Aalto University

CS-C4100 Digital Health and Human Behavior

Final Project: Network Study

Introduction

Datasets with an extensive array of attributes, derived from prolonged studies, are exceedingly uncommon. This rarity primarily stems from the formidable challenges associated with conducting such comprehensive investigations. However, the scarcity of such datasets implies that when one is successfully compiled, it becomes an invaluable resource for researchers.

In the years 2012 and 2013, the SensibleDTU project undertook the creation of a dataset enriched with attributes, involving approximately 1,000 participants [1]. Detailed in the section Dataset Description at page 4, this dataset predominantly delves into social behavior and aims to provide exhaustive participant descriptions. The attribute-rich nature of this dataset not only enables the exploration of a diverse range of inquiries but has also garnered considerable attention from researchers. Since its inception, scholars have delved into various research domains, encompassing human mobility, epidemiology, friendship dynamics, gender studies, and numerous other subjects.

Relevant Literature

This incredibly rich dataset has been explored by a number of researchers. Past publications approach the dataset from a variety of angles, asking questions about the nature of social behavior within a dense social environment. Some of the explored fields are presented below.

Human Mobility

The paper titled *"Tracking Human Mobility Using WiFi Signals"* focuses on WiFi data gathered from 63 selected participants [2]. Initially, the paper employs this data to identify WiFi routers. Leveraging this newly acquired information, researchers utilize the data from participant connections to each WiFi access point, effectively accounting for 80 % of mobility within the observed population. The paper suggests that the abundance of WiFi routers in the modern era enables the scalability of this outdoor positioning method across a broader population.

Another paper combines datasets from the Copenhagen Networks Study, The Reality Mining project, the Lifelog dataset, and the Mobile Data Challenge dataset conducted by the Lausanne Data Collection Campaign to explore human mobility patterns [3]. While conducting various analyses on the datasets, a key finding indicates that human mobility patterns undergo significant changes over time, but this change is relatively constant. This implies that the number of familiar locations an individual visits at any given time is a conserved quantity, approximately 25 locations. This conclusion contradicts previous scientific community findings regarding human mobility.

An additional study exclusively utilizes WiFi data from the Copenhagen dataset to analyze human mobility patterns using the concept of *stop-locations* [4]. A stop-location includes two timestamps indicating start and end times, along with a location. The paper introduces two innovative methods for inferring stop-locations based solely on WiFi data. The study concludes by validating the results of the two algorithms using an established method that relies on GPS data for inferring stop-locations.

Yet another paper from the dataset proposes a unique method for inferring person-to-person proximity using only WiFi data [5]. The authors aim to compare the lists of routers observed by two study participants simultaneously to determine if they were physically close. The paper concludes that WiFi availability lists over time can indeed serve as a strong indicator of physical proximity. The authors argue that this finding could potentially eliminate the need for future studies to include Bluetooth-enabled devices for gathering physical proximity data.

Friendship Studies

In various scenarios, friendships emerge as the most robust social connections among individuals. Consequently, directing analytical efforts towards these relationships has the potential to unveil more about an individual than studying all interactions indiscriminately. Researchers leveraged the Copenhagen Networks Study dataset to explore the accuracy of modeling or predicting friendships.

Shortly after the Copenhagen Networks Study, a paper utilized data from proximity sensors in smartphones to derive insights about friendships within participants' social groups [6]. Focused on 134 students during the academic year 2012-2013 over 119 days, the authors employed Bluetooth sensors to record devices within ten meters every five minutes. Signal strength data was used to estimate distances during face-to-face interactions. The challenge addressed was that physical proximity alone does not guarantee social interaction. The authors proposed a method to distinguish between social and non-social face-to-face interactions.

In a recent study titled *"Offline Behaviors of Online Friends"*, the Copenhagen Networks Study was analyzed to assess the accuracy of reconstructing the Facebook friendship graph, Facebook interaction network, and call/text message networks [7]. This study encompassed data from all 1000 participants over a full year. The researchers identified subtle behavioral differences between pairs interacting through various channels. They delved into the meaning of Facebook friendship status in relation to phone calls, text messages, and online interactions. Ultimately, the results demonstrated the feasibility of determining genuine friendship based on interaction data, despite potential noise from offline interactions among familiar strangers.

Gender Studies

As highlighted in the Available Data section on page 4, this dataset stands out due to the predominance of male participants. The significant disparity in the number of males compared to females in the study is reflective of the broader university population. Consequently, researchers have explored questions about social behavioral distinctions between males and females within this gender-imbalanced environment.

A specific paper investigates the possibility of distinguishing between male and female participants based on behavior, along with identifying the most predictive behavioral features [8]. This study utilizes data collected from participants' mobile phones and self-filled questionnaires. The authors observe personality trait differences consistent with contemporary psychology literature on gender disparities. However, their findings regarding human mobility challenge existing literature, revealing that, contrary to previous work, females tend to travel more on average than males. In terms of social networking, the study observes that females, on average, communicate more than males, though both genders exhibit similar levels of homophily, indicating a tendency to associate more with one's own gender. The analysis of the participants' Facebook social network reveals that females tend to occupy more central positions. Ultimately, the study utilizes these features to predict student gender, achieving a reasonably accurate model.

Another study uses data from the Copenhagen Networks Study to explore the potential to predict academic performance in a gender-imbalanced environment [9]. In the dataset subset used for this study, comprising 420 males and 120 females, the authors assert that this distribution aligns with the gender imbalance in the overall student population. The study finds that social indicators, such as the mean grade point average of peers or the fraction of low-performing peers, more accurately predict low performance in male participants compared to female participants.

Project Objectives

This project delves into the intricate dynamics of student social interactions through a comprehensive analysis of various communication channels—Bluetooth scans, phone conversations, SMS messages, and Facebook friendships. **The objective is to unravel patterns, correlations, and gender-based nuances within the social fabric of the student community**. The exploration begins by scrutinizing the daily count of Bluetooth scans, shedding light on potential periodic patterns in social interactions. Simultaneously, the network representing student connections is examined, providing insights into the structure and dynamics of these relationships.

Motivating the endeavor is a keen interest in understanding the intricacies of student interactions and communication preferences. **One central research question addresses how various communication channels contribute to the social network of students, with a focus on patterns, gender dynamics, and preferred modes of communication**.

Utilizing the Python programming language, specifically leveraging libraries such as `NetworkX`, `Pyvis`, `Numpy`, `Pandas`, `Matplotlib`, and `Seaborn`, a rigorous analysis approach has been employed. Degree distribution serves as a crucial tool in unraveling network structures, offering a comprehensive understanding of connectivity patterns.

The study explores interactions between male and female students, shedding light on notable aspects. Furthermore, it is confirmed that the Facebook friendships network exhibits a scale-free nature, and messages are identified as the preferred communication tool between students of the opposite gender.

The analyses conducted in this project have indeed demonstrated **the presence of periodic patterns in social interactions detected through Bluetooth scans**. Moreover, the differences between more and less crowded days are clearly visible when analyzing the network. Additionally, the **existence of correlations between the number of exchanged messages and calls has been confirmed**.

Conducting analyses while considering the gender of the involved students revealed that, **despite females being in the numerical minority, they have made a significant contribution to communication with students, highlighting specific differences in the analyzed networks**.

As expected, the network of Facebook Friendships exhibited characteristics typical of a scale-free network. Finally, it was observed that messages were the preferred communication tool among students of the opposite gender.

Problem Formulation

As mentioned earlier, numerous studies have been conducted on the Copenhagen Networks Study, each focusing on different aspects of the dataset. In this project, an attempt was made to address various questions:

Firstly, by analyzing the daily count of Bluetooth scans, we aimed to textbfhighlight any periodic patterns in social interactions, as well as potential differences in the network representing student connections.

Regarding the data extracted from phone conversations, we analyzed the textbfdifferences between the distributions of outgoing and incoming calls. In particular, efforts were made to highlight the main differences for each student between the calls made and received.

Concerning the data obtained from SMS messages, the degree distribution of the network representing messages exchanged by students was extracted. In particular, attempts were made to textbfextract possible correlations between the number of calls and messages exchanged over the four weeks of observations.

Subsequently, efforts were made to **understand whether the disparities between the number of male and female students had in any way influenced the number of messages exchanged for male and female students**. Additionally, in both the case of calls and messages, an analysis was conducted on how diversified the interactions with other students were, attempting to highlight any patterns for both male and female students.

An attempt was also made to analyze whether male students, as well as female students, prefer to converse on the phone with students of the same gender or the opposite gender.

Regarding the network of friendships on Facebook, particular attention was given to textbfthe degree distribution to validate the hypothesis that it was a scale-free network, as is the case with most social networks.

Finally, taking into consideration the interactions between SMSs, calls, and Facebook, an effort was made to textbfunderstand the preferred tool for communication between students of the opposite gender.

Dataset Description

General Overview

Creating an attribute-rich longitudinal dataset for social network analysis poses considerable challenges. Coordinating the participation of a substantial number of individuals is already demanding within shorter timeframes and becomes progressively more intricate over extended periods. A common strategy to address these challenges involves amalgamating datasets from diverse sources, artificially constructing a dataset that encompasses both longevity and the desired attributes.

While such combined datasets can offer utility, the ideal scenario involves the collection of a singular dataset from a unified source, encompassing the necessary attributes and temporal scope for deriving meaningful conclusions. In pursuit of this goal, the SensibleDTU project was initiated [1]. **This project aimed to gather social data from hundreds of students at The Technical University of Denmark (DTU) over an extended duration, ultimately culminating in the creation of the Copenhagen Networks Study**.

In the two iterations of the SensibleDTU study conducted in 2012 and 2013, an extensive volume of data was amassed from around 1000 participants, all of whom were students at the same university. Given their shared university affiliation, these participants exhibited high social density and frequent close proximity in their regular interactions.

Participants were administered comprehensive questionnaires encompassing various personality assessments. Additionally, periodic snapshots of Facebook data were collected for each participant, including details such as birthday dates, education history, friends lists, interests, likes, and political views. The dataset also incorporated school-related information like class schedules and academic performance. To monitor day-to-day social behavior, each participant was provided with a smartphone specially equipped to track activities such as WiFi connectivity on the DTU campus, phone calls, text messages, and close physical proximity to other devices in the study.

The granularity of the dataset varied depending on the attribute. For instance, Facebook data was scanned as a snapshot for each participant every 24 hours, while Bluetooth scans occurred on the study-provided mobile phones every five minutes to collect proximity data. Texts or phone calls sent or received on these devices were logged, including details like sender, receiver, timestamp in seconds, and whether the call was missed, along with the duration of the call.

The campus WiFi system was leveraged to track participants' locations, providing data every ten minutes about all devices connected to wireless access points on campus. This data included the MAC address of the access point, the building location of the access point, and student IDs connected to that access point. Student IDs were anonymized by associating them with participant IDs from the study.

The comprehensive dataset also encompassed an anthropological study involving a randomly selected group of approximately sixty students. Qualitative data collected during group activities, including group work, parties, trips, and other social events, aimed to explore the formation of different groups under varying conditions.

Available Data

While the complete dataset is vast, it is important to specify what data were accessible for this project [13]. The available data include the Bluetooth scans every five minutes as well as the signal strength (RSSI) of each scan. There are also gender, phone call and text message logs, and one snapshot of the Facebook friends network that was taken during the four week period.

As indicated in figure 1, the dataset comprises a total of 848 students, and it includes 614 males and 173 females, while 61 students didn't disclose their gender. Table 1 indicates total the number

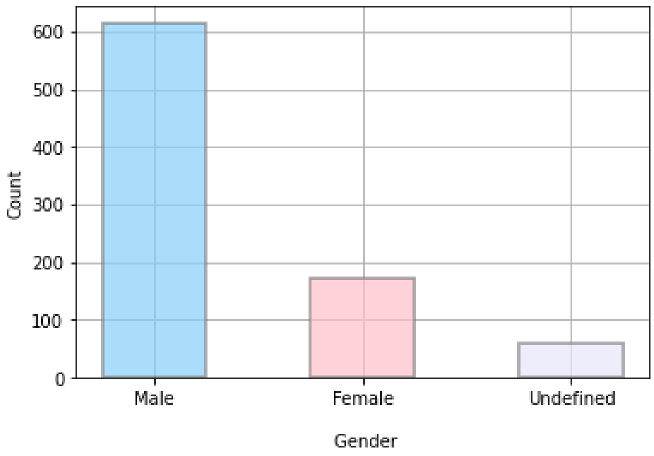

Figure 1: Gender Distribution

of data points between participants in the study, for each attribute.

	Text Messages	Phone Calls	Bluetooth Scans	Facebook Friendships
Count	24,333	3,234	2,426,279	6,429

Table 1: Quantity of Data

Table 2 indicates the number of students with at least one data point for each attribute. From the four weeks of data, a reasonable number of Bluetooth scans and Facebook friendships is observed, considering the number of participants. However, **the number of text messages and phone calls seems unusually low**.

	Text Messages	Phone Calls	Bluetooth Scans	Facebook Friendships
Count	568	525	824	800

Table 2: The Number of Students with At Least One Data Point

Considering the overall number of participants in the study, it's somewhat surprising that **only 568 of them sent at least one text message and 525 participated in at least one phone call**. One might have anticipated a higher volume of electronic communication among students during 2012 and 2013. This unexpected finding could be attributed to the prevalence of instant messaging platforms like WhatsApp in the specific region where the study occurred. Another possible explanation is that students may still own and regularly use additional mobile devices not covered by the study. While these sections of the dataset may not be as comprehensive as initially hoped, they nonetheless offer valuable insights for this project.

Methods

To analyze the data, the `Python` programming language was employed. The analyzed dataset did not contain N/A values and outliers. The only cleaning process was related to bluetooth empty scans denoted by user B=-1 and RSSI=0, as well as scans of devices situated outside the experiment marked by user B = -2, which have been eliminated from the dataset. Additionally, from the call dataset, all calls in which one of the two users did not answer, marked with duration=-1, were removed. Specifically, for a more accurate depiction of the networks, the degree distribution was considered, illustrating the distribution of the number of connections (degrees) that nodes in the network possess. This information reveals how many nodes have a specific number of connections.

The primary reason for utilizing the degree distribution is its capability to precisely portray a network, unveiling the overall structure, aiding in identifying network types, highlighting influential nodes (hubs), and predicting network behavior and resilience.

The main libraries employed for data analysis and visualization were `numpy`, `pandas`, `matplotlib`, and `seaborn`. Regarding network analysis and visualization, three specific libraries were utilized:

- NetworkX is a Python package designed for creating, manipulating, and studying the structure, dynamics, and functions of complex networks.
- Pyvis is valuable for network visualization due to its user-friendly interface, allowing the creation of interactive and visually appealing graphs (even though in this report the images are in png format, preventing interactive visualization of the network). It supports customization, provides dynamic updates, and facilitates easy sharing of visualizations in HTML format. This library is particularly useful for exploring and presenting complex relationships within networks.
- Plotly is a Python graphing library known for creating interactive and versatile visualizations. Its key strengths lie in interactivity, ease of use, support for various chart types, online collaboration, and integration with Dash for building interactive web applications.

Results

Results Discussion

Initiating the analysis of Bluetooth scans, a comprehensive examination of the daily scan count across 28 days was carried out to identify potential periodic patterns, which were indeed discerned, as showed in figure 2. Significantly, **more scans were recorded on weekdays (Monday to Friday) compared to the weekend (Saturday, Sunday)**. This conspicuous pattern can be logically attributed to the absence of classes during the weekend, creating an environment where weekdays become more conducive to heightened social interactions. This logical correlation seamlessly aligns with the primary focus of the dataset, which centers around monitoring the social dynamics of students at DTU.

In an effort to gain a more nuanced understanding of the relationship between the number of scans and the actual daily encounters experienced by each student, two distinct networks were constructed: one based on the day boasting the highest scan count (the second Thursday) and another reflecting the day with the lowest scan count (the first Sunday). The exploration of these networks, where node size serves as a proxy for the number of people encountered by individual students, illuminates the intricate dynamics of social connections within the studied population. This contrast is vividly portrayed in figure 3, where Sunday unveils a maximum node degree of 4 (indicating interaction with a maximum of four people), in stark contrast to Thursday, which exhibits a notable peak node degree of 26. Noteworthy is the observation that on Sundays, the majority of students engage with only one or two others, while on Thursdays, it is not uncommon to find students interacting with six or seven peers.

It is crucial to acknowledge that the dataset utilized for the experiments may not comprehensively represent digital communication within the population under study. **Notably, phone call logs reveal a surprisingly high number of students who did not initiate a single phone call during the entire 4-week observation period**. A similar trend is observed in text message logs. Furthermore, **the overall count of phone calls and text messages appears unexpectedly low, given the number of participating students**.

For each student, the number of outgoing calls (figure 4) and incoming calls (figure 5) was tallied. As previously mentioned, some students did not place any calls, while others received none. Conversely, certain students either received or initiated numerous calls. Specifically, **the student with the highest outgoing call count made 94 calls over the 28-day period, while the student with the highest incoming call count received 101 calls**.

Histograms in figures 6 and 7 further illustrate that, **for the majority of students monitored, both outgoing and incoming calls did not surpass 20 instances throughout the observation period**.

Regarding call duration, as showed in the histogram in 8 **most calls lasted less than 250 seconds** (approximately four minutes).

The same type of analysis was conducted to examine the exchange of messages among students. As noted in table 1, **the number of SMSs exchanged is much higher than the number of phone calls made**, as anticipated. This was easily predictable, considering that messages

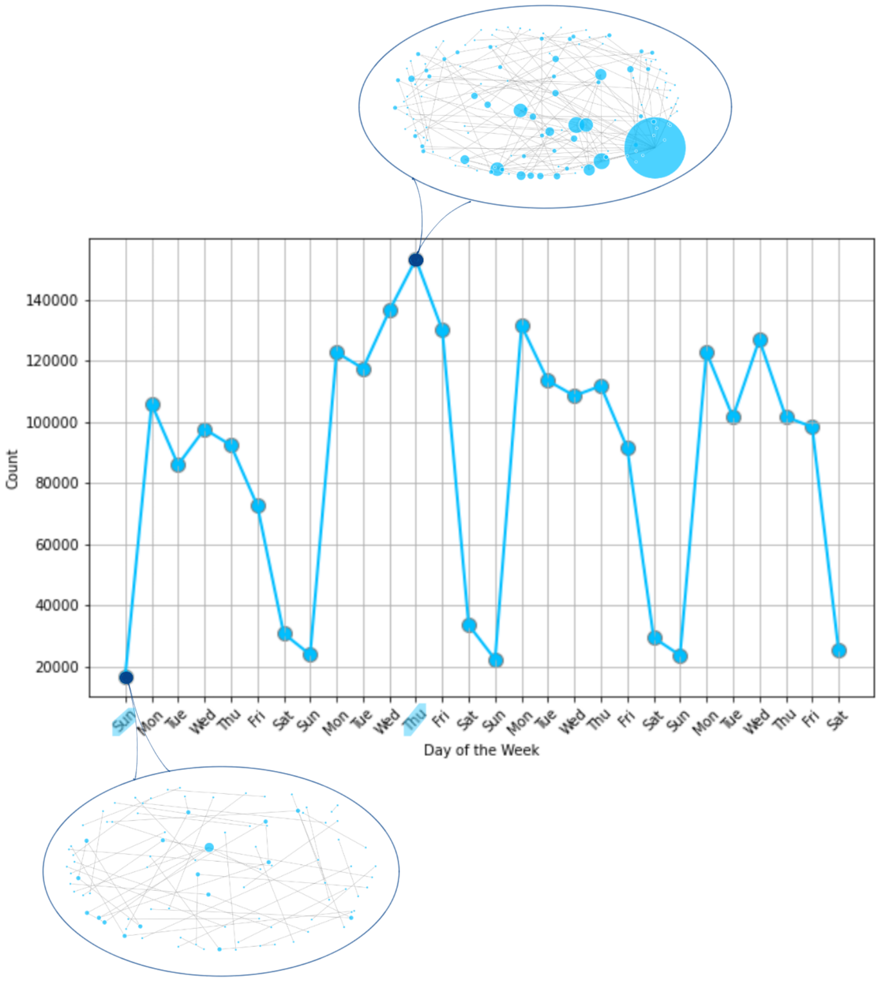

Figure 2: Total Bluethoot Scans Per Day and Network Student Visualization

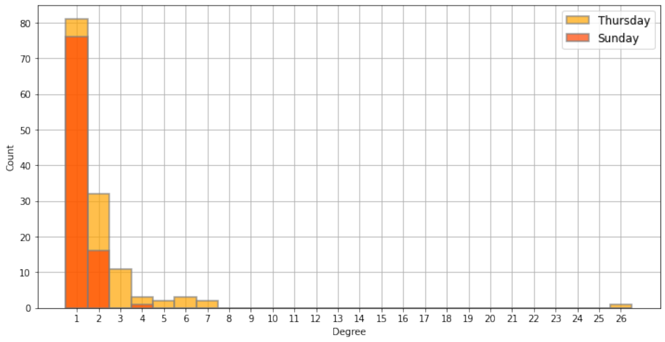

Figure 3: Degree Distribution of the Busiest and Least Crowded Day of the Period

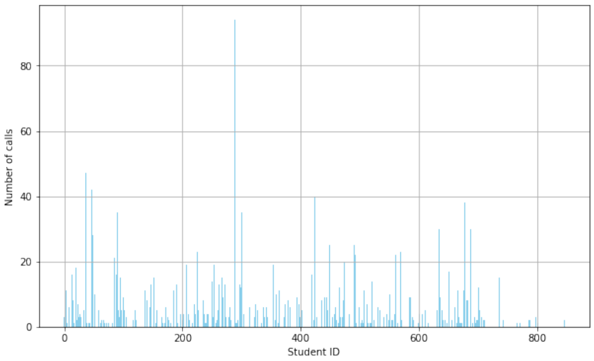

Figure 4: Outgoing Calls per Student

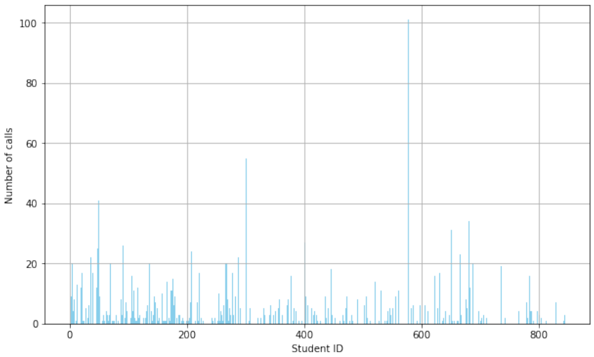

Figure 5: Incoming Calls per Student

represent a more instantaneous form of communication compared to phone calls, especially given the participants' age.

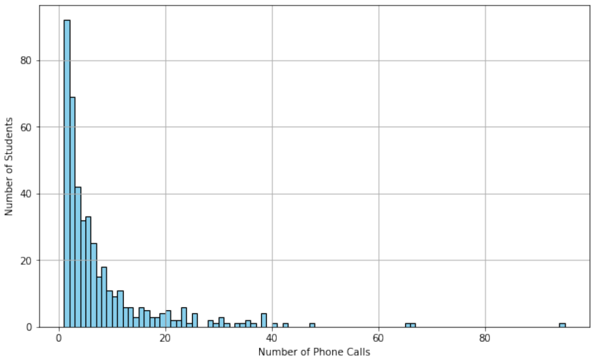

Figure 6: Outgoing Phone Calls Distribution

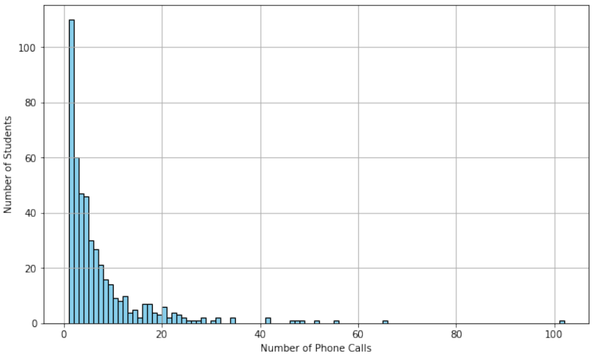

Figure 7: Incoming Phone Calls Distribution

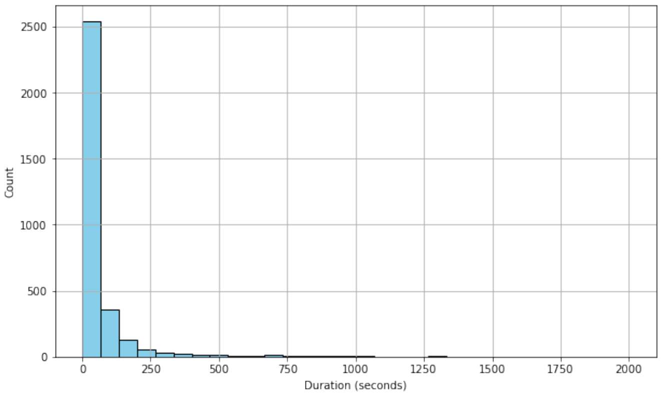

Figure 8: "Distribution of Call Duration"

In this case, **the highest number of messages sent by a student was 1448, while the highest number of received messages was 1412**. Similar to phone calls, there are students who either did not send any SMS or did not receive any.

In figure 11, the representation of the message network is depicted, along with the degree distribution shown in figure 12. It can be observed that, typically, **most students communicated via messages with a number of peers ranging from one to eight, with one student interacting with a maximum of twenty others**.

To ascertain potential correlations, figure 13 illustrates the trend of the number of calls and messages exchanged for each of the 28-days. A good correlation can be observed, especially in the second half of the period. It is important to note, however, that unlike the interactions derived from Bluetooth scans in 2, **there is no well-defined periodic trend for the exchange of calls and SMS**.

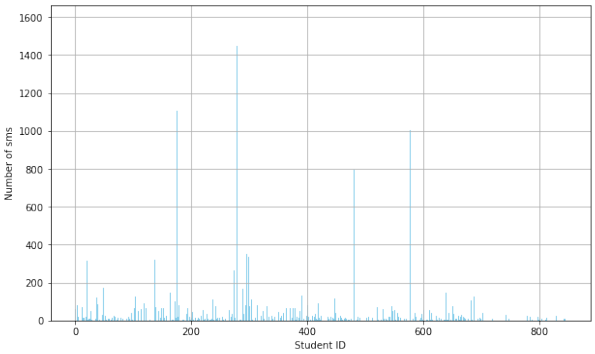

Figure 9: SMSs Sent per Student

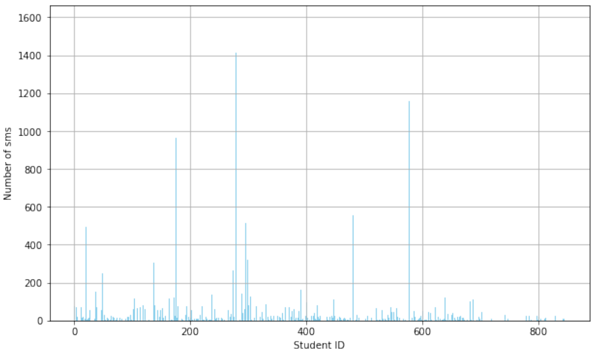

Figure 10: SMSs Received per Student

Until now, most considerations have not taken into account the gender of the students. As shown in figure 1, there is a significant discrepancy between the number of male and female participants. Therefore, it may be interesting to analyze whether this disparity could influence communication and social interaction among students in a more or less noticeable way.

Firstly, the gender of the participants was contextualized in relation to the number of messages sent and received, to determine which, among males and females, sent and received more

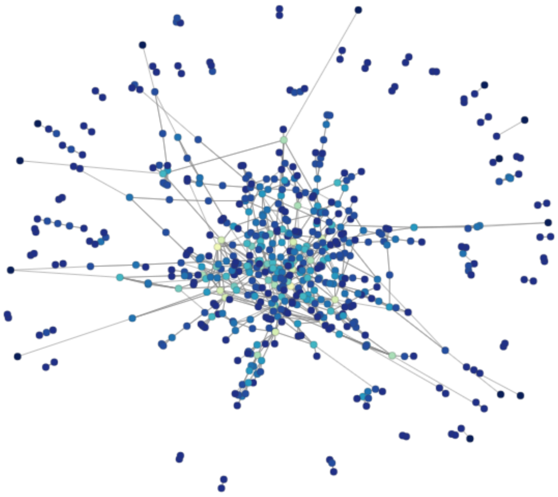

Figure 11: SMS Network Visualization

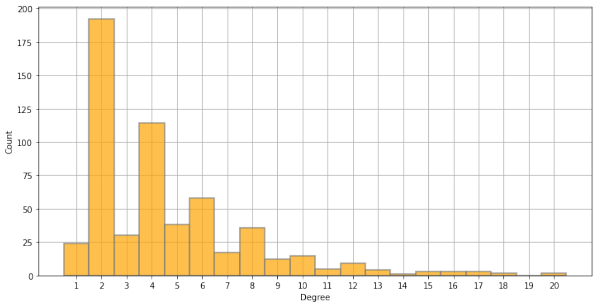

Figure 12: SMS Network Degree Distribution

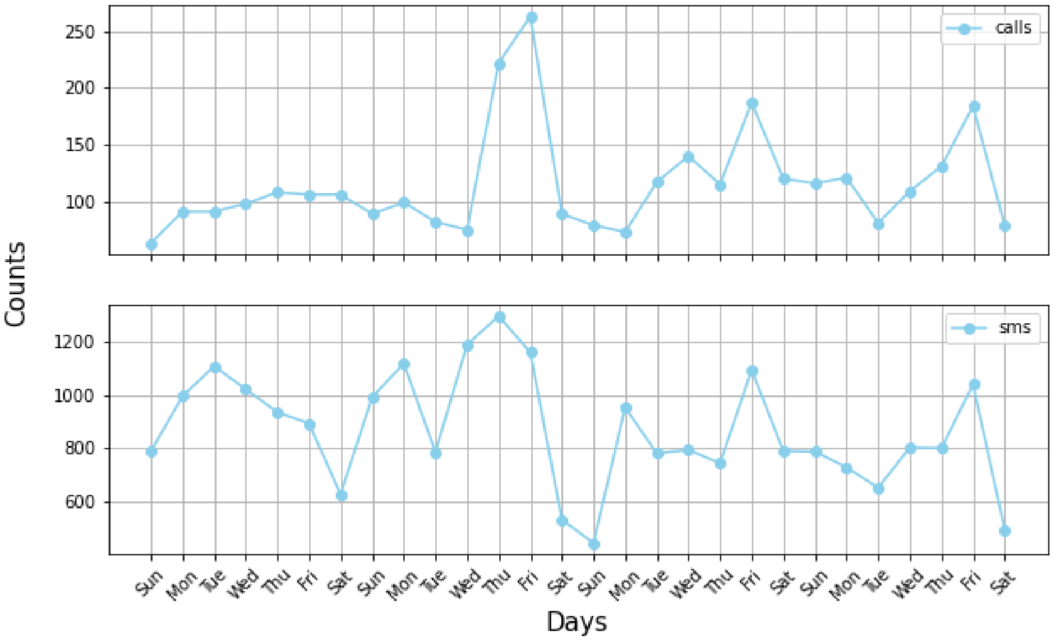

Figure 13: Number of Calls and SMSs per Day

messages. The histogram in figure 14 reveals that **male students both sent and received the highest number of SMSs**. This result was expected, given the higher number of male students participating in the data collection. However, it is interesting to note that the disparity between male and female students in the quantity of messages sent and received is not as pronounced as the disparity in the number of participants. This confirms that, **despite being fewer in number, female students made a significant contribution to message exchange and, consequently, social interactions**.

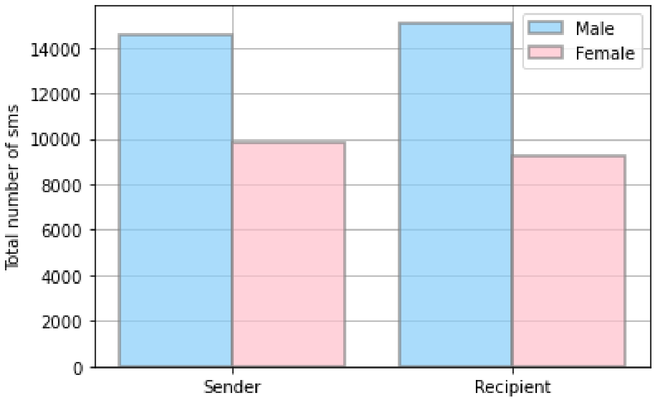

Figure 14: Gender-Specific Message Patterns: Sent vs. Received

Subsequently, a more detailed analysis was conducted on an individual scale to examine whether male or female students were more active in sending and receiving messages with their peers.

In the network visualization in 15, each node represents a student, and the node size is proportional to the number of different individuals to whom the student has sent SMS. From the network, **it is evident that female students have sent messages to a greater number of diverse individuals**. In particular, the most active female student sent messages to 20 different individuals, while the most active male student sent messages to only 17.

The same analysis was also performed concerning the number of different individuals from whom the student received SMS. In this case, as showed in figure 16 despite the majority of male students, **female students received more messages**. Specifically, the "most sought-after" female student received messages from 21 different individuals, compared to the "most sought-after" male student, who received messages from only 17."

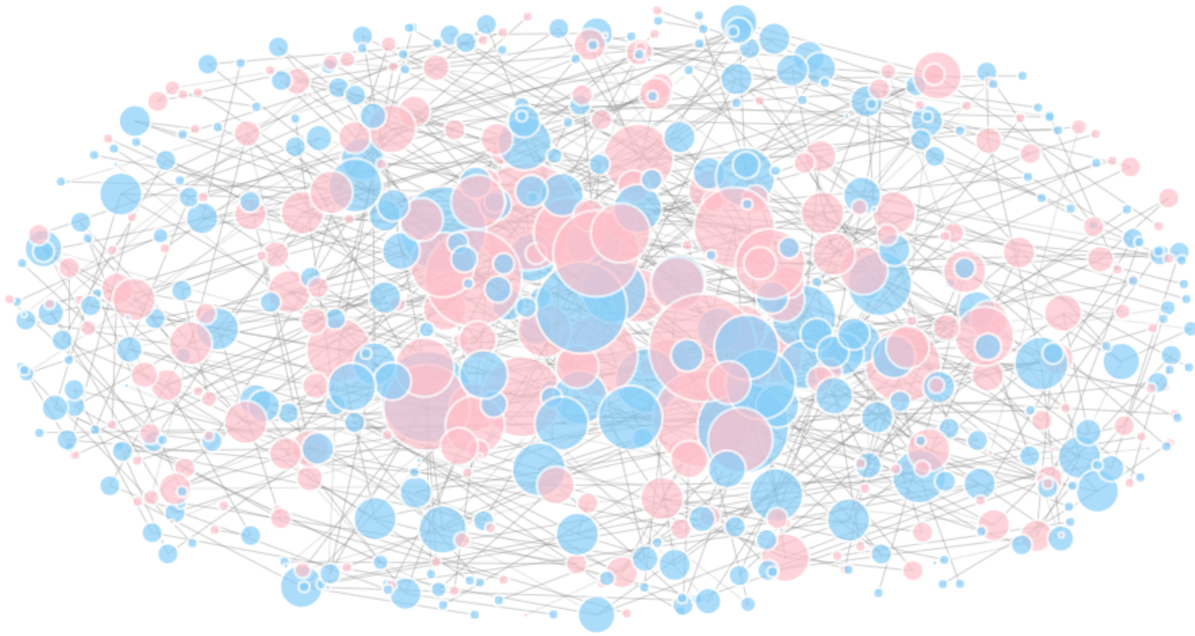

Figure 15: Gender-Specific Network Depicting the Number of Different Individuals to Whom a Message Was Sent

After observing the significant role played by female students in message exchanges, a similar analysis was conducted for phone calls. This time, in the network visualization shown in figure 17, each node still represents a student, but the size is proportional to the number of different individuals they called. **Once again, it was the female students who engaged in phone interactions with a greater variety of people**. In particular, the most active female student made calls to 22 different people over the course of four weeks, while the most active male student only called 17 different people.

The same analysis was also carried out for received calls. This time, as depicted in the network in figure 18, male students received calls from a more diverse set of people, reaching a peak of 22 individuals, compared to the 19 observed for female students.

From this analysis, it can be observed that, regarding phone calls, **female students tend to call a more diverse set of people compared to male students, who, in turn, receive calls from a more varied group of individuals**. It is interesting to note that despite their numerical disadvantage in this context, female students contribute noticeably to social interactions through phone calls as well.

An additional analysis conducted on the calls involved representing, for each gender, the breakdown of the individuals called based on gender, indicating the number of males and females. As can be observed from figure 19, for female students, the majority of calls (both outgoing and incoming) were made with male students, although there are instances where some female students exclusively interacted over the phone with other female students or only with male students over the four weeks.

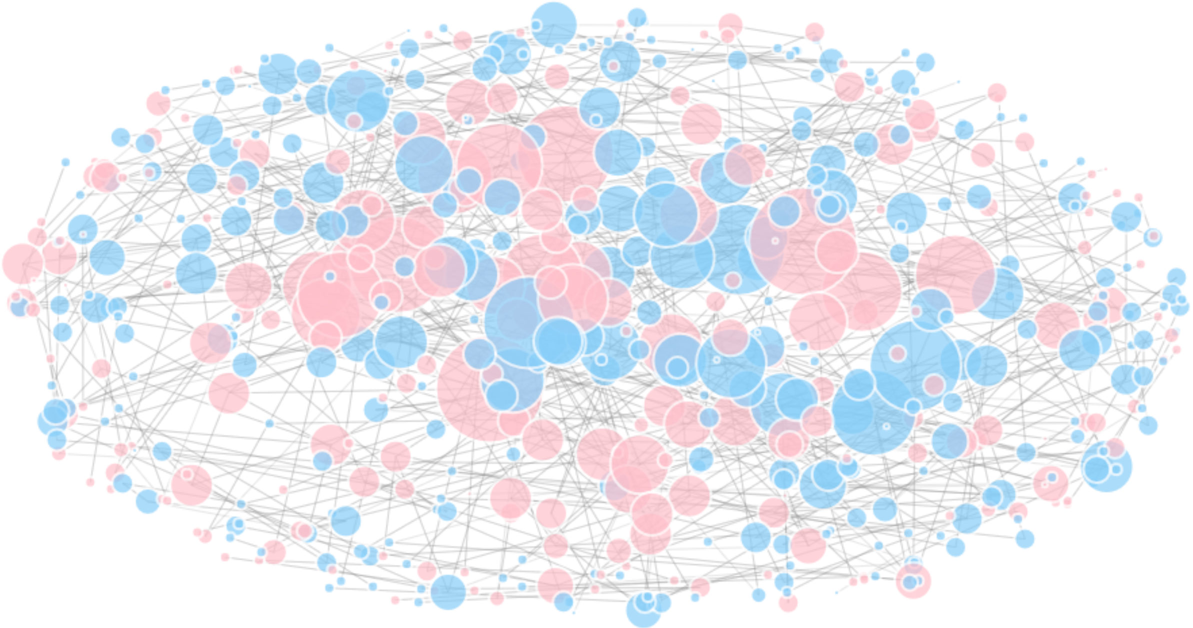

Figure 16: Gender-Specific Network Depicting the Number of Different Individuals from Whom a Message Was Received

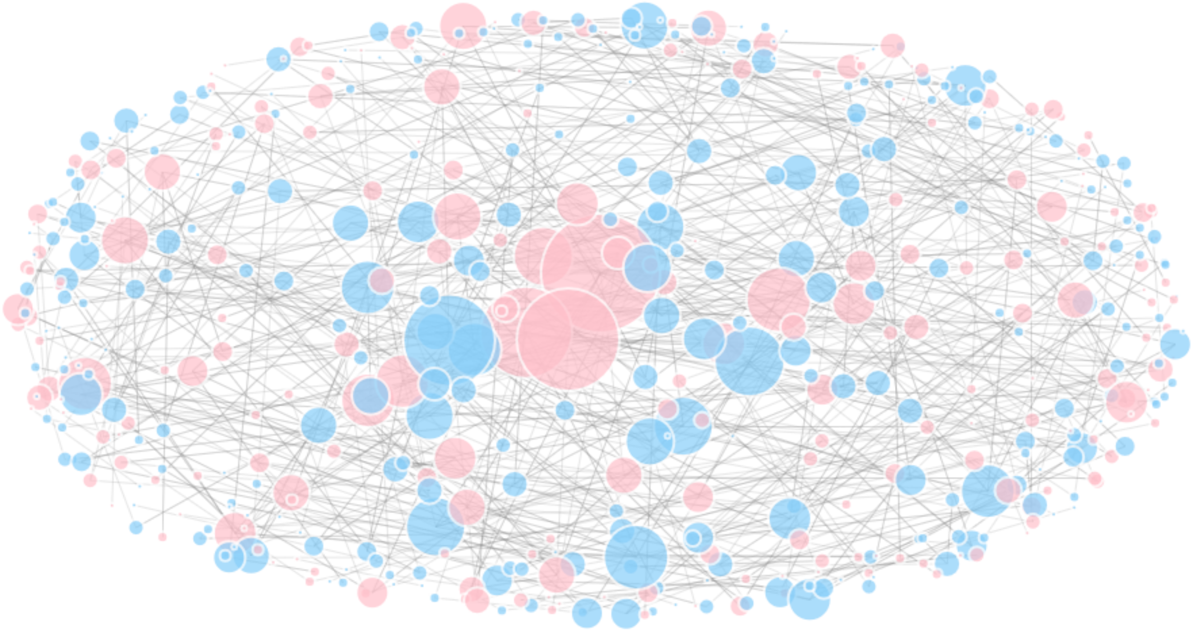

Figure 17: Gender-Specific Network Depicting the Number of Different Individuals to Whom a Call Was Made

Regarding the analysis conducted on male students, similarly, the majority of calls (both outgoing and incoming) were made with students of the same gender, as showed in figure 20. In this case as well, there are instances where male students exclusively interacted over the phone with other male students or only with female students.

One fundamental aspect in the study of Network Science concerns degree distribution, and to delve into it, the Facebook network has been considered. This network comprises the edge lists of all declared Facebook friendships that were established before the end of the observation period and remained intact until after the conclusion of the observation.

The degree distribution of the network, as depicted in the figure 21, has been compared with that obtained from a random graph containing the same number of nodes, with a connection probability of 0.5, in figure 22.

In traditional random networks, the distribution of node degrees tends to be rela-

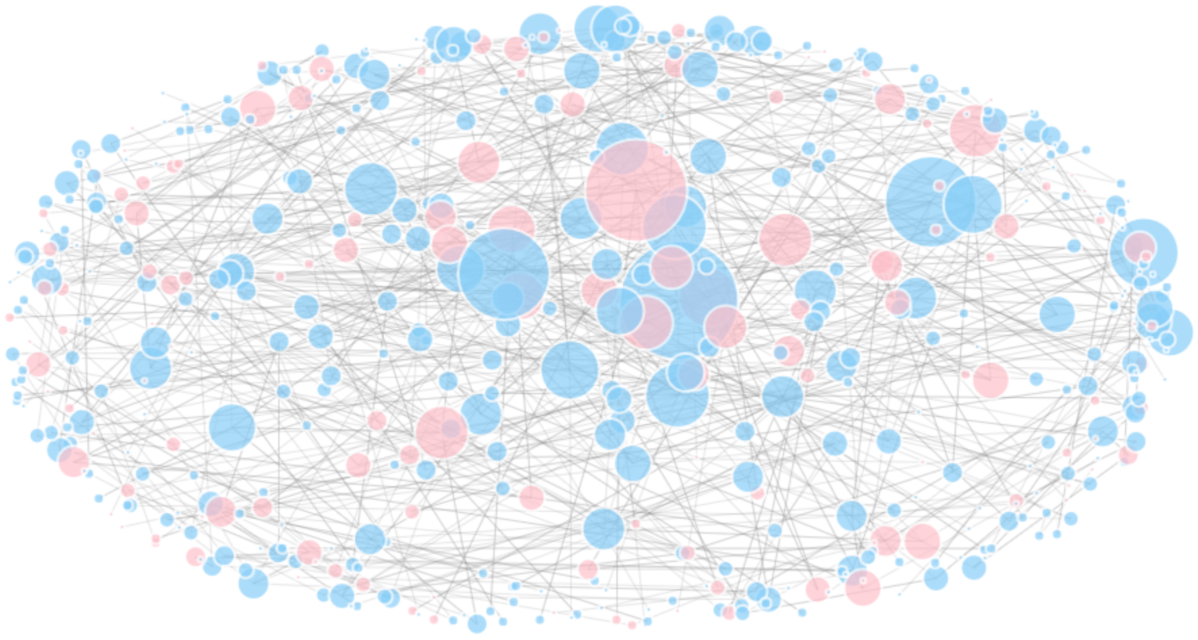

Figure 18: Gender-Specific Network Depicting the Number of Different Individuals to Whom a Call Wa Received

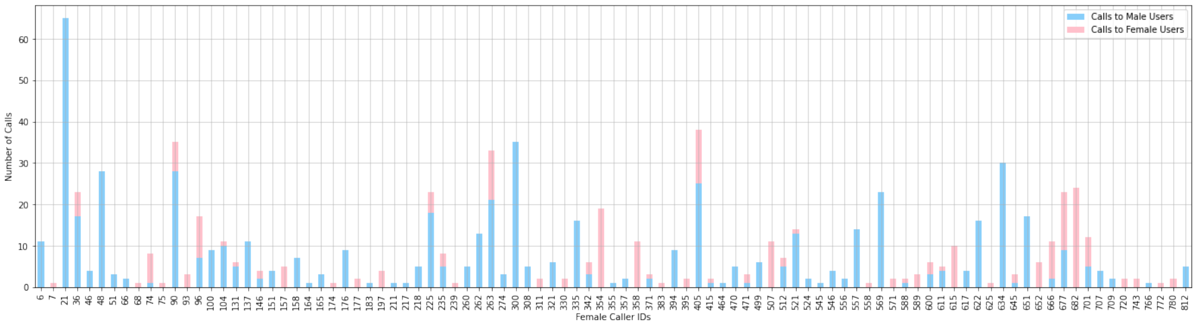

Figure 19: Individuals Called by Female Students Based on Gender

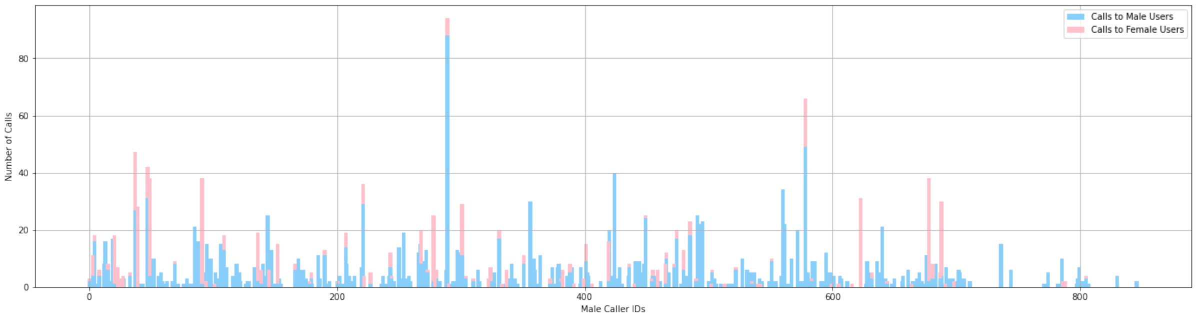

Figure 20: Individuals Called by Male Students Based on Gender

tively uniform, with most nodes having a moderate number of connections. The degrees of individual nodes cluster around the average degree of the network. This homogeneity in connectivity is a characteristic feature of random networks generated using models like the Erdős-Rényi model, where edges between nodes are formed independently with a certain probability [10]

On the other hand, real-world networks (like the Facebook one analyzed) often exhibit a markedly different pattern known as a skewed node-degree distribution. In such networks, most nodes have only a few links, representing a majority of individuals with a limited number of connections. However, in stark contrast, there exist a small number of nodes that are highly connected, forming what is referred to as hubs or influencers. These highly connected nodes play a crucial role in shaping the overall structure of the network.

This uneven distribution of node degrees, where the majority have few connections while a few have a disproportionately large number, is commonly described by a power-law or scale-free distribution [11]. In a power-law distribution, the probability of a node having a specific degree is inversely proportional to that degree raised to a certain exponent. This results in a heavy-tailed distribution, indicating the presence of a few nodes with exceptionally high degrees. Scale-free networks, characterized by this power-law distribution, are prevalent in various real-world systems, including our Facebook friendship network.

The emergence of scale-free networks is often attributed to the preferential attachment mechanism, where new nodes are more likely to connect to existing nodes with higher degrees [12]. This ”rich-get-richer” phenomenon over time leads to the formation of hubs, influencing the topology and resilience of the network. Understanding the characteristics of power-law distributions and scale-free networks is crucial for comprehending the underlying principles governing the structure and dynamics of complex systems observed in nature and society.

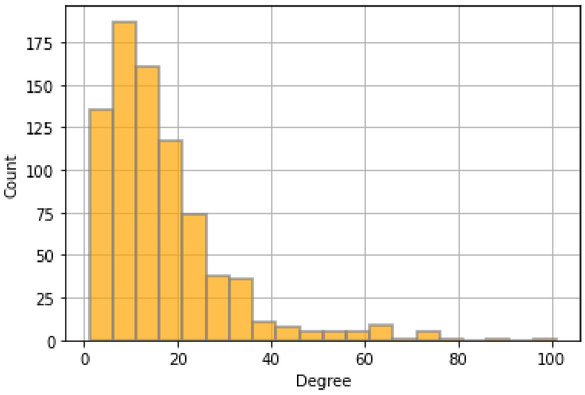

Figure 21: Fb Network Degree Distribution

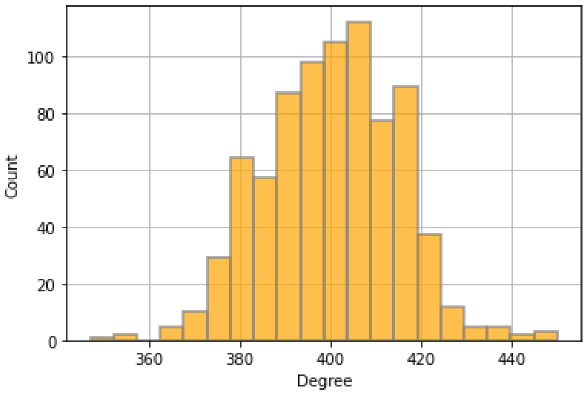

Figure 22: Random Graph Degree Distribution

Finally, in the last analysis, the types of communicative interactions among male individuals, female individuals, and mixed-gender interactions were examined.

As evident from the heatmap in 23, **female students predominantly favor messages as a means for their interactions, unlike male students who prefer engaging through calls and Facebook**.

Ultimately, social interactions between male and female students predominantly occur through messaging.

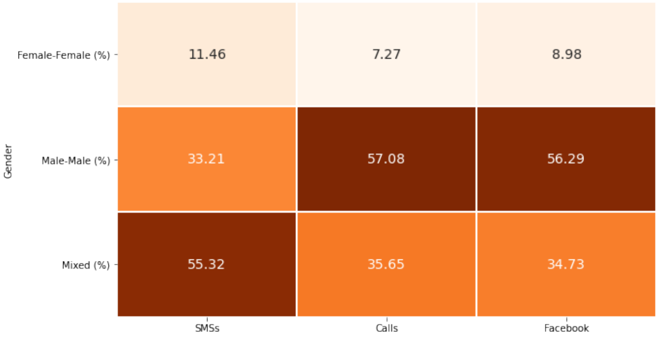

Figure 23: Distribution of Communication Types by Gender Combinations

Conclusion and Discussion

Wrapping Up

The analyses conducted in this project have indeed demonstrated the **presence of periodic patterns in social interactions detected through Bluetooth scans**. Moreover, the differences between more and less crowded days are clearly visible when analyzing the network. Additionally, the existence of correlations between the number of exchanged messages and calls has been confirmed.

Conducting analyses while considering the gender of the involved students revealed that, **despite females being in the numerical minority, they have made a significant contribution to communication with students, highlighting specific differences in the analyzed networks**.

As expected, the network of Facebook Friendships exhibited characteristics typical of a scale-free network. Finally, it was observed that messages were the preferred communication tool among students of the opposite gender.

Limitations and Future Works

In contrast to the comprehensive Copenhagen Networks Study dataset, there are specific aspects where the study's scope could be broadened.

It is evident that not all participants actively utilized their assigned smartphones throughout the observed four-week period. Some students experienced no phone calls, text messages, or proximity scans during this timeframe, suggesting potential variations in communication preferences. Factors such as a preference for instant messaging over traditional texting or frequent Bluetooth deactivation may contribute to these data gaps. To mitigate these gaps in future studies, several strategies could be explored. Monitoring instances when a student's device has Bluetooth disabled could help account for missing Bluetooth scans. Alternatively, considering methods to prevent all devices in the study from disabling Bluetooth could provide more comprehensive data. Additionally, acknowledging the possibility that students may use alternative smartphones for their daily communication needs, a pragmatic approach could involve developing software installable on a student's existing device, rather than providing a new one.

Furthermore, to delve into the dynamics of social groups rather than individual pairs, a prospective study could shift its focus to identifying group text messages instead of individual exchanges. While the current dataset may contain some text messages sent to groups, the lack of explicit differentiation makes it challenging to precisely discern them. A refined dataset should explicitly categorize text messages exchanged between individuals and those occurring within group settings.

References

[1] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann. Measuring large-scale social networks with high resolution. PLoS ONE, 9(4), 2014

[2] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann. Tracking human mobility using WiFi signals. PLoS ONE, 10(7), 2015.

[3] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann, and A. Baronchelli. Evidence for a conserved quantity in human mobility. Nature Human Behaviour, 2(7):485–491, 2018.

[4] D. K. Wind, P. Sapiezynski, M. A. Furman, and S. Lehmann. Inferring stop-locations from WiFi. PLoS ONE, 11(2):1–15, 2016.

[5] P. Sapiezynski, A. Stopczynski, D. K. Wind, J. Leskovec, and S. Lehmann. Inferring Person-to-person Proximity Using WiFi Signals. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologie, 2016.

[6] V. Sekara and S. Lehmann. The strength of friendship ties in proximity sensor data. PLoS ONE, 9(7):1–8, 2014.

[7] P. Sapiezynski, A. Stopczynski, D. K. Wind, J. Leskovec, and S. Lehmann. Offline Behaviors of Online Friends. In ICWSM 2014, 2018.

[8] I. Psylla, P. Sapiezynski, E. Mones, and S. Lehmann. The role of gender in social network organization. PLoS ONE, 12(12):1–21, 2017.

[9] P. Sapiezynski, C. Wilson, and V. Kassarnig. Academic performance prediction in a gender-imbalanced environment. In Proceedngs ofFATRECWorkshop on Responsible Recommendation at ACM RecSys, Como, Italy, 2017

[10] Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. Publ. math. inst. hung. acad. sci, 5(1), 17-60.

[11] Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. Scientific american, 288(5), 60-69.

[12] Piva, G. G., Ribeiro, F. L., & Mata, A. S. (2021). Networks with growth and preferential attachment: modelling and applications. Journal of Complex Networks, 9(1), cnab008.

[13] Sapiezynski, P., Stopczynski, A., Dreyer Lassen, D., Lehmann Jørgensen. The Copenhagen Networks Study interaction data. 2019. https://figshare.com/articles/dataset/The_Copenhagen_Networks_Study_interaction_data/7267433/1