R for Big Data
Pr. Charles Bouveyron

The final reports have to be uploaded in PDF (see detailed instructions below).

1 Objectives

The examination will be in two parts on the Enron email data.

1.1 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. The data set contains 3 different relational databases:

- 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

1.2 Shiny application

Create a Shiny application that allows to explore the Enron email data. A minimal application should contains:

- an exploration of the most active Enron employees in the email database,
- an analysis of the role of the different users, according to their status,
- an analysis of the temporal dynamic of the messages, in relationship with the public events (see on the web information about the Enron scandal),
- a basic analysis of the content of the messages.

1.3 R markdown

The application should be accompanied by a R markdown document explaining the different step of the analysis and providing some personal scientific comments on the data, based on statistical results.

2 Exam instructions

The final submission should contain:

- the Rmd version of the R markdown file, including all R codes,
- a (compiled) HTML version of the R markdown file, including all R codes,
- the Shiny application (app.R).

The three files should be uploaded on Moodle. **It is not necessary to re-upload the data set on Moodle!**