Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: SNA y redes Complejas

Complex networks and graph-mining

Study the Covid-19 pandemic effects on an academic electronic messaging network

---

Autor: Ivan Borrego García

Tutor: Jordi Nin Guerrero

Profesor: Jordi Casas Roma

Barcelona, January 8, 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	Evaluating the Covid-19 pandemic in academic environments using electronic messaging data
Nombre del autor:	Ivan Borrego García
Nombre del colaborador/a docente:	Jordi Nin Guerrero
Nombre del PRA:	Jordi Casas Roma
Fecha de entrega (mm/aaaa):	MM/AAAA
Titulación o programa:	Màster Universitari en Ciència de Dades
Área del Trabajo Final:	M2.979 - TFM - Àrea 2
Idioma del trabajo:	Inglés
Palabras clave	Complex Networks, Graph Mining, Covid-19 Confinement

Agradecimientos

I would like to especially thank my supervising teacher, Jordi Nin, for his guidance on this exciting journey through data. Also to my family, since without their support this project would not have been possible.

Abstract

During 2020, the Covid-19 pandemic has deeply affected the regular activity of every academic organization worldwide. This project aims to analyze the electronic messaging data from a Peruvian university during the months before and after the confinement establishments. By doing this, it is possible to assess how communication patterns evolve when travel and meetings were restricted and the campus initiated a lockdown.

Since email data is certainly a good indicator of the interactions within the community, the project will search in which ways it has adapted to the new situation and how work routines have been modified. To do so data will be studied at a broad level while trying to find new general trends. Aspects as the topology or the rumor spreading capacity within the network will be considered, as the metadata will help improve the understanding of the organizational aspects in the community.

Keywords: Complex networks, Graph mining, Network dynamics, Covid-19 Confinement, Organization lock down.

Resumen

Durante 2020, la pandemia de Covid-19 ha afectado profundamente a la actividad normal de organizaciones académicas en todo el mundo. Este proyecto tiene como objetivo analizar los datos de mensajería electrónica de una universidad peruana durante los meses previos y posteriores al establecimiento de confinamientos. Al hacerlo, es posible evaluar cómo evolucionaron los patrones de comunicación a medida que viajes y reuniones fueron restringidos y se inició el cierre del campus.

Dado que los datos de correo electrónico son sin duda un buen indicador de las interacciones dentro de la comunidad, el proyecto buscará de qué manera ésta se ha adaptado a la nueva situación y cómo se han modificado las rutinas de trabajo. Para ello, los datos se estudiarán en un nivel amplio mientras se intenta encontrar nuevas tendencias generales. Se considerarán aspectos como la topología o la capacidad de propagación de rumores, ya que sus metadatos ayudarán a mejorar la comprensión de los aspectos organizativos en la comunidad.

Palabras clave: Redes complejas, Minería de grafos, Dinámica de redes, Confinamiento.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The analysis and extraction of knowledge from great amounts of raw data is one of the pillars of data science [8]. The interpretation, the search of patterns, and the conversion of the findings into real, operational decisions, are among the most relevant areas in the study of data [1].

As shown in Figure 1.1, above raw *Data* we have the *Information* layer, that should be able to answer basic questions such as who, when or how many. *Knowledge*, for its part, is capable of resolve more elaborate questions, like how-to. Finally, *Wisdom* involves the exercise of judgment, that can be programmed and automated through a specific logic.

Figure 1.1: DIKW Pyramid

Data may come from many different data sources and it should be processed in accordance with the analytic needs. This project will deal with electronic message data [10]. Email data is considered a semi-structured data source because it has some defining or consistent characteristics but does not conform to a structure as rigid as is expected with a relational database. Specifically, while the actual content is unstructured (mail body and title), it does contain structured data such as name and email address of sender and recipient, time sent, etc.

We use this latter type of data to define the communication patterns used in an academic organization. Usually such patterns are studied under the umbrella of Network theory [12]. Such research field uses graphs as a basic representation of either symmetric or asymmetric relations between discrete objects such as individuals to build a complex network.

The treatment of complex networks applied to a social environment can result in an excellent way of obtaining knowledge capable to address the main questions of the project. In general, the *datification* of this kind of human interactions have proven to be an interesting and very fruitful field of information for any social related study [2]. Social network data allows a shift from modeling networks to the analysis of real-time network dynamics, as the replacement of small group studies by the analysis of social media platforms. The structure, patterns, and trends of data objects and their relations are often systematically visualized.

This project intends to find and understand how the restriction of movements and meetings have affected communications within an academic environment. The Covid-19 pandemic has had a great impact at all levels of organization, forcing a general shut-down in an attempt to control the disease spreading.

Therefore, it is interesting to study how daily tasks as traditional teaching in a face-to-face classroom, laboratory work, or administration services, have derived in alternative ways of communication and which patterns can be extracted from them. Also, the network can show how a strict global confinement affects to actual working hours and routines, and how the different departments within the university environment have reacted to that.

As we can see, this challenging situation may result in new protocols and trends affecting response times, work organization, and time management in general.

1.1 Objectives

The objectives for this project are varied.

The main objective of the project consists in detecting new trends in the behaviour of the university staff, to determine how the pandemic have affected internal protocols of work efficiency and time management.

To achieve this objective, an initial goal needs to be the transformation of raw data into a manageable, graph-like networks that facilitates the application of the necessary study techniques.

Besides, when networks are defined, we aim to study their topology and main characteristics at a global scale.

As a secondary objective it is possible to study if purely online academic environments are more likely to propagate fake news than face-to-face environments.

1.2 Planning

The total data consists of 10 monthly Excel documents (from december 2019 to september 2020) that contain anonymized data for every email sent. Specifically the dataset contains: Sender, Receiver(s), Date, Time, Subject, and Area + Type of Personnel for both sender and receiver.

The data is in need of processing and cleaning in order to adapt it to the objectives of the project. An adjacency matrix will be implemented to facilitate the extraction of knowledge. Their results will provide enough material to run a thorough study on them.

Parallel to this data processing, a detailed search and study of network theory documentation is conducted, in order to acquire the necessary knowledge and ideas to carry out the project. Principal search platforms are Google Scholar or Plos One, or also the material studied during the master's degree.

Table 1.1 shows an approximate weekly work plan.

Project	**Sep'20**				**Oct'20**				**Nov'20**				**Dec'20**				**Jan'21**			
Definition and Planning																				
Study of Related Work																				
Data Cleaning and Filtering																				
Data Analysis																				
Conclusions and Visualizations																				
Document Work Memory																				
Delivery of Results																				

Table 1.1: Weekly Work Plan.

Coding will be done in *Python*. *Pandas* and *numpy* libraries are used to process the data. *igraph* is the selected library to create and manage the network itself, while *matplotlib* is incorporated for data visualization. The memory of the project will be done using *LaTeX*.

The implementation and technical part of the project also faces some issues regarding space and processing time, due to the size of the working dataset. Sorting and cleaning algorithms will run on an Intel i7 with 4-cores @3.5Ghz that allow multi-threading. Installed RAM is 16Gb, and although hard drive space is not a problem, any file created will be compressed in order to drastically improve their reading time and optimize general usage.

Chapter 2

Preliminaries

Human interactions via information and communication technologies have been a field of interest for complex networks since they became relevant. The structure of these networks has been studied using the language of graph theory, a branch of mathematics. Literature about the structure and dynamics of complex networks is available [5, 15].

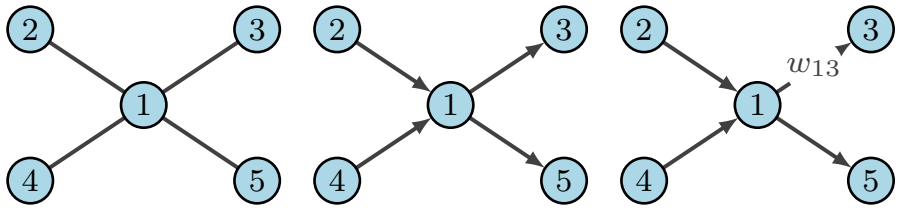
A graph is a mathematical abstraction consisting of a set of N nodes or vertices, connected by a set of E edges or links. Nodes are usually depicted as circles, and lines between them represent existing relationships. In a social network, nodes tend to be people, and connections map their interactions (email communications for example).

Networks can be classified according to several topological properties, where the most basic classification refers to the nature of the interactions. Networks can be directed or undirected, depending on whether the directionality of connections matters for the analysis and interpretation: it may be relevant to consider who sends/receives the messages, or it is possible to consider the communication reciprocal. As far as the interaction strength is concerned, links may be weighted or unweighted: if two individuals communicate frequently, their bond will be stronger (or heavier) than if they write each other only occasionally; a weighted network records the information of this frequency of interaction.

Every graph can be represented in a matrix notation through the so-called adjacency matrix A. This is a $N \times N$ squared matrix where the entries $a_{ij} = w_{ij}$ indicate the existence of a link of weight w_{ij} from vertex i to j. Adjacency matrices representing undirected networks are symmetric, $a_{ij} = a_{ji}$, while unweighted networks generate binary matrices, $a_{ij} \in \{0, 1\}$.

Adjacency matrices may represent a fully connected network, where all nodes are connected among them, or a network structure divided into different connected components. A connected component is understood as each section of a graph where all nodes are somehow interconnected. When having different components, usually we need to resort to study the principal connected component, that is, the largest one.

The most simple and studied property of a node or vertex in a graph is the connectivity, or degree, of a node i: k_i. This counts the number of edges connecting node i to other nodes in the network. Also, if the network is weighted, it is possible to add the amount in each of its edges to find the total strength of the node. In directed networks we can define them as in(out)-degree and in(out)-strength, where we count the number of incoming and outgoing edges and their corresponding weight. Figure 2.1 shows several examples of very basic graphs and their respective adjacency matrix.



0	1	1	1	1
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0

(a) Undirected

0	0	1	0	1
1	0	0	0	0
0	0	0	0	0
1	0	0	0	0
0	0	0	0	0

(b) directed

0	0	4	0	2
2	0	0	0	0
0	0	0	0	0
2	0	0	0	0
0	0	0	0	0

(c) weigthed

Figure 2.1: Network examples

Although connectivity is a local property of vertices, degree distributions often determine some important global characteristics of networks, and they help determine a classification according to the homogeneity of the degree distribution.

In graph theory, the redundancy of connections amongst the neighbours of a vertex i describes another important structural property of networks. It represents a measure of the degree to which nodes in a graph tend to cluster together, also known as clustering coefficient. Two versions of this measure exist: the global and the local. The global version was designed to give an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes.

The local clustering coefficient of a vertex (node) in a graph quantifies how close its neighbours are to being a clique (a complete graph) [22]. A neighbourhood N_i for a vertex v_i is defined as its immediately connected neighbours:

$$N_i = \{v_j : e_{ij} \in E \vee e_{ji} \in E\}$$

The local clustering coefficient C_i for a vertex v_i is then given by the proportion of links between the vertices within its neighbourhood divided by the number of links that could possibly exist between them. For a directed graph, e_{ij} is distinct from e_{ji}, and therefore for each

neighbourhood N_i there are $k_i(k_i - 1)$ links that could exist among the vertices within the neighbourhood (k_i being the number of neighbours of a vertex). The local clustering coefficient for directed graphs is given as

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

The global clustering coefficient is based on triplets of nodes. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties [14]. A triangle graph therefore includes three closed triplets, one centered on each of the nodes (n.b. this means the three triplets in a triangle come from overlapping selections of nodes). The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed). This measure gives an indication of the clustering in the whole network (global), and can be applied to both directed and undirected networks, also known as transitivity [20].

So, the global clustering coefficient C is defined as:

$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}}$$

Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tight groups characterised by a relatively high density of relationships. this likelihood tends to be greater than the average probability of a tie randomly established between two nodes [9].

Another relevant characteristic of a vertex is its betweenness centrality, a measure based on the concept of shortest path, l_{i_j}. For every pair of nodes i and j, the shortest path is the minimum of the possible paths starting at node i and ending at j. It corresponds to the number of edges comprising the path, and allows the categorization of networks in two groups: connected graphs, where there is at least one shortest path between all pairs of vertices (*strongly* connected, if directionality is taken into account); and non-connected graphs, which are made up of broken sub-graphs. The study of the biggest connected component in a network and the size distribution of lesser connected sub-graphs offers another dimension in the description of network topology.

Chapter 3

State of the art

The study of diffusion dynamics has a long tradition in the social sciences [17]. Most of these studies, however, are based on aggregated data of adoption rates, which are compatible with a number of individual-level mechanisms. The study of contagion dynamics often makes explicit the effects of network structure on adoption rates: the assumption is that information or behaviour diffuses in a population because adopters are exposed to previous adopters via their networks, which delineate the boundaries of their group of reference.

3.1 Epidemic and rumour models

The mathematics of epidemic spreading were originally developed in the fields of Medicine and Biology [3]. Their application to information cascades came through physicists and computer scientists, who found in epidemic spreading a valid metaphor of information propagation. This approach assumes that information travels through social networks as viral infections, and that personal interactions open the diffusion routes.

According to these models, contagion dynamics evolve following a simple scheme: at each time step, infected individuals propagate the contagion to susceptible neighbours with probability λ. Additionally, infected individuals can recover at a rate μ (as in the susceptible–infected–recovered, or SIR, models); or they can revert to the susceptible state with probability μ (as in the susceptible–infected–susceptible, or SIS, models). Pastor-Satorras & Vespignani [16] analytically established, for the SIS model, that the critical point (or epidemic threshold) in uncorrelated scale-free networks is given by $\lambda_c = \frac{k}{k^2}$, leading to $\lambda_c \to 0$ as $N \to \infty$ when $2 < \gamma \leq 3$. Tuning the infection probability, they can reproduce the most frequent cascades as well as match cascade size distributions.

A different approach to contagion goes deeper into the mechanisms that allow epidemic dynamics to unfold. Unlike what happens with viral epidemics, social contagion relies on the

effects of social influence, which is at the core of sociological research. Information, like viruses, can propagate with a single exposure, but the spread of behaviour often requires multiple exposure from multiples sources. Evidence of this type of contagion has been found in a number of online settings [13].

A question that has naturally followed from the study of these contagious dynamics is where seeds are in the network topology, that is, if the leaders of the process have a specific network position.

Across all these areas of application, disease spreading has become a rather usual benchmark to identify the network features —mainly degree and centrality measures— that perform better when identifying *outstanding spreaders*, the nodes in the network that trigger larger cascades.

Rumour dynamics have also been simulated in parallel to more general models of social influence. These models sprung directly from the canonical SIR, renaming the susceptible, infected and recovered classes (SIR) to ignorant (who has not heard the rumour yet), spreader (who knows the rumour and is *infecting* ignorants) and stifler (who also knows the rumour, but has decided not to spread it further). Although rumour models are often regarded as a simple mapping of its epidemic counterpart, some differences set them apart [4].

Networks can be used as a simulation model to study possible relationships between echo chambers and the viral spread of information. In these simulations an *echo chamber effect* can be found [19], i.e., the fact that the presence of an opinion within a polarized cluster of nodes in a network contributes to the diffusion of complex contagions. A synergetic effect is generated between opinion and network polarization on relation to the virality.

The probability that activation will spread to a majority of the network's nodes, given a certain distribution of network structures, is known as the *virality* of the diffusion. A cascade can initiate due to the activation of a randomly selected node and its neighbors. In each following time step, nodes that have more than a certain fraction —threshold— of their neighborhood activated themselves become activated. This continues until a steady state is reached. If at this point most of the network has become activated, the cascade is classified as successful.

3.2 Information Cascades

Information cascades are important dynamical processes in complex networks. An information cascade can describe the spreading dynamics of rumour, disease, memes, or marketing campaigns, which initially start from a node or a set of nodes in the network. If conditions are right, information cascades rapidly encompass large parts of the network, leading to epidemic spreading. Certain network topologies are particularly conductive to epidemics, while others decelerate and even prohibit rapid information spreading.

There are models that describe information cascades in complex networks [11] (with an emphasis on the role and consequences of node centrality) that obtain simulation results on sample networks revealing just how relevant the centrality of initiator nodes is on the latter development of an information cascade. It also defines the spreading influence of a node as the fraction of nodes that is activated as a result of the initial activation of that node. This shows that some centrality measures, such as the degree and betweenness, are positively correlated with the spreading influence, while other centrality measures, such as eccentricity and the information index, have negative correlation. A positive correlation implies that choosing a node with the highest centrality value will activate the largest number of nodes, while a negative correlation implies that the node with the lowest centrality value will have the same effect. This can be used used to study how information cascades help identify nodes with the highest spreading capability in complex networks.

3.3 Threshold models

Models of collective behavior are developed for situations where actors have two alternatives and the costs and/or benefits of each depend on how many other actors choose which alternative [7]. The key concept is that of *threshold*: the number or proportion of others who must make one decision before a given actor does so. Beginning with a frequency distribution of thresholds, the models allow calculation of the *equilibrium* number making each decision. The stability of equilibrium results against various possible changes in threshold distributions is considered. Stress is placed on the importance of exact distributions for outcomes. Groups with similar average preferences may generate very different results. So it is difficult to infer individual dispositions from aggregate outcomes or to assume that behavior was directed by common rules. Possible applications are to riot behavior, innovation and rumor diffusion, strikes, voting, and migration.

The first attempt to interweave cascading phenomena and complex networks built on previous work on the diffusion effects of interdependent decision-making [21]. In this article, Watts provides an analytic approach to discern the conditions under which global cascades may occur in structured sparse topologies. Using percolation methods, the model explores how network topology and individual thresholds interact in the spreading of behaviour.

3.4 Network analysis in a real-world organization

The Enron email corpus has been appealing to researchers because it represents a rich temporal record of internal communication within a large, real-world organization facing a severe and

survival-threatening crisis. Previous works [6] have explored the dynamics of the structure and the properties of the organizational communication network, as well as the characteristics and patterns of communicative behavior of the employees from different organizational levels. It was found that during the crisis period, communication among employees became more diverse with respect to established contacts and formal roles. Also during the crisis period, previously disconnected employees began to engage in mutual communication, so that interpersonal communication was intensified and spread through the network, bypassing formal chains of communication.

A major problem in social network analysis and link discovery is the discovery of hidden organizational structure and selection of interesting influential members. The work of Shetty and Adibi [18] used entropy models to identify the most interesting and important nodes in a graph. There we can see how entropy models on graphs are relevant to the study of information flow in an organization, as the results of two different experiments both based on entropy models are reviewed.

Chapter 4

Data Preprocessing

Data origin is the *Universidad del Pacífico*, a private peruvian university located in Lima and mainly focused in studies on economy, finances, marketing and international business. Besides its career degrees, it has several other sections, including a Business School, a Public Management School, and a Language Center. Its academic community of workers -including teachers, researchers and administration personnel- form the messaging network used in this study.

4.1 Extract, Transform and Load

Raw data comes directly from origin in the form of monthly excel files. Each file is composed of 9 columns as seen in table 4.1.

Attribute	**Format**
Emisor	the domain name is shown as it really is, while the username comes anonimized using a hexadecimal code.
Receptor	same as *Emisor*
FechaEnvio	YYYY-MM-DD date
HoraEnvio	hh:mm:ss time
TituloCorreo	string
AreaEmisor	string, discrete values
AreaReceptor	string, same discrete values as *AreaEmisor*
TipoPersonalEmisor	string, discrete values
TipoPersonalReceptor	string, same discrete values as *TipoPersonalEmisor*

Table 4.1: Original attributes.

Every field is mainly self-explanatory. *Area* and *Personnel Type* refer to the employee relationship the owner of the address (sender or receiver) has with the university.

Rows with no *Emisor*, *Receptor* or *FechaEnvio* are considered non-valid. *Receptor* can be a list, in which case they are being treated as different rows. The pair formed by *Area* and *TipoPersonal* for both *Emisor* and *Receptor* not always have values, but when they have, it is true for both of them. *HoraEnvio* and *TituloCorreo* will not have a significant impact in this study, so they just can be ignored.

Taking this conditions into account, a fitting pipeline is created to clean and adapt the monthly files into several data sets. These resulting data sets can be mainly separated into one comprehensive list of interactions -collecting *sender, receiver, date* info-, plus six more data sets created with the *address, area, type* attributes, and differentiated according to their domain name and if they have info in the area/type attributes or not. See Figure 4.3

4.2 Data facts

After this first treatment, we can count a total of $21,344,736$ interactions *Emisor-Receptor* going from December 2019 to September 2020, both months included. Next step is to form a list of email addresses with Area/Type information. The result contains 28564 unique addresses with info on their area and type, as seen in Figure 4.1

Figure 4.1: Addresses with info on their Area/Type of personnel

On the other side, we find $383,559$ unique addresses with no specific employee relationship to the university, as seen in figure 4.2

Since the resulting figures were pretty high, data provider was consulted in order to know if any kind of aliases list exists. The answer came late in the design process, with the analysis and code implementation already completed. Still, the list was not comprehensive, and the impact over the nodes that make up the study was minimal (the aliases reduced its numbers from the original $4,615$ to $4,525$). So the network figures seen in the project do not incorporate this alias list.

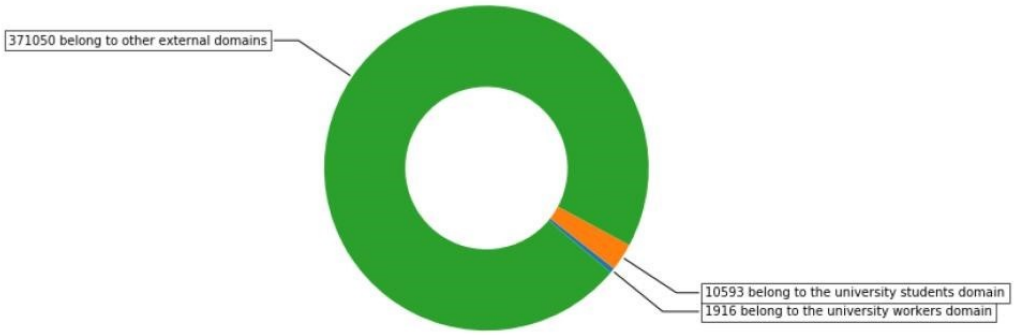


Figure 4.2: Addresses with no specific employee relationship to the university

Looking at the discrete fields of Area and Type we can see that there are 67 possible areas and 420 different types of personnel. Parallel to this high number of types, the university has a shorter list of just 6 types that summarizes all the others:

- Academic Services
- Campus Staff
- General Services
- Intern
- IT Staff
- Professor

It also includes the possibility of adding a 6-digit binary code that masks several characteristics for every type of job. Specifically, each flag of the code indicates if the position is an administrative, a researcher, a manager, a full-timer, an assistant or a coordinator. Some of them are not mutually exclusive. The dataframe in Figure 4.3 shows this modifications.

	Emisor	Receptor	FechaEnvio
0	EB1F535935EFC72D4AFBC577@outlook.com	A1F61FE0BE29EBC364C503CB@mailing.up.edu.pe	2019-11-30
1	1E9A1A2B0352D5BA0B9B9354@alum.up.edu.pe	2E8997E5F38013A626C9D021@up.edu.pe	2019-11-30
2	B6A785A03BBEC19FF7E73A53@mindpressive.com	07FD0AB0A4B218CC7776090E@up.edu.pe	2019-11-30
3	D69098AF12B1BC435E71B32B@up.edu.pe	D69098AF12B1BC435E71B32B@up.edu.pe	2019-11-30
4	0B21868B4D7E4D9390E03E9C@oechsle.pe	520B0E679CF4D6AB3CD2DDE9@up.edu.pe	2019-11-30
...	...	...	...
21344731	0B21BFF983F0F94BD463DC75@email.mckinsey.com	526C3EC28A40296A4BDCE266@up.edu.pe	2020-09-30
21344732	5B88CA1CFAAA095BDD3EA70F@valor.com.pe	BA40676ABE950AA8F781F2A7@up.edu.pe	2020-09-30
21344733	11B3B503E4754A150CD55C67@euromonitor.com	BD4EAF8286CD4E8A50B31422@up.edu.pe	2020-09-30
21344734	11B3B503E4754A150CD55C67@euromonitor.com	FC56FE8F2D8963DC89A5DF7C@up.edu.pe	2020-09-30
21344735	11B3B503E4754A150CD55C67@kayak.com	BCBBEF7906A3F89E5AA0EE42@up.edu.pe	2020-09-30

[21344736 rows x 3 columns]

Total mails sent

	Node	Area	Tipo	Tipog	Codi
0	D69098AF12B1BC435E71B32B@up.edu.pe	DPTO. ACAD. DE CS. SOCIALES Y	PROFESOR PRINCIPAL	PROFESSOR	010100
1	C2027A124A0732D0C3541E8B@up.edu.pe	DPTO. ACAD. DE ADMINISTRACIÓN	PROFESOR CONTRATADO A TIEMPO PARCIAL	PROFESSOR	010000
2	82B49CB309214A59FF2D7BB1@up.edu.pe	GESTIÓN DE LA INFORMACIÓN E IN	ADMINISTRADOR DE SERVIDORES	GENERAL SERVICES	100100
3	515191E869352A1C15B3E373@up.edu.pe	DPTO. ACAD. DE CS. SOCIALES Y	PROFESOR CONTRATADO A TIEMPO PARCIAL	PROFESSOR	010000
4	E44B8B38B19C84432308AB37@up.edu.pe	CENTRO DE IDIOMAS	COORDINADORA ADMINISTRATIVA - CIDUP	GENERAL SERVICES	100101
...	...	...	...	...	...
4610	884F16F79F7F17E9ADCCA251@up.edu.pe	ESCUELA DE POSTGRADO	ASISTENTE DE SECRETARÍA ACADÉMICA	ACADEMIC SERVICES	100110
4611	7E8BBA10354468F9876904BC@up.edu.pe	ESCUELA DE POSTGRADO	GERENTE DE MARKETING DE POSTGRADO	GENERAL SERVICES	101100
4612	C48C92190D87C0B74CDE13A8@up.edu.pe	ESCUELA DE GESTIÓN PÚBLICA	GERENTE DE PROYECTOS - EGP	GENERAL SERVICES	101100
4613	4CA29DAE7A6A66B732FE1471@up.edu.pe	COMUNICACIONES E IMAGEN INSTIT	JEFA DE IMAGEN INSTITUCIONAL	GENERAL SERVICES	101100
4614	9D3C7881F97DAE9AA17D2DBD@up.edu.pe	CENTRO DE IDIOMAS	INSTRUCTORA	CAMPUS STAFF	010100

[4615 rows x 5 columns]

Workers addresses with info

Figure 4.3: Dataframes resulting from preprocess

Chapter 5

Data Analysis

The size and complexity of the dataset required making some relevant decisions at this point. Taking into account the original purpose of studying changes in the behavior of the academic community, the focus will be set on the addresses that belong to the university workers domain and have a determined area and position. This attributes will be translated and reduced to one of the 6 unique general types established by the university, as they are enough for the study conducted in this project.

5.1 Preliminary steps

As we said, the main combination will be the addresses with a worker domain and info on their job position. The other combinations will be treated as super-nodes that collect all their inputs and outputs.

In a first attempt, the main group was also prepared to work as a different number of super-nodes. Several levels of granularity could be applied, from the individual nodes to a few super-nodes for any attributes (area, general type, or specific binary-mask code) combination possible. So, although this branch of code is finally not included in this project, it can be used for other studies on the same dataset.

Also, there was the question of the time sections that will be studied. The project can use daily, weekly, or monthly data, or any other measure that could be deemed appropriate.

Once assured that the memory space and the processing time needed to do the study were acceptable, the final decision was to use just the individual nodes, with no other super-nodes that could interfere with the network topology. This was aligned with the premises of the study, and allows to apply different possibilities on the *timeline* side.

So, the first detailed section in the study is aimed at daily data from the *worker domain + info on position* nodes.

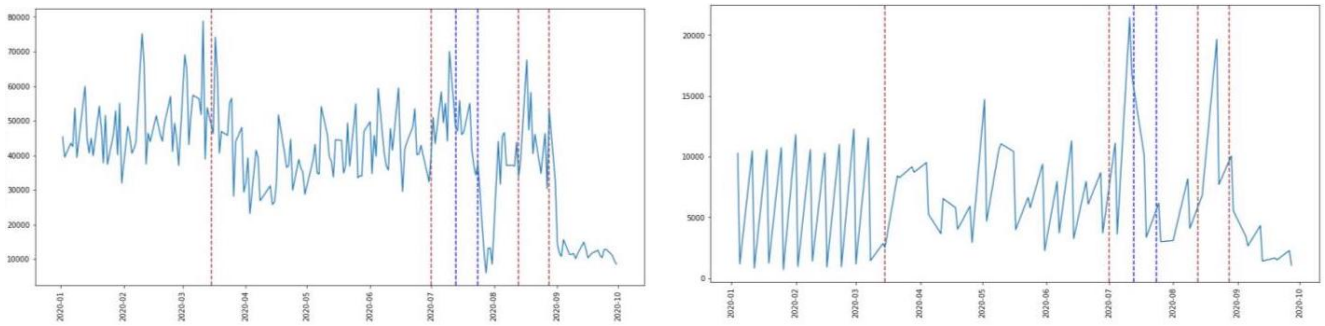
5.2 Daily analysis

A total of 274 graphs are generated, from the 1st of January to the 30th of September. Basic analysis, like total mails sent 5.1 or graph density 5.2, return some interesting insight.

The timeline has been divided in business days and weekends, since their data were obviously different. Four relevant dates regarding the pandemic have been indicated in red:

- March 15, 2020: The country declares a state on national emergency, Including mandatory social isolation.
- July 1, 2020: The quarantine officially ends.
- August 13, 2020: Return to confinement.
- August 28, 2020: Restrictive measures are extended for another month.

While blue marks represent *final exams weeks* (July 13, 2020) and *final exams reviews* (July 24, 2020).



Total mails sent - Business days Total mails sent - Weekends

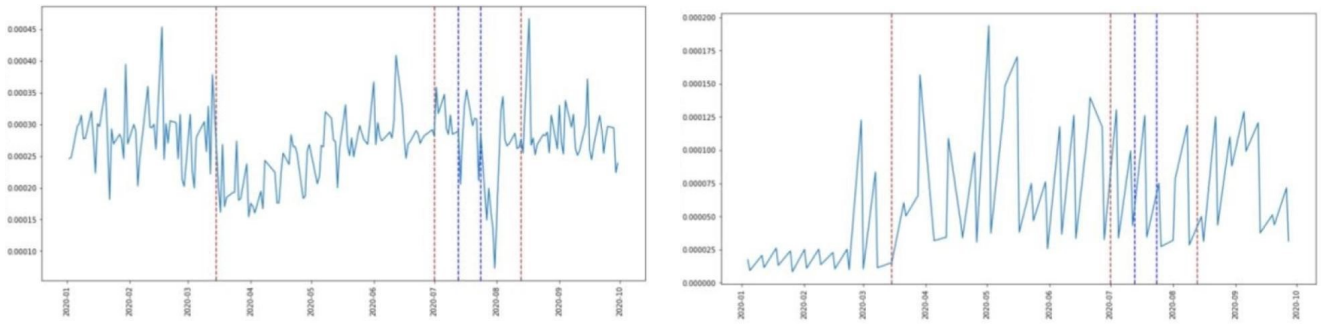
Figure 5.1: Daily number of mails sent

It is remarkable the impact that the first milestone (start of quarantine) has on both charts in Figure 5.1. While the raw number of emails on business days decline -and won't recover in a few months-, the weekends show a very different and more irregular trend, that maintain and increase the number of messages sent.

Also, it is interesting to see how the last milestone sets a plummet on both charts in Figure 5.1, probably pointing at a change in work methodology.

A second Figure 5.2 shows the evolution of the density, a measure representing the number of actual connections versus all possible connections defined as follows:

$$D = \frac{|A|}{N * (N - 1)}$$



Graph density - Business days | Graph density - Weekends

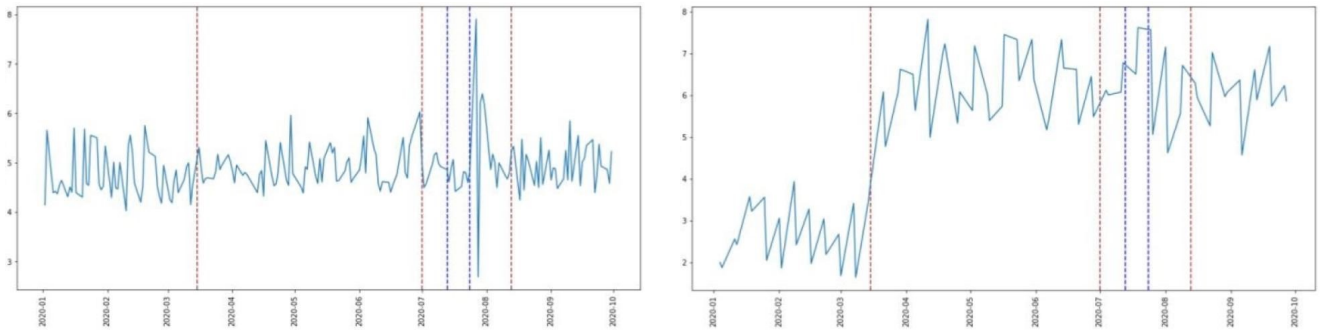
Figure 5.2: Graph density

where $|A|$ and N stand for the number of 1 in the adjacency matrix (*i.e.* the number of active edges) and the number of nodes respectively.

Observing Figure 5.2, we can see that it follows a similar general trend than the total sent charts, confirming the results. This measure is more graph-oriented, and does not show the decrease at the end of August, indicating that people kept in touch with the same number of contacts despite the decrease in the raw number of mails.

5.2.1 Average Path

Figure 5.3 refers to the *average path length* measure, that is, the average number of steps along the shortest paths for all possible pairs of network nodes. It allows to form an accurate idea on how efficient the transmission of information can be.



Average path length - Business days | Average path length - Weekends

Figure 5.3: Average path length

As we can see in Figure 5.3, the average path for business days remains fairly constant around 5 steps, which is a quite normal result that follows along the *small-world* theory of the six degrees of separation. The only remarkable trait is that the variance of the measure is lower during the beginning of the confinement, probably due to the necessity of reorganization.

On the other side, weekends network varies drastically since the moment of the first lockdown, doubling its average from around 2-3 to 6-7. So it is possible to infer that new actors are added to the network, since the number of edges must be higher. These new relations are highly clustered, meaning that workers communicate more frequently out of working days, but only with their closest collaborators.

5.2.2 Assortativity

The assortativity is defined by the preference of network's nodes to attach to others that are somehow similar to them. Figure 5.4 shows the daily assortativity for each one of the 6 different types of workers. Communities as *Campus Staff*, *General Services* or *Professor* reach their peak assortativity coefficient at 0.4 short after the first confinement. This increase can be interpreted as an augment of the self-organization in these specific areas due to the need to establish new ways of operating.

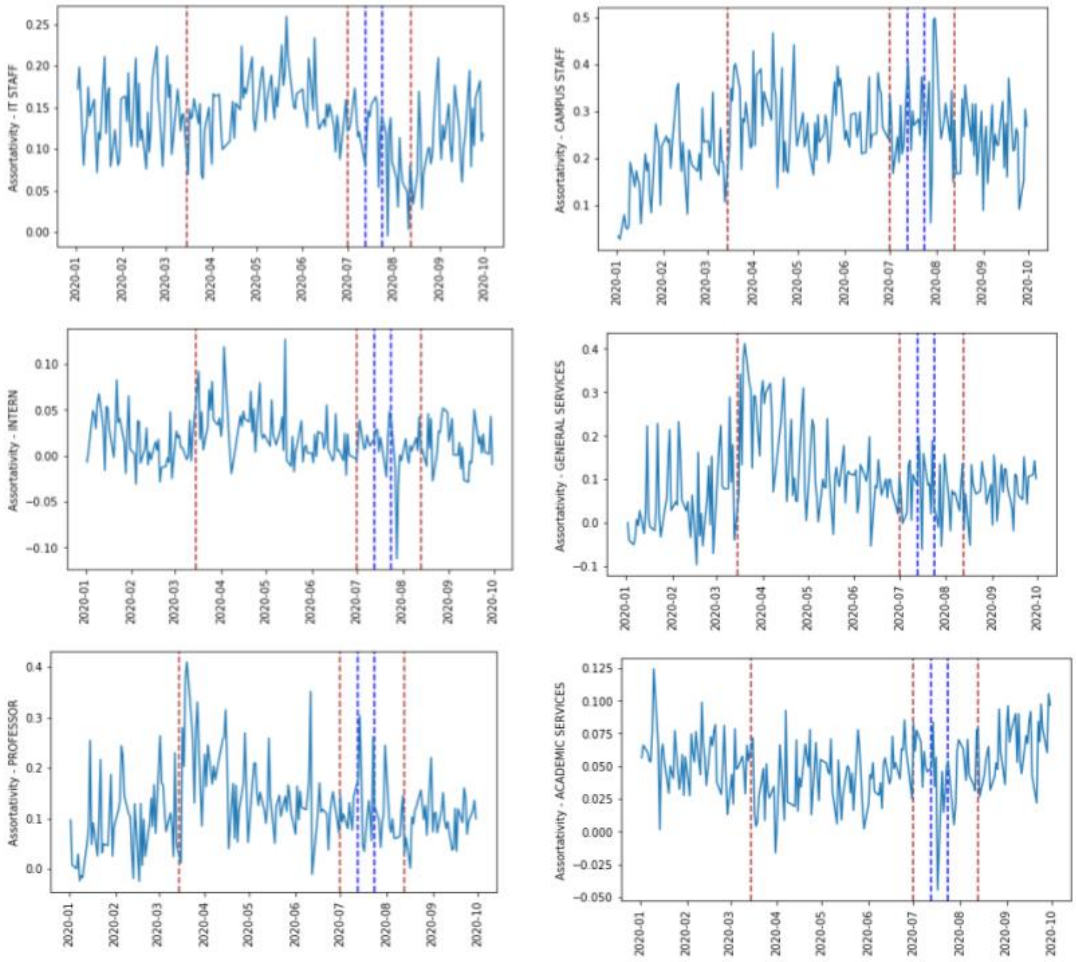
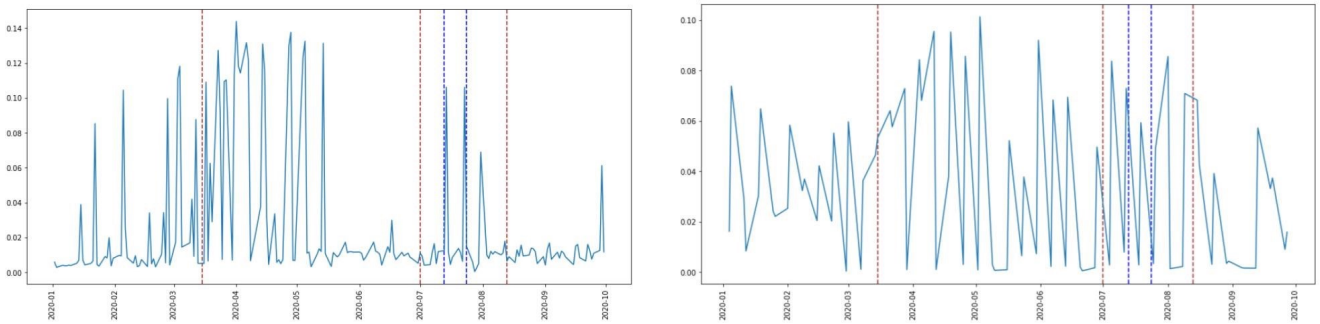


Figure 5.4: Assortativity

On the other side *Academic Services* personnel lower its coefficient to less than 0.05. This specific area needs to inform and coordinate others, hence its decrease in assortativity during lockdown times.

5.2.3 Global Clustering Coefficient

This global coefficient is designed to give an overall indication of the clustering in the network. Again, it will indicate how tight the relationships are within the several sections in the community.



Global clustering coefficient - Business days Global clustering coefficient - Weekends

Figure 5.5: Global clustering coefficient

As we can see in Figure 5.5, the global clustering coefficient for business days chart shows some interesting data. Although with heavy spikes, the daily graphs both immediately before and after the first confinement consistently exceed an 0.1 coefficient. Once the situation is normalized, the coefficient remains mainly below 0.02.

This decrease in the coefficient as the crisis settled is due to the augment of nodes in the network that only interact with their closest colleagues. This can be interpreted in the sense that relationships that took place at the working space turn into virtual once the new organization set in.

5.3 Periodic analysis

As aforementioned, another approach would be to study some specific periods of the year. Looking at the charts covered so far, we decided to create 4 periods of 30 days to look at their topological characteristics. The selected periods are, specifically, the days before the first confinement, after the first confinement, before the relaxation on the restrictions, and after the second confinement.

Figure 5.6 present these 4 networks. Nodes are ordered by type and degree to facilitate data visualization, so their position on the x axis may vary.

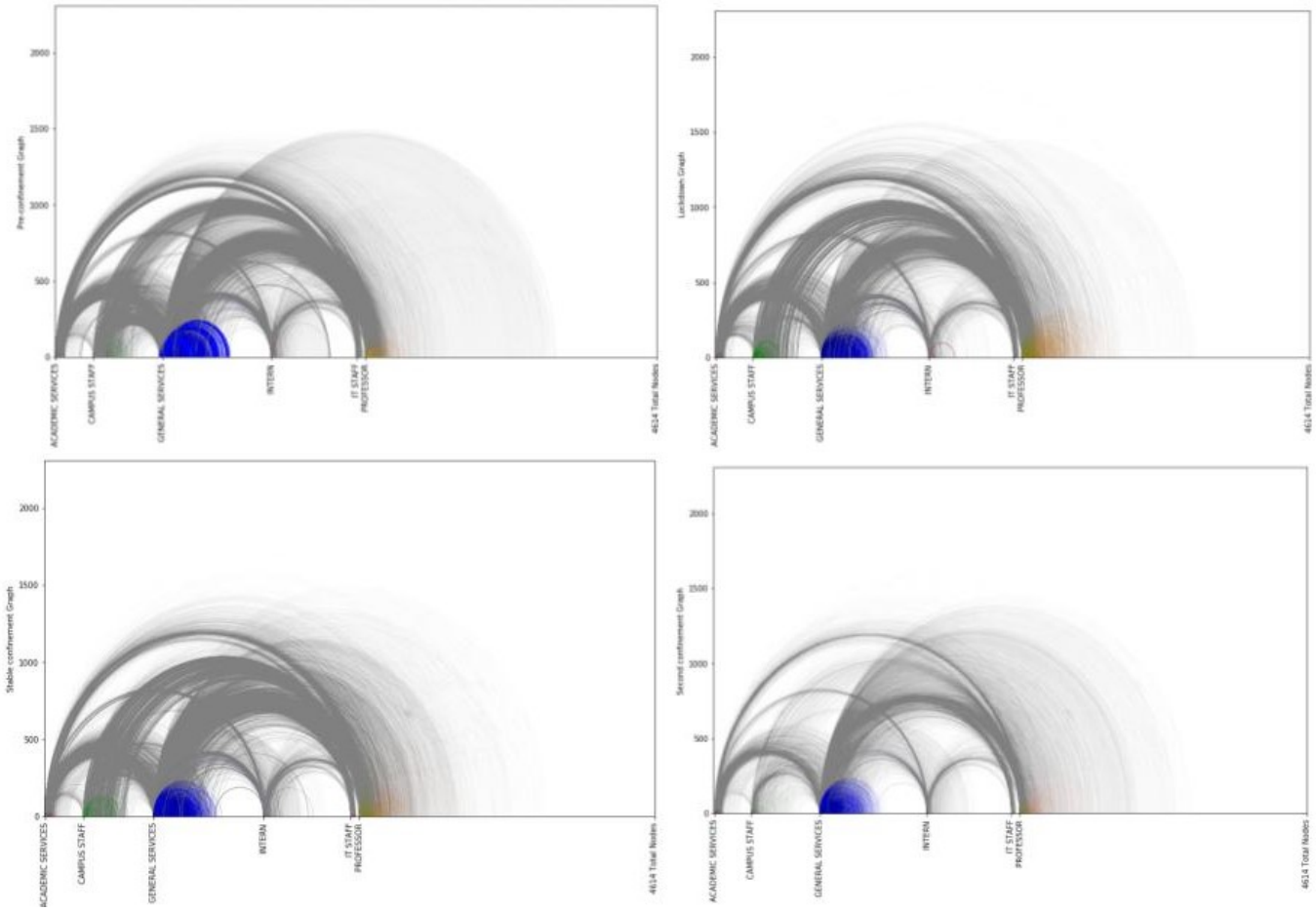


Figure 5.6: Representation of the 4 networks. Intra-group messages are marked in color

Several aforementioned hypothesis can be confirmed looking at the generated graphs. For example it is visible the augment on the assortativity for the *professor* group at the beginning of the first confinement, or the descent on the total number of messages sent during the second confinement, especially the interaction between *campus staff* and *professor*. This last fact may lead to an actual explanation on the fall in the number of messages sent, since the university is physically closed all campus-oriented activity is on hold.

In general we can see how the structure of the network remains the same. This indicates that people active through the year are actually the same, and that the main focus should be on edges and their characteristics over vertices.

5.3.1 Global network statistics

Global network statistics for each of the graphs are shown in Table 5.1. Data continue to confirm previous observations. We can see how the proportion of *nodes that send messages* versus *nodes that only receive* began at a 49.74% to increase to a 54.85% right after the first confinement. This confirms the assumption that people looked for other ways of communication to compensate the lockdown.

In this sense, it's interesting to see how considering a *weak type* relation, all the active nodes make up a connected graph. If the relation is strong, *Stable confinement* and *Second confinement* networks show a couple of nodes that only message between them, effectively making the graph unconnected.

	Pre-conf.	**Lockdown**	**Stable conf.**	**Second conf.**
Density	0.00197	0.00155	0.00201	0.00180
Avg. path length	3.034	3.241	3.294	3.282
Sending nodes	1548	1663	1608	1590
Receiving nodes	1564	1369	1497	1317
Total nodes	3112	3032	3105	2907
Global clustering coef.	0.03383	0.04753	0.03638	0.04388
Local avg. clustering coef.	0.2981	0.4844	0.3285	0.3804
3-cliques (triangles)	147135	79217	150421	128877
Avg. triangles per edge	95.64	51.49	97.78	83.77

Table 5.1: Network statistics

Also, the number of *3-cliques* (together with *local average clustering coefficient*) show how, when the lockdown initiated, the community reduced its interactions to the minimally necessary environment.

5.3.2 Degree distribution

Degree distribution is an important metric in the study of networks, as it defines which fraction of nodes in the network have a specific degree, as seen in Figure 5.7.

All four charts are represented in logarithmic scale, and as we can see the result is a straight descendant line. This indicates the function is ruled by a power law, which points to the fact that the network is *scale-free*. This kind of networks, typical among social networks, have several interesting characteristics such as:

- A strong robustness to failure. Major hubs are closely followed by smaller ones. These smaller hubs, in turn, are followed by other nodes with an even smaller degree. As a

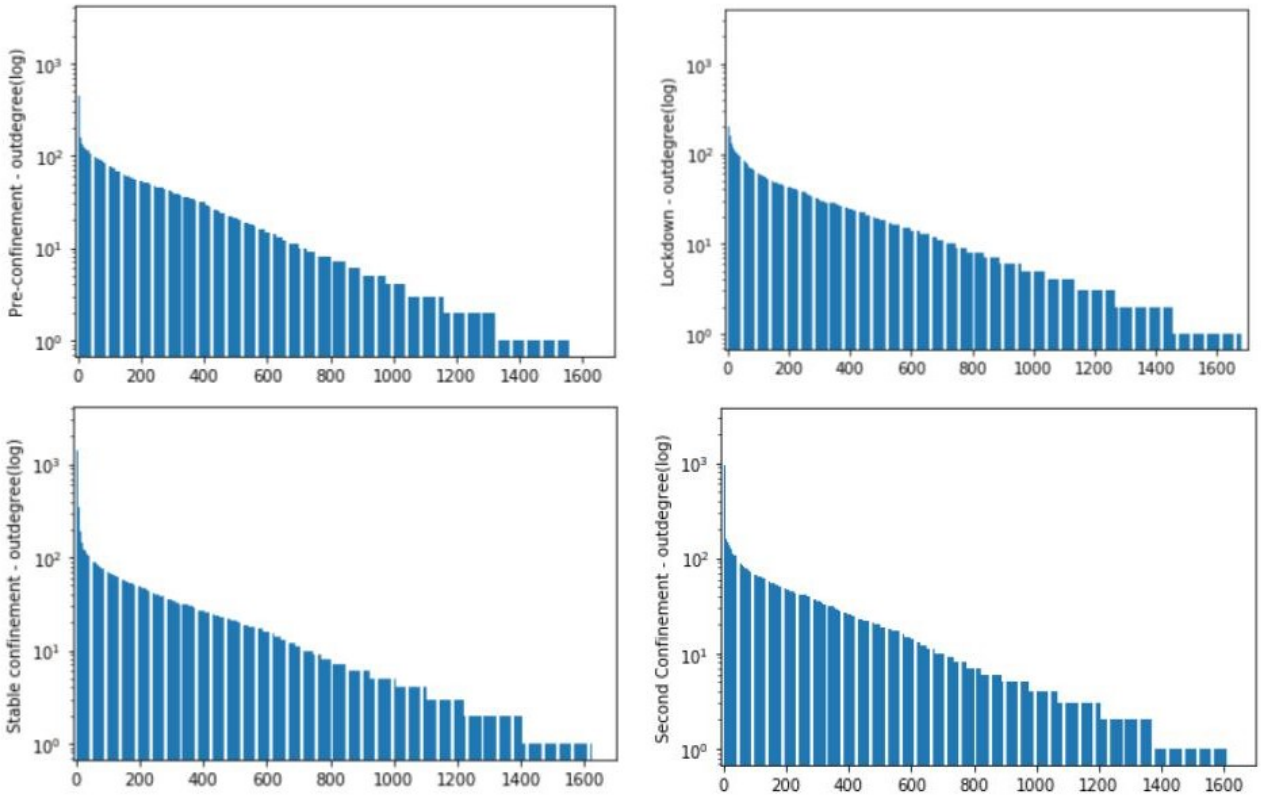


Figure 5.7: Degree distribution in logarithmic scale

result, the probability of failure of a high-degree node is really low, but in the case it fails, it may disconnect one or more sub-graphs.

- Clustering coefficient increases as the degree decreases. This implies that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through high-degree nodes.

- Immunization can be easily applied to the highest degree nodes -since there are few of them-, effectively reducing the spread of epidemics/rumors through the network.

5.3.3 Betweenness centrality

Betweenness centrality is a measure of the centrality a node has within a graph, and it is based on the number of shortest paths that go through it. The distribution of betweenness over the nodes is fairly similar through the 4 networks, as seen in Figure 5.8, giving a tipical logaritmic function for social networks, where degree is directly related to betweenness.

Nodes with a high betweenness centrality are considered critical in the robustness of the network. They usually represent actors that have a high level of organizing control over it.

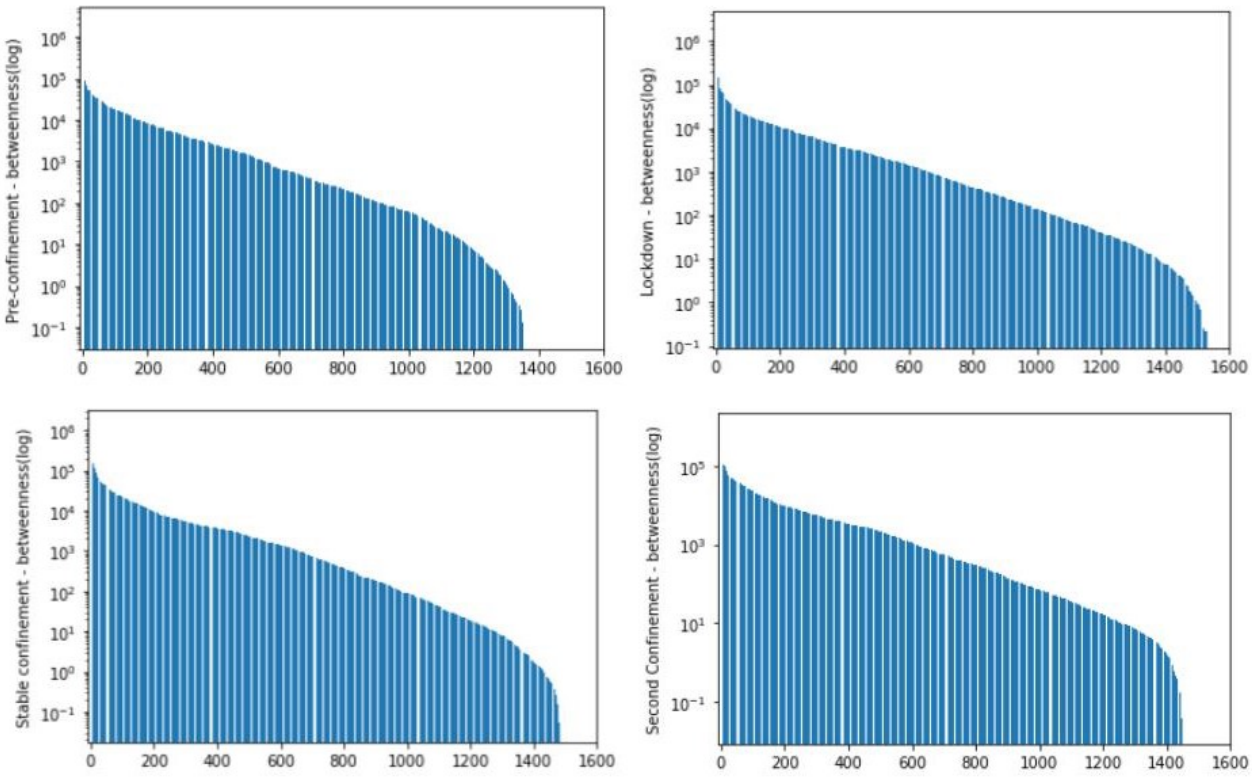


Figure 5.8: Betweenness centrality distribution logarithmic scale

Chapter 6

Network Dynamics

This section will conduct a study on rumor spreading over the 4 30-days period networks created in the previous chapter. This kind of dynamics return relevant insights on network topology, and open the way to the detection of interesting questions, as the existence of echo chambers or opinion leaders.

6.1 Spreading algorithm

The algorithm is defined by its initial parameters, as seen in Table 6.1.

Parameters	**Definition**
g - *Graph*	Representing communications during the corresponding 30-days period.
d - *Distance*	Intensity loss at every step.
p - *Probability of spreading*	Understood as the chance to spread the rumor for every communication
m - *Minimal implication*	Threshold below which there is no spreading.

Table 6.1: Parameter definitions

A first origin node -and its immediate neighbours- are randomly selected as the initial spreaders. While the algorithm ignores non-intervening nodes, and it is proven that *weak* relationships conform a connected graph, it is still possible to land in an *only-receiving* node (also known as *sink*). These first considerations are addressed at Algorithm 1.

Second part of the Algorithm 2 spreads the rumor from the core origin to its neighbours. Several factors are taken into account:

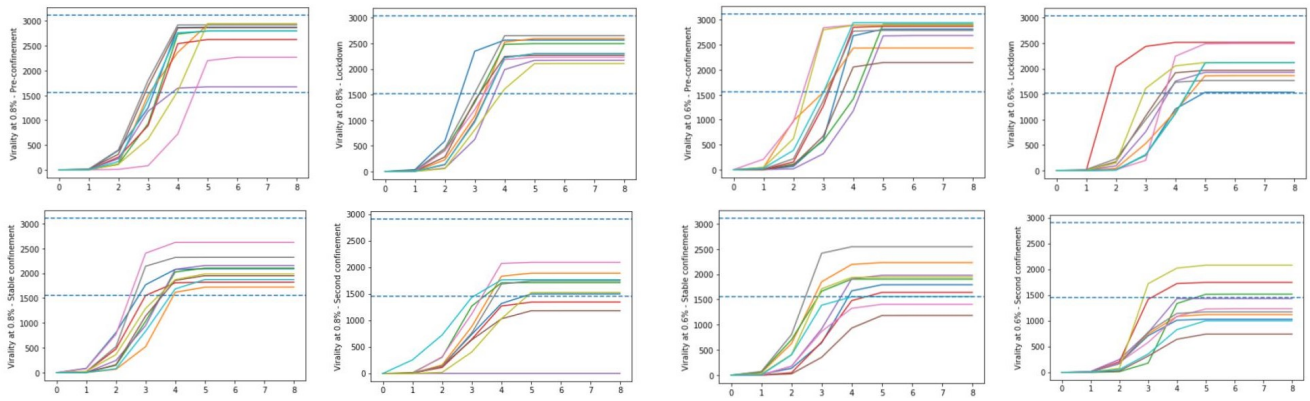
- How involved is the source node. Each node has a degree of implication in the rumor that goes from 0 to 1.

- Maximum implication degree is limited to 1, since it is possible to reach a higher level.
- If the implication degree is lower than the specified minimum, the node will not spread the rumor.
- Number of messages sent from source to target node.
- Probability p of spreading the rumor in each message.
- Distance to the origin of rumor.
- If the nodes have the same type, probability loss is not applied in order to reflect their closeness.
- If a node is the target of more than one spreader, its values are added.

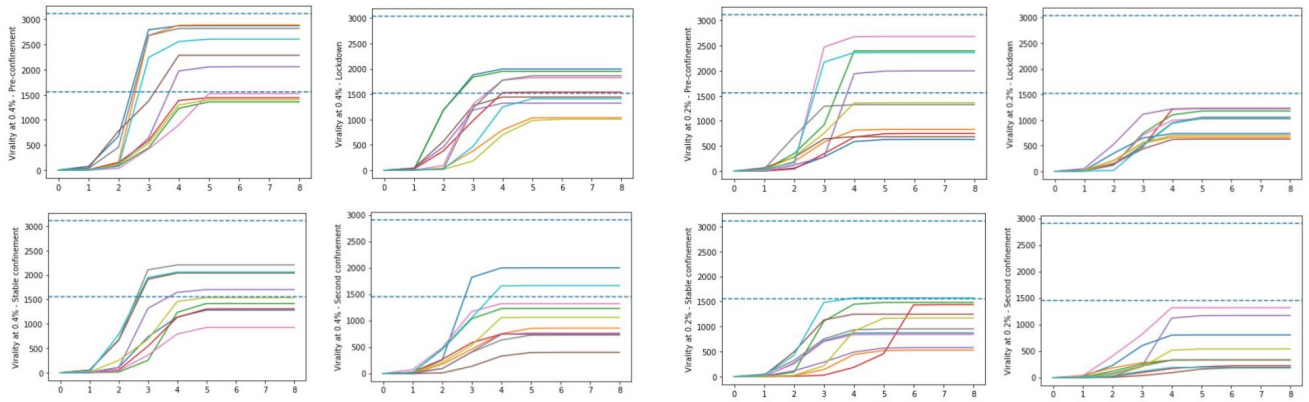
6.2 Virality

To carry out the study, loss due to distance d is set to 0.8, and the minimal threshold to spread m is set to 0.1, making the spreading probability p the one valor that can be tuned. In Figure 6.1 the algorithm is run ten times for every network, the result is how much the rumor spreads at every step until balance is reached. If at the last step the rumor surpasses the 50% mark it is considered successful.

The charts represent graphically how fast information travels through the networks. They show that under normal circumstances it takes not more than 5 steps from origin to reach a balance. Also, the general trends match the intuitive notion that the *global clustering coefficient* gave before, being the *Pre-confinement* network the less clustered, while *Lockdown* and *Second confinement* were the most clustered, thus indicating that it is harder for the rumor to reach a high percentage of nodes.



Virality with a spreading probability of 0.8 Virality with a spreading probability of 0.6



Virality with a spreading probability of 0.4 Virality with a spreading probability of 0.2

Figure 6.1: Viralities by graph and spreading probability

Algorithm 1 Pseudocode for the *spreading* algorithm (part 1, selection of the initiator node and its immediate neighborhood)

Input: Four parameters: g the original graph, d intensity loss at every step, p probability of spreading the rumor, m minimal level of implication required for a node to spread the rumor.

Output: Total level of implication for each step.

```
clu = g.clusters(mode =' WEAK')
na = []
for n in clu do
    if len(n) equals 1 then
        na.append(n[0])
    end if
end for
g.delete_vertices(na)
vm = np.zeros((len(g.vs), 9), dtype = np.float64)
ini = random.randrange(len(g.vs))
vm[ini, 0] = 1
re = []
nb = g.vs[ini].neighbors(mode =' OUT')
if len(nb) greater than 0 then
    for e in g.es.select(_source = ini) do
        g.delete_edges(e.index)
    end for
    for v in nb do
        if g.vs[ini]['Tipo'] equals g.vs[v.index]['Tipo'] then
            vm[v.index, 1] = 1
        else
            vm[v.index, 1] = 1 * p
        end if
    end for
else
    return [sum(x) for x in vm.T]
end if
```

Algorithm 2 Pseudocode for the *spreading* algorithm (part 2, all other steps until balance is reached)

```
re.append(1)
re.append(sum(vm[:, 1]) + 1)
i = 2
while (len(nb) greater than 0) and (d**i greater or equal than m)
and (sum(vm[:, i − 1]) greater than 0) do
    nbb = []
    for v in nb do
        nnb = g.vs[v.index].neighbors(mode =′ OUT′)
        for w in nnb do
            rat = g.es[g.get_eid(v.index, w.index, directed = True)]['Peso′] ∗ p
            g.delete_edges(g.es[g.get_eid(v.index, w.index, directed = True)].index)
            if rat greater or equal than m and vm[v.index, i − 1] greater or equal than
            m then
                if w not in nbb then
                    nbb.append(w)
                end if
                if g.vs[v.index]['Tipo′] equals g.vs[w.index]['Tipo′] then
                    vm[w.index, i] = (d ∗ ∗i) ∗ vm[v.index, i − 1] + max(vm[w.index, : i])
                else
                    vm[w.index, i] = rat ∗ (d ∗ ∗i) ∗ vm[v.index, i − 1] + max(vm[w.index, : i])
                end if
                if vm[w.index, i] > 1 then
                    vm[w.index, i] = 1
                end if
            end if
        end for
    end for
    nb = nbb
    re.append(sum([max(x) for x in (vm[:, : i + 1])]))
    i+ = 1
end while
for j in range(i, len(vm.T)) do
    re.append(sum([max(x) for x in (vm[:, : i + 1])]))
end for
return re
```

Chapter 7

Conclusions

2020 has been a year absolutely marked by the COVID-19 pandemic. Labor relations, like any other kind, have had to adapt to the new situation, navigating through all phases of the crisis in search of a *new normal* that can finally settle the disruptions created.

Along this project, it is possible to check how a specific, real, academic community has endured this process of rearranging through the first nine months of the year.

7.1 Technical conclusions

The treatment of networks and big amounts of data is a very interesting -and useful- branch of data mining. Since its associated optional courses can only be taken this semester -and I could not take them-, my intention was to improve my knowledge in the area while doing the project.

Concepts like *epidemic models* or *information cascades* were completely unknown me to at the beginning of the project. I find the study of graph theory, and how to extract information and classify networks, a centerpiece of what we currently understand as data mining.

The processing of big datasets and files with millions of rows have made me face challenges that I did not even considered as such at the beginning of the project. Issues like strict optimization of both code and data size to reduce processing time in a extremely effective way were totally new to me.

Also the library *igraph* for python has proven to be a most useful tool to any graph-oriented study. I consider having a good knowledge about it an advantage.

7.2 Personal conclusions

This has been my first contact with what could be considered a real, academical work. The amount of documentation to read and decisions to make can't be compared to any other practice

or exercise I have done before. The much greater scope on this project has forced me to plan the efforts and to calculate the total time invested in the project. In this area I have learned mostly from the mistakes made.

Also the guidance from my tutor has allowed me to have a look at how a professional data scientist actually develops his work on a day to day basis.

7.3 Future work

As reported in the data preprocessing chapter, there are some sections of the initial stage that finally have not been used in this project. Aspects like the addition of the binary flags attributes, or the incorporation of some of the supernodes to the network, can add several layers to the dimension of the study.

Also, the original attribute *subject* of the message -that was discarded at the first stage of the study-, can be an extremely interesting source of data if the goal is set towards the recognition and study of trending topics or fake news.

Other concepts that have been only superficially seen -and could be really interesting to study in depth-, are the detection of opinion leaders and how robust the network is to the loss of nodes.

Bibliography

[1] R. L. Ackoff. *Ackoff's Best.* John Wiley & Sons, New York, 1999.

[2] Enrica Amaturo and Biagio Aragona. Methods for big data in social sciences. *Mathematical Population Studies*, 26(2):65–68, 2019.

[3] N.T.J. Bailey. *The mathematical theory of infectious diseases and its applications.* Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[4] Javier Borge-Holthoefer, Raquel A. Baños, Sandra González-Bailón, and Yamir Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, 04 2013.

[5] J. Casas-Roma and C. Perez-Sola. *Análisis de datos de redes sociales.* Editorial UOC, Barcelona, 2016.

[6] Jana Diesner, Terrill L. Frantz, and Kathleen M. Carley. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Computational and Mathematical Organization Theory*, 11(3):201–228, October 2005.

[7] M. Granovetter. *Threshold models of collective behavior.* American journal of sociology, journals.uchicago.edu, 1978.

[8] Joel Grus. *Data Science from Scratch: First Principles with Python.* O'really Media, Inc., 2019.

[9] Paul W. Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 2(2):107–124, 1971.

[10] L Hughes. *Internet E-mail: Protocols, Standards and Implementation.* Artech House Publishers, 1998.

[11] Mahdi Jalili and Matjaž Perc. Information cascades in complex networks. *Journal of Complex Networks*, 5(5):665–693, 07 2017.

[12] Charles Kadushin. *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press, 2012.

[13] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. *Patterns of Cascading Behavior in Large Blog Graphs*, pages 551–556. Society for Industrial and Applied Mathematics, 2007.

[14] R.D. Luce and A.D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(1):95–116, 1949.

[15] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford; New York, 2010.

[16] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, Apr 2001.

[17] E.M. Rogers. *Diffusion of Innovations*. NY Free Press, New York, 2003.

[18] Jitesh Shetty and Jafar Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, page 74–81, New York, NY, USA, 2005. Association for Computing Machinery.

[19] P Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS ONE*, 13(9), 2018.

[20] S. Wasserman and K. Faust. *Structural Balance and Transitivity. In Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, Cambridge, 1994.

[21] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[22] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.