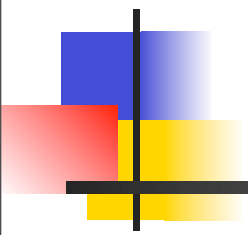
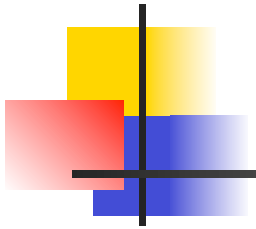


MPRI - Graph Mining

Mauro Sozio

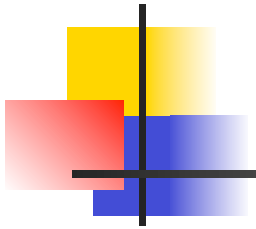
sozio@telecom-paristech.fr



About the course

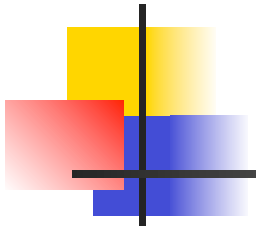
- Questions/Discussions are encouraged
- More fun if it is interactive!
- Proofs on the board, rest on slides

- Evaluation:
  - 80% based on a written exam
  - 20% based on a project
  - participation to discussions/questions



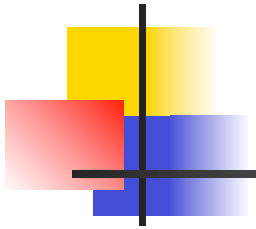
Logistic info

- Schedule:
  - Start: 19.12.18 End: 20.02.19 Exam on 27.02.19
  - Next class on 09.01.20
- Mauro Sozio, first 4 lectures, Pierluigi Crescenzi, remaining 4 lectures.
- Possibly some well known researchers in the field as guests



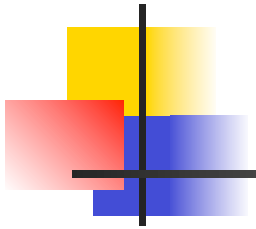
Exam

- Written Exam:
  - Exercises will be given throughout the course
  - The same or similar exercises will be asked at the exam (e.g. develop an algorithm, proof of correcteness/guarantees).
  - Closed-book exam (no notes/books)



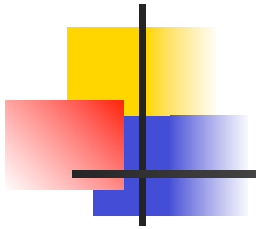
Project

- Project:
  - deadline: middle February 2019.
  - It will be announced on 09.01.2020, typically implement an algorithm we saw during the class or a related work.
  - Propose your own project and discuss it with the instructor! E.g. community detection, streaming algorithms, large scale algorithms, data analysis, approximation algorithms. Deadline 09.01.2020



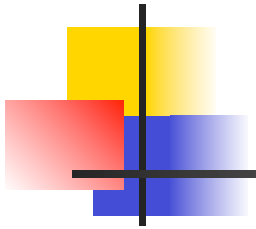
Internships

- Posted on the MPRI website:
  - Finding quasi-cliques (graphs that contain a constant fraction of the edges in a clique)
  - dynamic trajectory clustering
  - dynamic approximation algorithms for clustering
- They can be made more theoretical or practical depending on the interest of the students.



Topics

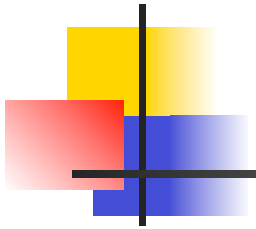
- The course gives an overview on some of the most prominent topics in graph mining:
  - PageRank
  - Community detection
  - Algorithms for finding dense subgraphs and cliques
  - Influence maximization in social networks
  - Streaming and dynamic algorithms
  - Scalable algorithms (e.g. adaptation to MapReduce/Spark)
  - Generative models for social networks
  - Distance distribution computation and centrality measures
  - Temporal graphs



Theory vs. Practice

- We will consider both:
  - simple heuristics with no guarantees that work well in practice
  - theoretical results using techniques such as LP rounding, greedy, amortized analysis, sampling

- Interplay between theory and practice: new algorithms -> more accurate studies -> interesting theoretical problems.



Types of networks

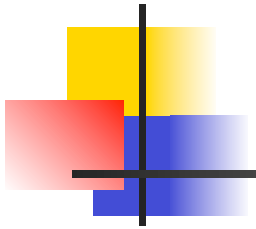
- Social networks
- Knowledge and information networks
- technology networks
- biological networks

Social Networks

- links denote social interactions
  - friendships
  - collaboration networks (actor and co-authorship networks)
  - phone call networks
  - email networks

High school dating network

Twitter follower graph



Knowledge and Information Networks

- nodes store information, links associate information
  - citation network
  - the Web
  - peer-to-peer networks
  - knowledge graph
  - blog networks

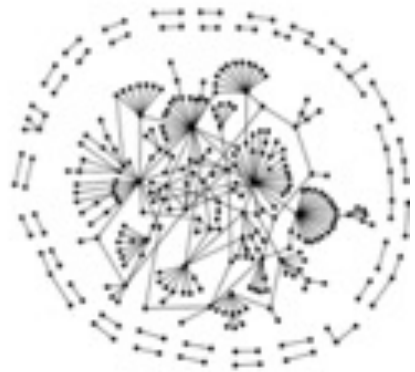
Technological networks

- Networks built for distribution of a commodity
  - the internet
  - power grids
  - telephone networks
  - transportation networks

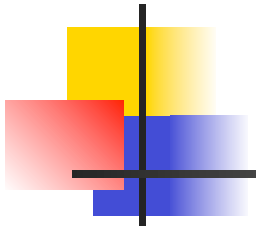
Internet network

Biological Networks

- Biological systems represented as networks:
  - protein-protein interaction systems
  - gene regulation networks
  - gene co-expression networks
  - neural networks

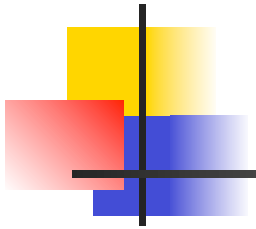


Yeast protein-protein interactions



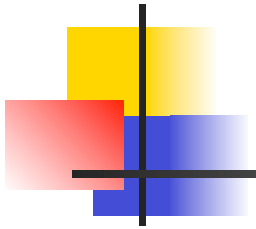
Network Science /Graph Mining

- Understand the topology of the networks and measure their properties.
- Study evolution and dynamics
- propose generative models
- devise algorithms to make sense of network data



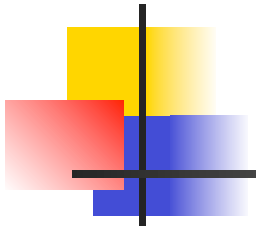
Graph Mining in the Past

- Graph datasets have been studied in the past, e.g. networks of highways, social networks.
- Graphs were small and static.
- Visual inspection can reveal useful information.



Graph mining now

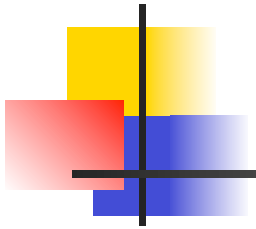
- More and larger networks are collected
- Contain thousands, millions, billions of nodes.
- Often difficult to visualize.
- More opportunities but more challenges (scalability issues, data evolve over time).
- More data might lead to different results...



Small-world experiment and six degree of separation

- Study conducted by Stanley Milgram in 1969.

- Questions:
  - What is the probability that two random people in the world know each other?
  - How many hops between them? (e.g. friend of friend of friend = 3 hops.)



Small-world experiment and six degree of separation

- Experiment:
  - Random people from Nebraska, Kansas,..., were sent a letter with the goal of forwarding it to a random person in Boston.
  - If the person knew that person then he/she could send him/her the letter directly.
  - Otherwise she could forward the letter to a relative or a friend who might know the person.
  - Some basic information about the target person were included.

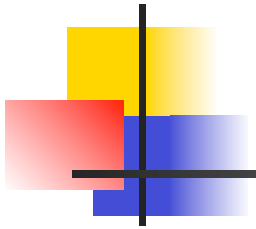
Small-world experiment and six degree of separation

One possible path of a message in the "Small World" experiment by Stanley Milgram.

Results:

- only 64 out of 296 letters reach the destination (some people refused to participate)
- among those reaching the destination, the average number of hops was ~5-6.

-> six degree of separation



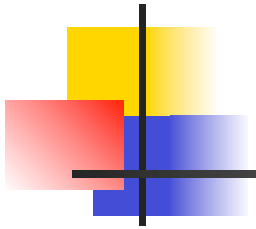
Six degree of separation in the BigData era

- Similar study on Facebook with more than 1 billion users!
- Sophisticated algorithms estimated the average path length between users: ~4!

References:

Travers, Jeffrey & Stanley Milgram. 1969. "An Experimental Study of the Small World Problem." *Sociometry*, Vol. 32, No. 4, pp. 425-443.

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, Sebastiano Vigna: Four degrees of separation. WebSci 2012:33-42



Acknowledgements

- Oana Balalau (INRIA)
- Hubert Chan (Hong Kong University)
- Maximilien Danisch (Sorbonne University)
- Aris Gionis (Aalto University, Finland)
- Ravi Kumar (Google Mountain View)
- Silvio Lattanzi (Google Zurich)
- Alessandro Panconesi (Sapienza Uni. Rome)
- Sebastiano Vigna (University of Milan)
- ...