Using Network Analysis to Study Human Behavior

1. Introduction

1.1. Setting and context

Studying the human behavior is a difficult task due to its inherent complexity (biology, environment, emotions, cognition), variability (people react differently to the same stimuli, moods change), and the observer effect which refers to the alteration of behavior when a person is watched (Farnsworth, B. 2025). All these factors make objective measurement hard. However, there are many ways to measure human behavior, involving a mix of direct observation, through surveys and questionnaires, and physiological/technological tools. A great approach that was recently introduced, is the network approach. This approach is described by Mkhitaryan et al., 2019 as "(…) the network approach represents observable indicators as nodes in a network and looks at the interrelationships (...)". This allows researchers to explore what might determine behavior and evaluate the importance of each variable in the studied system.

1.2. Literature Review

Smartphone use in today's world offers the advantage to acquire behavioral data in situ, representing a population's daily life. As Harari et al. (2016) mention, "(…) these data allow a fine-grained, continuous collection of people's social interactions, daily activities, and mobility patterns.". There is some literature regarding the use on network studies to study human behavior that gives us some insights of how smartphones and social interactions have a strong relationship on the present day. For example, Rotondi et al. (2017) found in their own analyzes that time spent with friends is worth less, in terms of life satisfaction, for individuals who use smartphones. On the other hand, Das et al. (2024) found that physical engagement is decreasing since people are tending to prefer interaction through social media. Finally, Aledavood et al. (2018) found a strong connection between the chronotypes of people and the structure of social networks that they form. These findings were confirmed that using smartphones is a great tool to collect people's interactions and make the study of human behavior easier.

1.3. Research Question

In this project I will work with The Copenhagen Networks Study interaction data, which was collected using smartphones and "(…) represents the multi-layer temporal network which connects a population of more than 700 university students over a period of four weeks." (Sapiezynski et al., 2019). Based on this dataset, the present study examines the following question: **How do physical proximity interactions relate to digital communication patterns, and what network structures emerge from these behaviors among university students?**

1.4. Contribution Summary

To address this question, the study analyzes the multilayer social network formed by the student's Bluetooth proximity, SMS, phone calls, and Facebook friendships. The analysis combines basic statistics with network-analysis methods that include weighted degree, betweenness centrality, temporal activity patterns, among others. The results reveal that the proximity network is highly dense and structured into two dominant communities, while SMS and phone call networks are much sparser. A moderate but significant association is observed between physical proximity and digital communication, suggesting that face-to-face interaction partially predicts online contact. Centrality measures identify a small set of students who serve as key social connectors within the community. These findings provide a multilayered idea of students' social behavior and show how data obtained with smartphones can capture, both, physical and digital dimensions of interaction.

2. Problem Formulation

Understanding how people interact across different dimensions of social behavior such as physical proximity, digital communication, and online social ties is a central challenge in behavioral network studies. While previous studies have clearly demonstrated that smartphone data capture meaningful aspects of social life, it is unknown just how strongly physical interactions align with digital communication, and how the structure of these interaction networks differ across types of interactions.

In this project, the research problem is to characterize and compare the physical and digital interaction networks of a student population by using Bluetooth proximity, SMS, phone calls, and Facebook friendship data from the Copenhagen Networks Study. This will be done through the examination of structural properties of each network, identifying community patterns, detecting the most influential individuals, and evaluating how can physical proximity predict students' digital communication behavior.

2.1. Scope

The problem is relevant and manageable since the dataset provides multilayer interaction traces for the same individuals (students) over the same time period and thus allows a direct comparison of behaviors across channels. At the same time, the analysis will address a less explores question regarding how consistent are social relationships across physical and digital networks, and what does that say about real-world social behavior? The present study aims to develop insights by approaching the data with network analysis techniques, that may contribute to a better understanding of human interaction patterns and their relationship with physical and digital communication.

3. Dataset Description

3.1. Dataset Introduction

The dataset used on this study was obtained from the Copenhagen Networks Study (Sapiezynski et al., 2019), which is a large population observational study that monitored more than 700 university students over a period of four weeks. Students were provided with smartphones that passively recorded multiple forms of interactions, forming a multilayer social network. Four types of interaction data were analyzed: physical proximity (estimated via Bluetooth signal strength), SMS, phone calls, and Facebook friendships. This was accompanied by the gender information of the students. Together, these datasets form a multilayer network capturing both physical and digital aspects of social behavior. A summary of the datasets is presented on Table 1.

3.2. Data & Features

Dataset	**Description**	**Observations** (number of rows)	**Individuals per dataset**	**Time Resolution**
Bluetooth proximity	Automatic scans of nearby devices, representing face-to-face interactions	3,245,378	692	5-minute intervals
SMS messages	Text messages exchanged between participants	24,333*	569	Based on event
Phone calls	Call events with timestamps and duration	3,600*	525	Based on event
Facebook friendships	A static social network showing existing friendship ties	6,429	800	Static
Gender	Demographic information used for group-level comparisons	800	800	Static

**Before cleaning*

Table 1. Summary of datasets

Below, there is a more detailed overview about each dataset's features and their meaning, including the data point meaning of each.

Datapoint meaning: a single scan indicating whether two devices were in physical proximity.

Bluetooth Proximity	
Feature	**Meaning**
timestamp	Timestamp of the proximity scan
user_a	Scanning device
user_b	Detected nearby device
rssi	Signal strength

Table 2. Bluetooth Proximity dataset features

Datapoint meaning: one text message sent from one user to another.

SMS	
Feature	**Meaning**
timestamp	Time message was sent
sender	User sending the SMS
recipient	User receiving the SMS

Table 3. SMS dataset features

Datapoint meaning: one phone call made from one user to another.

Phone Calls	
Feature	**Meaning**
timestamp	Call start time
caller	User making the call
callee	User receiving the call
duration	Duration of the call in seconds

Table 4. Phone Calls dataset features

Datapoint meaning: one friendship link that tells that one user knows another.

Facebook Friendships	
Feature	**Meaning**
user_a	Facebook user
user_b	Facebook friend

Table 5. Facebook Friendships dataset features

Datapoint meaning: one user and their gender.

Gender	
Feature	**Meaning**
user	Unique participant ID
female	Gender label (0 = male, 1 = female)

Table 6. Gender dataset features

3.3. Data Preprocessing

Before constructing the interaction networks, each dataset was cleaned to remove invalid observations and retain only the meaningful interaction events. These steps were guided both by the official data processing recommended documented in Sapiezynski et al. (2019), and the analysis goals of this project. Below, I describe the preprocessing approach separately for each data modality, followed by the methodological justification.

3.3.1. Bluetooth Proximity Data

Bluetooth scans are sampled every 5 minutes. This dataset contains the following observations:

- Empty scans are recorded with user_b = -1 and rssi = 0
- Devices outside the monitored population appear as user_b = -2 - Rssi = 0 can also indicate invalid readings

Bluetooth scans generate extremely large datasets and removing invalid or non-participant interactions is essential to prevent spurious edges and inflated degree centrality. These were the preprocessing steps: o Removed empty scans defined by which correspond to failed detection events and not real proximity

- o Removed scans where the device is outside the monitored population, which represent devices not belonging to the study population.
- o Kept only interactions between participants by only keeping the positive values of both users
- o Removed all datapoints that have invalid readings

3.3.2. SMS Data

Each row corresponds to one SMS message. The dataset contained a small number of duplicate rows. These duplicate rows would overestimate tie strength and inflate weighted degree. The preprocessing step consisted of:

- o Removed exact duplicate messages

3.3.3. Phone Call data

Each row represents one call event, including the duration. The dataset contains the following observation:

- The dataset contains missed calls encoded with a duration of -1.

Only meaningful communication events should contribute to tie strength in the call network. For this reason, the preprocessing steps consisted of:

- o Removed missed calls, which don't represent actual interaction and would artificially increase degree in the call network.
- o Removed calls with no duration, which represent instantaneous or failed connections and do not reflect real communication.

3.3.4. Facebook Friendship Data

The dataset lists static friendship relations and was recorded once. For this dataset, the preprocessing steps were:

○ Renamed inconsistent column names to maintain clarity when merging or constructing networks

- Removed duplicate friendship entries

3.3.5. Gender Data ○ Renamed inconsistent column names to maintain clarity when merging or constructing networks

- Created a dictionary mapping the user with its gender ○ Counted how many students were male and female

4. Methods

This project relies on network science and statistical analysis to study the communication patterns of university students in a physical and digital environment. The analyses were performed in Python, using pandas, NetworkX, NumPy, SciPy, and scikit-learn.

After preprocessing the data, interaction networks were constructed for each modality. In all networks, nodes represent individual students and edges (or links) represent a pair of students' interactions. For proximity, SMS, and call data, the networks were weighted: the weight of an edge corresponds to the total number of times the pair interacted in that modality. Since interactions are naturally bidirectional, each pair of users was represented as an undirected edge; this was done by sorting the user IDs within each pair to make sure that a unique dyad (pair of students) identifier is consistently assigned. The Facebook network was represented as an unweighted, undirected graph with static friendship ties. These graph representations formed the basis of all further analyses.

To investigate the structure of social interactions within the student population, a combination of descriptive statistics, correlation analysis, and advanced network science techniques was applied across the communication layers: Bluetooth proximity, SMS messaging, phone calls, and Facebook friendships. These methods allowed examination of both community-level network properties and individual-level behavioral roles, as well as relationships between physical and digital forms of interaction.

4.1. Basic Methods

The analysis began with basic descriptive methods, data visualization, and correlation analysis, which were essential for establishing an initial understanding of each network.

For all networks (*G_prox, G_sms, G_calls, G_fb*), node counts, edge counts, and network density were computed to quantify the size and sparsity of each interaction layer. Histograms were used to visualize the weighted degree distribution and edge weight distribution in the proximity network (*G_prox)*. This was crucial to characterize heterogeneity of student activity. Time series plots (line plots) were used to compare patterns of SMS messaging, calling, and proximity across the hours of the day.

To understand the relation between physical interaction and digital communication, some basic inferential methods were applied. The Pearson and Spearman correlations between weighted degrees of users in the proximity networks were calculated. Because the Pearson correlation captures the linearity of association while Spearman correlation captures the monotonic consistency in ranking, both measures could be referred to for a better interpretation.

A simple linear model was fitted to predict SMS weighted degree from proximity weighted degree, to assess if the proximity activity could possibly be a predictor of SMS activity.

4.2. Advanced Methods

Beyond descriptive and correlation methods, several advanced network analysis techniques were used, particularly focusing on the proximity network (*G_prox*) that provides the most complete representation of physical social interaction. Centrality measures quantified individual influence and behavioral roles within the network. Weighted degree centrality identified those students who are the most socially active in terms of total interactions, while betweenness centrality identified those who served as structural bridges, connecting different groups. Then, a comparison between these two measures was done using a scatterplot for further analysis.

Community structure was analyzed using greedy modularity maximization, a widely used method for partitioning networks into cohesive groups. Additional structural metrics including average shortest path length, connected components, and clustering coefficients were used to characterize the network's global organization.

Ego-network analysis of the most socially active student was used to intuitively visualize his or her immediate neighborhood and thus present a view of local network structure that centered around a given node.

4.3. Method Motivation

The methodological choices of this project were guided by the need to understand multilayered social dynamics present in a population, and to consider how different communication channels relate to one other.

Since all these datasets consist of pairwise interactions, the meaningful unit of analysis is the relationship between two individuals rather than the isolated event. For that reason, raw time-stamped data was transformed into undirected weighted graphs using NetworkX. Such representation allowed each edge to encode both the presence and strength of a social tie. The use of undirected graphs was justified because, in the majority of cases, the direction of initiation was not relevant, and what mattered was the presence of mutual interaction, rather than who initiated it.

A central tool in this transformation process was the NetworkX library. NetworkX was chosen because it offers a flexible, well documented framework for representing and analyzing complex networks using Python (NVIDIA, n.d.). Compared to alternative approaches such as writing custom graph structures or relying on visualization-oriented tools, NetworkX provides a programmable, reproducible, and highly extensible environment. It supports weighted graphs, efficient computation of centrality measures, community detection algorithms, and integration with pandas DataFrames, making it very suitable for the analytical needs of this project.

Basical analytical methods were employed first and mostly for data exploration and validation. These methods provide an important overview of the network size, density, and degree distributions, which help in verifying how reasonable the constructed graphs are. Because the question is at the intersection of behavioral data science and social network analysis, the correlation measures provided a useful bridge between traditional statistical methods and network-based insights.

Advanced methods in Network Science were applied to capture structural properties not observable through descriptive statistics alone. Specifically, centrality measures allowed for the identification of highly socially active individuals or those playing a structurally important bridging role in the network. Community detection using modularity maximization provided an objective way to uncover the social clusters that form naturally among students.

5. Results

This section presents the key findings that resulted from the multi-channel analysis of proximity, SMS, call, and Facebook friendship network. The results are interpreted in relation to the research question previously presented: *How do physical and digital communication channels relate to one another, and what do these interactions reveal about the underlying social structure of the student population?*

5.1. Key Findings and Interpretation

5.1.1. Physical Presence Forms a Dense, Unified, Small-World Network

An important result of the analysis is the structural contrast between physical proximity and digital communication networks. Real-world social networks are typically sparse because individuals maintain relationships with only a small subset of all possible partners. In contrast, the proximity network (*G_prox*) displays a very high density of approximately 0.33. In networks with hundreds of people, densities close to zero are the norm, even when the relationships are strong. This confirms that proximity captures ambient co-location rather than intentional social interaction, producing unusually dense connections. Moreover, it consists of one single connected component, meaning every student is indirectly reachable from any other through chains of physical co-location.

The degree and edge-weight distributions reveal a strong right skew (Figure 1), where most edges represent weak, brief, low-frequency encounters. Also, a small number of edges represent extremely strong ties, with the strongest pair (students 322 and 321) recording 5,766 proximity interactions. Moreover, a subset of individuals shows very high weighted degree (over 25,000 interactions), identifying them as potential super connectors.

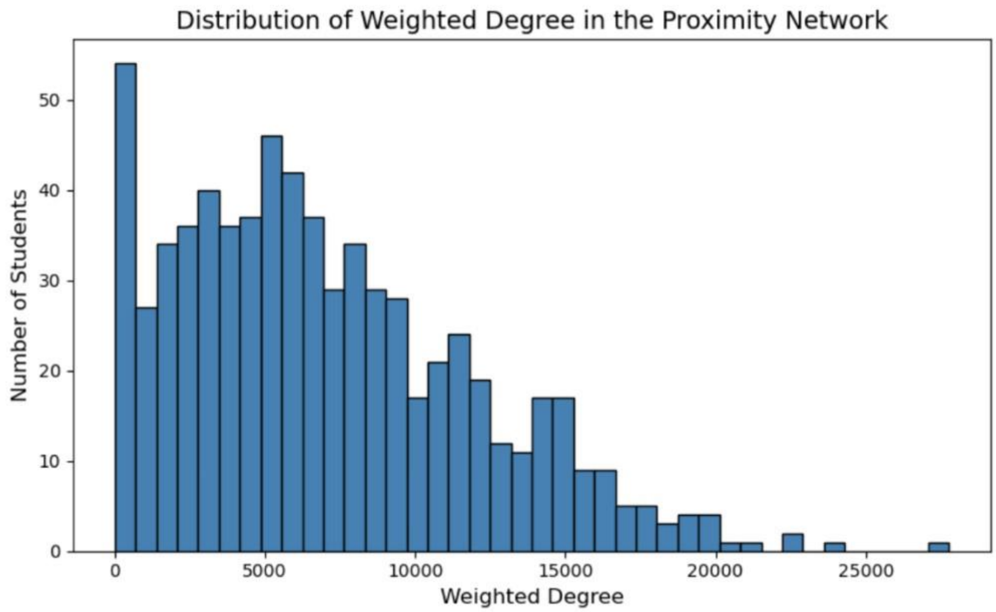

Figure 1. Weighted Degree Distribution in the Proximity network showing how much edges represent either weak or strong encounters.

The network exhibits a very short average path length (of about 2.08, meaning that every student is about edges from any other student), which is characteristic of small world networks. Taken all this into consideration, these results indicate that the physical environment, for example shared lectures or hallways, create a highly interconnects social circle. Physical proximity reflects ambient social exposure rather than social choice.

5.1.2. Digital Communication Networks are Sparse, Selective, and Fragmented

In contrast, both the SMS and call networks are sparse, with densities below 0.005 and are fragmented into many disconnected components: SMS network has 46 connected components and Call network has 54. Only small subsets of students maintain frequent digital communication, most edges have weights, and the networks do not form a unified structure.

Digital communication represents intentional social effort. Therefore, students choose only a small number of close contacts to text or call, forming selective clusters rather than wide structures. These digital networks capture stronger, more meaningful ties but miss the broader social exposure present in physical space.

5.1.3. Physical and Digital Interaction Levels are Largely unrelated

To examine the relationships between physical interactions and digital communication, the weighted degree of each user in *G_prox* was correlated with their weighted degree in *G_sms*. The results show a Pearson correlation of 0.140, a Spearman correlation of 2.81 and a linear regression R^2 of 0.02. Moreover, the scatterplot (Figure 2) reveals no linear trend and only weak monotonic structure.

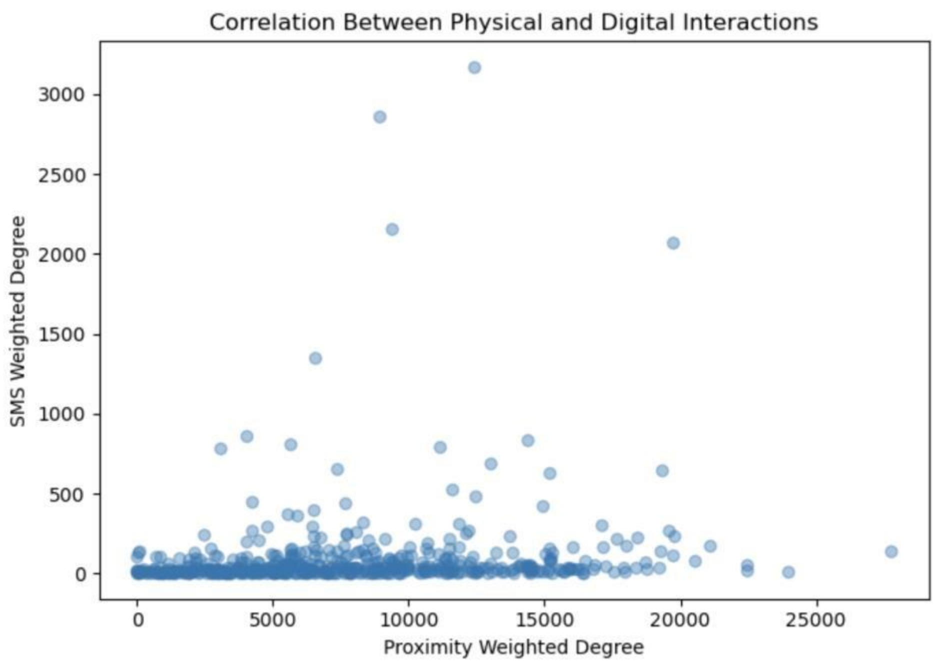

Figure 2. Scatterplot showing Correlation between physical and digital interactions, demonstrating that there is no linear trend.

Only 2% of the variability in SMS activity can be explained by proximity activity, demonstrating that physical presence and digital communication represent fundamentally different behavioral aspects. In other words, students who encounter many peers physically do not necessarily communicate widely digitally, and vice versa.

5.1.4. Structural Segregation and Social Roles

Community detection via greedy modularity maximization identified four communities, but two are very dominant (C0 and C1) which can be better appreciated in Figure 3. Community C0 has 353 students, and community C1 has 334 students, while the remaining two communities only have 3 and 2 students. These two large clusters are accountable for 99% of the student population.

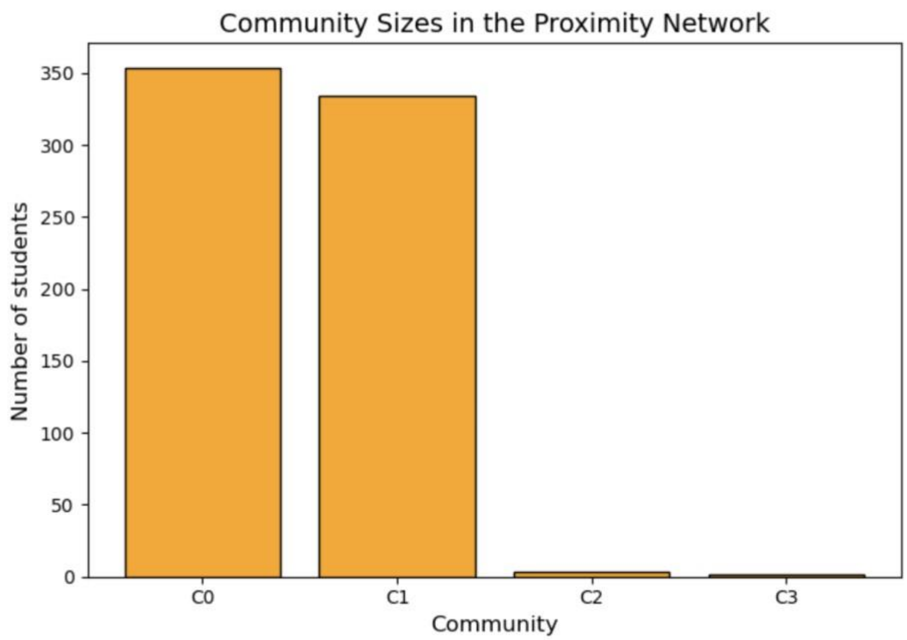

Figure 3. Histogram showing how two communities are responsible for 99% of the student population.

The physical social structure is organized around two major groups, likely reflecting structural forces, for example, housing or similar class schedules. Despite the entire network being connected, social physical interaction is strongly divided into two large blocks. This division provides important insights regarding how information or a virus for example, might spread through the population.

Centrality measures reveal that the students who are most socially active (highest weighted degree) are not the same as those who are most structurally influential (highest betweenness centrality). Student 100 is the most socially active, while student 414 is the most structurally important connector. This demonstrates that high activity does not imply influence. Other structural metrics, like the extremely low clustering coefficient (approx. 0.0013), further suggest a structure where connections are broad but rarely form tight, local groups.

5.1.5. Temporal Patterns

Temporal activity patterns reveal strong differences (Figure 4): Proximity activity peaks earlier (10:00-11:00 and 13:00-15:00), consistent with course schedules and campus movement. Meanwhile, SMS and call activity peaks later (17:00-20:00), suggesting a deliberate social activity after the daytime activities have finished.

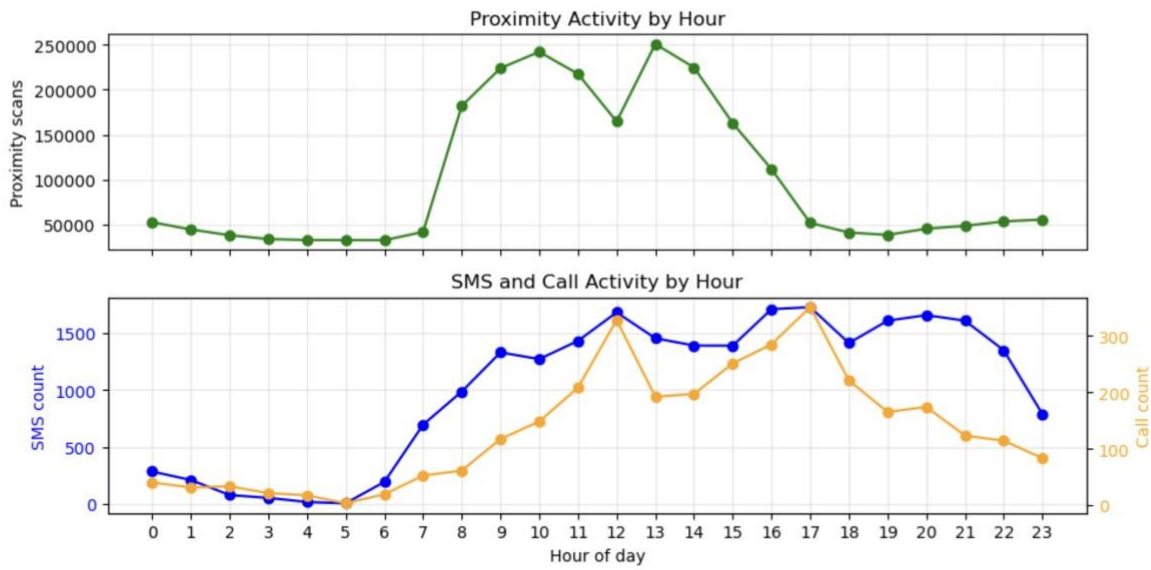

Figure 4. Plots showing the difference of activity on the different interaction channels.

6. Conclusion & Discussion

6.1. Implications of the Findings

The results presented in Section 5.1 highlight clear differences between physical and digital interaction patterns, the structural organization of the student population, and the temporal rhythms that shape the social behavior. These results have important implications for understanding social cohesion, communication strategies, and the design of environments, as well as the design of interventions.

6.1.1. Implications for Social Cohesion and Integration

The primary finding is the existence of two large, dominant communities (C0 and C1) that comprise 99% of the student body.

Despite the overall network being fully connected, this strong two-way division suggests a risk of structural segregation. Information exchange, collaboration, or even disease spread, will happen quickly within each community but may travel slowly between them. This natural clustering can limit integration across the student body.

A few individuals (for example the student with the highest betweenness, ID 414) serve as critical connectors. These individuals are essential for maintaining the cohesion of the overall network because they act as the primary links between the two large communities. Interventions focused on improving interactions between groups should involve these individuals rather than relying on the most “popular” or active students.

6.1.2. Implications for Communication Strategies

The extraordinary low relationship between physical and digital activity (Pearson = 0.14, R^2 = 0.02) indicates that these two modes of interaction reflect entirely different separate behaviors. Physical proximity reflects routine exposure due to shared schedules and spaces. On the other hand, digital communication reflects intentional social effort and stronger personal relationships. Therefore, communication strategies must adapt to these channels' strengths.

Use physical settings to broadcast information broadly and reach many people quickly (for example event notice posting). Use digital platforms to target stronger, more meaningful ties.

Moreover, students with very high weighted degree (for example student 100) are highly active but not structurally positioned to spread information across the entire population. Therefore, influence strategies must prioritize students who can bridge the separate communities.

6.1.3. Implications for Scheduling and Space Design

The distinct temporal patterns suggest that the physical environment is currently dominated by mandatory or institutional scheduling. The late afternoon and evening peaks in SMS and calls indicate that students reserve these hours for intentional communication. On the other hand, physical interactions are constrained by institutional schedules.

6.2. Limitations and Shortcomings

Several limitations should we considered when interpreting the results of this study. First, all networks were constructed as static and undirected, meaning that directionality (who initiates communication) and temporal patterns were lost during the process. For instance, one student may be consistently calling another who rarely calls back, but the current model treats this as a reciprocal relationship. This prevents the analysis of reciprocal relations, evolving relationships, or how the communities change over time.

Second, the interpretation of communities is limited by the lack of external context information. The two large proximity communities might be structured this way based on fixed schedules or residential arrangements. These factors are relevant but are not available in the dataset.

Also, the dataset suffers from boundary effects, which means that only interactions between study participants are recorded. Students who interact frequently with people outside the cohort may appear artificially isolated, which can alter centrality measures and not represent properly their true social activity.

Moreover, the gender analyses are limited by the highly imbalanced gender distribution (614 males vs 173 females). Apparent gender patterns may only reflect the numerical imbalance rather than true homophily (preference for same-gender interaction).

Finally, but very important, the study is correlational, not causal. While proximity and digital activity appear unrelated, we cannot determine why this occurs or which external factors shape the observed patterns. We cannot definitively establish the cause of these patterns. For example, the split of communities may be driven by non-reported, structural factors (rigid class schedules, academic major enrollment, or residential assignment) rather than pure social choice. These confounding variables limit the ability to attribute the structural findings directly to social preferences.

6.3. Future Steps

Future work should be done to improve this analysis in several ways. First, constructing dynamic network, for example weekly data collection, would reveal how communities and central individuals evolve over time, overcoming the limitations involving static approaches. Second, converting SMS and call interactions into directed networks would allow investigation or reciprocity. Thirds, applying formal statistical tests would improve and strengthen the validity of the findings. Moreover, integrating external metadata, such as study degree, class schedules, or residence aera, would help determine whether the two large proximity communities arise from structural constraints rather than social behavior. Finally, analyzing all communication layers as a multilayer network (combining all channels into a single structure) to understand how the overlap. This reveals which relationships and individuals exist across multiple channels and how the structure of the social network changes depending on the communication channel. Together, these steps would provide a more detailed, causal, and comprehensive understanding of student social dynamics.

7. References

Aledavood, T., Lehmann, S., & Saramäki, J. (2018). Social network differences of chronotypes identified from mobile phone data. SpringerOpen. https://doi.org/10.1140/epjds/s13688-018-0174-4

Das, B., Chanda, A., & Roy, R. (2024). The Impact of Smartphone use on Social Relationships among Young People. ResearchGate. https://www.researchgate.net/publication/380759822_The_Impact_of_Smartphone_use_on_Social_Relationships_among_Young_People

Farnsworth, B. (2025). Understanding Human Behavior - a Physiological Approach - iMotions. https://imotions.com/blog/learning/best-practice/understanding-humanbehavior/#what-influences-human-behavior

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science. Perspectives on Psychological Science, 11(6), 838–854. https://doi.org/10.1177/1745691616650285

Introduction to network analysis in Python. (2024). Trenton McKinney. https://trenton3983.github.io/posts/intro-network-analysis/

Ladd, J. R., Otis, J., Warren, C. N., & Weingart, S. (2017). Exploring and Analyzing Network Data with Python. Programming Historian. https://programminghistorian.org/en/lessons/exploring-and-analyzing-network-datawith-python

Mkhitaryan, S., Crutzen, R., Steenaart, E., & De Vries, N. K. (2019). Network approach in health behavior research: how can we explore new questions? Health Psychology and Behavioral Medicine, 7(1), 362–384. https://doi.org/10.1080/21642850.2019.1682587

What is NetworkX? (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/enus/glossary/networkx/

Rotondi, V., Stanca, L., & Tomasuolo, M. (2017). Connecting alone: Smartphone use, quality of social interactions and well-being. Journal of Economic Psychology, 63, 17–26. https://doi.org/10.1016/j.joep.2017.09.001

Sapiezynski, P., Stopczynski, A., Lassen, D. D., & Lehmann, S. (2019). Interaction data from the Copenhagen Networks Study. Scientific Data, 6(1), 315. https://doi.org/10.1038/s41597-019-0325-x

Singh, S., Holvoet, N., & Dewachter, S. (2021). A Relational Understanding of CoEducating and Learning: Information sharing and advice seeking behavior in a dairy cooperative in Gujarat, India. Journal of Co-operative Organization and Management, 9(2), 100150. https://doi.org/10.1016/j.jcom.2021.100150