Graph Mining
SD212
Introduction

Thomas Bonald, Alexandre Hollocou

2017 – 2018

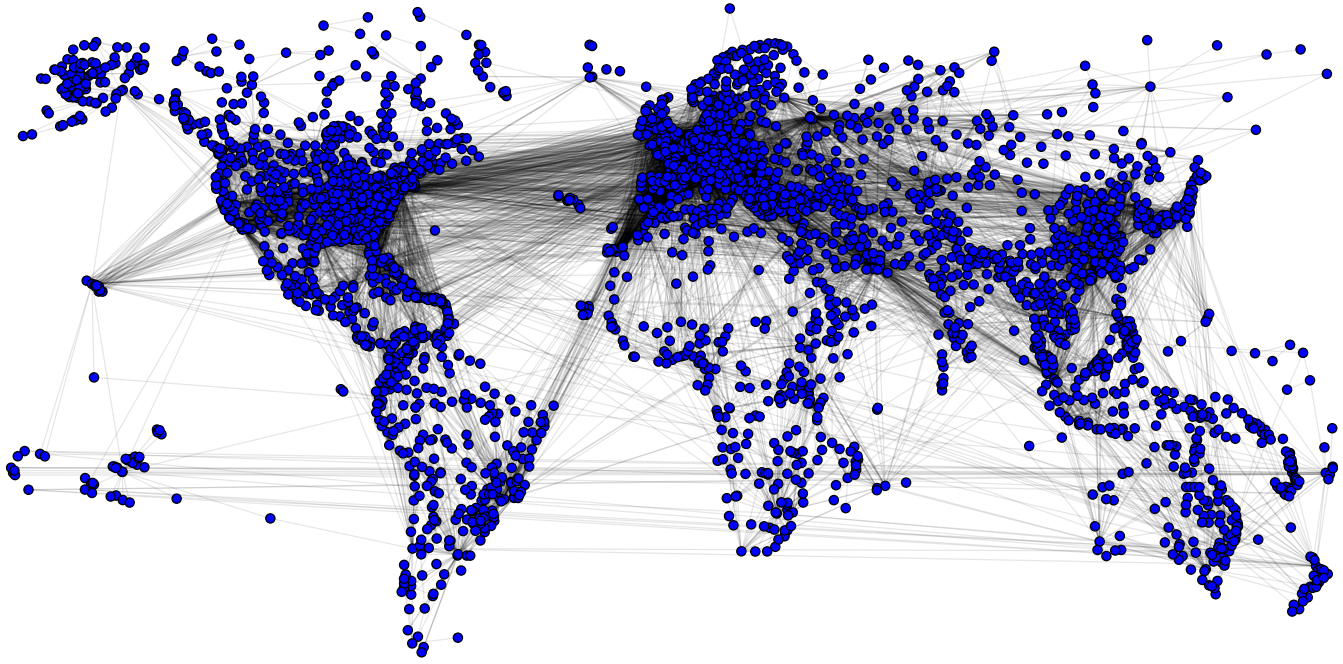
Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...

Main European highways

Graph data

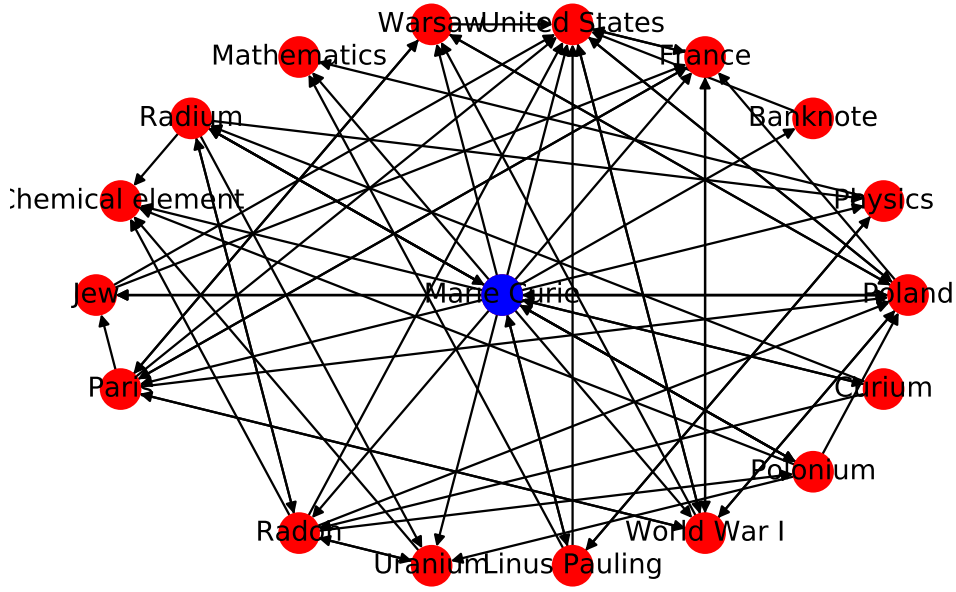
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...



International flights

Graph data

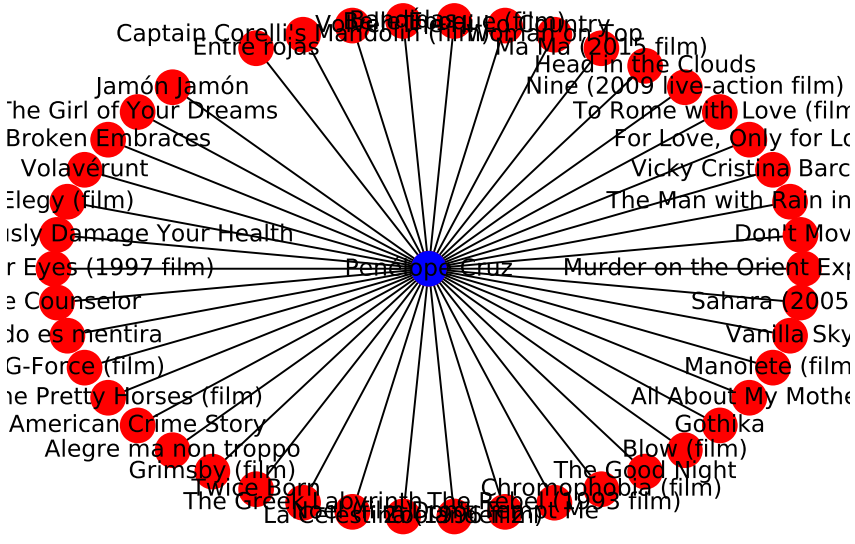
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...

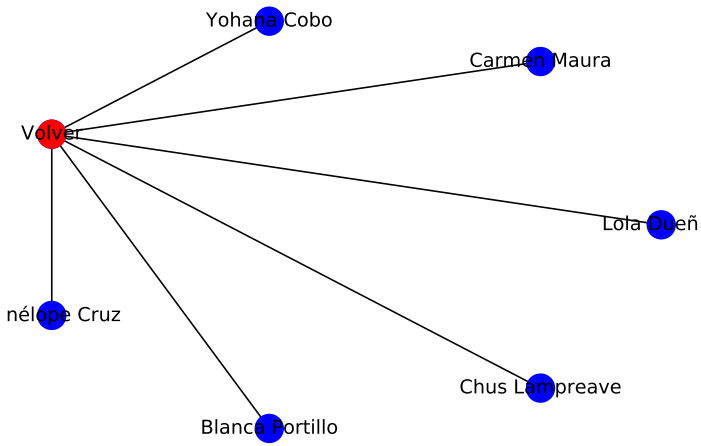


Extract from Wikipedia for Schools

Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...

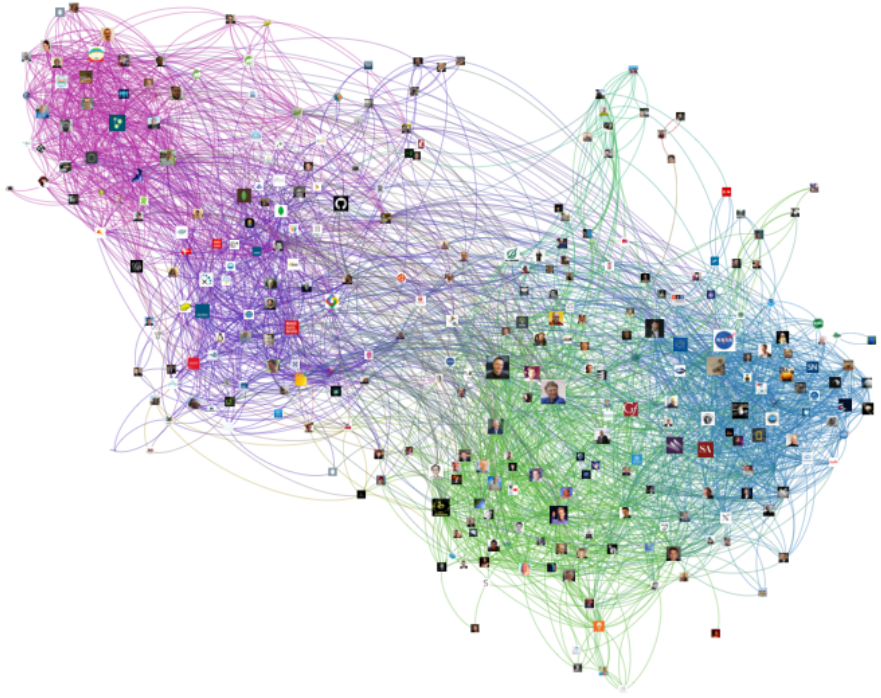




Extract from the movie-actor graph

Graph data

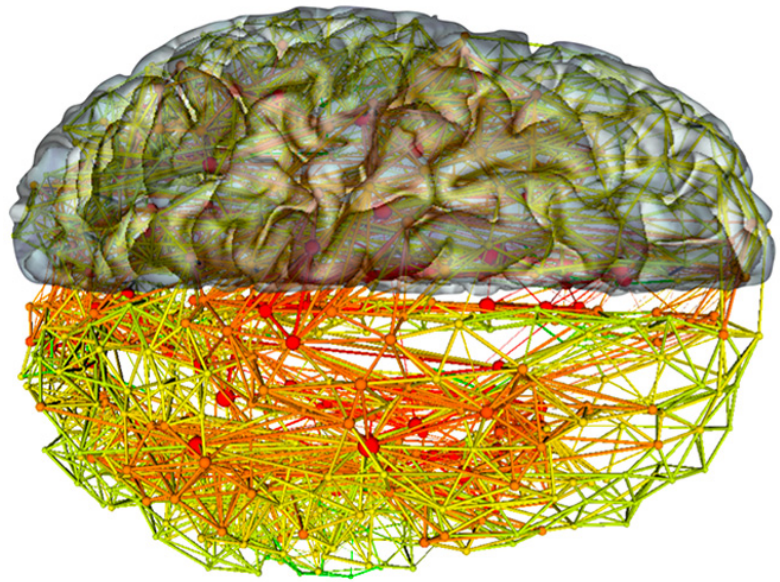
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...



Extract from Twitter
Source: AllThingsGraphed.com

Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...
- **Biology:** brain, proteins, phylogenetics, ...

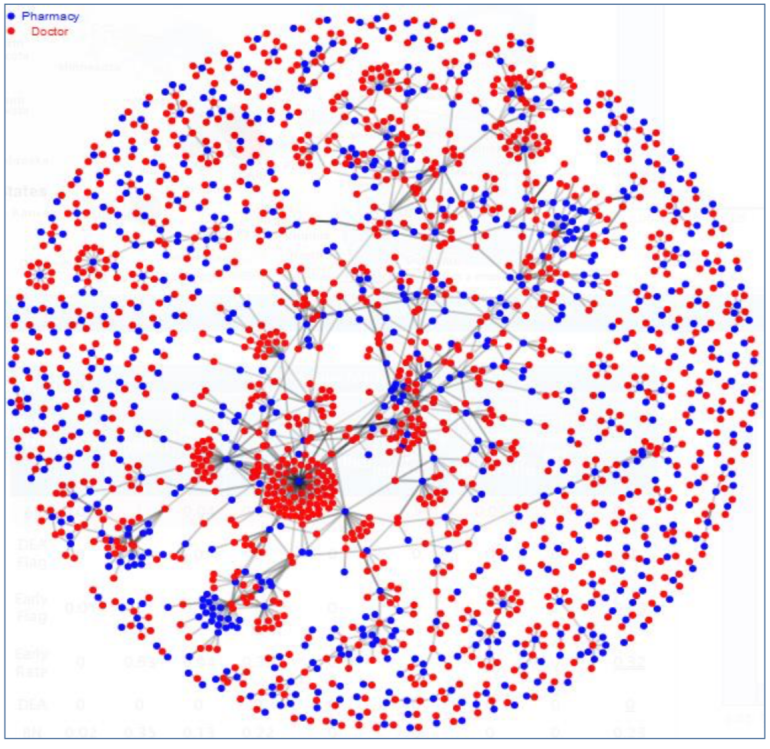


The brain network

Source: Wired

Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...
- **Biology:** brain, proteins, phylogenetics, ...
- **Health:** patient-doctor-pharmacy-drugs, ...



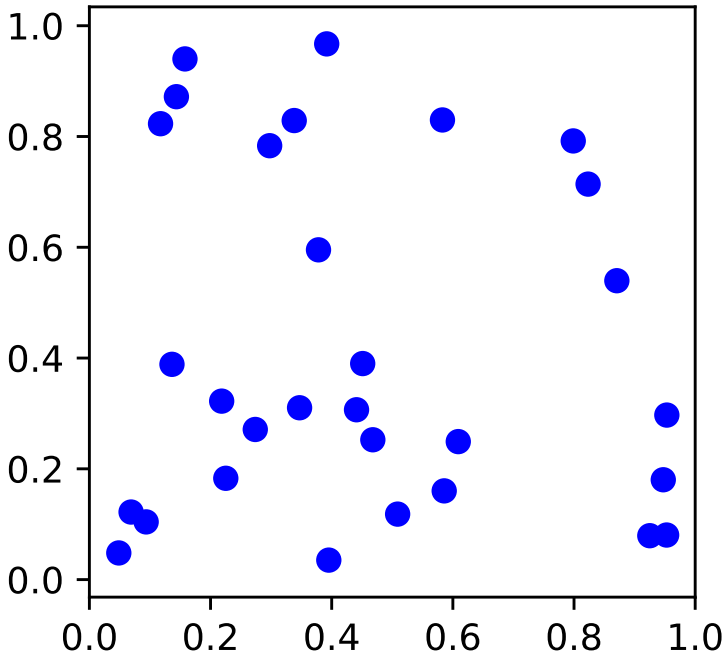
Pharmacy-doctor network
Source: IAAI 2015

Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...
- **Biology:** brain, proteins, phylogenetics, ...
- **Health:** patient-doctor-pharmacy-drugs, ...
- **Marketing:** customer-product, ...

Data as graph

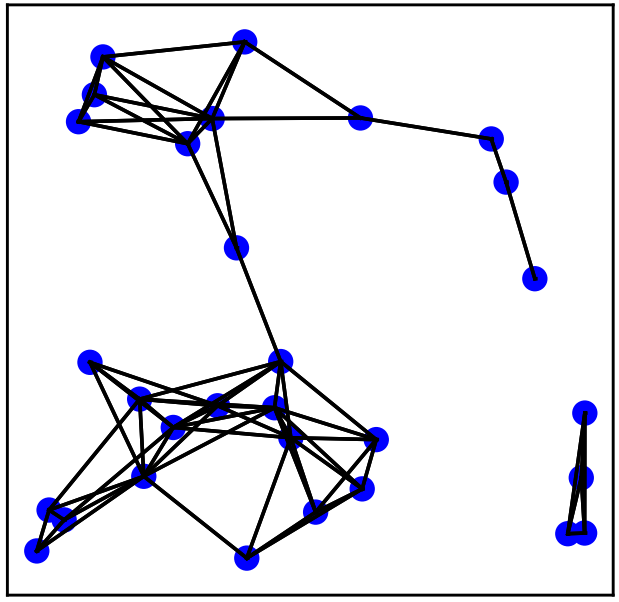
- Dataset $x_1, \ldots, x_n \in \mathcal{X}$
- Similarity measure $\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
- Graph of n nodes with weight $\sigma(x_i, x_j)$ between nodes i and j



Example: $\mathcal{X} = [0,1]^2$, $\sigma(x,y) = 1_{\{d(x,y)<1/4\}}$

Data as graph

- Dataset $x_1, \ldots, x_n \in \mathcal{X}$
- Similarity measure $\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
- Graph of n nodes with weight $\sigma(x_i, x_j)$ between nodes i and j



Example: $\mathcal{X} = [0,1]^2$, $\sigma(x,y) = 1_{\{d(x,y)<1/4\}}$

Motivation

- Information retrieval
- Content recommandation
- Advertizing
- Anomaly detection
- Security

Graph mining

Practical issues

- What are the most important nodes?
- What is the "distance" between two nodes?
- How to predict new links?
- How are organized nodes?
- How to predict labels?

Need for efficient techniques to **sample**, **rank** and **cluster** nodes

Outline of the course

1. Node and edge sampling
2. Random graphs
3. Small world
4. Betweenness centrality
5. PageRank
6. Clustering
7. Soft clustering
8. Spectral embedding

Each block = **lesson** (1TH) + **lab session** (1TH)

Lab sessions

Jupyter notebooks based on the Python `networkx` package
For each lab session, you are asked to answer

- **closed** questions
  e.g., what is the maximum degree of nodes in this graph?
- **open** question(s)
  e.g., interpret the result

You need to be connected to Telecom ParisTech network
You can work in (small) teams, but your answers must be **yours**
The deadline is the end of the lab session + **45'**
e.g., today 5:30 pm

Self-assessment, peer reviews

For open questions,

- you will be asked to rate **your** answer (self-assessement) and those of **3 other students** chosen at random and whose identities are not revealed (blind peer reviews)
- rating from **0** (no answer) to **3** (perfect answer)
- you will be given the **correct answer** and informed of the **review period**
- a dedicated Jupyter notebook

No review (and thus no open question) for the last lab

Evaluation

Based on lab sessions only
For each lab session from 1 to 7,

- **5 points** for closed answers
- **3 points** for open answer(s)
- **2 points** for the reviews

For the last lab session,

- **10 points** for closed answers

You will be informed of your results along the course
Your final mark (over 20) is the total number of points divided by 4

Your marks for each lab session

Closed answers

- You're informed online about the validity of your answer
- It is sufficient to have answer correctly **once** (on time) to a question to get the corresponding points
  False, False, True, False $\rightarrow$ True

Open answers (except the last lab)

- You will typically receive 4 marks (one from you, 3 from others)
- Your final mark is the average after **removing the extremes**
  $0, 1, 2, 2 \rightarrow 1.5$

Reviews (except the last lab)

- No reviews or late / bad reviews $\rightarrow$ **0 point**
- Some missing or bad reviews $\rightarrow$ **1 point**
- Good reviews, on time $\rightarrow$ **2 points**

Don't...

- miss the **deadlines** (lab + review)
- use a different login than **yours**
  e.g., name Marie Curie $\rightarrow$ login curiem
- **copy/paste** the answers of another student
- be **kind** or **nasty** in your reviews (just be fair!)

Pedagogical site

You will find

- the slides
- the lecture notes
- the notebooks for labs (with instructions)
- the notebooks for reviews

You will be notified by email for the review period