Learning with Complex Data

Part 1: Statistical Learning with Networks - 2 hours

1 General questions

1. Explain the rational behind the stochastic block model (SBM) and what are the advantages or limits of this model compared to other approaches.
2. Explain how one can select the appropriate number of groups in statistical clustering approaches.

2 Short project: analysis of the Enron email network

2.1 The data

We consider here a classical communication network, the Enron data set, which contains all email communications between 149 employees of the famous company from 1999 to 2002. The original data set is available at `https://www.cs.cmu.edu/~./enron/`. We chose this specific time window because it is the denser period in term of sent emails and since it corresponds to a critical period for the company. Indeed, after the announcement early September 2001 that the company was "in the strongest and best shape that it has ever been in", the Securities and Exchange Commission (SEC) opened an investigation on October, 31th for fraud and the company finally filed for bankruptcy on December, 2nd, 2001. By this time, it was the largest bankruptcy in U.S. history and resulted in more than 4,000 lost jobs.

The pre-processed data are provided in the `Enron.Rdata` file available at:

`https://plmbox.math.cnrs.fr/f/f22f43e715254c61b49b/?dl=1`

The data set contains 3 different relational databases (the last one is useless):

- employeelist: the list of the Enron employees and their email addresses,
- message: all emails exchanged between 1999 and 2002,
- recipientinfo: the recipients (TO, CC, BCC) of each message

2.2 Project objectives

The objectives of this project is to (minimum objectives):

- use your knowledge in R and Data Science to reconstruct the email network from the raw data,
- visualize the data using the visualization tool of your choice (packages network, igraph or sna),
- use the LSM, LPCM and/or SBM methods to analyze the network.

You can of course extend the analysis with other tools / methods and you can focus on a specific aspect / period of the data to make more original your study. All steps of the analysis should be commented and the results should highlight some specific and meaningful facts to understand the so-called "Enron scandal".

2.3 Project instructions

The result of the project should be a Rmarkdown document / notebook contaning:

- an explanation of your positioning for this project,
- the codes used for the project (with enough comments and description),
- the interpretation of the obtained results,
- a conclusion

The implementation of question #1 and the short project have to be uploaded **as a single zip file (Rmd + html files)** under the name "StudentName.zip", by **Thursday March 30, 6:00pm**, on:

`https://plmbox.math.cnrs.fr/u/d/a8665491eecf4c2e8a3c/`