Universitat de les Illes Balears

Time Series Analysis of Online Social Media

Oriol Artime Vila

Master's Thesis

Master's degree in Physics of Complex Systems

at the

UNIVERSITAT DE LES ILLES BALEARS

Academic year 2013-2014

Date ______ *Author signature* ______

UIB Master's Thesis Supervisor: Jose Javier Ramasco *Supervisor signature* ______

UIB Master's Thesis Co-Supervisor: Maximino San Miguel *Co-Supervisor signature* ______

Accepted by the Director of the Master in Physics of Complex Systems *Signature* ______

Aquesta tesi no se la dedico a ningú en especial.
Com diuen desde montañas del sureste mexicano,
Para todos, todo.

Abstract

Last years have witnessed fast and fruitful advances in the knowledge of human dynamics. It has had two main and parallel contributions. On the one hand, theoretical modeling using tools from Statistical Physics have been important in an area of knowledge studied mostly by social scientists. On the other hand, with the development and improving of computers and their softwares, it has been possible to collect and to process big amounts of data, having empirical results to prove or to reject the existing theories, and even to find new and unexpected behaviors.

In this master thesis we analyse a database containing information of Twitter users. We focus our attention on inter-event times, this is, the time elapsed between two consecutive occurrences of the same event. These events are tweets holding the condition of replies, which is a way to interact directly with other users. We consider communication in Twitter as an example of a correspondence phenomenon, showing that it has strong temporal heterogeneities, with events clustered together in very small time windows followed by long periods of inactivity.

In this work we characterize the bursty communication pattern by studying the inter-event time distribution. We move on analysing correlations in the time series, finding that data are correlated to old times, beyond the circadian rhythms.

We also use this empirical inter-event time distributions as an interacting rules for the voter model, showing that correlations enhanced the time to reach consensus.

Acknowledgements

Tot i que el que hi ha escrit aquí és una tesi, un treball acadèmic, un munt de gent hi ha contribuït, d'una manera o altra, directa o indirectament. Així, els agraïments van en diverses direccions.

En primer lugar, por supuesto es el académico. Quisiera agradecer sinceramente a Jose y Maxi por supervisarme durante estos últimos meses de trabajo, por su tiempo y dedicación. Su guía y sus consejos han hecho de esta tesis, en conjunto, un tiempo totalmente disfrutable, en el cual he tenido la oportundiad de aprender un arsenal de cosas, herramientas y técnicas nuevas y útiles. Crec que l'Antònia també mereix la seva part de glòria i ser aquí, per donar-me un cop de mà en els meus primers (i no tant primers...) passos amb el MongoDB. Finally, to all my master colleagues, which we shared this last year helping each other in a healthy and selfless manner.

Secondly, I would also like to reserve some lines to people that, without being related directly to this work, they precisely helped and inspired me at some point while I was not working. In a personalized way, this paragraph goes for Alison, Danis, Dimitra and Vasso.

Per últim, un carinyós agraïment a la meva família i amics a l'altra banda del mar, per ser una xarxa de seguretat sempre que ho he necessitat. Per fer-ho tot més fàcil. Pel seu encoratjament i suport, com també l'ajuda mostrada al llarg d'aquests últims mesos i al llarg de tota la vida. I especialment al meu avi Santiago, que fa anys que diu "Això no ho veuré", però sí que ho fa. Sense ell, avui potser sí que escriuria aquesta tesi, però jo seria una persona diferent i ho faria en unes condicions totalment diferents.

Contents

Chapter 1

Introduction

It is old the interest of scientists and scholars to model society. Back in the *XVI*th and *XVII*th centuries, mechanistic theories of the universe appeared, with Galilei and Newton as prime examples. It was using this mechanical and reductionist point of view that philosophers and thinkers like Thomas Hobbes made the attempt to create the so called 'calculus of society' [1,2]. The goal was clear, although nowadays their methodology could be questionable, developing by rigorous logic and reason a science of human interactions, politics and society, building the whole theory from irreducible and evident axioms.

From there on, and for many years, sociologists, thinkers and political scientists have developed their theories trying to explain social phenomena that involve interactions. The approach was traditionally theoretical, with understandably small data sets to prove their statements and conclusions.

It is during the mid of *XIX*th century that what is known nowadays as Statistical Physics was born. Originally developed to study gases formed by billions of molecules, it explains macroscopic properties, like pressure or magnetism, from microscopic variables, like the speed of molecules or their spin. It is based on a clever usage of statistics, postulating that what matters when dealing with a large number of identical particles is not the detailed behavior of each one, but the average of their quantities, as well as the extent of deviation from these averages.

In the first half of the *XX*th century, both relativity and quantum mechanics concentrated lots of efforts from the physicist community. However, from 1950s onwards, with the creation of faster and more powerful computers allowing for big simulations, together with the development of renormalization techniques, the field of Statistical Physics acquired prominence again.

It is then when bridges began to be found among the concepts and tools taken from this theory, historically devoted mainly to atoms, molecules and spins, and the kind of problem the modelling of society presented. Phenomena as different as the modelling of road traffic, collective movement in schools of fishes or flocks of birds or opinion spreading have been able to be explained in the same framework and through the same underlying principles. See the review [3] for a clear presentation of these and more examples.

Among the myriad of problems and questions addressed by this new branch of Science, Sociophysics, in this master thesis we focus on a specific one. Taking advantage of capability and power of computers and softwares, we investigate a network of users of the well-known online social networking service Twitter[1].

Recent research regarding communication and interaction among individuals goes in several directions. One of them is the study of correspondence patterns and the underlying mechanisms that drive them. From surface mail of letters [4, 5], passing through e-mail communications [6], to mobile phone calls [7]. The main results is that the communication patterns are inhomogeneous, normally leading to heavy tail inter-event time[2] distributions [6], although other shapes for the distribution have been found [8]. A power law inter-event time distribution is the signature of a bursty dynamics, which can be explained through decision-based queuing process [6]. The concepts of burstiness in time series and inter-event time distributions play an important role in this work and they are explained in more detail at the end of this chapter.

Another direction has been to study online social media and web analysis. It has been proven as a good source to furnish information of working and leisure time habits, providing a better understanding of several aspects of human social behavior [9–11]. In these works, the authors characterize the users through their way of navigation on the Internet, analysing the role of their activity, i.e., how many times they perform an action in a given range of time. From there, is possible to find temporal probability functions that provide useful information: the inter-event time distribution tells us what is the probability of an event be repeated between two times or the waiting time distribution, which tells us the probability of the reaction to a given action.

Online social networks, due to the growth of the presence of Internet, both in developed and undeveloped countries, have become a more and more important source of social data. Regarding Twitter, which is the social network we study here, the conducted research has been focusing on different aspects: from mobility dynamics [12] or language analyses [13, 14] to forecasting popularity of news [15] or political polarization [16].

Usage of Twitter varies a lot from user to user. In [17] it is estimated that some use it for reading news (3.6 %), others for self-promotion (5.9 %), or even for spam (3.8 %). However, they find that the largest percentages of content of tweets are either conversational (37.6 %) or tweets that simply are informal discourses that do not cover any functional topics of conversation (40.1 %). Although it is a study conducted in 2009, we assume that it is still qualitatively valid, proving Twitter as a good platform to interact virtually.

1.1 Burstiness in human dynamics

Twitter is just an example of one of the large number of available online social networks. The main reasons to choose it among the others are that it has a really big network of users, so statistical analyses can be done safely, and because it is one of the few that freely offers a percentage of its data (around 1% of it). It is also particularly convenient because we can download and store it in a periodically automatized way.

In spite of the existence of a big variety of online social networks, they have a common feature. It has been found that the way that people use them and how they interact is concentrating lots of actions in a relatively small range of time, combined with long

[1] https://twitter.com/

[2] Along the chapters we use indistinctly either inter-event times or IET.

periods of inactivity. Roughly speaking, this is what we call burstiness.

Bursty dynamics in temporal processes has been found as a universal phenomenon displayed by lots of systems, either social or not. Poisson processes have been traditionally used to model systems related to human actions, such as modelling traffic flows patterns or accident frequency [18]. These processes are characterised by a probability rate parameter λ, that is nothing but the expected number of events that occur per unit time, and they are defined by a probability distribution

$$P(k) = \frac{e^{-\lambda t}\,(\lambda t)^k}{k!}, \tag{1.1}$$

where $k \in \mathbb{N}$ is the number of events within the time interval $[0, t)$.

In contrast to this widespread assumption of Poissonian statistics, there is increasing evidence that time series are not so well-modelled using Poisson processes. Heavy tails for different distributions are collected in many social systems, specially if human actions are involved. The physical interpretation of a Poisson distribution is that when an event occurs, the probability of the following one decays as an exponential function in time, due to the fast decay of the tail of the exponential. It means that two consecutive events are regularly distributed in time, excluding the chance to wait long times of inactivity. On the other hand, the signature of heavy tail distributions is a much more heterogeneous pattern: many of the events are separated by short times while a non-negligible amount of them can be separated by a very long period of inactivity.

Power-like decays can be explained as a consequence of a decision-based queuing process. Typical queuing processes assume lists of tasks, plus someone or something to execute them. The standard and most well-known processes are two:

- In the first-in-first-out (FIFO) selection rule protocol, the oldest tasks remaining in the lists are the ones that are going to be executed. The time an item stays on the list is simply the sum of the time needed to execute all the ones that arrived before. An example of a this kind of process process could be the queue formed when buying in the butcher's shop. They are used widely, for instance as accounting techniques used in managing inventory and financial matters involving the amount of money a company has tied up within an inventory of goods, feed stocks, etc.

- In the other extreme we have the random execution protocol, where tasks are randomly chosen. In this case, the time an item has been on the list is not important, they all have the same probability to be executed at each time step.

Both protocols presented above are characterized by exponential waiting time distributions, in disagreement with what is usually found in empirical data. Barabasi proposed in [6] that in order to achieve a long tail distribution, we can think in a priority-dependent execution protocol in the following way. Let an individual have a list composed by L tasks. A priority x_i, $i = 1, \ldots, L$, obtained from a certain distribution $\rho(x)$, is assigned to each task, so we can order them from the highest to the lowest priority. At each time step, the individual chooses the highest priority task and executes it, i.e., deletes it from the list. Automatically, a new task is added, with an assigned priority drawn from the same $\rho(x)$. Note that this model ignores the chance to choose an extremely low probability task, so in order to take into account some randomness in the execution procedure, a parameter

$p \in [0, 1]$ is introduced. So at the time of executing a task, with probability p the highest priority task will be chosen to be executed and with probability $1 - p$ a random selected task will be executed. Therefore, tuning a single parameter we can recover Poissonian behavior in the limit $p \to 0$ or power law behavior when $p \to 1$.

Indeed, numerical simulations show that exponential distribution and power-law distribution in the two different limiting cases. Moreover, it is shown that the tail of the distribution $P(\tau)$ is hold for $L \geq 2$, meaning that length of the list, which may vary from individual to individual, it does not change the dynamics.

1.2 Inter-event Times and Time Series

Time series are defined in a quite flexible way, with few restrictions. They are formed just by the collection of measures of a certain magnitude along time. Such a general requirement make them a very useful tool when studying areas as diverse as economics, weather forecasting or electroencephalography, among many other examples. The analysis of the time series is a discipline by its own, comprising methods for analysing the data in order to extract meaningful statistics, as well as, it provides methods to predict futures values from the observed ones. Several concepts, methodologies and tools have been proposed to study different properties [19], such as the autocorrelation function, usage of filters to remove noise, spectra analysis, to name few of them. Technical details are beyond the scope of this work, so we stay at a simple level, taking some concepts, nomenclature and methods, explained along the following chapters.

When studying discrete time series, concepts like waiting time arise naturally, as it is seen before with the queuing process. In the same way, another important concept that can be defined is the inter-event time τ, which is the time elapsed for an action to be repeated. In the case of the priority queue of the last section, τ could be the time between the execution of the same task, assuming that we can, somehow, tag what are the tasks and that once a task is executed, it can be reinserted in the list later. For example, imagine we had a task that was to reply an email. At time t_1 we do it and, following the model, it is removed from the list. At some point, due to the execution of another different task, replying an email will be again inserted in the list. If at time t_2 is when we reply an email again, then the inter-event time is just $\tau = t_2 - t_1$.

Since we are dealing with an online social network, we have to define what an inter-event time is in this context. In Twitter, users can send messages, the so-called tweets, and there are different kinds of tweets, depending on their content. A standard tweet is called a status update, and if it contains the expression "*username*", then the tweet is called a reply or a mention. Mentions is just a tag in the content of the message, just citing the username of the user, whereas replies are a way of directly notifying *username* that has a message for him/her. Replies are interpreted as a directed and personal message to someone, and they are the ones we use for our analyses.

We want to collect inter-event times, so the temporal processes we look for are the following ones. At time t_0, user Ann replies to user Bob $A \longrightarrow @B$. We wait until, at t_1, the event is repeated $A \longrightarrow @B$, so we say that the inter-event time is simply $\tau_1 = t_1 - t_0$. If they interact again at t_2, the next inter-event time would $\tau_2 = t_2 - t_1$, and so on. Because we look for interacting users, we restrict the tweets to those ones that

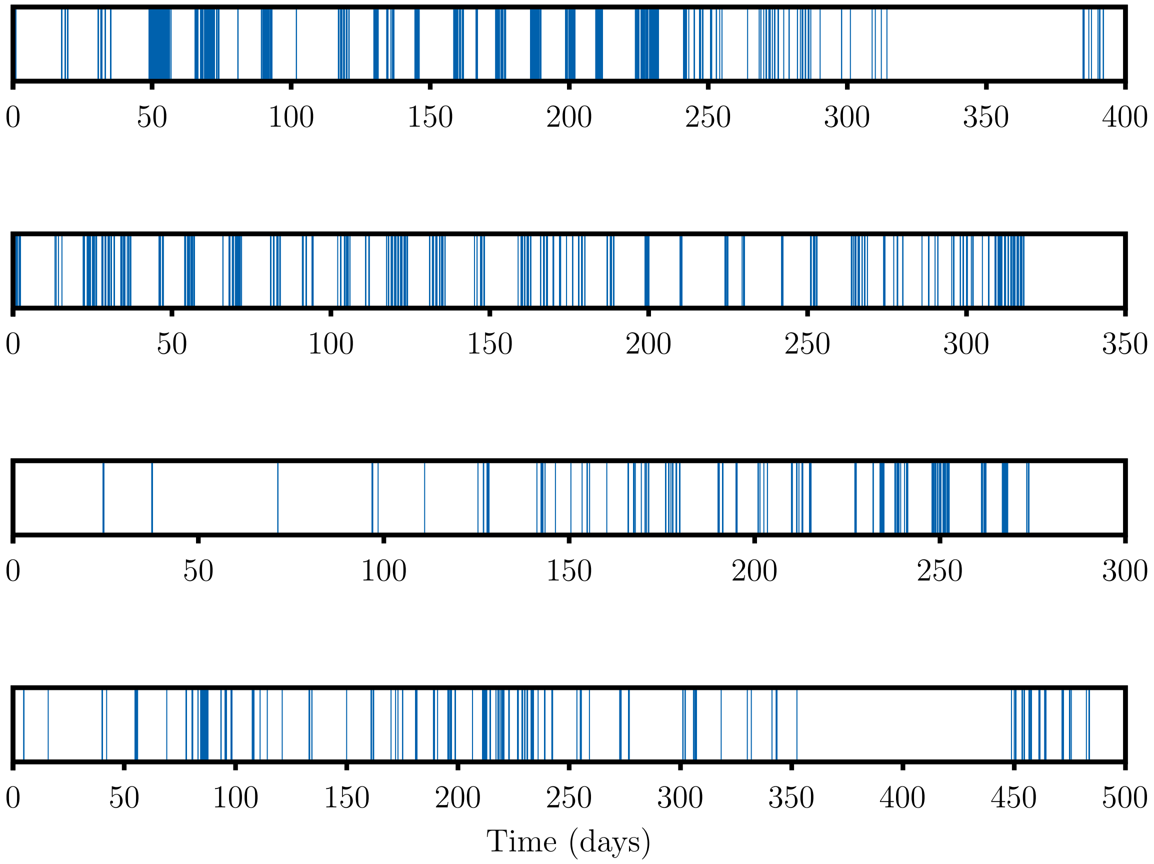


Figure 1.1: Activity patterns for four different pairs of users. Every vertical line correspond to a sent message, so inter-event times are the distance among two of them. The bursty behavior can be appreciated in the concentration of relatively narrow clusters of messages, separated by long periods of inactivity. The measure of time is relative to each pair of users, thus, $t = 0$ corresponds to their first interaction. In the figures are displayed 1914, 1198, 695 and 823 messages, respectively.

have the condition of replies, since they represent the most direct way of representing an interaction.

In Figure 1.1 we present 4 inter-event time patterns for different pairs of users. The IETs correspond to the horizontal distances between two vertical lines, that are the concrete moment when a message was sent. As can be appreciated, the activity does not follow a uniform tendency. It groups in clusters accounting for the burstiness explained before.

In order to convert the set of IETs in a discrete time series, we assign to each event an integer number, starting at 0. The vertical axis stands for the duration of the inter-event time τ. It is hold the relation $\tau_i = t_i - t_{i-1}$, for $i = 1, 2, \ldots$, where i corresponds to the event number and t_i is when the tweet i was sent. In Figure 1.2 we present the inter-event time series of the four pairs of users of Figure 1.1.

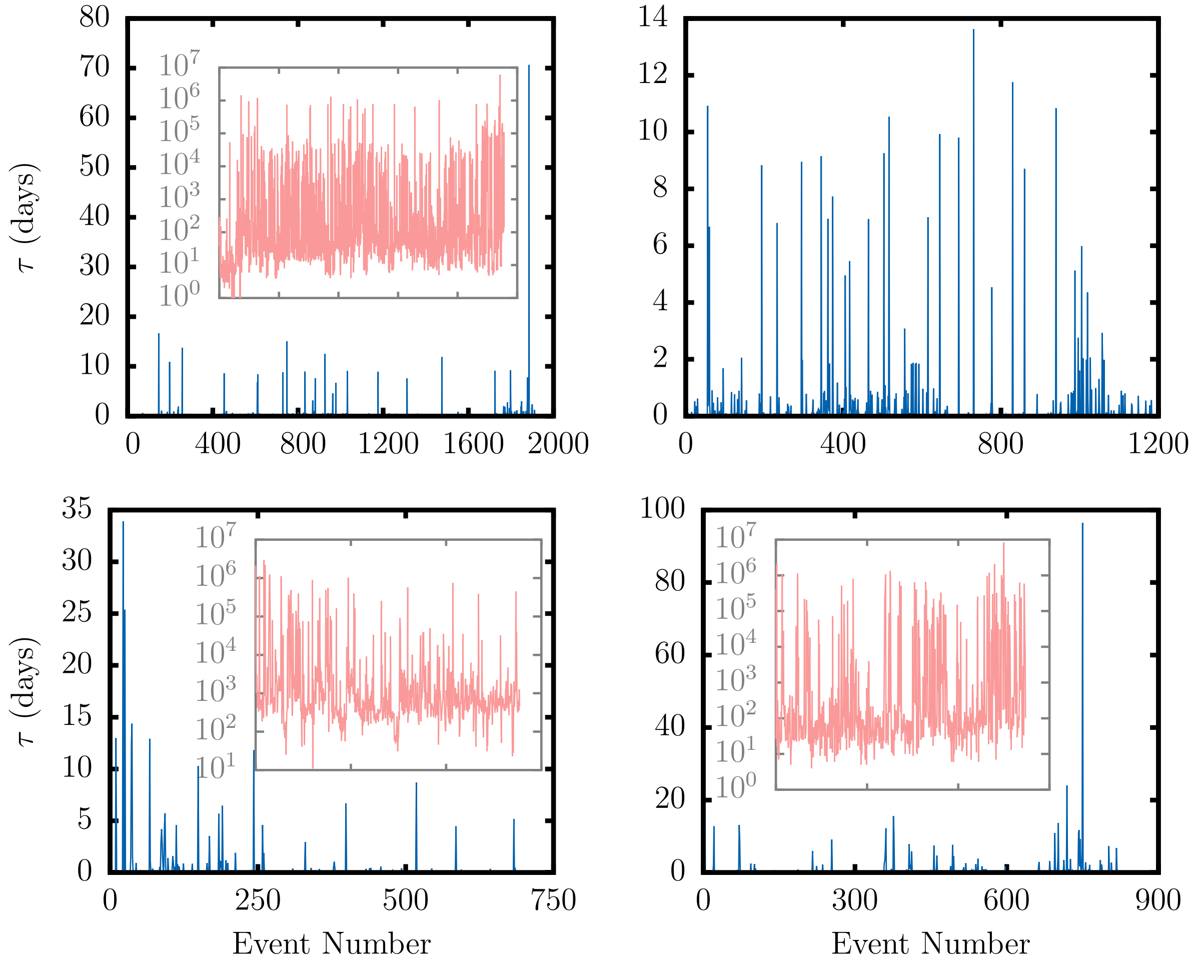


Figure 1.2: Time series corresponding to the users of 1.1. In the horizontal axis there is the counting of the event number. In the vertical axis, the value of the IET. Since the IETs span several orders of magnitude, the low ones are indistinguishable. We plot in the insets the same series series in logarithmic scale, to achieve a better appreciation of the shorts IETs. The unities of the insets are seconds and up to 7 orders of magnitude of variation are appreciated.

1.3 Organization and Goals

From the gathered data of Twitter, we are able to construct the network of relationship of the users. In the second chapter, we study this network as a static skeleton of interactions, without taking into account the temporal dimension. Links are present if users have ever interacted. By doing so, we are able to characterise some properties of the network.

We explained two different activity patterns, the Poissonian and the bursty ones, but we have not mentioned anything about issues such as memory or correlations of the series. Poisson processes are characterized by a constant probability rate of the events in a certain time interval dt, which is the signature of a memoryless series. This constant rate leads to an exponential decaying inter-event time distribution. However, the fact of having an IET distribution with a heavy tail means that we have lost the global constant rate for the intervals and it changes from one dt to another. In this case, we say that our series has memory.

Beyond these bursty IET distributions, another issue is the presence of correlations in the series. Whereas long-tail IETs distributions have been widely reported, the study of correlations and their impact in social time series is not so clear and well studied. These are the main question we address along chapter 3.

Once with the distributions obtained, in the chapter 4 we use them as an interaction rule in the voter model, showing how correlations affect the spreading of an opinion. Particularly, we focus on how much time is needed to reach consensus in a group of voters, comparing the cases of the standard, non-correlated and correlated updates.

Finally in chapter 5 we draw the conclusions of the work and briefly comment the expected further work to conduct.

Chapter 2

Network Analysis

2.1 Twitter and the data base

Twitter was created at 2006 and since then, it has not stopped growing. Due to mobile devices with Internet connection that turned out so popular in recent years, data generated from Twitter, and other online social media, has became a breeding ground to analyse and gain further insight into human behaviors.

Twitter is constantly generating vast amounts of data. Tweets can be downloaded from Twitter's API (Application Programming Interface), but technical difficulties arise at the time to store it, as well as, collecting tweets at real time is a non trivial task. The size matters too, yet each tweet is about 3 Kb, not a problem for few thousands of tweets, but easily it becomes difficult to handle as the number of tweets increases. We are then in the realm of Big Data. We deal with such quantity of data that it is difficult to process by means of traditional methods. To this end, we use MongoDB, a NoSQL software to work with data bases, specially useful when dealing with non-uniform data[1].

Since capabilities of databases are not infinite and with the aim of running the simulations in a sensible amount of time, we can not analyse all the generated data in Twitter. Instead, an algorithm is created to collect tweets under specific queries.

Twitter offers freely only the 1% of its total tweets, which we are constantly collecting. From them, we sift out the geolocalized tweets of the metropolitan areas of Barcelona, Zurich and London. Along a period of time, we took the most active users within these cities, which are the ones we follow and every two weeks, the database is updated with the last 200 tweets of the selected users. We have checked several times which are the most active geolocalized users in our cities, finding that most of them are always the same, so we do not include new users to the network and we keep gathering the tweets of the initial most active ones. The reason for choosing these cities and not other is simply because another project is running in parallel to this thesis involving these specific places. Note that the followed users were in these cities sending a geolocalized tweets at the moment of gathering, although this does not mean that they live in that cities (they could be tourists, for example) or that all their tweets are geolocalized (maybe they also use more frequently Twitter from their desktop computer).

[1] Twitter lacks data uniformity because from time to time its developers update it, so the inner syntax of tweets may change in these updates, becoming a problem in SQL data bases.

There is also a reason to deal with the geolocalized tweets only, instead of choosing random tweets coming from that 1%. Our main goal is to study interaction between users, thus, the probability of taking two random users that have been interacting between each other is much less than if they are selected within certain coordinates. By doing the collection of the data in the way we do, despite of tweets come from the 1% of the total, we achieve almost the 100% of the time line of the users we follow, besides ensuring a higher probability of interaction than in a random gathering. Thus, the less sparser the network of users is, the better for the analysis, in terms of computing time and accuracy of results. In Chapter 3 we discuss how this uncertainty in the completeness of the data base can affect the total results.

Taking into account what is explained above, the remaining data collection we work with consists of 215 962 496 tweets. However, since we are looking for inter-event time distribution, we have to add an extra query to select just the tweets holding the condition of replies. After filtering, there are 65 936 651 tweets left.

Instead of using a the whole data base of replies, we need to take a sub collection of it due to the time constrains at running our programs. Thus, the data base we work with corresponds to a subcollection of $6\,10^6$ randomly chosen reply tweets taken from the big collection. Moreover, in in this subcollection there are tweets that are repeated, i.e., tweets that were gathered more than once. After discard them we end up with 4 406 109 unique tweets.

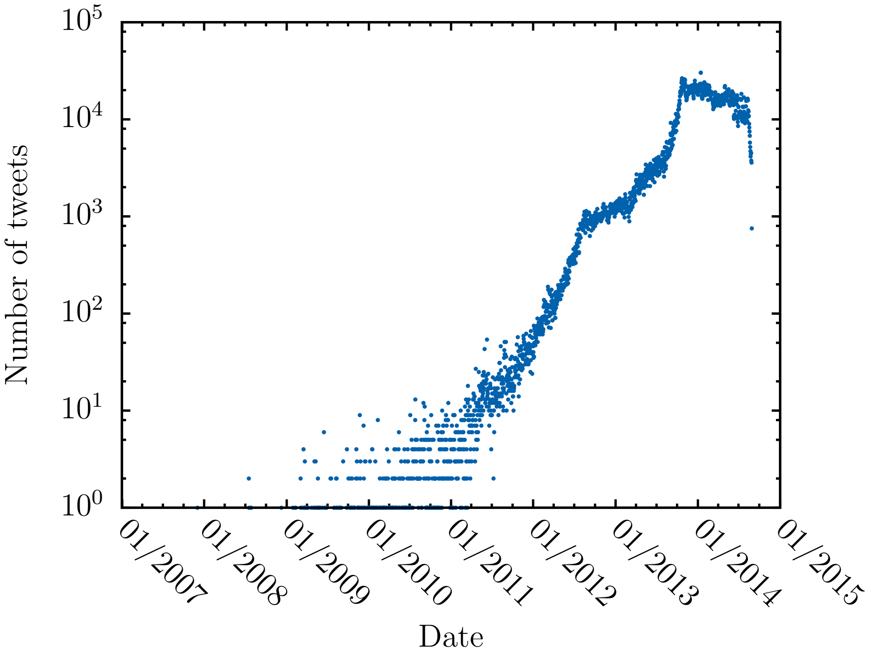


Figure 2.1: Number of tweets per day in the analysed data base. We can appreciate the variability of the data, note the log scale in the vertical axis.

The time range of the data collection, once filtered, comprehends tweets made from 1/12/2007 to 29/08/2014. In Figure 2.1 we present the number of tweets per day spanning the whole time range. Note that although the earliest and the latest tweets are separated by almost 7 years, they are not equally distributed in time, with the main contributions coming from mid of 2012. This inhomogeneity comes from the way we gathered tweets, the early ones corresponding to the last 200 of the first update of the data base.

Tweets stored in the data base come with a very detailed information. Below we present a random tweet from our data base with all the fields we have access to, in the JSON format.

```
{
        "_id" : ObjectId("52173b20fc56492cc673e3de"),
        "id" : NumberLong("353282484630327300"),
        "contributors" : null,
        "truncated" : false,
        "text" : "Gezellig avondje gehad, nu slapenn!",
        "in_reply_to_status_id" : null,
        "favorite_count" : 0,
        "source" : "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
        "retweeted" : false,
        "coordinates" : null,
        "entities" : {
                "symbols" : [ ],
                "user_mentions" : [ ],
                "hashtags" : [ ],
                "urls" : [ ]
        },
        "in_reply_to_screen_name" : null,
        "id_str" : "353282484630327300",
        "retweet_count" : 0,
        "in_reply_to_user_id" : null,
        "favorited" : false,
        "user" : {
                "follow_request_sent" : null,
                "profile_use_background_image" : true,
                "default_profile_image" : false,
                "id" : 269347942,
                "verified" : false,
                "profile_image_url_https" : "https://si0.twimg.com/profile_images/3724075001/99c54b3813810824addb883dfa43cd7f_normal.jpeg",
                "profile_sidebar_fill_color" : "C0DFEC",
                "profile_text_color" : "333333",
                "followers_count" : 101,
                "profile_sidebar_border_color" : "A8C7F7",
                "id_str" : "269347942",
                "profile_background_color" : "022330",
                "listed_count" : 0,
                "profile_background_image_url_https" : "https
```

```
                    ://si0.twimg.com/images/themes/theme15/bg.png",
                "utc_offset" : null,
                "statuses_count" : 2411,
                "description" : "Havo 4 sintjoris college / Waalre B1 / instagram: wouterbosman6",
                "friends_count" : 97,
                "location" : "",
                "profile_link_color" : "0084B4",
                "profile_image_url" : "http://
a0.twimg.com/profile_images/3724075001/99c54b3813810824addb883dfa43cd7f_normal.jpeg",
                "following" : null,
                "geo_enabled" : false,
                "profile_banner_url" : "https://pbs.twimg.com/profile_banners/269347942/1371993561",
                "profile_background_image_url" : "http://a0.twimg.com/images/themes/theme15/bg.png",
                "name" : "wouter bosman",
                "lang" : "en",
                "profile_background_tile" : false,
                "favourites_count" : 5,
                "screen_name" : "wouterbosman10",
                "notifications" : null,
                "url" : null,
                "created_at" : ISODate("2011-03-20T15:59:06Z"),
                "contributors_enabled" : false,
                "time_zone" : null,
                "protected" : false,
                "default_profile" : false,
                "is_translator" : false
        },
        "geo" : null,
        "in_reply_to_user_id_str" : null,
        "lang" : "nl",
        "created_at" : ISODate("2013-07-05T22:41:22Z"),
        "filter_level" : "medium",
        "in_reply_to_status_id_str" : null,
        "place" : null
}
```

We can access to everything related to the content of the tweet, as well as to the user account: the number of people the user is following, his/her profile image, the number of p that tweet has been marked as favourite, etc. Thus, we do not require most of the information; it is necessary to filter it, in order to keep only the fields which we are interested in. These are the time at which the tweets created[2], the identity of the user that sends the tweet and the identity of user that receives it. Moreover, as explained before, to ensure the direct interaction we look for, the reply field has to be activated

2.2 The network of users

We can construct an interaction network from the filtered tweets. Since we have both senders and receivers interacting through messages, the simplest kind of network representing them is the one where nodes are users. Edges between them are created if they have interacted. For the sake of simplicity, we stay at this simplified scheme since it is not a main goal of the work to study the network. Next steps could be to include a weight in the links, to account for the frequency of interactions, where the more two users have messaged them, the larger the weight is. Since the interactions are discrete and instantaneous in time, we could improve the description in a more advanced level by taking temporal effects [20], i.e., edges switching at the precise moment when the message is sent.

The network is formed by 358 273 different users and it has a total number of links of 542 435. Note that the number of links is bigger than the number of nodes, yet each node does not have lots of connections to the others, i.e., the network is sparse [3].

Since users come from different cities, nothing guarantees us that the network is composed by a single component. Indeed, in Figure 2.2 we show the number of components against their size. We have 15 755 non-connected subgraphs, on of them being a big cluster of around the half of the users.

Another interesting quantity of a network is the degree of its nodes, which corresponds to the number of neighbors a certain node has. There are three different kinds of degree:

- The undirected degree k: We measure it by considering an undirected network, where k is simply the number of links a user has with its neighbors, no matter who was the sender and who the receiver.

- The indegree k_{in}: To compute it we consider only as valid links the ones that correspond to tweets that are received from the others. In systems where ties are associated to some positive aspect, where most of Twitter users could be included, the indegree is interpreted as a form of popularity.

- The out-degree k_{out}: In this case we only take the links that represent sent tweets to other users. It is usually interpreted in social networks as a measure of gregariousness.

In Figure 2.3 we show the probability of the three kind of degrees for our network. The shape of $P(k)$ can tell us information of the kind of the network we are dealing with. For example, it is well-known that degree distributions for a Scale Free networks it is a

[2] Since the users can tweet world wide, it has been taken into account the time stamp from which was it sent.

[3] Note that the maximum value of possible links is $k_{max} = N(N-1)/2$, corresponding to a fully connected network.

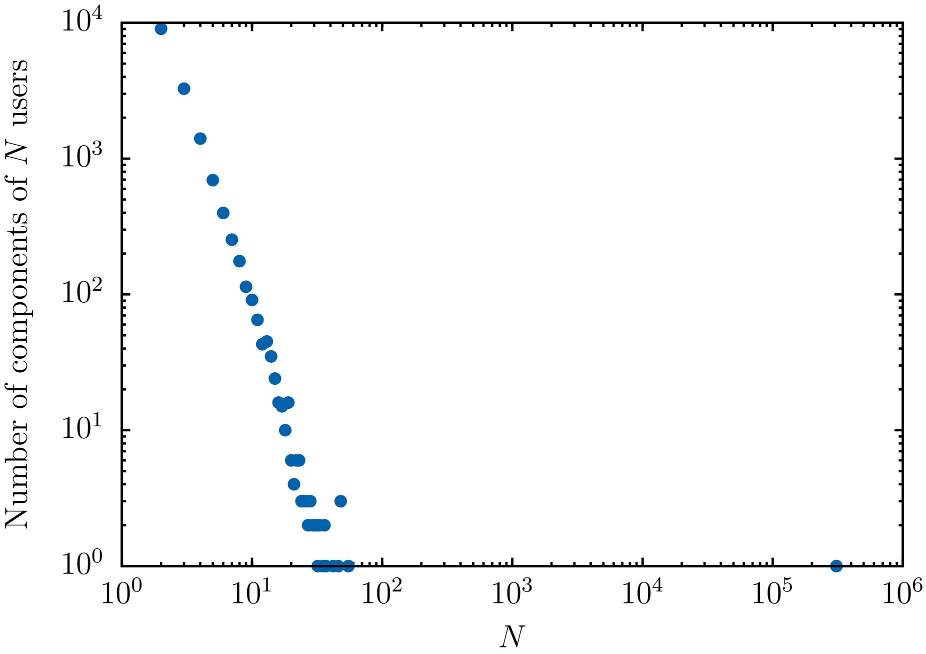


Figure 2.2: Number of non-connected components (subgraphs) of the total data.

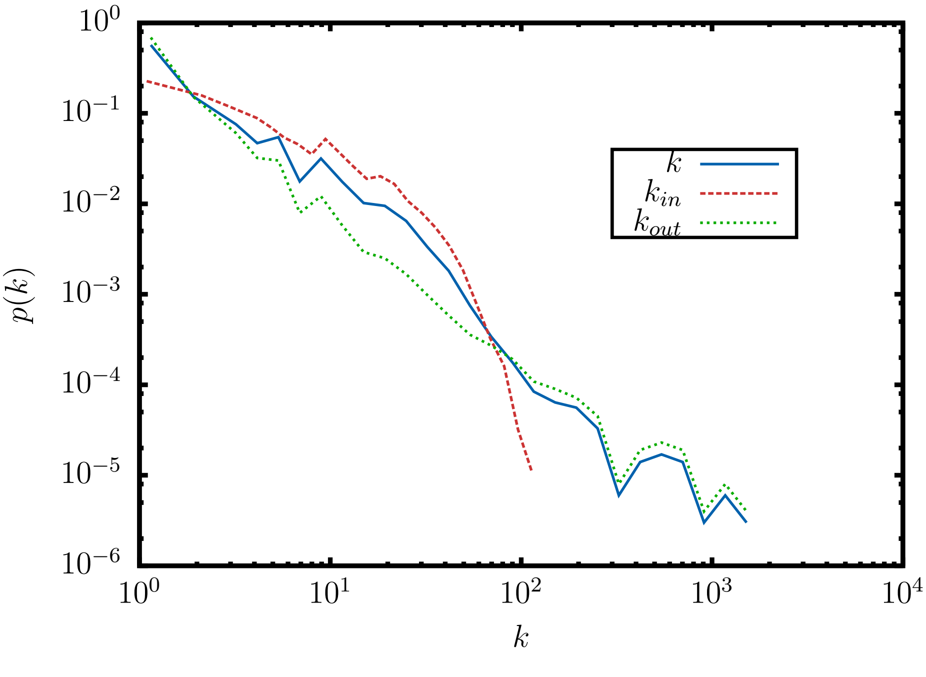


Figure 2.3: Degree distribution of the network of users, considering a link among two users if they have interacted. Scaling behavior is appreciated, which is a signature of scale-free networks

.

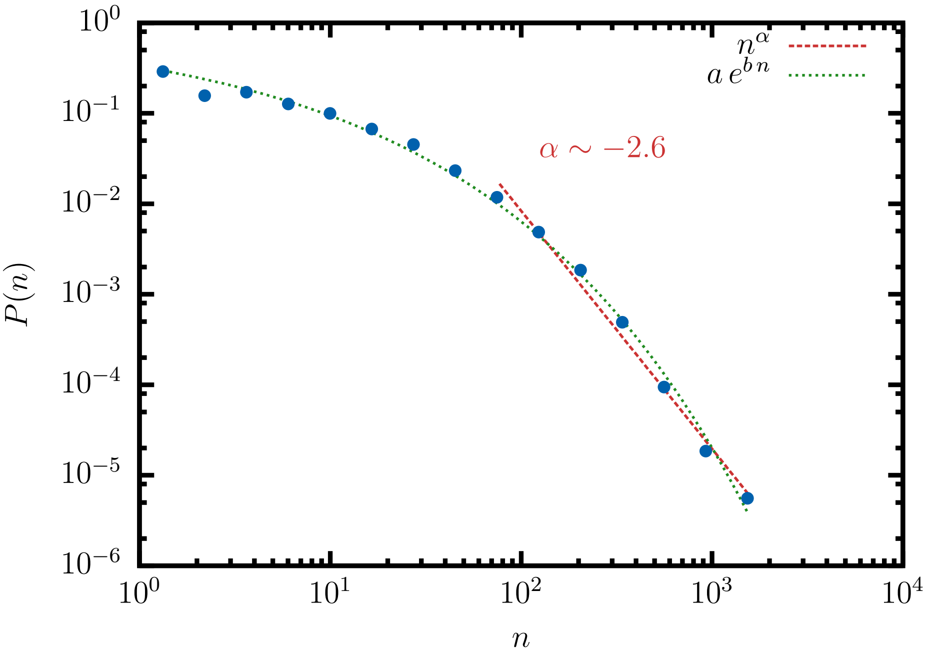


Figure 2.4: Probability of a pair of users to interact n times among them. The tail decreases power like with the number of performed operations n. As a help to the eye, we superimpose two fitted functions to the data, one for the tail of the distribution and one being an exponential fit in the whole range.

power law shape [21]. The degree distribution for the undirected degree shows a power like decay, with an exponent of $\alpha \sim 1.8$. It is a signature of a scale-free network, which it has associated a combination of few very well-linked nodes, the so-called hubs, with lots of barely connected nodes. Here we show, then, that our network of Twitter users interacting via replies seems to have a scale-free topology, with some deviations from the theoretical behavior.

Another quantity of interest when analysing social systems is the activity of the agents. Its is found that individuals frequently present very heterogeneous degrees of activity for the same task [11]. It may seem trivial to think that in a big group of individuals not every one have to be involved at the same pace or with the same intensity in doing a certain activity. Yet, it is not trivial to guess what is the distribution of this activity along the individuals and if it (if it does) affects the dynamics [22].

In Figure 2.4 we present the probability $P(n)$ of our data set, corresponding to the probability of having a number of performed operations n by a single user, i.e., the number of times he/she has interacted with another user via replies. As it can be seen, it is broad, standing for inhomogeneous activity from one user to another. Its tail decays powerlike $\sim n^\alpha$, with an exponent $\alpha \sim -2.6$. Nevertheless, this tendency does not hold for the whole range, since the data curve at low values of n, being a shifted power law or an exponential function better candidates to fit the range.

Chapter 3

Temporal analysis

Once we have constructed the static network of interactions, it is time to move forward and take into account the temporal dimension. In this chapter, we focus on the analysis inter-event times τ. To do so, we compute the time difference between replies of any pair of users. We obtain then the probability density function for the inter-event time distribution, finding a decreasing tendency, combined with circadian peaks. We also address out attention to correlation in the time series: how the duration of the current inter-event time is affected by the last ones. Specifically, we look first for the conditional probability for only one IET and then, we move on with the case for correlations at any IET.

3.1 Inter-event time distribution

Inter-event time distributions give a good first approach to human dynamics. Dealing with Twitter as a system of study, we consider as an event the reply of one user to another. Then, the inter-event time is the time elapsed between two of these replies. To compute the distribution we need to check how many times every pair of users of the network have interacted, tracking the time at which the messages were sent. With this information, we find the time differences between the closest ones in time and this corresponds to the duration of the inter-event time. Thus, the inter-event time distribution of a generic pair of users i and j gives us the probability of having two tweets with reply between them separated by a time τ.

In Figure 3.1 we show the inter-event time distribution $P(\tau)$ computed for all users. We encounter a pattern where maxima of probability are neatly repeated at the multiples of 24 hours, and lasts several months. As seen in the insets, there are also appreciated much smaller peaks between two maxima, corresponding to a 12 hours pattern. They are consequence of the so-called circadian rhythms, which are repeated oscillations of behaviors or phenomena along the time in humans, animals, plants, bacteria, etc, and their periods of the oscillations can be of diverse duration: daily, weekly, seasonal, annual,... Circadian rhythms are normally self-sustained, like sleeping cycles or tides. However, there are cases where there exist adjustment to external causes, for example to the daylight.

We plot in red circles the log-binned data, that although we lose resolution and we avoid fluctuations in the tail, it helps us to compute the exponent of the decay, as a

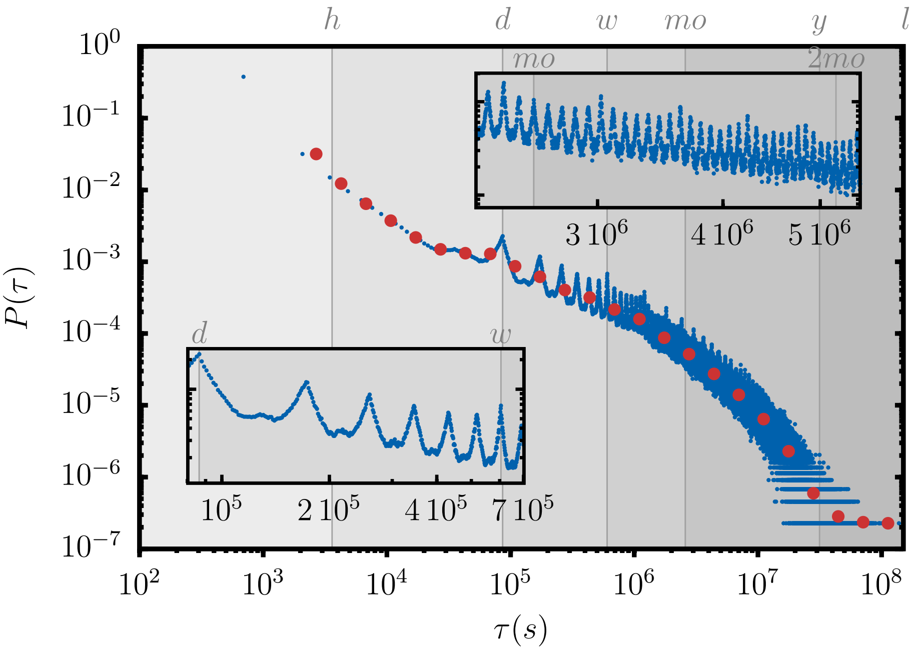


Figure 3.1: Inter-event time probability function. In the main plot is shown both the linear binned data, in blue, and the logarithmic binned data, in red. In the insets are displayed the circadian peaks along the first week and along the second month of interaction. The gray symbols above the plot stand for hour, day, week, month, year and lustum, respectively.

measure of characterization of the network. We address this issue at the end of this chapter.

We have obtained the inter-event time distribution, which gives information about the interaction patterns but it says nothing about correlations, on how the consecutive values are related between them.

3.2 Inter-event time correlations

In the study of temporal series, the next sensible step after studying inter-event times or waiting times distributions is to look for correlations in the data. This is, how do the current value of the series depend on the last ones ? While research on studying human time series has focused on issues such that the shapes of the distributions or trying to find relation between the individual and collective dynamics, we have many less examples of systematic analyses of correlations in human time series in the literature. A nice exception is [23], where they show that bursty time series are indeed correlated and that the autocorrelation function is not a good measure of these correlations.

In this section we study the possible correlations in our series of IETs. First we analyse the case of first-order correlation and then, we focus the attention on the higher-order ones.

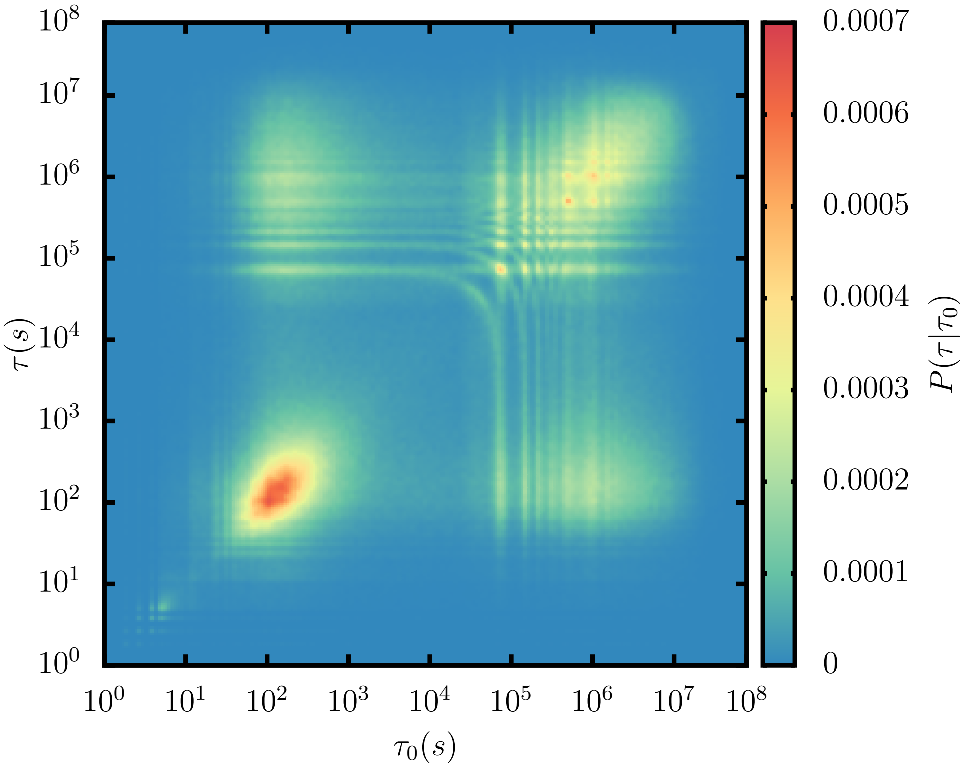


Figure 3.2: Heat map corresponding to the conditional probability $P(\tau|\tau_0)$.

3.2.1 Two inter-event times correlation

In this section we stay at Markovian level, i.e., the relation of just two consecutive IETs. Mathematically, this is done by computing the conditional probability $P(\tau|\tau_0)$, that stands for the probability of having an inter-event time of value τ if the last one was τ_0. By analysing all the consecutive IETs pairs (τ_0, τ), we are able to see if there exist correlations.

At first-order correlations, there are only two excluding possible scenarios: either there exist a dependence on τ_0 or not. By plotting the pairs (τ, τ_0) in a 2-dimensional plane, it is heuristically easy to see if the pairs are correlated. The signature of lack of correlation would be a non-dependence of the probability on τ_0, i.e., $P(\tau|\tau_0) = P(\tau)$ and we would see a scattered cloud of random points.

To proceed, we bin logarithmically the temporal axes, from 1 second to the maximum inter-event time. The number of bins is 150 per axis. Once all the values are assigned to a bin, we use a heat map, given by Figure 3.2, to represent the data. The colors are related to the probability $P(\tau|\tau_0)$ and its numerical correspondence is given in the tics of the color box.

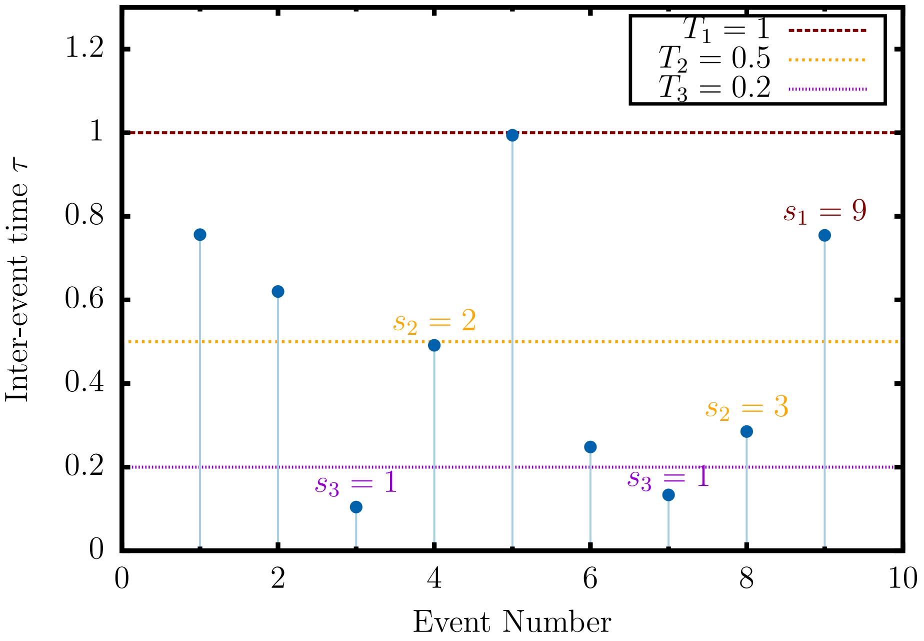


Figure 3.3: Toy time series for the explanation of how we count cluster sizes. We define a threshold and count how many IETs we have below that value, consecutively. On the top of the IET we display the value of the last cluster size.

Figure 3.2 strongly indicates the presence of correlations in the series of inter-event times. There are heterogeneities, and can be appreciated the existence of four "batches" of probability, separated by a cross of almost 0 probability. We see the tendency of the small IET to be followed by small ones, which is the highest probable area of the heat map. Also the large-large IET gathering looks more probable than the small-large and large-small ones. In the plot can be also appreciated horizontal and vertical lines corresponding to the circadian rhythms.

3.2.2 Going further than Markovian analysis

Once showed that there are correlations, at least, at first-order, we move on to study higher-order correlations.

To do so, we develop a method that benefits from the time series representation that we presented in Figure 1.2. We define several temporal thresholds, from which we compute how many consecutive IET in the series we have, lower than that threshold value. We present in Figure 3.3 a simplified example of how this counting works for a toy time series. In this case, we define 3 thresholds and, for each one, we count how many consecutive inter-event times are below the threshold value. For instance, for the threshold T_1, all the IETs are below it, so the size s of the cluster is as many points as we have in the series, in this case only one value, $s = 9$. For the second threshold T_2, we count again only the consecutive IETs below it, leading to two cluster sizes of $s = 2$ and $s = 3$. The same for the third threshold T_3, giving two cluster sizes of $s = 1$.

The expected result is that for high thresholds, there are few cluster sizes with large

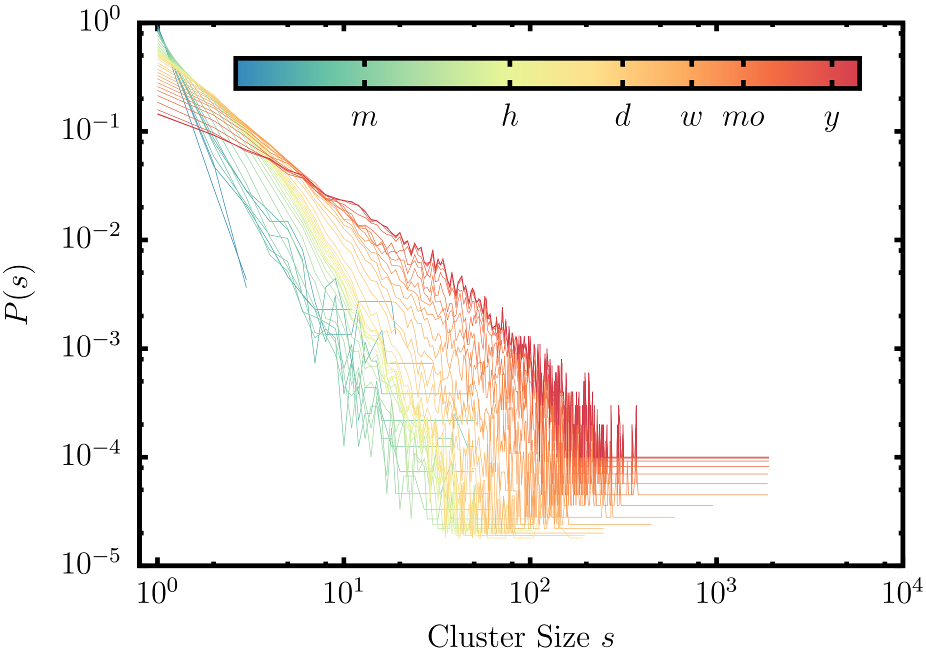


Figure 3.4: Probability of having a cluster of size s, in function of the value of the threshold, represented by the color of the line.

values, since most of the points fall below the threshold. On the contrary, for low values of the threshold, we expect to have lots of small cluster sizes, because most of the points are over the threshold, hence large s being unreachable.

In Figure 3.4 we compute the probability $P(s)$ of having a cluster of size s. We define 40 different thresholds, logarithmically spaced from $T_1 = 1s$ to $T_{40} \sim 10^8 s$, which approximately corresponds to the highest inter-event time. The value of the threshold is expressed by the color of the line, corresponding with the color box. As expected, it can be appreciated that the probability of having a low cluster size, when the threshold is small, is higher than if the threshold is large. At the same time, low threshold curves do not grow more than few unities of s. Note that there is a transition from low thresholds, where the probability of obtaining big clusters is strictly 0, to the region of large thresholds, where big clusters are of low probabilities but allowed.

What may be striking at first sight is that for large thresholds, the probability of finding small cluster sizes is low, but not too low. This contribution comes directly from the pair of users that only messaged each other very few times, so the maximum s is forced to be of the size of the number of interactions in that pair. In this way, users with low activity contribute to low s values, due to the limited number of times they interacted, not because the threshold.

Figure 3.4 shows what could be a scaling law. It seems that the time series of Twitter users displays correlations at different scales and that burstiness is a consequence of correlations beyond Markovianity. The presence of the long tails in $P(s)$ can be a signature of correlations to a big number of past events, i.e., to older and older times. If there is scaling it would mean that circadian rhythms do not affect the correlations, being the events correlated to longer times than the duration of the circadian rhythms.

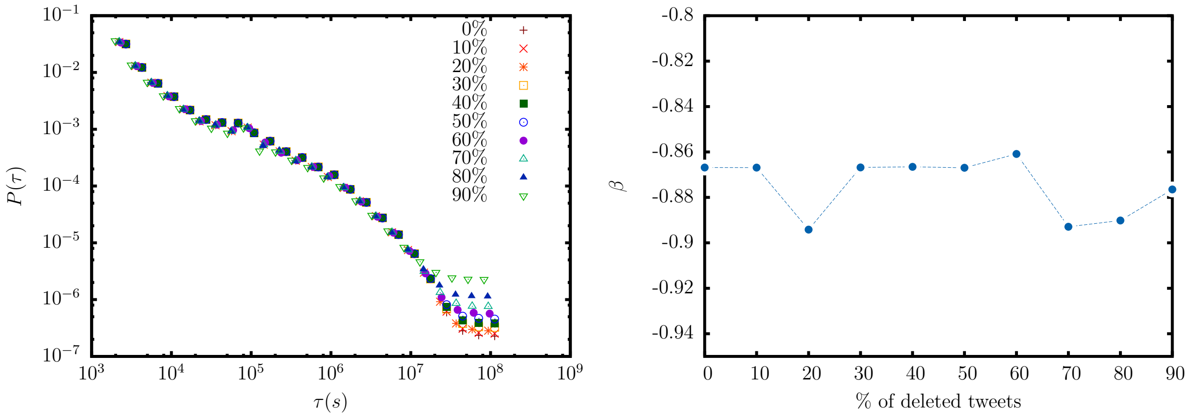


Figure 3.5: At the left-hand side, inter-event time distributions in logarithmic binning, with a percentage of deleted elements. At the right-hand side, the exponent of the decay, without taking into account the sporious flat tail.

Due to the limited amount of time, we did not have time to work deeply on the analysis of the correlations. The next step will be trying to find a scaling law for the different curves, fitting the exponent and gaining insight into the correlations across temporal scales.

3.3 On the completeness of the network

In Chapter 2 we presented the information of the network we are dealing with. We have over $6\,10^7$ replies, but we are working with a subset of it of about 10% of the total. Because simulating times are of the order of $\mathcal{O}(N^2)$, the time constrain does not allow us to investigate larger percentages of the network. The goal of this section is to see heuristically the effect of this incompleteness, by studying how the inter-event time distribution changes.

To this end, we randomly delete an increasing percentage of IETs of our series and we compare how their distribution and their decay exponent evolve with this removing. In Figure 3.5 we show the results. At the left-hand side there are plotted the distributions of the inter-event times, in logarithmic binning, of the removing percentages from 0% to 90%, by increments of 10%. It barely changes along the first decades and it is at the very end of the tail where there are discrepancies among the percentages curves. However, these differences are caused by the loss of total number of points at the moment of averaging at the binning, and not because the loss of points produces another type of functional behavior. Note we have a plateau that only changes its 'height'. At the right-hand side we present the change in the exponent of a fit of the form $y \sim \tau^{\beta}$ applied on the left distributions. As we increase the number of randomly erased tweets, it does not produce any appreciable tendency of change in β, it just fluctuates in a range smaller than 0.04 unities.

We remark that we are not trying to conclude that these distributions follow power laws, nor obtaining accurate fitting parameters for they decay. What we do is to find a parameter that can characterize the distribution, computed in the same way for all of them, and to test the robustness of the network to randomly vanishing of tweets. We believe that this heuristic evidence is sensible enough to affirm that the correlation analyses and the drawn conclusions in this chapter are valid for smaller sizes of the network, as well as, in an exercise of extrapolation, for bigger ones. In future work we want to develope the analysis with the complete network and see if it is still held this behavior.

Chapter 4

Modelling

The last two chapters have been based on an empirical approach: from a data base containing information about tweets, we computed static and temporal properties of the network of users. Possessing the IET distribution, we can insert it in dynamical models as an actual link activation distribution and compare if the standards results are modified. With this purpose, we choose the voter model and we modify its dynamics in order to include this empirical interacting rules.

4.1 The Voter Model as an ABM

The voter model belongs to the so called Agent-Based models (ABMs). Their main goal is to simulate the actions and interactions of elements of a system, called agents, which can be as diverse as ants, spins or humans. It is of crucial importance the concept of emergence, which represents the relation between the lower (micro) level, which is the scale at agents interact, with the higher (macro) level, where global complex behaviors arise, driven by simple rules.

ABMs are widespread and used in different disciplines, such as biology, economics and sociology. In this chapter we deal with the voter model, a model of social consensus. The agents, from here on called voters, are in any of two different states, that can represent an opinion on an issue. From this simple set up, several questions can be addressed. Will a network of N voters reach consensus if we let them to interact ? Or otherwise will a global state with several coexisting option prevail ? How do the topology of the network or the rules of interaction affect the final state ?

The model consists in the following. We construct a network, where each node corresponds a voter holding a given opinion, represented by either +1 or -1. Two voters connected by a link are called neighbors and interaction between them is allowed. At each simulation step, there is a state update, consisting in two ingredients: *1)* we pick a voter i with opinion x_i and *2)* we pick randomly a neighbor j with opinion x_j, from which i will imitate: $x_i \to x_i = x_j$.

We define an interface as a link that connects two nodes with different states. In order to characterize the order in the system, understood from a statistic mechanical point of view, we can introduce the average interface density

$$\rho = \frac{\text{Number of links between } +1 \text{ and } -1}{\text{Total number of links in the network}} = \frac{1}{\sum_{i=1}^{N} k_i}\left(\sum_{\langle ij\rangle}\frac{1-x_i x_j}{2}\right)$$
$$= \frac{1}{2N\langle k\rangle}\left(\sum_{i=1}^{N}\sum_{j\in neigh(i)}(1-x_i x_j)\right) \quad (4.1)$$

where $\langle k\rangle$ and $neigh(i)$ correspond to the average degree of the network and the set of neighbors of node i, respectively. In a consensus state, all voters end up with the same opinion, either $+1$ or -1, so the order parameter is 0 in that case.

4.2 Standard updates

Concerning the dynamical part of the model, an important property to consider is the kind of update we use. There are three standard updates and they all represent an interaction activity at a constant rate:

- Random asynchronous update (RAU): At each step, one voter is chosen randomly and is updated. In average, all voters are updated the same number of times. We define the time t as the number of steps over the total number of voters, N.
- Sequential asynchronous update (SAU): In this update, the voters are chosen following always the same sequence.
- Synchronous update (SU): Here, the update is made at the same time for every voter.

In Figure 4.1 we show the evolution of the order parameter of the equation (4.1) for different kinds of networks: fully-connected, random and scale-free (to know how to build them and some of their properties, see the review [21]). Consensus is always present, no matter what network or what update we are dealing with. However, as we increase the size of the network, the time to reach it grows with N. The system is constantly fluctuating, but because is finite, at some point one fluctuation is big enough to create a domain that conquer all the other ones. In the limit $N \to \infty$, we would not have consensus, we say, then, that the ordering is achieved due to finite size effects.

4.3 Inhomogeneous update

There is a need to go beyond the constant rate update, since heterogeneities produce qualitatively changes in the results [24], and they account for more realistic interactions than the standard updates.

Here we introduce a model that accounts for temporal inhomogeneities when voters copy each other, meaning that their copy rate is not constant. In chapter 3 we compute inter-event times, so as an equivalence to inter-event times in Twitter's case, we consider

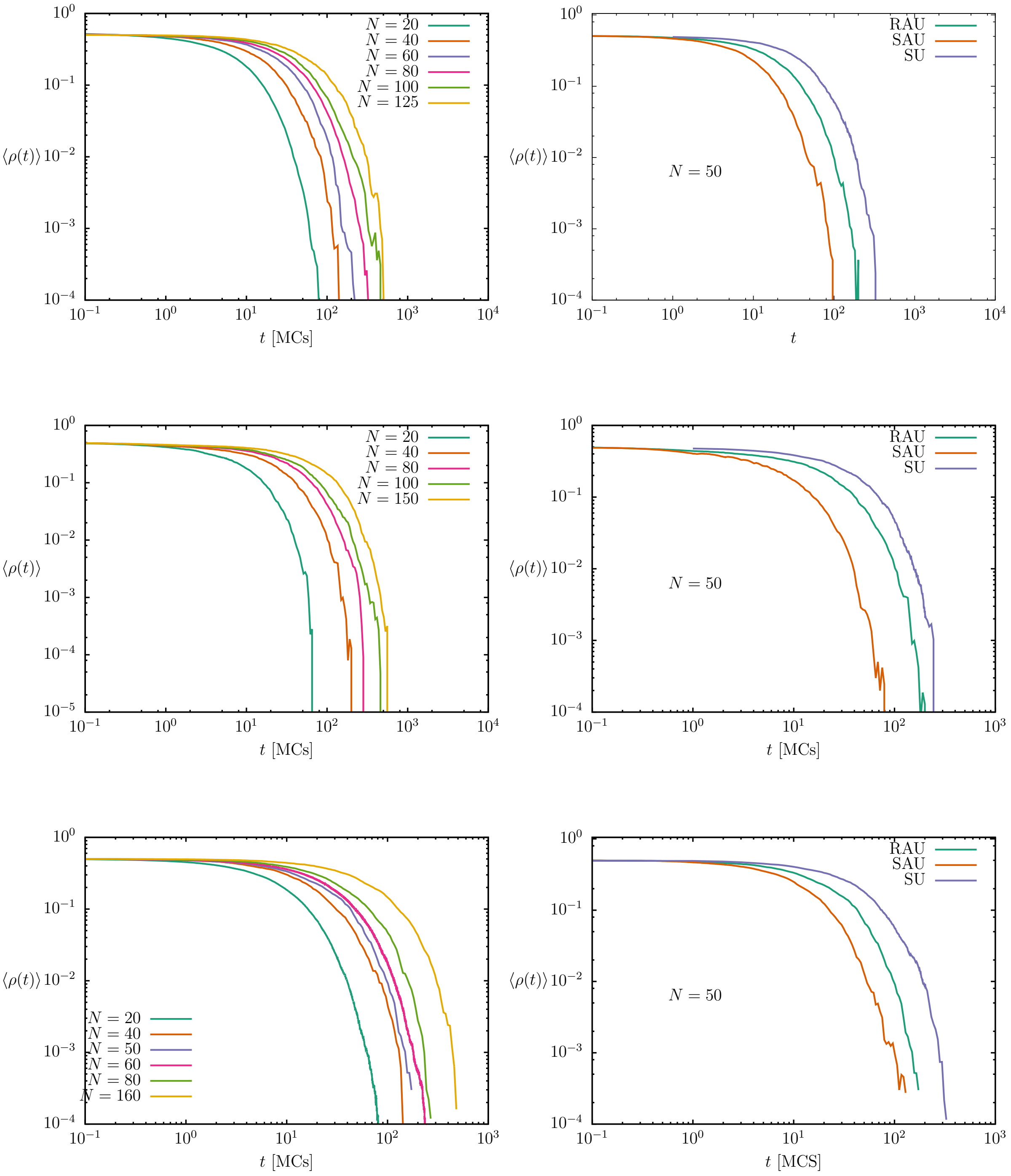


Figure 4.1: At the left hand-side column, the evolution of the order parameter for different system sizes. At the right hand-side column, the order parameter dependence on the kind of update. The topologies are random, scale-free and fully-connected networks, from above to below.

the time elapsed from one copy attempt to another as the IET. By defining the inter-event like this, we can plug our distribution of inter-event times into the model.

We set the network topology and the initial distribution of opinions, as well as, we also set an initial temporal distributions of the interactions. We give at each link a time of activation τ_l, with $l = 1, 2, ..., N_l$ where N_l is the total number of links. These τ's are drawn from the empirical IET distribution found in chapter 3. We let the system evolve, and when a global clock common for all the links is equal to some τ, then voters of that link interact, one copying to the other. We consider an undirected network, and since the IET is assigned to the links, there is no preferred direction to copy, so we choose randomly one voter to copy the other. Then, we draw another IET for this link and the system evolves until reaches the nearest IET, where the copying procedure is done again. Schematically, these are the steps:

1. We let the system evolve, with a global clock common for all voters. When this clock is equal to one of the times on the links, the link is activated and then one voter copies another's opinion.

2. A new inter-event time is drawn from the distribution and associated to the used link.

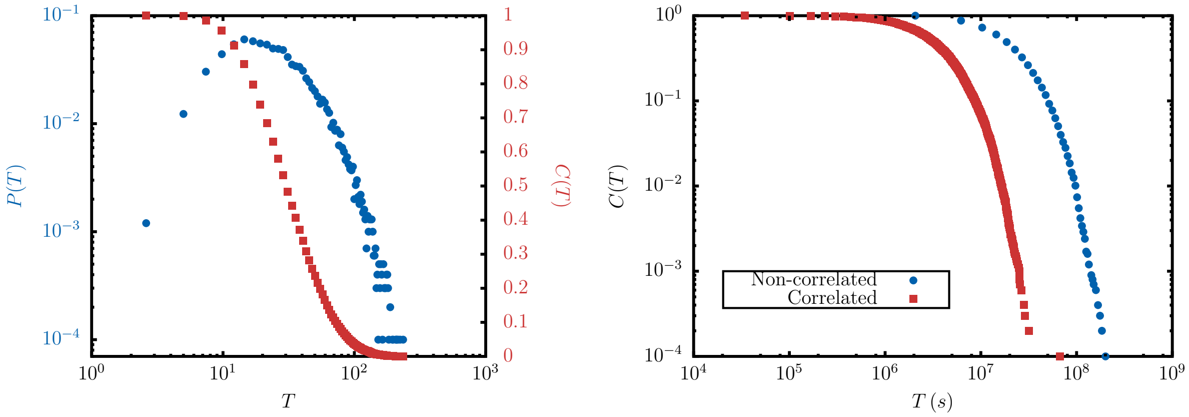
3. Repeat ad infinitum.

Note that in standard updates, at each Monte Carlo step we ensured a copy event, so averaging over samples is trivial. This is not the case here, since updates are produced at very different times, because from one sample to another the copy events are performed at different times of the global clock, since they are drawn from an IET distribution that comprehends several orders of magnitude. The drawback of the model, therefore, is that we can not perform averages easily, but, on the contrary, it is run on real time values, so it can predict how much time take the network of voters to reach the consensus. Thus, instead of characterizing the system with the order parameter $\rho(t)$, we study how much time the system needs to end up with a single opinion.

We have seen in Figure 2.3 the similarity in the degree between the network of our Twitter users and the scale-free network. Since we are taking the empirical IET distributions and using them in the dynamics, we choose a scale-free network of voters as topology for the simulations.

We present in Figure 4.2 two plots comparing the standard update with our model. The studied quantity is the time T needed to reach equilibrium. At the left-hand side figure we show both $P(T)$ and $C(T)$ which are the probability and the cumulative function, respectively. At the right-hand side of the Figure we show our dynamics implemented with the IET distributions of Chapter 3. We simulate first the dynamics without including correlations, where we just take a new activation time for the link from the distribution and let the system evolve. We also run the simulations allowing for correlations, using the 2 dimensional distribution of Figure 3.2.

The results for the standard update are the expected, since it can be appreciated a log-normal shape of T. This means the existence of a most probable time T^* with a non-symmetric distribution of times at lower and larger values than it. For lower values than T^* the probability changes abruptly and the minimum T is much closer to T^* than



(a) (b)

Figure 4.2: **Figure(a)**: In blue, corresponding to the left axis, the distribution of times to reach consensus in the case of standard update (simulated using RAU). A lognormal shape is appreciated. In red, corresponding to the right axis, the cumulative distributuion function. **Figure(b)**: Cumulative distribition functions for simulations using the empirical distribution of inter-event times, with and without correlations. There are 10 000 independent runs with a scale-free network of $N = 50$ voters and an attachment parameter of $m = 5$ (number of links a new node attaches to the already existing ones when it arrives to the network).

the maximum one. The exponential tail corresponds to the increasingly less and less probable times to reach consensus, away from T^*.

The results for our model, in comparison to the standard update one is that the shape of the distribution of $P(T)$ is exponential. The orders of magnitude of T can not be compared, because one is related to the MCS (at every update, the *time* increases a factor $1/N$), and in ours, it stands for the real time, in second, a group of N voters, interacting in the way explained above, would take to reach a consensus. It is worthy to note the role of correlations, which leads the system to reach faster consensuses than the non-correlated case. Recall that in the correlations we obtained, the most probable behavior is a short-short and fast-fast IET relation. This induces that two voters that have interacted are more probable to interact again sooner than in the case of the non-correlated IET distributions.

We see that the case of correlated in the case of IET the systems arrives at consesus faster than the non-correlated case. This is because it does not matter that two voters holding different opinions are connected by a link that will be activated in a long time, because the opinion may be spread from one to the other through a longer[1] path of bursty links. Thus, we show that correlations in the activation times of the links enhance the spreading of opinions.

A consequence the existence of these highways of bursty links connecting the network is that the more well-connected is a network, the faster the consensus is reached. To check this statement we run the model with the correlated IETs for several attachment parameters m of the scale-free network. We see in Figure 4.3 that, indeed, this behavior

[1] Long in the sense that it passes through more nodes to connect them.

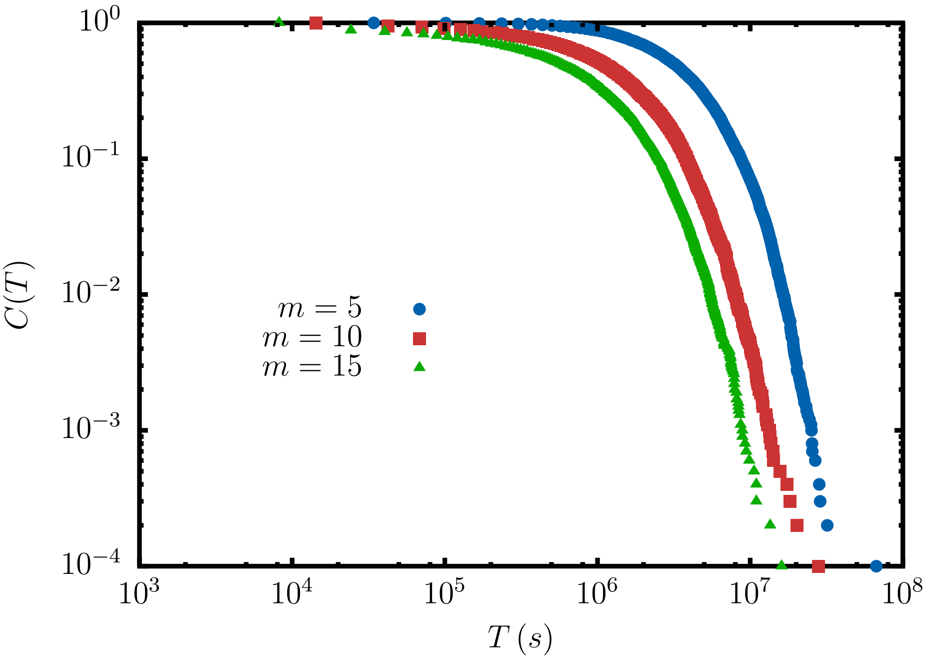


Figure 4.3: Cumulative density function for different parameters of the scaling free network.

is hold.

Chapter 5

Conclusions

People within society inevitably interact: helping and cooperating, fighting and conflicting, speaking and discussing to reach consensuses, following fashions,... Questions that Sociophysics addresses are whether we are able to find and explain mechanisms that make the behavioral patterns emerge and whether we can make some predictions on them. This is the main goal of this work, focusing on the interactions of users in an online social network.

In this work, we have studied several aspects of a network of Twitter users. We took a look firstly to the static network, i.e., the skeleton of the interactions, where we have not taken into account the time. We have been able to characterize the network by studying some of the properties of the nodes. Thus, we have seen that our tweets range a time interval close to 7 years, that the total network is non-connected, having only one big cluster and lots of contributions coming from isolated interactions involving few nodes. When studying the degree distribution, it seems that it has a long tail, allowing for the existence of very few well-connected nodes and for a large number of poorly-connected nodes. It is the scale-free signature, showing that our network of users holds it. We also see that users are heterogeneous in their activity, with a broad distribution with a long tail.

Moving forward to analyse inter-event time distributions, we have obtained two time scales that play a role. The short time scale, where circadian rhythms are present, can be appreciated for some months. At peaks multiple of $24h$ the users are more likely to interact than at any other time. For the first days there is also possible to detect little increments of probability located perfectly in the middle of these $24h$, accounting for $12h$ rhythms. The long time scale shows a decay of several orders of magnitude of IETs in the distribution, accounting for the burstiness of the series: lots of messages are separated by small times, while from time to time we find a long period of inactivity.

Long tails in human time series are well-known and frequently reported, but in literature is not usual the systematic study of their correlation. We addressed this question by studying correlations both at two inter-event times and at more-than-two inter-event times. It is found that probabilistically, consecutive IETs cluster in their size: low-low and large-large correlation is found. To go further than Markovian analysis we have developed a method to study higher order correlations, based on the study how inter-event times cluster together below a certain threshold. We have obtained that, indeed, there exist correlations much beyond Markovianity, meaning that correlations span for large times, larger than circadian rhythms.

In order to use the obtained distributions, we studied the voter model with an inhomogeneous update. We have presented a model where two nodes were interacting, i.e., copying other's opinion, only at a time drawn from the empirical IET distribution. Due to the time heterogeneity, instead of analysing the system with its order parameter, we have used the time needed to reach equilibrium to this end. We have simulated the model with distributions with and without correlations. The main difference between the standard update and the empirical distributions is that for the first, there is a log-normal distribution of times to reach consensus, while with the empirical ones, the shape is exponential. Besides, we have seen that consensus is reached faster with the correlated IETs than with the non-correlated ones, as a consequence of the highways of bursty links that can connect any two nodes, maybe not in the shortest way in terms of space, but with rapid spreading of information thanks to correlations.

Time is always worth but also limited, and when working with a non-very-flexible deadline, the time constrain seems to be heavier than normally. We have had to avoid to delve in some parts and to prioritize some other parts, trying to keep a balance along the different chapters.

There are left further studies we want to conduct. Although we have shown that the completeness of the network seems to not affect the results, at least, from our maximum user number to smaller numbers of them, it is left the analysis of bigger percentages of the network. We have done some characterization of the static network and we would like to extend it by considering the network as a weighted link one, which is a more realistic and precise approach than the one performed.

As we said, correlations in human time series have been barely studied in a systematic way. We would like to keep working on that, characterizing deeply how correlations act across time scales. In the modelling part, we want to study the dynamics on other networks, as well as, using real inter-event time distributions to analyse other models of spreading, like the infectious disease ones.

Appendix A

Time Conversions

One of the goals of this work is to characterize inter-event time distributions of Twitter users. These IETs involve several orders of magnitude of time and it is not always easy do the mental conversions. Here we present a table of time correspondences as a cheat sheet, with the aim of being useful when looking at the plots.

Other unities	Seconds	Powers of 10 s	Time
minute	60	10^2	$100s$
hour	3660	10^3	$16m$
day	84600	10^4	$2h\ 45m$
week	604 800	10^5	$1d\ 4h$
month	2 592 000	10^6	$11d$
year	31 536 000	10^7	$116d$
lustrum	157 680 000	10^8	$3y$

Bibliography

[1] T. Hobbes, *Leviathan, or the matter, forme and power of a commonwealth ecclesiasticall and civil.* Yale University Press, 1928.

[2] P. Ball, “The physical modelling of society: a historical perspective,” *Physica A: Statistical Mechanics and its Applications*, vol. 314, no. 1, pp. 1–14, 2002.

[3] C. Castellano, S. Fortunato, and V. Loreto, “Statistical physics of social dynamics,” *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[4] J. G. Oliveira and A.-L. Barabási, “Human dynamics: Darwin and einstein correspondence patterns,” *Nature*, vol. 437, no. 7063, pp. 1251–1251, 2005.

[5] R. D. Malmgren, D. B. Stouffer, A. S. Campanharo, and L. A. N. Amaral, “On universality in human correspondence activity,” *science*, vol. 325, no. 5948, pp. 1696–1700, 2009.

[6] A.-L. Barabási, “The origin of bursts and heavy tails in human dynamics,”

[7] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, “Uncovering individual and collective human dynamics from mobile phone records,” *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, p. 224015, 2008.

[8] Y. Wu, C. Zhou, J. Xiao, J. Kurths, and H. J. Schellnhuber, “Evidence for a bimodal distribution in human communication,” *Proceedings of the national academy of sciences*, vol. 107, no. 44, pp. 18803–18808, 2010.

[9] B. Gonçalves and J. J. Ramasco, “Human dynamics revealed through web analytics,” *Physical Review E*, vol. 78, no. 2, p. 026123, 2008.

[10] M. Meiss, J. Duncan, B. Gonçalves, J. J. Ramasco, and F. Menczer, “What’s in a session: tracking individual behavior on the web,” in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pp. 173–182, ACM, 2009.

[11] F. Radicchi, “Human activity in the web,” *Physical Review E*, vol. 80, no. 2, p. 026118, 2009.

[12] P. C. Maxime Lenormand, Antònia Tugores and J. J. Ramasco, “Tweets on the road,” *PLoS ONE*, vol. 9, p. e105407, 08 2014.

[13] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, "The twitter of babel: Mapping world languages through microblogging platforms," *PLoS ONE*, vol. 8, p. e61981, 04 2013.

[14] D. S. B Gonçalves *arXiv preprint arXiv:1407.7094*, 2014.

[15] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity.," in *ICWSM*, 2012.

[16] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter.," in *ICWSM*, 2011.

[17] R. Kelly, "Twitter study reveals interesting results about usage," *PearAnalytics. August 12th*, 2009.

[18] F. A. Haight, "Handbook of the poisson distribution," 1967.

[19] W. W.-S. Wei, *Time series analysis.* Addison-Wesley publ, 1994.

[20] P. Holme and J. Saramäki, "Temporal networks," *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[21] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[22] T. Zhou, H. A.-T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme, "Role of activity in human dynamics," *EPL (Europhysics Letters)*, vol. 82, no. 2, p. 28002, 2008.

[23] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész, "Universal features of correlated bursty behaviour," *Scientific reports*, vol. 2, 2012.

[24] J. Fernandez-Gracia, V. M. Eguiluz, and M. San Miguel, "Update rules and interevent time distributions: Slow ordering versus no ordering in the voter model," *Phys. Rev. E*, vol. 84, p. 015103, Jul 2011.