Graph Learning
SD212

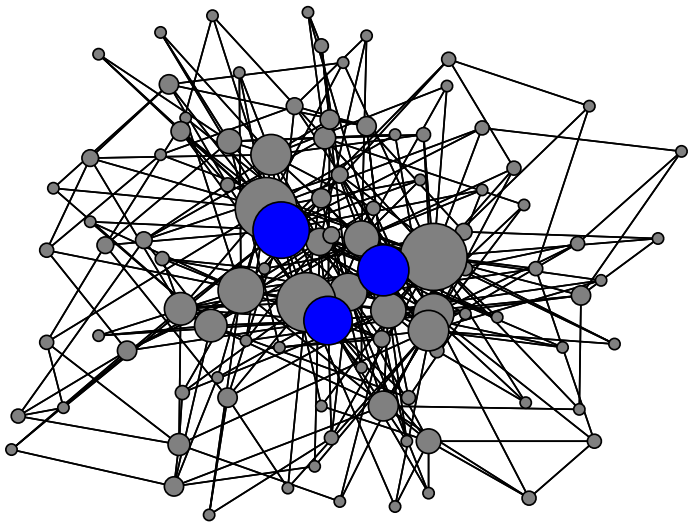
Thomas Bonald & Simon Delarue

2023 – 2024

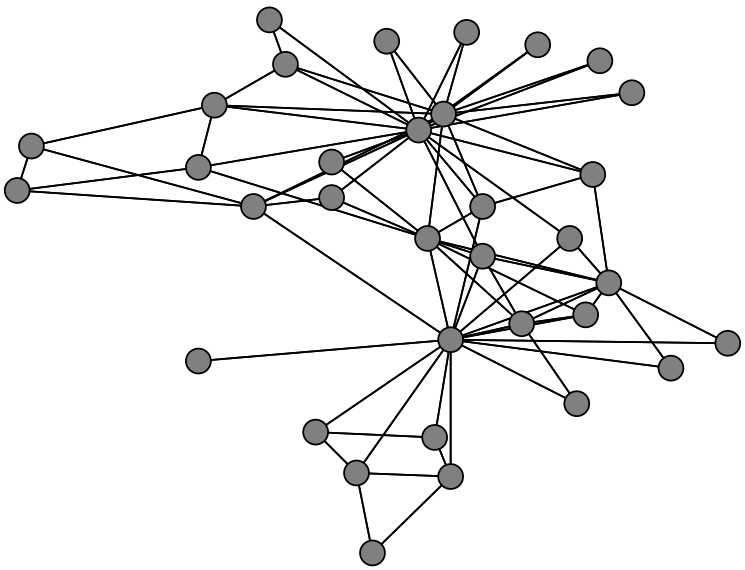
Graph data

Graphs describe **links** between objects:

- **Social networks** $\rightarrow$ contacts
- **Web** $\rightarrow$ hyperlinks
- **Knowledge bases** $\rightarrow$ facts
- **Documents** $\rightarrow$ references
- **Commerce** $\rightarrow$ transactions
- **Biology** $\rightarrow$ interactions

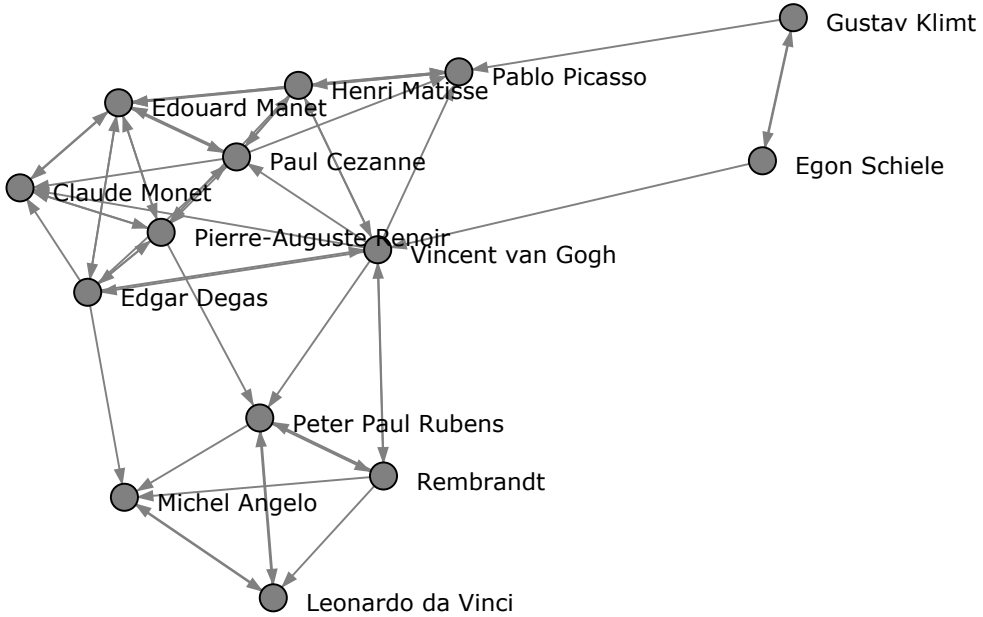


Graphs



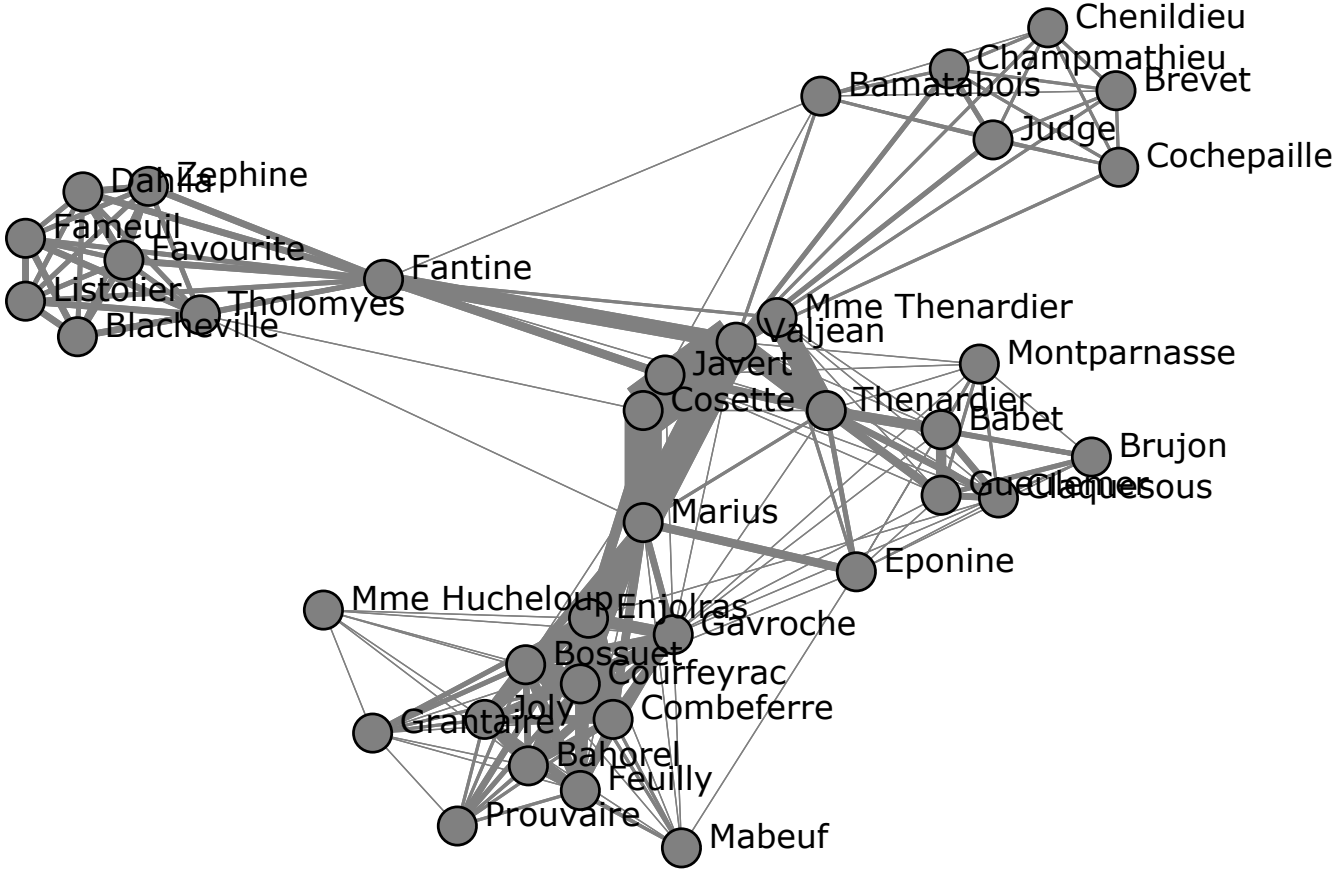
Karate club graph

Directed graphs



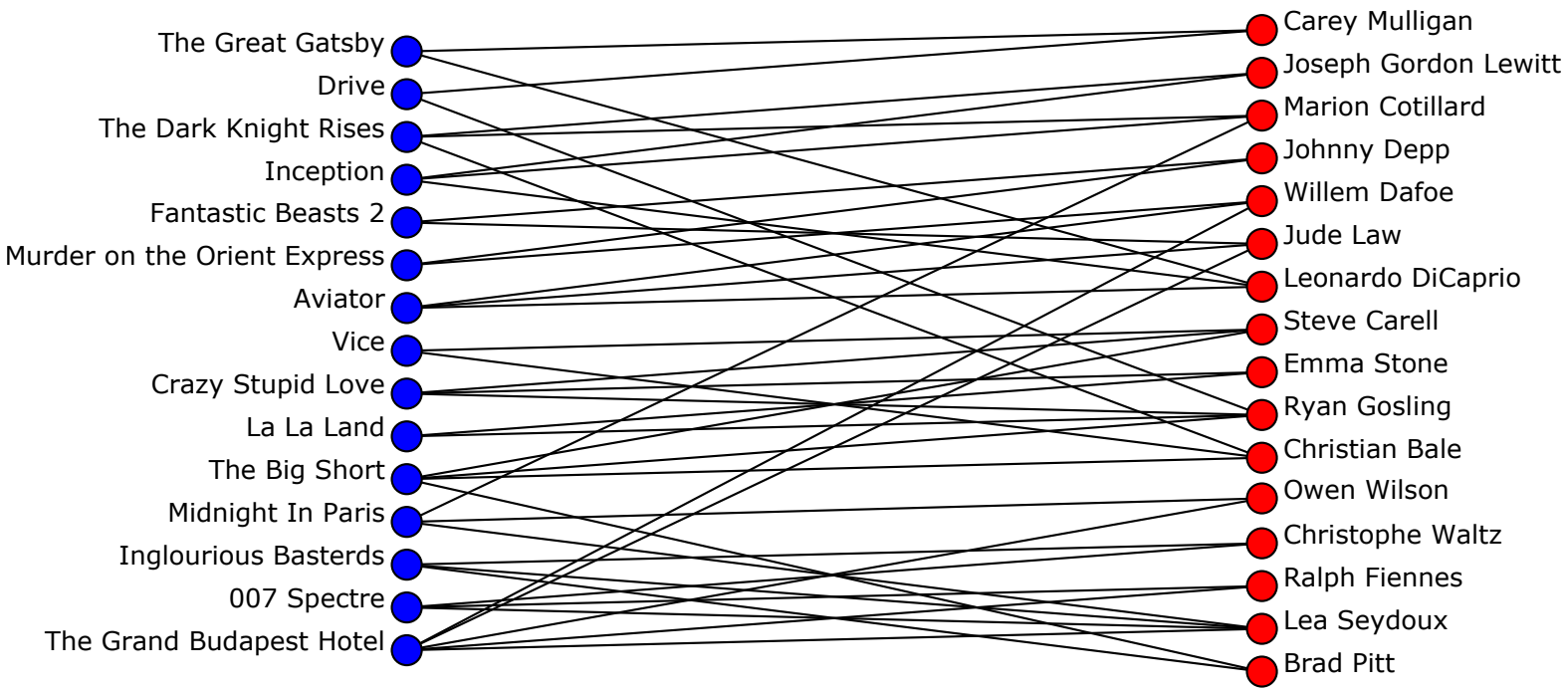
Links between some Wikipedia articles

Weighted graphs



Co-occurence of some characters of the novel Les Misérables

Bipartite graphs



Actors starring in movies

Tabular data

Source: Adult income dataset (Kaggle)

	gender	age	workclass	education	family	occupation
0	Male	40	State-gov	Bachelors	Not-in-family	Adm-clerical
1	Male	50	Self-emp-not-inc	Bachelors	Husband	Exec-managerial
2	Male	40	Private	HS-grad	Not-in-family	Handlers-cleaners
3	Male	50	Private	11th	Husband	Handlers-cleaners
4	Female	40	Private	Masters	Wife	Exec-managerial
...	...	...	...	...	...	...
31973	Male	20	Private	Bachelors	Not-in-family	Prof-specialty
31974	Male	60	Self-emp-inc	HS-grad	Husband	Transport-moving
31975	Male	60	Private	Assoc-voc	Husband	Craft-repair
31976	Female	50	Private	HS-grad	Wife	Adm-clerical
31977	Female	30	Private	Some-college	Other-relative	Machine-op-inspct

31978 rows × 6 columns

Tabular data as bipartite graph

	gender_ Female	gender_ Male	age_20	age_30	age_40	age_50	age_60	age_70	age_80	age_90	...
0	0	1	0	0	1	0	0	0	0	0	...
1	0	1	0	0	0	1	0	0	0	0	...
2	0	1	0	0	1	0	0	0	0	0	...
3	0	1	0	0	0	1	0	0	0	0	...
4	1	0	0	0	1	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
31973	0	1	1	0	0	0	0	0	0	0	...
31974	0	1	0	0	0	0	1	0	0	0	...
31975	0	1	0	0	0	0	1	0	0	0	...
31976	1	0	0	0	0	1	0	0	0	0	...
31977	1	0	0	1	0	0	0	0	0	0	...

31978 rows × 56 columns

```
> pd.get_dummies(dataframe)   # one-hot encoding
```

Large graphs are sparse

Dataset	#nodes	#edges	Density
Flights	2,939	30,500	$\approx 10^{-3}$
Amazon products	335k	925k	$\approx 10^{-5}$
Actors	382k	33M	$\approx 10^{-4}$
Wikipedia	12M	378M	$\approx 10^{-6}$
Twitter	42M	1.5G	$\approx 10^{-6}$
Friendster	68M	2.5G	$\approx 10^{-7}$

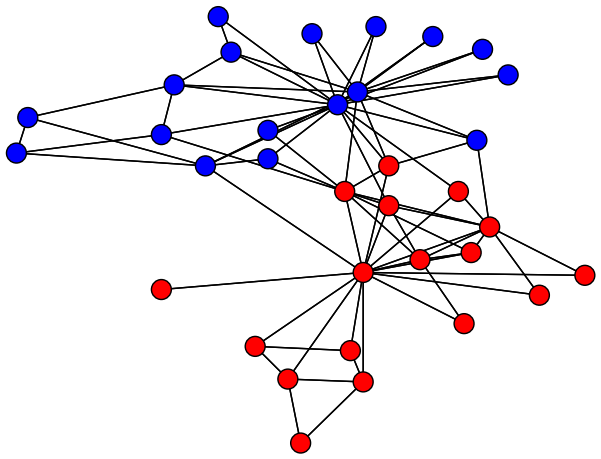
Machine learning on graphs

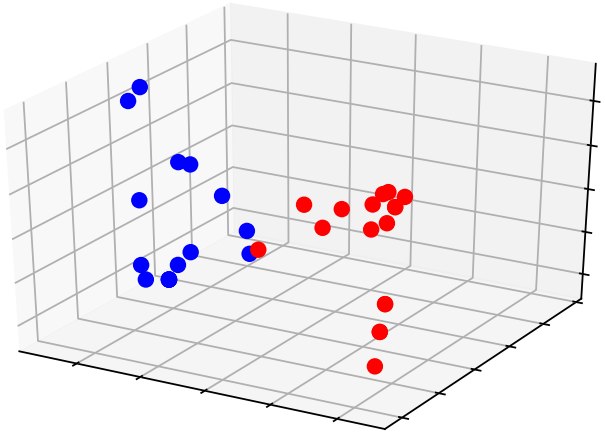
Supervised learning

- Classification
- Regression

Unsupervised learning

- Ranking
- Clustering
- Embedding
- Link prediction
- Anomaly detection





Outline of the course

1. Sparse matrices
   Graph structure
2. PageRank
3. Clustering
4. Hierarchical clustering
5. Heat diffusion
6. Spectral embedding
7. Graph neural networks

Each course = **lecture** + **quiz** + **lab**

Validation

There is one quiz per lecture:

- the deadline is the following **Tuesday** at 6pm
- there are at most **3 attempts**
- only the **last attempt** is considered

The exam itself is a quiz:

- in **limited time** (2h)
- with only **one attempt**

The final grade is based on:

- the lecture quizzes (40%)
- the exam (60%)

Labs are **not** graded

Attendance

You **must** attend the labs in person.

Your attendance might be checked.
Please sign **only** if you are on site.

You are allowed to miss **1 lab** over the 7.
You will get a penalty of **2 points** on the final grade for each additional absence.

Labs

scikit
network

A Python library for graph analysis

- **easy** to install `pip install scikit-network`
- **easy** to use `algorithm.fit(data)`
- **fast** and **memory-efficient**

Relies on **NumPy** and **SciPy** only
BSD license

Benchmark

NetworkX Network Analysis in Python igraph graph-tool

Test on the **Orkut graph** (3M nodes, 117M edges)

Memory

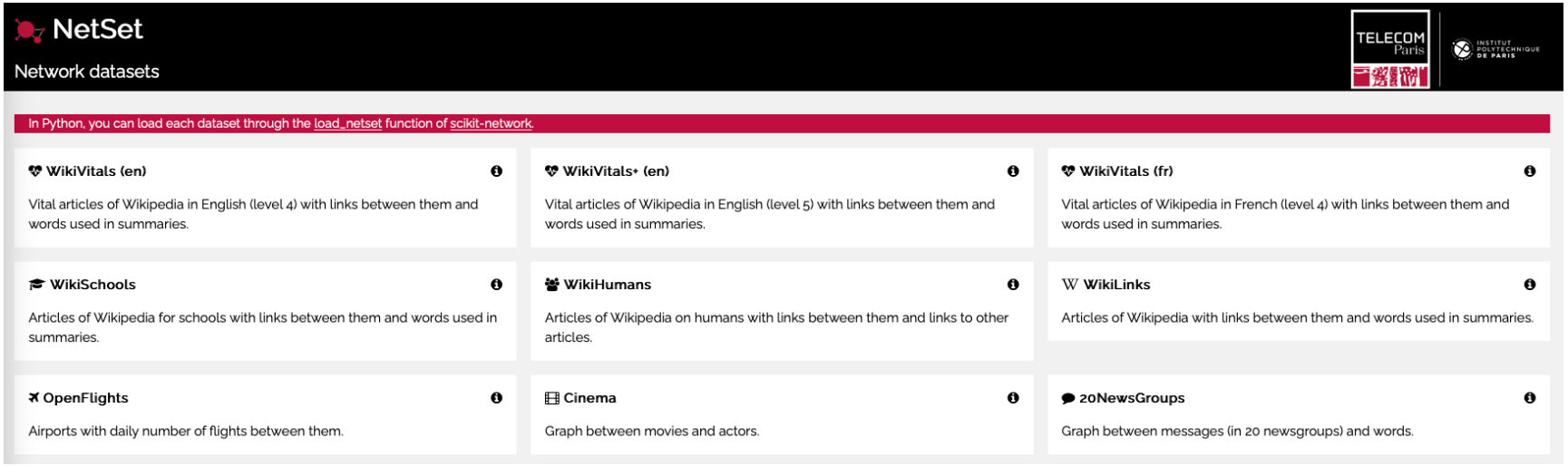
NetworkX	iGraph	graph-tool	scikit-network
	18G	10G	**1G**

Running time

	iGraph	graph-tool	scikit-network
PageRank	3 min 56 s	**45 s**	48 s
Louvain	33 min		**2 min**

The NetSet collection

A collection of network datasets maintained by Télécom Paris
https://netset.telecom-paris.fr



Easy to import with scikit-network!