UNIVERSITÀ DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA IN FISICA

CICLO XXXV

Network reconstruction and prediction of mobility and interaction patterns in social environments: the Parma University Campus as a case study

Coordinatore:
Chiar.mo Prof. Stefano Carretta

Tutore:
Chiar.ma Prof. Raffaella Burioni

Dottorando:
Andrea Guizzo

Anni Accademici 2019/20 - 2022/23

Abstract

Cities represent one of the most fascinating man-made complex systems. The study of the mobility and social dynamics occurring in urban areas has aroused particular interest in the scientific world. The emergence of the Covid-19 pandemic has led the scientific community to study and model social dynamics in public environments, to understand how they influence the epidemic process and to develop effective containment measures. The need to create a sustainable transport system to reduce CO2 emissions and make travel more efficient, has required the study and mathematical modelling of urban travel to determine how travel occurs.

However, mathematical modelling without real application turns out to be a partial and incomplete study. The use of a data-driven approach allows mathematical models to be studied and tested in real social environments using data from widely used electronic devices.

Within this framework, in this thesis we focus of data-driven processes of modeling mobility, transportation, and social dynamics in urban environments. We mainly focus on the study and analysis of the dynamics of group formation and reshuffling and user interactions in public environments, during the Covid-19 pandemic by going to evaluate the effect of containment measures. We used the empirical data from the groups to go test and demonstrate the effectiveness of contact tracing mechanisms, in particular sideward contact tracing, for suppressing epidemic spread. For the urban mobility dynamics part, we propose a model for simulation and prediction of the use of public transportation as a mode of transportation. We show the results of the data-driven models applied to the city of Parma and in particular the effect of the Scientific Campus on public transportation and the social dynamics occurring within it.

The city of Parma and its Scientific Campus represent a particular case for the study of urban and social mobility. The Scientific Campus is a strong center of attraction for users, connected to the city center and in particular to the train station by a few lines. This can cause an overcrowding of buses and transport delays that can lead users to prefer the use of private transport. Analogously, the University campus is subject to significant flows occurring entirely inside it among different buildings or different enclosed spaces. This high social mobility follows social gatherings and courses schedules, creating a temporal network of interactions among users, which is a relevant signal to map and rethink a clever use of the common areas. Moreover, the analysis and modeling of the internal flows has become of crucial interest for the University for the past two years, due to the Covid-19 pandemic.

The traffic flows towards and from the Campus and the social mobility inside the Campus can be studied by analysing massive datasets of GPS data collected

by users with mobile devices and data from WiFi connections of single users to the access points of the University WiFi Network.

Inside the Campus, we analyse the internal flows using WiFi data and we reconstruct the temporal connectivity patterns among users (simplexes), identifying large gatherings. In particular, for each phase of pandemic with different containment measures in the Campus, we were able to measure the probability distributions of groups and link activation at Wi-Fi Access Points, investigating how different areas are used in the presence of restrictions. We consider a recently proposed epidemic model on simplicial temporal networks and we use the measured distributions to infer the change in the reproduction number in the three phases. Our data clearly show that additional measures are necessary to limit the epidemic spreading in the total opening phase, due to the dramatic increase in the number of contacts

We present a new model for generating journeys on public transportation in the urban area, built on an Origin-Destination matrix and on daily activity profile extracted from the analysis of GPS data collected on a wide sample of users. With this model, we can generate scenarios for bus crowding, find routes and times where buses are likely to be overcrowded and find other possible issue. The main goal is to modify some parameters to build possible alternatives to improve public transportation and limit the overcrowding of bus lines to Campus.

Acknowledgements

First and foremost, I am deeply grateful to my family. I express my deep gratitude to my mother Tiziana, for supporting me in my life and studies, encouraging me to follow my passions and desires to achieve this goal. I also especially thank my sister Laura, for taking care of me during my childhood and always being there for me, especially during the most difficult times. Special mention also to my father, who cannot attend this milestone but of which he would be deeply proud.

I am profoundly grateful to my girlfriend Sandy, for always being sunny, instilling a lot of confidence in me and giving me confidence on the day we met. You broke into my life and now have a special place in my heart, I hope to have you by my side forever.

Thanks also to my friends and my college classmates, with whom I had a great time together and who allowed me to grow as a person. I also thank all the friends scattered around the world who have been part of my life even for a small moment.

Last but not least, I am deeply grateful to my supervisors, Prof. Raffaella Burioni and Prof. Alessandro Vezzani, for their guidance during these three years. They have guided me inside the world of research with valuable advice and discussions of scientific topics. In addition, they gave me the opportunity to have research experiences in different groups in which I was able to improve my skills and meet people who stimulated me to give my best. During these years, I had the opportunity to collaborate with fantastic people who contributed significantly in the realization of this thesis. First of all, I want to thank Prof. Vittorio Loreto for giving me the opportunity to intern with the sustainable city group of Sony CSL in their office for more than six months. A special mention to Bernardo Monechi and Bruno Campanelli who supervised me in the implementation of the work described in the last chapter of this thesis, I also thank Hygor Piaget and Matteo Bruno for all the breaks and happy moments spent in the lab. I also want to extend my sincere thanks to Claudio Castellano and Marco Mancastroppa for their collaboration and stimulating discussions during my very first research project.

Thank you Covid-19 for making me miss two years of conferences and summer school, I hope I never see you again.

Preface

This thesis was mainly developed at the Department of Mathematical, Physical and Computer Sciences of the University of Parma, in collaboration with the Italian National Institute for Nuclear Physics (INFN, Gruppo Collegato di Parma, Sezione Milano Bicocca).

The results discussed in this thesis are based on the following pubblications:

- Marco Mancastroppa, Andrea Guizzo, Claudio Castellano, Alessandro Vezzani, Raffaella Burioni, *"Sideward contact tracing and the control of epidemics in large gatherings"*, J. R. Soc. Interface.1920220048 (2022)
- Andrea Guizzo, Alessandro Vezzani, Andrea Barontini, Federico Russo, Cristiano Valenti and Raffaella Burioni, *"Simplicial temporal networks from Wi-Fi data in University Campus: the effects of restrictions on epidemics spreading"*, Frontiers in Physics, 10 (2022)

The original research presented in this thesis is the result of scientific collaborations with the Institute for Complex Systems (ISC-CRN) in Rome, and the Sustainable Group of Sony CSL in Rome.

Contents

List of Figures

Chapter 1

Introduction

The physical world, including both microscopic and macroscopic systems, as well as human activities, often consists of numerous interconnected elements. These interactions result in the emergence of collective and cooperative behaviors that cannot be fully understood by examining the individual components alone. The study of these **complex systems** spans multiple disciplines, from statistical physics to computer science, and from biology to economics [10, 73, 78, 79].

Many complex systems can be represented as networks, where nodes represent the elements and edges represent their interactions. However, in some systems, these interactions are not static or fixed, but instead evolve over time. Examples of such systems include the interactions between neurons, the public transport system, and social interactions. These dynamic systems are referred to as temporal networks, as they require explicit consideration of the temporal dimension. This is crucial, as the continuous destruction and recreation of edges over time greatly impacts the topological properties and introduces new temporal features [50–52, 73].

The interplay between network dynamics and dynamical processes taking place on the network is complex and significant. If the two processes occur at similar time scales, the dynamic process is greatly impacted by changes in the network structure. This is because the dynamic process follows the connections of the network, which are themselves evolving. Additionally, the dynamic process may induce adaptive changes in the network, further complicating the relationship between the two dynamics. To fully understand these systems, it is necessary to consider the temporal aspects of the interactions and the adaptive interactions between the dynamic process and the network. This leads to the need for adaptive temporal networks, as the dynamics of both the network and the dynamic process have a reciprocal impact on each other. The mathematical modeling of these systems is challenging due to the complex interplay and the non-trivial effects of adaptive mechanisms on the dynamic process [43, 44, 96].

The spread of *epidemic processes on the social interactions network* is a prime example of the interplay between network dynamics and dynamic processes. In this case, the time scale of the epidemic is comparable to that of social dynamics, and the outbreak of the disease leads to adaptive behaviors in the population, such as changes in behavior due to symptoms, increased risk awareness, and the implementation of control measures. To effectively tackle epidemics, it is important to consider both the temporal dynamics of social interactions and the adaptive behaviors induced by the disease [36, 43, 85].

The beginning of my PhD project coincided with the onset of the COVID-19 pandemic. This pandemic significantly impacted the approach to epidemic control: initially it required the rapid development of containment policies with the introduction of strict lockdowns to limit the spread of the virus and its consequences in terms of health overload and deaths [4, 28]. Simultaneously, it required the development of tracking strategies to keep the population active and avoid economic, psychological and social consequences of excessively restrictive measures [47]. In addition, the pandemic allowed for empirical data on the effectiveness of measures to restrict tracking strategies, implemented by different states in a heterogeneous manner [88]. These empirical data allowed us to demonstrate that contact tracing is a key strategy for mitigating disease transmission without restricting social activity [35, 49].

The modeling of social dynamics is further complicated by the higher-order nature of social interactions, which are often organized in groups and gatherings [15], leading to potential superspreading events (SSEs) of epidemics. These events play a key role in the spread of diseases like SARS-CoV-2, making it important to address them with control measures [11, 55]. Contact tracing is a strategy that can control SSEs without completely prohibiting large gatherings. The impact of contact tracing in gatherings has yet to be fully investigated. In this study, we assess its impact by modeling the transmission of SSEs on a simplicial adaptive activity-driven network with tracing applied on simplices [55]. Our results show that the traditional **forward** and **backward** tracing methods are enhanced by a third mechanism called **sideward** tracing, which occurs laterally by exploiting the simplicial structure of interactions. This mechanism is critical in tracing large gatherings and is especially useful in strategies targeting them. The model is also applied to an empirical dataset of gatherings at the University of Parma, which estimates the optimal size of gatherings to be traced to control the epidemic without disrupting teaching and research activities.

Besides epidemic models, another example of network dynamics interfacing with process dynamics is provided by mathematical modeling of human mobility. Human mobility is a complex and dynamic phenomenon that has been of great interest to researchers from various disciplines. Understanding human mobility patterns and behavior is crucial for a range of applications, such as urban planning [91, 114, 116], transportation engineering [39, 57, 107], public health [8, 22, 110], and social sciences [25]. Over the years, researchers have developed mathematical models to capture and analyze the patterns of human mobility.

The journey of human mobility modeling starts from early empirical models based on the assumption of random walk [9] or other simple statistical distributions. With the advancement of technology and the availability of large-scale mobility data, the focus shifted towards data-driven models that can better capture the real-world human mobility patterns. These models use various techniques, such as graph theory [72], network analysis [53], and machine learning [66], to extract meaningful insights from the data. In recent years, there has been a growing interest in developing mathematical models that can capture the social dynamics of human mobility, such as the influence of friends, family, and coworkers on an individual's mobility behavior. These models go beyond the traditional data-driven models by considering the underlying social relationships and connections between individuals.

Public transportation plays a crucial role in creating sustainable cities [74]. It

provides a reliable and efficient mode of transportation for large numbers of people, reducing the need for individual cars on the roads. This helps to reduce traffic congestion, air pollution, and greenhouse gas emissions [77]. Public transportation also has the potential to reduce social inequality [92,113] by providing access to job opportunities, education, and healthcare for those who may not have access to a private vehicle. Additionally, investment in public transportation infrastructure can lead to economic growth and job creation. Overall, public transportation is a crucial component in the development of sustainable cities that are both environmentally and socially responsible.

The development of models for describing commuting by public transportation in the city is crucial for a detailed description of the collective and individual behavior of the population. Situations of overcrowding, delays, and inefficiency of public transportation leads the citizen to choose private transportation resulting in congestion of the road network and increased pollution. Moreover, public transportation has been significantly impacted by the COVID-19 pandemic [34,71,101]. The virus has led to a decrease in demand for public transportation due to fear of exposure, travel restrictions, and work-from-home mandates. As a result, many transit systems have experienced decreased revenue and have had to make significant changes to their operations, including reducing service levels, implementing stricter cleaning protocols, and promoting social distancing measures on vehicles and in stations.

The pandemic has also accelerated the shift towards alternative modes of transportation [42], such as private vehicles and active transportation, which has resulted in increased congestion and emissions in some cities. This has led to calls for increased investment in public transportation as a way to support a sustainable and resilient transportation system that can withstand future disruptions. While the COVID-19 pandemic has brought significant challenges to public transportation, it has also provided an opportunity to rethink the role of public transportation in creating sustainable and resilient [30,108]. A careful planning of public transportation, becomes a key factor in the ecological transition of cities, limiting overcrowded situations during rush hours, giving access to public transportation to those who cannot afford private transportation, and limiting the spread of viruses by promoting social distance between passengers.

With the increase in available data sources, mathematical modeling of real phenomena, such as epidemic models or mobility models, is no longer sufficient, and the ability to apply the models on real cases using available data such as trajectories or attendance within the same building plays a key role in model validation. Data-driven approaches play a crucial role in testing mathematical models, as they provide real-world data that can be used to validate the model's predictions. A mathematical model can only be considered reliable if it can accurately reflect real-world scenarios and phenomena. The use of data-driven methods ensures that the model is tested and evaluated with actual data, rather than relying solely on theoretical assumptions. This leads to a more accurate and robust model, as well as a better understanding of the underlying relationships and mechanisms that govern the system being modeled. Additionally, data-driven approaches allow for the continuous improvement of mathematical models as new data becomes available, providing an ongoing process of validation and refinement.

The thesis is organized as follows. Chapter 2 review of the main concepts of data-driven approach with Wifi data, especially as a proxy for the study and analysis of

presences and social dynamics in public environments. In the final of the chapter, the dataset used in this work obtained from the University of Parma WiFi network, the case study of this thesis, will be briefly introduced. In the Chapter 3.1, we introduce the basic concepts of High-Frequency Located-based data as the best proxy for the analysis of human mobility, and then a first statistical analysis of the HFLB data for the province of Parma will be done. The chapters 4-6 are the results of original research. In the Chapter 4 we present the results described in Ref. [45], introducing a new formalism for the analysis and monitoring of spaces in public environments in which a WiFi network is presented as a proxy for the prevention and containment of infectious diseases, specifically going to study the case of the University of Parma and going to see the effects of restriction measures during three different phases of the COVID-19 pandemic that occurred in Italy. Chapter 5 is based on Ref. [69], introduces contact tracing on groups with a simplicial activity-driven network and we tested the effects of manual contact tracing using empirical data from groups in the University of Parma. Chapter 6 is based on work that is still under development and introduces a new mathematical model, which is based on empirical trajectory data, for predicting public transportation use in metropolitan areas. The model, in addition to reproducing the real scenario, allows for the creation of new scenarios with the goal of improving the efficiency of public transportation. We will use the metropolitan city of Parma and its University Campus as a case study.

Chapter 2

WiFi data-driven approach for social dynamics

The study of social dynamics in public environments has become increasingly important in recent years, as understanding how individuals move and interact in these spaces can provide valuable insights into urban design, public health, and other fields. One method for studying social dynamics in public spaces is the analysis of WiFi network data. WiFi network data can provide information on the presence and movement of individuals within a public space, as well as their interactions with their environment. The data can be collected by monitoring the signals emitted by WiFi-enabled devices, such as smartphones, as they connect to WiFi networks within a specific area.

Previous studies have shown that WiFi network data can be used to infer the flow of foot traffic in public spaces, the popularity of different areas, and the duration of time spent in specific locations. For example, WiFi network data can be used to identify areas of high congestion and low congestion, which can inform the design of public spaces to optimize the flow of people. Additionally, WiFi network data can be used to identify patterns of behavior that are specific to certain times of day, such as peak hours of foot traffic in a shopping district.

Furthermore, WiFi network data can be used in combination with other data sources, such as social media data, to gain a more comprehensive understanding of social dynamics in public spaces. For instance, social media data can be used to analyze the sentiment of people in a specific area, which can inform the understanding of the perception of the area. Overall, WiFi network data can provide valuable insights into social dynamics in public spaces, and can inform the design and management of these environments to promote safety, health, and well-being for all individuals.

2.1 The limits of WiFi data

Passive WiFi data collected at the access point (AP) level can provide a general understanding of people localization and co-location, however, it is not a precise measure. Factors such as signal strength and the number of connected devices can affect which AP a device connects to, resulting in devices in the same room connecting to different APs [94] or devices in different rooms connecting to the same AP. Therefore, passive WiFi data alone may not be sufficient for accurate

people localization.

While there are technologies available that improve WiFi-based localization, such as special sensing applications installed on devices or special sensors on APs, these are not widely available and deployed. Instead, this study aims to investigate the results obtained from WiFI infrastructure that is currently deployed in most buildings and public spaces, without relying on these additional technologies. It is acknowledged that due to localization errors, some of the results described in the next section may present miscounts. However, it is believed that in the context of large university buildings, which typically have large classrooms and halls [19, 60, 118], over long observation periods these errors will likely average out, as it is more likely to be connected to the AP in the room where the device is actually present.

The results of this study are based on the comparison between different time periods characterized by different opening and closure phases. Therefore, it is likely that possible errors are cancel-out, since the number of miscounts due to limitations of our technologies should affect the different phases in the same manner. Moreover, the advantages of using "standard" passive WiFi data, both in terms of widespread applicability and in terms of allowing long-term, indefinite observation periods outweigh the limitations. Despite data limitations, it is believed that the results and conclusions hold and the method provides reliable results.

2.2 A case study: WiFi data in Campus of University of Parma

The University of Parma, similar to many universities and public institutions, has equipped its buildings and spaces with a unified WiFi network (see Fig. 2.1), allowing all users to establish more than 10000 sessions per day. The ICT services office of the University of Parma is responsible for collecting all session data from the login management system, which manages all wireless access points and all users' requests for connection to the internet with their registered devices. The staff of the ICT service office are authorized to access files containing personal data and they carry out an anonymization process to protect the privacy of individuals.

2.3 Preprocessing and privacy-preservation mechanisms

All data collected from the unified University WiFi network is obtained by the "ICT services" office (IT) through the login management system. The IT extracts a tabular file, referred to as a log file, on a daily basis, where each row represents a connection start or end event. The log file contains user information, device information, and session information. To reconstruct users' connections, the following attributes are particularly relevant:

- **Username**: unique identifier for each user, the email address is used;
- **Type of user**: from the e-mail, we can identify the type of user: student, university staff or external guest;

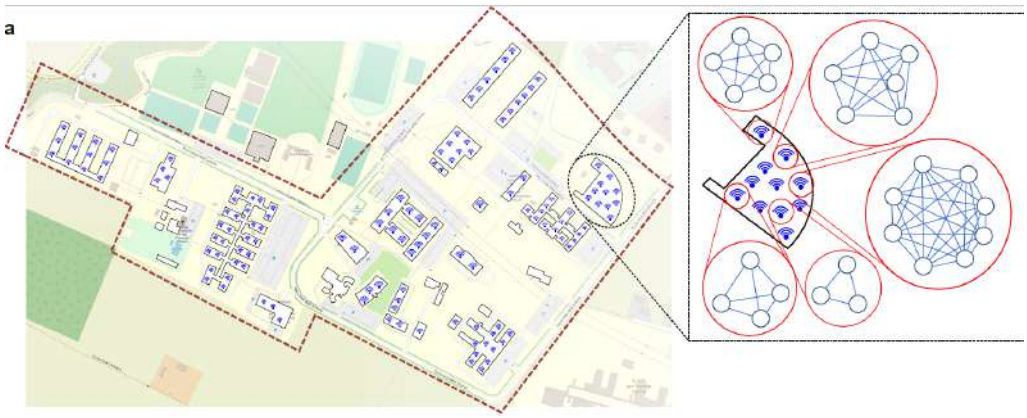


Figure 2.1: **WIFI data for gatherings at University of Parma.** The map of the University of Parma's Campus, known as the"Parco Area delle Scienze", shows the location of scientific departments with facilities for teaching and research activities such as classrooms, laboratories, study spaces, and libraries. It also schematically highlights the buildings where gatherings are recorded through WIFI access point data.

- **Calling device ID**: MAC address of device to distinguish all user's devices;
- **Device type**: this attribute allows distinguish user's device (computers, smartphones or tablets);
- **Called station ID**: MAC address of the AP to which the device wants to connect;
- **Status type**: this attribute indicates whether this accounting request marks the beginning of the user service (*Start*) or the end (*Stop*);
- **Date-time**: this attribute represent the day and the hour for the accounting request;
- **Session ID**: this attribute is a unique accounting ID to make it easy to match start and stop record in a log file.
- **Terminate cause**: this attribute indicates how the session was terminated, and can only be present in Accounting-Request records where the Status type is set to *Stop*.

From these data, it is possible to reconstruct all the connection sessions to the University Wi-Fi network, including their duration and location based on the APs maps.

To ensure compliance with privacy regulations, such as the European General Data Protection Regulation (GDPR) and the Regulation on Privacy and Electronic Communications, I have implemented privacy-preservation mechanisms. Specifically, I cannot use the log files directly as they contain personal information. I have conducted a Data Protection Impact Assessment (DPIA) to adhere to data minimization principles and ensure compliance with the latest Regulation on Privacy and Electronic Communications [23]. Recital 25 of the GDPR specifically addressing

WiFi monitoring, states: *Service providers have emerged who offer physical movements' tracking services based on the scanning of equipment related information with diverse functionalities, including ... ascertaining the number of people in a specific area ... for which the consent of end-users is not needed, provided that such counting is limited in time and space to the extent necessary for this purpose. Providers should also apply appropriate technical and organisations measures ... including pseudonymisation of the data ... erase it as soon it is not longer needed for this purpose. Providers engaged in such practices should display prominent notices located on the edge of the area of coverage informing end-users prior to entering the defined area that the technology is in operation ....*

2.4 Anonymization and aggregation data algorithm

I implemented a new algorithm with the aim of removing all sensitive data and replacing it with hexadecimal strings generated with a randomising algorithm. Each day, at midnight, a file is generated from the WiFi network logger system with all the data of the connections to the WiFi network that occurred that day, then the university's IT department starts the script for the anonymization process to generate the output file which then allows me to analyse the data for the study of space utilisation and social dynamics within the Parma University and its Scientific Campus. This anonymization script was completely **designed**, **implemented** and **tested** by me during my first year of my PhD. Once the testing phase was completed, the algorithm was provided to the IT department, which carries out the daily process of removing sensitive data and then forwards the file generated as output by the script to me. In this way, I only have access to the file without personal data in accordance with European GDPR laws. Each line in the logger file contains an event of starting or ending a connection and contains the following sensitive data: username, mail, mac address, and session id. The anonymization algorithm performs the following operations:

- **Username.** The algorithm checks if the row matches the first connection made by the user. If it is the first connection, the algorithm generates a randomly generated 16-digit hexadecimal string that replaces the original username and saves the username-string match to a dictionary. If the user has made connections previously, the hexadecimal string is retrieved from the dictionary in which it was saved in the previous anonymization process.

- **Type of users.** The algorithm add a new attribute called "type of user" where I save the email domain in order to distinguish the user between students, professor or external guest.

- **Calling station id.** This attribute corresponds to the MAC address of the user's device and should therefore be anonymized. As with the username, this field is replaced by a randomly generated 16-digit hexadecimal string that is later saved in a dictionary.

- **Session ID.** This attribute, which is used to join the two connection start and end lines, is generated by the logger system according to the username and calling

station id. Therefore, the algorithm processes and anonymizes the attribute as in the case of username and calling station id.

- **Consistency.** The algorithm preserves all correlations between attributes even in different anonymization processes by going to save attribute-random string dictionaries.

The anonymization algorithm just described, outputs a file similar to the file generated by the wifi network logger system, in which sensitive data are no longer contained. After a testing phase, I improved the algorithm to increase performance and then delivered the script to the IT department.

Since the dataset may contain connections that are occasionally interrupted for short periods due to various reasons such as weak signal or a user's device being in standby, I established a protocol for handling these instances. If two or more consecutive connections of a single user to the same AP are present and the interval time, denoted as τ, between these connections is less than 5 minutes, I considers it as a single connection, with the start time being the first connection and the stop time being the last connection.

After establishing the protocol for handling short connection interruptions, I I developed a script that calculates the various measures that are central to our analysis (all data are calculated first on the whole population and then separately for staff members and students):

- *Group Size*: the number of devices s_i connected to a given (i^{th}) AP for each 15 minutes. Devices must remain connected to the AP for the whole time.

- *Presence*: the total number of individual connected to the University WIFI during each 15 minutes interval.

- *Group Size Statistic*: the total number of 15 minutes clusters of a given size s present in the whole University during each day of observation.

- *Average number of link*: the average number of links $\langle s(s-1)/2 \rangle_i$ formed at the AP i^{th} during each day of observation. The average is computed over the 15 minutes intervals during the considered working day.

- *Different Link Number*: the daily number of different links l_i (i.e., unique pairs of users) connected to a given (i^{th}) AP for more than 15 consecutive minutes. If two users meet two or more times at the same point but at a different time, this is counted only once.

- *Statistic of Link duration*: the statistic of daily link duration in each access point (number of connection having a certain duration in unit of 15 minutes).

- *Total number of different links*: the total number $\mathcal{L}_d$ of different links formed daily in the whole university. If two users meet two or more times in the same day even at different APs, this is counted only once.

Any counters that fall below the threshold (5 individuals / 10 links in this specific implementation) are removed to prevent the potential disclosure of the presence or

absence of specific individuals or links. The resulting anonymized counts are the only data that are permitted to leave the IT department for further analysis.

The algorithm just described was implemented to anonymize the dataset extracted from the WiFi network management system of the University of Parma. However, with simple and quick modifications it is possible to adapt the algorithm to other WiFi management systems present in other universities or public environments.

2.5 Dataset

After the process of anonymization, the dataset analyzed in this study includes information from a sample of 696 wireless access points (APs), 19,749 users (broken down into 16,505 students, 1,968 staff members, and 1,276 external guests), and approximately 15,000 daily connections. The data covers a period of ten months, beginning on December 10^{TH}, 2020 and ending on November 7^{TH}, 2021. Due to the ongoing COVID-19 pandemic, this period can be divided into three distinct phases: a closing phase, a partial opening phase, and a total opening phase for the 2021/2022 academic year.

- **Closing phase.** During this phase, access to University buildings was restricted to staff, faculty, and students who were participating in laboratory activities. All classes were conducted remotely. This phase starts on 10^{TH} December 2020 ends on 21^{TH} February 2021 and starts again on 15^{TH} March 2021 and ends on 18^{TH} April 2021.

- **Partially opening phase.** During this phase, access to University buildings was limited to first year students to allow them to attend lessons in-person (about 25% of students enrolled in degree courses). This phase began on February 22^{ND}, 2021, ended on March 14^{TH}, 2021, and resumed again on April 19^{TH}, 2021, before finally ending in June.

- **Total opening phase.** During this phase, access to University buildings was granted only to those with a valid Green Pass, which was required to attend in-person classes. This phase starts on September 27^{TH} 2021 and ends with our dataset.

In Chapter 4 we will we will describe in detail the quantities just described and use the dataset to present in detail the study and analysis of university space occupation and social dynamics within the University of Parma.

Chapter 3

High-Frequency Location-Based data-driven for human mobility

The volume of digital traces that log human activities has seen a tremendous increase in recent years, thanks to the proliferation of mobile devices, remote sensing and the digitization of various human activities. This has allowed researchers to examine human behavior in social-technical systems at an unprecedented level of detail, providing a wealth of data to study the evolution and dynamics of large-scale complex systems such as social networks, urban systems, patent records, and trade networks. Complex systems refer to systems whose emergent features are more than what one would expect from the sum of their individual components. For instance, social networks are a prime example of a complex system, where people may look for information and connections in relatively similar manners, yet their nontrivial network of interactions gives rise to complex phenomena such as the emergence of communities, viral content, or echo chambers.

Similarly, human mobility, though seemingly simple to define as the desire to move from one place to another, features social, economic and temporal constraints that shape the collective mobility patterns in unexpected ways. The new digital sources that log the movement of people allow us to study and characterize how these interconnected factors affect individuals' daily mobility patterns with an exceptional level of detail. From GSM phone locations based on linked antenna data [17, 40, 65] to credit card transaction sequences [64], the recorded data has improved in temporal and spatial resolution, even to real-time GPS traces [81], enabling the modeling of urban traffic [9], exploration strategies [100], and statistics on visit sequences [83].

This wealth of data has opened the door for the application of advanced tools and techniques borrowed from natural language processing, information theory, and statistical physics, to studying and modeling human mobility patterns. This endeavor has not only led to the statistical description of human mobility [12, 37] and the modeling of short- and long-scale commuting [98], but also deepened our understanding of the strategies that individuals apply when exploring spaces [2]. By doing so, it is possible to uncover invisible boundaries that segregate people into different groups, featuring diverse accessibility to public transport [13], different visit rates to different venues [20, 29] and spatial segregation [93].

A particular lens through which to measure and characterize different behaviors in a population under investigation is to inspect their activity allocation strategies in social spaces [95] or spatial areas [1, 2]. The temporal, cognitive, and economic

constraints that apply to people's daily social interactions [75] and movements [83] deeply affect their statistical signatures. Besides the natural limitations given by cognitive capacities, other social and economic restraints such as income, education, and culture may impact how people explore their surrounding spaces [20, 29, 64].

A possible framework to describe such exploratory dynamics is that of the adjacent possible [59], which states that all the ideas, concepts or places one can visit are grouped into three separate regions: (1) the actual, accounting for all the tokens that users have already discovered and experienced; (2) the adjacent possible, encompassing all the concepts and tokens that are just one step away from the actual, and that can be easily reached or discovered; and (3) the distant possible, encompassing all the concepts and tokens that are farther away and that may require more effort or resources to reach or discover. This approach allows us to understand the limitations and possibilities of human exploration and innovation, and how these are shaped by the complex interplay of individual and collective factors.

Furthermore, the use of these digital traces has also allowed for the creation of detailed maps of human mobility patterns, which can be used to inform urban planning, transportation infrastructure development and emergency management. These maps can also reveal patterns of inequality in access to resources and opportunities, and can be used to design policies and programs that promote more equitable and sustainable outcomes.

In conclusion, the wealth of digital data on human activities has provided a valuable resource for understanding the evolution and dynamics of large-scale complex systems such as social networks, urban systems, patent records, and trade networks, and has led to a deeper understanding of human behavior and strategies. The use of advanced tools and techniques borrowed from natural language processing, information theory, and statistical physics has allowed us to uncover previously hidden patterns and trends, and to inform the design of more equitable and sustainable policies.

3.1 Geolocated data as a proxy for human mobility

The recent abundance of digital data capturing human mobility patterns has spurred a wide range of research aimed at understanding the underlying mechanisms that shape people's movement habits. This work specifically focuses on the spatial diversity of anonymous, opted-in users' behavior in urban and rural areas. Urban and extra-urban environments are characterized by a wide range of behaviors that are reflected in the varied ways citizens move through their surroundings, which are closely connected with individuals' socioeconomic characteristics. In recent years, many studies have attempted to uncover this link, by inferring these characteristics from historical data about human mobility.

One example of this approach is the use of call detailed records (CDRs)to predict the social status of individuals in Rwanda [14]. The study showed that a broad set of features describing individuals' call activity patterns, social networks, and mobility patterns could predict their wealth and whether or not they owned certain commodities. Such studies are particularly relevant in developing countries, where surveys from national statistics institutions are not readily available. Similarly,

has been demonstrated a link between socioeconomic indicators and mobility data, showing that features extracted from CDRs are good predictors of income per capita and accessibility to social deprivation [84].

Geolocated data from developed countries offer even more possibilities to understand the social structure of cities and the dynamics of their social groups. The combined use of points-of-interest data and human-flow-line-based data can reveal a great deal about how different social groups explore different parts of a city. For example, [76] shows how individuals living in poor and wealthy areas typically visit different locations in several cities across the United States. Economic inequalities can also be studied using geolocated credit card data, although this data is not typically available in many developed cities. For instance, [29] used credit card transactions and Twitter data to show how social interactions are segregated among different social groups, and [20] showed how credit card transactions can be used to identify individuals' socioeconomic characteristics.

Gender inequalities are also reflected in mobility patterns. For example, Gauvin et al. (2020) used CDRs in the metropolitan area of Santiago, Chile, to show that women tend to move less and explore a smaller subset of regions compared to men.

Geolocated data is also valuable from perspectives other than socioeconomic. These data also contain information about other aspects of the underlying structure of cities, such as land use, and the daily dynamics of individuals during normal and exceptional conditions.

One of the most valuable applications of geolocated data is the ability to extract so-called origin-destination (OD) matrices. An element of an OD matrix, M_{ij}, represents the number of individuals commuting from area i to area j during a typical day. Typically, national statistics institutes construct OD matrices by conducting periodic surveys of a given area's population, but this process is slow and costly. The same information can be inferred from CDRs [38, 80] at a lower cost and with higher temporal resolution.

While OD matrices are useful for transport planning and accessibility studies, geolocated data offers a more granular level of information. For example, it is possible to infer the mode of transportation an individual used during a trip. In Sadeghinasr et al. (n.d.), the authors combined CDRs with public transport data to cluster individual trips according to the mode of transportation used.

3.2 High-Frequency Location-Based data analysis

High-frequency location-based data, such as GPS coordinates and timestamps, collected via applications installed on GPS-enabled phones, have proven to be a powerful tool for understanding human mobility. These data sets, such as Microsoft's GeoLife data set, allow researchers to measure mobility at unprecedented levels of temporal and spatial resolution. However, in order to fully utilize these data sets, it is necessary to filter, analyze, and enrich them with various techniques.

One way to do this is to extract information on the purpose of trips and stops. A trip's purpose can be determined by analyzing the type of origin and destination locations, for example, whether a trip is a home-to-work commute. Similarly, the purpose of a stop can be inferred by examining the semantic category of the Point of Interest (POI) where the user makes the stop, such as a commercial center or restaurant.

To further enrich these data, additional information can be obtained by combining them with other data sources, such as demographic data, weather data, or transportation data. This can provide a more comprehensive understanding of the factors that influence human mobility, and enable more accurate predictions of future mobility patterns. Additionally, advanced data visualization techniques can be used to present the mobility data in a way that is easily understandable and actionable.

Overall, high-frequency location-based data is a valuable resource for understanding human mobility, and by filtering, analyzing and enriching these data with various techniques, researchers can gain deeper insights into individual behavior, and the factors that shape it.

3.2.1 Trajectory analysis

When working with location-based data, it is important to carefully consider the level of temporal and spatial resolution that is desired for the analysis. [63] conducted a comprehensive study of different spatial and temporal aggregation strategies, and found that there is a trade-off between a fine-grained resolution (both temporal and spatial) and the level of statistical significance that can be achieved. The smaller the areas and time bins used to aggregate the data, the fewer users will fall within one bin. This trade-off applies to both the stability of land-use identification and the identification of home and work locations, which are important for constructing origin-destination (OD) matrices of daily commuting.

Another important filter is to keep only users with relevant statistics, in terms of active days or daily behavior. The selection of users will depend on the focus of the research. For example, if the goal is to detect home-work commuting OD matrices, it may be desirable to keep users with more active days and retain their home and work stops [3, 56]. In contrast, if the goal is to reconstruct the population's response within an area to extreme natural events or massive popular happenings, it may be useful to include temporary visitors such as tourists in the analysis.

It is also important to note that, depending on the focus of the research, different methods can be used to define home and work locations, such as using the location where a user spends the majority of time or using the location of maximum activity. Therefore, it is crucial to carefully evaluate and choose the appropriate method based on the research question and data availability.

Once the set of events and users to be analyzed has been defined, the next step is to translate a sequence of anonymized unique IDs (UIDs) and latitude and longitude coordinate pairs (latitude, longitude) into a series of stops and trips. Stops are defined as a group of registered positions that are contiguous in space, meaning that the maximum distance between any pair of points in the group is less than a given distance D_s, for at least a given time T_s.

A possible approach for identifying stops is to use the k-medoids algorithm, a variant of the k-means clustering algorithm [6]. This method involves selecting one point in the sequence of a user and a radius D_s. All the points within distance Ds are grouped together as a cluster and their mean location is calculated, serving as the new starting point. This process is repeated until the mean point does not move anymore, resulting in the identification of the first user location. The points within distance D_s from this point are then removed from the total group, and the

process is repeated with the remaining points until all the user's locations have been identified.

The previous method for identifying stops does not take into account the temporal ordering of the logged events. This means that points relating to a single location may be spread out over different times of the day or year, depending on the time frame of the data set. This can lead to the inclusion of points from two different visits to two separate venues in the same area as belonging to one unique location.

To address this issue, one solution is to consider the points in the order in which they were recorded, as done in Project Lachesis [48]. This approach involves defining a minimum duration time of a stop T_s and a maximum diameter D_s that points belonging to the same stop can have. The algorithm starts by looking at the first point in the sequence at time t_i and then examines all the points up to time $t_i + T_s$. If the diameter (the maximum distance between all pairs of points within this time period) is larger than D_s, it is not considered a valid stop, and the process is repeated with the next point in the sequence. If the diameter is less than D_s, it is considered a valid stop, and points are added to the stop group until a point is encountered that increases the diameter above D_s. After that step, the stops between the first and last valid points are annotated, and the analysis begins again using the first non-valid point. The stop is then assigned to the medoid of the sequence, which is the point i that minimizes $d(i) = \sum_j \in Sd(i, j)$, where S is the set of points in the stop. The time of the stop is assigned to be within the initial and final points' timestamps.

There are more advanced techniques to filter out noise from the raw data, such as filtering points based on the speed between two consecutive records [115, 117]. These selection rules are often available in specialized libraries for ad hoc analysis, such as the one found in [82]. The resulting sequence of stops for each user can then be used to compute the temporal density of users in a specific area of the city or to infer land use [104], travel demand [103], or spatial route estimation [58] of different areas of the city.

3.2.2 Location analysis

Once the sequence of a user's stops has been identified, it may be useful to determine the user's locations by grouping the stops that belong to the same visit of a single location. For example, a user's workplace or a recreational venue. One method for doing this is the method defined in [48], which involves clustering all the individual stops of a user using the DBSCAN algorithm. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. It groups together points in a dataset that are close to each other in terms of distance, while also ignoring points that are isolated or located in sparse areas of the data.

The algorithm works by defining a neighborhood around each point, and then grouping together all points that have a sufficient number of nearby points within their defined neighborhood. This forms clusters of high-density points, and points that are not part of any cluster are considered noise. DBSCAN has several advantages over other clustering algorithms, such as being able to find clusters of arbitrary shape and being able to identify noise points. It does not require the number of clusters to be specified in advance and it does not require the scale of the data to be

specified.

DBSCAN requires two parameters: ϵ and s. ϵ is the radius around each point that defines its neighborhood, and s is the minimum number of points required to form a dense region. The algorithm starts with an arbitrary starting point that has not been visited, finds all points within ϵ distance of the starting point and if they have s within ϵ distance, a cluster is formed. At the end of the analysis, all stops belonging to the same location (cluster) are assigned a unique label, and the coordinates of the location are typically set at the medoid of the cluster's stops. This medoid is the point of the cluster that is the most centrally located among all the stops of the cluster. In this way, the DBSCAN algorithm returns a set of unique labels for each location visited by the user, providing a clear picture of the user's mobility patterns. A typical distance between two stops belonging to the same location should be used, usually around $\epsilon \approx 200m$. Additionally, the number of neighboring points (s) can be adjusted to filter out noisy locations by requiring a user to visit a location more than once for it to be considered a valid location.

The DBSCAN algorithm is a powerful density-based clustering method, but it has some limitations when it comes to identifying user's locations. One of the possible shortcomings of DBSCAN is that when stops are spatially dense, it may connect two locations by mistake. Additionally, DBSCAN analyzes the set of stops user by user, as mixing stops from different users may increase the density of stops in a region and create an artificial larger and wider cluster of stops for a single location.

To overcome these limitations, Aslak and Alessandretti proposed an alternative approach that uses the Infomap algorithm [7], developed by Rosvall and Bergstrom [90], which is typically used to detect communities of nodes in large networks, to cluster stops. In this approach, the stops are represented as nodes in a weighted undirected network. The weight of a link between two stops is set to the reciprocal of the distance between the two stops, provided that they are within a typical distance r_2 (otherwise there is no link between the two stops), which can be tuned to find larger or sharper locations. The closer two stops are in physical space, the more strongly they are connected in the network.

This approach has several advantages over DBSCAN, as it allows all the stops of a given dataset (even coming from different users) to be considered at once, and it appears to provide better and more refined locations than the DBSCAN method. By creating a weighted network of stops, it allows to better capture the relationships between stops, and by using Infomap it allows to find communities of stops that are likely to represent a user's location.

Once the user's stops have been identified and grouped into locations, the remaining points in the user's sequence that are not assigned to a stop can be considered as trips. These trips can provide valuable insights into road usage patterns, origin-destination (OD) matrices, and route-selection mechanisms. By analyzing the patterns in the trips, researchers can gain a better understanding of how users move through a given area.

Additionally, the sequence of stops from different users can be used to measure the dynamic activity profile of different areas. This includes the number of people found in a given district at a given time. These activity profiles can then be transformed through normalization procedures to compute a residual activity, which allows for the inference of land use in different areas by clustering areas with similar

temporal profiles together. Another approach is to use the similarity in land use between two districts based on the correlation matrix between the raw activity profiles. Clusters of areas with similar land use can then be found by applying the Infomap algorithm on the similarity matrix. This approach gives a better understanding of how land use is related to human activity and how it changes over time.

3.2.3 Enriching Locations

Once the points have been assigned to stops and each stop has been assigned to a location, it is useful to assign meaning to each location. The first step is to identify a user's most important places, such as their home and work locations. There are several methods to do this, but they all focus on determining where a user spends most of their time during the night (home) or day (work) periods. It's worth noting that the work location may correspond to a school for a student or a place where an individual spends most of their time during working hours (such as a recreational or employment center for an elderly or unemployed person).

A first possible method is the one proposed in [63]. In this method, the location visited during most of the daytime hours is assigned to the work location, while the location visited most of the nighttime hours is assigned to the home location. A more restrictive condition can also be applied as a filter and require that a fraction δ_h of the nighttime and daytime hours also be assigned to the home or work location for the location to be considered valid.

Another possible method [103] is to calculate the time spent at each location between 7a.m to 8p.m on workdays and the time spent at each location between 8p.m and 7a.m on nights. To limit the noise given by occasional visits, the location work may not be assigned to the user if the user does not visit that location at least once a week or if the location is too close to the home. The remaining locations may be assigned according to the category of service provided.

The second method has the advantage that it makes it possible to measure OD matrices by selecting trips based on home or work [3, 103, 119] and to characterize the mobility patterns of users and the purpose of these trips [40]. In addition, the distinction between daily and occasional commuting makes it possible to apply models based on [98] or gravitational law [3] and to describe both short- and long-scale statistical properties of human commuting [62]. We can predict the temporal activities of locations [31] or develop data-driven models generating synthetic trajectories of users [119] from the statistical feature of human commuting.

3.2.4 OpenStreetMap POI data

One way to add context to the identified stops is to classify them based on the type of location they represent. This can be accomplished by utilizing private data from companies that specialize in geolocation services and venues (such as Moro et al., 2019) or by utilizing open data sources such as Open Street Map (OSM). OSM is a non-relational database that is collectively populated by volunteer users from around the world. In this case, we will use data from the Overpass API of the OSM project [5]. Each venue listed in OSM has a mapping of key and value features that can be used to determine the type of location. For example, bus stops are listed as "highway.bus_sto" while hospitals are labeled as "amenity.hospital" or

"building.hospital". However, given the large number of possible key-value combinations (more than 800), it is necessary to simplify these attributes and map them to a more manageable set of categories. To accomplish this, we developed a custom mapping from these strings to nine well-defined categories:

- **Services Public**: all public service locations, including city hall, tax offices, and job centers;
- **Services Private**: all services catering to individual needs, such as legal, financial, and architectural services and so on;
- **Transport**: all major transportation hubs, including train and bus stations, metro stations, and airports, with the exception of local bus stops, which may be located too closely to other points of interest;
- **Shop**: all place where food is not sold such as jewellery, sport and so on;
- **Groceries**: all place where food and good for nutrition are sold, including supermarket;
- **Leisure**: all the rest of commercial places (restaurant, cinemas) and recreational venues;
- **Education**: all the school and universities found in the territory;
- **Health**: all the health buildings, such as hospitals, family care doctors and so on;
- **Industrial**: all locations that represent production sites and do not fall into the previous categories, such as car manufacturing plants or mining sites.

We use a custom-developed algorithm to transform each point of interest (POI) from the Open Street Map (OSM) database into a point in space with latitude and longitude coordinates. For non-point-shaped POIs, we use the centroid as the projected point to simplify the analysis and improve the efficiency of matching stops to locations. This approach eliminates the potential for overlapping polygons and ambiguity when assigning a stop to a location. Each stop that is not designated as home or work is assigned to the closest POI category. If the POI is within 200 meters of the user's stop, we retain that location category as the stop's meaning; otherwise, the stop cannot be associated with a specific location. We separate "Home" and "Work" as distinct categories that pertain to all stops a user makes in their home and work clusters of stops. This method allows us to tentatively identify the purpose of each user stop even when transaction records are not available [20].

3.3 A case study: trajectories analysis in Parma province

In this section we analyze the original dataset of HFLB data acquired from Cuebiq by Sony CSL group. Sony CSL acquired location data collected from the GPS signal of smartphones from Cuebiq covering all of Italy and France for the year 2017. The

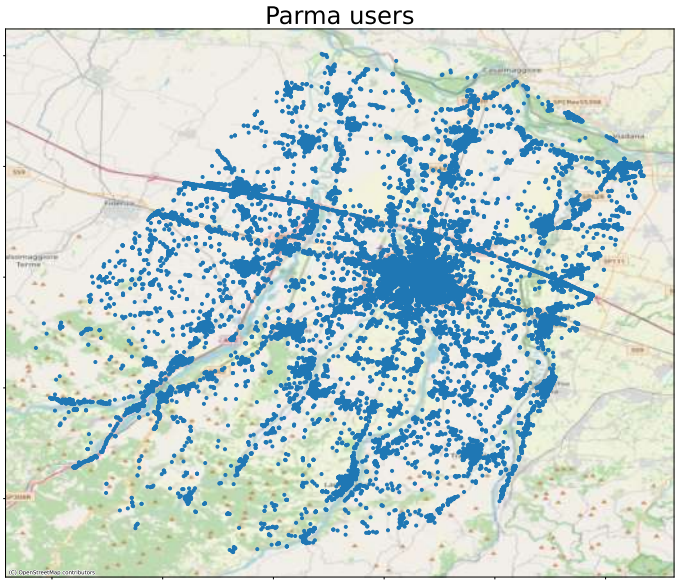


Figure 3.1: **Location of users from HFLB data** The plots show all locations recorded by 105682 users. We can observe an high point density along the main street of the province.

sustainable city group of Sony CSL provided me the data for the province of Parma, from which I performed a preliminary analysis of the statistics of the number of users, location of their work and home locations in order to improve my ability to manage large amounts of data and to design a process to filter the data to keep only users with good statistics.

The entire dataset, collects data from 105682 different users who at least once passed through the province of Parma and had their locations recorded. Fig. shows the locations recorded by Cuebiq. It can be seen that location registration is very dense in the city of Parma, in the population centers of the towns located in the province. In addition, a high sampling of positions can be seen along the main arterial roads of the road network and in particular, along the highway. Of these users, not all are residents of the province, some have only made a temporary passage through the area of interest for reasons of work, vacation, etc. Our goal is to mathematically model commuting trips that occur in the Parma area, so we are only interested in users who live in the entire province and with high location registration. In this way it is possible to go and get a statistical analysis of users that is not affected by fluctuations due to tourists, outside workers, or from users with a low registration rate. For this reason, we calculated:

- **activity:** for each user, is the number of distinct days on at least one record of their location;
- **activity on working day:** for each user, is the number of distinct working days in which they have at least one record of their location;
- **home day:** number of days in which the user has at least one record in the location identified as “home”;
- **work day:** number of days in which the user has at least one record in the location identified as “work”;

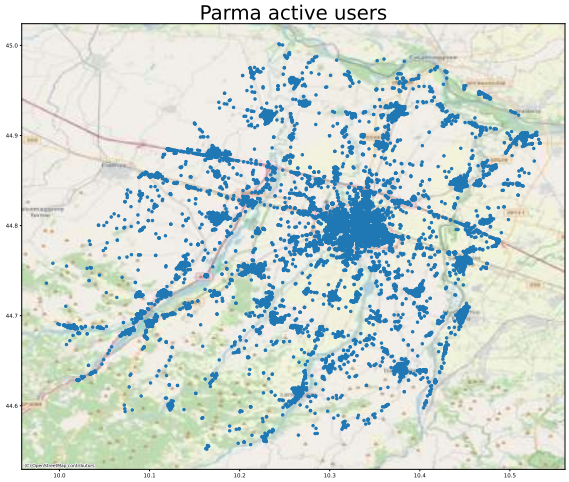


Figure 3.2: **Location of active users from HFLB data** The plots show all locations recorded by 80636 users with at least 15 working days of sampling in province of Parma. The density point has decreased along the main roads like highway or "via Emilia".

From these measures, I eliminated users with activity on work days less than 15 days. By analyzing the quantities described above, I observed a correspondence between work days and active days, so it was sufficient to fit a filter only on active days. The result is a dataset of spatial coordinates related to 80636 users, which were shown in Fig. 3.2. Point density has decreased throughout the province of Parma compared to Fig. 3.1 and in particular, along major roads such as the "via Emilia" and the highway. From this reduced dataset, it is possible to derive a statistical analysis of the province's residents' trips that is good and limits errors due to inactive users that do not allow commuting trips to be identified. From this dataset, we extract all the quantities needed for the model that will be described in Chapter 6.

After selecting the most active users, I did a robustness analysis of the process of determining the location of the users' home described in Sec. 3.2.3 I compared the Census population data for the neighborhoods of the city of Parma and compared them with the users who were assigned a home location within the neighborhoods. The HFLB dataset consists of a fraction of all the residents in the city, so I normalized the population obtained from the census data and the HFLB data to compare them with each other. In Fig. 3.3 we can see that the method for assigning home location has a distribution of users among different neighborhoods in good agreement with the distribution given by the census data. The method has an overestimation of population in the city center and other neighborhoods, probably due to students or people who do not result with the population census. In other neighborhoods, on the other hand, we can see an underestimation of the population but still a value that we can consider good for our objectives. Moreover, Fig. 3.4 shows that the population data obtained from the HFLB data are correlated with Pearson's correlation factor $r^2 = 0.78$ with the population data provided by the census. From Fig. 3.3 and Fig. 3.4, we can say that the method used to determine home and work leases is a good one for the province of Parma and does not require further implementation.

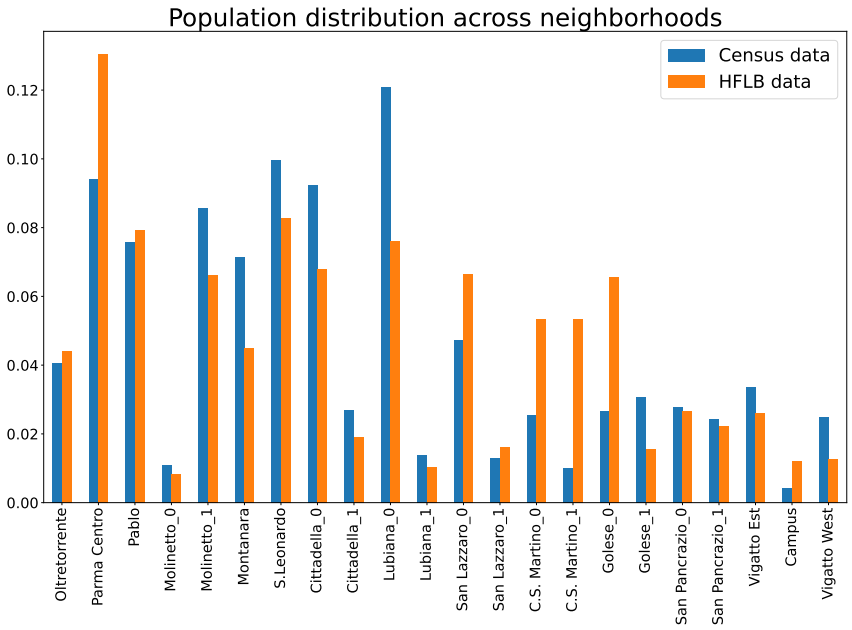


Figure 3.3: **Comparison of population distributions in Parma city.** We compared the population distribution obtained from assigning "home" locations to users in the HFLB dataset with the population distribution given by the Census data. We can see that the two distributions are similar to each other but with an overestimation of the population in the city center.

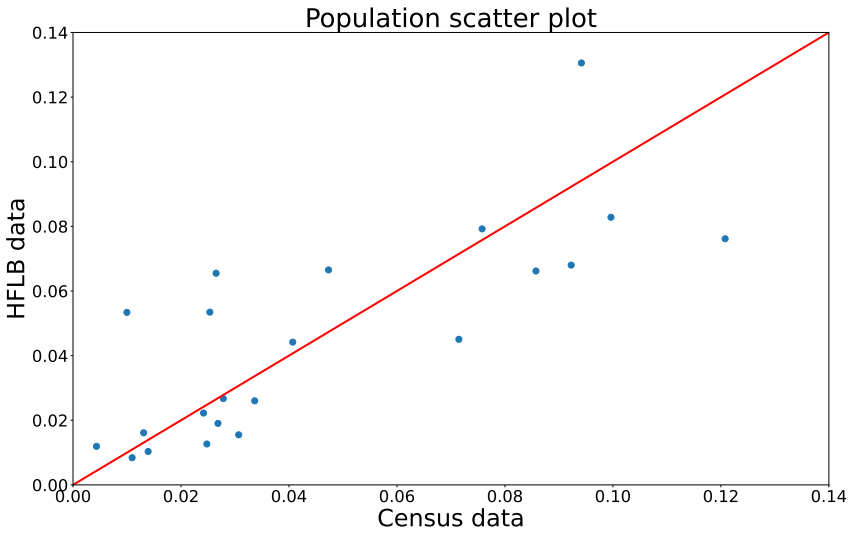


Figure 3.4: **Scatter plot of population distributions in Parma city.** In relation to the population distribution shown in Fig. 3.3, we have reported a scatter plot of the census data and the data derived from home location assignment from the HFLB data. The data are well correlated and have a Pearson correlation factor of $r^2 = 0.78$. This agreement between the two distributions allows us to verify that the method of assigning rent house and jobs is very efficient with our dataset.

Chapter 4

Analysing Campus occupancy and social dynamics via passive WiFi sensing

In this chapter, considering the problem of pandemic control, we use WiFi data from the University Campus in Parma (Italy); described in detail in Sec. 2.2, to analyze patterns of large gatherings and high-traffic areas on the campus. We analyze anonymized data of devices connected to the access points (APs) of the network to track usage patterns in public areas across the campus. Our focus is on the probability distribution of large gatherings and the number of unique device connections (links) in specific areas of the campus, which may indicate potential risks for dangerous contacts. We examine three distinct phases, each characterized by different containment measures, including a closing phase, a partial opening phase, and an open phase, where University activities resumed almost fully. Based on our analysis, we rank the APs based on their potential danger and examine how different groups of users utilized the spaces during the three phases.

4.1 Attendance data and temporal behavior

From the connection sessions data, we can estimate the time evolution of attendance at a single access point (AP), in a specific University building or area, or across the entire University campus. To achieve this, we extract the number of users connected to each AP as a function of time, every minute. The sum of all AP's temporal series corresponds to the total presence for the entire University or for a specific area within the University. Our algorithm also extracts the dynamic behavior of attendance for different user groups, such as students, faculty and staff, and external guests. In Fig. 4.1, Fig. 4.2 and Fig. 4.3, we graphically represent attendance data at the University during different phases of closures. We find that while the presence of structured staff remains relatively stable, the presence of students increases significantly in the final phase of reopening.

Fig 4.4 illustrates the daily attendance patterns at the University during the closing (panel **a**), partial opening (panel **b**), and total opening (panel **c**) phases. The attendance curve shows a sharp increase in the morning, a decrease during lunchtime, an increase in the afternoon, and a decrease in the evening for all three phases. In the closing phase, staff members have the highest attendance, which remains similar in

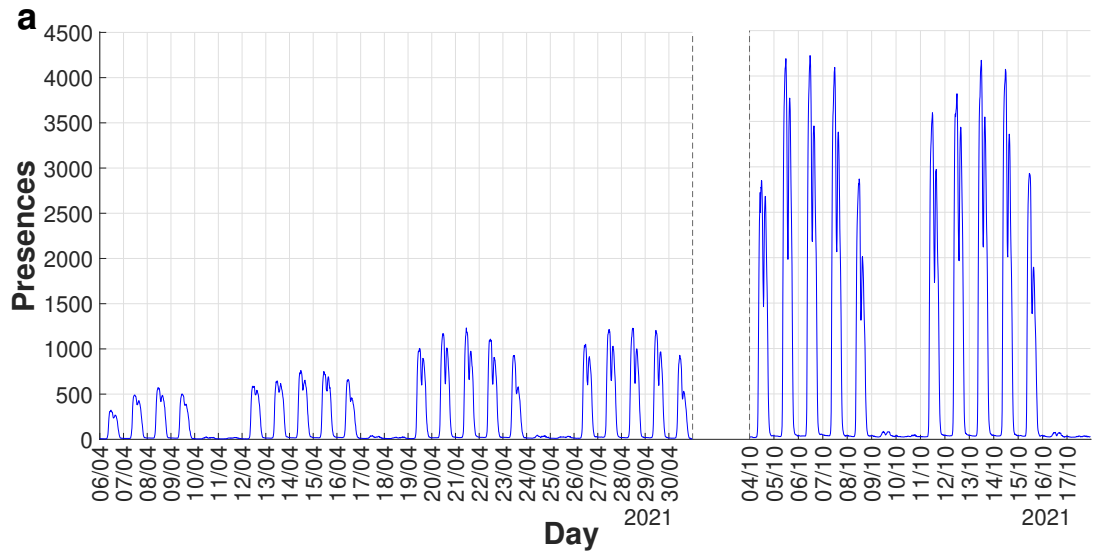


Figure 4.1: **Temporal evolution of attendance at Parma University.** The plots show the dynamics of attendance for all users during the different phases of closures. The first two weeks depict the closing phase, the middle two weeks depict the partial opening phase, and the last two weeks depict the total opening phase. The data also displays the lower attendance during weekends and holidays.

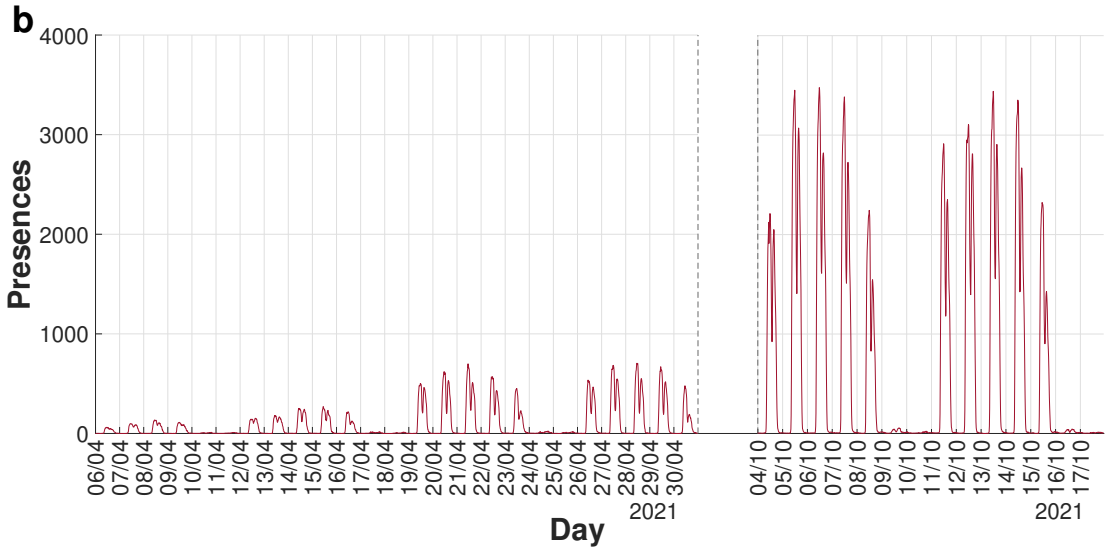


Figure 4.2: **Temporal evolution of attendance for students at Parma University.** The plots show the dynamics of attendance students (Panel B) during the different phases of closures. The first two weeks depict the closing phase, the middle two weeks depict the partial opening phase, and the last two weeks depict the total opening phase. We can see a significant increase in attendance during these phases. The data also displays the lower attendance during weekends and holidays.

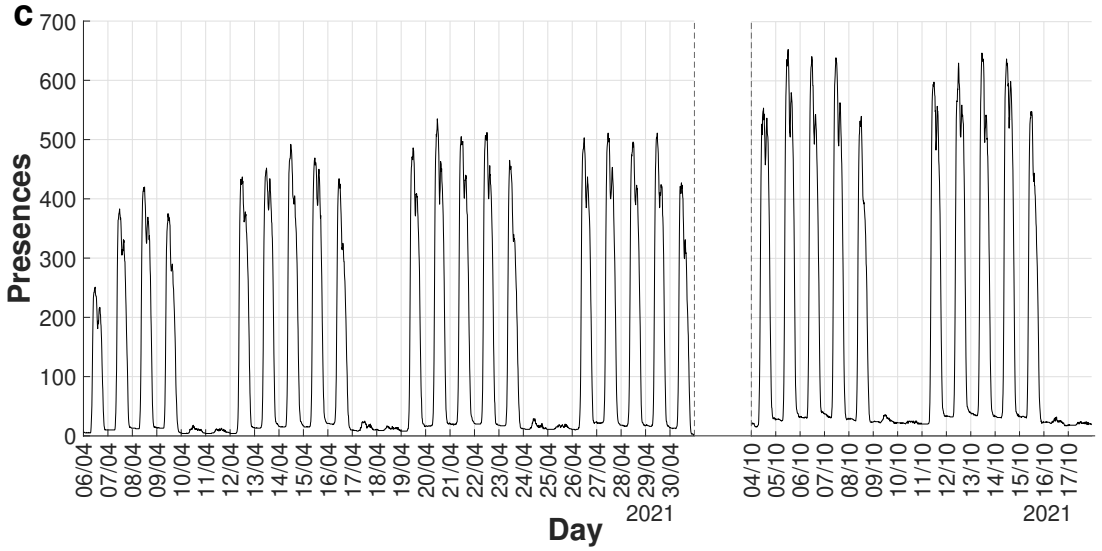


Figure 4.3: **Temporal evolution of attendance for structured staff at Parma University.** The plots show the dynamics of attendance for structured staff (Panel C) during the different phases of closures. The first two weeks depict the closing phase, the middle two weeks depict the partial opening phase, and the last two weeks depict the total opening phase. We can see that the attendance of structured staff remains relatively consistent across all phases. The data also displays the lower attendance during weekends and holidays.

the partial opening phase. However, in the total opening phase, student population dominates the attendance patterns. Additionally, the student population displays a fluctuating pattern with peaks during morning and afternoon classes, while the presence of staff members is more consistent throughout the working hours.

4.1.1 Calibration

One limitation of using WiFi data to measure the number of users is that it only captures individuals who are connected to the WiFi network, which may lead to an underestimation of the total number of people present. This is because some people may not be connected to the WiFi network or may be using their own mobile data instead. In ordr to obtain a calibration for a WiFi data, we compare the attendance data with data obtained from the badge access to a specific building (the Physics building). For this calibration we compare the temporal behavior of attendance for structured staff in the partial opening phase and only for the data from Physics department. Indeed, only in this case the use of the badge was compulsory. Fig 4.5 shows an underestimation for the attendance given by WiFi data, about by a factor 2 compared to the attendance given by badges data. We also compare WiFi data from APs near a specific classroom, "Aula Newton" within the Physics building, to online seat reservations for face-to-face lessons during the partial opening phase. We find a similar underestimation of the WiFi data by a factor of 2 when compared to the online seat reservations.

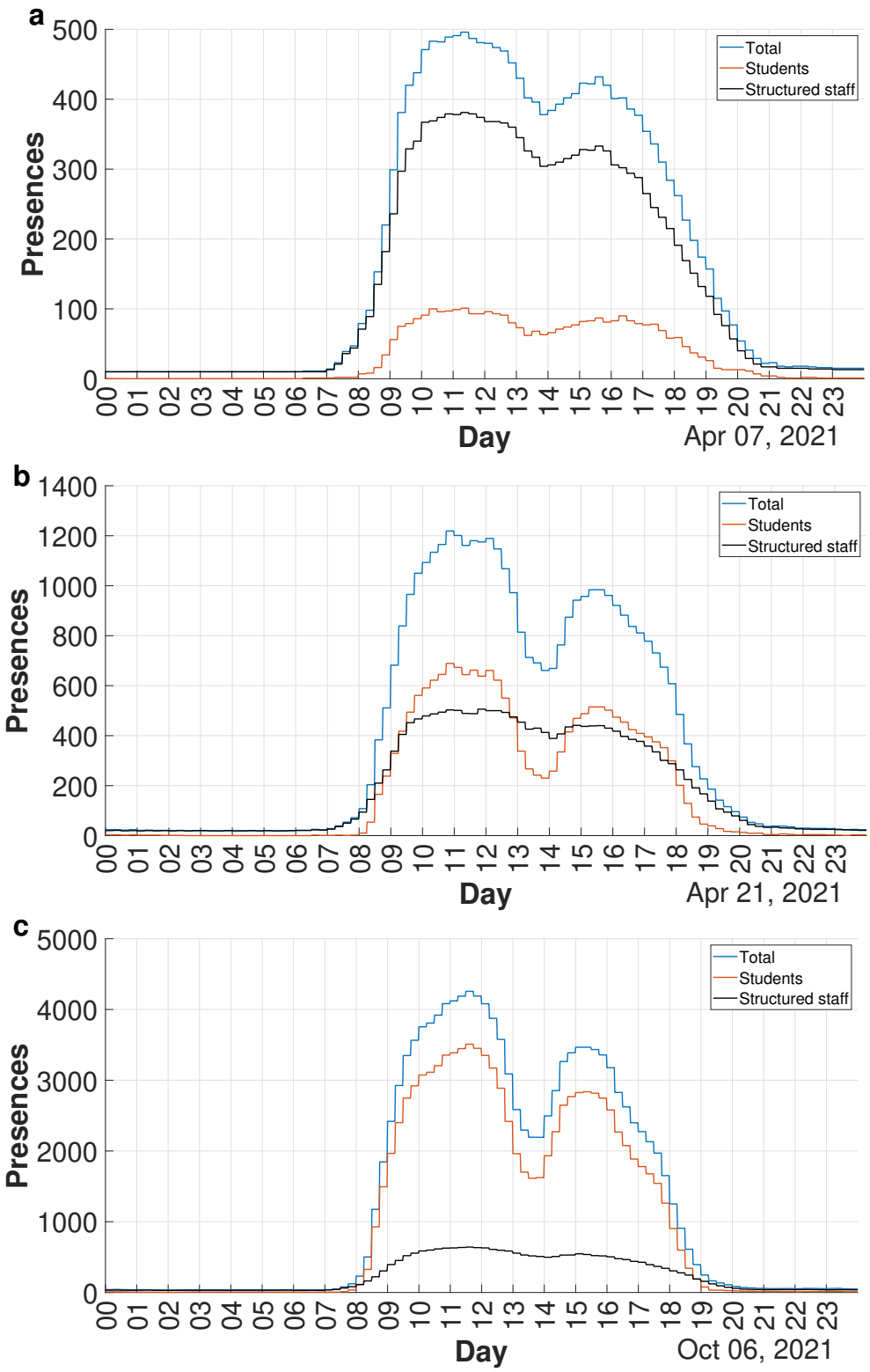


Figure 4.4: **Presences in Parma University during a typical working day. a** Temporal evolution of attendance during a typical working day in the closing phase. During this phase, staff members have the highest presence and student attendance is minimal. In panel **b**, we see that students and structured staff have a similar level of presence, while in panel **c**, we see that in the total opening phase, attendance is primarily driven by the student population.

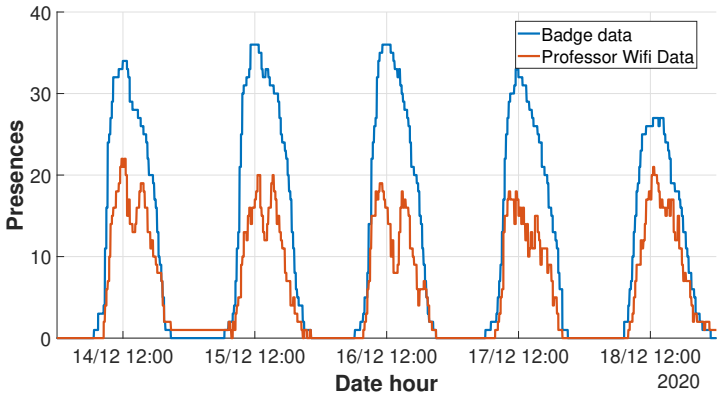


Figure 4.5: **WiFi data calibration.** The temporal evolution of attendance for structured staff, as obtained from badge reader (blue line) and WiFi data (red line), is compared during a typical working week. The attendance data collected through WiFi is found to be underestimated by approximately a factor of 2 when compared to the attendance data provided by badge access.

4.2 The simplex size distribution

During the Covid-19 pandemic, controlling the spread of the virus through limiting large gatherings of people has been a key focus, particularly in situations where tracking attendance is challenging. One innovative solution to this problem is the use of WiFi data. This data can provide a natural way of monitoring the presence of large groups (simplices) within a certain area by considering the number of people simultaneously connected to the same access point (AP).

To analyze this data, the study first extracted the group size distribution (P(s)) from WiFi data connections during working hours (from 8.00 am to 7.00 pm). In social systems, the duration of face-to-face contacts typically displays a broad distribution with complex behaviors. However, the infection process occurs on a characteristic timescale, which for COVID-19 has been estimated to be around 15 minutes. Therefore, the study defined a simplex of size s as a group of s people connected to the same AP for at least 15 minutes. This is the same time interval used in contact tracing apps.

To ensure that the results were not influenced by this choice, the study also verified that the main results were qualitatively independent of the time interval by varying it from 5 to 30 minutes. For each AP, the study split working hours into 15-minute intervals and found the number of users connected to that AP for the entire time interval. This number corresponds to the simplex size. Disconnections from a single AP shorter than 5 minutes were discarded from the data set.

The study also distinguished the three phases of the pandemic period with different levels of restriction (closing, partially opening, and total opening phases) and found three different distributions of group sizes (Fig. 4.6). The results clearly showed that a broader distribution of P(s) was observed in the opening phases and the maximum group size grew from $s_max^C = 36$ to $s_max^PO = 58$ and to $s_max^TO = 174$ in the closing, partial opening, and total opening phases respectively. Furthermore, the distribution P(s) in the total opening phase was found to be compatible with a power law $f(s) \propto s^{-\nu}$ with an exponent $\nu \approx 1.65$.

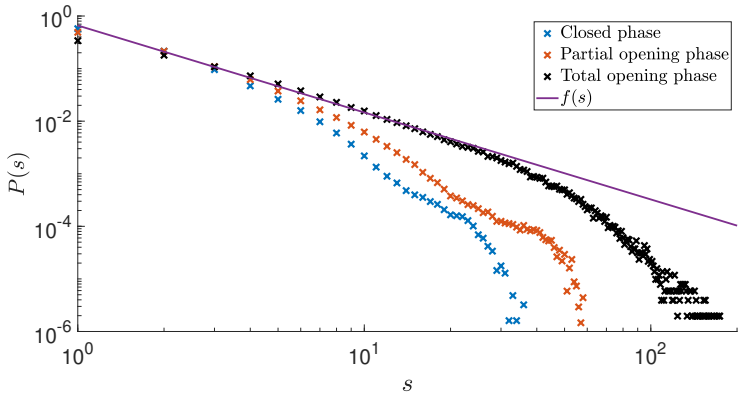


Figure 4.6: **Group size distribution in the University buildings.** $P(s)$ is the probability to find a group of s people that are connected to the same AP for almost 15 consecutive minutes. Plot are reported in log-log scale and we plot the group size probability distribution for each phase with different restriction regimes. The distribution $P(s)$ in the total opening phase is compatible with a power law $f(s) \propto s^{-\nu}$ with an exponent $\nu \approx 1.65$.

The study also plotted the different contribution of University staff and students to the group size distribution (Fig. 4.7). The results showed that the restrictions in the closing and partial opening period mainly affected the behavior of the student population, whose simplex distribution was strongly modified in the different phases, while for staff members, the differences were very limited.

In conclusion, this study demonstrates the potential of WiFi data in providing valuable insights into the presence of large groups during the Covid-19 pandemic and how these patterns change over time as restrictions are implemented. The findings can inform public health policy and decision-making during pandemics, as well as better understand how different populations are affected by restrictions. Overall, this study illustrates the effectiveness of WiFi data as a powerful tool for monitoring and understanding the spread of infectious diseases, and in informing public health policy and decision-making during pandemics.

4.3 Effects of restrictions on the reproduction number of the epidemic model on simplicial temporal network

In the previous section, we derived and analysed probability distributions $P(s)$ of group size. These distributions can be used to quantitatively determine the measure of the effects of the restriction measures implemented by the University of Parma to mitigate the spread of Covid-19 in University environments during the pandemic. In this section, starting from a Susceptible-Infected-Recovered (SIR) model on activity driven simplicial networks [69, 87], we quantify the impact of the closure measures implemented by the University during the Covid-19 pandemic by estimating the value of the reproduction number R_0. The same analysis was performed in the section 5.4 for the model including CT strategies.

The model we are using describes the evolution of an interaction network where

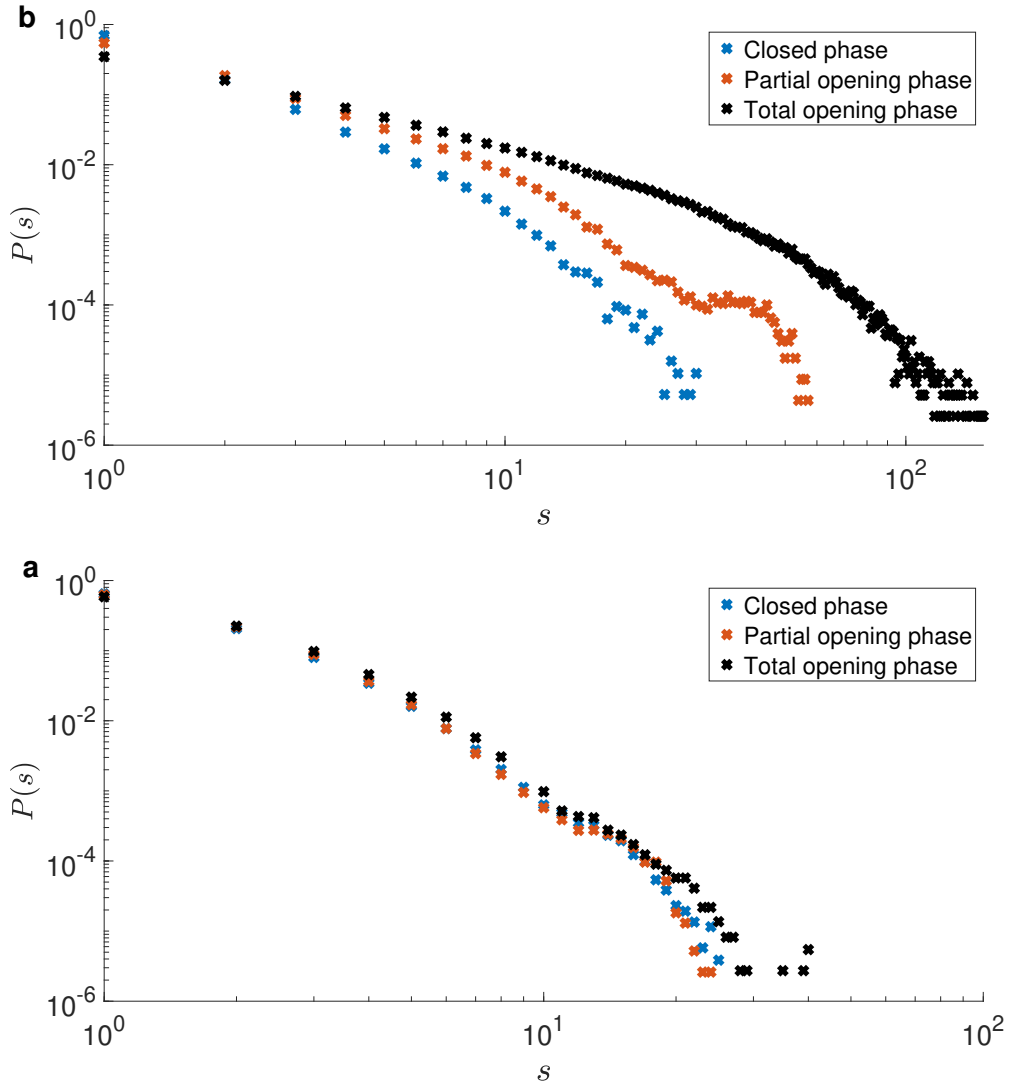


Figure 4.7: **Group size distributions comparison.** As shown in panel **a**, the contribution of university staff to the group size distributions in the three phases with different restriction regimes is depicted. The impact of the restrictions in the closing and partial opening phase is minimal on the staff population. In contrast, panel **b** illustrates that the restrictions have a significant impact on the behavior of the student population, with their group size distribution being greatly altered across the different phases.

simplices (clusters of individuals) are activated at a rate a. When a simplex of size s is active, s nodes are chosen randomly to participate, creating $s(s-1)/2$ interactions. Once the interaction is complete, the cluster is dissolved and the process is repeated. The distribution of simplex sizes, $P(s)$, captures the heterogeneity in the size of gatherings. Each susceptible node within the cluster has a probability of λ of being infected by the infected nodes within the same cluster. Additionally, each node has a chance to recover at a rate μ. The spread of the epidemic is determined by the basic reproduction number R_0. If $R_0 < 1$, the epidemic will not spread and only a limited number of individuals will be infected. However, if $R_0 > 1$, there will be an exponential increase in the number of infections, potentially infecting a significant portion of the population. Because the network is constantly reshuffled as simplices are recreated, a mean field approach can accurately describe the evolution of the system:

$$\begin{aligned} \partial_t S(t) &= -S(t) \int asP(s) \left[1 - (1 - \lambda I(s))^{s-1}\right] ds \\ \partial_t I(t) &= -\mu I(t) + S(t) \int asP(s) \left[1 - (1 - \lambda I(s))^{s-1}\right] ds \\ \partial_t R(t) &= \mu I(t) \end{aligned} \tag{4.1}$$

In Eq.s 4.1, the term $\mu I(t)$ represents the recovery rate of the process $I \to R$ and the integral describes the infection process. The activation rate of susceptible nodes into a cluster of size s is represented by $asS(t)P(s)$ and the probability that a susceptible node activated in a cluster of size s becomes infected by one of the remaining $s-1$ individuals of the cluster is represented by $\left[1 - (1 - \lambda I(s))^{s-1}\right]$. The integral is calculated over all possible cluster sizes. The stability of the solution where all nodes are susceptible (i.e. $S = 1$, $I = 0$, and $R = 0$) can be studied by linearization. Specifically, by linearizing the second of Eq.s 4.1 we can analyze the stability of the solution. In particular, the linearization gives

$$\partial_t I(t) = (-\mu + a\lambda\langle s(s-1)\rangle)I(t). \tag{4.2}$$

with $\langle s(s-1)\rangle = \int s(s-1)P(s)ds$ and the solution with $I = 0$, $R = 0$ and $S = 1$ is stable only if $\mu > a\lambda\langle s(s-1)\rangle$ and the basic reproduction number reads: :

$$R_0 = \frac{a\lambda\langle s(s-1)\rangle}{\mu}. \tag{4.3}$$

Determining the parameters in Eq. 4.3 can be challenging, but the expression for R_0 can be used to make a preliminary assessment of the impact of changes in network connectivity on the spread of the epidemic. The equation states that R_0 is proportional to the number of connections within the system. However, it should be noted that this approach does not take into account the correlations and memory effects that are present in real-world temporal networks, as it randomly reshuffles the individuals in a cluster at each time step, which is a simplification commonly used in activity-driven models. Eq. 4.3 can be used to compare the basic reproduction number among different phases, such as the closing, partial opening, and total opening phases (as shown in Fig. 5.4). By analyzing the effect of the different size distribution $P(s)$ in these phases, we can estimate the change in R_0 from the closing to the partial opening phase:

$$\frac{R_{0,PO}}{R_{0,C}} = \frac{\langle s(s-1)\rangle_{PO}}{\langle s(s-1)\rangle_{C}} \approx 2.63 \tag{4.4}$$

while going from the partial opening to the total opening phase implies:

$$\frac{R_{0,TO}}{R_{0,PO}} = \frac{\langle s(s-1)\rangle_{TO}}{\langle s(s-1)\rangle_{PO}} \approx 13.03 \tag{4.5}$$

It is clear from our analysis that the ratio of the basic reproduction number between the total opening phase and the partial opening phase is much larger than the ratio between the partial opening phase and the closing phase. This suggests that while tracing measures may be effective in controlling the spread during the transition from the closing phase to the partial opening phase [69], they may not be as effective in the transition to the total opening phase. However, it is important to note that during the total opening phase, a significant portion of the population had been vaccinated (about 80% in Italy) and access to University buildings was restricted to only those who had been vaccinated or tested. These additional measures, along with the use of masks and social distancing, were crucial in limiting the spread of the virus in the total opening phase, due to the significant increase in potentially dangerous contacts. We found that our estimates for the ratios in Eq.s (4.4,4.5) remain consistent when using different time intervals to define the simplices in our dataset. Specifically, when using a time interval of 5 minutes, we obtain ratios of 2.91 and 12.36, and when using a time interval of 30 minutes, we obtain ratios of 2.78 and 12.47 for the ratios in Eq. (4.4) and (4.5) respectively. This confirms the robustness of our results.

4.4 Access Points daily links

So far, we have analyzed the sizes of groups that form in the University and compared the distributions in the three phases with different restrictions. This data can also be used to identify the most critical areas in the University where large gatherings are more frequent. However, it is important to note that the formation of large groups is not the only relevant information to determine if an AP is critical. For example, if large gatherings are due to face-to-face lessons, we expect that the groups are stable for the entire duration of the lesson and that the contagion can be effectively traced (e.g. in the partially open period, an online seats reservation procedure was active). On the other hand, there could be places (such as an atrium) where small groups form but they are continuously reshuffled. These places are typically dangerous because they host a high number of different contacts which are very difficult to trace.

To address this, we introduce a different characterization of the APs to identify places where a large variety of contacts may occur. In particular, we define the "daily links" l_i of an AP as the number of contacts per day formed for more than 15 consecutive minutes between two different users in the same location. The index i labels the different APs. In this framework, if two users meet two or more times at the same point but at a different time, this is counted only once. Again, in counting users pairs, we have considered only the working time.

This new characterization of APs will help in identifying not only the places where large groups are formed, but also the places where a high number of different contacts may occur, which are typically difficult to trace. This information can be used to inform public health policy and decision-making during pandemics and to better understand how different populations are affected by restrictions.

To evaluate the effectiveness of the simplex size and daily link measures in identifying critical locations, the study focuses on two specific APs; one located in a classroom of the teaching building and the other in the atrium of the Physics Department. Panel **c** of Fig 4.8 shows that in the partial opening period, the size distribution of the two APs, $P_i(s)$, is similar. However, the average number of links, $\langle s(s-1)\rangle_i$, at the atrium is about 50

On the other hand, in the partial opening phase, the distribution of the number of different links per day, $\tilde{P}_i(l)$, is very different in the two APs. In particular, the average number of links per day is almost four times larger in the atrium than in the classroom. This highlights the much more variable behavior of contacts in the atrium. Panels **g** and **h** also show that in the atrium, the occupancy is almost constant during working hours, while in the classroom, people assemble mainly during lessons.

The data also indicate that people use the two areas differently in the periods examined. In the atrium, the average link number grows from $\langle l\rangle_i \approx 73$ in the closure period to $\langle l\rangle_i \approx 270$ in the partial opening phase, reaching $\langle l\rangle_i \approx 445$ in the total opening phase. In contrast, in the classroom, the average link number rapidly grows from $\langle l\rangle_i \approx 0.7$ in the closure period to $\langle l\rangle_i \approx 70$ in the partial opening phase and then up to $\langle l\rangle_i \approx 2030$ in the total opening phase.

The fact that different locations are used differently in the different restriction regimes is further confirmed in Figure 4.9, which plots the distribution of the duration of daily contacts between pairs of individuals in the different restriction regimes in the atrium and the classroom. All distributions display a typical exponential decay with a characteristic timescale and, as expected, the contacts in the atrium are typically shorter than in the classroom. Furthermore, as restrictions are removed, the duration of the connection in the classroom increases, as one would expect due to the larger number of lessons; however, in the atrium, the contact time becomes shorter, highlighting a potential problem in monitoring gatherings in common areas.

4.5 Space monitoring and resource optimisation

The analysis of group sizes and daily links provide different insights into identifying critical areas during the Covid-19 pandemic. To effectively monitor potential hotspots, it's important to consider both measures. In order to find the areas that need to be monitored first, we characterized each access point (AP) by the average number of different links per day, $\langle l\rangle_i$, and the average number of links present in groups, $\langle s(s-1)\rangle_i$. In figures 4.12 and 4.13 we show the APs with the highest $\langle l\rangle_i$ and $\langle s(s-1)\rangle_i$ in the partial and total opening period, respectively. This ranking highlights the importance of considering both measures, as there are cases where an AP's position changes significantly between the two rankings. For example, the atrium of the Physics Department ranks 16th according to $\langle s(s-1)\rangle_i$ and 9th according to $\langle l\rangle_i$, while a classroom moves from 24th to 60th position with the link ranking.

The classification by average daily links of the APs appears to be the most relevant for determining the critical areas on the university campus, as a high number of daily links implies a continuous reshuffling of users that could lead to superspreading events. It's also worth noting that the location of critical APs completely changes between the partial and total opening phases, due to the fact that some

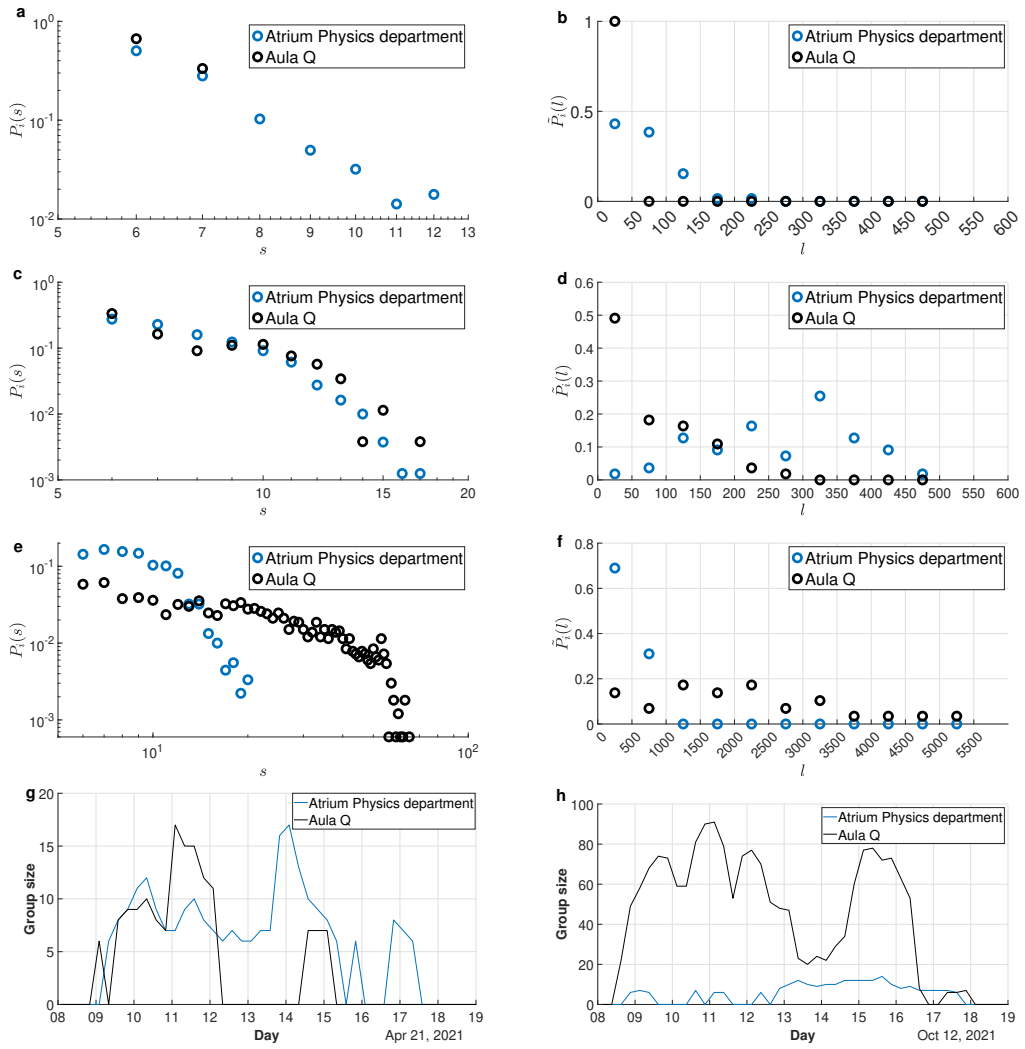


Figure 4.8: **Comparison between different APs of the simplex size and daily links measures.** We compare the group size and daily link distributions of two APs installed in different types of areas: one placed in a classroom, and the other in the atrium of the Physics Department. In panels **a** and **b**, the group size distribution and daily link during the closing phase are plotted for the AP placed in the classroom. It shows that small groups are formed due to restrictions. Panels **c** and **d** show the distributions obtained during the partial opening phase for the same APs. The group size distribution is similar for both APs, but the daily link distributions are different, indicating that the groups in the classroom are more stable than in the atrium, where groups are continuously reshuffled. Panels **e** and **f** show that large gatherings are formed in the classroom due to the return of face-to-face lessons. Panels **g** and **h** show the time evolution of the group size for the two APs during a typical working day in the partial opening and total opening phases respectively. It's worth noting that for small groups with size $s < 6$ the group size is set to 0.

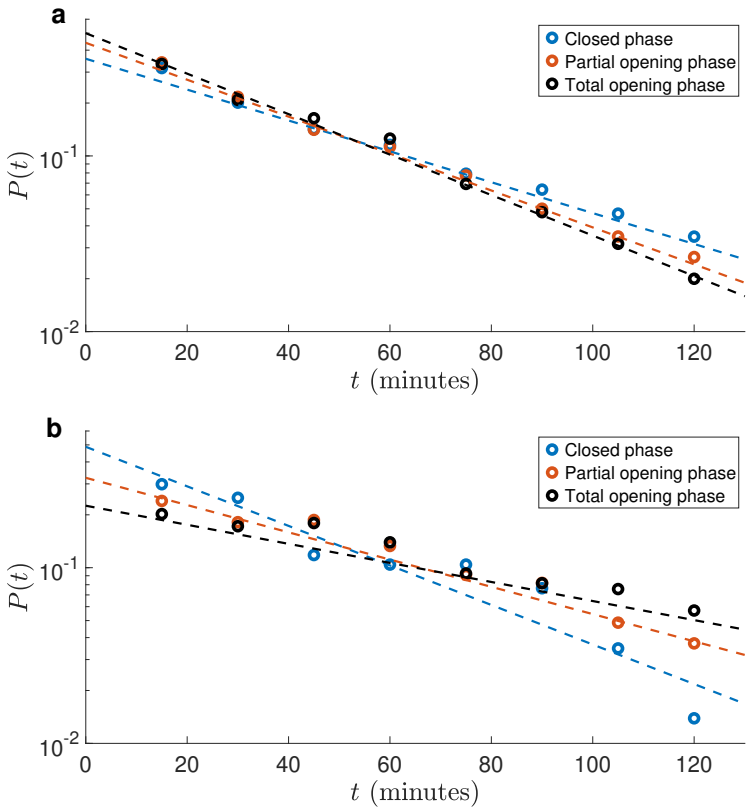


Figure 4.9: **Contact time distribution.** The probability, $P(t)$, that two users are in contact for a time t during working hours per day is plotted in Panel **a** and **b** for the AP placed in the atrium of the Physics Department and classroom respectively, for the three different restriction regimes. The distribution of contact time is found to have an exponential decay, $f(t) \propto e^{-t/\tau}$, but with different constants τ. For the contact time distributions related to the AP in the atrium of the Physics Department, the constants are $\tau_C = 49 \pm 2, m$, $\tau_{PO} = 41 \pm 1, m$ and $\tau_{TO} = 38 \pm 1, m$, while for the AP placed in the classroom, the constants are $\tau_C = 39 \pm 5, m$, $\tau_{PO} = 56 \pm 4, m$ and $\tau_{TO} = 80 \pm 7, m$. It is noteworthy that the contact time decreases as the restrictions are removed in the atrium of the Physics Department while in the classroom the contact duration becomes longer in the opening periods.

teaching activities were still held remotely during the partial opening period. This is highlighted in Figure 4.10, where university buildings are color-coded according to the number of critical APs present in figures 4.12 or 4.13 for the partial and total opening periods. This information can be used to effectively target resources and interventions to the areas that need it most.

Figure 4.11 illustrates the distribution of $\langle l \rangle_i$ and $\langle s(s-1) \rangle_i$ among the different APs in the university. Both quantities have been normalized by their average value across all APs, in order to allow for comparison on the same scale. The data shows that there is a slow decrease in the distribution at large values, which correspond to the critical APs identified in Figures 4.12 and 4.13, as indicated by the vertical dashed lines in the figure. These critical APs correspond to locations with high numbers of daily links and/or large groups, which are potential risk factors for the spread of infection. Overall, the distribution of data in Figure 4.11 highlights the importance of monitoring both the average number of different links per day and the average number of links present in relevant groups in order to effectively identify the critical areas on the university campus.

4.6 Beyond simplicial temporal networks

The basic reproduction number R_0 that we estimate in section 4.3 is based on a temporal network model that utilizes simplices and a mean field approach without memory [69, 87]. This means that the model assumes that all the $s(s-1)/2$ links formed in a group of size s occur between distinct individuals, with a complete reshuffling of connections at each time step. However, our data indicates that links between the same users tend to repeat at the same location (i.e. AP) and may also occur at different locations, leading to an overestimation of the number of contacts. Additionally, we have observed that the repetition of links exhibits different timescales in various restriction regimes and across different locations. In light of this, contact reshuffling may play a significant role in the spread of epidemics, even when comparing different phases of restrictions.

In the previous section, we analyzed the distribution of links formed between individuals at each access point (AP) on a university campus. This information gives us insight into the number of unique pairs that form connections at a specific location over a period of time. However, to fully understand the impact of contact reshuffling across different phases, it's important to also consider the total number of unique pairs formed across the entire university on a daily basis. This data, which has been calculated by the university's ICT services, is presented in Sec. 2.3.

When comparing the average number of unique pairs formed per day during the partial opening phase to the closing phase, we find:

$$\frac{\langle \mathcal{L}_d \rangle_{PO}}{\langle \mathcal{L}_d \rangle_C} \approx 3.70 \tag{4.6}$$

The ratio of these averages is larger than the ratio of total links formed in the same comparison. This suggests that not only are group sizes smaller during the closing phase, but there is also a more limited diversity of contacts. This could be due to the reduced presence of students and the resulting decrease in inter-departmental interactions and use of common areas. On the other hand, staff members tend to

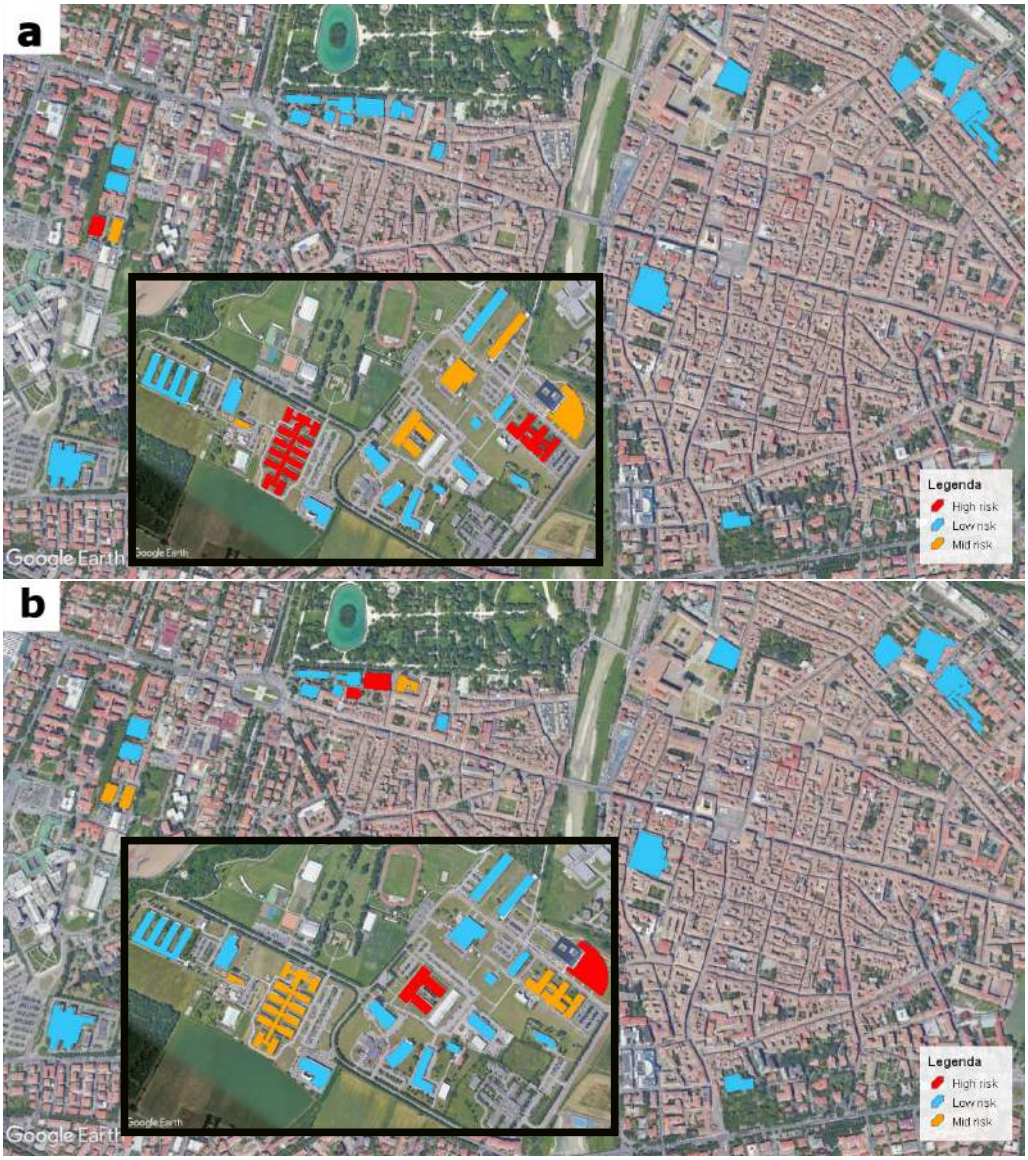


Figure 4.10: **Geographical map of buildings with critical APs.** Map of Parma with the University buildings color-coded based on the number of critical APs inside. In panel **a**, the buildings are ranked based on the APs in Figure 4.12 during the partial opening period. In panel **b**, the buildings are ranked based on the APs in Figure 4.13 during the total opening period. Buildings with no critical APs are represented in light blue, buildings with one or two critical APs are represented in yellow, and buildings with more than two critical APs are represented in red. The return of face-to-face lessons has a significant impact on the location of critical APs, with buildings such as the Economics department becoming a critical area.

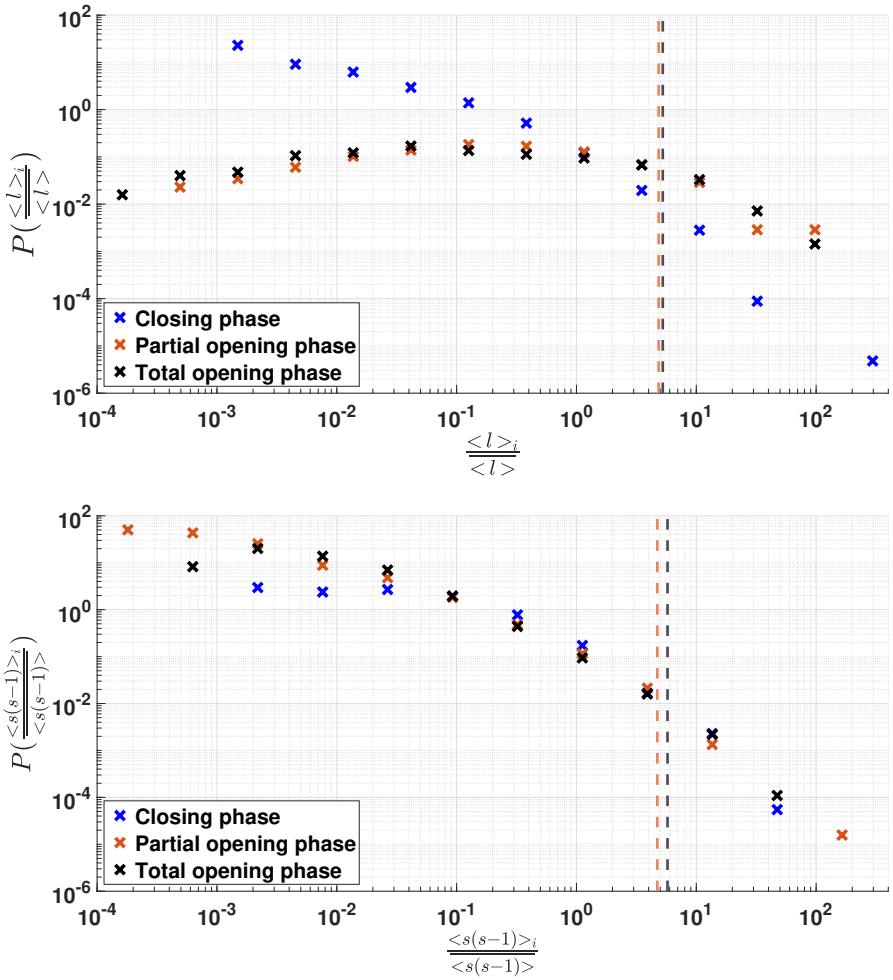


Figure 4.11: **Links distribution among the different AP.** The plots depict the distribution of the average number of different links per day, denoted as $< l >_i$, and the average number of links present in groups, denoted as $< s(s-1) >_i$, among all the available APs. To facilitate comparison, both measures are normalized by their overall average across all APs. The critical APs, as identified in Figures 4.12 and 4.13, can be identified as the APs with values above the threshold indicated by the dashed lines. Panel **a** illustrates that the distribution of the data is notably different during the closing phase compared to the opening phases, emphasizing that groups tend to be more stable during the closing phase.

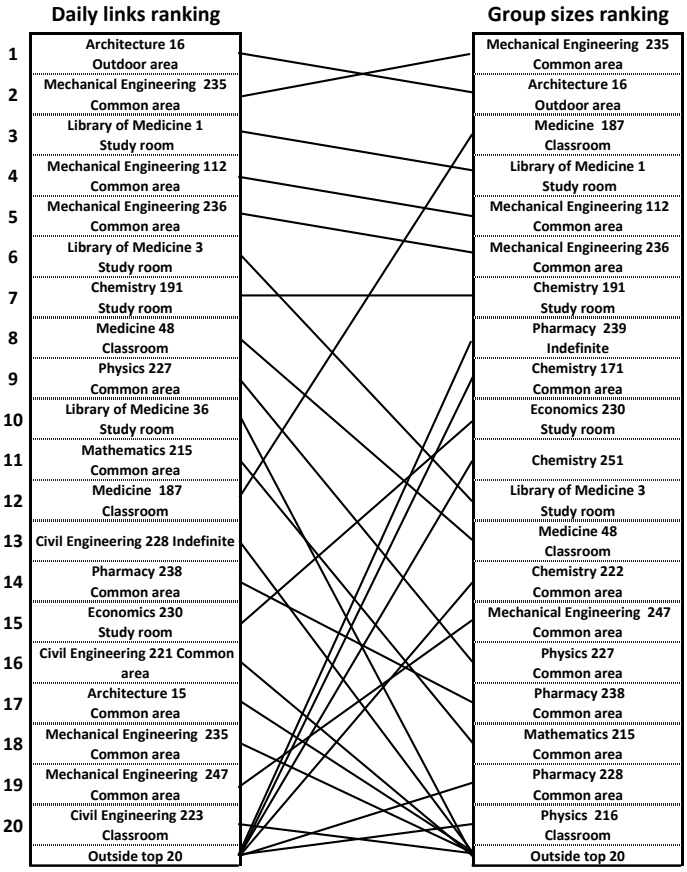


Figure 4.12: **Rankings during partial opening phase.** Comparison of the top twenty APs in the ranking by average daily links and the ranking by average group size during the partial opening phase.

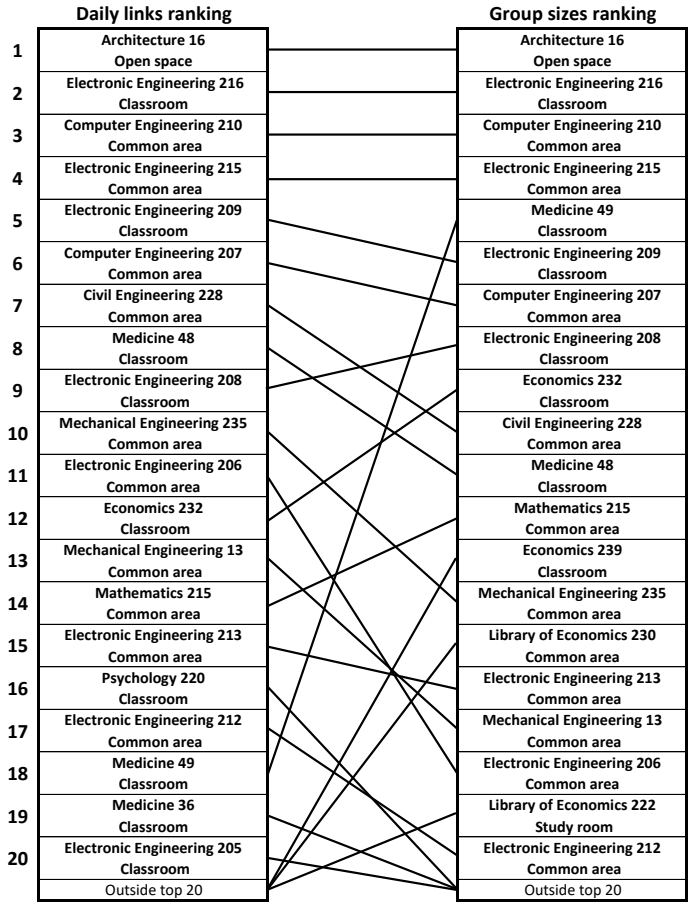


Figure 4.13: **Rankings during total opening phase.** Comparison of the top twenty APs in the ranking by average daily links and the ranking by average group size during the total opening phase.

primarily interact with those within their own labs, leading to a more stable pattern of contacts.

When comparing the total opening phase to the partial opening phase, the ratio of average number of unique links per day is:

$$\frac{\langle \mathcal{L}_d \rangle_{TO}}{\langle \mathcal{L}_d \rangle_{PO}} \approx 12.98. \tag{4.7}$$

This is similar to the ratio observed in the total number of links formed, indicating that the effect of link reshuffling versus link repetition is consistent across these phases and that there is only a general increase in the total number of contacts. We also observe that during these phases, the presence of students and face-to-face lessons dominates, leading to a similar usage of university spaces. However, as previously noted, during the closing phase, laboratory activity tends to result in a more stable pattern of link formation.

Our results indicate that the time interval used to define simplices does not significantly affect our conclusions. The estimates for the ratio of the average number of different pairs and average number of different links per day in the two periods remain robust even when the time interval is changed. For example, when a 5-minute interval is used, the ratios in Eq. (4.6) and (4.7) are still 3.56 and 12.50 respectively, and when a 30-minute interval is used, the results are still 3.64 and 13.33. This further supports the validity of our findings.

Furthermore, our analysis of the distribution of links among the different APs further supports the conclusion that the closing phase is characterized by a distinct behavior compared to the opening phases. Figure 4.11 illustrates that the distribution among the different APs are very similar in the two opening phases, indicating that different areas are occupied in a similar way and that the reshuffling mechanism is analogous. However, in the closing period, the distributions in Figure 4.11 are significantly different even after the global rescaling. This confirms that the closing phase is characterized by a much more stable pattern of contacts as the reshuffling mechanism is significantly suppressed. Additionally, the fact that the location of critical APs changes drastically in the two periods, as shown in Figure 4.10, further highlights the distinct behavior of the closing phase.

Chapter 5

The control of epidemics in large gathering

In this chapter, we propose a model that seeks to understand the spread of a contagious disease, such as SARS-CoV-2, on a network of individuals. The model is based on a compartmental model, which accounts for the main stages of a contagious disease, including asymptomatic and presymptomatic transmission. The model also takes into account the temporal dynamics of social contacts and the higher-order nature of interactions by modeling the gatherings as "simplices" . Additionally, the model considers an adaptive behavior in which symptomatic individuals are immediately isolated and unable to participate in gatherings, thus preventing them from propagating the infection. The text that follows describes the details and assumptions of the model and the results obtained through its application. In the previous chapter, we introduced a new formalism for monitoring the use and formation of groups within public environments. Using the results obtained in the previous chapter, we test our results on synthetic simplex size distributions and on an empirical dataset for gatherings in a University Campus.

5.1 Epidemic model

In this model, individuals are represented as nodes in a temporal network. They interact with one another by participating in gatherings, which are modeled as "simplices." Each node in a gathering is in contact with all the other nodes, creating a fully connected cluster. The network evolves in discrete time steps, with simplices activating at a rate of a, known as the "simplex activity." When a $(s-1)$-simplex (i.e. a group of size s) is active, s nodes are randomly chosen to participate, resulting in $s(s-1)/2$ interactions. Then, the links between nodes are broken and the process is repeated. The size of the simplices, s, is determined by a distribution $\Psi(s)$, which models the diversity in the size of gatherings. Interactions between pairs of individuals correspond to simplices with $s = 2$. This activity-driven simplicial temporal network model [70, 86, 105] takes into account the temporal dynamics of social contacts and the higher-order nature of interactions [11, 18, 87]. The proposed model for understanding the spread of disease is an enhanced version of the Susceptible-Infected-Recovered (SIR) model. In addition to the standard SIR classifications, this model includes further distinctions based on the stage of infection. These distinctions are made according to whether an individual is presymptomatic

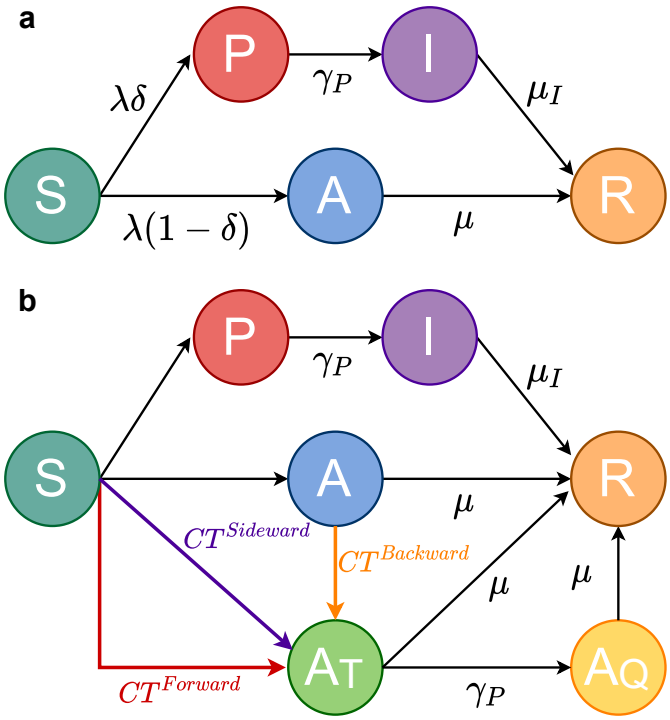


Figure 5.1: **The epidemic compartmental model with and without tracing.** The first model **a**, is a representation of the spread of disease without considering contact tracing. The second model **b**, builds upon the first model by including the effects of forward, backward and sideward contact tracing. The rates of infection and contact tracing are discussed in further detail in the accompanying text and supplementary information.

(P), symptomatic (I), or asymptomatic (A). The pathogen spreading dynamics without tracing is shown in Fig. 5.1**b**. If we consider the cantact tracing mechanism, we have two new stages of infection A_T and A_Q, which are traced asymptomatic and quarantine asymptomatic respectively. The allowed transitions among these states are shown in Fig. 5.1**b**. The model also takes into account changes in social behavior depending on the health status of individuals. We do not consider here memory [102] nor burstiness effects in the dynamics [70].

5.2 Contact tracing mechanism

In this model, we consider a traditional, non-app-based contact tracing (CT) process [68] on gatherings (or simplices) with the goal of identifying infected asymptomatic individuals. Asymptomatic individuals A are the most difficult to track since, as they have no symptoms, they may continue to have contact with other nodes (people) and to detect them it is necessary for a susceptible contact S to go into a presymptomatic P or symptomatic I state. It's important to note that since individuals in state I are instantaneously isolated, they do not participate in gatherings. Only individuals in state P or A may spread the infection. When a presymptomatic individuals develops symptom $P \to I$, the contact tracing mechanism is activated.In this case, each gathering he/she has participated in during the previous T_{CT} days is traced as a whole, with a probability $\epsilon(s)$. Each node belonging

to a traced gathering is tested, and if found in the asymptomatic infected (A) state, is isolated.

The probability of successfully tracing a gathering (or simplex) can depend on its size s, as people typically only remember some of the gatherings they have joined. Some gatherings are easily fully traced (e.g. school classes, workplace meetings) while others are not easily reconstructed (e.g. interactions on public transportation, shops, restaurants). The dependence of $\epsilon(s)$ on the gathering size s allows for modeling tracing strategies targeted at groups of a given size.

To account for CT in the model, two additional compartments are introduced for asymptomatic individuals (Fig. 5.1**b**): traced asymptomatic (A_T) and quarantined asymptomatic (A_Q). An A_T node is asymptomatic and infective, just like an A individual, but has been identified as having been in contact with a presymptomatic individual who remembers the gathering where the contact took place. When the presymptomatic develops symptoms, the A_T node enters quarantine and becomes an A_Q node, which is isolated and hence does not participate in gatherings.

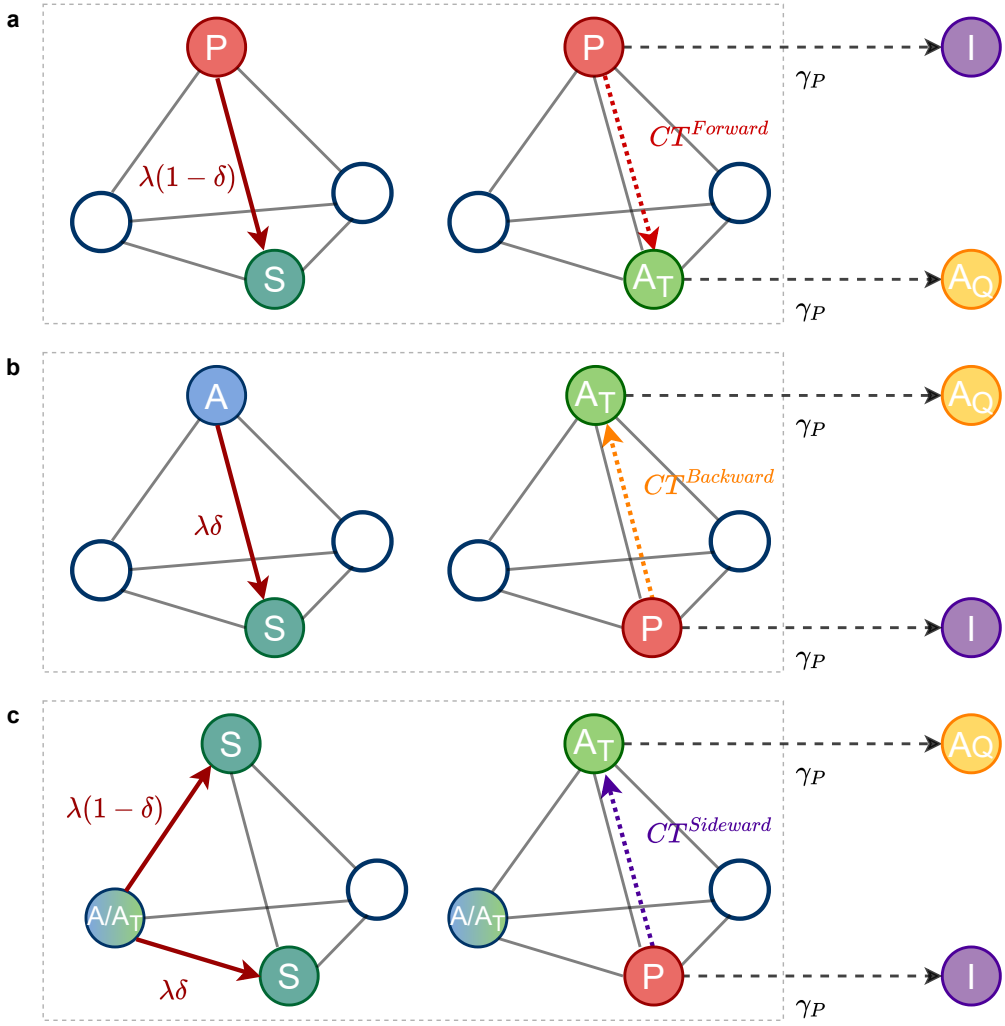
For simplicity, it is assumed that tracing and isolation occur instantaneously after the presymptomatic node becomes symptomatic. Hence the transition $A_T \to A_Q$ occurs with the same rate γ_P as the appearance of symptoms.

In this work, in addition to the forward and backward CT mechanisms, already in place for pairwise interactions [16, 32, 61, 68], contact tracing is reinforced by a third tracing mechanism called sideward CT, which is specific of simplices traced as a whole.

- **Forward CT.** This CT mechanism is activated when a presymptomatic individual develops symptoms and goes on to detect all asymptomatic individuals infected when the activating node of the tracing was in the presymptomatic state (see Fig. 5.2**a**). If infection events occur, the node switches from the susceptible state S to the asymptomatic state A_T and is traced.

- **Backward CT.** The process of backward contact tracing (CT) is activated when an asymptomatic individual infects a susceptible individual, making them presymptomatic (see Fig. 5.2**b**). In this scenario, the asymptomatic individual is traced, becoming a traced asymptomatic individual (A_T), along the contact that produced the infection. However, the tracing goes in the opposite direction of the infection event, as the asymptomatic individual is already infected when they enter the simplex. This means that the tracing occurs after the infection event and after the individual has potentially infected other nodes.

- **Sideward CT.** Sideward contact tracing (CT) is a mechanism that identifies asymptomatic individuals who have been infected by *other* asymptomatic individuals, by utilizing the presence of a third node that develops symptoms (as seen in Fig. 5.2**c**). This CT mechanism requires the presence of at least three nodes in the simplex: an asymptomatic A (or traced asymptomatic A_T) node that infects a susceptible node that will become asymptomatic, and at least one other susceptible node that will become presymptomatic due to contact in the simplex. When the presymptomatic individual develops symptoms and moves into the symptomatic state, the CT mechanism is activated and the asymptomatic individual is tracked. The process of sideward CT requires



d

	Forward CT	Backward CT	Sideward CT
Involved Nodes	P S → A_T	A S → P	A/A_T S → A_T S → P
Traced Node	**S → A_T**	**A → A_T**	**S → A_T**
Conditions	Tracing following infection link. Existing ∀ s	Tracing following backwards infection link. Existing ∀ s	Tracing following a link different from the infection link. Existing for s ≥ 3

Figure 5.2: **Contact tracing mechanisms in simplices: forward, backward and sideward CT. a** Scheme of the forward CT. **b** Scheme of the backward CT. **c** Scheme of the sideward CT:The CT process can be activated in cases where the infected node is either A or A_T. In addition, in this scenario, backward CT may also occur on node A when activated by $S \to P$, but this process is not shown in the scheme for simplicity. Additionally, sideward CT can be activated if the asymptomatic and symptomatic infections are caused by different asymptomatic individuals (or traced asymptomatics) who participate in the same simplex. While this process is considered in the model and produces a term that is quadratic in the density of asymptomatic nodes, for simplicity, only the first-order term is considered in the scheme, which is the only one that survives the linearization for the calculation of the epidemic threshold. **d** Summary of the main features of the three CT mechanisms.

a minimum of three individuals in a simplex, and is only activated when the simplex size is equal to or greater than 3. This type of contact tracing does not involve tracing the direct transmission of an infection, but rather identifying asymptomatic individuals who may have been infected by other asymptomatic individuals through the presence of a third node that develops symptoms. It is important to note that in the same simplex, backward CT can also be activated on the infector if they are an infected asymptomatic A.

In Appendix A, we have derived mean-field equations for the evolution of an epidemic with pre-symptomatic and asymptomatic transmission on simplicial activity-driven networks. In the linearized mean-field equations, the forward, backward and sideward CT are described by the following equations:

$$C^{Forward} = \lambda(1-\delta)a\langle\epsilon(s)s(s-1)\rangle P(t) \tag{5.1}$$

$$C^{Backward} = a\left\langle\epsilon(s)s\left[1-(1-\lambda\delta)^{s-1}\right]\right\rangle A(t) \tag{5.2}$$

$$C^{Sideward} = \lambda(1-\delta)a\left\langle\epsilon(s)s(s-1)\left[1-(1-\lambda\delta)^{s-2}\right]\right\rangle [A(t)+A_T(t)] \tag{5.3}$$

These are the three terms given by the three contact tracing mechanisms, in particular sideward CT is described by a term composed of multiple factors. In eq. 5.3 in the linearized regime, a is the simplex activation rate, the probability to take part to the simplex for the susceptible node is proportional to s and, $(s-1)[A(t)+A_T(t)]$ is the probability that among the remaining $(s-1)$ nodes there is an asymptomatic (or traced symptomatic) node. The term $\lambda(1-\delta)$ represents the probability that the infection occurs and is asymptomatic, while the term $\left[1-(1-\lambda\delta)^{(s-2)}\right]$ is the probability that at least one of remaining $(s-2)$ nodes is infected with manifestation of symptoms.

We obtain a set set of 5 linearized equation and it can be written as:

$$\begin{bmatrix}\partial_t A_Q(t)\\ \partial_t I(t)\\ \partial_t P(t)\\ \partial_t A(t)\\ \partial_t A_T(t)\end{bmatrix} = J \begin{bmatrix}A_Q(t)\\ I(t)\\ P(t)\\ A(t)\\ A_T(t)\end{bmatrix} \tag{5.4}$$

J is the Jacobian matrix of this set of 5 linearized equation:

$$J = \begin{bmatrix} -\mu & 0 & 0 & 0 & \gamma_P \\ 0 & -\mu_I & \gamma_P & 0 & 0 \\ 0 & 0 & -\gamma_P+\beta & \beta & \beta \\ 0 & 0 & \Delta\left(1-\frac{\langle\epsilon(s)s(s-1)\rangle}{\langle s(s-1)\rangle}\right) & -\mu+\Delta-\Gamma-\Phi & \Delta-\Phi \\ 0 & 0 & \Delta\frac{\langle\epsilon(s)s(s-1)\rangle}{\langle s(s-1)\rangle} & +\Gamma+\Phi & -\mu-\gamma_P+\Phi \end{bmatrix} \tag{5.5}$$

$$J = \begin{bmatrix}\mathbb{A}(2\text{x}2) & \mathbb{C}(2\text{x}3)\\ \mathbb{O}(3\text{x}2) & \mathbb{B}(3\text{x}3)\end{bmatrix} \tag{5.6}$$

where:

$$\Phi = \lambda(1-\delta)\frac{\overline{n}}{\langle s(s-1)\rangle}\left\langle \epsilon(s)s(s-1)\left[1-(1-\lambda\delta)^{s-2}\right]\right\rangle$$
$$\Gamma = \frac{\overline{n}}{\langle s(s-1)\rangle}\left\langle \epsilon(s)s\left[1-(1-\lambda\delta)^{s-1}\right]\right\rangle$$
$$\beta = \lambda\delta\overline{n}$$
$$\Delta = \lambda(1-\delta)\overline{n}.$$

5.3 The effects of contact tracing strategies in the epidemic threshold

In the previous sections, we introduced a new epidemic model based on SIR model on activity-driven network. Three different CT methods were added to this model: forward, backward and a new method called sideward CT. The purpose of contact tracing is to mitigate the spread of the epidemic by tracking and isolating infected individuals. The three different mechanisms contribute differently to limiting the epidemic and their effects depend on the structure of the interaction, i.e. the distribution of simplices. In this section, we will compare the results and performance of these three methods. For this comparison, we assume that the probability to be traced is equal for simplices of any size, i.e. $\epsilon(s) = \epsilon$, $\forall s$. Additionally, we assume that the number of average contacts per individual and per time unit $\overline{n} = a\langle s(s-1)\rangle$ is constant, so that different distributions $\Psi(s)$ correspond to the same number of interactions, arranged in simplices of different size. We evaluate the impact of each contact tracing CT strategy by calculating the epidemic threshold when only symptomatic individuals are isolated, then when each CT mechanism is active separately, and finally when all CTs are active. In Figures 3a-3b, we consider all simplices to have the same size $\overline{s}$, with $\Psi(s) = \delta(s-\overline{s})$. In Figures 3c-3d, we use an exponential distribution $\Psi(s) \sim e^{-\beta s}$, for $s \in [2, \infty)$. Lastly, in Figures 3e-3f, we use a power-law distribution $\Psi(s) \sim s^{-(\nu+1)}$, as observed in real systems [97], with $s \in [2, s_M]$.

The effectiveness of isolation of symptomatics and forward CT does not depend on the distribution of simplex sizes $\Psi(s)$, while the effectiveness of backward and sideward CT is highly dependent on it. Sideward CT is most effective in large simplices, where there is a higher chance of lateral infection, while it is not effective for pairwise interactions ($s = 2$). In large simplices, sideward CT can trace and isolate all new asymptomatic individuals at the time of infection, preventing the spread of the epidemic and explosive outbreaks of SSEs. Conversely, backward CT is more effective in small simplices, as in large clusters with many contacts, it only traces the source of infection and other simultaneous contagions may go undetected.

When all CT mechanisms are active, the combination of backward and sideward tracing results in a non-monotonic behavior as a function of simplex size. The epidemic threshold exhibits a maximum, where CT is most effective. In networks with a single simplex size or with a sharp exponential distribution, the size corresponding to this maximum is around 100 nodes. However, for broader distributions with large clusters dominating transmission even at small average simplex sizes, tracing is most effective when the average simplex size is around 10. The position of this

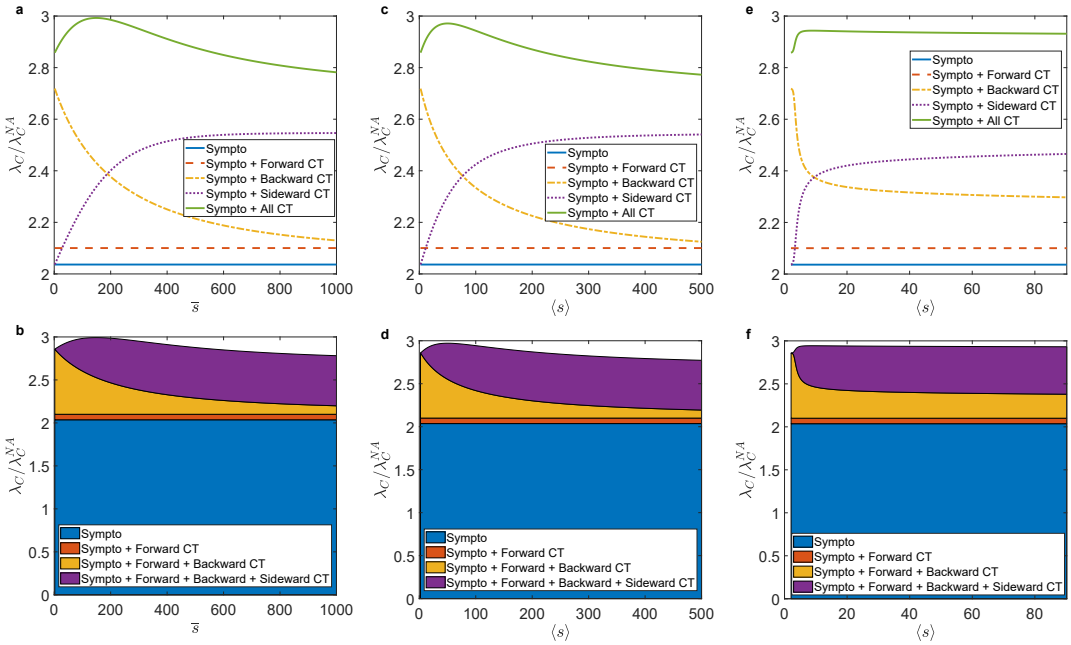


Figure 5.3: **The effects of forward, backward and sideward contact tracing.** Analysis of the impact of contact tracing on the epidemic threshold. The ratio of the epidemic threshold λ_C with symptomatic isolation and CT activated to the non-adaptive case threshold λ_C^{NA} is plotted as a function of $\overline{s}$ for a constant simplex size distribution (Panel **a** and **b**), an exponential distribution (Panel **c** and **d**) and a power-law distribution (Panel **e** and **f**). The activation of each CT mechanism alone and all mechanisms combined is also shown. The average simplex size is varied by changing β and ν in the exponential and power-law distributions, respectively. The probability of tracing a simplex is held constant at $\epsilon(s) = 0.3$ for all simplex sizes.

maximum also depends on the fraction of asymptomatic nodes, with a larger fraction being more effectively traced through sideward tracing and a smaller fraction through backward tracing.

5.4 Contact tracing in a real setting: Parma University Campus as case study

The spread of an epidemic, in the presence of large simplices, is mainly driven by SSEs events with many infections in a few but large groups. This suggests that we should focus the tracking mechanism in places where large groups are formed, such as concerts, cinemas or public spaces. In particular, the university represents a public space where every day, thousands of students and staff gather to work or teach. In this section we will quantitatively test these hypotheses by comparing the impact of the three CT mechanisms in the real case of the University of Parma (Italy). Our findings suggest that a focus on large simplices may be an effective strategy for reducing the spread of the epidemic.

We use the probability distribution $\Psi(s)$ measured empirically, using data obtained from WiFi network, described in detail in the previous chapter in Sec. 4.2. The WiFi network, by measuring the number of simultaneous connections to Access

Points (APs), represents a good proxies for the gatherings. We focus on gatherings that have been deemed epidemiologically significant, lasting for more than 15 minutes as per guidelines by the ECDC [33]. It is known that indoor transmission of SARS-CoV-2 can occur within the spatial range covered by WIFI access points [41], making WIFI data a useful tool in supporting traditional contact tracing methods [21]. To this end, we analyzed WIFI data from the University of Parma and calculated two separate distributions of $\Psi(s)$. The distributions of $\Psi(s)$ for the two time periods, obtained from WIFI data at the University of Parma, are heterogeneous as seen in other datasets [97]. The distribution during the closure period has a smaller upper limit due to restrictions on activities with many people, such as in-person classes. This is reflected in an increase in the probability of simplices with $s = 0$ or $s = 1$ (i.e. 0 or 1 individual connected to the AP). The changes in the distribution of simplex sizes, as represented by $\Psi(s)$, have a significant effect on the epidemic threshold. When only isolating symptomatic individuals is implemented, the epidemic threshold increases during a period of restricted activities (e.g. closure) by a factor

$$\frac{\lambda_C^{closure}}{\lambda_C^{opening}} = \frac{\langle s(s-1)\rangle_{closure}}{\langle s(s-1)\rangle_{opening}} \simeq 2.63, \tag{5.7}$$

so that $\lambda_C^{closure}/\lambda_C^{NA} \simeq 5.35$. This reduction in activities comes at a cost of stopping most teaching and working activities. Contact tracing aims to maintain activities while still controlling the epidemic. We evaluate the effect of the contact tracing strategies during the partial opening period and compare the mitigation achieved through tracing with that achieved through closure.

We investigate the effectiveness of different contact tracing (CT) strategies by modeling them through different values of $\epsilon(s)$. One strategy focuses on large simplices, with size $s \geq s^*$, and is represented by $\epsilon(s) = \theta(s - s^*)$, where $\theta(x)$ is the Heaviside step function. This approach is compared with a uniform tracing strategy, where all simplices are traced, and a strategy that targets only small simplices, represented by $\epsilon(s) = \theta(s^* - s)$. We keep the resources allocated for tracing constant by setting the average fraction of traced nodes to $\epsilon^* = \frac{\langle \epsilon(s)(s-1)\rangle}{\langle s-1\rangle}$. Results in Fig. 5.4**b** show that targeting large simplices is the most effective strategy, as it requires tracing only a small fraction of nodes to achieve a significant increase in the epidemic threshold. In contrast, tracing only small simplices requires tracing almost all gatherings to achieve a comparable result. Tracing efforts targeted at large gatherings can produce similar results to full closures if $\epsilon^* \gtrsim 0.47$, as shown in Fig. 5.4**b**. This requires tracing at least all simplices with $s \geq 6$, representing 16.1% of simplices with $s \geq 2$. Full closures may be a more drastic measure than necessary, particularly in the case of a highly transmissible variant, with a basic reproduction number of $R_0 \approx 4.5$, even in the presence of additional mitigation measures such as the use of face masks [24, 26, 99, 109, 111, 112]. A partial reopening of activities while implementing targeted tracing efforts on large gatherings can still effectively control the spread of the virus.

As illustrated in Fig. 5.4**b**, to keep the epidemic below the threshold, it is sufficient to trace on average a manageable fraction of nodes per simplex, specifically $\epsilon^* \gtrsim 0.27$, which corresponds to tracing all simplices with $s \geq 9$, representing only 6.2% of simplices with $s \geq 2$. The lower panels of Fig. 5.4 demonstrate the various contributions of forward, backward and sideward CT for the different strategies. It is evident that backward tracing provides the most significant increase in the thresh-

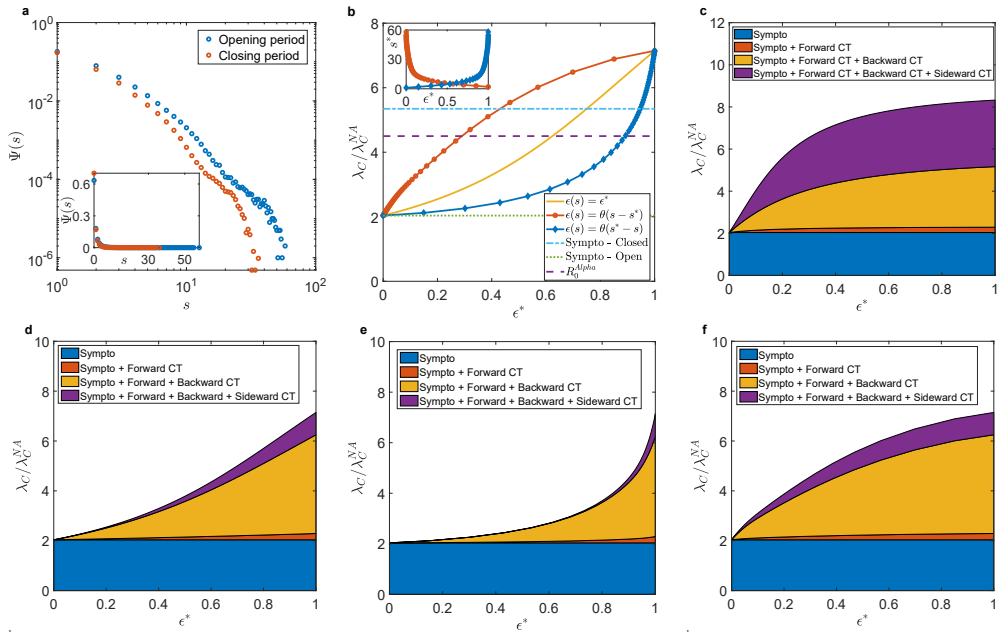


Figure 5.4: **Tracing strategies on empirical University gatherings data. a** We plot the distributions $\Psi(s)$ obtained via WIFI data for the University of Parma during the partial opening and closure periods. The main plot is in log-log scale, while in the inset the same distributions are plotted in linear scale. **b** We plot the ratio λ_C/λ_C^{NA} as a function of $\epsilon^* = \frac{\langle \epsilon(s)(s-1)\rangle}{\langle s-1\rangle}$, where λ_C^{NA} is the epidemic threshold in the non-adaptive case and λ_C is the threshold when isolation of symptomatic individuals and CT are activated. We consider the empirical $\Psi(s)$ during the partial opening period and the three CT strategies. The three horizontal lines are: the ratio $\lambda_C^{opening}/\lambda_C^{NA} \approx 2.04$ when only symptomatic individuals are isolated during the partial opening period (dotted); the value of $R_0 = \lambda/\lambda_C^{NA} = 4.5$ for a variant of SARS-CoV-2 (dashed); the ratio $\lambda_C^{closure}/\lambda_C^{NA} \approx 2.63\lambda_C^{opening}/\lambda_C^{NA} \approx 5.35$ when only symptomatic individuals are isolated during the closure period (dot-dashed). In the inset it is plotted s^* as a function of ϵ^* for the two targeted tracing strategies. **c** We plot the ratio λ_C/λ_C^{NA} as a function of ϵ^* for symptomatic isolation and activating progressively all CT mechanisms, considering $\Psi(s) \sim s^{-(\nu+1)}$, with $s \in [2, 200]$ and $\nu = 1.5$ and implementing a tracing strategy targeted at large simplices. **d** Same of panel **c** with the empirical distribution $\Psi(s)$ during the partial opening and implementing a uniform tracing strategy. **e** Same of panel **d** implementing a tracing strategy targeted at small simplices. **f** Same of panel **d** implementing a tracing strategy targeted at large simplices. In all panels the parameters are fixed as discussed in Methods.

old. However, in the strategy targeted at large simplices, sideward CT also plays a significant role in controlling the spread of the epidemic, particularly when large simplices are driving transmission. It's worth noting that during the partial opening period, several restrictions were still in place, such as reducing the maximum capacity of a classroom by a factor of 4 [106], which explains the relatively small cut-off $s_M \approx 60$ in the $\Psi(s)$ distribution. For broader distributions with larger cut-off (corresponding to a full opening of the University), sideward tracing is expected to provide the primary contribution, as shown in Fig. 5.4**c**, where we plot the different contributions to targeted tracing for a power-law distribution $\Psi(s) \sim s^{-(\nu+1)}$, with $\nu = 1.5$ and a cut-off $s_M = 200$.

Chapter 6

Data-driven prediction of public transport use

So far, we have presented the results for the mathematical modelling of social dynamics in public settings and for the development of a new activity-driven epidemic model with contact tracing mechanism. These studies occurred due to the onset of the Covid-19 pandemic, which coincided with the start of this PhD cycle. Now that the situation is returning to normal, we have returned to the mathematical modelling of urban mobility. In this chapter, we present the work of my last year of my PhD, which took place in collaboration with the Sustainable Cities group of Sony CSL Rome led by Vittorio Loreto.

In this chapter, we propose a data-driven framework to estimate the use of public transport in the metropolitan area of Parma by citizens, estimate the number of users for each single bus ride and find situations where the public transport service offered is not very efficient, such as cases of bus overcrowding. This model, has the possibility of creating scenarios to evaluate the impact of certain interventions, such as new bus routes or new bus schedules, to improve the public transport service, limit service costs and, above all, limit overcrowding during peak hours.

6.1 Data description

The model exploits several open data and private data sources to build a baseline scenario representing a typical day of commuting in the Parma Metropolitan Area. In this paragraph, we present all types of data needed for the public transport occupancy simulation model.

6.1.1 Parma metropolitan area

First of all, we need to define the area of interest for the study of public transport in Parma. TEP, the company that provides the public transport service, operates throughout the province of Parma with two different types of lines:

- urban buses, which operate only within the municipality of Parma:
- interurban buses, which connect the city to the entire provincial territory.

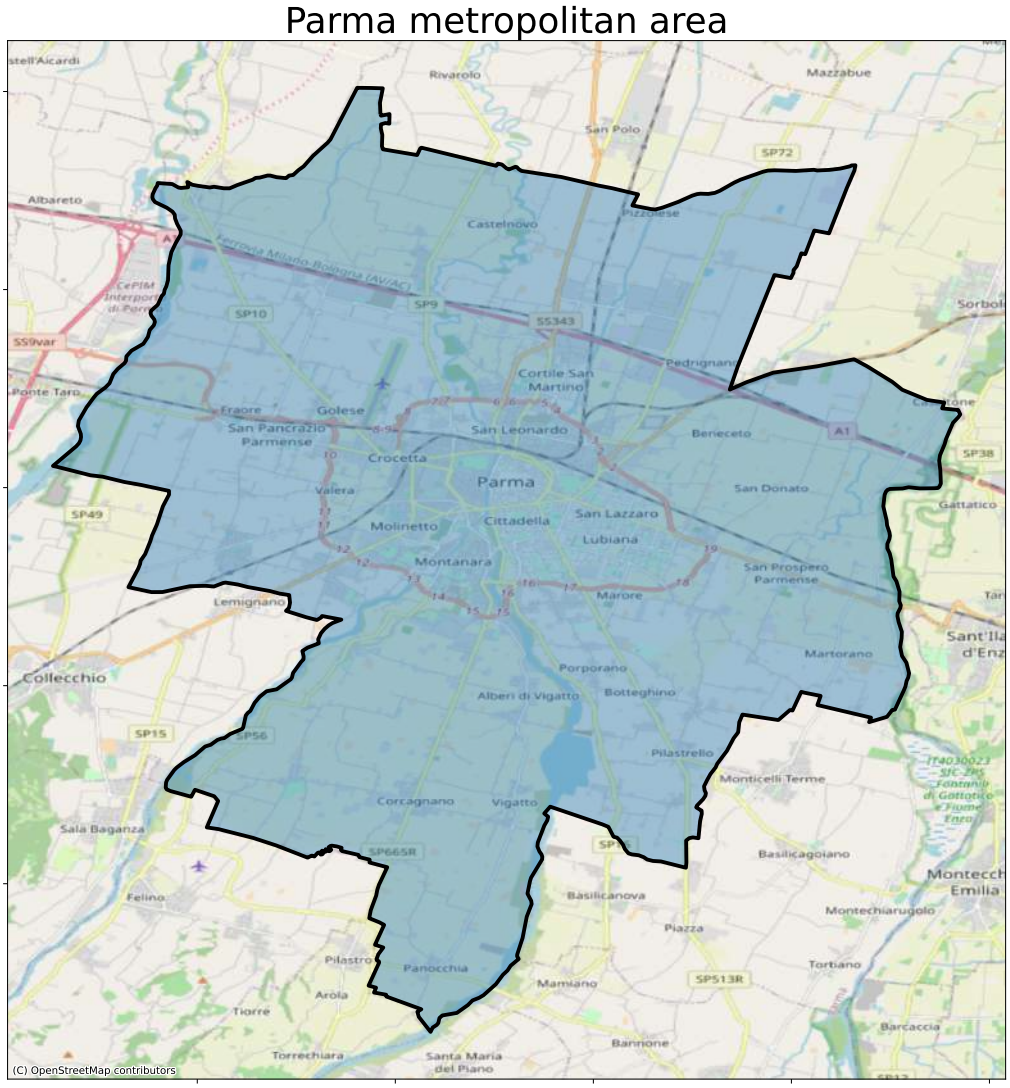


Figure 6.1: **Parma metropolitan area.** In the figure we can see the metropolitan area of Parma highlighted in blue and delimited by the black line. As can easily be seen, the zone covers the municipality of Parma some neighbouring municipalities.

The aim in this chapter is to evaluate the impact of commuting users on public transport in the city of Parma. Many people live in the municipalities bordering the city but have their workplaces in the city, leading to the creation of commuting flows into the city in the morning rush hour and outflows from the city in the evening. For this reason, limiting ourselves to assessing the model only in the municipal area is very limiting and the entire commuting territory of the city must be analysed. We decided to use the metropolitan area of Parma defined by OECD[1] as the area of interest for the model Fig. 6.1. OECD defines the metropolitan area of a city as the union of two zones: the first is the city (or city centre) and the second is the commuting zone, i.e. the zone integrated socioeconomically with the city. The metropolitan area was subdivided into zones in order to construct an Origin-Destination matrix M of trips that will be determined by a careful statistical analysis of the HFLB data. We therefore divided the territory into 21 zones (see in

[1] `https://doi.org/10.1787/9789264174108-en`

Fig. 6.2); 20 zones were defined by the province of Parma by subdividing the city into macro-areas each of them self-sufficient, while one zone was inserted specifically for the location area of the University Scientific Campus. This zone was inserted specifically since the objective of this work is to predict the use of public transport to the campus.

6.1.2 Census data and economic activities

To reconstruct the Parma Metropolitan Area's socio-economic structure, we used data from the ISTAT, the Italian National Statistical Institute. The data we collected comes from the 2011 national survey that mapped the population and the economic activities. The mapping is at the level of census areas, i.e. small territorial units varying in dimension to have an almost uniform population with respect to one another. The number of employees of each economic activities is recorded within each census areas. Moreover, economic activities are divided according to the ATECO (ATtività ECOnomiche) code that can be used to identify the kind of activity (e.g., manufacturing, agriculture, finance). We aggregated the 1321 census areas contained in the metropolitan area into 21 defined in the previous paragraph. This aggregation is necessary in order to obtain datasets from different types of data that can be combined with each other.

6.1.3 Open Street Map Data

Open Street Map (OSM) data gathers much geographical information that can be used for multiple purposes [46]. In this work, we exploited OSM in a twofold way. Firstly we used OSM collected OSM Points-of-Interest in the Parma Metropolitan Area. We assigned to each POI a label p based on the category it belongs to between *restaurant/leisure/services/health*, discarding the others. In the following, we will refer to the category of a POI with $kpoi \in \{restaurant, leisure, services, health\}$. Secondly, we used the OSM street network and Open Source Routing Machine (OSRM) [54] to compute travel times on foot between public transport stops.

6.1.4 GTFS data

TEP has provided the metropolitan area's public transportation data in General Transit Feed Static (GTFS) format. GTFS is a standard for public transport schedules developed by Google[2] to encourage administrations and public transport companies to release data in a uniform format. GTFS data contains all the trips (T) performed by public transport services within an area, a trip representing a bus, tramway or metro train performing its services between a starting stop s and ending stop s^*. Trips are divided in connections (C), i.e. a public transport vehicle moving from one stop to another. TEP's data covered all the public transport services within the Metropolitan Area, including extra-urban buses.

[2] `https://developers.google.com/transit/gtfs/`

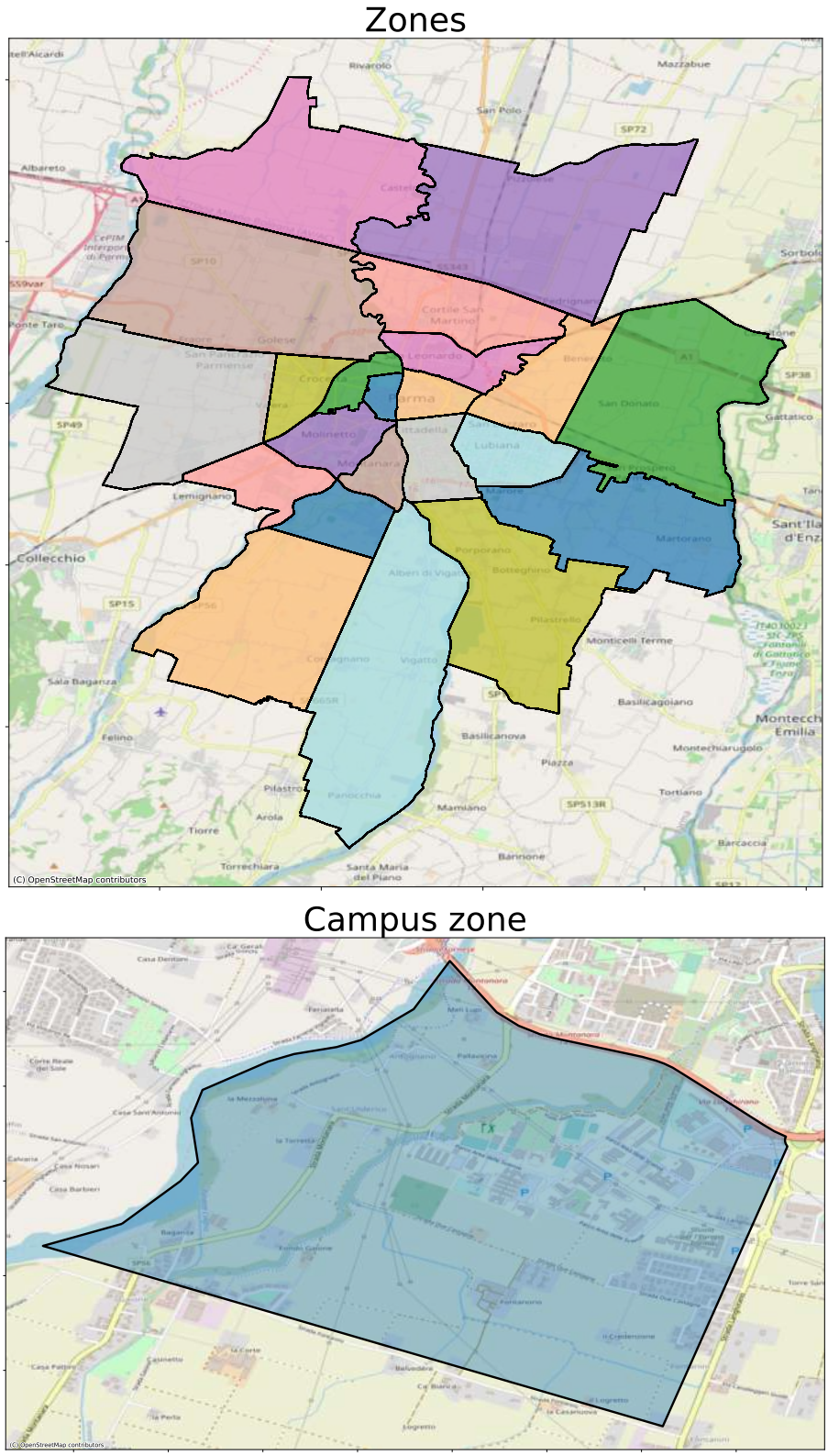


Figure 6.2: **Parma commuting zones.** The first figure shows the subdivision of the metropolitan area, while in the second figure we can see a detail of the area containing the science campus.

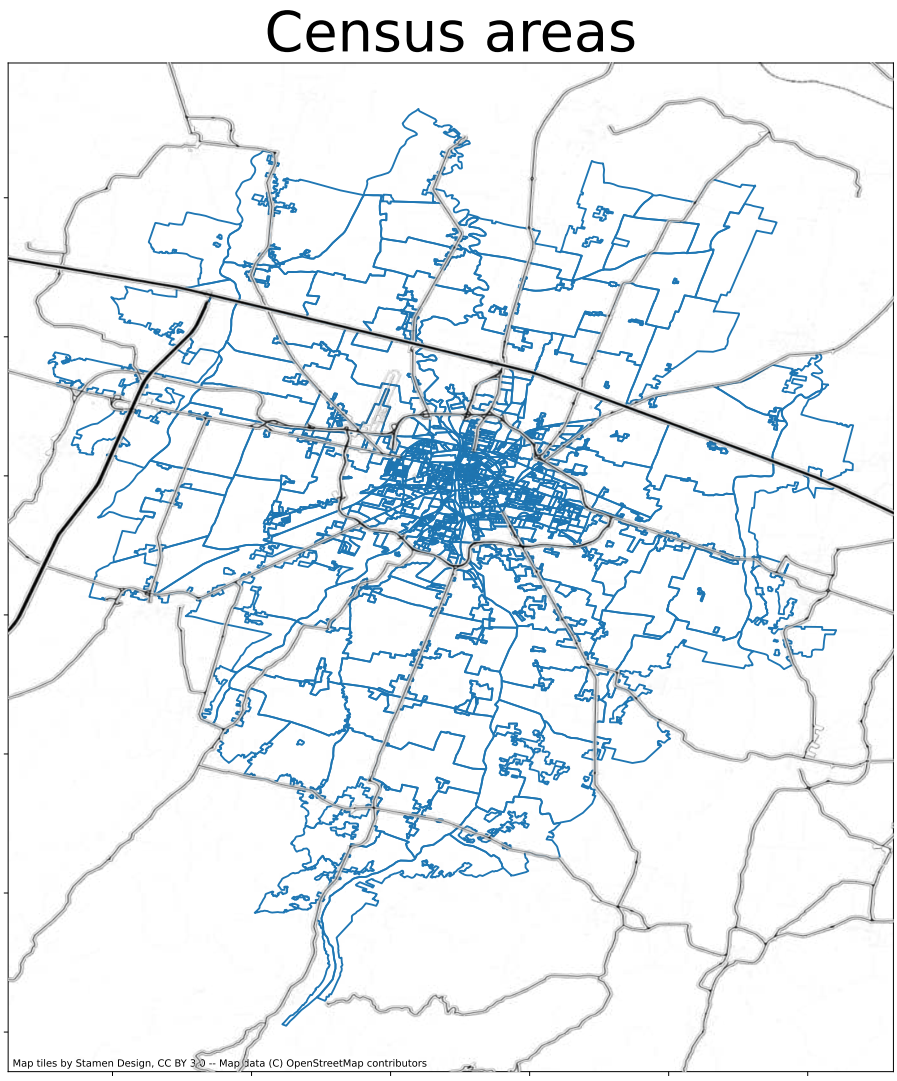


Figure 6.3: **Census area.** The census areas contained in the metropolitan area are shown in the figure. These areas have been aggregated between loso in order to obtain the same subdivision used in Fig. 6.2.

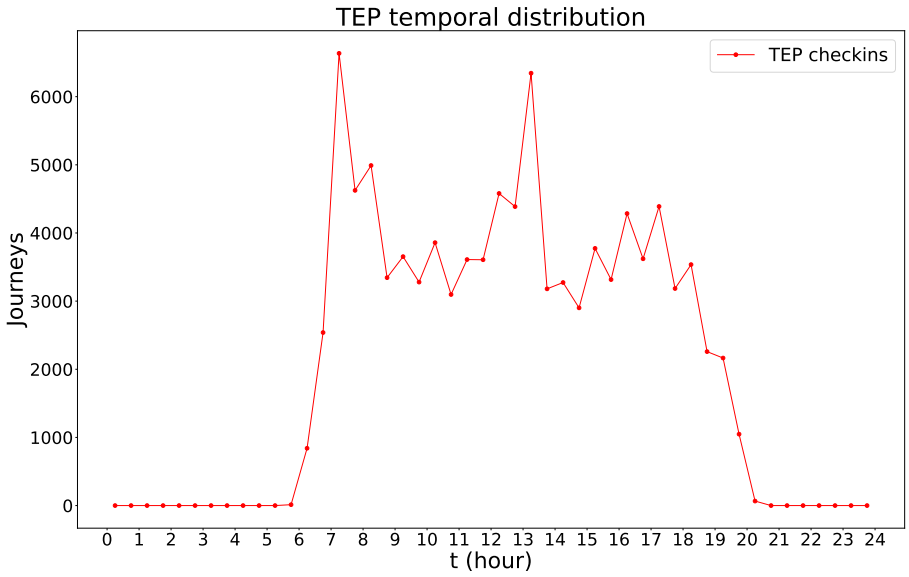


Figure 6.4: **Bus users temporal distribution.** Time trends of users of the urban public transport service in Parma during a working day. We can see that there is a high volume of users in the morning and at lunchtime caused by students going to school or university.

6.1.5 Bus access data

TEP's administration has collected information on the number of users who use public transport as a means of transport. In 2015, it started collecting data on the number of people getting on and off at each stop and on each urban bus route on working days from 6 a.m. to 8 p.m. TEP provided us with two different datasets:

- total number of travellers on each run of each city line. For each urban bus line l, this dataset provide $N(l,t)$, where N is the number of users who took the route l that started its journey at time t;
- daily number of check-ins and check-outs on each stop of each urban bus route. For each urban bus line l, this dataset provide $r(l,s)$ and $d(l,s)$ are respectively the users boarding and alighting on line l at stop s.

From these two datasets, we can obtain the time distribution of public transport users on a typical working day. In Fig. 6.4, we can see that there is high bus utilisation during the morning rush hour (7-9 am) and during the lunch break. In the case of the morning, the high volume of users of the public transport service is due to the combination of users going to work, combined with the high number of students going to school or university. On the other hand, the peak of users at around 1.30 p.m. is given solely by students returning home once their teaching activity has ended. It is easy to see that in the peak hours of the evening, the time when people usually return home, the volume of users is much lower than in the morning. This leads us to think that the largest users of the service are students, while workers tend to prefer the primary means of transport for commuting trips. These two datasets were provided for two uses: the first to go through the process of calibrating the model, the second to go through the validation of the model's result. This will be discussed in the following paragraphs.

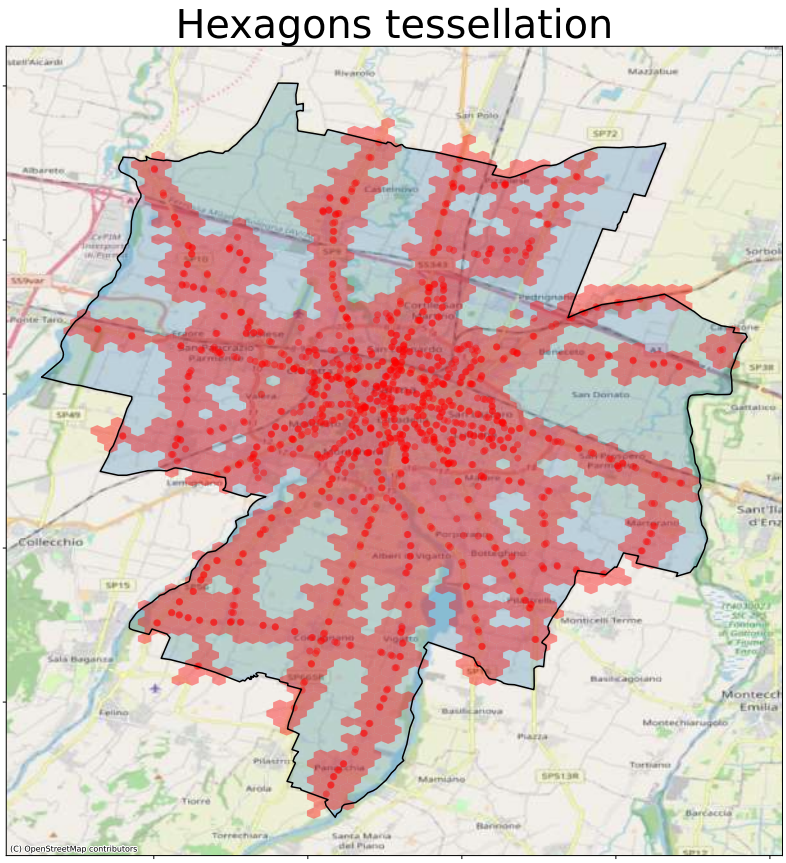


Figure 6.5: **Parma hexagons tessellation.** We can see that the public transport service with its stops (red dots) does not cover the entire territory. However, after checking with ISTAT population data, only 8% of the population is not reached by any bus line provided by TEP.

6.1.6 Citychrone++

The model is based on the subdivision of the territory into hexagons to generate scenarios. This subdivision is carried out by the Citychrone++ platform and divides the metropolitan area into hexagons of equal size, this size being defined by the user as an input parameter for the code. The algorithm, creates a tessellation of hexagons of the territory and only keeps the hexagons that are reachable by public transport or on foot from the nearest stop with a maximum walking distance defined by the user. This distance is defined by the user by setting the maximum walking time and the average walking speed. For the Parma area, we defined a maximum walking time of $t_{walking} = 10min$ and an average walking speed of $\overline{v}_{walking} = 5km/h$. In addition to these parameters, the Citychrone++ algorithm requires as input the GTFS data and the size of the hexagons set as $hex_{side} = 200m$. In output we have an ordered dataset of hexagons: for each hexagon we know the bus stops contained within it and all other neighbouring prime hexagons. Fig. 6.5 shows the tessellation of the metropolitan area with hexagons obtained as output from Citychrone++ and it can be seen that not in all the area the transport service is guaranteed. In fact, there are many suburbs in agricultural areas where buses do not run with the risk of not providing the service to many users who would then be forced to use private transport. For this reason, we verified that the population outside the hexagons (not

served by buses) is a negligible percentage compared to the population served. By cross-referencing the hexagon dataset with the ISTAT population dataset, we found that only 8% of the population living in the metropolitan area are not provided with public transport services to their homes, so we can say that the coverage of bus stops is very good throughout the area.

6.2 Mathematical model

The aim of this model is to predict the occupancy level of public transportation in a city in the real case and in new scenarios where some of the input parameters of the model were changed. For example, one could assess the impact of smartworking, the impact of changing bus routes or schedules. This model predict the occupancy at the connection level c, where we use the term "connection" to indicate a movement of a public transport vehicle between two scheduled stops, without intermediate stops where passengers can get it or out. Connections can, in a sense, be therefore seen as the "atoms" of which public transportation schedules are made of.

The first step for constructing scenarios is to reproduce a baseline model representing the system's functioning in normal conditions. The final result of the model is a set of multi-modal trajectories on public transport, whose purpose of the trip is know. Conceptually, our approach can be articulated in two steps:

1. First, we generate a synthetic/simulated population of public transportation users, each with its own origin and departure zones, departure time, and purpose of the journey.

2. We calculate for each of them the shortest path from origin to destination, using a slightly modified version of Connection Scan Algorithm (CSA) [13, 27] as implemented in Citychrone++ framework, which has been developed between 2018 and 2021 by Sony CSL Paris and Sapienza university of Rome.

Once we have the paths, expressed as lists of connections, it becomes trivial to calculate how many passengers there are on every connection, how many get in each stop, and so forth. In this way, we can obtain a double information:

- **Individual behaviour**: from each trip, we can identify the best-performing route once origin, destination and departure time have been defined, identify points where a change of line is needed and the theoretical arrival time.

- **Collective behaviour**: for each connection c between two consecutive stops we have the number of users, the time course of the number of users on a typical working day.

In Fig. 6.6 we can observe a schematic representation of mathematical model. In blue square we represent all empirical data from HFLB and TEP dataset. A calibration step, necessary to eliminate outliers, will be introduced to the model.

6.2.1 Input data

The characteristics of the simulated users' population are to be derived from empirical data. To this end, the following four input dataset are required for the generation of synthetic journeys with public transportation:

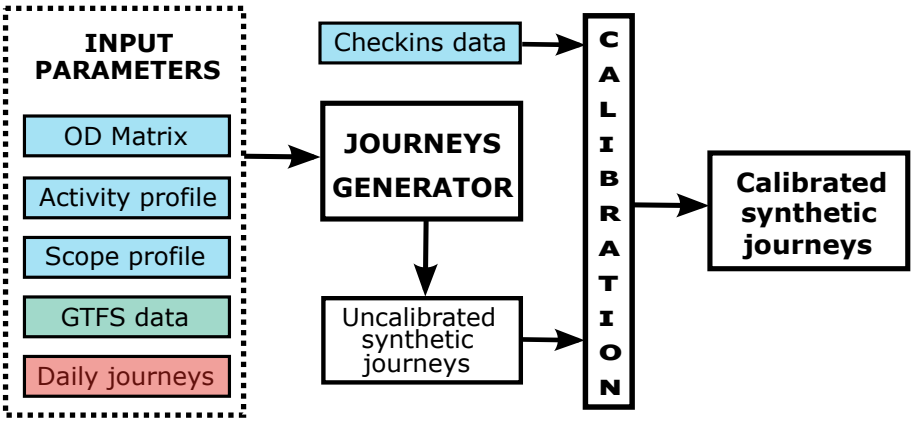


Figure 6.6: **Mathematical model scheme.** Diagram of the mathematical model for generating synthetic trips. The parameters in blue are empirically derived data from the various datasets available, the parameters in red is a user-chosen parameter, and in green is bus data. Included in the diagram is the calibration step that will be introduced later.

1. The **OD matrix** M of the interest area, i.e. a matrix where, assuming a division of the city into a number of zones, element i, j denotes a number of journeys between zone i and zone j. Only a relative weights matter, because the model allows the user to fix the total volume to a desired value and rescale the matrix accordingly.

2. The **activity profiles** $\tau(i, j, t)$ of the city, i.e, assuming a division of the city into zones (which may be different from the ones where the OD matrix is defined), the fraction of users traveling between i and j for every pair of zones i, j as a function of time t. In Fig. 6.10 we can see and example of activity profile between the Campus zone and other zones.

3. The **scope distributions** of the city, where by "scope" we denote the purpose of the journey, described by attaching a Point-of-Interest category label to the starting and ending points. In other words, given a pair of zones, the probability distribution that a journey between the two will have given starting and ending POI labels. Again, the zone division does not need to be the same as the other datasets.

4. The schedule of the city's transportation services.

We calculate the first three element using High Frequency Location Based Data (HFLBD) obtained from Cuebiq. To do this, we consider only relevant users (users with almost 10 acitivty days) and only the journeys during the working day, all trips data from weekend journeys were discarded. Although the goal of the model is to describe trips by public transportation, the trajectory dataset holds contains all trips that occur by all types of transportation because from the HFLB dataset available to us, it is not possible to determine the mode of transportation by which the trip was made. Note that, in principle, the input 1 and input 2 could be simply combined in a time-dependant OD matrix. The OD matrix that we derived from the HFLB data, was obtained from all trips that are made by the most active users that occur on weekdays. Thus, the OD matrix also takes into account the trips of

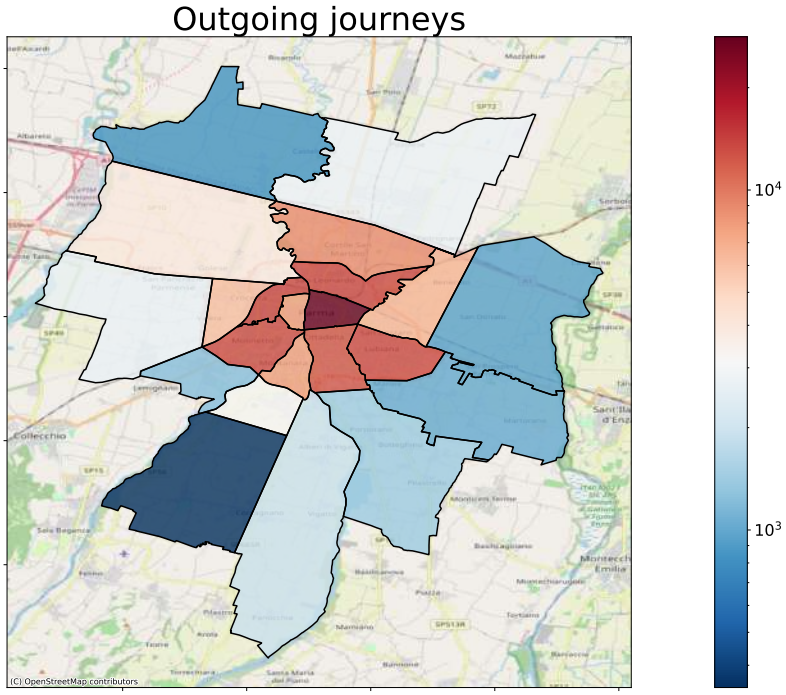


Figure 6.7: **Outgoing journey obtained from GPS data** In this figure we have shown the number of trips starting from zone i to one of the remaining zones obtained from the analysis of GPS trajectories. It can be seen that the highest number of trips start from the city centre, while in the suburbs the trip statistics drop dramatically.

users living outside the Parma metropolitan area who come to Parma for study or work. In Sec. 3.3, we verified that the algorithm for determining the stop location defined as "home" succeeds in identifying the population in the different zones of metropolitan area and these data agree with the census population data with a Pearson correlation coefficient $r^2 = 0.78$. This correlation allowed us to verify the good representation of the population data but this does not allow us to say that the data obtained for the OD matrix of trajectories has the same proportionality of representation. For this reason, combined with the problem of commuters outside the metropolitan area, the OD matrix data were not rescaled with the population bias but we considered. In Fig. 6.7 we have plotted the 21 zones with a colour according to the number of trips to destinations in other zones. As can be seen, the largest number of trips from the HFLB data analysis starts in the city centre, while there is less traffic in the neighbouring areas. However, the areas we are considering have very different population densities: in the city centre, where there is a high population density, it is easier to have a high number of trips, whereas in the suburbs with fewer users, the number of trips recorded by the data drops significantly. To test how population density affects the statistics of the GPS data, we normalised the number of outgoing trips obtained from the HFLB data with the population of each zone obtained from the census data (see Fig. 6.8). The data, normalised with population, show that population density does not affect the travel statistics too much. The only exception concerns the Campus area: here the number of trips is much higher than the number of people living in that area. This effect is certainly due to the activity that takes place within the university, which attracts a large number of students every day.

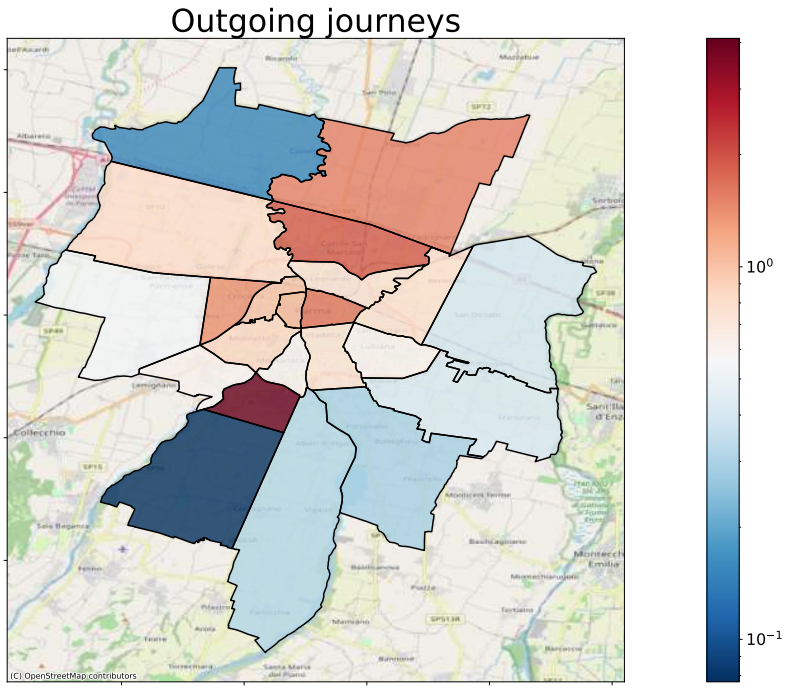


Figure 6.8: **Normalized outgoing data.** In this figure we have shown the number of trips starting from zone i to one of the remaining zones obtained from the analysis of GPS trajectories. It can be seen that the highest number of trips start from the city centre, while in the suburbs the trip statistics drop dramatically.

In fig. 6.9 the OD matrix elements of column j = Campus zone were reported, i.e. all trips from all zones to the University Campus zone. In Fig. 6.10 we can see and example of activity profile between the Campus zone and other zones, in particular the city center zone and the zone where the train station is located. We can immediately see how the campus is a centre of attraction in the morning during peak hours. In contrast, in the evenings, there is a high flow of trips leaving the campus and heading to the centre or the station. Fig. 6.9 shows the elements of the OD matrix with destination j = Campus zone. The main flows to the campus come from the city neighbourhoods, with the largest flow coming from the centre. On the other hand, flows from the peripheral areas of the metropolitan area are close to zero. In Fig. 6.10 we can see and example of activity profile between the Campus zone and three other zones: city centre, the area containing the train station (San Leonardo) and a residential district (Lubiana). In all cases, we can see that the temporal distribution of incoming trips to the campus, has a peak during the morning rush hour 7-9 a.m., while at other times of the day there is an even distribution. In contrast, outbound trips from the campus to their respective areas are more likely to occur in the evening, from 5-7 p.m. From these distributions, we can see that the campus area is an area of attraction for work or school activities and, moreover, with a volume of residents within it compared to the volume of student workers.

Once the input parameters derived from the empirical data have been defined, the user must define the number of daily users U_d, i.e. the number of trips by public transport. This number should be obtained from empirical data and/or chosen to improve the agreement between model and realty. In our case, we obtain the daily numbers of journeys from the checkins dataset provided by TEP described in section

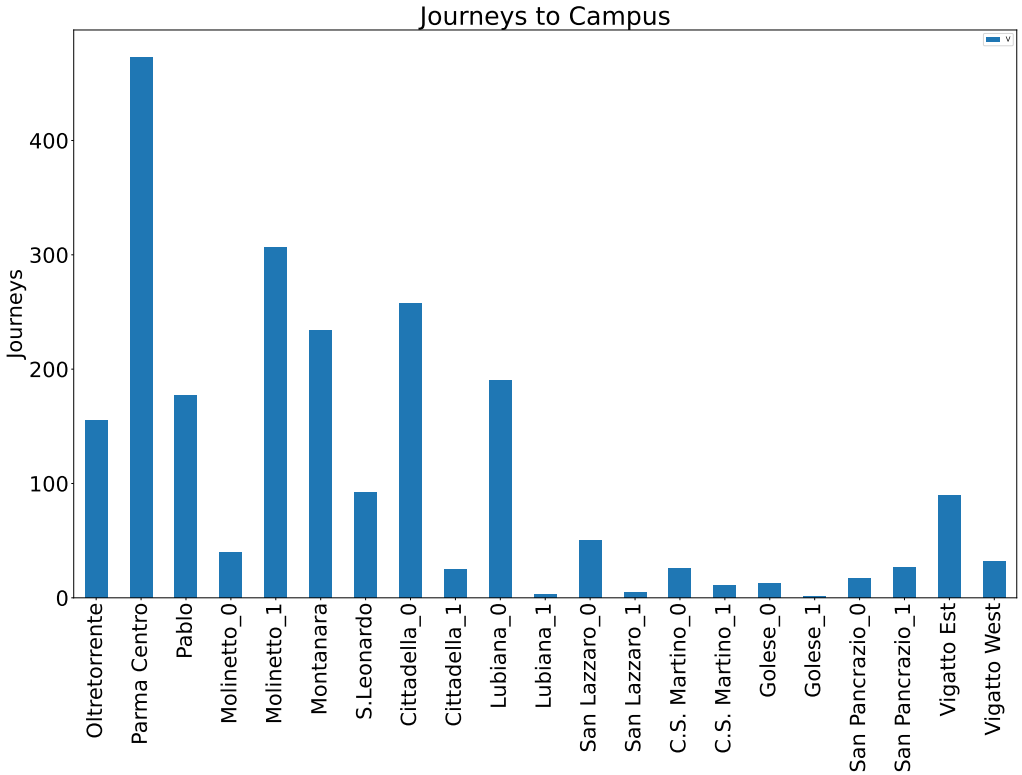


Figure 6.9: **OD matrix element with destination the Campus area.** In this figure we plot the activity profiles between Campus zone and other three zones: the city center, the zone with train station (San Leonardo) and another neighborhood. The green line represent the temporal distribution of journeys to Campus, on the other hand the orange line is the outflow from Campus zone. For each zone, we can observe an high volume of journeys to Campus during the rush hours in the morning.

6.1.5 and corresponds to the integral of the curve in Fig. 6.4.

6.2.2 Journeys generation

The OD matrix is then rescaled accordingly, and the number of journeys between each pair of zones is fixed. These are generated by calling the Citychrone++ library, using a slightly modified CSA algorithm [27]. While the basic version merely requires that a user can arrive at the departure stop before departure time in order to be able to get a connection c, our modification introduces a vehicle-dependent *buffer time* ω_t, i.e. the user must arrive a few minutes earlier to be able to catch a bus ω_t^B, metro ω_t^M or tram ω_t^M. In metropolitan area of Parma, public transport is only by bus, so we have defined a buffer time only for bus $\omega_t^B = 10$minutes. This choice was made after experimenting with the algorithm as a penalty for bus line change. In fact, changing the line is a loss of time caused by bus delays, traffic or other external factors that are not considered in the theoretical GTFS schedules.

The generation algorithm performs the following steps in sequence:

1. the number of trips for each zone pair is rescaled with the total number of trips U_d;

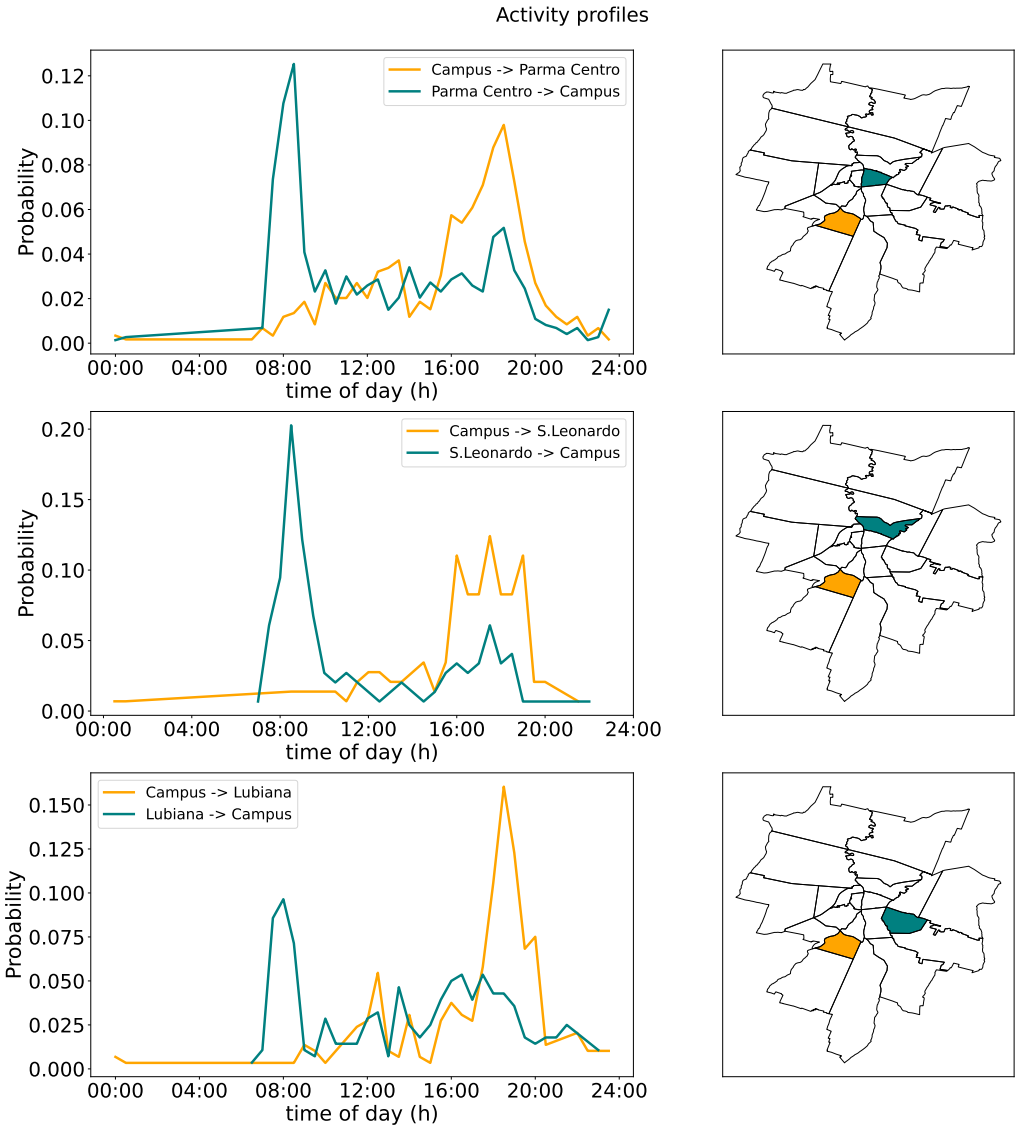


Figure 6.10: **Activity profile between campus and other zones.** In this figure we plot the activity profiles between Campus zone and other three zones: the city center, the zone with train station (San Leonardo) and another neighborhood. The green line represent the temporal distribution of journeys to Campus, on the other hand the orange line is the outflow from Campus zone. For each zone, we can observe an high volume of journeys to Campus during the rugh hours in the morning.

2. for each O-D zone pair, the following steps are iterated for each trip::
   (a) chosen with uniform probability distribution, a starting stop among all stops within the O-zone with uniform probability distribution;
   (b) chosen with uniform probability distribution,an arrival stop among all stops within the O-zone with uniform probability distribution;
   (c) chosen the starting time t_i from the activity profile $\tau(O, D)$ between O-D zones extracted from HFLB data;
   (d) chosen the scope of trips from the scope distribution $\sigma(O, D, t_i)$ between O-D zones extracted from HFLB data;
   (e) calculates the fastest route between the two stops with the Connection Scan Algorithm;
3. return a list of journeys.

The simplified scheme in Python of the synthetic trip generation algorithm is defined as follows:

```
def jouneys_generator(time_distr, scope_distr,\
                      total_trips):
  synthetic_journeys = []
  for O,d in zones(Parma):
    journeys = total_trips * normalized OD matrix
      for i in range(journeys):
        origin_stop = np.rand(stops(O))
        destination_stop = np.rand(stops(D))
        start_time = np.cumsum(time_distr(O, D))
        scope = np.cumsum(scope_distr(O, D, start_time))
        jounery = CSA(origin_stop, destination_stop,\
                      start_time, scope)
        synthetic_journeys.append(journey)

  return synthetic_journeys
```

Note that, in case of journeys that begin and end in the same zone, or happen between neighboring zones, the shortest path may simply be walking from origin to destination. Since the total number of journeys and the input distributions are meant to represent the users of public transport, in this case the journeys is rejected and another one is generated.

The final output of the procedure is a table with the following information for each journey: origin and destination zones (in terms of all the zone divisions involved: OD matrix, activity profiles and scopes), origin and destination stop, time of departure, scope, connections and check-ins. The connections are a list of ids, which can be used to retrieve information about each connection (departure and arrival times, departure and arrival stops, route, operating agency and so forth) in the Citychrone++ output files. The check-ins are a list of every time the user checks into a new line. This is not useful for computing metrics, but it is useful for validating against empirical data, as it can be compared with ticket validation data.

6.2.3 Model calibration

Dataset obtained from TEP discused in Sec. 6.1.5, contains all the check-ins and check-outs of urban lines in Parma obtained from measurements carried out in 2015. These allow us to perform a very precise comparison between the model's results and real data. While there seems to be a clear linear relationship between predicted and real chek-ins (see next section), there are many outliers, even in important/crowed station like bus stops in train station. Under assumption that HLFB data are representative of the real travel volumes, we must conclude that the distributions obtained from them do not fully capture the behavior of public transportation users. In order to improve the agreement, we therefore define a calibration procedure as follows.

To start, we define a set of *contraints*, derived from real data. A constraint consist of a boolean function of journey (for example: whether the journey passes through a certain stop or certain bus route), and a target T_i of journeys for which the function must evaluate to true. The constraint i is *satisfied* if:

$$r_i = \frac{T_i - X_i}{X_i} < 5\% = \text{tol} \tag{6.1}$$

where X_i is the observed number in the model output. The user can choose a different tolerance value if desired. The model framework supports user-defined constraints by inheriting from the Constraint class defined in the code. When $r_i < -\text{tol}$, we say that a constraint is over-represented. Likewise, a constraint is under-represented when $r_i > \text{tol}$.

Calibrating the system is an iterative process, where at each iteration, journeys are replaced with new ones, generated as to not alter the existing distributions: if a certain amount of journeys between zones O and D is removed, the replacements will have the same origin and destination zones, new departure/arrival stops in those same zones, and new departure time and scope chosen from the same distributions. The user can stop the process when a certain number of iterations has been performed, or when a certain number of constraints are satisfied.

The general idea is that the journeys we want to replace are the ones that contribute to underrepresented ones. During each iteration, we start by calculating the following delete score for each journey J:

$$D_J = \left\langle \gamma_i(J) \left[1 - \min\left(1, \frac{T_i}{X_i}\right) \right] \right\rangle_{i|\gamma_i(J)=1} \tag{6.2}$$

where $\gamma_i(J)$ is 1 if J participates to i, 0 otherwise, and D_J defaults to 0 if J participates to no constraint. We then mark each journey for deletion (but do not delete it yet) with a probability given by D_J. This means that contributing to an over-represented constraints increases the probability of deletion, by an amount proportional to how much the constraint is over-represented.

We then turn our attention to the underrepresented journeys. For each of them, we look at the journeys contributing to it, and note the set S_i of OD zone pairs they are defined on. Then, we consider the set S_i^J of *all* journeys whose origin and destination zone pairs are in S_i, and which contribute to no underrepresented constraint. We randomly choose $T_i - X_i$ of them and mark them for deletion. In case the size of S_i^J is less then $T_i - X_i$, all the journeys in are marked. If S_i^J is

empty, the constraint is deemed unsatisfiable and is removed from the constraint list. It is clear that a journey might be marked for deletion multiple times. this is intended and beneficial, as it would mean that multiple constraints can be improved by replacing a single journey. Finally, the set of journeys marked for deletion is replaced with another set of newly generated journeys.

In our case study, from real check-ins data provided by TEP, we defined two different types of constraints:

- constraint on the total number of passengers for each bus line;
- constraint on the total number of users for each stop.

In the first constraint, a total number of users per bus line is set, i.e. the number of users taking a given bus line without taking into account the stop where check-in takes place. In the second constraint, on the other hand, a total number of check-ins per stop is imposed, without taking the bus line into account. Other constraints could be imposed, such as those on the number of check-outs, but in order not to impose too many constraints, it was decided to start the calibration with only these two types of constraints and add further constraints later if necessary. We therefore have a total of 358 initial constraints, 35 of which are unsatisfiable and were therefore removed from the list and we iterated the calibration operation 40 times. Our constraint definition turns out to be very weak and does not allow for a fine tuned travel calibration. There may be several ways to achieve a fine-tuned calibration, such as using commuting data and census data to do a coefficient expansion analysis of the OD matrix. Using these datasets could go a long way toward improving the performance of the calibration process and could allow us to satisfy more data. These analyses were not carried out during my thesis because I had access to the commuting data at the end of the project and did not have time to derive the expansion of OD coefficients derived from the HFLB data by using the census data merged with the commuting data to derive the OD matrix from them. However, this analysis option remains a possible solution for a future implementation of the model. The results are reported in the following section.

6.3 Scenario definition

So far we have been concerned with reproducing the behaviour of the the real system in normal conditions. Beyond that, we are interested in simulating what-if scenarios, especially in relation to the impact of the Campus on public transport. With the creation of new scenarios, it is possible to modify the public transport service offered (such as schedules or bus routes) or change the university lecture timetable in order to better distribute the arrival of students and evaluate how these measures affect the filling of each bus route. Another interesting study could be the evaluation of scenarios during the COVID-19 pandemic or similar future events, and public transport is a place where many infections can occur. Indeed, during the pandemic period, limiting gatherings was one of the non-pharmaceutical measures implemented to limit the spread of the virus. With the creation of what-if scenarios, it is possible to assess the distance between users in the buses and check whether the social distance of 1.5m introduced by the government at the beginning of the pandemic.

POI category	Discard	ATECO Codes
Tranports	1	H, N, S
Leisure	1	I, R, S
Groceries	1	H, I
Public Services	1	D, E, O, P, Q, S, U
Education	1	P, M
Private Services	1	D, E, F, G, H, J, K, L, M N, S, T
Health	1	Q
Shop	1	G, H, I, S
Religion	1	S
Industries	1	A, B, C, D

The easiest way to define a new scenario is to "switch off" certain classes of points, either completely (e.g. complete closure of schools) or partially (partial closure or forcing restaurant to close at 10 p.m.). This is achieved by post-processing the calibration results to:

a) change of journey time for a part of journeys with a certain scope given by the POI of arrival;

b) delete those journeys that are no longer possible since their starting/ending POI category is no longer active;

c) in case of POI categories that are only active during part of the day, moving the journey to time when POIs are active.

Let us consider a new scenario in which we define the strategies to be implemented, such as changing the start time of lessons. To each POI category defined in Sec. 3.2.4, we assign a series of ATECO codes (used by the ISTAT to categorize economic activities) in the following way: where the middle columns of the table defines the probability of discarding a journey to/from POI category. The correspondence between POI and ATECO codes is included in order to assess the impact of smartworking introduced for certain categories of workers during the COVID-19 pandemic. For each of them, we can introduce a fraction of home workers, using data from the Italian government. In addition to defining the rejection probability, we can define different scenario *bands*, specifying variations in opening/closing times:

- Band 0: No variation.
- Band 1: Postpone 25% of journeys to education POIs starting between 7:30 a.m. and 9 a.m. by one hour.
- Band 2: Postpone 50% of journeys to education POIs starting between 7:30 a.m. and 9 a.m.: 25% by one hour and 25% by two hours.
- Band 3: Postpone all journeys to education and leisure POIs by two hours.

Further bands can be defined, these are the ones we have tried to apply so far.

Once the probability of discarding and different bands have been defined, the synthetic trip dataset obtained after calibration is modified by removing or moving

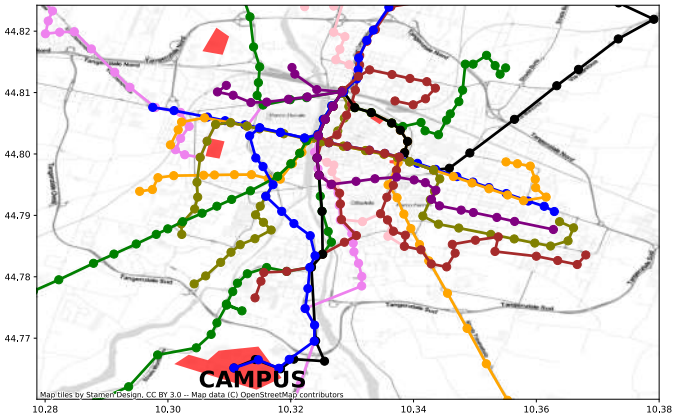


Figure 6.11: **Urban bus lines in city center** We plot the route of all urban lines of Parma public transportation. In particular, the campus is connected to the city centre by only two lines (blue and black lines) which have a very similar route.

the selected trips. In this way, it is possible to evaluate the filling of the means of transport and assess how the changes have affected public transport.

In the cases just described, scenarios are generated by modifying the trip dataset obtained after calibration. An alternative to have new scenarios is to change the public transport service, such as the timetable or the addition of new trips or new lines. In this case, the creation of new scenarios is more complicated because it means going to modify the GTFS files to re-evaluate all the routes of journeys generated by the CSA [27] using the new connection dataset. The implementation of these scenarios is currently only theoretical and will be defined more precisely in the next step of the research activity.

6.4 A case study: Parma mobility prediction

In this section, we apply the model to the metropolitan area of Parma and present the results obtained. The work is not yet complete, at the moment we have only stopped at the step of calibration and validation of the set of synthetic trips.

We have applied the model to the metropolitan area defined in Sec. 6.1.1, within which there are both urban and suburban bus lines. The urban lines are depicted in Fig. where each different colour represents a different line. In particular, we are interested in mobility by public transport to the Campus. The Campus is connected to the city centre by only two lines, the “7” and the “21” (see Fig. 6.11), which are the two lines to which we will refer in more detail. The algorithm generates a set of 200,000 public transport trips from empirical data obtained from the HFLB data of the GPS signal of mobile devices. As an output, we obtain a dataset of trips, each of which is given the departure time, departure and arrival zone, connection list (i.e. list of connections between two successive stops) with the corresponding bus line. From this dataset, and in particular from the list of connections for each trip, it is possible to derive the route taken for each trip generated, to know the bus line on which the user boarded and whether any line changes were made during the trip. Fig. 6.12 shows four examples of trips, two of which have the University Campus as their destination and two of which have a change of line. The dot indicates the starting stop, while if the line is represented by two colours, it means that the user

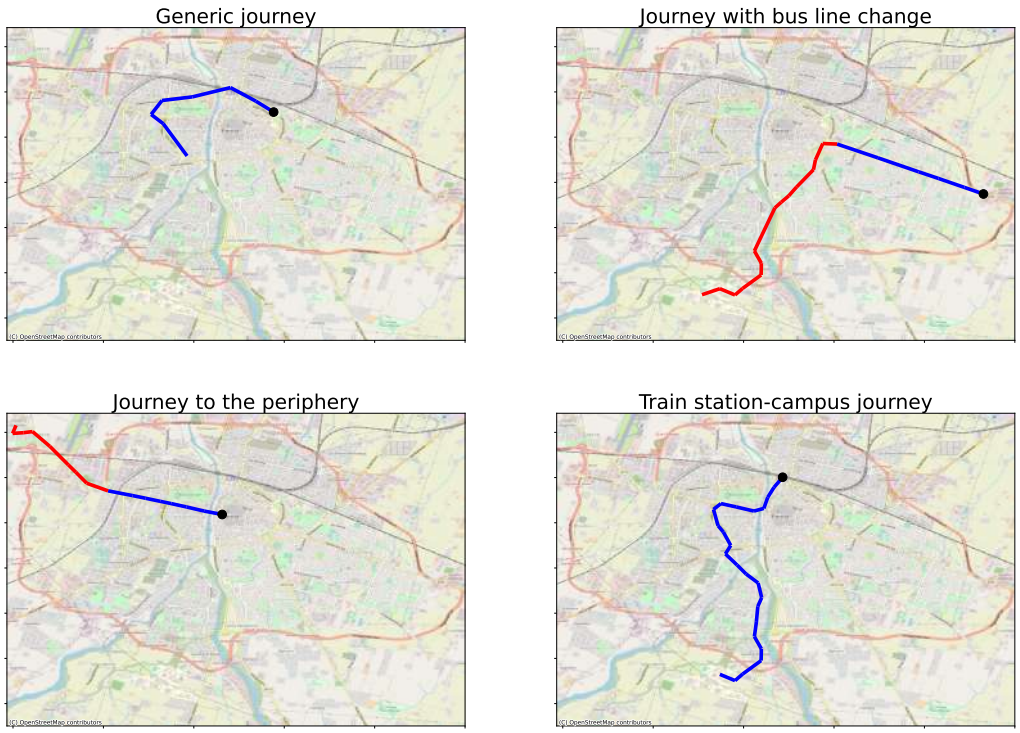


Figure 6.12: **Synthetic journeys generated by the algorithm.** In this figure we plot four example of synthetic journeys generated by the algorithm before the calibration process. The black point represent the starting stop, while if the trips is represented by two colours, it means theat the user had to make a bus line change. In particular, the bottom right panel shows a typical trip to the University Campus from train station with bus line " 7". In the top right panel shows a journey from

had to make a line change. In Fig. 6.12, the panel at the top right shows that the fastest trip by public transport from the eastern periphery of Parma to the Campus requires a change of line, an operation that results in an extension of travel time. The university campus is a hub for thousands of users who come from all over the urban (and extra-urban) area of Parma but is only served by two bus lines that do not manage to cover the entire urban territory, forcing users to change lines with a consequent loss of time (as in bottom right panel in Fig. 6.12). From the dataset of synthetic journeys, we understand that one of the possible scenarios that could be created would be one in which there are more than two urban lines leading to the Campus and that would improve the coverage of the city's urban territory and verify whether more direct trips to the campus would be created without the change of line, in order to offer a more efficient service.

After generating the first dataset of trips, without a calibration step, we compared the time distribution of the trips obtained from the dataset of real checkins, with the dataset of checkins from the dataset of trips generated by the model. In Fig. 6.14 we immediately notice that the two distributions differ from each other: the model underestimates morning users by about 2000 travellers, underestimates the number of users taking the bus during the lunch hours and overestimates users making commuting trips during the evening rush hour by about 3000. The calibration step, which replaces the trips that do not meet the imposed constraints, improves the output of the model, almost completely eliminating the underestima-

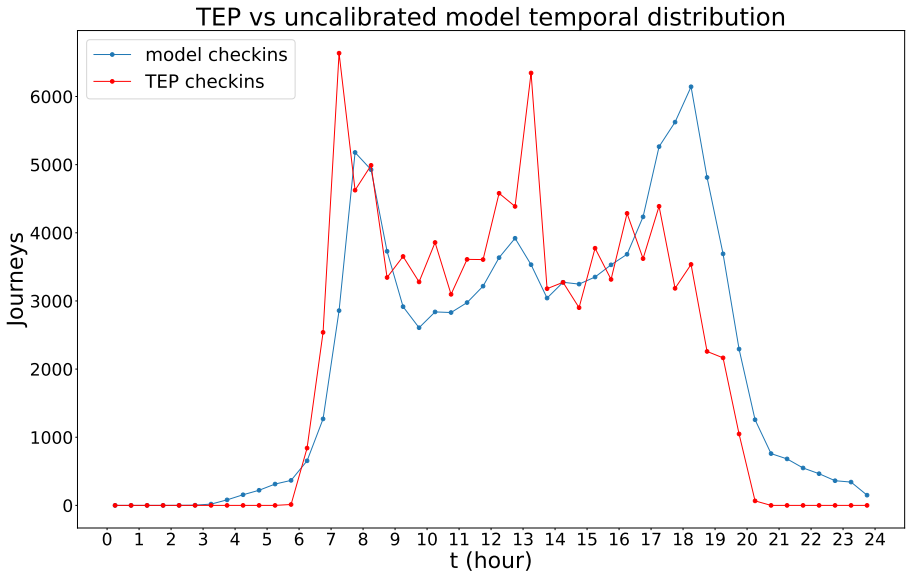


Figure 6.13: **Temporal distribution of synthetic journeys before calibration**In this figure, the distribution of actual trips (taken from tep data) is compared with the time distribution of synthetic trips before calibration. We can see that the model underestimates users during the morning and lunchtime peak hours, while it predicts a peak of users from 5 p.m. to 7 p.m. in disagreement with the real data.

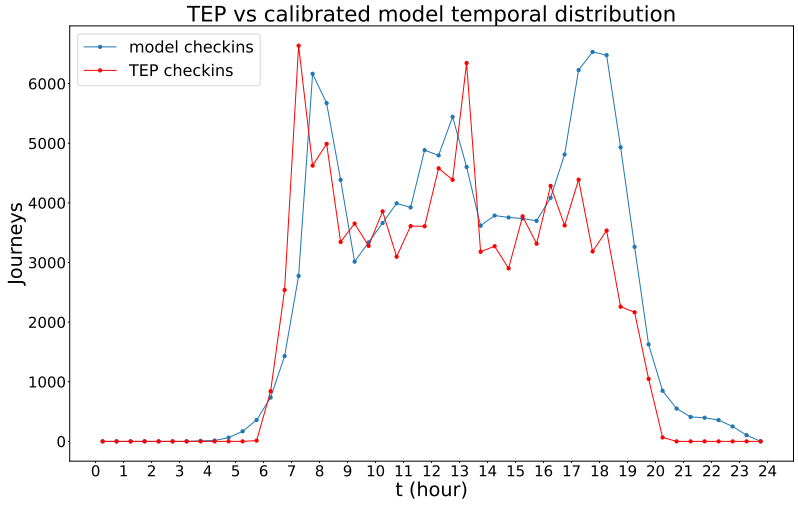


Figure 6.14: **Temporal distribution of synthetic journeys after calibration** In this figure, the distribution of actual trips (derived from the tep data) is compared with the time distribution of synthetic trips after calibration. We can see that calibration improved the prediction of passengers in the hours where the model underestimated the number of users, while it continues to overestimate users in the evening hours.

tion of users during the 7-9 a.m. and 12.30 p.m. to 2 p.m. time slots. However, even after calibration, the model still overestimates the number of people using public transport as a means of transport during the evening peak hours. This discrepancy between model and reality leads to an error in the prediction of bus line occupancy during the evening, finding overcrowding situations that do not occur in reality. The difference between model and real data could be due to three different error factors:

1. Error in checkins measurements made by TEP. Measurements of checkins and checkouts were carried out in a manual manner during working days in an undefined period of 2015. This measurement was done in a punctual manner by taking measurements of checkins on different lines on different days. No statistics were made on multi-day access, which could lead to an underestimation of public transport users during the evening rush hour. A measurement of checkins on several days, resulting in a calculation of average accesses could limit the effects of daily fluctuations of users.

2. Error on empirical data derived from HFLBD data. All empirical data derived from the HFLB data used as input in the generation model were carried out by performing a statistical analysis on the entire trip dataset. However, the model only deals with the generation of public transport trips. The overall dynamics of trips may not correspond to the dynamics of public transport trips and this could be one of the factors for the difference between the actual and synthetic time distribution.

3. Real data from checkins and the HFLD data refer to two different years, 2015 and 2017 respectively. From 2015 to 2017, the collective behaviour of commuting travellers may have changed to make greater use of public transport for journeys home in the evening.

Of these three points, the only thing that can be done in a short time is to review the statistical analysis performed on the travel dataset from the HFLB data. The idea I am developing is to create a process of analysing the empirically obtained trips to determine the means of transport by which they occur. In this way, it is then possible to select the subset of only those trips that were associated with the use of public transport and obtain a new OD matrix and temporal distribution for this specific dataset. This process is under development and very complicated, requiring careful analysis of HFLB data to create a universal workflow.

In addition to the problems just described for the temporal distribution of synthetic trips, we nevertheless continued the analysis of the data obtained and analysed the effectiveness of the calibration process. The aim of calibration is to improve the agreement between actual data and model output. To verify the performance of the calibration, I performed a comparison between the checkins data provided by TEP and the checkins data given by the model for each bus stop. Fig. 6.15 shows two scatter plots comparing real data and model data, the first with the data before calibration, the second after calibration. We immediately notice how the scatter plot before calibration shows that the real and synthetic checkins data are very well correlated with each other with a Pearson correlation coefficient $r^2 = 0.58$. After calibration, on the other hand, there is a good correlation between observed data from the model dataset with the target values set with the constraints, in which case the Pearson correlation coefficient becomes $r^2 = 0.91$. From this we can see that

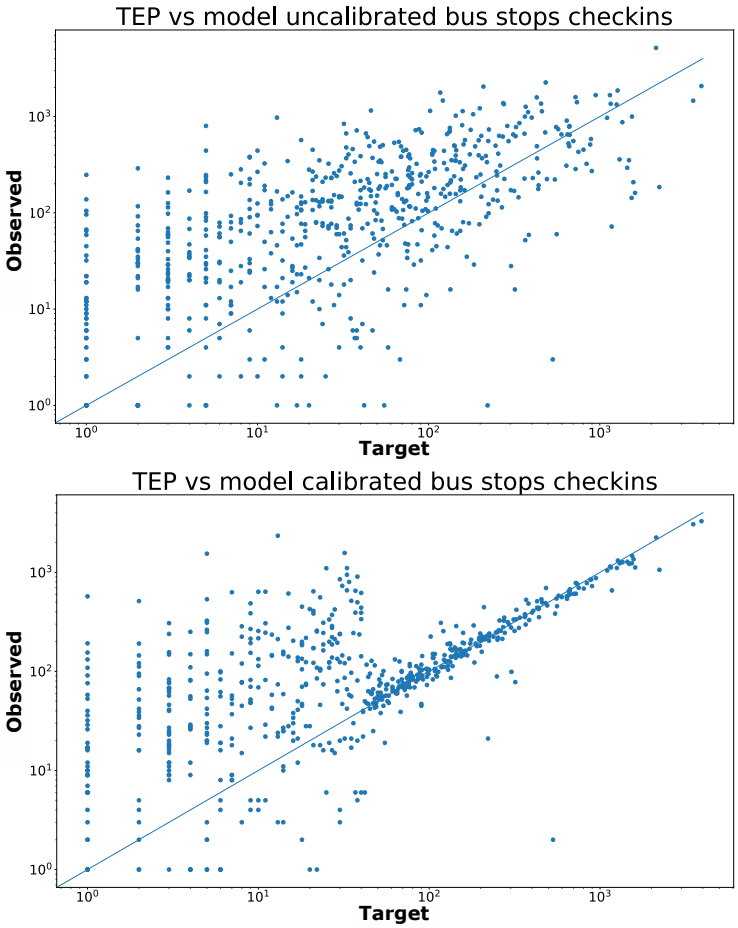


Figure 6.15: **Observed vs target checkins for each stop** Scatter plot of the number of checkins per bus stop between real data and data obtained from the synthetic dataset. In the top panel, we have the comparison between TEP data and data before calibration, while in the bottom panel the comparison is between real data and post-calibration data. It can easily be seen that calibration but greatly improve the correspondence between the model and reality for stops with a high number of checkins.

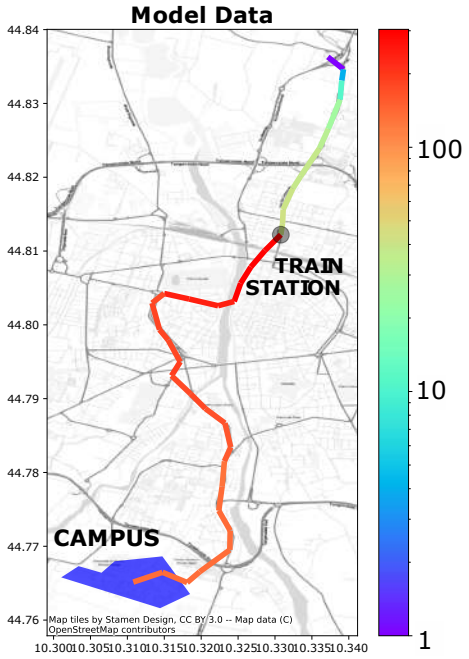


Figure 6.16: **Bus route filling in morning rush hour.** The picture shows the model's prediction of the number of passengers on the "7" line leaving at 8.10 a.m. in the direction of the University Campus. It can be seen that before the train station, the model predicts about 10-15 people, while after the station the predicted number of people increases by a factor of 10 (the colorbar is represented on a logarithmic scale) until the end of the run.

the calibration process is very efficient and should not require further development steps.

To conclude, at the moment we have only verified on a few runs how the number of passengers varies within the buses along the entire line. An example is shown in Fig. 6.16 where the number of users per connection on the 8.10 a.m. run of the line "7" in the direction of the University Campus is shown. In this case we see a very borderline case with an overcrowded situation with more than 120 users simultaneously expected on the route between the railway station and the University Campus. This number of users is too high for the capacity of the bus and therefore in the real case some users would have to wait for the next bus. Furthermore, the overcrowding would lead to people not maintaining the interpersonal distance that was introduced by the government during the Covid-19 pandemic. Another detail that can be noted is that the bus is partially crowded before arriving at the stop near the train station, which leads us to think that the users of this particular route arrive in Parma from neighbouring cities and regions on their way to the Campus. Better management of the bus and its timetable based on train schedules with a high number of people could improve the efficiency of transport.

At the moment, work on this line of research has stopped at this point. We are trying to refine the statistical analysis of the HFLB data in order to find the subset of trips by public transport, to see if we can get a better match with the time distribution obtained from the TEP surveys. The next step will be to implement the algorithm for creating the scenarios and the analysis of them.

Chapter 7

Conclusion

This thesis work focused on the study of data-driven approaches for the study and modeling for the reconstruction and prediction of patterns of mobility and interactions in social environments, with a particular focus on the study of social dynamics occurring on the university science campus in Parma.

In the first part of this thesis, we reviewed the fundamentals of the theory and application for data-driven use of the two main types of data that we had available to develop the scientific research presented in this paper: data from the WiFi network and High-Frequency-Located-based (HFLB) data. We initially introduced the data that can be obtained from WiFi network, describing their characteristics, advantages and disadvantages of using this type of data to monitor social dynamics. We saw that WiFi network can be found installed in a wide type of environments, particularly in public environments such as the University. Finally, we described the dataset at our disposal provided by the University of Parma related to the Covid-19 pandemic period, its preprocessing steps to eliminate sensitive data and to filter out incomplete or unnecessary data for our goal.
Next, we reviewed the foundations of High-Frequency-Located-Based data, the advantages and disadvantages of these data for the empirical description of human mobility. We introduced techniques used to perform an initial analysis of geolocation data with the goal of including a process to filter noise and to be able to enrich GPS data with other datasets such as, POIs data obtained from Open Street Map. This introduction was followed by a brief analysis of the dataset that Sony CSL provided to me related to the province of Parma dating back to the year 2017.

In the second part of the thesis, we formulated a set of general strategies to monitor the presence of users, the formation of groups, and their temporal evolution in restricted areas using only WiFi network connection data. We introduced two different quantities: the typical group size and the average number of links between different users formed in areas reached by the WiFi signal through Access Points (APs). These quantities allow us to classify areas according to the size of the groups formed in them and according to the reshuffling that occurs during a working day. We applied these measures to study the number of attendances, social dynamics and use of spaces within the University of Parma and, in particular, the Scientific Campus. The available dataset allowed us to perform the analysis during the Covid-19 pandemic in which there were three phases with three distinct containment measures. We classified the areas of the Campus to determine the areas

that are potentially more dangerous due to the risk of infection. In addition, we used measures of group size and average number of links to effectively monitor the different use of spaces by different classes of users, such as students and staff.

Within an approach to epidemic spreading on simplicial temporal networks, using the size distribution of the simplices as an input parameter to the theory, our analysis provides a specific estimate of the dramatic increase in the value of the reproduction number that occurs in the fully reopened phase of the University due to the increase in contacts.

In addition to application to epidemic models, WiFi data provide an easy-to-use tool for obtaining the structure and evolution of groups of people moving and connecting in the same environments. In particular, the measurements obtained from WiFi make it possible to find the presence of more complex sub-structures nested other than fully connected simplices.

Then, we considered a compartmental model on simplicial activity-driven networks, specific for transmission of the SARS-CoV-2 pathogen, based on an SIR model which takes into account asymptomatic and presymptomatic transmission. We developed this compartmental model and extended it by introducing manual contact tracing mechanisms. The proposed model, allows us to clearly estimate the effect of CT mechanisms on reproduction number, going to show that efficient contact tracing allows us to limit the spread of the disease. We used empirical data of the simplex size distribution, obtained from connections to the WiFi network of Parma University, as input to the epidemic model and analyzed the effects of CTs an application of the model on a real case.

In the final part of the thesis, we set out the research carried out in my last year after the end of the pandemic. We introduced a new data-driven model for predicting urban mobility using public transportation. The model makes it possible to describe individual behavior by going to determine the trajectories of the individual user who moves on the public transport network and collective behavior by going to see the filling conditions of each means of transport. We applied the model in the Parma metropolitan area, with the aim of describing the effect of the Science Campus on the filling of bus routes. We found that the high number of people using the campus lead to bus overcrowding during peak hours. The model is still under development but was formulated to be able to create alternative scenarios to find strategies to improve the efficiency of public transportation.

This model has some critical issues due to the input parameters empirically derived from actual trajectories. Parameters were derived from a complete set of trajectories, in which bicycle, walking, and private transportation trips are included. Improving the trip analysis by including a process for determining and selecting the mode of transportation would greatly improve the model and its description of urban mobility with public transportation.

The approach proposed in this thesis, in the study of social dynamics, urban mobility and epidemic models, opens up promising new directions in the study of complex systems and the use of data-driven method to apply models in real cases. Promising prospects are in the direction of modeling trajectories inside the University Campus, creation of scenarios for improving public transportation, and modeling and incentivizing car sharing to make urban transportation more sustainable.

Appendix A

Derivation of epidemic threshold via mean field approach

In this chapter we present the evaluation of the epidemic threshold derived from the derivation of the mean field equations for the epidemic model described in chapter 5, evolving on an adaptive-driven network in the presence of simplicial interactions. The containment measures are implemented as contact tracing of asymptomatic nodes with its forward, backward and sideward implementations. The epidemic model proposed is a Susceptible-Infected-Recovered (SIR) model, with further distinctions for the stage of the infection. The distinctions are based on the presence of symptoms (P - infected presymptomatic, I - infected symptomatic, A - infected asymptomatic), on tracing and isolation (A_T - asymptomatic traced and A_Q - asymptomatic quarantined) and they model changes in social behaviour depending on nodes' health status. We do not consider here memory [102] nor burstiness effects in the dynamics [70]. The allowed transitions among these states are shown in Supplementary Fig. A.1, which coincides with Fig. 5.1**b** of the main text and which we reproduce here for simplicity.

The infection and CT events are discussed in the main text, together with the CT mechanisms. An individual can be infected with a probability of δ to develop symptoms or $(1-\delta)$ without symptoms. A manual implementation of contact tracing (CT) is also considered [68], where all the individuals in the gathering that a symptomatic individual participated in are traced and the tracing is effective with probability $\epsilon(s)$ depending on the size of the gathering. The text further explains the specific conditions under which infections and tracing transitions take place within a gathering

Considering the necessary and sufficient (minimum) conditions for them to occur in a simplex, the transitions of infection and tracing are described below.

$$P + S \xrightarrow{\lambda\delta} P + P \tag{A.1}$$

$$P + S \xrightarrow{\lambda(1-\delta)\epsilon} P + A_T \tag{A.2}$$

$$P + S \xrightarrow{\lambda(1-\delta)(1-\epsilon)} P + A \tag{A.3}$$

A susceptible node S can be infected by a presymptomatic node P in a variety of ways. If the infection is symptomatic, the susceptible node will become a presymptomatic P as per Eq.(A.1). However, if the infection is asymptomatic, the

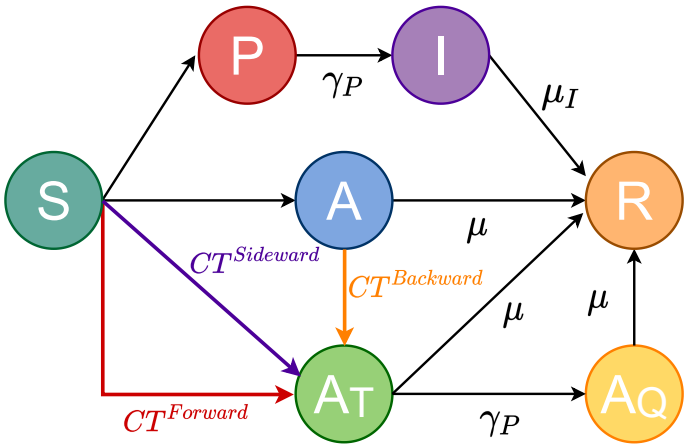


Figure A.1: **Epidemic model with contact tracing.** We plot the scheme of the compartmental epidemic model together with the transitions for CT. The rates for the infection events and for the CT mechanisms are not indicated here but they are detailed in the text.

contagion can be traced through the use of forward CT activated by the infector P, and the susceptible node will become a traced asymptomatic A_T as per Eq.(A.2). In the event that the simplex is not traced, the susceptible node will become an asymptomatic A as per Eq. (A.3).

$$A + S \xrightarrow{\lambda\delta\epsilon} A_T + P \tag{A.4}$$

$$A + S \xrightarrow{\lambda\delta(1-\epsilon)} A + P \tag{A.5}$$

$$A_T + S \xrightarrow{\lambda\delta} A_T + P \tag{A.6}$$

A susceptible node S can be infected by an asymptomatic A (or traced asymptomatic A_T) node with symptomatic infection and thus becomes presymptomatic P. The asymptomatic infector can be traced by backward CT, activated by the newly infected P, becoming traced asymptomatic A_T (Eq. (A.4)), otherwise if the simplex is not traced the infector does not change status (Eq. (A.5)). In principle also a traced asymptomatic infector can be traced with backward CT but this has actually no consequences (Eq. (A.6)). In the event of Eq. (A.4) both individuals change state.

$$A + S \xrightarrow{\lambda(1-\delta)\epsilon} A + A_T \quad \text{if in the same simplex exists} \quad A + S \xrightarrow{\lambda\delta} A + P \tag{A.7}$$

$$A + S \xrightarrow{\lambda(1-\delta)(1-\epsilon)} A + A \quad \text{if in the same simplex exists} \quad A + S \xrightarrow{\lambda\delta} A + P \tag{A.8}$$

$$A + S \xrightarrow{\lambda(1-\delta)} A + A \quad \text{otherwise} \tag{A.9}$$

$$A_T + S \xrightarrow{\lambda(1-\delta)\epsilon} A_T + A_T \quad \text{if in the same simplex exists} \quad A_T + S \xrightarrow{\lambda\delta} A_T + P \tag{A.10}$$

$$A_T + S \xrightarrow{\lambda(1-\delta)(1-\epsilon)} A_T + A \quad \text{if in the same simplex exists} \quad A_T + S \xrightarrow{\lambda\delta} A_T + P \tag{A.11}$$

$$A_T + S \xrightarrow{\lambda(1-\delta)} A_T + A \quad \text{otherwise} \tag{A.12}$$

Finally, the spontaneous transitions are:

$$P \xrightarrow{\gamma_P} I \qquad A \xrightarrow{\mu} R \qquad I \xrightarrow{\mu_I} R \tag{A.13}$$

$$A_T \xrightarrow{\gamma_P} A_Q \qquad A_T \xrightarrow{\mu} R \qquad A_Q \xrightarrow{\mu} R \tag{A.14}$$

which account for spontaneous recovery, spontaneous symptoms development and isolation of traced individuals.

We apply a *mean-field* approach, which is exact since local correlations are continuously destroyed because of link reshuffling. Therefore, the epidemic thresholds of the SIR and SIS (Susceptible-Infected-Susceptible) models coincide [102]. This allows us to determine the threshold by considering the mean-field equations for the SIS version of the dynamics.

The probability of a node belonging to a specific compartment of the epidemic model, as outlined in Supplementary Fig. 5.1**b**, is analyzed. Equations for the time-based progression of these probabilities are developed, considering the network dynamics, epidemic spread, and CT dynamics. Using a mean-field approach, the epidemic dynamics is portrayed through the use of probabilities, which are:

- $P(t)$ for a node to be infected presymptomatic at time t;
- $I(t)$ for a node to be infected symptomatic at time t;
- $A(t)$ for a node to be infected asymptomatic at time t;
- $A_T(t)$ for a node to be asymptomatic traced at time t;
- $A_Q(t)$ for a node to be asymptomatic isolated at time t;
- $S(t) = 1 - P(t) - I(t) - A(t) - A_T(t) - A_Q(t)$ for a node to be susceptible at time t.

The network dynamics is described by the activation of simplices, which occur at a Poissonian rate of a. The size of each simplex is determined by the distribution $\Psi(s)$. Additionally, each node (individual) is assigned an attractiveness parameter b_i, which determines their likelihood of participating in social interactions. As a result, nodes join active simplices with probability based on their attractiveness $p_{b_i} \propto b_i$ [67, 68, 89]. To simplify the problem, the following assumptions were made:

- All susceptible nodes have the same attractiveness $b_S = b$. In this way, they participate equally in active simplices;
- symptomatic and quarantined asymptomatic are isolated and do not participate in simplices ($b_I = b_{A_Q} = 0$);
- presymptomatic, asymptomatic and traced asymptomatic, behave like susceptible nodes ($b_P = b_A = b_{A_T} = b_S = b$);
- we consider the problem in thermodynamic limit.

The probability $P(t)$ that a node is in the presymptomatic state evolves according to the following equation:

$$\partial_t P(t) = -\gamma_P P(t) + \int ds \Psi(s) a s P_S(t) Z_s(t) \delta \tag{A.15}$$

On the right-hand side of the equation there are two terms: the first term takes into account the process of symptom development, the second term takes into account the process of infection of susceptibles. In particular, the latter is given by the product of the activation rate a of a simplex of size s, the probability $sP_S(t)$ that one susceptible node participates in it, the probability $Z_s(t)$ that at least one of the other $(s-1)$ nodes infects the susceptible one and the probability δ that the infection is symptomatic. Both terms are averaged over the size of the simplex $\int ds \Psi(s)$.

Let X be one of the compartment states of the model, the probability that a node of a simplex belongs to compartment X is:

$$P_X(t) = X(t) \frac{b_X}{\bar{b}(t)} \tag{A.16}$$

Therefore $P_I(t) = P_{A_Q}(t) = 0 \, \forall t$ and $P_X(t) = \frac{X(t)}{S(t)+P(t)+A(t)+A_T(t)}$ for $X = S, P, A, A_T$.

The probability $Z_s(t)$ of at least one of the other $(s-1)$ nodes infecting the susceptible node is

$$Z_s(t) = 1 - \xi(t)^{s-1} \tag{A.17}$$

where $\xi(t)$ is the probability that a node in the simplex does not infect the susceptible one. We can rewrite this quantity as

$$\begin{aligned} \xi(t) &= P_S(t) + (1-\lambda)[P_P(t) + P_A(t) + P_{A_T}(t)] \\ &= 1 - \lambda \frac{P(t) + A(t) + A_T(t)}{S(t) + P(t) + A(t) + A_T(t)} \end{aligned} \tag{A.18}$$

Thus, the complete equation for the time evolution of $P(t)$ is:

$$\begin{aligned} \partial_t P(t) = -\gamma_P P(t) + a \frac{S(t)}{S(t) + P(t) + A(t) + A_T(t)} \cdot \\ \cdot \delta \left\langle s \left[1 - \left(1 - \lambda \frac{P(t) + A(t) + A_T(t)}{S(t) + P(t) + A(t) + A_T(t)} \right)^{s-1} \right] \right\rangle \end{aligned} \tag{A.19}$$

where we indicate with $\langle f(s) \rangle = \int ds \Psi(s) f(s)$: the first term on the right hand side accounts for spontaneous recovery and the second term for symptomatic infections in simplices.

The equation for the probability $I(t)$ that a node is in the symptomatic infected state is trivially (eq. A.20) and contain one term for the spontaneous recovery and one term for spontaneous symptoms development.

$$\partial_t I(t) = -\mu_I I(t) + \gamma_P P(t) \tag{A.20}$$

In the other hand, the equation for the probability $A(t)$ to be in the untraced or non quarantined asymptomatic state is the most complex:

$$\begin{aligned} \partial_t A(t) = -\mu A(t) &+ \int ds \Psi(s) a s P_S(t) Z_s(t)(1-\delta) \\ &- \int ds \Psi(s) C_s^{Forward}(t) \\ &- \int ds \Psi(s) C_s^{Backward}(t) \\ &- \int ds \Psi(s) C_s^{Sideward}(t). \end{aligned} \tag{A.21}$$

The equation on the right hand side accounts for the dynamics of the epidemic model. The first term represents the rate at which individuals recover spontaneously. The second term represents the rate of contagion in active simplices, taking into account the activation rate a, the probability that a susceptible node $P_S(t)$ participates in the simplex, and the probability that at least one other node in the simplex infects the susceptible node. The third, fourth, and fifth terms represent the rates of forward, backward, and sideward CT, respectively. All terms are averaged over the size of the simplex, using the distribution $\Psi(s)$. The evaluation of each CT term is similar to the evaluation of the infection term. After we evaluate separattely each term.

A.1 Forward CT

The forward CT, as previously discussed, traces an asymptomatic individual who was infected by a presymptomatic node that will develop symptoms and activate CT. Therefore, the forward CT term accounts for the activation rate a of a simplex of size s, the probability $sP_S(t)$ that a susceptible node participates in the simplex, and the probability $F_s(t)$ that at least one of the other $(s-1)$ nodes is a presymptomatic node P who infects the susceptible node with an asymptomatic infection $(1-\delta)$. The simplex is traced with probability $\epsilon(s)$.

$$F_s(t) = 1 - k(t)^{s-1} \tag{A.22}$$

is the probability that at least one of the other $(s-1)$ nodes is presymptomatic and infects the susceptible one. $k(t)$ is the probability that a node in the simplex is not in the presymptomatic state or, if they are presymptomatic, they do not infect the susceptible node.

$$k(t) = P_S(t) + P_A(t) + P_{A_T}(t) + (1-\lambda)P_P(t) = 1 - \lambda \frac{P(t)}{S(t) + P(t) + A(t) + A_T(t)} \tag{A.23}$$

Thus, we obtain:

$$\begin{aligned} C^{Forward} =& a\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}(1-\delta)\cdot \\ & \cdot\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{P(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle \end{aligned} \tag{A.24}$$

A.2 Backward CT

$$C^{Backward} = \int ds\Psi(s)asP_A(t)\epsilon(s)W_s(t) \tag{A.25}$$

The backward CT term accounts for the probability that an asymptomatic individual who infects a susceptible node and produces a symptomatic infection will be traced. This is determined by the activation rate a of a simplex of size s, the probability $sP_A(t)$ that an asymptomatic node participates in the simplex, and the probability $W_s(t)$ that at least one of the other $(s-1)$ nodes is a susceptible node who is infected by the asymptomatic individual with a symptomatic infection. The simplex is traced with a probability

$$W_s(t) = 1-\phi(t)^{s-1}. \tag{A.26}$$

The backward CT term accounts for the probability that at least one of the other $(s-1)$ nodes in a simplex of size s is a susceptible node that is infected by an asymptomatic individual with a symptomatic infection. The probability that a node in the simplex is not a susceptible node infected with a presymptomatic infection is represented by $\phi(t)$.

$$\begin{aligned} \phi(t) &= P_P(t)+P_A(t)+P_{A_T}(t)+(1-\lambda)P_S(t)+\lambda(1-\delta)P_S(t) \\ &= 1-\lambda\delta\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)} \end{aligned} \tag{A.27}$$

Thus, we obtain:

$$\begin{aligned} C^{Backward} =& a\frac{A(t)}{S(t)+P(t)+A(t)+A_T(t)}\cdot \\ & \cdot\left\langle \epsilon(s)s\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle \end{aligned} \tag{A.28}$$

A.3 Sideward CT

$$C^{Sideward} = \int ds \Psi(s) a s P_S(t)(1-\delta) H_s(t) \epsilon(s) K_s(t) \tag{A.29}$$

The sideward CT traces an asymptomatic individual who is infected by another asymptomatic (or traced asymptomatic) individual, in the presence of a susceptible node that is infected with a symptomatic infection and subsequently activates CT. The sideward CT term accounts for the activation rate a of a simplex of size s, the probability $sP_S(t)$ that a susceptible node participates in it, the probability $H_s(t)$ that at least one of the remaining $(s-1)$ nodes is infected asymptomatically (or traced asymptomatically) and infects the susceptible node with an asymptomatic infection $(1-\delta)$, and the probability $K_s(t)$ that among the remaining $(s-2)$ nodes, at least one of them is a susceptible node infected with a symptomatic infection in the simplex. The simplex is traced with probability $\epsilon(s)$, similarly to the cases of forward and backward CT:

$$H_s(t) = 1 - h(t)^{s-1} \tag{A.30}$$

where $h(t)$ is the probability for a node of the simplex not to infect the susceptible node unless he/she is P.

$$\begin{aligned} h(t) &= P_S(t) + (1-\lambda)(P_A(t) + P_{A_T}(t)) + P_P(t) \\ &= 1 - \lambda \frac{A(t) + A_T(t)}{S(t) + P(t) + A(t) + A_T(t)} \end{aligned} \tag{A.31}$$

Analogously

$$K_s(t) = 1 - \phi(t)^{s-2} \tag{A.32}$$

where $\phi(t)$ is given by Eq. (A.27).
Thus, we obtain:

$$\begin{aligned} C^{Sideward} &= a \frac{S(t)}{S(t)+Y(t)}(1-\delta) \cdot \\ &\cdot \left\langle \epsilon(s) s \left[1 - \left(1 - \lambda \frac{A(t)+A_T(t)}{S(t)+Y(t)} \right)^{s-1} \right] \left[1 - \left(1 - \lambda\delta \frac{S(t)}{S(t)+Y(t)} \right)^{s-2} \right] \right\rangle \end{aligned} \tag{A.33}$$

where $Y(t) = P(t) + A(t) + A_T(t)$.
For the evolution of $A(t)$, the complete equation is:

$$\begin{aligned}\partial_t A(t) = &- \mu A(t) + a\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}(1-\delta)\cdot \\ &\cdot\left\langle s\left[1-\left(1-\lambda\frac{P(t)+A(t)+A_T(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle \\ &-a\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}(1-\delta)\cdot \\ &\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{P(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle \\ &-a\frac{A(t)}{S(t)+P(t)+A(t)+A_T(t)}\cdot \\ &\left\langle \epsilon(s)s\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle \\ &-a\frac{S(t)}{S(t)+Y(t)}(1-\delta)\cdot \\ &\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{A(t)+A_T(t)}{S(t)+Y(t)}\right)^{s-1}\right]\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+Y(t)}\right)^{s-2}\right]\right\rangle\end{aligned} \tag{A.34}$$

where the first term on the right-hand side accounts for spontaneous recovery, the second term accounts for asymptomatic infections in simplices, and the third, fourth, and fifth terms account respectively for forward, backward, and sideward contact tracing. The equation for the probability of being in the asymptomatic traced state ($A_T(t)$) is:

$$\partial_t A_T(t) = -(\mu+\gamma_P)A_T(t) + C^{Forward} + C^{Backward} + C^{Sideward} \tag{A.35}$$

Thus substituting:

$$\begin{aligned}\partial_t A_T(t) =& -(\mu+\gamma_P)A_T(t)\\ &+a\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}(1-\delta)\cdot\\ &\cdot\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{P(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle\\ &+a\frac{A(t)}{S(t)+P(t)+A(t)+A_T(t)}\cdot\\ &\cdot\left\langle \epsilon(s)s\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+P(t)+A(t)+A_T(t)}\right)^{s-1}\right]\right\rangle\\ &+a\frac{S(t)}{S(t)+Y(t)}(1-\delta)\cdot\\ &\cdot\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{A(t)+A_T(t)}{S(t)+Y(t)}\right)^{s-1}\right]\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+Y(t)}\right)^{s-2}\right]\right\rangle.\end{aligned} \tag{A.36}$$

Finally, the probability of being in the quarantine asymptomatic state is given by the following equation

$$\partial_t A_Q(t) = -\mu A_Q(t) + \gamma_P A_T(t) \tag{A.37}$$

where we can observe two terms: first term accounts for spontaneous recovery and the second term account for isolation of traced asymptomatic nodes.

To conclude, we obtain a set of 5 coupled differentaial non-linear equations:

$$\partial_t P(t) = -\gamma_P P(t) + a\frac{S(t)}{S(t)+Y(t)}\delta\left\langle s\left[1-\left(1-\lambda\frac{Y(t)}{S(t)+Y(t)}\right)^{s-1}\right]\right\rangle \tag{A.38}$$

$$\partial_t I(t) = -\mu_I I(t) + \gamma_P P(t) \tag{A.39}$$

$$\begin{aligned}\partial_t A(t) =& -\mu A(t) + a\frac{S(t)}{S(t)+Y(t)}(1-\delta)\left\langle s\left[1-\left(1-\lambda\frac{Y(t)}{S(t)+Y(t)}\right)^{s-1}\right]\right\rangle\\ &- C^{Forward} - C^{Backward} - C^{Sideward}\end{aligned} \tag{A.40}$$

$$\partial_t A_T(t) = -(\mu+\gamma_P)A_T(t) + C^{Forward} + C^{Backward} + C^{Sideward} \tag{A.41}$$

$$\partial_t A_Q(t) = -\mu A_Q(t) + \gamma_P A_T(t) \tag{A.42}$$

with

$$C^{Forward} = a\frac{S(t)}{S(t)+Y(t)}(1-\delta)\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{P(t)}{S(t)+Y(t)}\right)^{s-1}\right]\right\rangle \tag{A.43}$$

$$C^{Backward} = a\frac{A(t)}{S(t)+Y(t)}\left\langle \epsilon(s)s\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+Y(t)}\right)^{s-1}\right]\right\rangle \tag{A.44}$$

$$\begin{aligned} C^{Sideward} &= a\frac{S(t)}{S(t)+Y(t)}(1-\delta)\cdot \\ &\cdot\left\langle \epsilon(s)s\left[1-\left(1-\lambda\frac{A(t)+A_T(t)}{S(t)+Y(t)}\right)^{s-1}\right]\left[1-\left(1-\lambda\delta\frac{S(t)}{S(t)+Y(t)}\right)^{s-2}\right]\right\rangle \end{aligned} \tag{A.45}$$

where $S(t) = 1-P(t)-I(t)-A(t)-A_T(t)-A_Q(t)$ and $Y(t) = P(t)+A(t)+A_T(t)$.

This set of equations allows for an absorbing state where the entire population is susceptible. To determine the conditions under which the absorbing state is stable, we perform a linear stability analysis around it. We use λ as a control parameter and the epidemic threshold is represented by λ_C.

To simplify the analysis, we disregard second-order terms in probabilities, resulting in a set of 5 linearized differential equations:

$$\partial_t P(t) = -\gamma_P P(t) + \lambda\delta a\langle s(s-1)\rangle[P(t)+A(t)+A_T(t)] \tag{A.46}$$

$$\partial_t I(t) = -\mu_I I(t) + \gamma_P P(t) \tag{A.47}$$

$$\begin{aligned} \partial_t A(t) = &-\mu A(t) + \lambda(1-\delta)a\langle s(s-1)\rangle[P(t)+A(t)+A_T(t)] \\ &- C^{Forward} - C^{Backward} - C^{Sideward} \end{aligned} \tag{A.48}$$

$$\partial_t A_T(t) = -(\mu+\gamma_P)A_T(t) + C^{Forward} + C^{Backward} + C^{Sideward} \tag{A.49}$$

$$\partial_t A_Q(t) = -\mu A_Q(t) + \gamma_P A_T(t) \tag{A.50}$$

where the linearized CT terms:

$$C^{Forward} = \lambda(1-\delta)a\langle\epsilon(s)s(s-1)\rangle P(t) \tag{A.51}$$

$$C^{Backward} = a\left\langle\epsilon(s)s\left[1-(1-\lambda\delta)^{s-1}\right]\right\rangle A(t) \tag{A.52}$$

$$C^{Sideward} = \lambda(1-\delta)a\left\langle\epsilon(s)s(s-1)\left[1-(1-\lambda\delta)^{s-2}\right]\right\rangle[A(t)+A_T(t)] \tag{A.53}$$

We will consider the average number of links established by an individual per unit time $\overline{n} = a\langle s(s-1)\rangle$ as a constant. By exploding this term in the linearized equations, we obtain

$$\partial_t P(t) = -\gamma_P P(t) + \lambda\delta\overline{n}[P(t)+A(t)+A_T(t)] \tag{A.54}$$

$$\partial_t I(t) = -\mu_I I(t) + \gamma_P P(t) \tag{A.55}$$

$$\begin{aligned} \partial_t A(t) = &-\mu A(t) + \lambda(1-\delta)\overline{n}[P(t)+A(t)+A_T(t)] \\ &- C^{Forward} - C^{Backward} - C^{Sideward} \end{aligned} \tag{A.56}$$

$$\partial_t A_T(t) = -(\mu+\gamma_P)A_T(t) + C^{Forward} + C^{Backward} + C^{Sideward} \tag{A.57}$$

$$\partial_t A_Q(t) = -\mu A_Q(t) + \gamma_P A_T(t) \tag{A.58}$$

where the linearized CT terms are:

$$C^{Forward} = \lambda(1-\delta)\overline{n}\frac{\langle\epsilon(s)s(s-1)\rangle}{\langle s(s-1)\rangle}P(t) \tag{A.59}$$

$$C^{Backward} = \frac{\overline{n}}{\langle s(s-1)\rangle}\left\langle\epsilon(s)s\left[1-(1-\lambda\delta)^{s-1}\right]\right\rangle A(t) \tag{A.60}$$

$$C^{Sideward} = \lambda(1-\delta)\frac{\overline{n}}{\langle s(s-1)\rangle}\left\langle\epsilon(s)s(s-1)\left[1-(1-\lambda\delta)^{s-2}\right]\right\rangle [A(t)+A_T(t)] \tag{A.61}$$

The set of linearized equations can be written as:

$$\begin{bmatrix}\partial_t A_Q(t)\\ \partial_t I(t)\\ \partial_t P(t)\\ \partial_t A(t)\\ \partial_t A_T(t)\end{bmatrix} = J\begin{bmatrix}A_Q(t)\\ I(t)\\ P(t)\\ A(t)\\ A_T(t)\end{bmatrix} \tag{A.62}$$

J is the Jacobian matrix of this set of 5 linearized equation:

$$J = \begin{bmatrix}-\mu & 0 & 0 & 0 & \gamma_P\\ 0 & -\mu_I & \gamma_P & 0 & 0\\ 0 & 0 & -\gamma_P+\beta & \beta & \beta\\ 0 & 0 & \Delta\left(1-\frac{\langle\epsilon(s)s(s-1)\rangle}{\langle s(s-1)\rangle}\right) & -\mu+\Delta-\Gamma-\Phi & \Delta-\Phi\\ 0 & 0 & \Delta\frac{\langle\epsilon(s)s(s-1)\rangle}{\langle s(s-1)\rangle} & +\Gamma+\Phi & -\mu-\gamma_P+\Phi\end{bmatrix} \tag{A.63}$$

$$J = \begin{bmatrix}\mathbb{A}(2\text{x}2) & \mathbb{C}(2\text{x}3)\\ \mathbb{O}(3\text{x}2) & \mathbb{B}(3\text{x}3)\end{bmatrix} \tag{A.64}$$

where

$$\Phi = \lambda(1-\delta)\frac{\overline{n}}{\langle s(s-1)\rangle}\left\langle\epsilon(s)s(s-1)\left[1-(1-\lambda\delta)^{s-2}\right]\right\rangle$$

$$\Gamma = \frac{\overline{n}}{\langle s(s-1)\rangle}\left\langle\epsilon(s)s\left[1-(1-\lambda\delta)^{s-1}\right]\right\rangle$$

$$\beta = \lambda\delta\overline{n}$$

$$\Delta = \lambda(1-\delta)\overline{n}.$$

The condition for the stability of the absorbing state, where all individuals are susceptible, is determined by analyzing the eigenvalues of the Jacobian matrix, J. This matrix is separated into two blocks, $\mathbb{A}$ and $\mathbb{B}$, and it is sufficient to study the eigenvalues of the latter block because the eigenvalues $\xi_1 = -\mu$, $\xi_2 = -\mu_I$ of block $\mathbb{A}$ are all negative. The threshold for the stability of the absorbing state is obtained by numerically diagonalizing the $\mathbb{B}$ matrix and ensuring that all eigenvalues are negative. The threshold value is represented by λ_C.

A.4 Limit case

In this section we have derived the mean filed equations for epidemic model for a general case, with an arbitrary $\Psi(s)$ and with a completely general $\epsilon(s)$. This

generalisation, allows to introduce complex effect s such as heterogeneity in simplices size and CT strategies.

Determining the epidemic threshold λ_C requires solving the stability conditions numerically, as the conditions are complex in structure. However, some limit cases can simplify the equations and allow for the explicit calculation of the threshold in analytic form. It is important to note that even if $\epsilon(s) = 1$ for all s, the epidemic threshold remains finite due to the presence of presymptomatic infections and the delay in isolating traced nodes.

A.4.1 Non-adaptive case (NA)

In the non-adaptive scenario, we assume that infected individuals do not change their behavior and continue to interact with others as if they were susceptible, with $b_I = b_S = b$. This means that there is no CT implemented and $\epsilon(s) = 0$ for all simplex sizes. Additionally, we set $\gamma_P/\mu = 1$. With these assumptions, we can derive the critical condition through the following equation:

$$-\mu + \lambda a \langle s(s-1) \rangle = 0$$

In this case, we obtain an explicit form for the epidemic threshold λ_C and reproduces the results obtained previously in [87]:

$$\lambda_C^{NA} = \frac{\mu}{a\langle s(s-1)\rangle} = \frac{\mu}{\overline{n}} \tag{A.65}$$

For only pairwise interactions, this results, reproduces the results obtained in [67, 68, 89].

A.4.2 Isolation of only symptomatic nodes

Here we consider the scenario where only symptomatic individuals are isolated upon symptom onset, meaning that no contact tracing is implemented. As a result, $\epsilon(s) = 0$ for all s, and $b_I = 0$. Under these assumptions, we obtain the following equation for the critical condition

$$-\gamma_P \mu + \lambda \overline{n}(\delta\mu + (1-\delta)\gamma_P) = 0$$

So we obtain an explicit form for the epidemic threshold λ_C:

$$\lambda_C^{sympto} = \lambda_C^{NA} \frac{\frac{\gamma_P}{\mu}}{\delta + (1-\delta)\frac{\gamma_P}{\mu}} \tag{A.66}$$

We note that, for pairwise interactions $\Psi(s) = \delta(s-2)$, this results reproduces the results obtained in [68], and for $\gamma_P/\mu = 1$ it reproduces the NA case (Eq. (A.65)).

A.4.3 Homogeneous case

Here we consider the case in which the size of simplices is homogeneous, i.e. $\Psi(s) = \delta(s-\overline{s})$, with a constant probability for a simplex to be traced with CT, i.e. $\epsilon(s) = \epsilon$.

If we set $\overline{s} = 2$, meaning only pairwise interactions, we obtain a quadratic equation in λ for the critical condition.

$$\frac{\lambda^2}{\mu^2}\overline{n}^2\delta^2\epsilon + \frac{\lambda}{\mu}\overline{n}\left(\delta + (1-\delta-\delta\epsilon)\frac{\gamma_P}{\mu}\right) - \frac{\gamma_P}{\mu} = 0$$

The equation can be solved and we obtain the same results obtained in [68]:

$$\lambda_C^{\overline{s}=2} = \lambda_C^{NA}\frac{2\frac{\gamma_P}{\mu}}{\delta + (1-\delta-\epsilon\delta)\frac{\gamma_P}{\mu} + \sqrt{(\delta + (1-\delta-\epsilon\delta)\frac{\gamma_P}{\mu})^2 + 4\delta^2\epsilon\frac{\gamma_P}{\mu}}} \tag{A.67}$$

In the case with $\overline{s} \to \infty$, it means to consider only the activation of simplices in wich all node partecipate, we obtain a linear equation in λ:

$$\lambda[\delta\mu\overline{n}(\gamma_P + \mu) + \overline{n}\gamma_P(1-\delta)(\mu + \gamma_P(1-\epsilon))] - \gamma_P\mu(\mu + \gamma_P) = 0$$

The equation can be solved and we obtain:

$$\lambda_C^{\overline{s}\to\infty} = \lambda_C^{NA}\frac{\frac{\gamma_P}{\mu}(\gamma_P + \mu)}{\delta(\gamma_P + \mu) + \gamma_P(1-\delta)(1 + \frac{\gamma_P}{\mu}(1-\epsilon))} \tag{A.68}$$

The maximum epidemic threshold λ_C^{max} is obtained when all simplices are considered and all individuals are traced at their infection. This is achieved when $\overline{s} \to \infty$ and $\epsilon = 1$. In this scenario, both asymptomatic individuals infected by A or A_T and those infected by P are traced through sideward and forward CT respectively. However, the epidemic threshold is still finite as traced individuals are isolated with a delay of τ_P and because of the presence of presymptomatic infections.

$$\lambda_C^{max} = \lambda_C^{NA}\frac{\frac{\gamma_P}{\mu}(\gamma_P + \mu)}{\gamma_P + \delta\mu} \tag{A.69}$$

Bibliography

[1] L. Alessandretti, U. Aslak, and S. Lehmann. The scales of human mobility. *Nature*, 587(7834):402–7, 2020.

[2] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann, and A. Baronchelli. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–91, 2018.

[3] L. Alexander, S. Jiang, M. Murga, and M. C. González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–50, 2015.

[4] V. Alfano and S. Ercolano. The efficacy of lockdown against covid-19: A cross-country panel analysis. *Applied Health Economics and Health Policy*, 18:509–517, 2020.

[5] J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. *OpenStreetMap in GIScience: Experiences, Research, and Applications*. Springer, Cham, CH, 2015.

[6] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–86, 2003.

[7] U. Aslak and L. Alessandretti. Infostop: Scalable stop-location detection in multi-user mobility data, 2020. arXiv preprint, arXiv:2003.14370.

[8] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[9] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[10] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[11] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.*, 874:1–92, 2020.

[12] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini. Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(05):05001, 2010.

[13] I. Biazzo, B. Monechi, and V. Loreto. General scores for accessibility and inequality measures in urban areas. *Royal Society Open Science*, 6(8):190979, 2019.

[14] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–6, 2015.

[15] G. Bonaccorsi. Economic and social consequences of human mobility restrictions under covid-19. In *Proceedings of the National Academy of Sciences*, volume 117, page 15530–15535, 2020.

[16] W. J. Bradshaw, E. C. Alley, J. H. Huggins, A. L. Lloyd, and K. M. Esvelt. Bidirectional contact tracing could dramatically improve covid-19 control. *Nat. Commun.*, 12(1):232, Jan 2021.

[17] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[18] G. Cencetti, F. Battiston, B. Lepri, and M. Karsai. Temporal properties of higher-order interactions in social networks. *Sci. Rep.*, 11(1):7028, Mar 2021.

[19] G. Chen and D. Kotz. *A Case Study of Four Location Traces*. https://digitalcommons.dartmouth.edu/cs_tr/246/, 2004.

[20] Clemente, L.-O. Riccardo, T. Miguel, X. Matias, V. Sharon, Bapu, and M. C. González. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature Communications*, 9(1):1–8, 2018.

[21] S. M. Cobb. Harvard to track affiliates' wi-fi signals as part of contact tracing pilot. *The Harvard Crimson*, 2020. 08-02-2020.

[22] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLOS Medicine*, 4(1):1–16, 01 2007.

[23] Council of the European Union. *Regulation on Privacy and Electronic Communications*. https://data.consilium.europa.eu/doc/document/ST-5008-2021-INIT/en/pdf, 2018.

[24] N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, C. C.-. W. Group, C.-. G. U. C.-U. Consortium, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, and W. J. Edmunds. Estimated transmissibility and impact of sars-cov-2 lineage b.1.1.7 in england. *Science*, 372(6538), 2021.

[25] P. Deville, C. Song, N. Eagle, V. D. Blondel, A.-L. Barabási, and D. Wang. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, 113(26):7047–7052, 2016.

[26] L. Di Domenico, G. Pullano, C. E. Sabbatini, P.-Y. Boëlle, and V. Colizza. Impact of lockdown on covid-19 epidemic in île-de-france and possible exit strategies. *BMC Med.*, 18(1):240, Jul 2020.

[27] J. Dibbelt, T. Pajor, B. Strasser, and D. Wagner. Connection scan algorithm. *ACM J. Exp. Algorithmics*, 23, oct 2018.

[28] L. Domenico, G. Pullano, C. Sabbatini, P.-Y. Boelle, and V. Colizza. Impact of lockdown on covid-19 epidemic in ile-de-france and possible exit strategies. *BMC Medicine*, 18:240, 2020.

[29] X. Dong, A. J. Morales, E. Jahani, E. Moro, B. Lepri, B. Bozkaya, C. Sarraute, Y. Bar-Yam, and A. Pentland. Segregated interactions in urban and online space. *EPJ Data Science*, 9(1):20, 2020.

[30] L. Downey, A. Fonzone, G. Fountas, and T. Semple. The impact of covid-19 on future public transport use in scotland. *Transportation Research Part A: Policy and Practice*, 163:338–352, 2022.

[31] K. D'Silva, A. Noulas, M. Musolesi, C. Mascolo, and M. Sklar. Predicting the temporal activity patterns of new venues. *EPJ Data Science*, 7:1–17, 2018.

[32] A. Endo, Q. J. Leclerc, G. M. Knight, G. F. Medley, K. E. Atkins, S. Funk, A. J. Kucharski, et al. Implication of backward contact tracing in the presence of overdispersed transmission in covid-19 outbreak, 2020.

[33] European Center for Disease Prevention and Control. Mobile applications in support of contact tracing for covid-19. https://www.ecdc.europa.eu/sites/default/files/documents/covid-19-mobile-applications-contact-tracing.pdf, 2021. Accessed on: 08-14-2021.

[34] R. Fernández Pozo, M. R. Wilby, J. J. Vinagre Díaz, and A. B. Rodríguez González. Data-driven analysis of the impact of covid-19 on madrid's public transport during each phase of the pandemic. *Cities*, 127:103723, 2022.

[35] C. Fraser, S. Riley, R. Anderson, and N. Ferguson. Factors that make an infectious disease outbreak controllable. In *Proceedings of the National Academy of Sciences*, volume 101, page 6146–6151, 2004.

[36] S. Funk, M. Salathe, and V. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of The Royal Society Interface*, 7:1247–1256, 2010.

[37] R. Gallotti, A. Bazzani, and S. Rambaldi. Erratum: 'towards a statistical physics of human mobility'. *IJMPC*, 24(2):1392001, 2013.

[38] Q. Ge and D. Fukuda. Updating origin–destination matrices with aggregated data of gps traces. *Transportation Research Part C: Emerging Technologies*, 69:291–312, 2016.

[39] S. Goh, K. Lee, J. S. Park, and M. Y. Choi. Modification of the gravity model and application to the metropolitan seoul subway system. *Phys. Rev. E*, 86:026102, Aug 2012.

[40] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–82, 2008.

[41] J. González-Cabañas, A. Cuevas, R. Cuevas, and M. Maier. Digital contact tracing: Large-scale geolocation data as an alternative to bluetooth-based apps failure. *Electronics*, 10(9), 2021.

[42] B. Gramsch, C. A. Guevara, M. Munizaga, D. Schwartz, and A. Tirachini. The effect of dynamic lockdowns on public transport demand in times of covid-19: Evidence from smartcard data. *Transport Policy*, 126:136–150, 2022.

[43] T. Gross and B. Blasius. Adaptive coevolutionary networks: a review. *Journal of The Royal Society Interface*, 5:259–271, 2008.

[44] T. Gross and H. Sayama. *Adaptive networks*. Springer, 2009.

[45] A. Guizzo, A. Vezzani, A. Barontini, F. Russo, C. Valenti, M. Mamei, and R. Burioni. Simplicial temporal networks from wi-fi data in a university campus: The effects of restrictions on epidemic spreading. *Frontiers in Physics*, 10, 2022.

[46] M. M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, oct 2008.

[47] T. Hale. A global panel database of pandemic policies (oxford covid-19 government response tracker. *Nature Human Behaviour*, 5:529–538, 2021.

[48] R. Hariharan and K. Toyama. Project lachesis: Parsing and modeling location histories. In M. Egenhofer, C. Freksa, and H. Miller, editors, *International Conference on Geographic Information Science*, page 106–24, Berlin/Heidelberg, DE, 2004. Springer.

[49] J. Hellewell. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8:488– 496, 2020.

[50] P. Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88:234, 2015.

[51] P. Holme and J. Saramaki. Temporal networks. *Physics Reports*, 519:97–125, 2012.

[52] P. Holme and J. Saramaki. *Temporal Networks*. Springer, Berlin Heidelberg, 2013.

[53] T. Hossmann, T. Spyropoulos, and F. Legendre. A complex network analysis of human mobility. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 876–881, 2011.

[54] S. Huber and C. Rust. Calculate travel time and distance with openstreetmap data using the open source routing machine (osrm). *Stata Journal*, 16(2):416–423, 2016.

[55] I. Iacopini, G. Petri, A. Barrat, and V. Latora. Simplicial models of social contagion. *Nat. Commun.*, 10(1):2485, Jun 2019.

[56] S. Jiang, J. Ferreira, and M. C. Gonzalez. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2):208–19, 2017.

[57] W.-S. Jung, F. Wang, and H. E. Stanley. Gravity model in the korean highway. *Europhysics Letters*, 81(4):48005, jan 2008.

[58] H. Kanasugi, Y. Sekimoto, M. Kurokawa, T. Watanabe, S. Muramatsu, and R. Shibasaki. Spatiotemporal route estimation consistent with human mobility using cellular network data. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). IEEE*, page 267–72, 2013.

[59] S. Kauffman. *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, Oxford, UK, 1993.

[60] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, pages 1–13, 2006.

[61] S. Kojaku, L. Hébert-Dufresne, E. Mones, S. Lehmann, and Y.-Y. Ahn. The effectiveness of backward contact tracing in networks. *Nat. Phys.*, 17(5):652–658, May 2021.

[62] M. Lenormand, A. Bassolas, and J. J. Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–69, 2016.

[63] M. Lenormand, T. Louail, M. Barthelemy, and J. J. Ramasco, 2017. Is spatial information in ICT data reliable? arXiv.

[64] M. Lenormand, T. Louail, O. G. Cantú-Ros, H. Miguel, A. Ricardo, J. Murillo, M. Barthelemy, S. Miguel, Maxi, and J. J. Ramasco. Influence of sociodemographic characteristics on human mobility. *Scientific Reports*, 5:10075, 2015.

[65] T. Louail, M. Lenormand, O. G. Cantú-Ros, H. Miguel, F.-M. Ricardo, R. Enrique, J. J., and M. Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4:5276, 2014.

[66] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo. A survey on deep learning for human mobility. *ACM Comput. Surv.*, 55(1), nov 2021.

[67] M. Mancastroppa, R. Burioni, V. Colizza, and A. Vezzani. Active and inactive quarantine in epidemic spreading on adaptive activity-driven networks. *Phys. Rev. E*, 102:020301, Aug 2020.

[68] M. Mancastroppa, C. Castellano, A. Vezzani, and R. Burioni. Stochastic sampling effects favor manual over digital contact tracing. *Nat. Commun.*, 12(1):1919, Mar 2021.

[69] M. Mancastroppa, A. Guizzo, C. Castellano, A. Vezzani, and R. Burioni. Sideward contact tracing and the control of epidemics in large gatherings. *J. R. Soc. Interface.*, 19:20220048, May 2022.

[70] M. Mancastroppa, A. Vezzani, M. A. Muñoz, and R. Burioni. Burstiness in activity-driven networks and the epidemic threshold. *J. Stat. Mech. Theory Exp.*, 2019(5):053502, may 2019.

[71] A. D. Marra, L. Sun, and F. Corman. The impact of covid-19 pandemic on public transport usage and route choice: Evidences from a long-term tracking study in urban area. *Transport Policy*, 116:258–268, 2022.

[72] H. Martin, N. Wiedemann, D. J. Reck, and M. Raubal. Graph-based mobility profiling. *Computers, Environment and Urban Systems*, 100:101910, 2023.

[73] N. Masuda and R. Lambiotte. *A Guide to Temporal Networks*. World Scientific, Europe, 2016.

[74] P. Miller, A. G. de Barros, L. Kattan, and S. Wirasinghe. Public transportation and sustainability: A review. *KSCE Journal of Civil Engineering*, 20(3):1076, 2016.

[75] G. Miritello, R. Lara, M. Cebrian, and E. Moro. Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3(1):1–7, 2013.

[76] E. Moro, A. Pentland, D. Calacci, and X. Dong. Atlas of inequality, 2019.

[77] P. Munoz and B. Cohen. Sharing cities and sustainable consumption and production: Towards an integrated framework. *Journal of Cleaner Production*, 134, 07 2015.

[78] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[79] M. Newman, A.-L. Barabasi, and D. Watts. *The structure and dynamics of networks*. Princeton University Press, 2006.

[80] Y. Nie, H.-M. Zhang, and W. Recker. Inferring origin–destination trip matrices with a decoupled gls path flow estimator. *Transportation Research Part B: Methodological*, 39(6):497–518, 2005.

[81] L. E. Olmos, S. Çolak, S. Shafiei, M. Saberi, and M. C. González. Macroscopic dynamics and the collapse of urban traffic. *Proceedings of the National Academy of Sciences*, 115(50):12654–61, 2018.

[82] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini. Scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data, 2019. arXiv.

[83] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(1):1–8, 2015.

[84] L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Dat Science and Analytics*, 2(1–2):75–92, 2016.

[85] R. Pastor-Satorras, C. Castellano, P. Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87:925–979, 2015.

[86] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Sci. Rep.*, 2(1):469, Jun 2012.

[87] G. Petri and A. Barrat. Simplicial activity driven model. *Phys. Rev. Lett.*, 121:228301, Nov 2018.

[88] F. Pinotti. Tracing and analysis of 288 early sars-cov-2 infections outside china: A modeling study. *PLOS Medicine*, 17:1–13, 2020.

[89] I. Pozzana, K. Sun, and N. Perra. Epidemic spreading on activity-driven networks with attractiveness. *Phys. Rev. E*, 96:042310, Oct 2017.

[90] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–23, 2008.

[91] H. D. Rozenfeld, D. Rybski, J. S. Andrade, M. Batty, H. E. Stanley, and H. A. Makse. Laws of population growth. *Proceedings of the National Academy of Sciences*, 105(48):18702–18707, 2008.

[92] M. Ruiz, J. M. Segui-Pons, and J. Mateu-LLadó. Improving bus service levels and social equity through bus frequency modelling. *Journal of Transport Geography*, 58:220–233, 2017.

[93] R. J. Sampson. *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press, Chicago, IL, USA, 2012.

[94] P. Sapiezynski, A. Stopczynski, D. K. Wind, J. Leskovec, and S. Lehmann. Inferring person-to-person proximity using wifi signals. *Proc. ACM Interactive, Mobile Wearable Ubiquitous Technologies*, 1(2), 2017.

[95] J. Saramäki, E. A. Leicht, E. López, S. G. Roberts, F. Reed-Tsochas, and R. I. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–7, 2014.

[96] H. Sayama. Modeling complex systems with adaptive networks. *Computers & Mathematics with Applications*, 65:1645–1664, 2013.

[97] V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *Proc. Natl. Acad. Sci. U.S.A.*, 113(36):9977–9982, 2016.

[98] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[99] R. Sonabend, L. K. Whittles, N. Imai, E. S. Knock, P. N. Perez-Guzman, et al. Evaluating the roadmap out of lockdown: modelling step 4 of the roadmap in the context of b.1.617.2. https://www.gov.uk/government/publications/imperial-college-london-evaluating-the-roadmap-out-of-lockdown-modelling-step-4-of-the-roadmap-in-the-context-of-b16172-delta-9-june-2021, 2021. accessed on: 06-29-2021.

[100] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–21, 2010.

[101] A. Tirachini and O. Cats. Covid-19 and public transportation: Current assessment, prospects, and research needs. *Journal of Public Transportation*, 22(1):1–21, 2020.

[102] M. Tizzani, S. Lenti, E. Ubaldi, A. Vezzani, C. Castellano, and R. Burioni. Epidemic spreading and aging in temporal networks with memory. *Phys. Rev. E*, 98:062315, Dec 2018.

[103] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162–77, 2015.

[104] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, page 1–8, 2012.

[105] E. Ubaldi et al. Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation. *Sci. Rep.*, 6(1):35724, Oct 2016.

[106] University of Parma. Coronavirus: all information for the university community upyeard in real time. https://www.unipr.it/coronavirus, 2020. accessed on: 07-17-2021.

[107] C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451, 2006.

[108] R. Vickerman. Will covid-19 put the public back in public transport? a uk perspective. *Transport Policy*, 103:95–102, 2021.

[109] E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, et al. Assessing transmissibility of sars-cov-2 lineage b.1.1.7 in england. *Nature*, 593(7858):266–269, May 2021.

[110] B. Wang, L. Cao, H. Suzuki, and K. Aihara. Safety-information-driven human mobility patterns with metapopulation epidemic dynamics. *Scientific Reports*, 2(1):887, Nov 2012.

[111] World Health Organization. Report of the who-china joint mission on coronavirus disease 2019 (covid-19). https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf, 2020. accessed on: 06-29-2021.

[112] World Health Organization. Tracking sars-cov-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/, 2021. accessed on: 06-29-2021.

[113] A. J. Yeganeh, R. P. Hall, A. R. Pearce, and S. Hankey. A social equity analysis of the u.s. public transportation system based on job accessibility. *Journal of Transport and Land Use*, 11(1):1039–1056, 2018.

[114] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 186–194, New York, NY, USA, 2012. Association for Computing Machinery.

[115] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, page 312–21, 2008.

[116] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, page 89–98, New York, NY, USA, 2011. Association for Computing Machinery.

[117] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, page 791–800, 2009.

[118] M. Zhou, M. Ma, Y. Zhang, K. SuiA, D. Pei, and T. Moscibroda. Edum: Classroom education measurements via large-scale wifi networks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 316–327, New York, NY, USA, 2016. Association for Computing Machinery.

[119] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González. Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities. *Transportation Research Record*, 2526(1):126–35, 2015.