POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Big data analysis for social network

Supervisors

Prof. Danilo GIORDANO

Prof. Paolo GARZA

Candidate

Nicolò TRENTACOSTE

October 2021

Summary

For years, social networks have been responsible for the enormous production of data exchanged on the network [1]. The number of people connected to social networks has changed over the years and their habits connected to them [2].
The Politecnico di Torino used passive probes to capture internet traffic for 5 years, from 2013 to 2018, collecting 218TB of compressed text TCP log data relating to connection flows to a central node of the network in Italy.
The main topic of this thesis is to perform big data analytics to analyze this data and study users' behaviour evolution over the 5 years. Firstly, raw logs are filtered by using big data analytics to eliminate the interaction caused by the social buttons used by the users that do not correspond to an actual active session on Facebook (separation between voluntary and involuntary visits).
Then, by using heuristics, user flows are aggregated in sessions describing single users' visits. These sessions are then analyzed to understand how the amounts of data and habits of users towards Facebook have changed over the years, comparing the takeaway with historical data in the literature.
The main takeaways regard: how the number of visits over the years has changed, the average size per visit, the duration of visits, how the daily / weekly pattern of visits has changed, how the time of first visits and the rush hours have changed, how the frequency of visits to the social network has changed.
The results show a tendency to anticipate the first connection to the platform, a longer time window for the most intense sessions and, few changes for the seasonal pattern of use.
From the point of view of the visit frequency, the results show an increase in the average number of daily visits per user and a decrease in the average time between two consecutive sessions.
These latest results, combined with the previous ones, confirm a hypothesis of compulsiveness of users towards the use of the Facebook platform.

Acknowledgements

A Giustina e Vincenzo che mi hanno supportato e sopportato in questo lungo e tortuoso percorso, non facendomi mancare mai nulla.
A Roberta per la felicità trasmessami ed il suo smisurato sostegno in ogni circostanza.
Ai miei coinquilini e ai miei grandi amici.
Voglio inoltre ringraziare il Politecnico di Torino e i Professori Giordano e Garza per l'opportunità, le competenze e le conoscenze che mi sono state fornite.

Grazie di cuore, Nicolò

Table of Contents

List of Figures

Chapter 1

Introduction

In this first chapter we introduce the thesis work, explaining the choice of the topic, the use case taken into consideration and, similar works related to social networks and big data.

1.1 Impact of social networks

For years now, social networks have become part of our lives, influencing them in a more or less implicit way.
The worldwide spread of social networks has its roots in social reasons such as stay in touch with friends, find funny or entertaining content, or share photos, videos and opinions with others [3].
Internet users spend an average of 2 hours and 22 minutes per day on social networking, while 16 to 24 years old users spend a median of 3 hours a day on social media [4].

This intense use has had both positive and negative social consequences.
Among the positive consequences there is that every person with marginal points of view can express his opinion and understand that he is not alone and, when these people meet via social media, they can do things like create publications, share opinions and create online communities that reinforce their view of the world and then break into the common world.
Without social media, social and ethical problems, environmental and political problems would have minimal visibility.
The better visibility of the problems has shifted the balance of power from the hands of the few to the masses.
Social studies show that social networks such as Facebook, Instagram and WhatsApp have made us more social, making us better as individuals and as a society,

being in contact with realities foreign to us [5].

On the other hand, the incredible visibility and networking power given by social networks has brought with it negative consequences.
Although social networks have increased awareness on niche issues, this has not always translated into effective actions on those issues, which instead has resulted in complaining and discussing those issues without leading to real change.
By clicking "like" and "share" buttons people are able to publicly declare their support for a charity in social media, but it can actually make them less likely to give a real support to the cause later [6].
Moreover the ease with which it is possible to read and make viral news on social networks makes it easier to be deceived by fake news.
About the 20% of American adults receive political news primarily through social networks.
Those who get their political news primarily through social networks tend to be less informed and more likely to be exposed to fake news than people who get their news from traditional sources [7].

1.2 Use case: Facebook

The number of users that Facebook has gathered in these years has growth exponentially, reaching 2.85 billion monthly active users in the first quarter of 2021.
Users are considered active if they have logged into Facebook during the past 30 days.
With this numbers, Facebook is the biggest social network worldwide [8].
Being Facebook the most used social network, it is also the one we chose to analyze and study.

Over the years Facebook has changed, the target to which it was addressed and the type of usable content has changed, as well as the type of protocol used and its network infrastructure.
This changes have also influenced the habits of users towards Facebook which are in turn changed over time [9].

Facebook was born in November 2003 from a project by the Harvard student Mark Zuckerberg. The original name was Facemash and its purpose was to vote for student photos on campus.
The following year it spread to other universities, and new features such as the friend request and the wall, where each user and his friends could post public messages, were introduced.

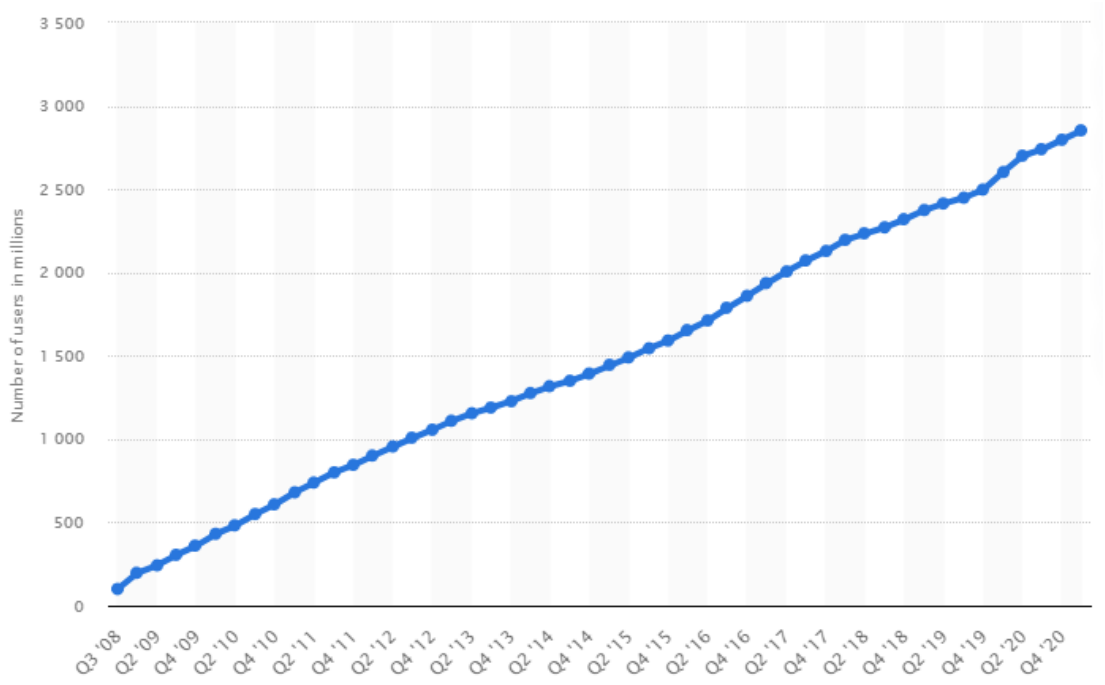


Figure 1.1: Number of monthly active Facebook users worldwide

In February 2009, after having spread also among high school students and having received various investments, Facebook introduces the "like" button, through which users could leave positive feedback on the contents published by other users.
In October 2012, after acquiring Instagram, Facebook reaches one billion monthly active users, becoming the most used social network in the world. Five years later in June 2017 Facebook reaches 2 billion monthly active users.
The huge audience of Facebook and the data collected becomes a privacy problem worldwide, in the following years several companies will exploit the data of millions of users to make behavioral analysis and influence decisions through social media.
In March 2017 one of the biggest scandals related to Facebook data comes to light: Cambridge Analytica, a company that through the information of more than 87 million Facebook users, uses quiz tools and personalized advertisements to influence the behavior of voters and the result of the elections.

The impact of Facebook in social relationships and the possibility of using the platform and its data to influence the habits and behavior of users becomes increasingly clear, the same year the New York Times brings to light another scandal [10].
The New York Times report showed how Facebook made agreements with major mobile device vendors to share large amounts of users' personal data with them. This scandal, along with others, took Facebook to court and eventually Facebook

put more than 30.000 people working on the platform's privacy to make it more secure.

Along with the functional change, the infrastructure and type of network communication to Facebook has changed over the years in order to manage billions of users and provide a service that is easily usable and distributed all over the world.
Initially Facebook was a website hosted on a single server which was interfaced via Rest API, later migrating to different data centers around the world the type of connection to Facebook has been modified, creating different flows for each user and using different network protocols and different APIs.

The purpose of this thesis is also to understand how users habits have changed over the years and in which way.

1.3 Related Works

In this section we will discuss and analyze the main papers exploited for our analysis.
In particular, we will discuss the two papers published by Facebook researchers regarding the analysis of user sessions and, software tools such as frameworks and languages used to conduct the studies.
Moreover we will also present works related to the topics covered such as the psychological impact of social networks and the big data analysis associated with them.

1.3.1 Academic papers

The two main papers [11] [12] used were produced by Facebook's research and development department.
These two papers analyze the online sessions of Facebook users, giving a definition of active session per user and defining how data are exchanged from an edge device to Facebook services.
These two papers were useful in defining a heuristic suitable for grouping user flows into sessions and were used to validate the results obtained, through the metrics they used.

The first paper was useful for defining a temporal heuristic to group flows into sessions based on the duration of the sessions.

The second paper was useful for defining a flow and session dimension heuristic, based on the number of packets exchanged from the edge to Facebook, to

initialize a session.

The combination of the two heuristics mixed with the exploratory analysis, made it possible to create a heuristic for grouping the flows into sessions and, to validate the choices made.

Moreover, beyond these two papers, we will see two papers that present psychological addiction to Facebook and three other papers that address the analysis of social network data from the point of view of big data.

1.3.2 Understanding Short-term Changes in Online Activity Sessions [11]

The online activity of users is normally characterized by patterns that reflect the circadian rhythm and work commitments with patterns that therefore vary between night and day and during the week.

In this paper, Facebook researchers analyzed data from their platform to find out more about users' time patterns, digging into a very small level of detail: the single user's session.

The purpose of this study is to predict the duration of sessions and actions performed by users, to improve the online user experience, using the content cache more efficiently, informing the caching algorithms to anticipate the need for certain information by users, especially for mobile devices in places with poor internet connection.
From the first minutes of a session it is possible to understand from the characteristics of the session whether it will be long or short.
Furthermore, the duration of the session, together with other factors, allows to understand when the user will return to visit the platform. Within a session, the time spent by a user on different contents depends on different characteristics and attributes such as age, number of friends, time spent on the platform since the session started, etc.
The researchers also found that actions such as likes, comments and shares are not evenly distributed across sessions.

This paper was particularly useful in defining "active session", a metric used in the development of the heuristic that groups user flows into sessions.

Human behavior is in fact very heterogeneous, and to control this heterogeneity, Facebook researchers have segmented the time series of user activities into

sessions.
A possible definition of a session is that which begins when a person opens Facebook and ends when he closes it. However, this means that if a user opens Facebook a few seconds after closing it, two different sessions are counted, or if a user leaves the browser open with Facebook without doing anything all day, only one long session is counted.

Since this definition does not reflect the actual use of the platform, an "active session" is empirically defined as a series of consecutive interactions with pauses no longer than 10 minutes.
In other words, an active session consists of all interactions that occur within 10 minutes of previous interactions.

Previous research [13] had shown that 10-15 minutes was an appropriate threshold for building a session. Also, even using slightly different values, the result did not change.

1.3.3 Internet Performance from Facebook's Edge [12]

In this paper, Facebook researchers made a large-scale granular analysis on 10 days of data exchanged by users of the platform on all PoPs in the world, measuring the performance in the TCP and HTTP layers of the sessions, identifying spatial and temporal variations.
However, capturing insights into achievable informations from passive traffic measurements is challenging given that most objects served by Facebook are small.

The purpose of this research is to understand if it is possible to improve performance by incorporating performance information into Facebook's routing decisions, with the ultimate goal of improving the user experience.

The researchers found that throughput is not a useful metric in session identification, as HTTP sessions can be idle for most of their lifetimes.
The study shows that most user sessions have low latency (median minRTT<40ms), and that therefore Facebook's default routing is already great.

While the previous paper was useful for defining a temporal heuristic, this paper was useful for defining a user flow-to-session dimension heuristic, and it was useful to understand what could be the metrics to be analyzed in the analysis of the sessions.

For our thesis work it is important to note that in their studies they always

consider connection made by a client to Facebook services, opened by an exchange of 40 packets in a single session, and they did not consider throughput as an useful metric in session analysis.

1.3.4 Psychology of Social Networking - Facebook addiction

The ever-present nature of Facebook has resulted in a growing body of literature suggesting its addictive potential [14].
The most recent research [15] in the cyberpsychological field identifies four main aspects of the experience in social networks: communication, identity, presence and relationships.

From a psychosocial point of view, social networks can be defined as a digital space, giving users the ability to manage the social network of relationships and their social identity. Furthermore, being able to build virtual and real connections at the same time creates a new reality that is more malleable and dynamic than previous social networks.
The union of real and virtual relationships leads to three paradoxes:

- Through the use of social networks it is possible to change our social identity and the perception that people in our network have of us, but external intervention by our network can influence our identity and reputation even more easily.
- Through social networks it is possible to highlight certain characteristics of our person but this exposes users to privacy problems by recomposing the various traces left by our virtual identity.
- Social networks close the gap between close friendships and acquaintances, allowing you to expand the network of acquaintances. However, this lack of differentiation allows us to behave and appear to acquaintances the same way we behave with close friends, with all the problems that this can bring.

These three paradoxes highlight how the use of social networks must be responsible for improving interpersonal relationships, otherwise it could cause damage to image that cannot be easily solved.

From a psychological point of view it has been shown [16] that the use of Facebook leads to a form of addiction to the gratification it provides.
Research has shown that the main rewards coming from Facebook are due to the maintenance of relationships, the distraction it causes to pass time and from

entertainment, the habitual and excessive use of Facebook is motivated by the desire to escape from negative moods. Some of these rewards are more common in women and young people.

Researchers have found that Facebook addiction is stronger in individuals who suffer from loneliness, anxiety or depression and find social support or a way to pass the time in Facebook. The improvement of the produced mood (mood alteration) leads to a deficiency in self-regulation.
In severe cases this can lead to addiction with serious negative life consequences.

1.3.5 Big data and social media

Along with the explosion in the growth of social networks, the amount of data generated has exploded.
Big data searches, which refer to these large datasets, provide information in different domains, especially in complex applications and social networks.
The management and analysis of large amounts of data associated with social networks is particularly complex, due to the unstructured nature of the data collected.
Several researches [17] have tried to analyze and process data from social networks to understand complex networks.
In particular, there are four areas of big data applied to social networks that have been studied and analyzed:

- Data storage and processing: Efficiency in data retention supports intensive data access and queries.
- Extraction of information and prediction of web content: Big data analyzes allow us to better understand the interests of users to make more accurate predictions on user behavior.
- Protection of privacy and security: it is important to be able to model a system for preventing the spread of disinformation from massive data.
- Multilayer networks and multiparty systems: the new supply of big data applied to social networks provides a set of entities that interact in complicated patterns on multiple dimensions towards different systems in architectures with separate services.

Big data and social media analytics have become of great importance for understanding patterns in data so that market intelligence can be improved.
Big data are used to monitor social media for market growth and brand management [18].

Businesses struggle to monitor customers as extensively as possible, so it becomes important to monitor people's online behavior for their success. Social media analysis thus becomes the analysis of users' online behavior. The phenomenon of big data applied to social media thus creates a new area called "sentiment analysis". The aim is to understand what people say and share every day, in order to improve marketing operations based on feelings.

Furthermore, social networks are characterized by velocity, volume, value, variety, and veracity, the 5 V's of big data. The frameworks for the analysis of big data in Social Network Analysis (SNA) are therefore used [19].
Big data analytical approaches are divided into two categories:

- Content-oriented: which focuses on analyzing the contents of social networks such as text, videos and images. The aim is to discover hidden similarities and relationships in the contents to be able to transform them into structured data on which to do further analysis. Analyzes are also conducted that use Natural Language Processing (NLP) to extract the insights and opinions that are perceived by the public.

- -Network-oriented: analyzes social network data based on nodes or entities and their relationships within social networks. Information about users and nodes can be extracted to predict the rate of news dissemination and how people are affected within the platform. We also use clustering techniques on the type of user to group people who have common characteristics to be exploited for business and marketing occasions.

Despite the numerous researches, methods and applications of big data on social networks, several challenges remain open to be able to manage these huge amounts of data and produce value.

Chapter 2

Dataset

In this chapter we will describe and analyze the dataset used for our study.

2.1 Dataset

The Telecommunication Networks Group of the Politecnico di Torino has created a tool called Tstat (TCP STatistic and Analysis Tool) cite tsat: web which uses a passive monitoring infrastructure of a national ISP - Internet Service Provider in Italy - on a tier- 2 node of the backbone, connected to hundreds of customers and peering Autonomous systems and some Autonomous systems providers [20].
The passive sniffer captures and analyzes real-time traffic through the use of either the libpcap library or Endace DAG cards from different points at the edge of the ISP's network.
The traffic is processed directly in the ISP's points-of-presence (PoPs), and is relayed to the monitoring probes. There is no type of aggregation and customers are associated with a fixed IP address which is anonymized by the probes.
Each probe exports a single entry for each TCP / UDP stream with statistics per stream.
Traffic from the various parts of the network is processed by the passive traffic analyzer mounted on board the probes.
The streams are considered terminated if the packet with the RST flag is observed or if the default timeout of Tstat is triggered.
Each record row contains classic monitoring fields such as IP, port, number of packets and number of bytes. In addition, data is extracted from the payload on application-level information such as ALPN, TLS handshakes, domain name of servers contacted and other essential data for identifying the service.

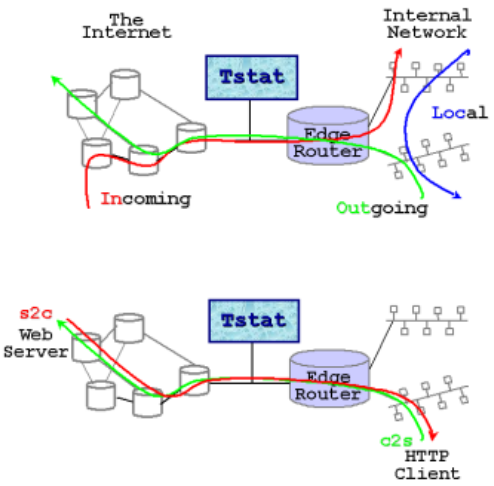


Figure 2.1: TSTAT position on the network.

The probe used collected the data of 10,000 residential ADSLs over 5 years (2013-2018), the records were created, anonymized and stored in the probe, and then moved daily to a centralized data center for historical data. The final dataset was 31.9 TB of anonymized and compressed stream logs (about 247 billion streams).

Started as an evolution of TCPtrace, Tstat analyzes either real-time captured packet traces, using either common PC hardware or more sophisticated ad-hoc cards such as the DAG cards.
Beside live capture, Tstat can analyze previously recorded packet-level traces, supporting various dump formats, such as the one supported by the libpcap library, and many more.
Assuming that, both forward and backward stream of packets, so that it can correlate them to infer advanced measurement indexes.
For example, if both TCP data and ACK segments can be analyzed, Tstat rebuilds each TCP connection status looking at the TCP header in the forward and backward packet flows.
The bi-directionality of the TCP flow analysis allows the derivation of novel statistics (such as, for example, the congestion window size, out-of-sequence segments, duplicated segments, etc.) which are collected distinguishing both between clients and servers, (ie, hosts that actively open a connection and hosts that reply to the connection request) and also identifying internal and external hosts (ie, hosts located inside or outside the measurement point).

This methodology has been applied to infer traffic characteristics at the transport layer, and in particular to the TCP and RTP / RTCP protocols.

The fields of the complete TCP logs are 44 standards metrics + 55 advanced sets of metrics not always captured and elaborated.
Data are saved as blocks of txt files, one block contains a set of files that group together 5 minutes of connections.
The first line of each txt file acts as a header and describes in a contracted way (by means of an abbreviation followed by the number of the field e.g. c-ip: 1) the content of each field.
Tstat creates a set of TXT flow files: log tcp complete, log tcp nocomplete, complete udp logs, complete mm logs, complete video logs, complete skype logs, complete chat logs and log chat messages. Ad each row of streams corresponds to a data stream and each column (data separated by spaces) is associated with a specific measure.

```
#c_ip:1 c_port:2 c_bytes_uniq:3 c_pkts_all:4 c_msn:5 c_msn_a:6 c_msn_d:7 c_msn_n:8 c_msn_u:9 c_msn_y:10 s_ip:11 s_port:12 s_bytes_uniq:13 s_pkts_all:14 s_msn:15 s_msn_a:16 s_msn_
d:17 s_msn_n:18 s_msn_u:19 s_msn_y:20 first_tm:21 durat:22 type:23 ver:24 c_isint:25 flowid:26 T:27 con_t:28
2.235.40.212 60046 176 7 0 0 0 0 0 0 64.4.45.209 1863 350 6 0 0 0 0 0 0 1363639548.923968 0.712 1 MSNP21 0 42180 T 32
10.17.17.67 60929 156 5 0 0 0 0 0 0 157.56.122.81 1863 352 6 0 0 0 0 0 0 1363639576.871537 0.181 1 MSNP21 0 103242 T 32
2.225.2.194 50974 150 7 0 0 0 0 0 0 65.55.64.254 1863 266 6 0 0 0 0 0 0 1363639591.383350 0.781 1 MSNP18 0 134435 T 32
10.174.171.137 33956 1607 26 0 61.145.124.95 5222 4225 21 0 1363639565.152109 33.597 2 UNK 0 77775 T 128
2.225.128.96 49913 267 7 0 0 0 0 0 0 157.56.122.81 1863 372 6 0 0 0 0 0 0 1363639601.690369 0.594 1 MSNP21 0 156038 T 32
10.17.17.67 52158 4362 23 0 0 0 0 0 0 157.56.192.32 1863 9442 26 2 0 0 0 0 0 1363639577.015579 78.457 2 MSNP21 0 103612 T 32
10.174.98.102 52475 261 7 0 0 0 0 0 0 157.56.123.81 1863 353 6 0 0 0 0 0 0 1363639656.819801 0.311 1 MSNP21 0 275002 T 32
2.225.30.15 50474 3423 54 6 63.111.29.135 5222 28116 62 8 1363639534.318737 133.482 3 UNK 0 9872 T 128
```

Figure 2.2: An example of how the log file is structured.

Chapter 3

Methodologies

In this chapter the important aspects of the thesis work are clarified and defined in a more formal way.
In addition, some of the fundamental concepts will be explained and what are the questions we want to answer/what are the analysis we want to do.

3.1 Methodologies and expected results

The key points of the thesis work can be summarized in two categories, the methodological contributions and the result contributions.
Methodological contributions:

- Identification of regular expressions to identify the flows of the social network.
- Distinguish between real visits and social buttons (voluntary and involuntary visits).
- Aggregate flows into sessions.

Once the methodological contributions have been implemented, it is possible to move on to the result ones on the new data obtained.
Result contributions:

- The increase in the number of visits over time.
- The increase in the average size per visit.
- Duration of visits.
- How daily/weekly pattern has changed over time.
- How the time and type of first visit has changed.

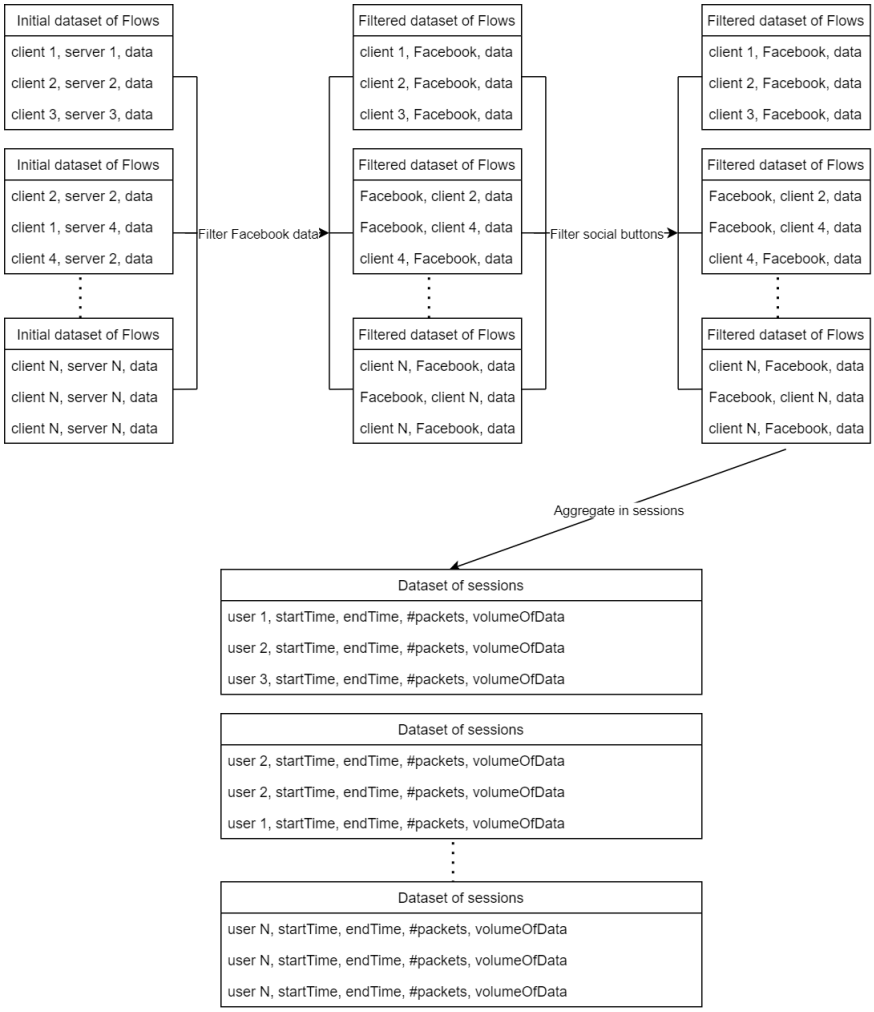


Figure 3.1: Scheme of methodological contributions.

- How has the frequency of visits changed (compulsive?).

To answer these questions, different analysis have been implemented at different levels, based on the heuristics we have developed.
Both solutions will be presented in the following chapters.

3.2 Flows and sessions

The concepts of flow and session are those around which the thesis work is based. The data used comes from TCP communications, a transport layer packet network protocol.

- By the term **flow** we mean the set of all data exchanged, encapsulated in packets, between a client and a server in one direction only. An action made by a client towards a server can be broken down into different flows, depending on the network architecture and the implementation of the protocol in the platform.
  the initial dataset is made up of the flows of users captured in the network and the responses they receive.

- With the term **session** we mean the set of all flows in both directions that group the client-server interactions in a given time span of a user. A single session can contain various actions such as image views, navigation within the platform, interactions with other users, etc.

3.3 Data extraction

The raw data acquired by Tstat must then be mapped and filtered to identify flows to and from Facebook.
Through the use of Spark, regular expressions are applied to full logs to map raw data into useful metrics. Raw data are mapped 3 times and filtered to get useful Facebook related metrics.

The first map takes place on the partitioned data which is mapped via mapPartitions (which converts each partition of the source RDD into multiple elements of the result - possibly none), for extraction.
For each log line, i.e. TCP flow, 22 fields relating to client-server communication are extracted.
The service name is also extracted, based on the most reliable information can be FQDN or SNI or SSL.
If the client is a real client, i.e. inside of the ISP network and the server is outside the ISP network, the data are returned to the main process to check byte to remove possible errors caused by the probe.

The TCP streams with the relative useful fields to which a name has been assigned, are now mapped again twice to obtain the name of the service used and the protocol used.

To obtain the name of the service, a dictionary is used which contains the main FQDN (Fully Qualified Domain Name) keywords for the main services. The service name is extracted and appended at the end of the previous data, if the service is not available flag the flow with "other".
To obtain the name of the protocol used for communication, the previously extracted fields are analyzed. In particular the type of transport protocol (TCP / UDP) and the connection type (Bitmap stating the connection type as identified by TCPL7 inspection engine) are analyzed. The protocol used for the communication is appended at the end of the previous data.

```
1.363639532101944E12,tcp,93.50.228.40,31.13.86.16,www.facebook.com,8192,1377,4039,-,29.720009,24.348783,474.251,facebook,tls,10,15
1.363639531913129E12,tcp,93.50.228.68,31.13.64.32,api.facebook.com,8192,1213,2687,-,36.062348,34.10053,683.215,facebook,tls,12,10
1.363639531156505E12,tcp,93.50.2.71,31.13.81.23,www.facebook.com,8192,1394,1782,-,40.38024,46.011199,1570.353,facebook,tls,13,14
1.363639532799366E12,tcp,93.50.2.253,79.140.80.81,sphotos-a.ak.fbcdn.net,0,0,0,-,41.183882,0.0,128.136,facebook,other_tcp,2,1
1.363639531275297E12,tcp,10.160.16.214,69.171.248.16,ect.channel.facebook.com,8192,1821,5625,-,146.423272,27.0131,1791.705,facebook,tls,9,12
1.363639532658219E12,tcp,93.50.2.218,193.247.166.50,fbcdn-profile-a.akamaihd.net,8192,787,2672,-,10.097663,109.518683,714.718,facebook,tls,9,9
1.363639532688697E12,tcp,2.225.99.195,31.13.64.24,m.facebook.com,8192,1573,4836,-,32.668747,23.909536,789.513,facebook,tls,11,16
1.363639532214985E12,tcp,2.225.30.167,31.13.64.32,api.facebook.com,8192,1614,3964,-,22.590247,49.230692,1509.961,facebook,tls,15,15
1.363639532586285E12,tcp,10.174.197.247,31.13.64.7,api.facebook.com,8192,1379,7581,-,28.537787,41.459927,1197.446,facebook,tls,13,16
1.36363953158195E12,tcp,2.225.30.137,31.13.64.32,api.facebook.com,8192,4990,2754,-,22.527582,221.407953,2246.581,facebook,tls,18,18
```

Figure 3.2: An example of how the extracted data are structured.

Finally, the mapped data are filtered to extract only the flows to and from Facebook, requiring that the field relating to the service is precisely that of the social network Facebook.

The data thus extracted are cleaned of possible errors made by the probe and normalized using milliseconds as the unit of measurement of time and bytes as the unit of measurement of the dimension.
The following are extracted and saved as csv: date, IP, size, number of packets and throughput for the flows to and from Facebook, divided by year.
The extraction of data relating to Facebook is necessary for the construction of the dataset on which the analyzes will be carried out. Initially, small data analyzes will be carried out on small portions of data for small time intervals.

The analyzes made on a few clients will give suggestions on the possible composition of traffic and hypotheses will be created from those which will then be applied on a larger scale to verify them.

A first study is done on a few clients in one week time windows to analyze possible weekly patterns. In the example shown it is possible to see the amount of data exchanged by the flows to Facebook of 3 clients with different traffic classes, one that exchanges a large amount of traffic and ranks in the top 10% of the traffic exchanged (90th percentile), one that trades an average amount of traffic and is

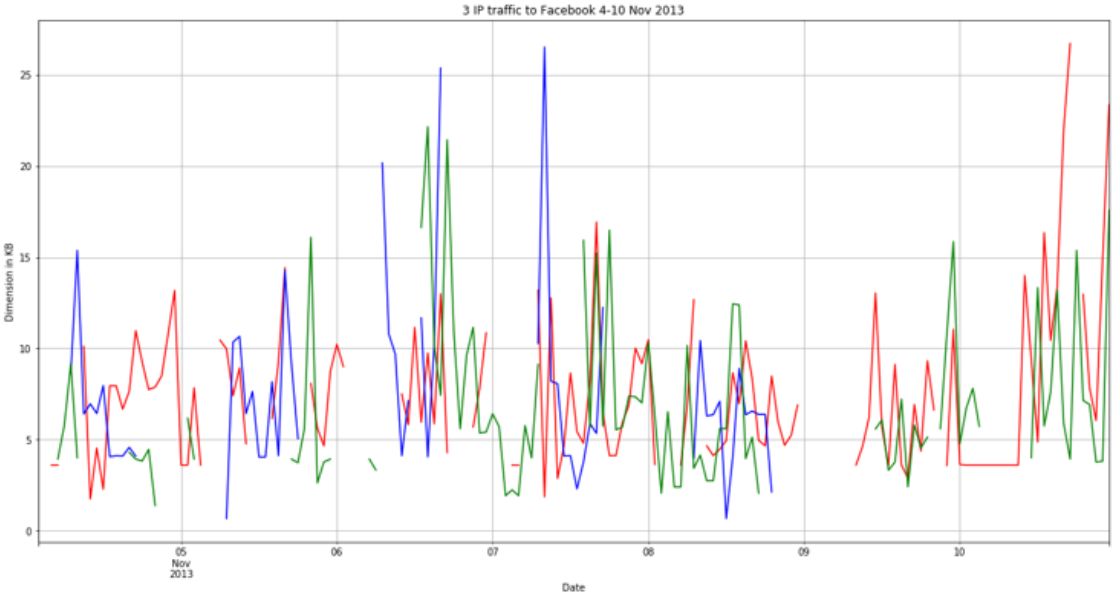


Figure 3.3: Plot of one week of extracted data. Three different class of traffic divided by volume (red: 90th percetile of volume, green: 60th percentile of volume, blue: 20th percentile of volume).

in the top 40% (60th percentile) and, the last one that trades a small amount of traffic and is in the top 80% (20th percentile).

3.4 Data filter

In this section we will talk about data filtering. The extracted data, although filtered, do not provide additional information to that already in our possession. It is therefore necessary to transform them, changing the format, structure, or values of data, in order to obtain new metrics to formulate an heuristic and draw conclusions.

3.4.1 Social buttons and real visits

The TCP flows to and from Facebook may not be caused by "real" visits but by interactions of users with social buttons [21] e.g.: like, share, comment, cross-site actions, auto-play of videos.

To aggregate the flows into sessions and analyze the visit patterns, it is necessary to eliminate the contributions due to the social buttons from the flows.
To filter the data relating to social buttons we develop a heuristic given by the

exploratory data analysis made on the data.

We consider the logs with a duration and a size between the first and the 99th percentile (to eliminate dirty or negligible data) and then we analyze the flows in several days of several years.
Starting with a small data analysis, drawing conclusions, validating them on a test set, and then applying the best solution to the entire dataset.

3.4.2 Exploratory data analysis

The main data of the flows in our possession are size, duration, number of packets and throughput.
Although, as we will see below, the number of packets and throughput are very influential from the point of view of the sessions, as far as flows are concerned, this is not true. In fact, the throughput and the number of packets are not indicative on the single flow to exclude the actions due to the social buttons, given the different nature and interaction / implementation of the social buttons themselves.
The exploratory analysis therefore focused more on the size and duration of the flows.
We expect that the contribution due to social buttons is not, volumetrically speaking, higher than 25% - 30% of the real traffic on Facebook.

The first step in formulating a heuristic is to consider data with a duration between the first and the 99th percentile. This is because below the first percentile there are flows with a negligible duration (e.g. <10 ms) and above the 99th percentile there are excessively large flows due to errors in the transcription of the probe.

The dimension of the flows is first analyzed on these data. We use the ECDF plot (Empirical Cumulative Distribution Function) which represents the division function of the empirical measure of a sample, which provides an unbiased and consistent estimator of the analyzed distribution.
From the ECDF of the flow size it is clear that flows with a size of less than 4.6kB, equal to 20% of the traffic volume, may be due to social buttons.
Furthermore, connections to Facebook by a user require a TLS handshake flow with an average size of 4.5kB [22].

The other parameter we analyzed was the duration of the flows to Facebook.
The minimum duration in the sample of data analyzed was between 5 and 10 seconds per flow.
We therefore decided to consider a duration of less than 10 seconds as an interaction due to social buttons.

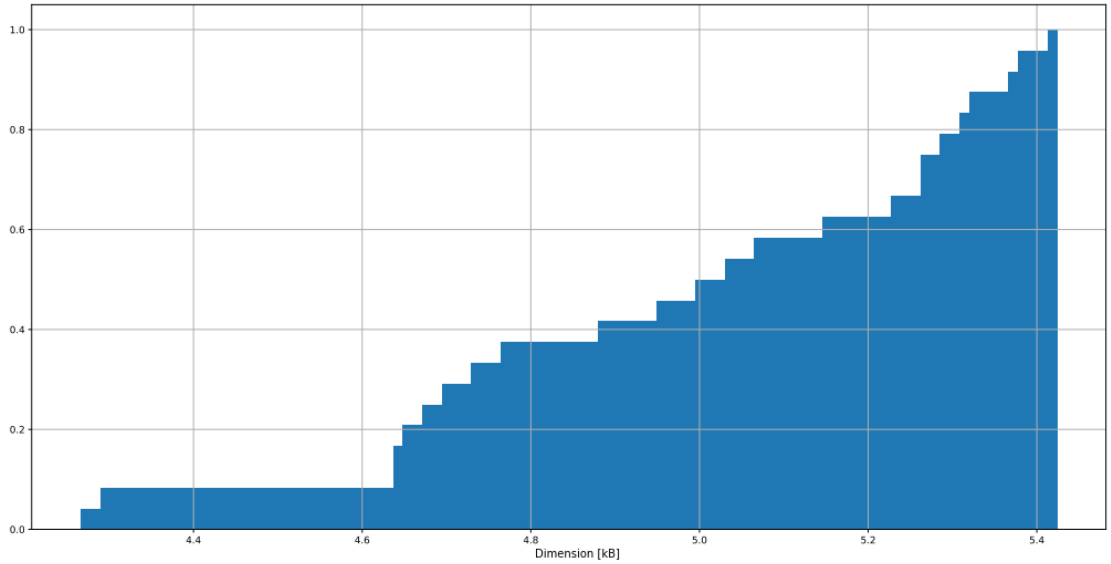


Figure 3.4: ECDF of dimensions of flows to Facebook.

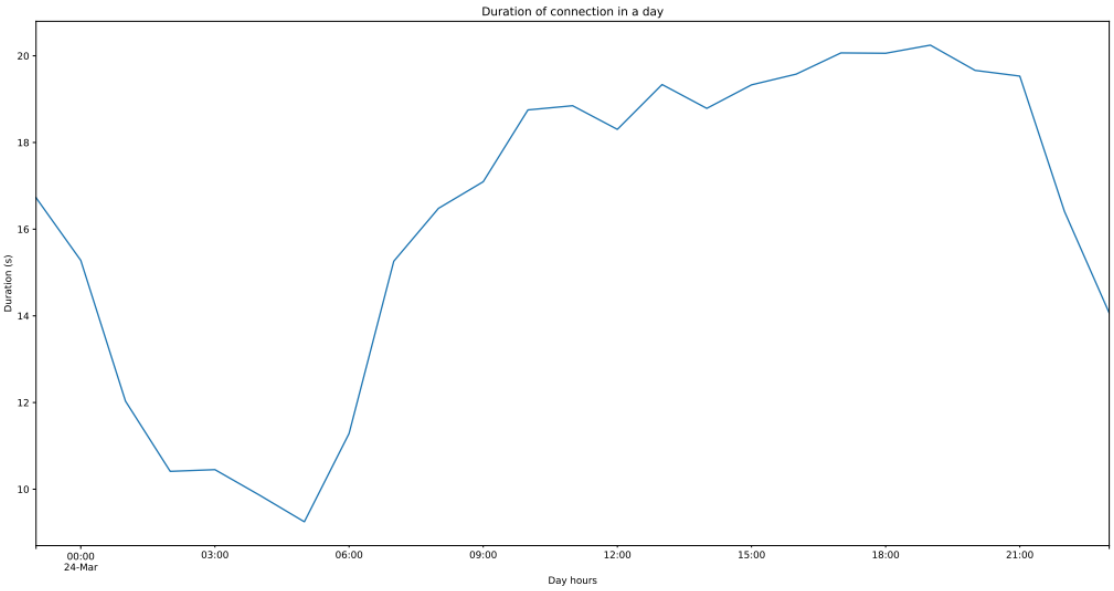


Figure 3.5: Plot of the durations of flows to Facebook.

3.4.3 Validation

To validate the assumptions we considered a maximum reduction in traffic volume of around 25-30% and we decided to analyze the daily and hourly patterns to see if they coincided with those in the literature.

The parameters in our possession (size, duration, number of flow packets and throughput) were considered as hyperparameters in a machine learning algorithm and we tried to find the optimal combination by applying a grid search.

The flow packets and the throughput initially considered were then discarded from the heuristic search because a flow could exchange an arbitrary quantity of packets and in any case not be considered part of the social buttons.
In addition, many of the flows had similar throughput and are not indicative of the use of social buttons.
The heuristics tested using these two parameters truncated much of the traffic and filtered up to 40% of the total traffic.

The gridsearch applied to the size and duration of the flows considered 4.6, 4.8 and, 5.2 kB for the size of the flows and, 5, 10 and, 15 seconds for the duration of the flows.

The heuristics that used 4.8 and 5.2kB as a threshold for the size of the streams were overly "aggressive" and filtered out more than 30% of the traffic.
We considered valid the heuristic with 4.6KB of size threshold and 10 seconds of duration of the streams.

We validated the results obtained in the small data field on larger data samples. The reduction in traffic volume after filtering the data was 26%. And the result was considered valid until the tests for subsequent experiments.

The heuristic presented above is the one that gave the most satisfactory results.
After filtering the data, we apply big data heuristics over the entire years, and analyze the daily and hourly patterns of client traffic through the use of heatmaps.
The heatmaps produced are normalized on the median of traffic per month and on the number of days/hours relating to the months (to avoid data unbalanced by the lack of certain time windows).

The daily patterns that can be analyzed after the application of the heuristic are quite consistent with the data in the literature: first connection at 8am and peak traffic around 10pm.
These results together with the not excessive reduction in the volume of traffic indicate that our heuristic is not exaggeratedly filtering even important data and could be valid.

However, we remain willing to change it if in subsequent experiments some analysis should have an unexpected value.

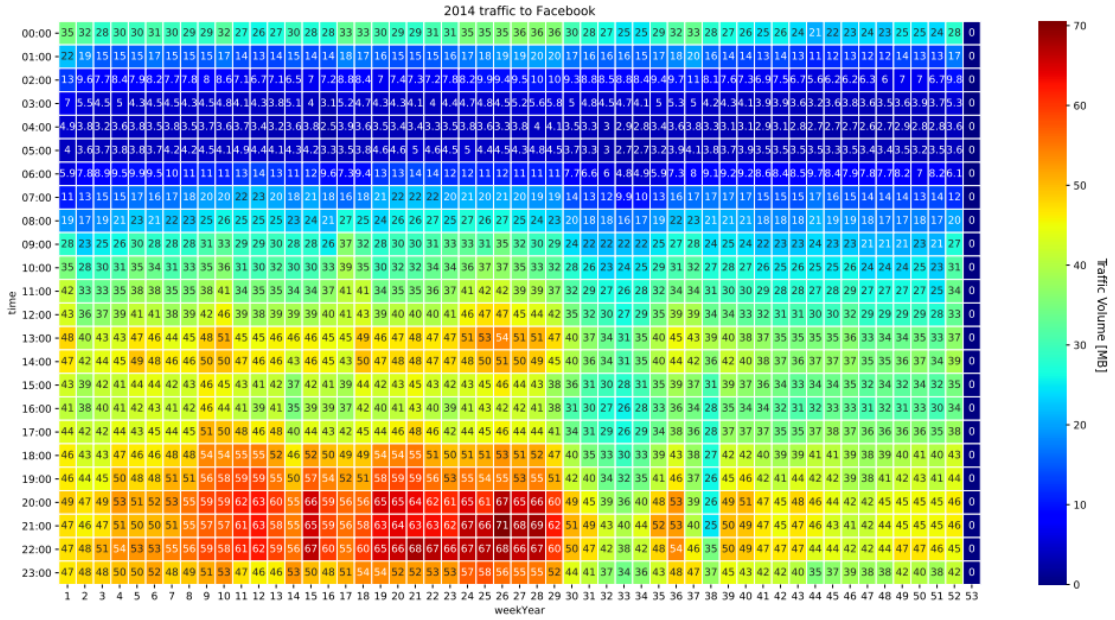


Figure 3.6: Heatmaps of per capita average traffic in 2014 - No social buttons.

3.4.4 Preliminary results

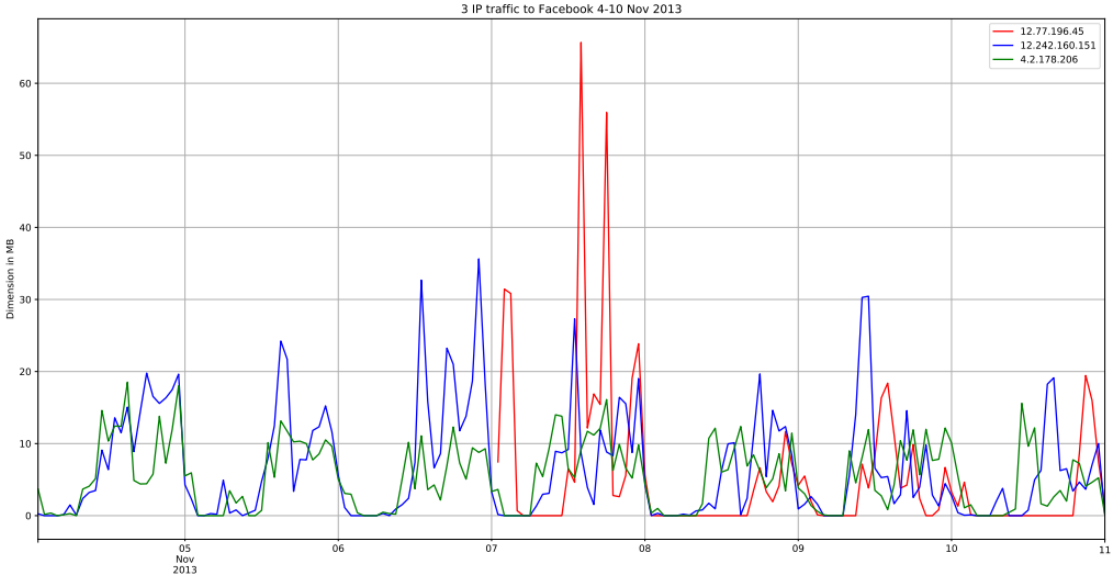


Figure 3.7: An example plot of the traffic generated by three clients in a week after filtering the social buttons.

The results obtained, although satisfactory, are considered valid until subsequent

applications and in any case re-evaluated if they do not comply with future results.

3.5 Tool

For the development of this thesis work different tools were used.
Some tools as platforms made available by the Polytechnic of Turin, others as frameworks and libraries made available by the language.
In this section we will present the main tools used, what they are, how they work and how they have been useful.

3.5.1 Python

Python [23] is a high-level programming language, first released publicly in 1991 by its creator Guido van Rossum, a Dutch programmer currently operating at Dropbox.
It derives its name from the comedy Monty Python's Flying Circus by the famous Monty Python, which aired on the BBC during the 1970s.
Python represents one of the main technologies of the core business of giants like Google (YouTube is heavily based on Python) and ILM.

Convenient, but also simple to use and learn, Python, in the intentions of Guido van Rossum, was born to be an immediately understandable language. Its syntax is clean and slim as well as its constructs, very clear and unambiguous. Logical blocks are built by simply aligning the lines in the same way, increasing the readability and consistency of the code even if different authors work on them.

Python supports several programming paradigms, such as object-oriented (with support for multiple inheritance), imperative and functional, and offers strong dynamic typing.
It comes with an extremely rich built-in library, which together with automatic memory management and robust exception handling constructs makes Python one of the richest and most comfortable languages to use.

Python is a pseudocompiled language: an interpreter takes care of analyzing the source code (simple text files with the extension .py) and, if syntactically correct, of executing it. In Python, there is no separate compile step (as in C, for example) that generates an executable file from source.
Being pseudo-interpreted makes Python a portable language. Once a source has been written, it can be interpreted and executed on most of the platforms currently used, whatever the operating system is, the presence of the correct version of the interpreter is enough.

These characteristics have made Python [24] the protagonist of a huge spread all over the world.
This is because it guarantees the rapid development of applications of any complexity in all contexts: from the desktop to the web, passing through video game development and system scripting.

All the scripts used for the thesis work, from data processing to graphical presentation, were written in Python.

3.5.2 Jupyter Notebooks - Jupyter Hub

The name of the project "Jupyter" derives from the three basic programming languages, Julia, Python and R.
Jupyter Notebook [25] is an application based on the client-server model of the non-profit organization Jupyter Project founded in 2015. It allows the creation and sharing of web documents in JSON format, which follow a pattern and an ordered list of input/output cells.
Among other things, these cells offer space for codes, markdown texts, mathematical formulas and equations or multimedia content (rich media).

The created Jupyter documents can be exported as HTML, PDF, Markdown or Python documents or alternatively they can be shared with other users via e-mail, Dropbox, GitHub or your Jupyter Notebook.
The two core components of Notebook Jupyter are a set of different kernels (interpreters) and the dashboard. Kernels are small programs that process language-specific requests (requests) and react with related responses. A standard kernel is IPython, a command line interpreter that allows you to work with Python. Over 50 kernels provide support for other languages such as C ++, R, Julia, Ruby, JavaScript, CoffeeScript, PHP or Java.
The dashboard serves on the one hand as a management interface for individual kernels and on the other as a central for creating new Notebook documents or for opening existing projects.

JupyterHub is a multi-user server that includes a proxy, which connects several Jupyter Notebook instances together.
It can be hosted in the cloud or on premise and allows the use of a common Notebook environment. The server administrator manages common access to individual documents at will (an authentication method can be implemented), while individual users can concentrate on their tasks.

The thesis work was carried out on the JupyterHub of the Polytechnic of Turin. Users spawn servers that are hosted in any worker machine of the BigData@PoliTO Cluster.
When spawning a server, users request specific amounts of memory and CPU cores for their notebooks. The distribution of resources in the BigData @ PoliTO cluster is managed with Kubernetes.
Several programming languages and environments are supported in the BigData@PoliTO Cluster, including Python, Java, R and Octave. Jupyter Notebooks can also interface with Hadoop services, for example, firing Spark jobs directly from users' notebooks to interact with TBs of data stored in HDFS and/or CEPH distributed file systems.

All popular Python packages are supported by the BigData@PoliTO Cluster, and the distribution of packages and virtual environments is managed with Anaconda.

3.5.3 Pandas

Pandas [26] is a Python language library useful for data analysis.
The name pandas is the contraction of "panel data" or data set.
Pandas provides a set of functions and methods for working with datasets of various kinds. In particular, it has implemented functions for managing dataframes, timeseries and, to plot data.

The convenience of pandas compared to other frameworks/libraries is the speed and ease of use. However, these facilities also lead to negative consequences such as the possibility of not being able to work in a distributed manner and the large amount of RAM used, which can reach up to 8 times the real size of the loaded data.

In this thesis work Pandas is used on small portions of data or on the result of pre-processing of data made with other frameworks to do data analysis.

3.5.4 Spark - PySpark

Spark [27] is a unified analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing.
It's the de-facto standard unified analytics engine and largest open-source project in data processing.

Spark uses clusters of machines to process big data by breaking a large task into smaller ones and distributing the work among several machines.
The secret to Spark's performances is parallelism.

Each parallelized action is referred to as a job. Each job is broken down into stages, which is a set of ordered steps that, together, accomplish a job.
Tasks are created by the driver and assigned a partition of data to process. These are the smallest unit of work.

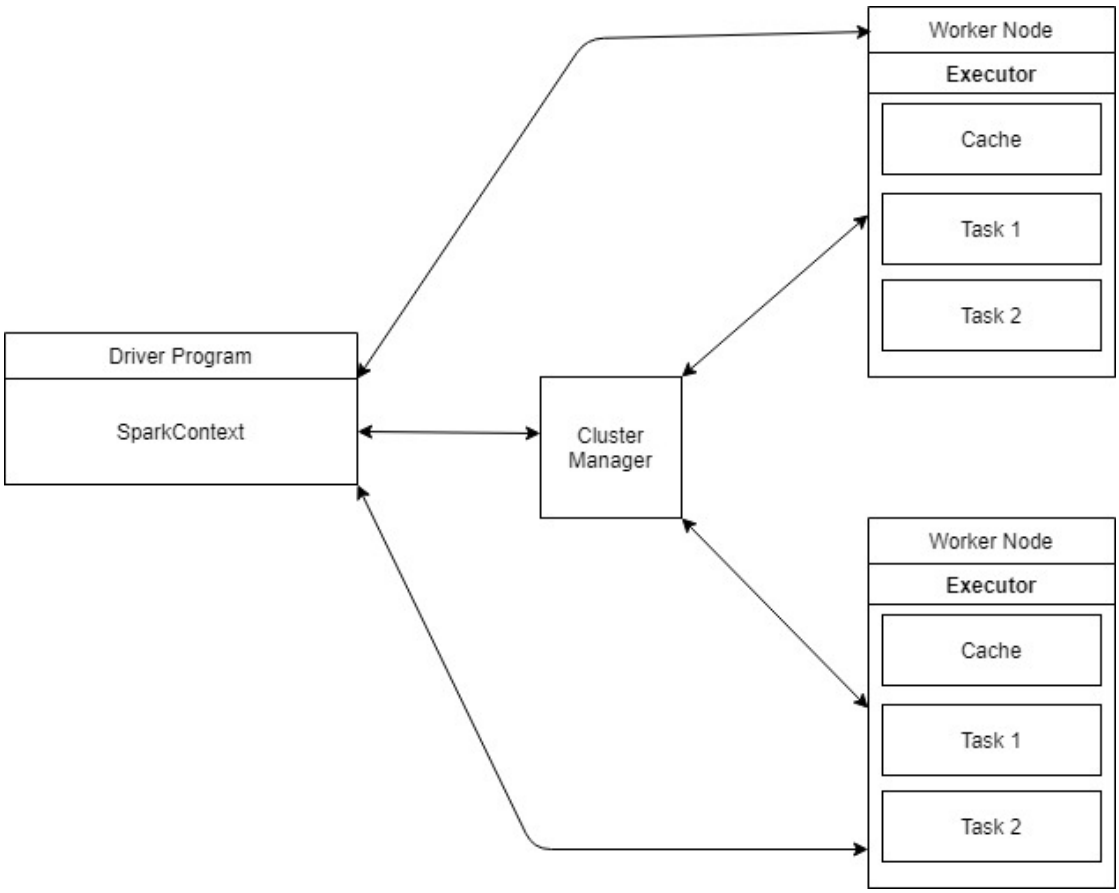


Figure 3.8: Apache Spark Architecture

In the master node, there is the driver program, which drives the application.
The written code behaves like a driver program.
Inside the driver program, the first thing to do is to create a Spark context. Let's assume the Spark context is a gateway to all Spark features. Everything done on Spark goes through the Spark context.
The Spark context works with the cluster manager to handle various jobs.

The driver program and SparkContext take care of the execution of the work within the cluster.
A job is split into multiple activities that are distributed across the worker node.

Whenever an RDD is created in the Spark context, it can be deployed across multiple nodes and can be cached.

The worker nodes are the slave nodes whose carrying out is essentially performing the activities.
These tasks are then performed on the partitioned RDDs in the worker node and then return the result to the Spark context.
Spark Context accepts the work, splits the work into tasks, and distributes them to the worker nodes. These activities work on the partitioned RDD, perform operations, collect results, and return to the main Spark context.

Apache Spark has a well-defined layered architecture where all components and layers of spark are loosely coupled. This architecture is integrated with various extensions and libraries. The Apache Spark architecture is based on two main abstractions namely the RDD and the DAG.

RDDs (Resilient Distributed Dataset) are the building blocks of any Spark application.

- **Resilient**: Fault tolerant and able to reconstruct data in the event of a failure.
- **Distributed**: Data distributed across multiple nodes in a cluster.
- **Dataset**: Collection of data partitioned with values.

It is a layer of abstract data on distributed collection. It is of an immutable nature and follows lazy transformations. The data in RDD is split into individual on a key. RDDs are able to recover quickly from any problem as the same data blocks are replicated across multiple executor nodes.

Direct Acyclic Graph (DAG).
The client sends the spark user application code. When application code is sent, the driver implicitly converts user code containing transformations and actions into a cyclic graph logically called DAG. converts the logical graph called DAG into a physical execution plan with many phases. After converting to a physical execution plan, it creates physical execution units called tasks at each stage. Then the activities are grouped and sent to the cluster.

The main feature of Apache Spark is its in-memory cluster computing which increases the processing speed of an application. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It is designed to cover a wide range of workloads such as batch applications, iterative

algorithms, interactive queries, and streaming.

Spark provides several high-level APIs to work with it, written in different languages. The API for Python is called PySpark which combines the computing power of Spark with the coding ease of Python.

In our thesis work all the elaborations made on big data were done with PySpark.

3.5.5 Hive

Hive is a datawarehousing framework for distributed processing of large amounts of data. It fits perfectly into the Big Data scenario and was designed by Facebook to facilitate the writing of MapReduce tasks on HDFS, Hadoop's distributed file system.

The innovative feature of Hive is the adoption of a SQL-like syntax, called HiveQL, for the definition of the operations to be performed on the data. This has made it an easy tool to approach both for programmers and for operators from other sectors.

Hive has evolved a lot, entering into close relationship with the universal Big Data platform, Apache Spark, opening up to various file formats and storage systems other than HDFS and depopulating on Cloud platforms.

Hive acts as a relational database and in this we can find a certain assonance with the choice of SQL as a query language. Hive allows to manipulate the files available to the repositories (HDFS) by means of "virtual" databases that it creates in a portion of space, called metastore.
The metastore is managed through a relational database which by default is Apache Derby, but can be replaced with others.
The metadata that will represent the link between the repository files and the metastore tables that will be stored.

In our thesis work Hive was used to provide data summary, query and analysis.

3.5.6 Hue

Hadoop Hue is an open source UX/UI for Hadoop components. The user can access Hue right from within the browser and it enhances the productivity of Hadoop developers.

Hue is developed by Cloudera and is an open source project. Through Hue, the user can interact with HDFS and MapReduce applications.
Users do not have to use command line interface to use Hadoop ecosystem to use Hue.

Hue contains within it several components through which a user can use the Hadoop ecosystem.

- **HDFS browser**: Working with Hadoop Ecosystem one of the most important factors is the ability to access the HDFS browser through which users can interact with HDFS files interactively.
  This provides such an HDFS interface through which all requests can be made over HDFS.
- **Job Browser**: The Hadoop ecosystems are made up of many jobs and work developers need to know which job is currently running on the Hadoop cluster and which job has completed successfully and which one has errors.
  Through the Job browser it is possible to access all information relating to the job directly from within the browser.
- **Hive Query Editor**: The Hive Query Editor allows users to write Hive SQL queries directly within the editor and the result can be shown in the editor.
  The hue editor makes query data easier and faster.
  The user can write SQL type queries and running these queries can produce MapReduce jobs by processing data and the job browser can be controlled by the browser even when it is running.
  The result of the query can be shown in the browser.

In our thesis work, Hue was used to explore data from browsers, see partitions and query results.
It was also useful for analyzing job execution.

Chapter 4

Methodological Results

This chapter explains and analyzes the process that was necessary to formulate a heuristic for grouping the individual flows of each user into sessions of the same user to Facebook.
It also introduces the methodological results, that is to say the ones related to the validation of the methodology and the transition from flows to sessions.

4.1 Raw logs to flows

As mentioned in the previous chapters we have applied a heuristic to identify the flows from the raw logs.
Filtering the data to eliminate possible errors due to the probes and identifying common patterns in the flows to and from Facebook through regex such as

```
service_keywords = {
    "facebook":
        ["\.facebook\.com$", "\.fbcdn\.net$", "\.facebook\.net$",\
        "^fbcdn", "^fbstatic", "^fbexternal", "\.fbsbx\.com$"]
                    }
```

Given the large amount of data, we have exported the fields in chunks divided by years and sorted by date.
The exported fields were the date in milliseconds, protocol used, the IP address of the client and the server, the service that made the connection, a boolean to represents if the connection is over tls, application layer protocol of the client and of the server, two boolean to check if the client and the server are internal to the network connection, the size, the average rtt, the duration and the number of packets of the flow.

/ user / s257667 / 2013.csv / **part-00001-c27a062e-3a34-44b8-97f2-5c3ecfbfb4ea-c000.csv**

```
1.36364312544274E12,tcp,93.50.209.71,195.10.50.177,photos-g.ak.fbcdn.net,0,0,0,-,3.20734,61.253437,5618.839,facebook,other_tcp,4,2
1.36364312548373 9E12,tcp,93.50.209.71,195.10.50.150,photos-h.ak.fbcdn.net,0,0,0,-,3.162342,61.063447,5579.798,facebook,other_tcp,4,2
1.36364307961006 3E12,tcp,93.50.210.37,69.171.235.16,4-pct.channel.facebook.com,8192,1620,5259,-,172.586685,103.230135,51453.984,facebook,tls,12,13
1.36364312548730 9E12,tcp,93.50.209.71,195.10.50.150,photos-h.ak.fbcdn.net,0,0,0,-,3.030348,65.094245,5579.735,facebook,other_tcp,4,2
1.36364312198712 6E12,tcp,2.225.128.133,193.247.166.64,fbcdn-profile-a.akamaihd.net,8192,989,3357,-,22.483917,466.833829,9081.755,facebook,tls,14,12
1.36364312547510 9E12,tcp,93.50.209.71,193.247.166.66,profile.ak.fbcdn.net,0,0,0,-,10.60647,60.165992,5597.058,facebook,other_tcp,4,2
1.36364312546674 6E12,tcp,93.50.209.71,193.247.166.67,fbcdn-profile-a.akamaihd.net,0,0,0,-,10.415979,64.685266,5610.02,facebook,other_tcp,4,3
1.36364312203091 3E12,tcp,2.225.128.133,193.247.166.56,fbcdn-profile-a.akamaihd.net,8192,989,3789,-,29.858507,387.158642,9052.438,facebook,tls,17,14
1.36364312755027 5E12,tcp,93.50.211.63,31.13.86.16,api.facebook.com,8192,1848,3767,-,37.658647,36.186794,3546.14,facebook,tls,15,15
1.36364312201904 8E12,tcp,2.225.128.133,193.247.166.56,fbcdn-profile-a.akamaihd.net,8192,989,3901,-,30.283486,462.206926,9080.782,facebook,tls,16,13
1.36364308794191 6E12,tcp,10.161.160.197,195.10.8.43,profile.ak.fbcdn.net,1,419,2873,-,31.10563,22.816239,43160.175,facebook,http,6,6
1.36364307533642 2E12,tcp,10.161.160.197,195.10.8.43,profile.ak.fbcdn.net,1,854,5600,-,33.809655,103.096138,55767.246,facebook,http,9,8
1.36364303680461 E12,tcp,10.161.160.197,195.10.8.43,profile.ak.fbcdn.net,1,3869,25504,-,39.29437,28.326574,94299.961,facebook,http,28,33
1.36364311217668 7E12,tcp,10.160.49.83,31.13.64.33,m.facebook.com,1,3172,42306,-,34.575108,342.20726,18938.505,facebook,http,34,51
1.36364312200737 8E12,tcp,2.225.128.133,193.247.166.56,fbcdn-profile-a.akamaihd.net,8192,989,3586,-,30.334983,566.619669,9107.881,facebook,tls,17,13
1.36364311298358 9E12,tcp,2.225.128.133,31.13.86.17,m.facebook.com,8192,5615,55465,-,30.633823,111.120778,18142.82,facebook,tls,47,66
1.36364312748032 4E12,tcp,10.160.8.173,195.10.8.66,profile.ak.fbcdn.net,0,522,0,-,50.79546,166.734663,3646.681,facebook,other_tcp,6,4
1.36364310977742 9E12,tcp,2.225.69.80,193.247.166.73,fbcdn-profile-a.akamaihd.net,8192,1220,7239,-,12.100197,119.375031,21349.794,facebook,tls,17,15
1.36364310978169 E12,tcp,2.225.69.80,193.247.166.73,fbcdn-profile-a.akamaihd.net,8192,1220,6957,-,20.703482,114.333913,21353.179,facebook,tls,16,15
1.36364310978016 9E12,tcp,2.225.69.80,193.247.166.73,fbcdn-profile-a.akamaihd.net,8192,1220,6829,-,22.722932,109.798204,21356.644,facebook,tls,16,14
```

Figure 4.1: Raw data to flows.

The results obtained, however, are still not very usable as they have a lot of fields that we don't use and present the contributions from the social buttons and therefore flows not given by "real" visits.

This data was then aggregated, filtered and cleaned to obtain only the date, the client IPs, the size, the number of packets and the duration of the flows.

/ user / s257667 / 2013Filtered.csv / **part-00000-a328f783-b759-4ccb-9ae2-df94c9c5c22d-c000.csv**

```
2013-03-18T21:45:32.617+01:00,10.174.164.67,15.0791015625,51,14.966977
2013-03-18T21:45:34.760+01:00,10.174.199.77,54.6650390625,122,13.482806
2013-03-18T21:45:36.336+01:00,2.235.40.182,53.259765625,74,12.011897999999999
2013-03-18T21:45:34.107+01:00,10.174.105.114,58.435546875,79,15.457605
2013-03-18T21:45:36.215+01:00,2.238.101.127,163.37109375,180,13.477274999999999
2013-03-18T21:45:36.743+01:00,10.174.192.94,112.6787109375,192,13.184381
2013-03-18T21:45:36.157+01:00,10.174.171.50,15.3125,30,13.917669
2013-03-18T21:45:38.555+01:00,2.238.101.81,57.8955078125,88,11.945227000000001
2013-03-18T21:45:39.413+01:00,10.162.160.61,26.1328125,49,11.356224000000001
2013-03-18T21:45:33.001+01:00,10.162.160.61,12.541015625,38,17.903636
2013-03-18T21:45:38.707+01:00,2.238.101.28,36.2041015625,87,14.337677
2013-03-18T21:45:36.961+01:00,10.162.185.69,12.5224609375,47,16.259376
2013-03-18T21:45:34.822+01:00,10.162.170.3,55.3251953125,93,19.627857
2013-03-18T21:45:36.149+01:00,10.174.171.50,13.1259765625,27,18.325423999999998
2013-03-18T21:45:34.817+01:00,10.162.170.3,43.8515625,72,20.291103
2013-03-18T21:45:40.982+01:00,93.50.211.214,25.0458984375,74,14.175256
2013-03-18T21:45:41.354+01:00,10.174.112.171,16.8955078125,64,13.879342000000001
2013-03-18T21:45:38.006+01:00,10.174.171.50,96.8017578125,112,17.310706
```

Figure 4.2: Raw data of flows filtered.

4.2 Flows to sessions

After validating the heuristic to eliminate social buttons, we looked for a heuristic to aggregate the flows into sessions.
A session is the set of flows relating to the use of Facebook by a user on an ongoing basis.

We initially analyze the traffic (size, duration, number of packets, throughput) of a few clients over several days in relation to the flows produced.
An initial heuristic is formulated which is then applied and validated on multiple clients in longer time intervals.

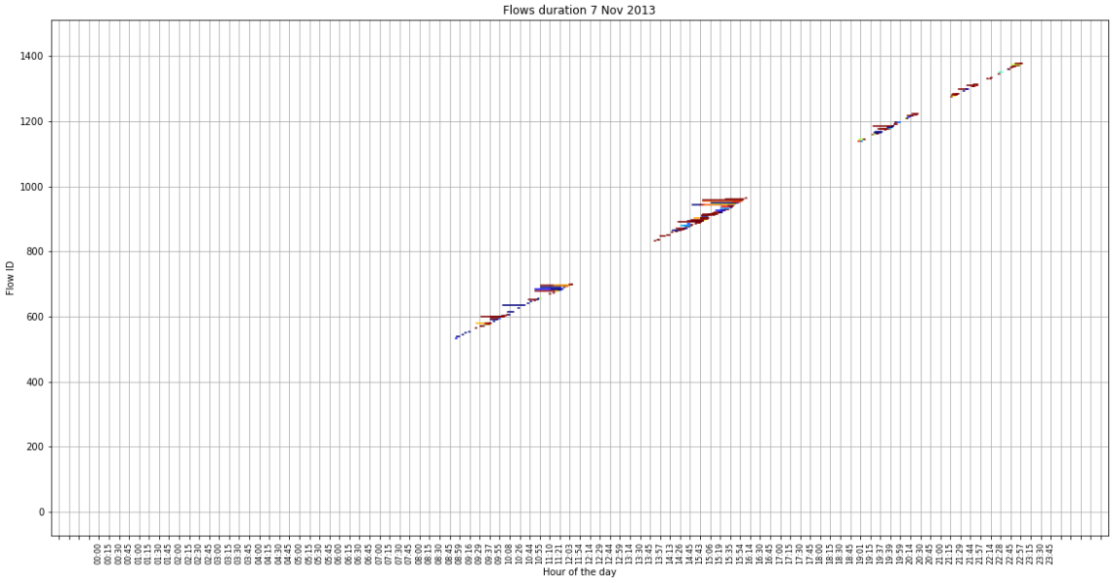


Figure 4.3: Per user flows in one day.

The increase in flows over the years corresponds to a shorter duration of the flows. This can be motivated by a change in the infrastructure used by Facebook and a change in the type of protocol used for data transmission more efficiently.
In fact, over the years Facebook has gone from the de facto standard TCP protocol, used for decades on the internet to their implementation of the QUIC [28] protocol, called FB Zero [29] [30].

The protocol change was implemented to optimize network performance in a more secure way, so as to provide users with a better experience in the services offered.
QUIC has shown significant improvements in several metrics, including request

errors, tail latency, response header size, and numerous others that meaningfully affect the experience of people using Facebook.
As a first step in formulating the heuristics, we overlap the typical connection patterns found previously and described in the literature with the set of user flows, after excluding the use of social buttons.
By associating the frequent connection times with the analyzed flows, we have seen how a session can be identified with 4 or more successive flows with an average duration between 4 and 10 minutes in an average time window of 20 minutes.

We then analyzed the behavior of the set of flows in relation to the duration and quantity of data exchanged (traffic volume and number of packets), looking for analogies with possible sessions.

4.3 Used heuristic

To the exploratory studies made on the data in our possession, we have added and enriched the heuristics with the contribution provided by the research carried out by the engineering department of Facebook [31].
In particular, we used the definition of active session given by Facebook [11], according to which the interactions of a user towards Facebook are to be considered as two distinct sessions if there is an inactivity period of more than 10 minutes between the actions performed in sessions.

Figure 4.4: Facebook Research about distance in time between sessions.

Another important research in the definition of our heuristic was the definition of a session to Facebook as flows that initially exchange 40 packets (Facebook handshake) and whose average duration ranges from 5 to 15 minutes [12].
Therefore, combining the studies made on our data with the considerations taken from the two reference papers, we formulated the heuristics and applied it on 100 clients in different weeks.
Applying the heuristics we noticed a separation of the flows into clusters similar to the sessions.
The first form of validation is done by comparing the distributions of the size and duration of the flows before and after applying the heuristics.

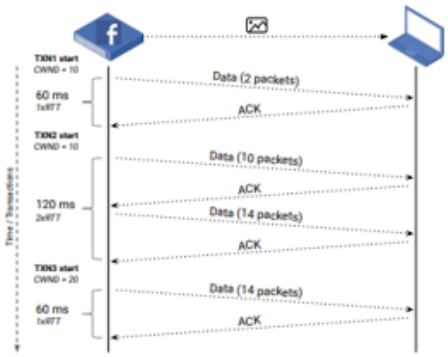
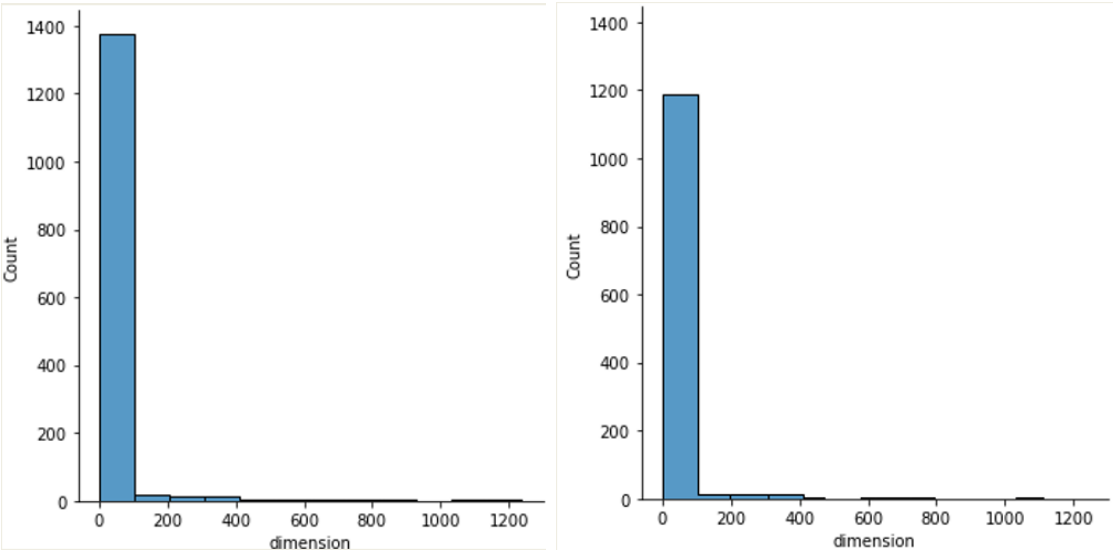


Figure 4.5: Facebook Research about data exchanged to start a session.

Figure 4.6: Flows grouped into sessions after applying the heuristic.

There are no relevant differences that could suggest excessive data loss or a wrong heuristic.



(a) Distribution of flows dimension before heuristic.

(b) Distribution of flows dimension after heuristic.

Figure 4.7: Differences in distributions of flows dimensions.

Chapter 5

Social Network Evolution

In this chapter we present the results obtained, analyzing how user habits have changed in different types of time patterns.

5.1 Results

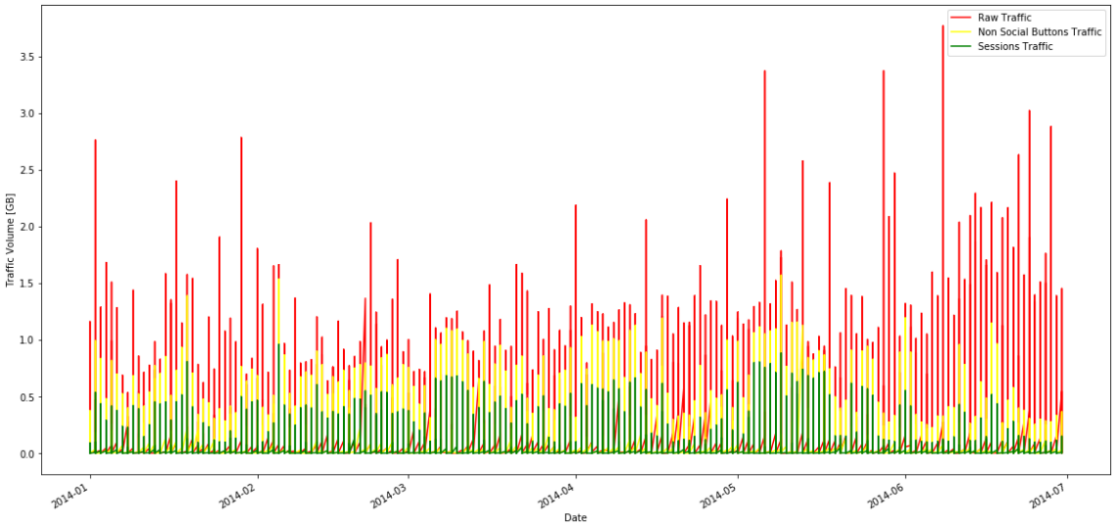
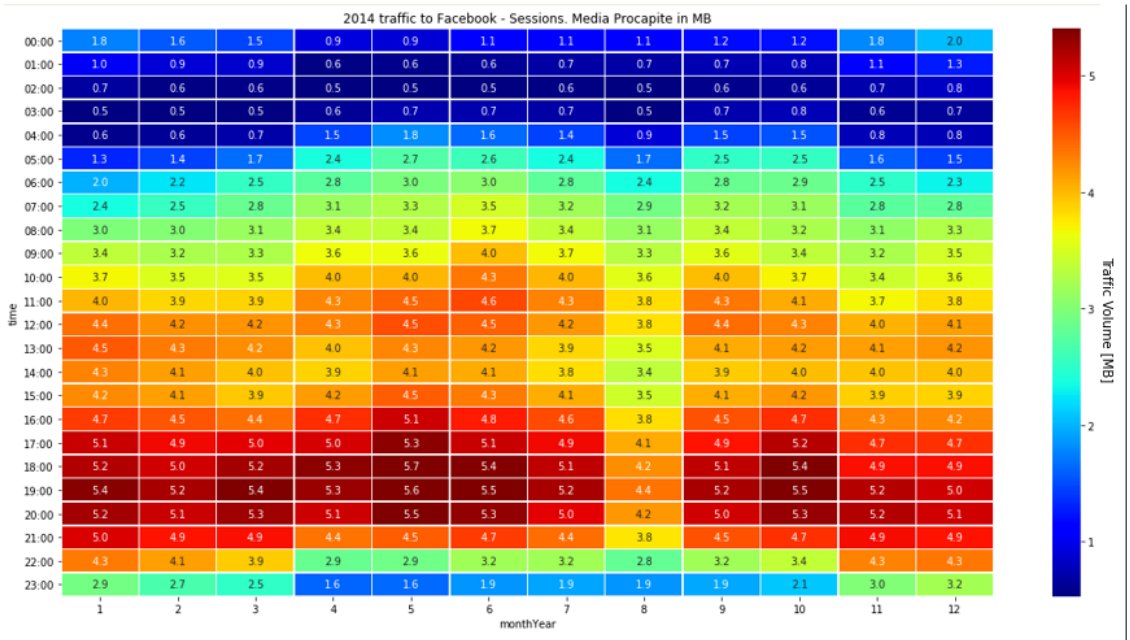


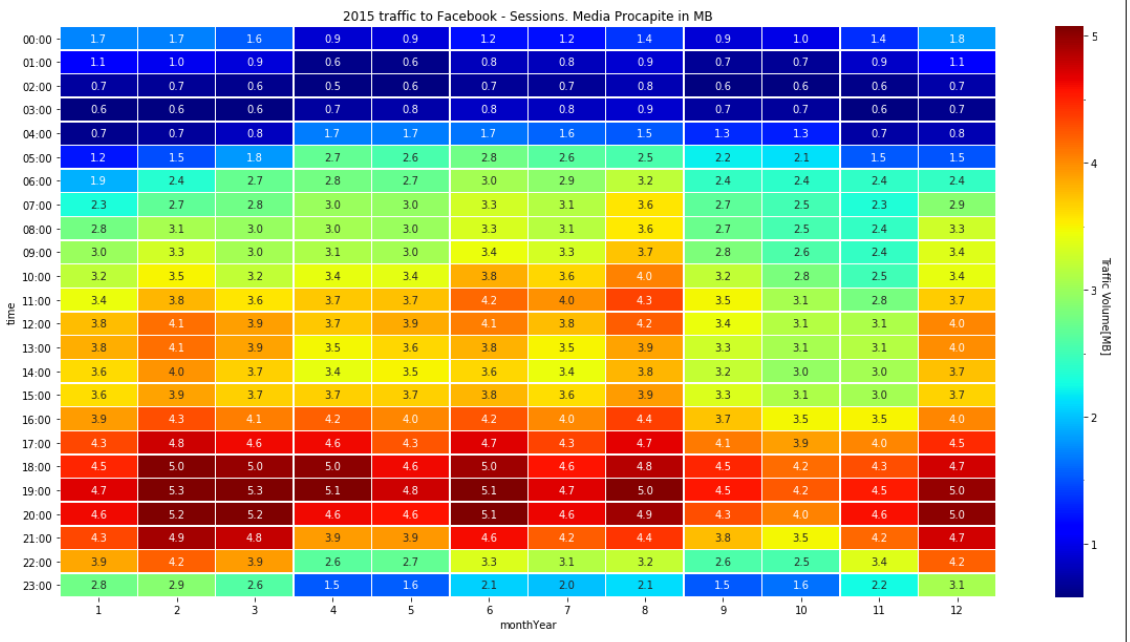
Figure 5.1: Differences in traffic to Facebook (Raw data, No social buttons data, Aggregated into sessions data).

After applying the heuristics we checked how the traffic volume changed.
The expected result was a smaller amount of traffic for each subsequent heuristic applied, as the amount of filtered traffic was increasing.

As mentioned in the previous paragraphs, the elimination of social buttons led to a reduction in the volume of overall traffic of about 26%, while the grouping into sessions filtered a further 5% of the traffic volume.

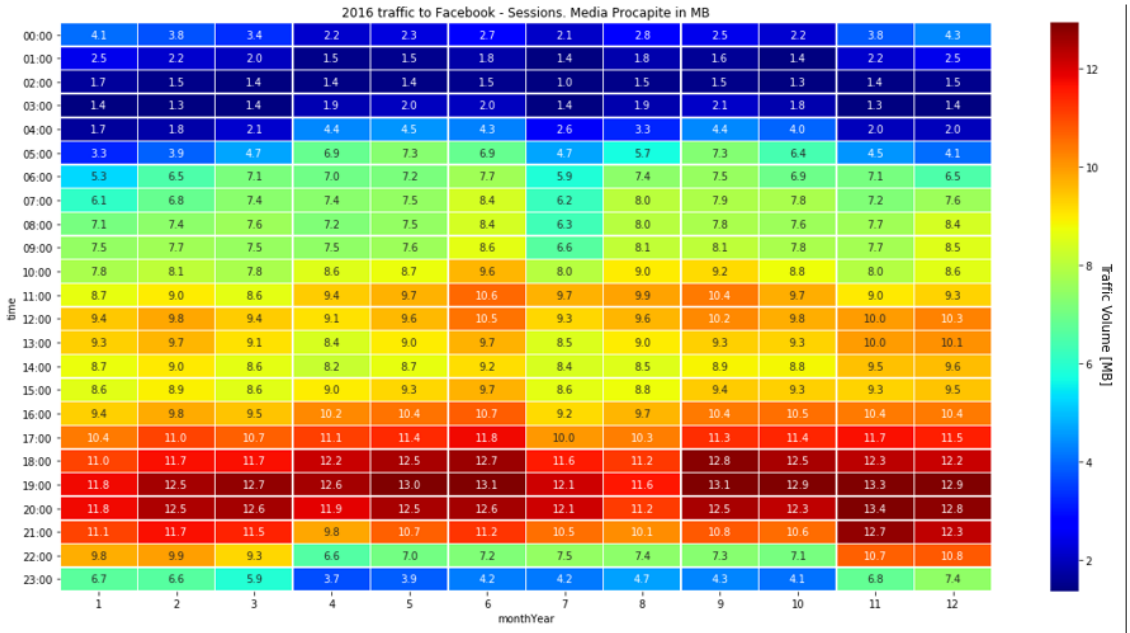


(a) Per capita average volume of sessions to Facebook in 2014.

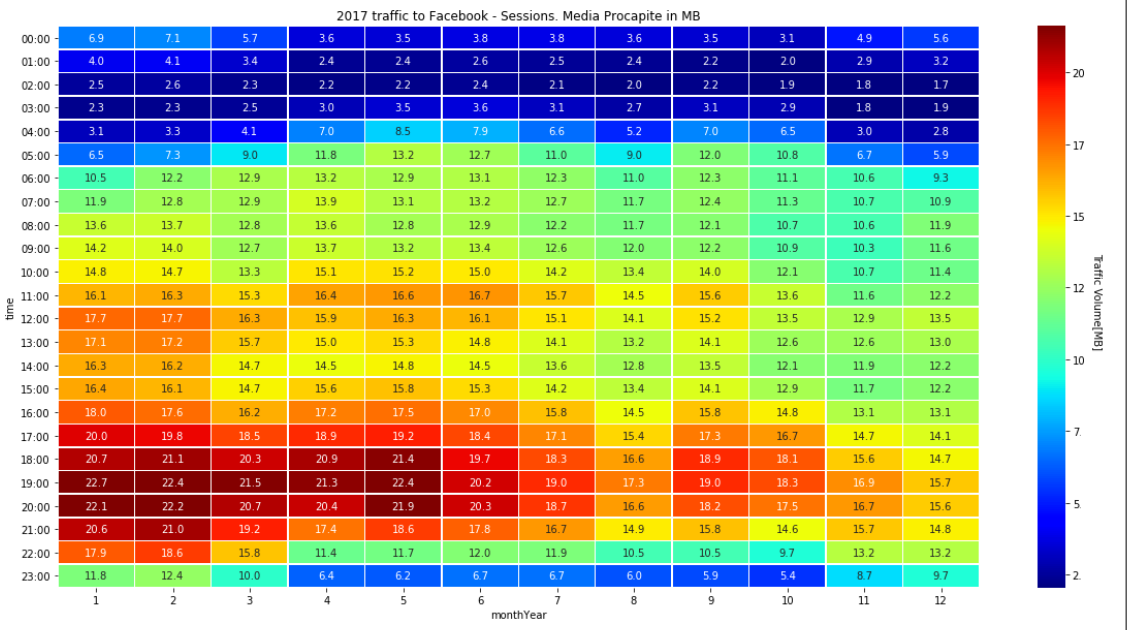


(b) Per capita average volume of sessions to Facebook in 2015.

Figure 5.2: Differences in average sessions volume over the years



(c) Per capita average volume of sessions to Facebook in 2016.



(d) Per capita average volume of sessions to Facebook in 2017.

Figure 5.2: Differences in average sessions volume over the years

Over the years, the average size of traffic over time in relation to the number of users increases, 2016 and 2017 appear to be the years with the greatest amount of "real" traffic to Facebook, with hourly values up to 7 times higher than 2013.

We have motivated this growth with the increase in the amount of media present in the platform, the increase in quality and consequently the weight of the media. Furthermore, Facebook also has heavily promoted video in its feed generally, reaching 8 billion daily video views per day in November 2015 (latest public data) [32].

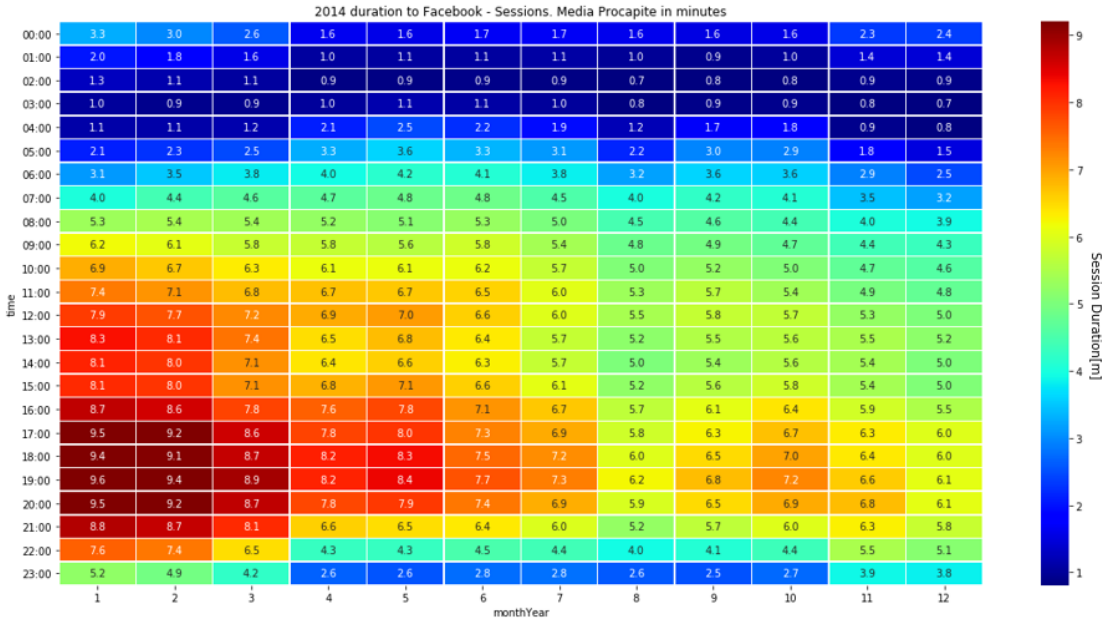
From a seasonal point of view, in 2013 the months with the most traffic were April and May with the peak traffic from 6pm to 10pm (50% / 70% of traffic more than the average-median month).
In 2014 the months with the most traffic were May and June with the peak traffic from 6 to 7 and from 20 to 22 (20% / 40% of traffic more than the average-median month).
In 2015 the months with the most traffic were January, February, March with the peak traffic from 6 to 7 and continuous in the remaining hours until 22 (20% more traffic than the average-median month).
In 2016, the months with the most traffic were from January to May with the first connection presumably from 6 to 7 (20% more traffic than the average-median month).
In 2017 the months with the most traffic were from January to May with the first connection presumably around 5/6 (20% / 30% of traffic more than the average-median month).
In 2018, having only the first 5 months, the data analyzed applying mean and median are not very significant.

Considering the amount of data exchanged in the years, we can conclude that the months in which the most data are exchanged are those ranging from January to May.

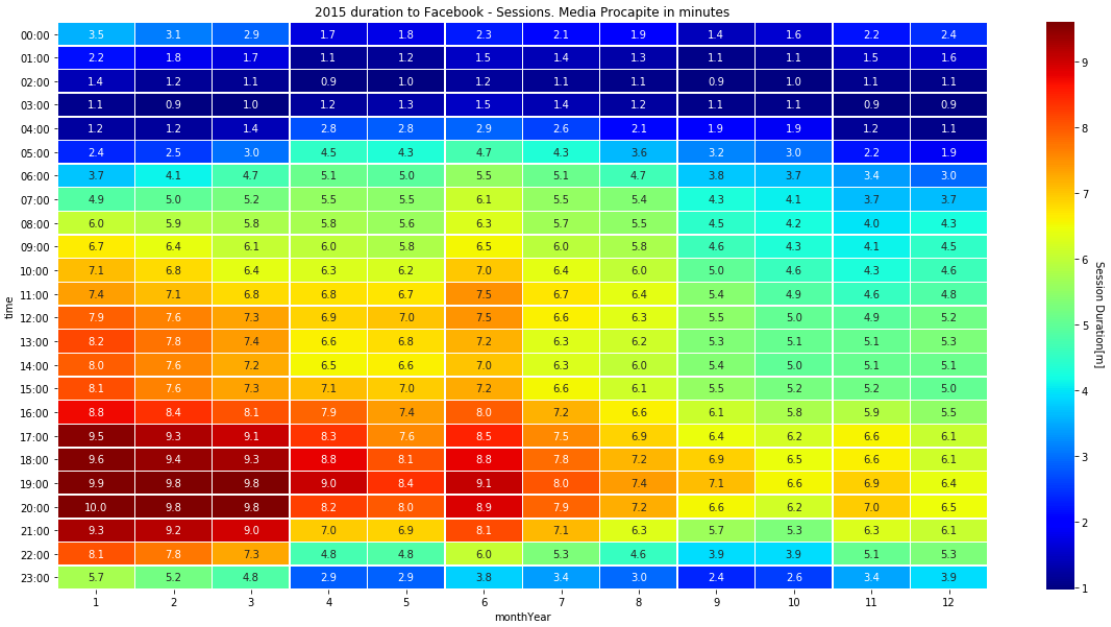
On the other hand, from the weekly time point of view, the traffic to Facebook in 2013 increases between 7am and 8am after the night.
The greatest traffic occurs from the 14th to the 28th week, from 8pm to 10pm.
The day with the most traffic is Monday.
Traffic to Facebook in 2014 increases between 7am and 8am after night time.
The greatest traffic occurs from the ninth week to the 29th, from 8pm to 10pm.
The day with the most traffic is Monday.
Traffic to Facebook in 2015 increases between 7am and 8am after night hours.
The greatest traffic occurs from the second week to the 24th, from 8pm to 10pm.
The day with the most traffic is Monday.
Traffic to Facebook in 2016 increases between 6am and 7am after night hours.
The greatest traffic occurs until June and after September, from 8pm to 10pm.

The day with the most traffic is Monday.
Traffic to Facebook in 2017 increases between 6am and 7am after night hours.
The greatest traffic occurs from the first to the 27th week, from 7pm to 10pm.
The day with the most traffic is Sunday.
Traffic to Facebook in 2018 increases between 6am and 7am after night hours.
The greatest traffic occurs from the 16th to the 19th week, from 7pm to 10pm.

From these data we conclude that 2016 saw a change in trend towards the first connection to Facebook, going from 7-8am of previous years to 6-7am starting from 2016.
From 2017, the peak of evening traffic occurs an hour earlier, from 7pm to 10pm, compared to the peak recorded since 2013 which occurred from 8 pm to 10 pm.
On average, the day with the highest traffic is Monday, since 2013.
In 2017, traffic on Sunday almost reaches the traffic of Monday.

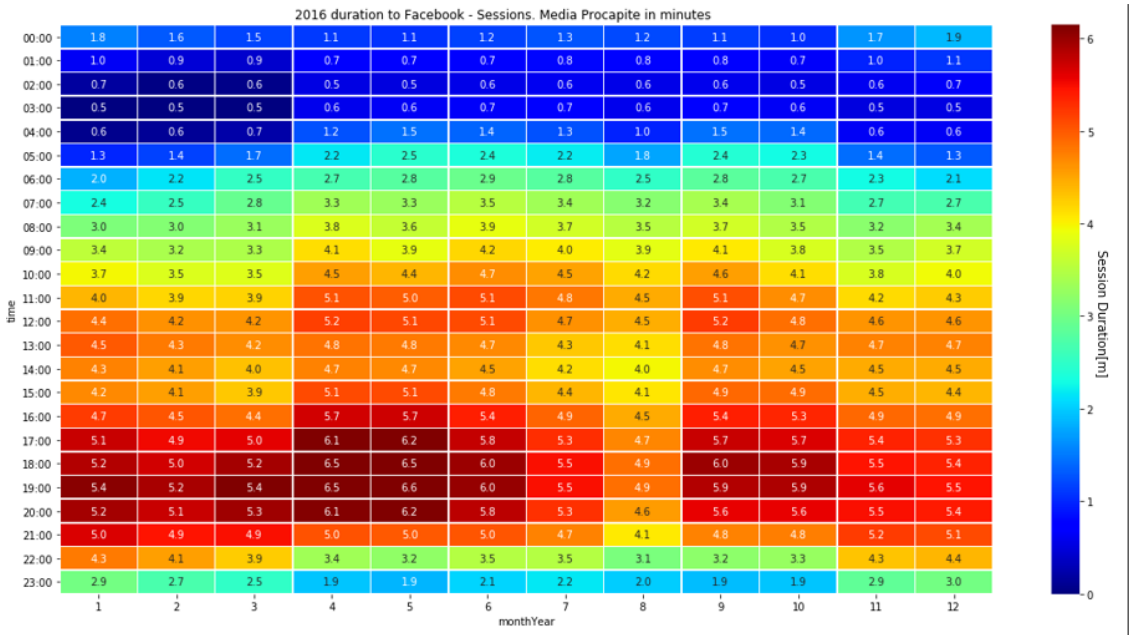


(a) Per capita average duration of sessions to Facebook in 2014.

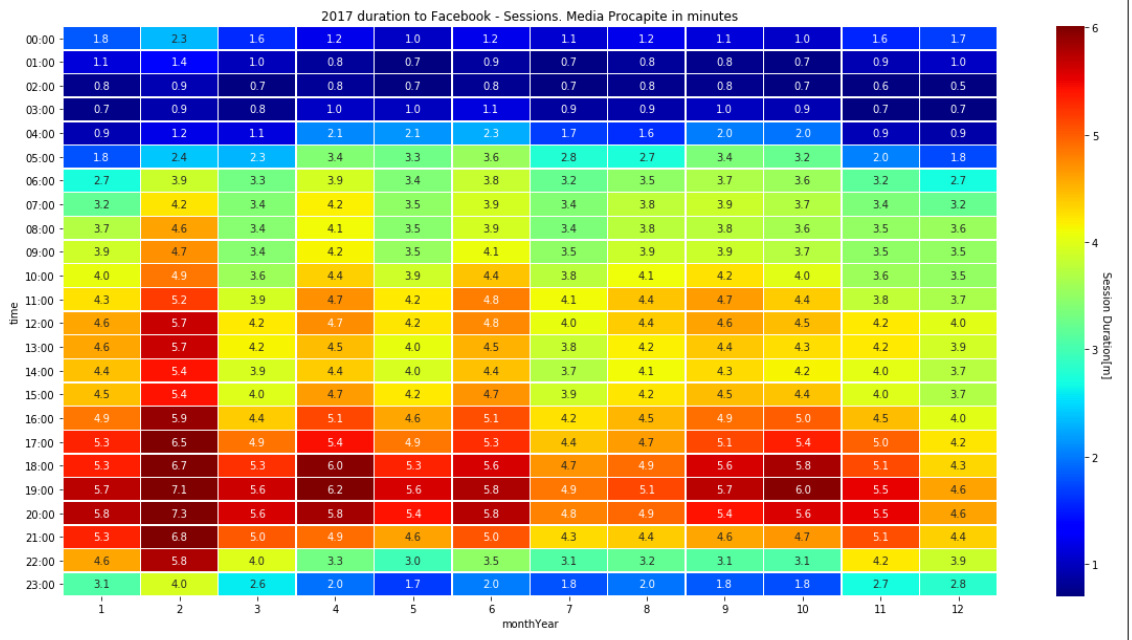


(b) Per capita average duration of sessions to Facebook in 2015.

Figure 5.3: Differences in average sessions duration over the years



(c) Per capita average duration of sessions to Facebook in 2016.



(d) Per capita average duration of sessions to Facebook in 2017.

Figure 5.3: Differences in average sessions duration over the years

While the average session volume increases over the years, the average session duration decreases slightly over time.
Although this behavior may be contradictory, it was a cue to deepen the analysis and hypothesize that it is an index of compulisiveness in visits, which are therefore shorter but with more traffic.

As for the annual traffic trend, the duration of the most "intense" sessions also extends from a time window initially between 4pm and 8pm, up to 10pm.
From a seasonal point of view, the months with the highest session duration are in the first half of the year, decrease in the summer and increase again at the end of the year.
This behavior, which may seem strange, was motivated by reasoning about the nature of the data we hold.
The data treated as mentioned above comes from a probe that detects residential data from Piedmont. During the summer, thanks to holidays and being away from home, most of the Facebook traffic is exchanged by mobile data or residential ADSL of holiday homes and not by residential ADSL in Piedmont.

Further analysis were carried out to understand how the first/last visit pattern has changed over time.
To analyze these patterns we analyzed user sessions at a smaller granularity, grouping by days of the week and weeks over the years.
In particular, referring to the applied heuristics and analyzing the sessions we noticed how the first session which occurred on average around 8AM in 2014, moved to 7AM in 2017, and the last view which in 2014 took place around 8PM and ended at 10PM, it moved at 7PM until 22PM.

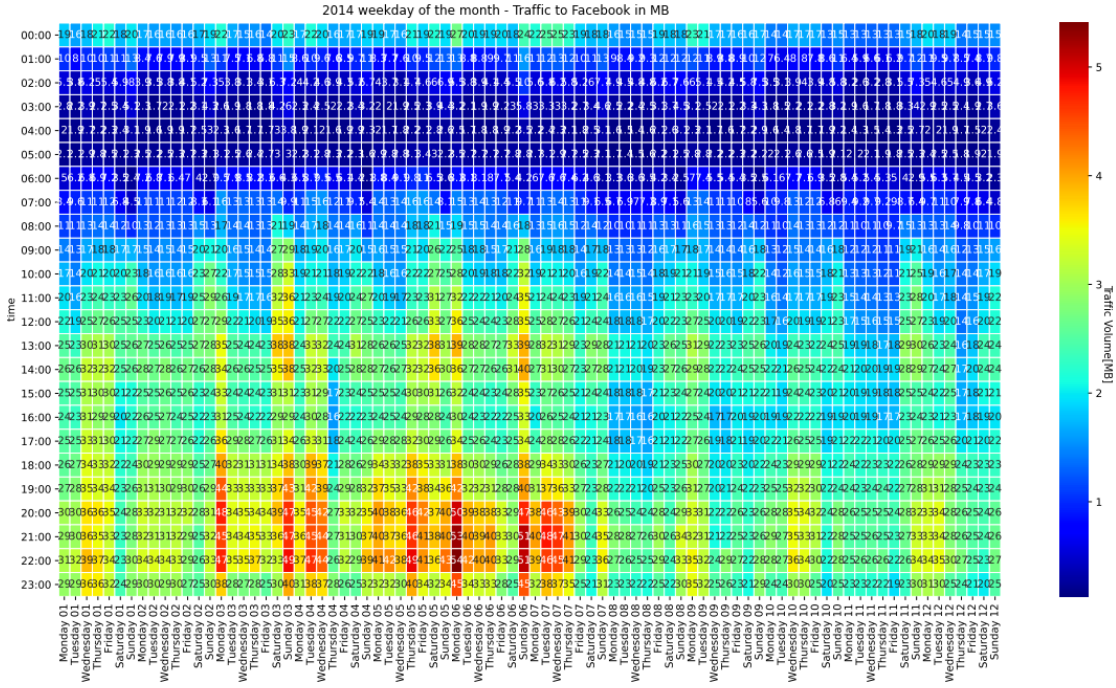
As mentioned previously, the data found suggests a compulsive use of Facebook. Although it is now a fact that Facebook causes addiction and compulsive use, it is interesting to prove it with the data we have [33] [34] [35].
We therefore analyzed the average number of sessions per user per day. The number of visits over the years increases, going from an average of 14 visits per day in 2013-2014, up to 17-18 visits in 2018.
This, combined with the results obtained previously, shows a compulsive form of visits to the platform.

The confirmation and verification of the hypotheses we have formulated and the results previously obtained is achieved by calculating the average distance between two successive sessions over the years and, showing how the monthly average moves.

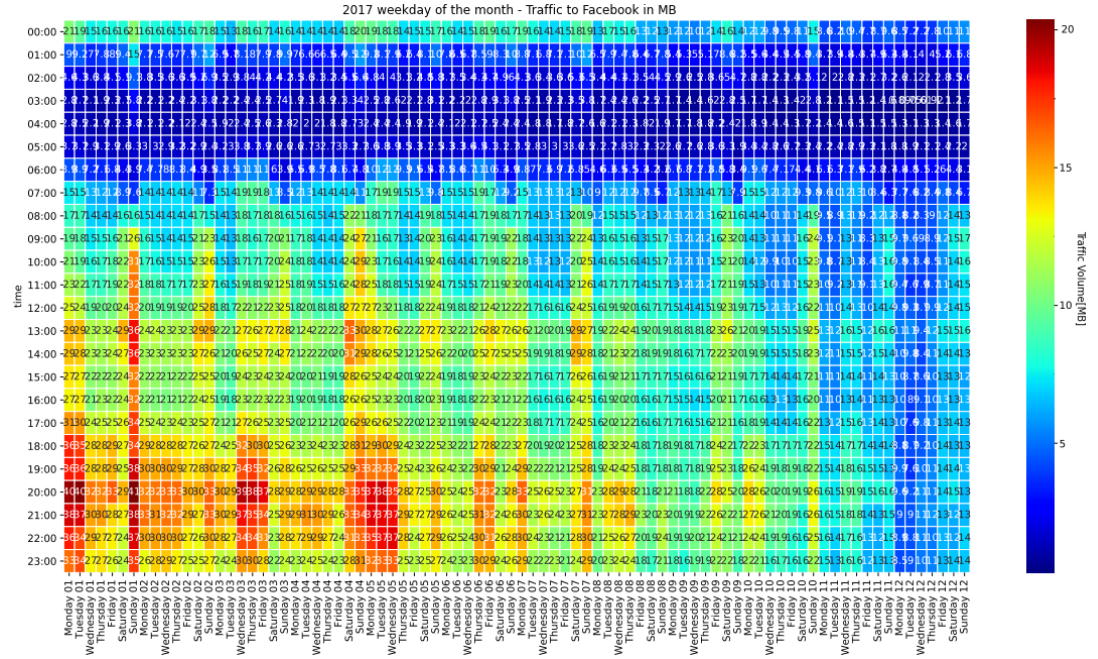
The average time between two sessions over the years decreases, this confirms that even from our data it is possible to detect what has been confirmed by the literature, that is to say that over the years Facebook users have developed a compulsive form of use of the platform, connecting more and more frequently, downloading more data, albeit with slightly shorter sessions.

All of our studies and research are largely confirmed by a series of research conducted by the Wall Street Journal [36] these days which based on internal research on Facebook documents, employee discussions and draft presentations by internal managers, show the main problems that the platform can bring to users.
An example of these problems is the very "toxic" addiction that many teen girls have towards Instagram.
Teenagers are developing problems like eating disorders or killing thoughts due to their time on Instagram, populated only by content that aims to showcase the best moments, and a pressure to look perfect.
More than the 40% of Instagram users are under the age of 22, and around 22 million young people log into the platform daily.
On average, a young person in America spends 50% more time on Instagram than on Facebook.

Teens said they wanted to spend less time on the platform, but many failed



(a) Per capita average volume of session grouped per day of month in 2014.



(b) Per capita average volume of session grouped per day of month in 2017.

Figure 5.4: Differences between first and last connection on average sessions - Fine granularity.

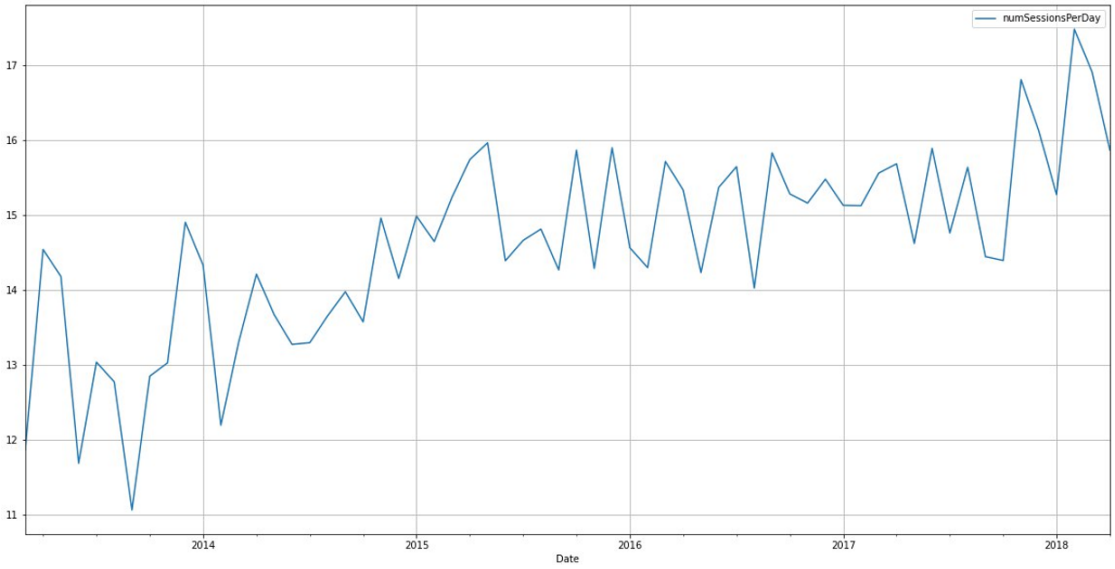


Figure 5.5: Per user average number of sessions per day during years.

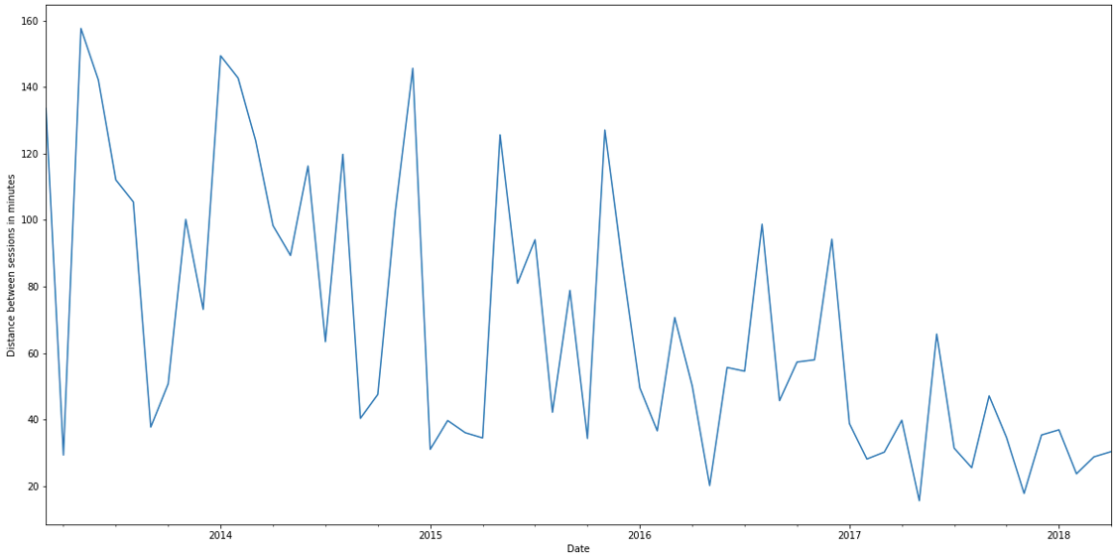


Figure 5.6: Per user average distance between sessions during years.

due to lack of self-control in doing so.
A manager of the Instagram researchers told his colleagues that, according to the documents, users often feel "addicted" and know that what they watch hurts them, but they do not feel able to stop as in some kind of addiction to a drug.

Chapter 6

Conclusions

This chapter summarizes the thesis work done, starting from the beginning up to the conclusions.
Possible points of reflection are also proposed regarding possible future works related to the result or the procedure of the thesis.

6.1 Findings

By analyzing the large quantities of TCP logs of Facebook users, it was possible to identify the social network flows, divide the involuntary visits from the volunteers and aggregate the voluntary flows into sessions per user.

After defining some heuristics for the data analysis we obtained several result contributions.
From 2013 to 2018, the number of visits increases over time.
The average size per visit also increases, although the average length of visits decreases slightly.
We also saw how the first visit time was brought forward by about an hour during the day and the time window of maximum use was extended by about two hours.
Finally, we have confirmed on our data several social and psychological studies that have found in Facebook a form of addiction that causes compulsive use.
In fact, we have seen how the number of visits per day increases and the time interval between two subsequent visits decreases.

6.2 Future works

The usefulness of the thesis with regard to future works can be discussed from two different points of view.
The first concerns the results obtained. In fact, further studies of a social-psychological nature can be carried out on the data grouped into sessions, as well as predictions on the future of the use of Facebook.
Another point of view is that concerning the study procedure used and the heuristics. The same process that we applied, unless the infrastructure/protocols of Facebook changes, can be used for new datasets or future Facebook data.

Furthermore, with the same dataset, one could think of analyzing the traffic of other social networks and see how they have behaved over the years compared to Facebook.
Obviously, for the analysis of other platforms, the heuristics would have to be changed, however the analysis procedure used would remain valid.

Bibliography

[1] Jonathan Magnusson. *Social Network Analysis Utilizing Big Data Technology*. 2012 (cit. on p. ii).

[2] Godefroy Dang Nguyen and Virginie Lethiais. «Impact des réseaux sociaux sur la sociabilité». In: *Réseaux* n° 195.1 (2016), p. 165. DOI: 10.3917/res.195.0165. URL: http://dx.doi.org/10.3917/res.195.0165 (cit. on p. ii).

[3] Worldwide. GlobalWebIndex. Statista. *Most popular reasons for internet users worldwide to use social media as of 3rd quarter 2020*. Dec. 2020. URL: https://www.statista.com/statistics/715449/social-media-usage-reasons-worldwide/ (cit. on p. 1).

[4] Nick G. *Social Media Marketing Statistics and Trends to Know in 2020*. Feb. 2021. URL: https://review42.com/resources/how-much-time-do-people-spend-on-social-media/ (cit. on p. 1).

[5] Adams Oluwadamilola Kemi. «Impact of Social Network on Society: A Case Study of Abuja». In: *American Scientific Research Journal for Engineering, Technology, and Sciences* 21 (2016), pp. 1–17 (cit. on p. 2).

[6] Kirk Kristofferson, Katherine White, and John Peloza. «The Nature of Slacktivism: How the Social Observability of an Initial Act of Token Support Affects Subsequent Prosocial Action». In: *Journal of Consumer Research* 40.6 (2014), pp. 1149–1166. ISSN: 00935301, 15375277. URL: http://www.jstor.org/stable/10.1086/674137 (cit. on p. 2).

[7] J. BAXTER OLIPHANT AMY MITCHELL MARK JURKOWITZ and ELISA SHEARER. *Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable*. July 2020. URL: https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/ (cit. on p. 2).

[8] H. Tankovska. Statista. *Number of monthly active Facebook users worldwide as of 1st quarter 2021(in millions)*. Apr. 2021. URL: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (cit. on p. 2).

[9] Ivory Sherman Andrew Greiner Seth Fiegerman and CNN Business Tiffany Baker. *FACEBOOK AT 15: HOW A COLLEGE EXPERIMENT CHANGED THE WORLD*. Feb. 2019. URL: https://edition.cnn.com/interactive/2019/02/business/facebook-history-timeline/index.html (cit. on p. 2).

[10] MICHAEL LaFORGIA GABRIEL J.X. DANCE NICHOLAS CONFESSORE. *Facebook Gave Device Makers Deep Access to Data on Users and Friends*. June 2018. URL: https://www.nytimes.com/interactive/2018/06/03/technology/facebook-device-partners-users-friends-data.html (cit. on p. 3).

[11] Farshad Kooti, Karthik Subbian, Winter Mason, Lada Adamic, and Kristina Lerman. «Understanding Short-term Changes in Online Activity Sessions». In: (Apr. 2017). URL: https://research.fb.com/publications/understanding-short-term-changes-in-online-activity-sessions/ (cit. on pp. 4, 5, 32).

[12] Brandon Schlinker, Ítalo S. Cunha, Yi-Ching Chiu, S. Sundaresan, and Ethan Katz-Bassett. «Internet Performance from Facebook's Edge». In: *Proceedings of the Internet Measurement Conference* (2019) (cit. on pp. 4, 6, 32).

[13] Daqing He, Daqing, Göker, and Ayse Goker. «Detecting Session Boundaries from Web User Logs». In: Jan. 2000 (cit. on p. 6).

[14] Anindita Chakraborty. «Facebook Addiction: An Emerging Problem». In: *American Journal of Psychiatry Residents' Journal* 11.12 (2016), pp. 7–9. DOI: 10.1176/appi.ajp-rj.2016.111203. eprint: https://doi.org/10.1176/appi.ajp-rj.2016.111203. URL: https://doi.org/10.1176/appi.ajp-rj.2016.111203 (cit. on p. 7).

[15] Giuseppe Riva, Brenda Wiederhold, and Pietro Cipresso. *The Psychology of Social Networking Vol.1. Personal Experience in Online Communities*. Aug. 2016. ISBN: 9783110473780. DOI: 10.1515/9783110473780 (cit. on p. 7).

[16] Tracii Ryan, Andrea Chester, John Reece, and Sophia Xenos. «The uses and abuses of Facebook: A review of Facebook addiction». In: *Journal of Behavioral Addictions JBA* 3.3 (2014), pp. 133–148. DOI: 10.1556/jba.3.2014.016. URL: https://akjournals.com/view/journals/2006/3/3/article-p133.xml (cit. on p. 7).

[17] M.T. Thai, W. Wu, and H. Xiong. *Big Data in Complex and Social Networks*. Dec. 2016, pp. 1–242. ISBN: 9781315396699. DOI: 10.1201/9781315396705 (cit. on p. 8).

[18] Nadir Zanini Vikas Dhawan. *Big data and social media analytics*. URL: https://www.cambridgeassessment.org.uk/Images/465808-big-data-and-social-media-analytics.pdf (cit. on p. 8).

[19] Sepideh Bazzaz Abkenar, Mostafa Haghi Kashani, Ebrahim Mahdipour, and Seyed Mahdi Jameii. «Big data analytics meets social media: A systematic review of techniques, open issues, and future directions». In: *Telematics and Informatics* 57 (Mar. 2021), p. 101517. DOI: 10.1016/j.tele.2020.101517. URL: http://dx.doi.org/10.1016/j.tele.2020.101517 (cit. on p. 9).

[20] Martino Trevisan, Alessandro Finamore, Marco Mellia, Maurizio Munafo, and Dario Rossi. «Traffic Analysis with Off-the-Shelf Hardware: Challenges and Lessons Learned». In: *IEEE Communications Magazine* 55 (Mar. 2017), pp. 163–169. DOI: 10.1109/MCOM.2017.1600756CM (cit. on p. 10).

[21] Facebook. *Facebbok for Developers - Plug-in social*. URL: https://developers.facebook.com/docs/plugins/ (cit. on p. 17).

[22] Pedro Miranda, Matti Siekkinen, and Heikki Waris. «TLS and energy consumption on a mobile device: A measurement study». In: *2011 IEEE Symposium on Computers and Communications (ISCC)*. 2011, pp. 983–989. DOI: 10.1109/ISCC.2011.5983970 (cit. on p. 18).

[23] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995 (cit. on p. 22).

[24] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697 (cit. on p. 23).

[25] Thomas Kluyver et al. «Jupyter Notebooks – a publishing format for reproducible computational workflows». In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90 (cit. on p. 23).

[26] Wes McKinney et al. «Data structures for statistical computing in python». In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56 (cit. on p. 24).

[27] Matei Zaharia et al. «Apache spark: a unified engine for big data processing». In: *Communications of the ACM* 59.11 (2016), pp. 56–65 (cit. on p. 24).

[28] Jana Iyengar and Martin Thomson. *QUIC: A UDP-Based Multiplexed and Secure Transport*. RFC 9000. May 2021. DOI: 10.17487/RFC9000. URL: https://rfc-editor.org/rfc/rfc9000.txt (cit. on p. 31).

[29] Subodh Iyengar and Kyle Nekritz. *Building zero protocol for fast, secure mobile connections (2017)*. URL: https://engineering.fb.com/2017/01/27/android/building-zero-protocol-for-fast-secure-mobile-connections/ (cit. on p. 31).

[30] Yang Chi Matt Joras. *How Facebook is bringing QUIC to billions*. URL: https://engineering.fb.com/2020/10/21/networking-traffic/how-facebook-is-bringing-quic-to-billions/ (cit. on p. 31).

[31] *Facebook Engineering*. URL: https://engineering.fb.com/ (cit. on p. 32).

[32] *The rise (and fall) of autoplay video*. URL: https://digiday.com/media/state-of-autoplay-video/ (cit. on p. 39).

[33] Matthew KO Lee Christy MK Cheung Zach WY Lee. «Understanding Compulsive Use Of Facebook Through The Reinforcement Processes». In: *ECIS 2013 Proceedings at AIS Electronic Library (AISeL)* (2013) (cit. on p. 44).

[34] Anindita Chakraborty. «Facebook Addiction: An Emerging Problem». In: *American Journal of Psychiatry Residents' Journal* 11.12 (2016), pp. 7–9. DOI: 10.1176/appi.ajp-rj.2016.111203. eprint: https://doi.org/10.1176/appi.ajp-rj.2016.111203. URL: https://doi.org/10.1176/appi.ajp-rj.2016.111203 (cit. on p. 44).

[35] Tracii Ryan, Andrea Chester, John Reece, and Sophia Xenos. «The uses and abuses of Facebook: A review of Facebook addiction». In: *Journal of Behavioral Addictions JBA* 3.3 (2014), pp. 133–148. DOI: 10.1556/jba.3.2014.016. URL: https://akjournals.com/view/journals/2006/3/3/article-p133.xml (cit. on p. 44).

[36] *The Facebook Files A Wall Street Journal investigation*. URL: https://www.wsj.com/articles/the-facebook-files-11631713039?mod=bigtop-breadcrumb (cit. on p. 44).