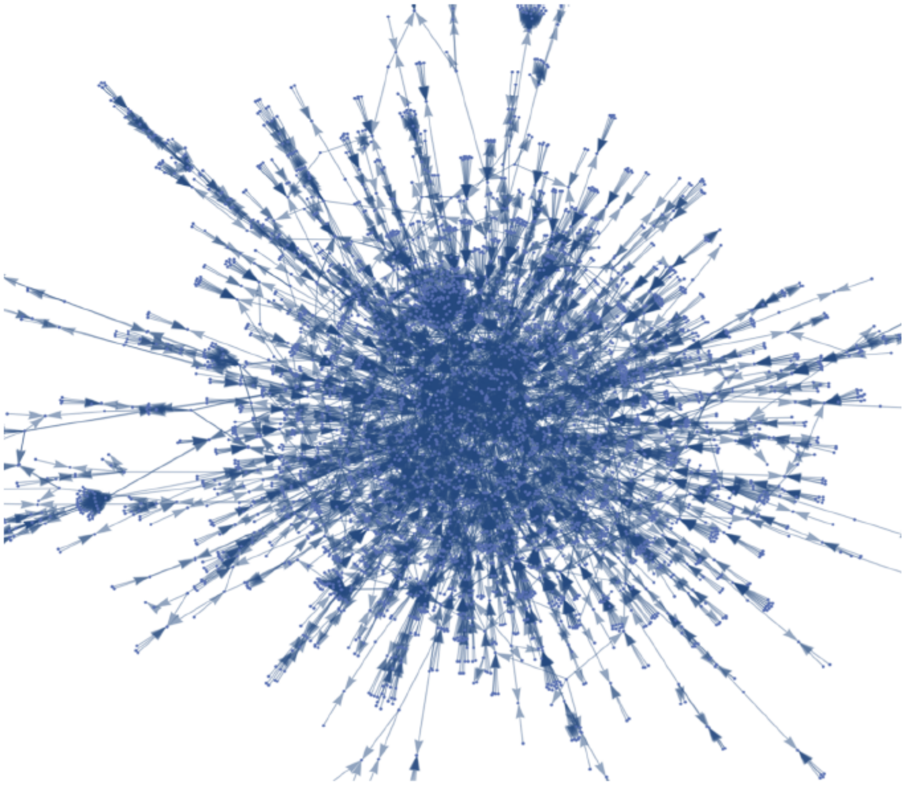
Recurrent patter in written communication

Giacomo Bassi, Olmo Notatianni.
Supervisor: Prof. Paolo Biscari
Politecnico di Milano - Physcis Engineering (Bs)

May 2021 - September 2021

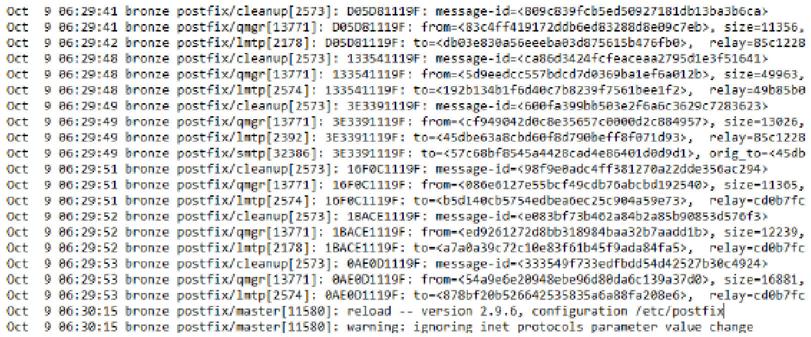


This thesis aims to identify recurrences in written communication in the context of big data. The methods used are those peculiar to statistics applied to a large database. The study case proposed is the exchange of electronic mail within a large organization, for example a university, unknown to us. The database therefore consists in tens of millions of rows 1, each of them insignificant when taken individually due of the variety of information contained. Therefore, the work was developed around few main questions: How to effectively represent a large volume of data? What processing strategy is efficient, in terms of the information obtained, given the time and computing power required? Which "quantities" are meaningful to investigate patterns and correlations inside the database?

To tackle the problem several *Wolfram's Mathematica* programs were developed during the thesis months. The main algorithm consisted in a two stage analysis. The first one considers the activity of all the addresses over a week of activity, while the second considers the whole six months span but only for a selected pool of addresses and their contacted network. Those addresses are called *Users* and their selection combines quantitative considerations, resulting from the first stage analysis, and a stochastic drawing process. Moreover to characterize exchange of emails we operate a re-clock employing a different notion of time, referred as *User's proper time.* The *User's proper time* span between two events is now measured in the number of emails sent by that *User* between event one and event two. This *decision-based* definition was argued to give a more natural way of measuring time. It in fact allows to account for the natural difference between night's and day's activities, Fig.4a. Overall this re-clock comes useful when studying interactions between *Users*, with a focus on *Users's conversation.* For each *User*, a *conversation* is define as the portion of emails sent (received) following emails received (sent) from (to) another address, namely the *Partner* of the conversation.

Starting from a database made of rows, Fig.1, after a cleaning and date reorganization a first characterization is obtained as in the histograms in Fig.3 and Fig.4. Network representations like in Fig.2 are also proposed as a graphic way to understand the database. Further in the analysis, the correlation between the volume of exchanged emails and the reply time delays in the *conversations* was investigated, allowing to draw quantitative graphs as contained in Fig.**??** and Fig.**??**. Lastly, the reply time perception within a *conversation* between pairs of *Users*, Fig. 6, is investigated with unexpected results.

In the next section, the results are briefly presented through their related figures. As a note, the figures were not recreated for this document, but simply cut out from the original document.



```
Oct  9 06:29:41 bronze postfix/cleanup[2573]: D05D81119F: message-id=<809c839fcb5ed50927181db13ba3b6ca>
Oct  9 06:29:41 bronze postfix/qmgr[13771]: D05D81119F: from=<83c4ff419172ddb6ed83288d8e09c7eb>, size=11356,
Oct  9 06:29:42 bronze postfix/lmtp[2178]: D05D81119F: to=<db03e830a56eeeba03d875615b476fb0>,  relay=85c1228
Oct  9 06:29:48 bronze postfix/cleanup[2573]: 133541119F: message-id=<ca86d3424fcfeaceaa2795d1e3f51641>
Oct  9 06:29:48 bronze postfix/qmgr[13771]: 133541119F: from=<5d9eedcc557bdcd7d0369ba1ef6a012b>, size=49963,
Oct  9 06:29:48 bronze postfix/lmtp[2574]: 133541119F: to=<192b134b1f6d40c7b8239f7561bee1f2>,  relay=49b85b0
Oct  9 06:29:49 bronze postfix/cleanup[2573]: 3E3391119F: message-id=<600fa399bb503e2f6a6c3629c7283623>
Oct  9 06:29:49 bronze postfix/qmgr[13771]: 3E3391119F: from=<cf949042d0c8e35657c0000d2c884957>, size=13026,
Oct  9 06:29:49 bronze postfix/lmtp[2392]: 3E3391119F: to=<45dbe63a8cbd60f8d790beff8f071d93>,  relay=85c1228
Oct  9 06:29:49 bronze postfix/smtp[32386]: 3E3391119F: to=<57c68bf8545a4428cad4e86401d0d9d1>, orig_to=<45db
Oct  9 06:29:51 bronze postfix/cleanup[2573]: 16F0C1119F: message-id=<98f9e0adc4ff381270a22dde356ac294>
Oct  9 06:29:51 bronze postfix/qmgr[13771]: 16F0C1119F: from=<086e0127e55bcf49cdb76abcbd192540>, size=11365,
Oct  9 06:29:51 bronze postfix/lmtp[2574]: 16F0C1119F: to=<b5d140cb5754edbea6ec25c904a59e73>,  relay=cd0b7fc
Oct  9 06:29:52 bronze postfix/cleanup[2573]: 1BACE1119F: message-id=<e083bf73b462a84b2a85b90853d576f3>
Oct  9 06:29:52 bronze postfix/qmgr[13771]: 1BACE1119F: from=<ed9261272d8bb318984baa32b7aadd1b>, size=12239,
Oct  9 06:29:52 bronze postfix/lmtp[2178]: 1BACE1119F: to=<a7a0a39c72c10e83f61b45f9ada84fa5>,  relay=cd0b7fc
Oct  9 06:29:53 bronze postfix/cleanup[2573]: 0AE0D1119F: message-id=<333549f733edfbdd54d42527b30c4924>
Oct  9 06:29:53 bronze postfix/qmgr[13771]: 0AE0D1119F: from=<54a9e6e20948ebe96d80da6c139a37d0>, size=16881,
Oct  9 06:29:53 bronze postfix/lmtp[2574]: 0AE0D1119F: to=<878bf20b526642535835a6a88fa208e6>,  relay=cd0b7fc
Oct  9 06:30:15 bronze postfix/master[11580]: reload -- version 2.9.6, configuration /etc/postfix
Oct  9 06:30:15 bronze postfix/master[11580]: warning: ignoring inet protocols parameter value change
```

Figure 1: A sample of the tens millions of rows composing the six months logbook in the database. As visible in the last two rows, postfix protocol manager records auxiliary messages unimportant for the analysis. Single emails are distributed on more than one row; the data needs to pass from a row-representation to an email-representation.

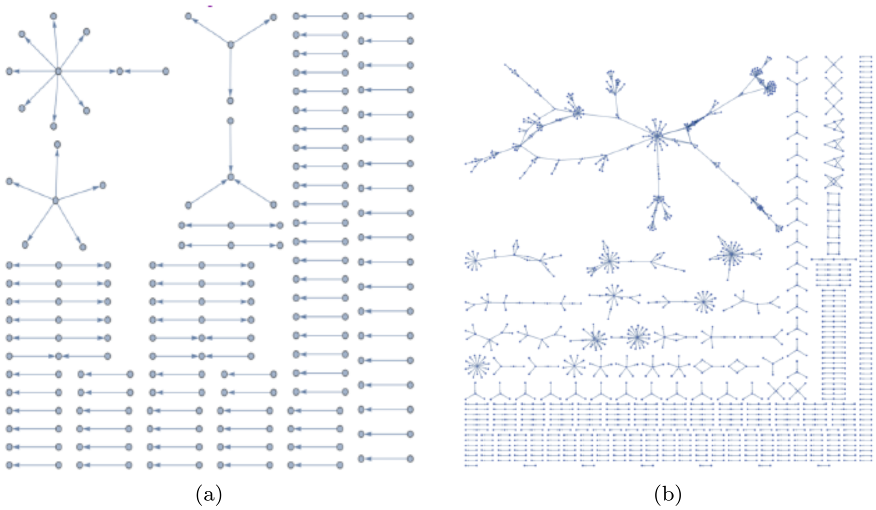
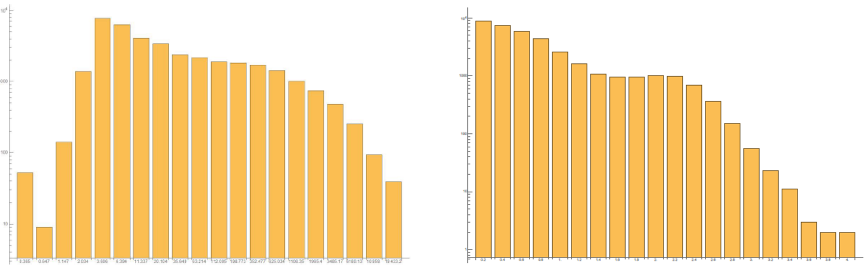
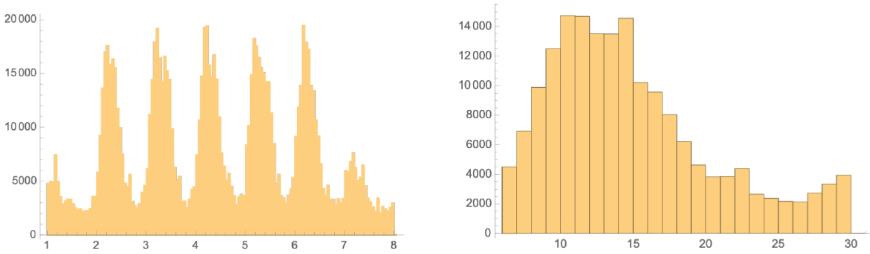


Figure 2: As the figure in the title page, simple network graphs, computed by *Wolfram's Mathematica*, were used to have a graphical representation of the database. Limited pool emails were to be chosen due do graphic computation capabilities. (a) An example of an 100 emails network. (b) 1000 emails network.



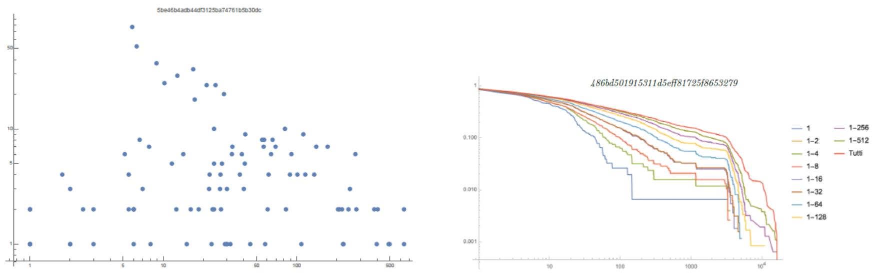
(a) (b)

Figure 3: After a cleaning algorithm and the reorganization of the data, a first characterization of one week out of six months of the database is possible. Log-log histograms are used to organize *Users* upon one of their quantifiable characteristics. (a) *Users* are distributed by the mean size, in Kilobytes, of the email sent. Assuming more real *Users* than virtual ones, e.g. machine addresses for auxiliary messages, the distribution suggest a threshold around 3 Kb of mean size after which the data is distributed seemingly by two power laws separated by a plateau. (b) *Users* are distributed by their *Activity*. A quantity, defined in the work, that counts, with a logarithm weight to account for "spam" emails and similar, the sent and received emails. The two power laws and a plateau distribution is again retrieved.



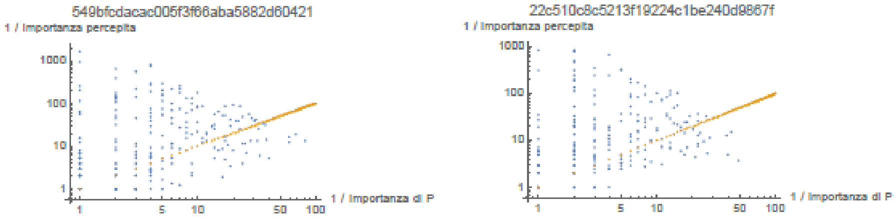
(a) (b)

Figure 4: A very meaningful characterization of one week of data comes from distributing the messages by the time recorded in the logbook. (a) On a seven days interval, day and night activity comes clearly different and this allows a characterization of week days against weekends. Given that the rows in the logbook only contain the day and the month, this is an example of a mean to retrieve a new piece of data, the year, simply comparing the email distribution with a calendar. (b) On a 24h interval, a deeper characterization of email distribution on a working day is possible. In the case where we had no information on the hour labeling the histogram bars, the label could be retrieved with arguments based on general assumptions regarding a daily routine.



(a) (b)

Figure 5: To gain insights of the kind of communication in our network, we investigated the correlation for each *User* the volume of a *conversation* with its mean reply time, calculated as the *User's proper time*. (a) In a log-log scale with the reply time on x-axis and the emails volume on y-axis, a single *User*'s *conversation* diagram presents characteristics recurring independently by the choice of user. Considering the *Partners* with whom the *User* exchanges the most emails, top of the figure, a descending behavior of the distribution is noticed. (b) Inverse cumulative distribution functions (ICDF) give other perspectives on the topic.The probability of replying to the *Partners* after x *User's proper time unit* is graphed for a different pool of *conversations*, including more and more *conversations* from the largest, with the most number of emails, to the tiniest. Again to probability seems to drop the soon the more emails the *conversations* contains. Finally to asses if the correlation is infact monotonic a Spearmen correlation test was implemented only on the most contacted *Partners* but it gave too low p-value to reject the hypothesis of independence.



(a) (b)

Figure 6: The use of an asymmetric clock between *Users* to investigate their *conversations* raises unexpected effect. For example all the selected *Users* counts more Partners to whom they respond in average quicker than they receive an answer. Being the yellow line the bisector of the graph and so the line for which the *User* feels to reply as quick as he receives answer. This was only pointed out in the thesis, but it might have unveiled a, to be better defined, pattern.