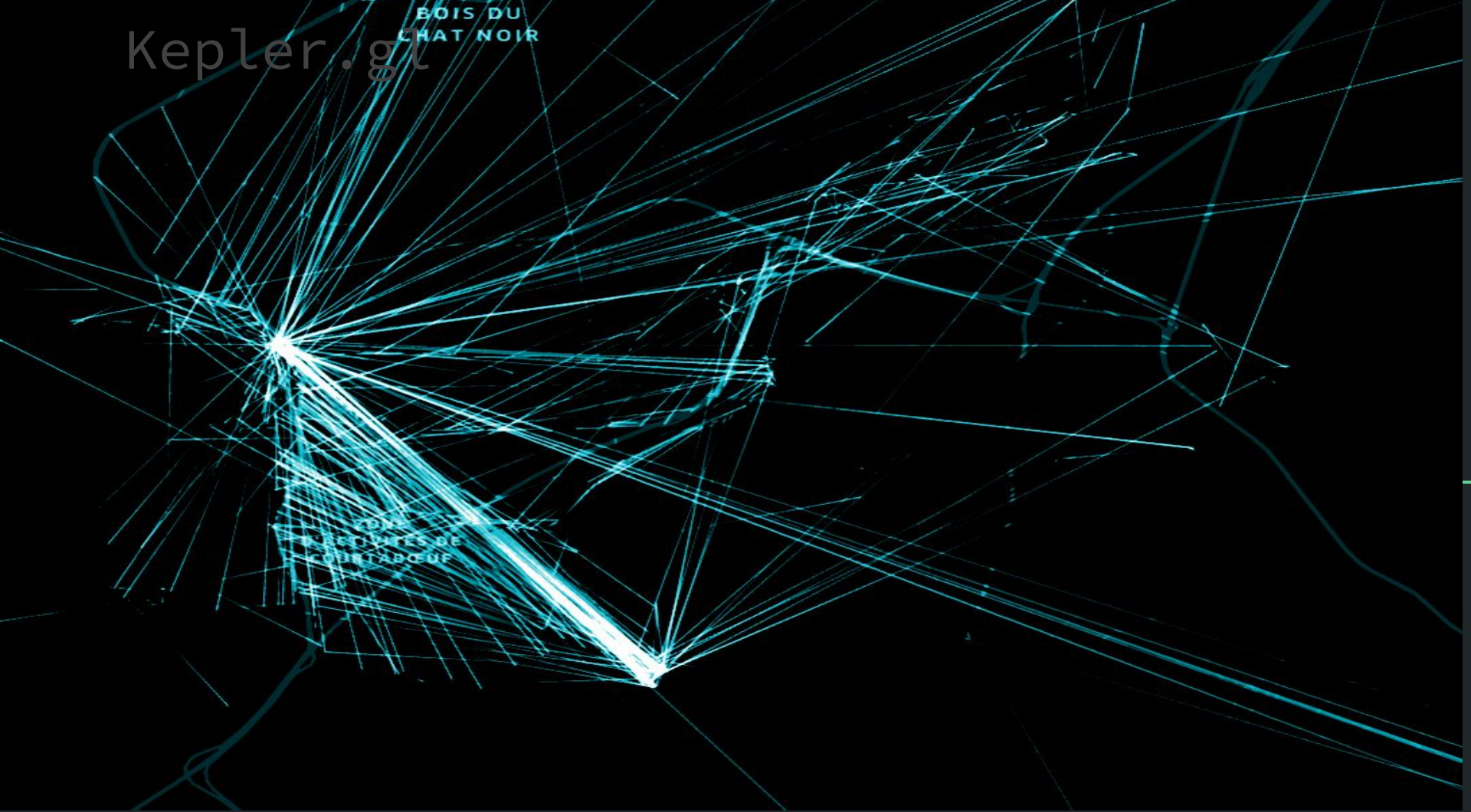
www.worldpop.org spatial population analysis
hdx platform

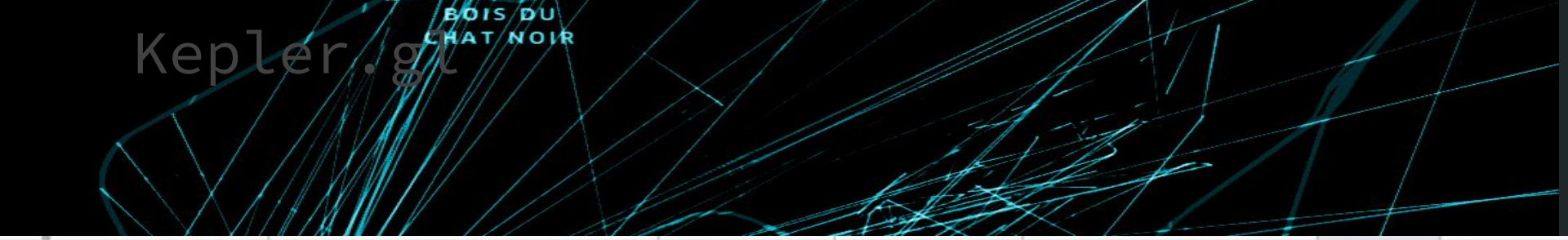
www.kepler.gl for mobility data visualisation (browser)
Mapbox python integration

Open source tools, python and R
https://www.python-graph-gallery.com/

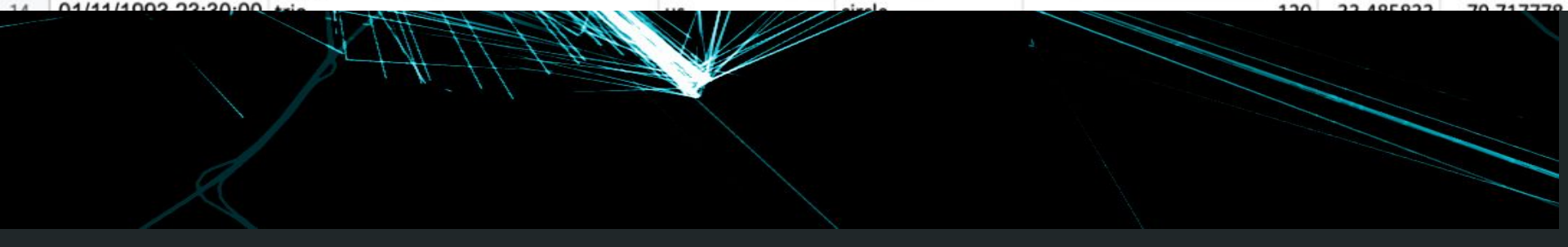
Packages:
Scikit https://github.com/scikit-mobility/scikit-mobility
Matplotlib, cartopy - simple plotting,
Folium - online plotting,
Geopandas - python package,
libpysal - spatial distribution,
Osmnx - python package for analysis of openstreetmaps
Spagetti https://github.com/pysal/spaghetti




BOIS DU
CHAT NOIR
Kepler.gl



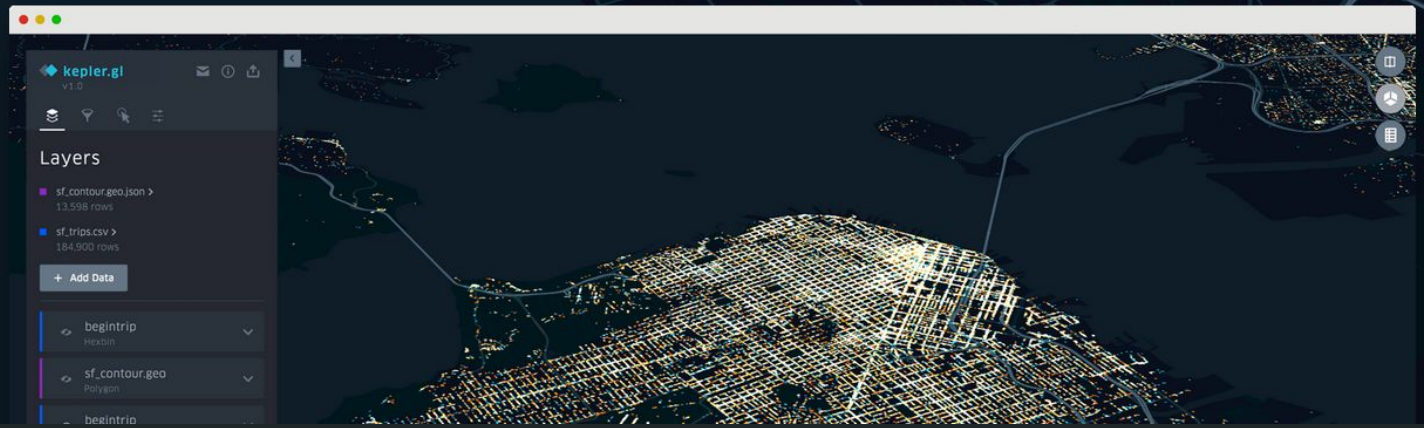
1	Date_time	city	country	UFO_shape	length_of_encounter_seconds	latitude	longitude
2	01/03/1993 12:00:00	prescott	us	diamond	900	34.540000	-112.467778
3	01/03/1993 13:00:00	chattanooga	us	sphere	300	35.045556	-85.309722
4	01/06/1993 14:00:00	pittsburgh	us	sphere	15	40.440556	-79.996111
5	01/06/1993 21:00:00	san jose (snell rd / blossom hill rd)	us	circle	300	37.339444	-121.893889
6	01/06/1993 22:00:00	billings	us	light	20	45.783333	-108.500000
7	01/07/1993 00:00:00	phoenix	us	oval	60	33.448333	-112.073333
8	01/07/1993 03:30:00	katy	us	sphere	15	29.785556	-95.824167
9	01/08/1993 17:00:00	warrenton	us	circle	300	38.713333	-77.795556
10	01/08/1993 21:30:00	tillamook(lees camp)	us	light	900	45.456389	-123.842778
11	01/09/1993 22:00:00	bethel (albany township)	us	light	120	44.404167	-70.791111
12	01/10/1993 20:00:00	delaware	us	light	7200	40.298611	-83.068056
13	01/10/1993 20:00:00	pryor	us	light	1800	45.429722	-108.532500
14	01/11/1993 22:30:00	tri...	us	circ...	120	33.485833	-70.717778



Kepler.gl is a powerful **open source** geospatial analysis tool for **large-scale** data sets.

GET STARTED

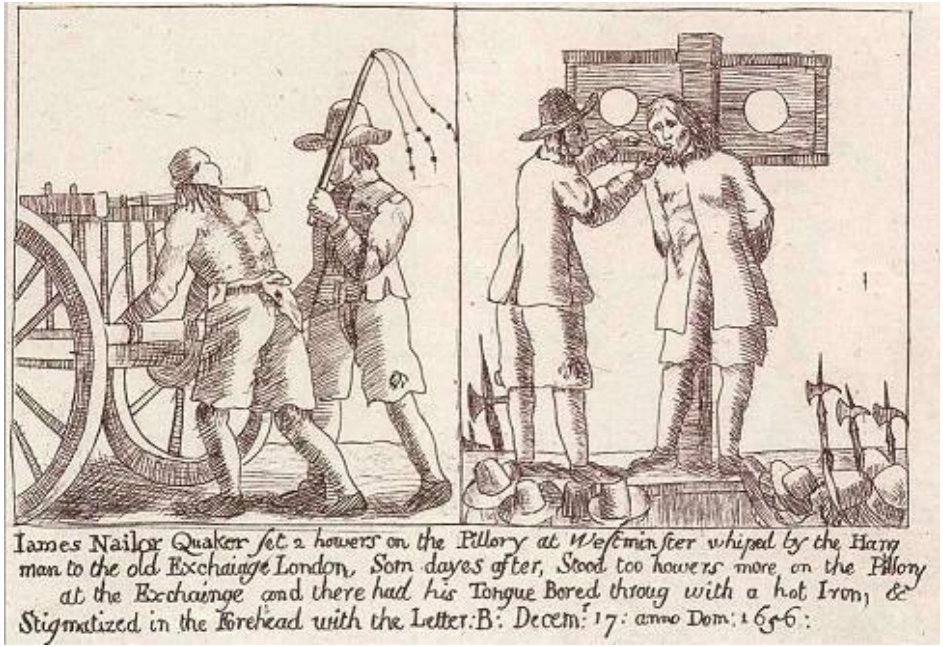
GITHUB



Quick check-in

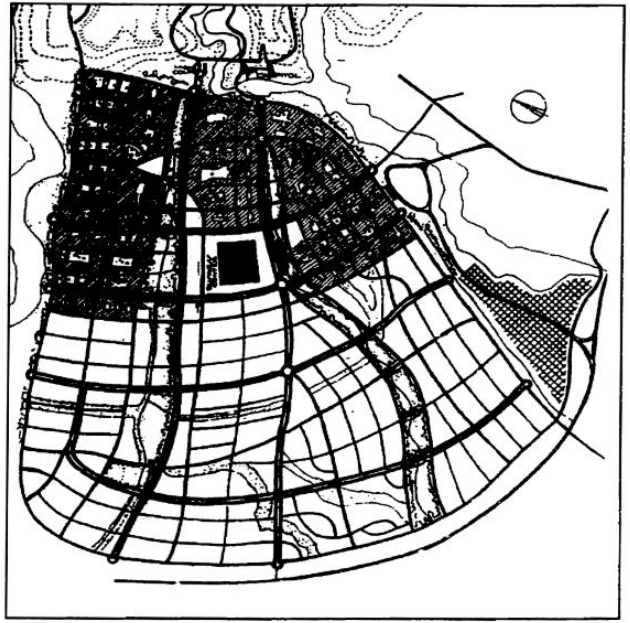
Can networks tell a new story about your data?

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

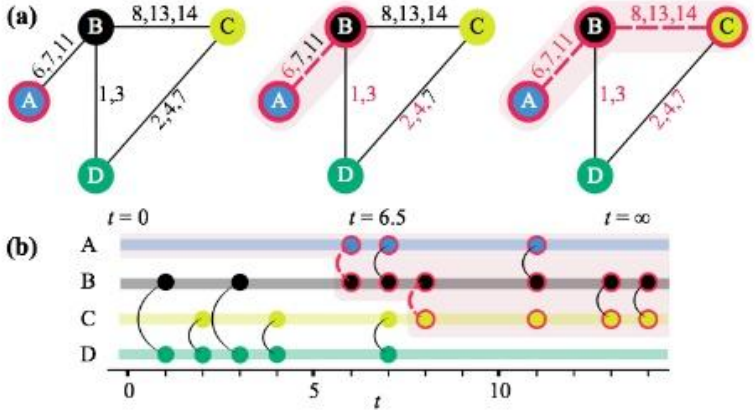


Networks in time and space

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

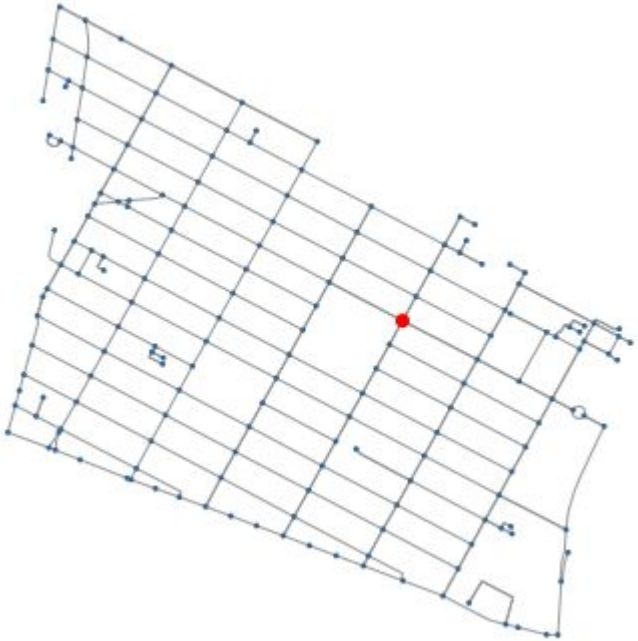


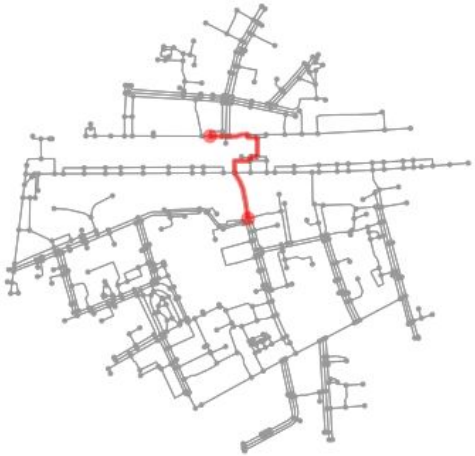
Temporality matters: reachability issue



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

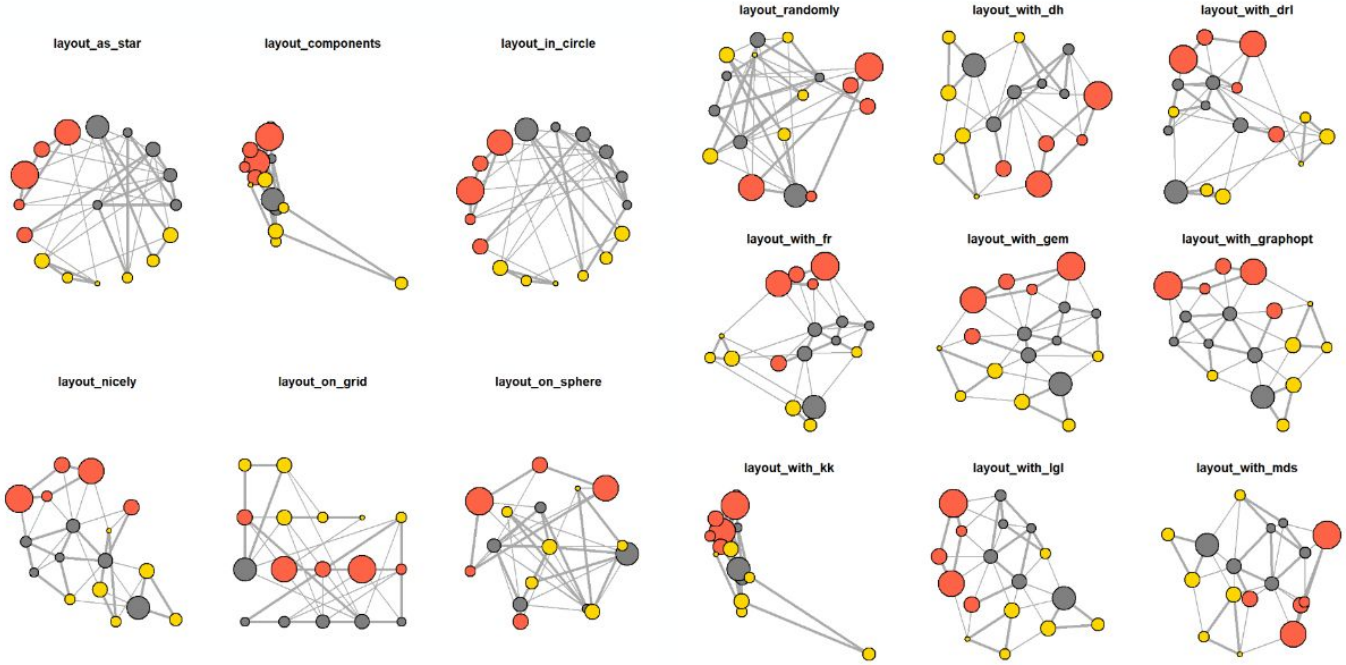
Networks in time and space





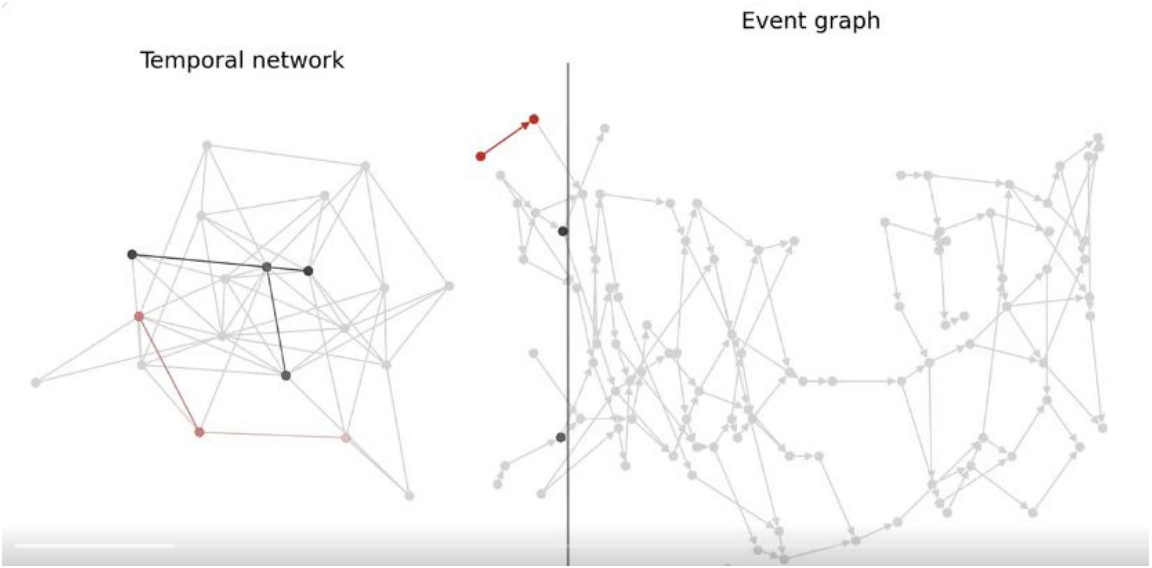
Osmnx for spatial networks analysis
https://arxiv.org/abs/1010.0302

Networks layout



Networks sonification

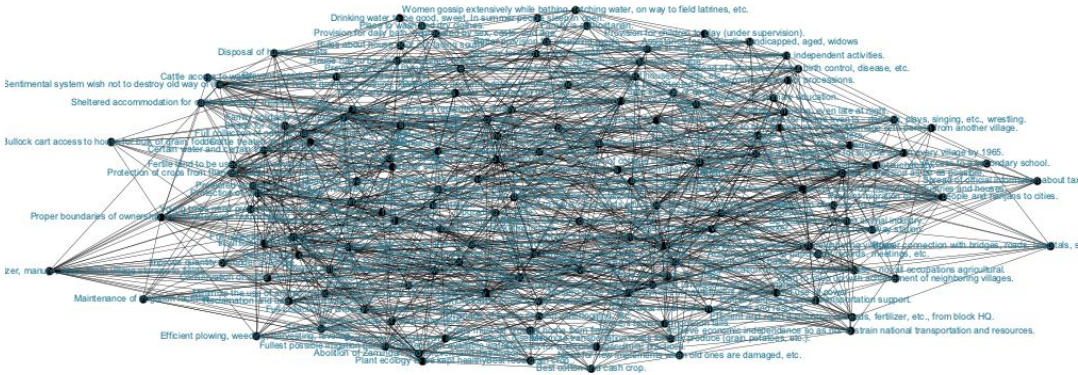
Directed percolation in temporal networks PRR 2022



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/

What we will look at in network science?

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

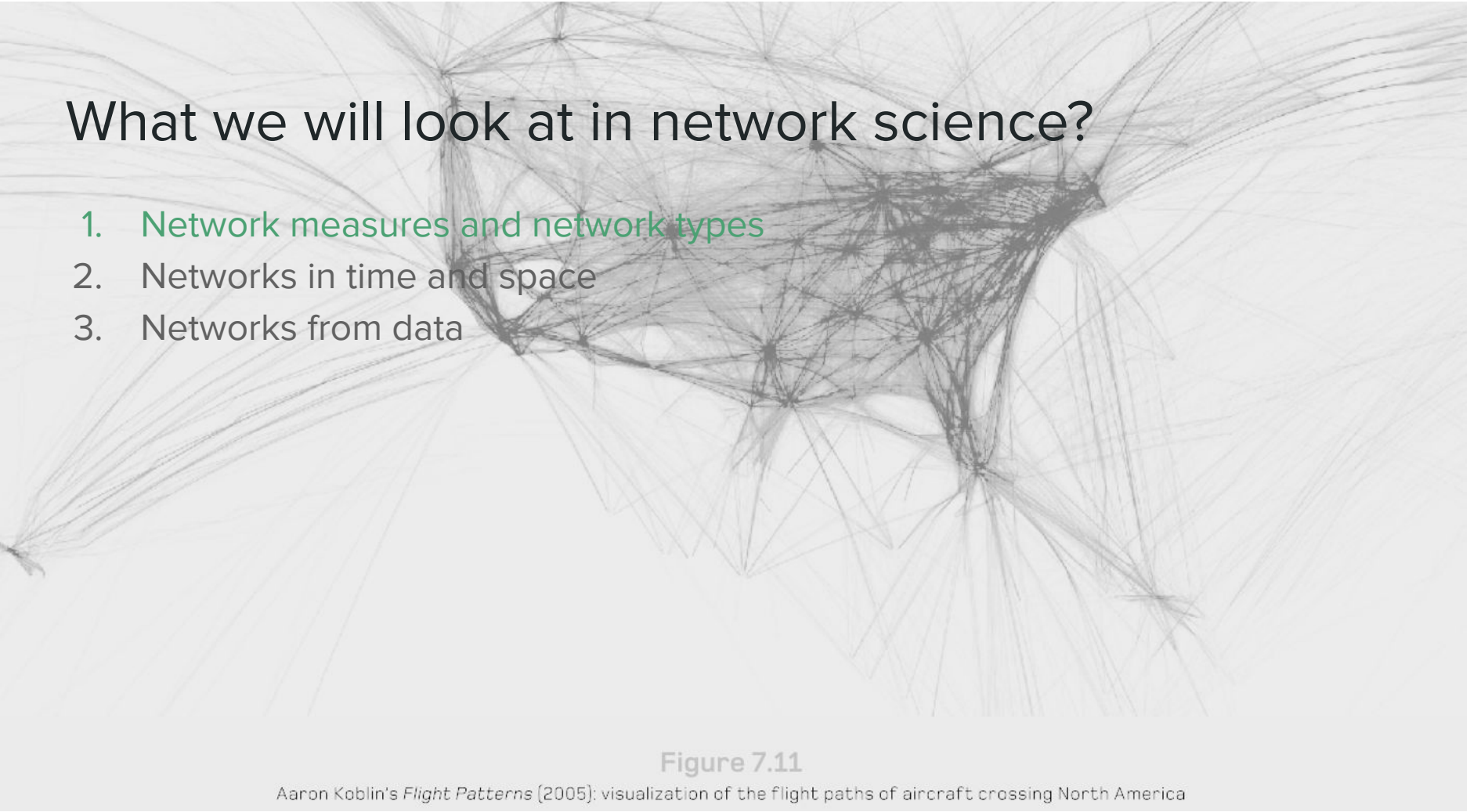


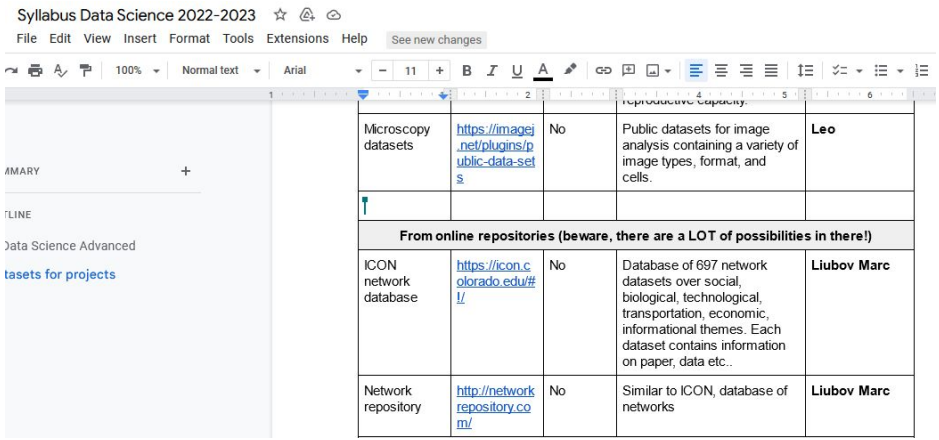
Figure 7.11

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

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

100% Normal text Arial 11

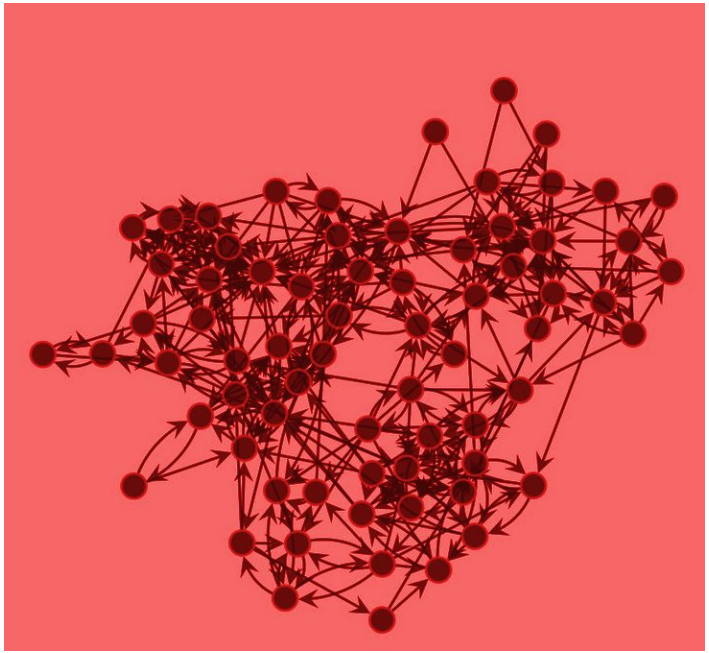
MMARY

TLINE

Data Science Advanced

tasets for projects

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**



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