Mathematical models of digital epidemiology in social networks

by

David Martín-Corral Calvo

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematical Engineering

Universidad Carlos III de Madrid

Tutor: Esteban Moro Egido

December 2023

"The creature that wins against its environment destroys itself."

– Gregory Bateson[1]

[1]Gregory Bateson. English anthropologist, social scientist, linguist, visual anthropologist, semiotician, and cyberneticist. Quote extracted from the book 'Steps to an Ecology of Mind'.

To my parents, Maite and Pepe, for giving life back to my soul, experiences and an education based on effort, learning, freedom and love to nature. To my sisters, Maite and Clara, for dealing with my internal conflicts. To Esteban Moro for believing in me, encouraging me to give my full potential, making me grow as a person and allowing me to learn at the highest level with the best, among them Manuel García-Herranz, Manuel Cebrián, Nick Obradovich, Alberto Aleta, Yamir Moreno, Alex Vespignani and Alex Pentland. To all my family, friends and acquaintances who express emotional detachment and unconditional love towards me. To nature for working the miracle of my existence and the rest of the living beings in the universe. We are nobody without the whole.

A mis padres, Maite y Pepe, por dar de nuevo la vida a mi alma, experiencias y una educación basada en el esfuerzo, el aprendizaje, la libertad y el amor a la naturaleza. A mis hermanas, Maite y Clara, por soportar mis conflictos internos. A Esteban Moro por creer en mí, alentarme a dar todo mi potencial, hacerme crecer como persona y permitirme aprender al más alto nivel con los mejores, entre ellos Manuel García-Herranz, Manuel Cebrián, Nick Obradovich, Alberto Aleta, Yamir Moreno, Alex Vespignani y Alex Pentland. A todos mis familiares, amigos y conocidos que expresan desapego emocional y amor incondicional hacia mí. A la naturaleza por obrar el milagro de mi existencia y del resto de seres vivos en el universo. No somos nadie sin el todo.

Abstract

HUMAN behavior and epidemics are intricately interconnected and this PhD thesis aims to enhance current mathematical epidemiological models by integrating human behavior through digital traces, leading to a more comprehensive understanding and prediction of viral epidemic spread. It addresses the limitations of existing models in capturing dynamic human behavior and proposes a novel approach that leverages alternative data sources, including social media and mobility data, as crucial variables in epidemic modeling.

Firstly, the thesis begins by providing an overview of communicable diseases, their interactions with human behavior, and the potential threats posed by human behaviors in the emergence of pandemics. It delves into the history of epidemiology, emphasizing the role of social networks in comprehending information dissemination and interventions targeting human behavior. Additionally, the thesis reviews the current state of computational and digital epidemiology, exploring novel data streams and advanced mathematical models for better understanding and prevention of epidemics.

Subsequently, the thesis addresses the limitations of existing Early warning epidemiological systems (EWES) based on official data and presents a novel approach that utilizes social media data to indirectly observe human behaviors and detect early outbreaks of Influenza-like Illness (ILI) using highly connected individuals. Through statistical machine learning models, behavioral insights are extracted from millions of Twitter posts, identifying highly connected nodes as effective sensors for early warnings. This approach could significantly improve the detection of ILI and COVID-19 outbreaks at scale while respecting privacy considerations.

Moreover, the thesis advances mathematical epidemiological modeling by incorporating real-world mobility data to construct social contact matrices, capturing

the intricate patterns of human contact. By utilizing GPS data, a data-driven approach is presented to simulate human mobility patterns and social interactions on a larger scale. The methodology is applied to understand the effects of different lockdown strategies during the second COVID-19 wave and to investigate the dynamics of infections and the impact of control measures in a metropolitan area like Boston.

In addition, the thesis extends the previous methodology by introducing a temporal component and constructing daily-level contact matrices, enabling the observation of feedback loops between social behaviors, infections, and the impact of Non-Pharmaceutical Interventions (NPIs) over time. This approach facilitates a granular understanding of viral spread, its spatial and temporal characteristics, detection of Super-spreading events (SSE) and evaluates the effectiveness of NPIs in controlling COVID-19 in various metropolitan areas across the United States.

In conclusion, this thesis presents the findings and discusses their implications for epidemic modeling. It also identifies future research directions derived from this work, emphasizing the potential for further advancements in modeling epidemics by incorporating human behavior through digital traces.

Resumen

La conducta humana y las epidemias están intrincadamente interconectadas y esta tesis doctoral tiene como objetivo mejorar los modelos epidemiológicos matemáticos actuales mediante la integración de la conducta humana a través de rastros digitales, ayudando a una comprensión y predicción mejor de la propagación de epidemias virales. Se abordab las limitaciones de los modelos existentes en la captura de la conducta humana y propone un enfoque novedoso que aprovecha fuentes de datos alternativas, como las redes sociales y los datos de movilidad, como variables cruciales en la modelización de epidemias.

En primer lugar, la tesis comienza proporcionando una visión general de las enfermedades transmisibles, sus interacciones con la conducta humana y las amenazas potenciales que plantean los comportamientos humanos en la aparición de pandemias. Se adentra en la historia de la epidemiología, haciendo hincapié en el papel de las estructuras sociales en la comprensión de la difusión de información y las intervenciones dirigidas a la conducta humana. Además, la tesis revisa el estado actual de la epidemiología computacional y digital, explorando nuevas fuentes de datos y modelos matemáticos avanzados para un mejor entendimiento y prevención de las epidemias.

Seguidamente, la tesis aborda las limitaciones de los sistemas de alerta temprana epidemiológica (EWES) existentes basados en datos oficiales y presenta un enfoque novedoso que utiliza datos de redes sociales para observar indirectamente los comportamientos humanos y detectar brotes tempranos de la gripe (ILI) utilizando sensores altamente conectados socialmente. A través de modelos estadísticos de aprendizaje automático, se extraen percepciones conductuales de millones de publicaciones de Twitter, identificando usuarios altamente conectados como sensores efectivos para alertas tempranas. Este enfoque podría mejorar significativamente la

detección de brotes de ILI y COVID-19 a gran escala, al tiempo que respeta las consideraciones de privacidad.

Además, la tesis avanza en la modelización epidemiológica matemática mediante la incorporación de datos reales de movilidad para construir matrices de contactos sociales, capturando los patrones intrincados del contacto humano. Utilizando datos de GPS, se presenta un enfoque basado en datos para simular los patrones de movilidad humana e interacciones sociales a gran escala. La metodología se aplica para comprender los efectos de diferentes estrategias de confinamiento durante la segunda ola de COVID-19 e investigar la dinámica de las infecciones y el impacto de las medidas de control en áreas metropolitanas, en este caso Boston.

Finalmente, la tesis amplía la metodología anterior al introducir un componente temporal y construir matrices de contacto a nivel diario, permitiendo la observación de los bucles de retroalimentación entre los comportamientos sociales, las infecciones y el impacto de las intervenciones no farmacéuticas (NPIs) a lo largo del tiempo. Este enfoque facilita una comprensión detallada de la propagación viral, sus características espacio-temporales, la detección de eventos de super dispersión (SSE) y evalúa la efectividad de las NPIs en el control de COVID-19 en diversas áreas metropolitanas de Estados Unidos.

En conclusión, esta tesis presenta los hallazgos y discute sus implicaciones para la modelización de epidemias. También identifica futuras direcciones de investigación derivadas de este trabajo, enfatizando el potencial de nuevos avances en la modelización de epidemias al incorporar la conducta humana a través de rastros digitales.

Published and Submitted Content

Below is a list of publications in which I have either authored or co-authored:

- **Social Media Sensors to Detect Early Warnings of Influenza at Scale** [1].

  I authored this article, where I took responsibility for designing the research, conducting the study, analyzing the data, and writing the manuscript. Our primary technical contribution to the research was the application of a social sensors methodology for outbreak detection. Significantly, we demonstrated, for the first time, the utilization of this methodology in the context of a biological-informational epidemic.

  This article is fully featured in chapter 2 of this thesis, where it showcases the findings and insights derived from our research.

  Access to the pre-print version of the article at Medrxiv.

  All material from this source included in the thesis is indicated through typographical means and an explicit reference.

- **Effectiveness of social distancing strategies for protecting a community from a pandemic with a data-driven contact network based on census and real-world mobility data** [2].

  I authored this technical report, which served as our initial pre-print publication focusing on social distancing and its impact on COVID-19 during the first lockdown. Our research built upon previous studies by utilizing mobility data as a proxy for human mobility. Notably, we were the first to incorporate real mobility data into modeling the spread of viral agents.

  Our contributions to this research encompassed various key aspects. We played a crucial role in designing the study, conducting the research, analyzing the

data, discussing the results, and subsequently writing the manuscript. This first technical report laid the foundation for the subsequent two articles, which are partially featured in chapter 3 and chapter 4 of this thesis.

Access to the technical report at MIT Connection Science Site.

All material from this source included in the thesis is indicated through typographical means and an explicit reference.

- **Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19** [3].

  I coauthored an article where I played a significant role in various aspects of the research. Specifically, I contributed to the research design, made partial contributions to the study, and actively participated in the analysis and discussion of the results. Additionally, I played a crucial role in editing the manuscript.

  Our main contributions centered around the analysis of mobility data, the construction of synthetic populations using both mobility and census data, and the development of temporal contact matrices. These components served as essential data inputs for agent-based model and the definition of Non-pharmaceutical strategies.

  A partial and modified version of this article is included in chapter 3 of this thesis, where it has been incorporated to provide relevant insights and support the overall argument.

  Access to the article at Nature Human Behaviour.

  All material from this source included in the thesis is indicated through typographical means and an explicit reference.

- **Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas** [4].

  I coauthored an article in which I made significant contributions to various aspects of the research process. Specifically, I played a role in designing the study, conducting the research, analyzing the extensive datasets of mobility data, and actively participating in the writing of the manuscript. Our primary

technical contribution focused on the analysis of large-scale mobility data, constructing synthetic populations, and developing temporal contact matrices from the mobility data, which served as crucial inputs for the agent-based model.

In chapter 4 of this thesis, certain sections of this article have been partially included, albeit with some modifications to fit the context and flow of the thesis.

Access to the article at PNAS.

All material from this source included in the thesis is indicated through typographical means and an explicit reference.

- **The unequal effects of the health-economy tradeoff during the COVID-19 pandemic** [5].

  I coauthored an article in which I played a role in designing the research and made a partial contribution to the study. Our primary technical contribution focused on analyzing mobility data, constructing synthetic populations, and generating temporal contact matrices from the mobility data. These components were utilized as data inputs for both an epidemiological agent-based model and an economic agent-based model.

  In the conclusions of my thesis, specifically in chapter 5, this article is cited as a potential future work, which has since become a reality and been accepted by the journal Nature Human Behaviour.

  Access to the pre-print version of the article at Arxiv.

  All material from this source included in the thesis is indicated through typographical means and an explicit reference.

Other Research Merits

Non-scientific publications

Here is a list of non-scientific publications related to this thesis:

- **More people, more fun: The scaling of events in cities** [6].

  I wrote an online article that explored the phenomenon of super-linearity in social events production and the diversity found in cities worldwide. This study was one of the initial projects I undertook as part of my thesis, aiming to gain a better understanding of fundamental concepts related to complex social systems and it is very related to chapter 3 and 4 of this thesis.

  Access to the online article at Medium.

- **Viviendo en una sociedad enferma** [7].

  I authored a book in Spanish that was published in May 2022, presenting my personal reflections on the current health status of our society, drawing from the knowledge gained during my thesis. In chapter 1, certain sections of the book have been partially modified and translated into English.

  Access to the book at libros.com.

Research impact

Our article featured in Nature Human Behavior [3] stands as one of the top 43 most cited papers affiliated with Universidad Carlos III de Madrid, amassing over 750 citations on Google Scholar. Moreover, here is a list of some online articles that have covered our work, raising awareness internationally and locally about our work:

- The effectiveness of social distancing strategies in the face of an epidemic, Medical Xpress, March 27, 2020.
- Measuring the Effectiveness of Coronavirus Social Distancing Policies, Wall Street Journal, April 10, 2020.
- More waves of virus cases could follow the first, health experts warn, The Dallas Morning News, April 10, 2020.
- La receta de un científico español para evitar los rebrotes y una segunda ola de COVID-19 , La Razón, August 9, 2020.
- Un científico toledano y su receta para evitar rebrotes de coronavirus, ABC, August 10, 2020.
- Un estudio muestra que los test sin rastreo eficaz de contactos son insuficientes para frenar la oleada, El País, August 12, 2020.
- Modeling the impact of testing, tracing, and quarantine, MIT News, September 8, 2020.
- Nociones cuánticas para entender a la sociedad actual, Tribuna de Toledo, December 13, 2022.

Contents

List of Acronyms

ABM	Agent-Based Model.
BMA	Boston Metropolitan Area.
CBSA	Core Based Statistical Areas.
CCPA	California Consumer Privacy Act.
CDC	Centers for Disease Control and Prevention.
CDRs	Call data records.
CSS	Computational social science.
EIOS	Epidemic Intelligence from Open Sources.
EWES	Early warning epidemiological systems.
GBD	Global Burden of Diseases.
GDP	Gross Domestic Product.
GDPR	General Data Protection Regulation.
GFT	Google Flu Trends.
GOARS	Global Outbreak Alert and Response Network.
ICU	Intensive care unit.
ILI	Influenza-like Illness.
IOA	Integrated Outbreak Analytics.
IPTC	International Press Telecommunications Council.
NCDs	Non-Communicable Diseases.
NLP	Natural language processing.
NPIs	Non-Pharmaceutical Interventions.
POI	Points of interest.
SEIR	Susceptible, exposed, infected and recovered.
SIR	Susceptible-infected-recovered.
SSE	Super-spreading events.

1

Introduction

"The epidemiological methods may be scientific, but its objectives are often thoroughly human."

– Alex Broadbent[1]

THE spread of viral agents is significantly influenced by human behavior, making it essential to develop improved epidemiological systems that can effectively understand and predict the spread of epidemics. To achieve this, it is crucial to incorporate human behavior by leveraging digital traces, which provide valuable insights into how individuals interact and perceive their own health status and react accordingly. By integrating these behavioral aspects, we can enhance the effectiveness of detection systems for future epidemics.

However, current mathematical epidemiological models predominantly utilize static and over simplified frameworks that inadequately capture the dynamic nature of human behavior. These models often overlook the availability of data sources that can capture the variability in people's movements, perceived symptoms, and responses to public health measures. Moreover, they commonly employ simplistic analytical approaches, assuming unrestricted interactions between individuals, or rely on static mobility data to approximate behavioral interactions across different groups and regions. Furthermore, these models often solely rely on official and hospital data, limiting their ability to capture real-time dynamics and the underlying contact network structure that drives virus transmission. As a result, these models

[1] Alex Broadbent. South African philosopher. Quote extracted from the epidemiological magazine 'The Epi Monitor'.

fail to provide timely and accurate insights into the trajectory of epidemics.

In this thesis, we present a novel approach that harnesses alternative data sources, including social media and GPS digital traces, to access real-time information on symptoms and social contacts for millions of individuals. By incorporating these data sources, we demonstrate how human behavior can be integrated as a crucial variable in the mathematical modeling of viral epidemics. This innovative methodology provides a more comprehensive understanding of epidemic dynamics and offers valuable insights for effective epidemic control and management. The main focus of our research revolves around the hypothesis that incorporating novel data sources that capture human behaviors can significantly enhance mathematical epidemiological models.

In this introduction, we begin by providing a comprehensive overview of the fundamental knowledge and conceptual framework required to grasp the significance of communicable diseases and their interactions with human behaviors. We delve into the ways in which our own behaviors as humans can pose a threat to society, potentially resulting in the emergence of new pandemics. Additionally, we outline our understanding of human behavior in the context of epidemics, shedding light on the factors that influence the transmission and impact of infectious diseases on human health. We emphasize the critical importance of studying these intricate relationships from both public health and economic perspectives.

Secondly, we briefly review the history of ancient and modern epidemiology. How social networks play a crucial role in epidemiology, helping us understand how information spreads within a complex contact network system. Some social-driven interventions aim to change human behaviors and interactions to reduce the spread of information. Third, we present the current state of the art in the computational and digital epidemiology field, covering everything from novel data streams to advanced mathematical epidemiological models that could help us understand, assess, and prevent future epidemics at scale. Furthermore, we present the research purpose and questions as guidelines for this thesis. Finally, at the end of this chapter, we provide a concise introduction to the subsequent chapters that comprise this thesis.

1.1 The recurrent problem between humans and viral epidemics

1.1.1 Humans and pathogens

Across the annals of history, humanity has confronted a multitude of viral epidemics, including very fearful plagues with millions of deaths, caused by pathogens, like viruses and bacteria, such as influenza, smallpox, measles, salmonella, malaria, yellow fever, typhus, and cholera. The first known epidemic dates back to 1,200 B.C. in Babylon, which may have been caused by the influenza virus [8]. Over time, many more influenza pandemics have occurred, with the virus exhibiting increased virulence due to genetic mutations [9]. Smallpox is another disease that has plagued humanity for centuries, with evidence of its existence dating back to the 3rd century B.C. in Egypt [10]. Smallpox is caused by the bacteria Yersinia pestis and has been responsible for numerous deadly plagues throughout history, claiming hundreds of millions of lives over thousands of years [11].

Over the past few decades, numerous new infectious diseases have emerged, including HIV infections [12], SARS [13], Lyme disease [14], Escherichia coli O157:H7 (E. coli) [15], hantavirus [16], dengue fever [17], West Nile virus [18], chikungunya virus [19], Ebola virus [20], Zika virus [21], and SARS-CoV-2 [22]. Additionally, there are diseases that have reemerged after experiencing a significant decline, like measles [23] or monkeypox [24]. The reemergence of diseases can be attributed to the emergence of new strains of known diseases or changes in human behaviour. Some examples of reemerging diseases include malaria, tuberculosis, cholera, pertussis, influenza, pneumococcal disease, and gonorrhoea. Globalisation and climate change have increased the likelihood of the emergence and spread of infectious diseases on a global scale. The COVID-19 pandemic has served as a warning of the potential threats that humanity may face in the future.

The severity of viral epidemics has been influenced by a diverse array of factors. Among these, several common factors emerge as crucial determinants shaping the spread and impact of these epidemics. These factors include the characteristics of the viral agent, its transmission dynamics, the susceptibility of a population to the viral agent, population density and mobility, the socioeconomic and cultural factors of the

population, Non-Pharmaceutical Interventions (NPIs), and finally, viral epidemics are influenced by global factors, including international travel, globalization of trade and commerce, connectivity, and coordination among countries for surveillance.

In addition to previous factors, our current civilization has influenced and transformed the environment, resulting in their destruction or loss of environmental equilibrium [25]. The impact has been so significant that there is a more than 95% probability that human activities in the last 50 years have altered the equilibrium of the planet, contributing to changes in ecosystems and all the sentient beings within them [26, 27].

The consequences of altering the equilibrium of our ecosystems are difficult to predict. However, changes in temperature and precipitation patterns are leading to a higher number of extreme weather events, such as droughts, floods, and extreme temperatures [28]. These events can have significant impacts on higher probabilities of global pandemics from zoonotic emergent pathogens [29].

Studies have shown that zoonotic risks are higher in forested tropical regions undergoing land-use changes, where wildlife biodiversity is high [30]. Changes in land use have also resulted in increased interactions between people, livestock, and wildlife reservoirs of zoonotic diseases [31]. In addition, a reduction of 1.4% in the Amazonian forest area has been estimated to increase the incidence of malaria by 1% [32]. Furthermore, Ebola outbreaks located on the fringes of the rainforest biome have been significantly associated with forest losses in the previous two years [33].

The combined impact of globalization and loss of equilibrium within ecosystems has only made matters worse, contributing to the rise in the distribution and prevalence of infectious diseases [34, 35], more severe influenza seasons [36, 37] and increase the risk of zoonotic pandemics such as COVID-19 [29]. An additional social threat is that pandemics also cause fear, political turmoil and the rise of authoritarianism [38].

Hence, there exists an urgent requirement for better mathematical models of epidemic spreading and improved Early warning epidemiological systems (EWES) to mitigate the risks posed by infectious diseases and gain deeper insights into the underlying factors driving viral outbreaks. By exploring the potential of human behavioral data in comprehending and mitigating the impact of viral epidemics, this thesis seeks to make significant contributions towards fostering more resilient

societies.

1.1.2 Health-related behaviours

Human health is determined by various factors, including biological, behavioral, sociocultural, economic, and ecological factors [39, 40]. For instance, several infectious diseases heavily rely on human behaviors for their transmission.

Sexually transmitted infections such as HIV/AIDS, syphilis, gonorrhea, and chlamydia are primarily spread through unsafe sexual practices, including unprotected sex and having multiple partners. Respiratory diseases like influenza, the common cold, and COVID-19 spread through respiratory droplets, and close contact in crowded spaces without proper hygiene practices increases the risk of transmission. Foodborne and waterborne illnesses such as salmonella, E. coli, norovirus, and hepatitis A result from consuming contaminated food or water due to inadequate hygiene during preparation and storage. Vector-borne diseases like malaria, dengue fever, Zika virus, and Lyme disease depend on human behavior in terms of exposure to vector habitats and the use of preventive measures such as mosquito nets and tick bite prevention. Nosocomial infections acquired in healthcare settings, including healthcare-associated infections and antibiotic-resistant bacteria, are influenced by poor hand hygiene, insufficient sterilization of equipment, and inappropriate prescription of antibiotics among medical practitioners.

Gaining a comprehensive understanding of the role of human behavior, such as mobility, sexual practices, and hygiene, in the transmission of infectious diseases is paramount. This understanding is essential for comprehending the patterns and mechanisms of pathogen spread [41, 42], as well as for devising and implementing effective NPIs. However, most of mathematical models to understand the transmission of infectious diseases are based on simple mathematical equations that have strong assumptions about how humans behave in a given population. For example, the fact that transmission is equally probable or homogeneous across different groups in the same geographical settings [43, 44].

These models are referred to as compartmental models, categorizing the population into different compartments. One fundamental example is the SIR model, consisting of three key compartments, S, I, and R, representing Susceptible, Infectious, and Recovered individuals, respectively. The model operates as follows:

- S represents the count of susceptible individuals. When a susceptible individual comes into contact with an infectious individual, they contract the disease and transition to the infectious compartment (I).

- I represents the number of infectious individuals. These are people who have been infected and can transmit the disease to susceptible individuals.

- R represents the count of removed individuals. This includes those who have recovered from the disease and moved to the recovered compartment, as well as those who have died. The assumption is made that the number of deaths is negligible in comparison to the total population.

The SIR model is built upon strong assumptions, assuming that every individual interacts with the entire population. The differential equations that dictate the dynamics of this nonlinear viral process are as follows:

$$\frac{dS}{dt} = -\frac{\beta IS}{N} \tag{1.1}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I \tag{1.2}$$

$$\frac{dR}{dt} = \gamma I \tag{1.3}$$

Equation 1.1 represents the rate of change of susceptible individuals over time. It depends on the transmission rate (β), which signifies how easily the disease spreads from infectious to susceptible individuals, the number of infectious individuals (I) and the number of susceptible individuals (S), divided by the constancy of the population (N). As infectious individuals come into contact with susceptible ones, the disease spreads, reducing the number of susceptible individuals.

Equation 1.2 describes the rate of change of infectious individuals. The first term represents new infections as susceptible individuals become infected (βIS), divided by the population size (N), and the second term represents the rate at which infected individuals recover or die (γI), indicating how quickly infected individuals recover or exit the infectious state.

Equation 1.3 governs the rate of change of recovered individuals. It is solely influenced by the recovery rate (γ) multiplied by the number of infectious individuals (I) transitioning to the recovered state.

Compartmental models, such as the SIR model, provide valuable average insights

into disease transmission within a population. However, their assumptions regarding human behaviors remain notably simplistic. For instance, they assume that the rate of contact between individuals and the probability of transmission are constant. This simplification is primarily due to the lack of fine-grain real-world behavioral data available at the time these models were developed. They also assume that all individuals in the susceptible state are exposed to those in the infected compartment (well-mixed hypothesis). However, this exposure is, again, mediated by how humans encounter and interact in real time, creating a heterogeneous and dynamic mixing that depends on human behavior.

In the current era, characterized by exponential growth in user-generated data due to the digital revolution, an unprecedented opportunity has emerged to observe and quantify human behavior on a large scale. The rise of web searches, micro-blogging posts, and geolocated data sources has enabled the measurement of human behavior and its impacts on public health. In this thesis, we harness these novel data streams to quantify human behavior, categorizing our observation methods into two groups: indirect and direct. For instance, posting content on a social network about personal health symptoms carries implicit information about one's health status. Similarly, individuals socializing in a restaurant, along with their digital footprints of mobility patterns before and after the encounter, provide insights into the potential for viral disease transmission.

In chapter 2, we make indirect observations of human behavior trough social media fingerprints to build proxies from ILI-related posts that highly correlate with health-related behaviors and health effects. In chapter 3 and chapter 4, we directly observe human behavior through real-world mobility data from GPS signals to build contact matrices, derive the effects of human activity on the overall health of the population and the healthcare system.

1.1.3 Burden of infectious diseases

Top infectious diseases in order of disability-adjusted life-years per 100,000 inhabitants, according to the 2019 Global Burden of Diseases, are neonatal disorders, lower respiratory infections, diarrheal diseases, HIV/AIDS, Tuberculosis and Malaria. Infectious diseases have experienced a decrease of 55.94% between 1990 and 2019, with 131.74 deaths per 100,000 in 2019 [45]. However, six infectious diseases

were among the top ten causes of burden in children [45] and the COVID-19 pandemic has imposed a significant global burden, leading to substantial health-related consequences and a loss in the value of life, particularly within developed economies [46]. Furthermore, the interaction between non-communicable diseases (Non-Communicable Diseases (NCDs)) and infectious illnesses, which are not transmitted directly from person to person, is gaining recognition as a significant factor amplifying the burden. The role of NCDs in exacerbating viral epidemics is substantial. These diseases heighten individual susceptibility, contribute to the development of comorbidities, trigger immune dysfunctions, and give rise to syndemics. This, in turn, widens the pool of individuals vulnerable to infections and exacerbates the strain on healthcare systems, as was evident during the COVID-19 pandemic.

In terms of economic burden, infectious diseases' economic impact is hard to measure. It affects so broadly to the economy that it is not fully possible to measure the impact. However, the COVID-19 epidemic has affected so broadly and globally that we have been able to measure its economic impact worldwide, by country and by industry. For instance, let us see some examples of financial impacts due to zoonotic infectious disease events beyond the public health sector. From 2013 to 2014, Gross Domestic Product (GDP) growth in Liberia decreased from 8.7% to 0.7%, due to Ebola and lowering commodity prices, and GDP growth in Sierra Leone decreased from 5.3% to 0.8% [47]. GDP growth in Guinea in 2015, predicted at 4%, fell to 0.1%. [47]. In Mexico, the H1N1 outbreak in 2009 affected tourism by US $ 2.8 billion [48]. In Somalia, RVF 1998–2002 outbreak impacted livestock export losses by US$ 435 million [49]. In Malaysia during the 1998–1999 Nipah epidemic, there was a loss of tax revenue of about US $105 million [50]. In Ghana, Liberia, and Sierra Leone, the 2013–2015 Ebola epidemic created a loss of investors' confidence of nearly US $600 million [47]. A final example is the 2003 SARS global epidemic, which created to airline companies a loss of more than US $7 billion+ [51].

However, in 2020 the coronavirus pandemic reached almost every country around the world, with a major impact on our economies, from governments, businesses and citizens. Let us see some examples of the early impact of a global pandemic like COVID-19. The FTSE dropped 14.3% in 2020 [52], its worst performance since 2008. Most of the countries entered into recession. For example, the International Monetary

Fund estimated that the global growth contraction for 2020 was nearly -3.5% [53], the worst since the Great Depression of the 1930s. Unemployment rates increased across major economies [53]. According to the accountancy giant EY, consumer behaviours changed significantly. Making that 67% customers were not willing to travel more than 5 kilometres for shopping [54]. It is a fact that a global pandemic has major impacts on our health and economies, and we are the only ones responsible for this situation. Without understanding the behavioral entanglement between health, jobs, NPIs and public health policies, it is impossible to devise better policies that minimize the impact of the disease while alleviating the effect in our society and economy of those policies.

1.2 Epidemiology and human health-related behaviours

Many definitions about epidemiology have been proposed, but there is a definition that captures the spirit and principles of this scientific field:

*Epidemiology is the **study** of the **distribution** and **determinants** of **health-related states or events** in **specified populations**, and the **application** of this study to the control of health problems* [55].

Based on the previous definition. Epidemiology is a scientific, systematic and data-driven discipline looking for unbiased collection, analysis and interpretation of data. Epidemiology is also concerned with the distribution, frequency and patterns of health states or events in a population by defining time, place, risk factors and individuals. Epidemiology also searches for determinants, the causes and effects of health-related events. This links with the main purpose of this thesis, to highlight the importance of human behaviours and their digital traces in epidemiology.

Nowadays epidemiology requires more interdisciplinary teams to provide the full potential of our current technological advances to move forward the boundaries of our knowledge and start building or updating epidemiological systems based on big data approaches. The main drivers are two. Firstly, the availability of vast amounts of behavioral and social data generated by our modern society provides a valuable resource for analysis. Secondly, recent advancements in data storage, processing, computing, and extraction techniques serve as essential tools for extracting insights from these novel data streams. Together, these factors enable public health

decision-makers to gain a deeper understanding of epidemics, facilitate explanation and prediction of their spread, and ultimately aid in the development of effective preventive measures.

1.2.1 Origins of epidemiology

Epidemiology's roots are nearly 2500 years old. Hippocrates, the Greek physician, is known as the father of medicine [56], was the first to attempt to explain disease events from a rational rather than a supernatural viewpoint, examining the relationship between the occurrence of a disease, the environment and host factors such as behaviours [57]. He was also the first to make the distinction between *"epidemic"* and *"endemic"* [58]. Diseases that are "visited upon" a population, are epidemic, contrary to those that "reside within" a population, which are endemic. However, the word epidemiology was not used to describe the study of disease epidemics until 1802 by the Spanish historian and physician Joaquín de Villalba y Guitarte in his book "Epidemiología Española" [58].

In the early years of epidemiology, pioneers such as John Graunt, Daniel Bernoulli, and William Farr made significant contributions to the field. Graunt, in 1662, conducted the first study quantifying birth, death, and disease patterns, noting disparities between genders, urban-rural differences, and seasonal variations [59]. Bernoulli, in 1766, developed the first mathematical model in epidemiology, linking susceptibility to endemic diseases with the force of infection and life expectancy [60]. Farr, often referred to as the modern father of medical statistics and epidemiological surveillance, systematically collected and analyzed mortality statistics in Britain, including studying smallpox epidemics and demonstrating the bell-shaped curve for disease outbreaks [61].

During the early 1900s, further advancements in mathematical epidemiology occurred. Ronald Ross discovered the transmission of malaria through mosquitoes [62], while Anderson Gray McKendrick and William Ogilvy Kermack developed the Kermack-McKendrick theory [63], a general epidemic model, constructed using ordinary differential equations described in equations 1.1, 1.2, and 1.3. This theory established the foundation for Susceptible-infected-recovered (SIR) models, which classify populations into susceptible, infected, and recovered individuals, as discussed earlier in subsection 1.1.2. These mathematical models

marked the beginning of modern mathematical epidemiology, although they relied on estimations and approximations due to limited granular behavioral health data at the time.

It is worth noting that these early models followed a top-down approach, assuming uniform behavior within the population due to the lack of detailed individual-level data. However, they set the stage for future advancements that would incorporate more nuanced and individual-specific factors to enhance the accuracy and granularity of epidemiological models.

After World War II, epidemiology started to focus on NCDs and saw an explosion in research of methods and theoretical aspects of epidemiology. These methods were used to identify links between health-related events and behaviours, environment and even attitudes. Two studies are worth mentioning that used epidemiological methods to chronic diseases like the linking between lung cancer and smoking [64], and the Framingham study, which characterized cardiovascular diseases and many others until our days [65]. For instance, a 2007 study used sophisticated social network methods to model obesity as an infectious disease. This study comprises data from over 32 years. Christakis and Fowler [66] were able to model the spread of obesity within households and the community, thanks to their longitudinal and very detailed data. In the late XX century, public health decision-makers applied epidemiological methods to eliminate smallpox outbreaks worldwide [67].

To tackle all these problems, the latest approach in epidemiology is the usage of behavioural data at scale from big data sources. Health systems and health organizations initiatives, like the Global Outbreak Alert and Response Network (GOARN) [68] from WHO that is composed of 250 technical institutions and networks globally and projects like the Integrated Outbreak Analytics (IOA) [69], the Epidemic Intelligence from Open Sources (EIOS) [70] and the Epi-Brain [71] that respond to acute public health events. This network is already moving in a double direction of incorporating early warnings from big data, social sciences techniques and behavioural data into epidemic response systems [72] to control outbreaks and public health emergencies across the globe and have a holistic understanding of outbreak dynamics.

In the following section, we are going to introduce the field of computational social science, which is the leading field in analysing human behaviours at scale

using social networks. As we have stated in this introduction, human health has a crucial component attached to individual and social behaviours. We believe in a bottom-up approach. We can incorporate behavioural data to advance epidemiological models and treat each individual within a social complex system as an independent agent contributing to the spreading information within a viral process. We still have the same hypothesis as Hippocrates, to understand the relationship between human behaviours, the environmental conditions and the human health. Despite we are using novel data streams and advanced computational models that Hippocrates never dreamt of.

1.2.2 Computational social science

The study of how our behaviors influence potential contacts and subsequent social interactions has changed dramatically with digital tools. Social behaviors encompass all interactions that occur among individuals, whether they are strangers, relatives, members of the same or opposite sex, or different generations. Together, these interactions form a complex social network that drives the social dynamics of the agents involved. This social network is comprised of nodes connected by dyadic ties that are based on homophily or affiliation, resulting in a sophisticated social structure that includes both offline and online interactions. During each interaction, energy and information - such as ideas, opinions, behaviors, viruses, or material transactions - are exchanged or transmitted.

The concept of social networks was first mentioned in the late 1890s in the early theories of social groups [73, 74]. In the 1900s, Simmel was the first to describe the nature of networks and their impact on social interactions in small groups [75]. During the 1930s, there was a surge in research and applications of social networks across multiple fields such as psychology, sociology, anthropology, and mathematics [76, 77], laying the groundwork for social network theory as the first social and psychological theories based on quantitative analysis of social interactions and community networks. It is worth mentioning the work of Jacob L. Moreno, who developed a systematic recording and analysis of social interactions in schools and companies, creating the first visualizations of social networks based on human interactions called sociograms [78]. Throughout this thesis, we will refer to sociograms as contact matrices.

During the 1960s and 1970s, the field of social network analysis began to shift its focus towards political and community networks, social movements, and more in-depth social network analysis [73]. Notably, Stanley Milgram conducted groundbreaking experiments that linked human behavior with social networks. His small-world experiment [79] showed that human social networks exhibit a small-world structure, where nodes have sparse connections that enable them to reach nearly every other node in the network with just a few steps. This property makes ties in small-world networks non-local, which has important implications for the spread of information over social networks. Therefore, network characterization is critical for understanding how information spreads over social networks.

The explosion of data and computational power during the 1990s and 2000s brought significant advancements in social network analysis and its applications. New mathematical models and methods emerged to analyze the vast amounts of online data, known as digital traces, generated by the internet's growth. These advancements enabled researchers to observe human behavior and social interactions at a macro-level scale. This period of research yielded important discoveries about the properties and dynamics of large-scale complex networks [80, 81] and provided insights into how they behave as macro-organisms, offering a better understanding of the social complexity of our universe, environment, and society.

In 2009, the interdisciplinary field of Computational social science (CSS) was officially established, based on the premise that we live our lives as agents in a network [82]. This new scientific field aimed to provide a solid framework for leveraging the capacity to collect and analyze data at scale to analyze our society [82]. Thanks to vast amounts of data, more computational power, new mathematical methods in network analysis, agent-based models, and statistical learning models, it has become possible to simulate, measure, understand, and predict individual and group behaviors like never before. The field has found applications in areas ranging from social inequality to the spread of infectious diseases, making it a highly relevant area of research.

Over the past decade, the field of CSS has witnessed an explosion of publications, thanks to new observational data on human behavior, experimental designs, and large-scale simulations that were previously impossible to research. This expansion of knowledge has greatly surpassed the expectations of the scientific community [83].

However, during this period, studies on how to influence human behavior, particularly in the areas of voting [84–86], emotional-personality [87, 88], and information virality [89, 90], have raised concerns about the ethical implications of CSS, data sovereignty, and the impact of socio-technical systems on individuals and society. In 2020, a paper by prominent scientists discussed the obstacles and opportunities facing CSS after a decade of progress. They emphasized the need for better data-sharing paradigms and improved research ethics to address legal and social implications [83].

CSS is a rapidly evolving field that has already shown promise in improving human health outcomes. However, several obstacles still impede its progress, such as data privacy concerns and the reluctance of private organizations to share data. In this thesis, we aim to contribute to the field by addressing some of these challenges and expanding the number of health applications aimed at preventing epidemics.

We recognize that two critical determinants of human health are human behaviors and the environment, and we focus on analyzing social networks from the perspective of CSS to gather information about the system and viral agents spreading within social environments. Our goal is to develop better mathematical methods for quantifying, explaining, forecasting, and defining public health policies that can maximize the health of populations.

1.2.3 The new data-driven behavioural epidemiology

The field of epidemiology has not been immune to the effects of the digital revolution that has taken place over the last few decades. There are two main reasons for this disruption. Firstly, the vast amounts of user-generated data that can now be stored at low costs, allowing epidemiologists to incorporate novel data streams such as search queries [91], microblogging [92–94], forums [95] and mobility data [96], which are collectively known as novel data streams. Secondly, new mathematical and computational models, such as statistical machine learning or agent-based models, have been developed and have proven to be effective in detecting outbreaks, simulating epidemic dynamics, evaluating the impact of travel on the transmission of global pandemics, monitoring the use of pharmaceuticals, and exploring links between higher temperatures and emotions, among many other use cases that we will delve into in the next section.

Two new branches of epidemiology, digital epidemiology and computational epidemiology, have emerged as a result of the aforementioned advances. Digital epidemiology, as described by Salathe et al. [97], focuses more on novel data streams, whereas computational epidemiology, as described by Marathe et al. [98], focuses more on the computational side. The main objective of both digital epidemiology and computational epidemiology is to explain and predict the patterns of diseases, health dynamics, and their causes, just like traditional epidemiology. However, what sets these new fields apart from traditional epidemiology is the data and methods used. They are simply new tools that have been developed to tackle the same old problems. The use of novel data streams and computer models helps to gain a better understanding of the spatiotemporal diffusion of diseases through populations [97, 98].

The COVID-19 pandemic has highlighted the significant potential of utilizing human behavioral data from digital sources within the field of epidemiology. This data can be effectively harnessed to comprehend and mitigate the spread of infections [41, 99]. In the subsequent sections, we will delve into several compelling examples that exemplify the practical application of this approach. While the discussion encompasses a broad range of aspects, our primary focus centers on three key components: the collection and analysis of human behavioral data, the integration of this data into models to enhance the description of human interactions, and the utilization of mathematical epidemiological models to simulate the spread of infections.

Digital traces as proxies of human health-related behaviors

The digital revolution has brought about spectacular advances in natural science, fueled by the vast amounts of user- and sensor-generated data available. In recent decades, the widespread adoption of communication and information technologies has resulted in nearly every person on Earth owning a mobile phone that generates data. As a result, almost everything we do and say leaves behind a digital trace that can be stored and analyzed. This abundance of data has led to a growing recognition of the importance of behavioral data in modeling tools for the healthcare system [100–103], as human behavior is a significant factor in determining human health.

Data generated from novel data streams contain epidemiologically relevant information about human behaviors, beliefs, and health status, which can be used to extract meaningful information and potentially understand and prevent disease dynamics at scale. These novel data streams come in various forms and sizes, and processing them can be a challenge. However, some potential use cases have already been proven, such as observing spatiotemporal human behaviors during an outbreak [104], detecting unusual respiratory diseases in remote areas [105], estimating near real-time influenza cases [91], and assessing the population's response to a vaccination campaign [106]. These early promising applications led to the founding of the field of digital epidemiology [97]. Furthermore, several articles highlight the need for behavioral data [100–103] to improve modeling tools in the healthcare system, as human behavior plays a significant role in human health.

The earliest and most famous example of a data-driven epidemiological model is the Google Flu Trends (GFT). It leveraged symptomatic user-generated search queries to predict and track Influenza-like Illness (ILI) [91]. However, this novel concept faced several problems, as discussed in articles such as [107–109]. The main issue was the private ownership of the underlying data used by GFT, which prevented independent replication and assessment of the epidemiological models [110]. This highlighted the need for open access to data to enable transparency and reproducibility in data-driven epidemiology research.

Indirect observation of the human health-related behaviours

When direct observation data is not available, researchers must look for indirect observations of the phenomena they aim to measure. Non-traditional data sources, such as web search queries and visits [91, 111–114], weather data [115], social media [92–94], or the monitoring of multiple digital traces simultaneously [116], have proven to be complementary and sometimes advantageous to traditional health monitoring systems. Online user activity offers benefits such as a wider spatial and demographic reach, as well as the ability to monitor populations with limited access to health services [114].

However, these data streams usually require preprocessing to make them ready for analysis since they deal with unstructured data. The preprocessing stages can range from simple keyword searching based on dictionaries with target keywords that are

signals of the phenomena in question to more advanced Natural language processing (NLP) and machine learning techniques to extract more detailed information. We will delve into the specifics in the following section.

As we stated above, the earliest and most famous example of a data-driven epidemiological model is GFT, which leveraged symptomatic user-generated search queries to predict and track ILI [91]. Social media traces from micro-blogging platforms, such as Twitter, have also proven to be good indirect observations of flu epidemics [92, 117, 118]. Our proposed EWES in chapter 2 utilizes advanced NLP and machine learning techniques to process indirectly observed unstructured data from the micro-blogging site Twitter and transform it into validated ILI-related behavioural posts from users.

Direct observation of human health-related behaviours

Along with the emergence of GFT, several other online-based epidemiological applications were developed in different countries, such as "Flutracking" [119], "Influweb" [120], "Flu Near ou" [95], "Influenzanet" [121], and "Grippenet" [122]. These applications were the first to use internet-based participatory syndromic surveillance of ILI developed by the academic sector. Syndromic surveillance involves gathering medical signs and symptoms of a syndrome, as well as individuals' perceived health. The purpose of these applications was to collect perceived health information from online users actively. They tested the feasibility of such systems and their ability to detect risk factors based on web-based data mining, which had a revolutionary impact in many areas. These applications could greatly impact how we monitor global health outcomes and human behaviours [97].

Furthermore, mobile phone data provides a significant and novel data stream that offers direct and objective observations of human behavior, including mobility and social interactions in the physical world [96]. This data can be collected in the form of call data records (CDRs) or global positioning system (GPS) data. CDRs contain information about the location of the mobile tower used to connect to the mobile network, while GPS data contains more detailed information about an individual's position based on latitude and longitude coordinates on the globe [123]. Both forms of mobile data provide spatiotemporal information about an individual's location. However, CDRs lack the granularity of GPS data, which can pinpoint an individual's

exact position in time, with an error of just a few meters.

Numerous use cases have demonstrated the reliability of mobile phone data in creating more detailed and objective models of human mobility [96]. Additionally, mobile phone data has been used to monitor and model outbreaks of infectious diseases [124–127]. During the COVID-19 epidemic, the applications of mobile phone data have expanded rapidly to include building contact networks [2, 3], see chapter 3, and understanding its spread [4, 128–131], see chapter 4.

The importance of social networks in health problems

As evidenced by a wide range of studies, social networks have been shown to play a crucial role in the spread of ideas, opinions, behaviors, and infectious diseases. In the early 2000s, social network analysis gained significant attention following research on preventing the HIV pandemic and other sexually transmitted or blood-borne infections. The idea of assessing risk-potential in human health through social network analysis gained traction due to a growing awareness of the interconnectedness of people and their health [132, 133].

However, the most notable work in this area has been done by Fowler and Christakis. They analyzed multiple and diverse datasets to study the impact of interpersonal influence on human health and the spread of various phenomena, such as obesity, smoking, alcohol consumption, loneliness, drug use, depression, sexuality and sexual orientation, cooperation behavior, happiness, and influenza [134]. They categorized these phenomena into three categories: behaviors, affective states, and germs. Additionally, they demonstrated that current network statistics methods are suitable for analyzing human behavior phenomena.

Albert-Lázló Barabási also applied network science methods to health to understand how symptoms, diseases, and genes interact [135–137]. He showed that a network-based approach to understanding human diseases' complexity and interconnection is feasible and enriching.

One intriguing sociological paradox that is relevant to this thesis is the friendship paradox, which can assist in monitoring disease outbreaks. This paradox refers to the phenomenon where most people have fewer friends than their friends have on average. Mathematically, it means that the mean number of friends of friends is always greater than the mean number of friends of individuals [138]. This phenomenon has

also been observed on social media platforms such as Twitter, giving rise to a related paradox called the virality paradox. The virality paradox states that, on average, your friends receive more viral content than you do [139]. This paradox can be useful in detecting outbreaks of viral information processes in social structures both online and offline.

If we select random individuals in a network who tell us their best friends, they will likely have more friends and be closer to the center of the network. During a viral outbreak, those closer to the center of the network are more likely to get infected earlier than those in the periphery due to a higher centrality. This strategy has been used to select human sensors in a social network and detect early outbreaks of a biological epidemic, such as ILI, before the peak [140]. Similarly, using friends as sensors has been shown to be effective in detecting early signs of informational outbreaks on social media platforms like Twitter [141]. This approach has several advantages, including computational feasibility and greater respect for data privacy, as it does not require monitoring the entire population of a network.

Despite the fact that online social networks mirror offline social networks [142–144], and that Twitter can be used to monitor and predict the seasonal ILI epidemic [92–94], the link between informational epidemics in an online social network and a biological epidemic in the physical environment has not been fully explored until now [1], chapter 2.

As highlighted throughout the introduction of this thesis, any information-based reaction-diffusion process occurs over a network. Whether it is encoded information containing an idea, opinion, behaviour, or a chain of RNA that a virus spreads, it is crucial to model such phenomena and understand their dynamics. Thus, the link between social networks and epidemiology is strong and complementary. One of the earliest models that measured the spread of a disease using social network analysis was in response to the AIDS epidemic in the early 1990s [145]. This work demonstrated the importance of assortative structured mixing, which is the tendency of nodes with the same attributes to link to each other, in detecting significant features of an epidemic. In the 2000s, several research groups started working on modeling the spread of a disease over social networks more realistically [146–148].

Data-driven contact structures as proxies of human social dynamics and interactions

Building on those initial approaches to incorporate real contact networks in the spreading of information, we propose a novel approach for building static and temporal contact structure datasets that enable the realistic modeling of the diffusion-reaction of COVID-19. Novel digital streams provide us with the ability to actively and passively collect data from complex networks, allowing us to observe and measure human dynamics and interactions at an unprecedented scale. The spread of viral diseases occurs within these complex networks, making them a key piece of the puzzle for realistically modelling the dynamics of a viral epidemic. Without them, it would be impossible to feed advanced epidemiological models. Thus, the construction of contact matrices, as a directed graph, is one of the primary tasks in representing human interactions. While advances in network science have allowed for the modelling of the spread of communicable diseases in greater detail over the last two decades, many models still lack a detailed representation of network heterogeneity and its features. Therefore, detailed heterogeneity within contact structures has been shown to be effective in modelling the impact of epidemic outbreaks [149, 150]. The first approaches to contact matrices were used during the 1980s and 1990s to study sexually transmitted diseases [132, 151].

The complexity of social mixing can vary from simple to more realistic, leading to several approaches to modelling the contact patterns of a population. The choice of approach depends on the availability of data. When only the average number of contacts per individual is known, a homogeneous mixing assumption can be made [152]. Group interaction mixing can be implemented when the average number of contacts per age group is available. If the full contact distribution for each individual is known, contact networks between individuals can be constructed [153]. However, this approach may not be feasible for an entire population due to data scarcity [153]. In such cases, multilayer networks can be used to model heterogeneity and mixing patterns between different social groups more effectively [153]. This provides a comprehensive representation of the social contact system and a better understanding of where the pathogen is spreading, such as schools, households, communities, and workplaces.

Infectious disease models have traditionally relied on a priori contact assumptions, which often lack empirical basis. Initial attempts to improve contact

pattern modelling involved indirect observations through surveys, which provided insights into contact patterns [152]. This approach enabled us to start building more sophisticated contact models using multilayer networks [153] to evaluate interventions. Despite these advancements, limitations still exist when access to all necessary information about connectivity between age groups and populations is unavailable.

Our first study to rely on direct observation through GPS data was conducted during the first wave of COVID-19 [2], using detailed mobility and sociodemographic data from Cuebiq and the US Census, respectively. These data allowed researchers to model a subpopulation of approximately 100,000 individuals in the Boston area using a multilayer network that captured social interaction patterns within different layers (i.e., community, households, and schools) to feed a data-driven Susceptible, exposed, infected and recovered (SEIR) model with an Agent-Based Model (ABM). The adult population was based on mobility data, while the children population was generated synthetically based on census data to infer the structure of social contacts [44]. Synthetic populations are commonly used in ABM when sociodemographic data is scarce [44, 154]. The current trend towards privacy makes it impossible to obtain data from an entire population, so census data is used to build synthetic populations along with real-world mobility data to create high-resolution contact matrices for ABM simulations of COVID-19 infection dynamics at metropolitan levels [3, 4].

This approach enables researchers to define and measure NPIs to prevent the healthcare system from becoming overwhelmed [3], chapter 3, or to quantify where infections are occurring based on policies and human behavior [4], chapter 4.

Computational epidemiological modelling

Recent advances in computing have created exciting new opportunities for combining computational thinking and traditional epidemiology. Computational models are used to understand the space-time dynamics of epidemics and assess intervention strategies, ranging from pharmaceutical interventions such as vaccination campaigns and anti-virals to NPIs such as social distancing and non-essential closures. Additionally, there is a complex interplay between human behavior, public policies, the economy, and epidemics, which can be understood using computational techniques and models. Therefore, computational modeling

provides a potent tool to gain insights into the workings of such complex systems.

Computational techniques and models offer a wide range of possibilities to study the dynamics of epidemics. They can range from simple descriptive analyses derived from large databases [136, 155] to more complex generative models, such as ABM, which simulate the spread of disease via social interactions on a complex social network [156–158]. These models can be used not only to study infectious diseases but also other phenomena where information is transmitted, such as behaviors, ideas, or internet memes, or even the diffusion of innovation [159]. Moreover, the populations of interest can vary depending on the disease, including humans, animals, plants [160–162], and even computers [163]. The interactions that are modelled depend on the infectious agent and the population of interest, and they may range from physical proximity for aerosol-borne diseases, sexual contact for sexually transmitted infections, to insect patterns for vector-borne diseases [98].

The Reed-Frost model, developed in 1930 [164], was the first stochastic epidemic model. It is a simple chain binomial and iterative model that predicts how an epidemic will behave over time until no infected individuals remain. This model only requires a set of initial parameters, such as the population size, the number of individuals already immune, the initial number of cases, and the probability of contact, which correspond to the basic reproduction number, R_0, which is the expected number of cases generated by one case. The first numerical implementation of this model was in 1952 [165]. By the end of the 20th century, we began to understand how epidemics spread in scale-free networks [163, 166], and large-scale agent-based models were developed to simulate the diffusion of HIV [167] and incorporate behavioral data [100] in the first decade of the 21st century. These models allow us to run millions of simulations, understand the diffusion and reaction of a pathogen in greater detail, and create "what-if" scenarios to mitigate the spread of the disease. During the second decade of the 21st century, ABMs were tested operationally and proved to be effective during the COVID-19 pandemic in 2020. ABMs can now represent entire populations with real-world mobility and census data, creating "what-if" scenarios to test feasible strategies to reopen society and the economy without overwhelming the healthcare system [2, 3].

To improve the accuracy and robustness of epidemiological models, it is crucial to integrate all available data sources. When dealing with limited data, generative ABMs are useful, but in the healthcare industry, there is a wealth of data that is often

fragmented across different systems. The adoption of big data and machine learning techniques has enabled the integration of this data, leading to impactful use cases in healthcare [168]. This transformational process allows for access to data from entire populations and its application in epidemiological models and studies [169, 170]. However, to build even better, more accurate, and robust EWES, it is necessary to incorporate novel data streams generated outside the healthcare system. The COVID-19 epidemic has accelerated this shift towards improving EWES.

Machine learning techniques are increasingly being used in epidemiology for two purposes. First, to automate data processing pipelines and structure data from various sources such as social media [171, 172], medical records, and images [173–175]. Natural language and image processing models are common tools in digital and computational epidemiology [176]. Second, with the large amount of current data and advances in deep learning techniques, new use cases are emerging, such as predicting the probability of having cancer or other health conditions based on proxies and biomarkers [177, 178]. Therefore, automated processing pipelines based on machine learning techniques are essential for effective early warning systems. These pipelines allow data to be harvested from multiple sources in a standardized manner, with minimal human intervention to reduce errors. This is crucial because the processed and structured data will feed domain-specific epidemiological models that explain and predict health outcomes at individual and population levels. Data-driven and computational epidemiological methods are valuable for building early warning systems that are sensitive to detect real health events, specific enough to avoid false positives, representative by accurately observing health events over time, timely in reporting health events, simple in reporting outcomes, flexible in reporting new health events, and acceptable to healthcare stakeholders and decision makers [179].

In chapter 2, we applied machine learning techniques to automate the data processing pipeline of unstructured micro-blogging data. Moreover, we used statistical learning techniques to explain and predict sensor data based on mobility, network, and content traits. To validate our results, we developed a simple mathematical agent-based model. In chapter 3 and chapter 4, we focused on acquiring the necessary data to feed agent-based models with real-world human mobility data.

1.3 The research purpose of this thesis

The relationship between humans and epidemics is an age-old one, influenced by human behaviors that modify the environment and, in turn, impact human health. It is a two-way relationship that requires constant monitoring to prevent human epidemics. Epidemiology has always relied on data to achieve this goal. However, the advent of user-generated data provides an opportunity to observe human behaviors at an unprecedented scale. Against static or homogeneous approaches to incorporate human behavior to understand epidemic spreading, we can now have the possibility to understand spreading using real-time data-driven mathematical models that include the feedback between epidemics and behavior. This, in turn, can help update the epidemiologist's toolbox of data sources and methods to better model and quantify epidemics at scale. By leveraging novel data streams and advanced social network methods, we can reduce health risks and healthcare costs for taxpayers more effectively. With this in mind, the following research questions will guide this thesis:

- How can novel data streams be utilized to observe human health-related behaviors with greater accuracy and scale?
- To what extent can novel data sources that observe human health-related behaviors enhance the modeling and quantification of epidemics?
- How well do social sensors from digital platforms capture biological processes in the environment, and how can they be optimized for this purpose?
- How reliable is mobility data as a proxy for human social interactions that drive viral biological processes in the environment, and how can it be better integrated into epidemiological models?
- What are the mechanisms through which human behaviors influence the course of viral biological processes, and how can these mechanisms be incorporated into modeling frameworks?
- How can the integration of human behavioral data into mathematical modeling methods enable more granular and accurate predictions of disease transmission dynamics?

Given the novelty, applicability and timely character of our research, we hope this thesis contributes to data-driven behavioural epidemiology, building on the foundational work of giants from multiple disciplines, including medicine, network science, psychology, sociology and computer science. Our research focuses on the interplay between infectious agents in the environment, human behavior, and its impact on the health of millions of individuals. We use novel passive data streams outside of the healthcare system to make direct and indirect observations of these phenomena. Furthermore, we apply advanced computational modeling techniques and social network methods to explain, simulate, nowcast, and predict the complex interaction between the environment, human behavior, and human health.

In 2009, a *Science* editorial [180] highlighted the need for real-time epidemiology and social science research. We believe that this thesis could expand the field and contribute to improving real-time and behavioral epidemiology, particularly in light of the COVID-19 pandemic.

Now, that we have made a deep introduction to the background problem and the state of the art, let see a brief introduction to the coming chapters. In chapter 2, we address the issue of existing EWES being based solely on official data or other limited sources, failing to consider the varying roles played by different individuals in the spread of information. To overcome this limitation, we propose a novel approach that indirectly observes human behaviors through the analysis of real-world social media data using mathematical models. Specifically, we leverage highly connected users, particularly those with high out-degrees on Twitter, to detect early outbreaks of ILI in the physical world, thereby eliminating the need to monitor the entire population. In addition, we showcase the utilization of statistical machine learning models to extract behavioral ILI-related insights from millions of Twitter posts. Through our research, we have identified which high out-degree users are most likely to serve as effective sensors for obtaining early warnings before random users on Twitter or official ILI-related cases in Spain. This approach not only enhances operational efficiency but also respects privacy, as it does not require the collection of extensive amounts of personal data. By incorporating our findings, current EWES for ILI or COVID-19 can be significantly improved and updated, even with limited resources, while ensuring the protection of citizens' data privacy.

In chapter 3, we advance the current state of the art in epidemiological modeling

by incorporating real-world mobility data to construct behavioral contact matrices that enhance both the traditional "well-mixed" approach of spreading within populations. Traditionally, epidemiological models have relied on simplistic assumptions of homogeneous mixing, assuming that individuals interact uniformly within the entire population. However, this fails to capture the complex patterns of human contact and potential variations in disease transmission dynamics. We detail our research on directly observing human behaviors to model and quantify viral biological processes in the environment through the use of real-world GPS data sources. We present a novel approach that utilizes massive datasets of human mobility data from mobile GPSs to construct data-driven contact matrices along with census data, which are then integrated into mathematical models through an ABM. The uniqueness of our approach stems from the utilization of spatiotemporal digital traces obtained from GPS, enabling us to accurately simulate human mobility patterns and capture social interactions at scale through contact matrices.

While the initial primary objective of this study was to demonstrate the validity of our approach and provide real-world contact matrices to feed ABMs to study epidemic spreading, our research happened during the first COVID-19 wave, and we applied our methodology to understand the effects of those initial and future lockdowns in the different waves. We fine-tuned a previous ABM that we were using to understand the spread of influenza across different social stratifications. We integrated our know-how, data, and novel approaches to work within an international scientific collaboration to address the challenge of reconnecting our societies after the lockdown without overburdening hospitalization systems. Additionally, we explored the forensic capabilities of our approach for understanding the dynamics of COVID-19 infections at a granular level, such as where and how they occur. Our data and mathematical models also allowed us to investigate potential effects of lockdowns and contract tracing strategies in the second wave. Through our research, we demonstrate the efficacy of our approach in enhancing the accuracy and realism of epidemiological modeling.

In chapter 4, we extend the methodology proposed in chapter 3 by introducing the temporal component and build contact matrices at daily level. These matrices allow us to observe the feedback loops between social behaviors and infections, as well as the impact of NPIs on social behaviours and infections over time. The

methodology uses temporal mobility data to construct social contact matrices, which incorporate behaviour changes in a population in an ABM that simulates the spread of an infectious disease in the environment. Our approach offers a granular understanding of the spread of viral agents, such as COVID-19, by incorporating a dynamic temporal component that empowers us to identify physical Points of interest (POI) with high probability of Super-spreading events (SSE). Additionally, we evaluate the actual effectiveness of NPIs in controlling the spread of diseases and assess their impact in various metropolitan areas, such as the New York and Seattle Metropolitan areas.

In the conclusions, chapter 5, we present our findings and discuss the implications of our work in this thesis. Furthermore, we identify potential avenues for future research that can be derived from our study. Additionally, we provide an epilogue, chapter 6, with some failed projects that are not in this thesis and personal reflections on the invaluable learning experiences gained throughout the journey of this PhD thesis, that was heavily shifted during the times of war against the COVID-19 epidemic.

2

Epidemic Social Sensors: Harnessing Early Signals for Infectious Disease Outbreak Detection through Social Media Data

"Everything is interaction and reciprocal."

– Alexander Von Handbult[1]

2.1 Introduction

INDIRECT observations of human health-related behaviors through novel data sources, such as microblogging sites, offer valuable insights into modeling and quantifying biological processes, including seasonal Influenza-like Illness (ILI). Detecting early signs of viral outbreaks poses a challenging yet critical task for public health, given the exponential nature of their spread. Social media data streams reflect real-world human behaviors, making them a promising resource for obtaining early warning signals of viral outbreaks. Social networks can provide two different aspects of human behavior relevant to disease spreading. On one hand, they provide real-time, society-wide alerts of how people feel, react, or anticipate the spreading of

[1] Alexander Von Handbult. German Polymath, Geographer and Naturalist. Quote extracted from the book 'Kosmos'.

the illness. At the same time, it can give use some intuition about the offline social network structure in which infections occur. Specifically, central nodes within social networks have been utilized as social sensors in both biological and informational diffusion processes, contributing to the early detection of contagious outbreaks. Yet, the effectiveness and value of social sensors derived from an information-biological viral process as early warning signals of viral outbreaks, particularly seasonal ILI, have not been sufficiently established.

In this chapter, we employ machine learning methods to process a comprehensive dataset covering three years of social media activity in Spain. We demonstrate the feasibility of extracting human health-related behaviors associated to ILI-related mentions and using highly central users, particularly those with a high out-degree on Twitter, as sensors to detect early warning outbreaks of ILI in the physical world, without the need to monitor the entire population. Furthermore, we explore additional behavioral and content features that differentiate these early sensors on social media, moving beyond centrality as the sole criterion.

While high centrality on Twitter emerges as the most distinctive characteristic of these sensors, we also find that they are more likely to engage in discussions related to local news, language, politics, and government compared to other users. Our novel approach enables the detection of a smaller and more efficient set of social sensors for epidemic outbreaks, ensuring operational efficiency and respecting privacy by minimizing the need for extensive data collection.

For the first time, we demonstrate that indirect observation of social sensors from an information-biological viral process, specifically individuals posting ILI-related first-person tweets, aids in the early detection of viral outbreaks. We propose a new approach to explain and predict a biological epidemic, such as the flu, by utilizing the informational epidemic on Twitter and leveraging the network's topology to identify super-sensors in the network that can be used to monitor biological viral processes in the environment.

In the following sections, you will find an updated version of the article *Social Media Sensors to Detect Early Warnings of Influenza at Scale* [1], where we provide further details on our methodology and present our findings.

2.2 Background & Hypotheses

For many viral diseases, the early detection of when and where an outbreak will appear is critical. Public administrations responsible for public health management face public health risks such as the Avian flu [181], Zika [21], SARS [182, 183], Ebola [184, 185] or the latest SARS-COV-2 [186, 187] that can cause millions of deaths in a short period of time at global scale [163]. Traditional health surveillance systems require monitoring and detecting symptoms or case incidence in populations. However, their precision sometimes needs to be improved by the size and delayed testing methods on those populations. Combining those data sources with others about people's mobility, the spatial spreading structure of the disease, and even other data sources seem like a promising venue to establish appropriate warning models in the early epidemic stage [116]. Novel data streams like related web search queries and web visits [91, 111–114], weather data [115] or monitoring multiple digital traces at the same time [116] have proven to be complementary and even advantageous to traditional health monitoring systems. In the same way, social media traces have been demonstrated to be a good proxy for digital epidemiological forecasting models of ILI [92–94]. Online user activity exhibits some benefits like broader spatial and demographic reach or monitoring populations that have no easy access to health services [114].

Since some viruses are transmitted by contact on face-to-face social networks, epidemiological methods that exploit the network structure are more effective in detecting, monitoring, and forecasting contagious outbreaks [188, 189], since they allow to anticipate more accurately the transmission dynamics. Furthermore, these methods can help public health decision-makers to enhance the adoption of public health interventions [190] like social distancing, vaccination, or behavior change campaigns, identifying those individuals most likely to get infected and spread an infectious disease or behavior (e.g., super-spreaders), or which places are more likely to be visited by those individuals [4]. This allows more efficient vaccination campaigns [191] when the vaccination of an entire population is not possible or recommended.

The key idea behind using high-connected individuals to monitor epidemic spreading is that they are more likely to be reached by the infection. In general, human social sensing, when carefully selected, can help predict and explain social

dynamics better [142, 192, 193]. In the absence of complete detailed data about contact networks, simple approaches like the friendship paradox [139] can be used to identify more connected and central individuals (sensors) in the network that can give early signals and anticipate the spreading of information, behavior or disease before it reaches a significant fraction of the population. In particular, the friendship paradox has already been found advantageous to identify sensors for detecting influenza [140, 194, 195] or COVID-19 [196]. In social media, a previous study demonstrated the detection of global-scale viral outbreaks of information diffusion [141] by monitoring high-degree users on Twitter.

In this chapter, we address the question of how we can use sensors for information propagation in online social media to get better early warning signals of a biological epidemic. We hypothesize that social media connectivity and activity are a proxy of social interactions in the real world. Thus, highly-connected users in social media (online sensors) also mirror highly-connected individuals (offline sensors) in the physical contact network. This hypothesis is based on the wealth of literature showing that online networks mimic offline contacts' connections, similarity, and spatial organization [142–144]. Furthermore, we study if it is possible to identify better social media sensors automatically based on their centrality (degree) and mobility, and content behavior. We found that social media sensors can serve as early warning predictors of the exponential growth of an epidemic several weeks before the peak. The current global pandemic threads make it vital to improve the efficiency of Early Warning Epidemiological Systems (EWES) by using operationally efficient methods to anticipate the exponential growth of a virus in a community, region, or country without compromising the citizens' privacy. Our method provides such a system in a fully privacy-preserving framework.

2.3 Data & Methods

Data collection

We extracted Twitter data through their streaming API [197] that allowed us to collect data programmatically on the Spanish mainland. The official ILI rate data was extracted through a web crawler built ad-hoc for the web of the Institute Carlos III of

Health since there was no access to the raw data from an open data portal or a programmatic interface.

ILI-related keywords based search and tweets classification

To get ILI-related mentions from users in the social media platform, we first filtered tweets by keeping those that mentioned simple terms like "flu" or other ILI-related words (See *Appendix A*). After that, we only kept first-person ILI related mentions to exclude general or not directly-related posts like 'The Spanish flu was an unusually deadly influenza pandemic'. This was done using Natural Language Processing methods. We applied a text classifier, using a `scikit-learn` implementation [198]. We handpicked and labelled a set of 7836 tweets to train our classifier, containing 3918 for true positive (first-person) tweets and 3918 for true negative tweets. Using that labelled data our classifier achieved an accuracy of (~ 0.94). We then applied our classifier to identify first-person mentions the remaining tweets (See *Appendix A*, section 1, for more details about our pipeline). After this process we ended up with $N = 19696$ users and 23975 tweets classified as first-person ILI-related post.

ILI-related post time series

We added up and normalized the number of weekly users mentioning the flu by the number of the total number of users in the system, we followed equation

$$\hat{x}_{users,t} = \frac{x_{\text{ILI Users},t}}{x_{\text{Total Users},t}}, \tag{2.1}$$

where t is the week. This time series is shown in Figure 2.1, together with the prevalence of ILI cases.

Centrality features

Each tweet has information about the out-degree (followees), $d_{out,i}$, and in-degree (followers), $d_{in,i}$, for each Twitter user i posting it. We used them as proxies of the network centrality for each user. We tested out several aggregated centrality features for the selection of sensors. We calculated the weekly total, mean, median, maximum and minimum out-degree of individuals before and after the peak making

first-person ILI related mentions to test which centrality metric had more explanatory power. We found that the weekly total out-degree was the best centrality metric to apply (See *Appendix A*, section 2 for further details). The weekly total out-degree is defined by

$$D_{T,t} = \sum_{i \in \Omega_t} d_{out,i,t} \tag{2.2}$$

where Ω_t is the set of users that made a ILI-related mention at week t.

Sensors are selected as the group of users with $d_{out,i} > 1,000$. For that group we also define the time series of their centrality as

$$D_{S,t} = \sum_{i \in \Omega_t^*} d_{out,i,t} \tag{2.3}$$

where Ω_t^* is the set of users in the sensor group that made a ILI-related mention at week t.

Linear autoregressive model

The following equation represents a linear autoregressive model for explaining and nowcasting the dependent variable, I_t, being the Official ILI rate for each week. $D_{T,t}$ are total weekly out-degree for the whole twitter population, and $D_{S,t}$, are total weekly out-degree for the whole sensor population. We followed

$$I_t = \beta_0 + \beta_1 I_{t-1} + \sum_{\delta \geq 0} (\alpha_\delta D_{T,t-\delta} + \gamma_\delta D_{S,t-\delta}) + \epsilon_t. \tag{2.4}$$

Agent-based model of ILI disease and information diffusion

To understand our empirical findings, we compare them with the simulations of epidemic spreading on a physical and online network through an agent-based model (ABM). We model the offline (physical) contacts using a random heavy-tailed network. Specifically, we created a synthetic population of $N = 150k$ agents with are connected through a scale-free network with degree distribution $P(k) \sim k^{-3}$ obtained through the Barabasi-Albert model. [199]. The network was built using the R package `igraph` [200].

At the same time, we supposed that each of the agents participates in a social media

platform. Our hypothesis is that the online degree of the agents is related to the offline degree in the complex network. To account for some variability, we assumed that the degree in the social media platform was modified by a random uniform distributed number (See *Appendix A*, section 4 for more details). Thus, the degree in the social media platform is given by $d_{out,i}^{\text{Twitter}} = d_{out,i}^{\text{Offline}}(1 + \nu_i)$, where ν_i is a random number uniformly distributed between 0 and 1. This way we account for potential variability between offline and online degrees.

We simulate the ILI spreading using a simple Susceptible-Infected-Recovered (SIR) epidemic model. In particular, at each time-step *t*, the infectious (I) agents can transmit the disease to their susceptible (S) neighbors in the contact network with probability β, see Equations (1.1)-(1.3). If the transmission is successful, the susceptible node will move to the (I) state. An individual will move independently to the recovery (R) state with a probability α. We initialized the model with two initial infected seeds. After getting infected, we assumed that the agent immediately posted an ILI-related tweet on the social media platform. In our model, we considered a user to be sensors if she has an out-degree in the platform higher than four times the average degree in the Barabasi-Albert model. We also calibrated the time unit in this model so that the epidemic curves have a similar time scale as the real ILI rate (See *Appendix A*, section 4 for further details on the simulation's parameters).

User traits

To characterize the different traits of Twitter users, we analyzed the tweets of each user during a time window of 30 days before the initial event. For the sensor group, we selected individuals with an out-degree $d_{out,i} \geq 1000$ and that made at least a ILI-related mention during the weeks $-15 \leq t \leq -2$ before the peak of the epidemic. The initial event is their first post with the ILI-related mention. For the control group, we picked individuals that made an ILI-related mention after the $-15 \leq t \leq -2$, then we picked a random post of them as initial event in weeks $-15 \leq t \leq -2$, before the peak of the epidemic. Using that 30 days period we computed different Mobility, Content, and Network traits to characterize each user

Mobility traits

We worked out the mobility pattern from a user by looking at geolocations from tweets. To characterize their mobility we used the radius of gyration [201] which measures the size of the area covered while moving around:

$$R_g^i = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_{mean}).} \tag{2.5}$$

Content topics

We extracted topics from users tweets' texts. To this end, we use the TextRazor classifier trained against the IPTC newscodes [202] which classify each tweet into approximately 1400 high-level categories organized into a three-level tree hierarchy. Each tweet is give a probability to contain such topic. Thus each user is characterized by a content vector of n topics

$$C^i = \{C_1^i, C_2^i, \ldots, C_n^i\} \tag{2.6}$$

where the components C_m^1 are the aggregated probability of topic m in all her tweets.

Network traits

Apart from the out-degree for each user i we also took into account the total user activity in the social network platform, by computing the number of tweets generated during the period of observation, this variable is called number of posts.

Linear logistic regression model

The following equation represents a linear logistic regression model for explaining the probability of an individual being a sensor by different features, where $\{M^i\}$ are the mobility features (we only consider the radius of gyration variable, R_g), $\{N^i\}$ the group of network variables, out-degree, $d_{out,i}$, and number of posts, and $\{C^i\}$ is the group of content variables for each individual i. Our model is

$$\Pr(i \in \Omega^*) = \text{logit}^{-1}[\beta_0 + \sum_l \alpha_l M_l^i + \sum_n \beta_n N_n^i + \sum_m \gamma_m C_m^i] \tag{2.7}$$

where Ω^* is the set of users defined as sensors, and $\text{logit}^{-1}(x) = e^x/(1+e^x)$. In the model, each individual variables in the different groups is standardized to have zero mean and unit variance.

2.4 Results

We used social media traces obtained from the micro-blogging site Twitter, where we collected more than 250 million tweets from December 2012 to April 2015 on Spain's mainland. Using Natural Language processing techniques, we only included first-person ILI-related posts, summing up a population of 19696 users with at least one first-person ILI-related mention, which comprised a total of 23975 tweets (*Appendix A*, section 1 discusses our methodology). We also made use of official ILI cases from the surveillance system for influenza in Spain (ScVGE) [203] managed by the Instituto Carlos III de Salud [203]. This system reported weekly ILI cases in Spain for each province with two weeks of delay in the state of the seasonal flu epidemic based on the current European Union proposal that regulates ILI surveillance [204]. Our dataset of official ILI cases ranges from December 2012 to April 2015 and includes three different seasons of influenza outbreaks in Spain.

Figure 2.1 shows a generalized ILI season from the average of ILI cases and ILI-related mentions for the three seasons. ILI cases and ILI-related mentions time series have a Pearson correlation of 0.87 (CI $[0.79, 0.93]$ and $p_{value} < 0.001$). Since different outbreaks happen at different times of the year, we have shifted each influenza outbreak to the time of its peak. We can see that ILI-related mentions precede the official ILI cases at the beginning of the growth stages before the peak. Previous studies have proved this [92–94]. Mentions of the outbreak in social media seem to precede the exponential growth in the total population. ILI-related posts peak at -15 weeks could be related to the start of the cold season and users mixing ILI symptoms with cold symptoms, stating that they are suffering from ILI. We found a similar pattern in Google trends data.

Validating high-degree individuals as sensors

However, here, we want to go a step further. Can we subset the users posting ILI-related posts to get better earlier warnings about the outbreak than monitoring

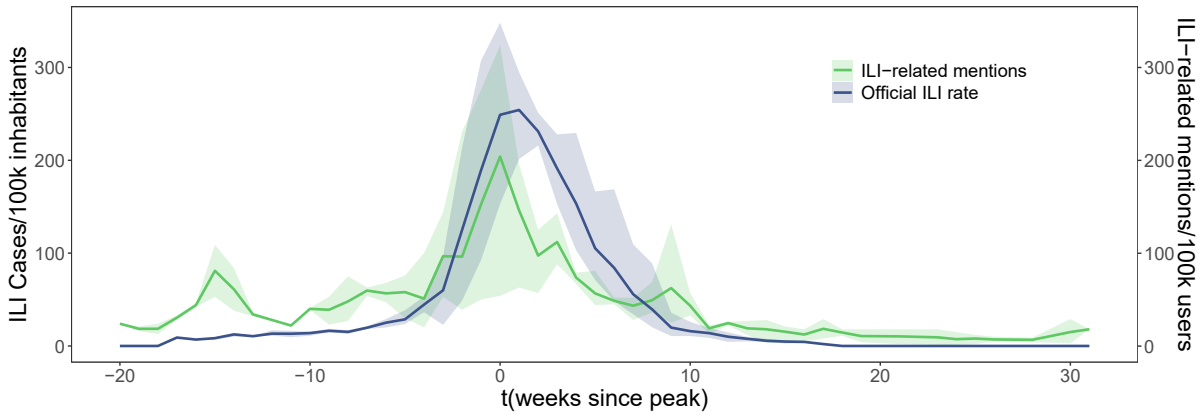

Figure 2.1: Average ILI season. An average ILI season centered to the peak for Spain mainland from December 2012 to April 2015. Horizontal axis is the temporal axis that measures weeks since peak. Primary vertical axis (left) is the Official ILI rate prevalence, cases per 100k inhabitants. Secondary axis (right) is the Twitter ILI-related mentions prevalence, mentions per 100k users. Lines are average weekly incidences for Official ILI rates (Blue) from Instituto Carlos III de Salud and first-person ILI-related mentions rate (Green) from Twitter. Bands are their confidence intervals. Figure reproduced from [1].

the whole social network platform? Similarly to [140], and [141], high-degree users could be better than the average individual on the platform. To test whether high centrality or degree correlates with early signals, we measure the total weekly out-degree, D_t, of users having social ILI-related mentions before and after the peak. Figure 2.2 shows distributions for D_t before the peak, after the peak and for the whole season. There is a statistically significant difference in the mean ($p_{value} < 0.01$). The average total weekly out-degree is 31108 (Confidence Interval, CI $[21539.03, 40677.32]$) before the peak, while it is only 14373 (CI $[11202.94, 18455.78]$) after the peak. The difference is also present in extreme values. We modelled large values of D_t as power laws with an exponent of 2.56 (CI $[2.51, 2.62]$) for the whole period. For the weeks before the peak, it follows an exponent of 2.10 (CI $[1.91, 2.29]$). Finally, for the weeks after the peak, it follows an exponent of 2.86 (CI $[2.48, 3.25]$). Thus, on the aggregated level, we indeed see that the users in social media that have ILI mentions before the peak have more social connections than after the peak. This result signals the possibility of using high-connected users as potential early sensors.

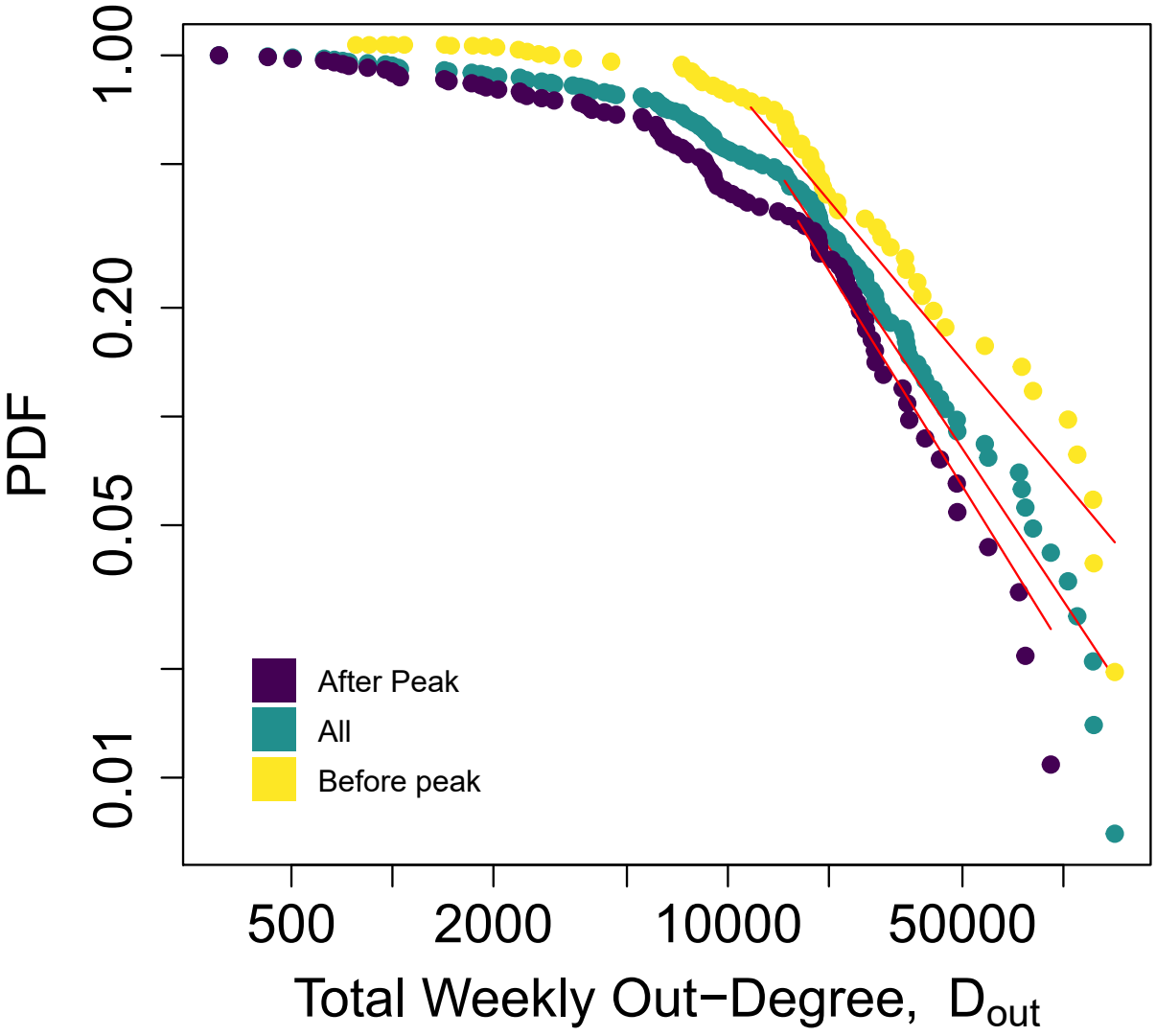

Figure 2.2: Total weekly out-degree before and after the peak. Total weekly out-degree, D_t, power law distributions for the whole season (Green), weeks before (Yellow) peak and weeks after peak groups (Purple). Horizontal axis, total weekly out-degree, D_t. Vertical axis probability distribution functions. Figure reproduced from [1].

This result is robust against other aggregated degree centrality variables (see *Appendix A*, section 2). For selecting sensors, we selected each individual with an out-degree greater that 1000 (see *Appendix A*, section 3).

Figure 2.3.A compares Twitter's cumulative ILI-related mentions of our control and sensor groups against the official ILI-related cases. As we said before, the activity in social media for both the control and sensor groups anticipates the cumulative

incidence of ILI cases by one or two weeks. For each user i we define t_i^{post} as the time in which she has an ILI-related post on social media. Figure 2.3.B shows confidence intervals for ILI-related posting times for each group and ILI season, relative to the peak $t_i^{post} - t^{peak}$. For all ILI seasons, the control group has an average ILI-related posting time of $\Delta t_C = \langle t_i^{post} - t^{peak} \rangle_{i \in C}$ = -5.35 (CI $[-5.54, -5.17]$) weeks before the peak. The sensor group has an average ILI-related posting time of $\Delta t_S = \langle t_i^{post} - t^{peak} \rangle_{i \in S}$ = -6.72 (CI $[-7.42, -6.02]$) weeks before the peak. This yields that sensors are posting on average $\Delta t_S - \Delta t_C = -1.37$ (CI $[-2.08, -0.64]$ and $p_{value} < 0.01$) weeks before the control group, during the exponential growth phase, between 8 to 4 weeks for all seasons. In more detail, the 2012-2013 season has a $\Delta t_S - \Delta t_C = -0.62$ (CI $[-1.58, -0.84]$ and $p_{value} > 0.1$) weeks, the 2013-2014 season has a $\Delta t_S - \Delta t_C = -2.46$ (CI $[-3.45, -0.36]$ and $p_{value} < 0.01$) weeks, and the 2014-2015 season has a $\Delta t_S - \Delta t_C = -1.54$ (CI $[-2.45, -0.63]$ and $p_{value} < 0.01$) weeks. As we can see, the ILI-related mentions of sensors could anticipate the epidemic's growth by 1 or 2 weeks with respect to other users in the platform.

Autoregressive models with sensors and its theoretical validation

To quantify statistically how valid our sensors in social media could be in a potential EWES model, we built an autoregressive model that considered different epidemiological and social media features (see *Methods* section). The models considered different combinations of the total number of weekly ILI cases at time t, I_t, the total weekly out-degree of all users from the social media platform ($D_{T,t}$) that posted ILI-related mentions, and the total weekly out-degree of the subset of those users in the sensor group ($D_{S,t}$). We have also considered different temporal week lags, $t - \delta$, for each variable to test their potential role as early warning signals. As a baseline, we have considered a model that only incorporates the ILI cases and their autoregressive power at $t - 1$. As we see in Table 2.1, that simple model is already quite accurate in explaining the evolution of the weekly ILI rate. On top of that baseline model, we built four others, including the degree centrality of all users and the sensor group at different lags. For each model, we predict the I_t number of ILI-related cases using the information of the I_{t-1} cases and the total out-degree of all users and sensors with ILI-related mentions at time t and $t - \delta$. We ran all models using a step-wise approach to keep only statistically significant regressors for

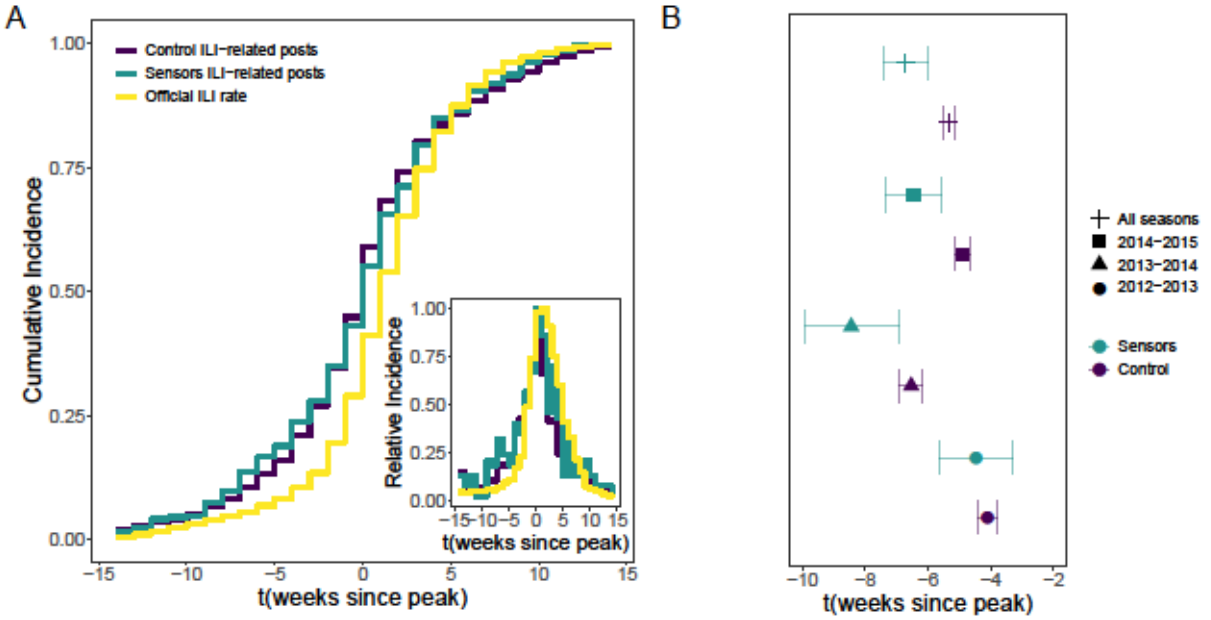

Figure 2.3: Cumulative incidence between real ILI, all Twitter, and only Twitter sensors. Empirical cumulative distribution differences in official ILI cases (Yellow), control ILI-related mentions on Twitter (Purple), and sensor ILI-related mentions on Twitter (Green). Horizontal axis measures weeks since the peak on ILI cases. Top-Inset: We also show the box-plot for each group ILI-related mentions time before the peak, including their median (vertical thick line) and mean (red circle) Bottom-Inset: weekly incidence for ILI cases, control ILI-related mentions and sensor ILI-related mentions on Twitter. Figure reproduced from [1].

$\delta = 1, 2, 3, 4$. Due to multicollinearity problems between variables, we also monitor the variance inflation factor (VIF) for each to choose the best δ. Results in Table 2.1 and Figure 2.4A quantitatively show the importance of social media ILI-related mentions, especially those from the sensor group. As we can see, the predicting power (adjusted R^2) on next week's official ILI rate after incorporating social media mentions increases significantly (and we also reduced collinearity), especially at three- or four-week lags. In all those cases, the total degree of sensors at time T and time $t - \delta$ has a significant regression coefficient and role (in R^2) in the prediction. That is, social sensors can help anticipate official ILI cases three to four weeks before, a result consistent with previous similar analyses of ILI contagious outbreaks in small settings [140] or of information spreading in social media [141]. We also note that the signs of the variables of all users and sensors have different effects. For example, a

	Official weekly ILI rate, I_t				
	I	I+T+S $\delta = t-1$	I+T+S $\delta = t-2$	I+T+S $\delta = t-3$	I+T+ S $\delta = t-4$
	(1)	(2)	(3)	(4)	(5)
I_{t-1}	0.924***	0.789***	0.856***	0.849***	0.800***
	(0.041)	(0.047)	(0.049)	(0.045)	(0.044)
$D_{T,t}$		0.717***	0.634***	0.580***	0.561***
		(0.0003)	(0.0002)	(0.0002)	(0.0002)
$D_{T,t-1}$		−0.281***			
		(0.0003)			
$D_{T,t-2}$			−0.344***		
			(0.0003)		
$D_{T,t-3}$				−0.313***	
				(0.0002)	
$D_{T,t-4}$					−0.217***
					(0.0002)
$D_{S,t}$		−0.443***	−0.393***	−0.348***	−0.339***
		(0.0004)	(0.0003)	(0.0003)	(0.0003)
$D_{S,t-1}$		0.211***			
		(0.0005)			
$D_{S,t-2}$			0.227***		
			(0.0004)		
$D_{S,t-3}$				0.186***	
				(0.0003)	
$D_{S,t-4}$					0.132***
					(0.0003)
Constant	0.000	0.000	0.000	0.000	0.000
	(4.627)	(4.093)	(4.071)	(4.123)	(4.516)
Observations	87	87	86	85	84
R^2	0.854	0.925	0.932	0.935	0.929
Adjusted R^2	0.852	0.920	0.928	0.931	0.924
Maximum VIF	NA	15.16	9.36	6.77	**5.28**
Residual Std. Error	34.092	25.042	23.951	23.529	24.741
F Statistic	497.040***	199.556***	218.885***	226.755***	203.163***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 2.1: Empirical ILI regression models. Regression table with normalized beta coefficients for each group of variables, Official (I)LI, (T)witter and (S)ensors, where X_t are weekly ILI related variables for each group. $D_{T,t}$ and $D_{S,t}$ are weekly total out-degree variables from Twitter (T) and Sensors (S). Table reproduced from [1].

higher total degree of sensors at times $t - \delta$ predicts more ILI-related cases (positive coefficient) at time t for $\delta > 0$, but a smaller number of cases (negative coefficient) for $\delta = 0$. As we will see below, this apparent contradiction comes from the high auto-correlation of the time series of ILI-related cases and the total degree of users.

We investigated the predicting power of high-degree sensors in a synthetic model to validate that sensors anticipate ILI cases because social media connectivity mirrors social connections in the real world. Specifically, we built a base agent-based susceptible-infected-recovery (SIR) epidemic spreading on a random network mimicking real (face-to-face) social contacts between people (see *Methods* for details about the network and simulations details). Apart from their physical contacts, we also assumed that each person has acted on a social media platform and that the degree in both the real and online networks are correlated moderately. Assuming that agents post on social media when they are infected, we also constructed the time series $\hat{D}_{T,t}$ and $\hat{D}_{S,t}$ for the model and their autoregressive fits as in Table 2.1. Our results once again show that high-degree agents (sensors) carry some predicting power on the epidemic spreading.

Furthermore, the coefficients for the different models show the same regression structure as the empirical models in 2.1, see Figure 2.4A. We can see that both coefficient structures are nearly the same, including their magnitude and signs. Although this is not direct proof of our hypothesis that the online and offline centrality of real users is similar, it shows that under that assumption, we not only get that the effect of sensors is the same as we found in our empirical analysis, but even the structure of coefficients (magnitude and sign) is similar. These results support the idea that sensors in an informational epidemic that mirrors a biological epidemic are also sensors of a biological epidemic, like ILI, that we can trace on Twitter.

Identification of sensors beyond out-degree

So far, we have seen that high out-degree users in social media can be early sensors of ILI cases. However, can we identify a better group of sensors beyond high degrees by looking at other traits? Are individuals that signal the epidemic's early stages defined just by their centrality degree, or do they have other behavioral or content traits? To do that, we define a sensor functionally as every user who posts an ILI-related tweet from fifteen weeks to two weeks before the epidemic's peak ($-15 \leq t \leq -2$). On the other hand, a control user was a random user who did not talk about ILI during the same period. (see *Methods* contextual features for more details).

To characterize users' content, behavior, and network traits in both groups, we analyzed every tweet they posted 30 days before their first ILI-related tweet (sensors)

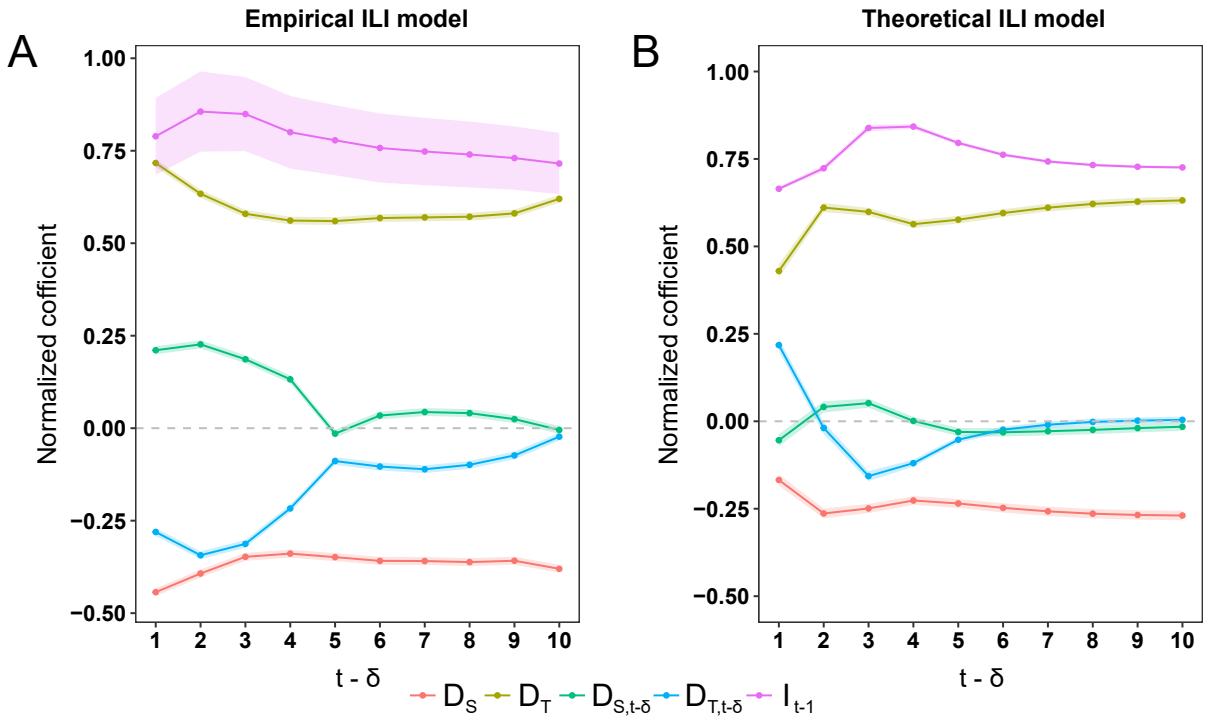

Figure 2.4: Results for the Empirical and Theoretical ILI auto-regression models. (A) Normalized coefficients for the different autoregressive models for I_t, see Eq. (2.4) for different time lags δ. Model regressors for each δ are the number of cases one week in the past I_{t-1}, the total out-degree at time t, $D_{T,t}$, the total out-degree at time $t-\delta$, $D_{T,t-\delta}$, and the total out degree of sensors at time t, $D_{S,t}$ and at time $t-\delta$, $D_{S,t-\delta}$. We show the normalized coefficient and their confidence intervals (shaded area). (B) Same as in (A) but for the agent-based model of ILI disease and information diffusion. Figure reproduced from [1].

or a randomly chosen tweet (control). Specifically, we identify three groups of traits for each user. Firstly, we extract the content of each user's tweets and classify them into topics like sports, politics, entertainment and many other categories using the TextRazor classifier (see *Methods*). Secondly, since our tweets are geolocalized, we extract the mobility features of each user, in particular, the radius of gyration, which measures the size of the area covered while moving around [201]. The radius of gyration could proxy the number of different and diverse people the user is in daily contact. Thus it might serve to estimate potential exposure to infected people [205]. Lastly, we also use their activity (number of posts) and, as before, their out-degree in the social network.

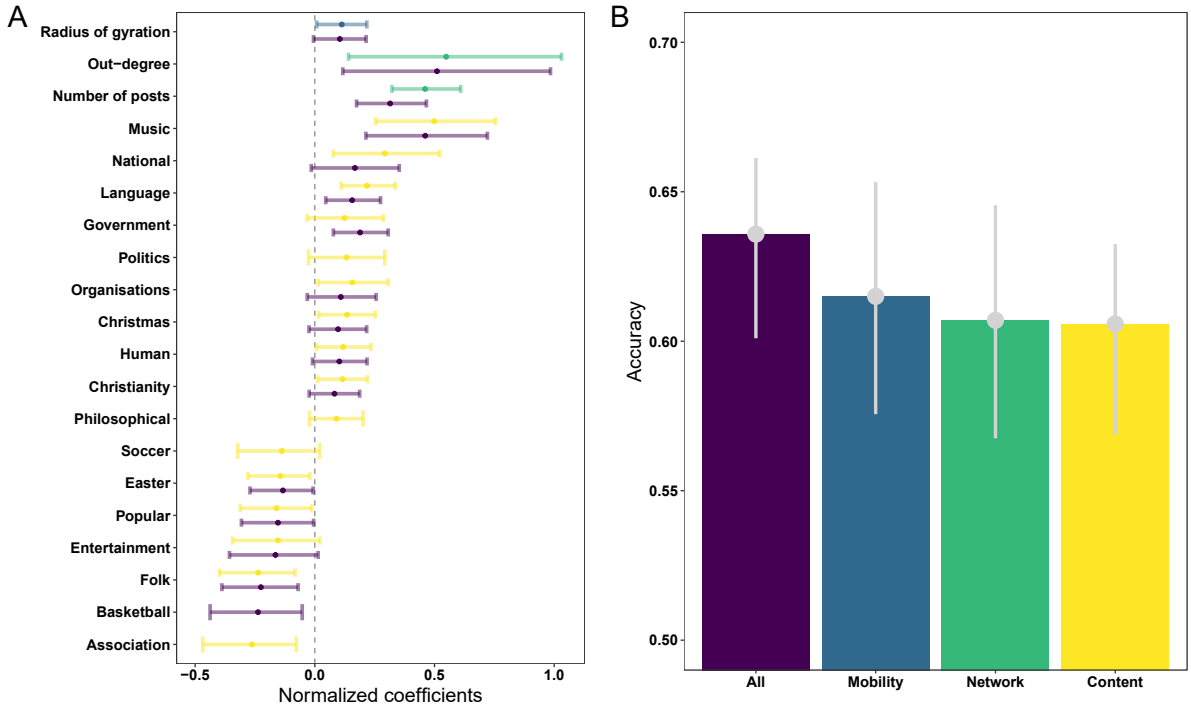

Figure 2.5: Super-sensors prediction models. (A) Normalized beta coefficients from logistic models for each factor for explaining sensors by content topics they posted (yellow), their network features (green), their mobility by the radius of gyration (blue) and all group variables together (purple). Horizontal axis measures the normalized coefficients from the logistic regression models. Vertical axis labels are variables. (B) Accuracy metrics for each group of variables: topics (yellow), network (green), mobility (blue) and all variables (purple). Vertical axis measures models accuracy. Horizontal axis represents each group variable model. Figure reproduced from [1].

To test how relevant those groups of traits are to define a sensor, we developed a straightforward logistic regression model (see *Methods*) to classify users into the sensor or control groups using different variables. As we can see in Figure 2.5, the accuracy of our models is above the primary level (0.5). While Network and Content groups independently achieve similar accuracies (~ 0.61) than the Mobility group (~ 0.62), we get better accuracy, including all types of traits (~ 0.64). This result signals that even different traits carry complementary information about who could be sensors in the social media platform. To understand this further, we looked into each trait's (normalized) coefficients in our model. As shown in Figure 2.4.A, the most crucial variable to predict a user in the sensor group is still the out-degree in the

social network, even after controlling for the number of posts. This is important because it shows that our simple method of using high-connected Twitter users as sensors works much better than other traits. We also see a small but significant effect on the radius of gyration, meaning, all things equal, users that move further are more likely to be sensors. Regarding the content, we see a structure of topics that users in the sensor group are more likely to discuss, like National, Language, Politics, and Government. On the contrary, users that talk about Sports, Popular topics, or Entertainment are less likely to be in the sensor group. This finding could signal and be related to other unobserved user traits like income or educational attainment level, which also are known to be related to the activity in social media [206] and amount of real offline contacts [207].

2.5 Discussion

Early warning epidemiological systems (EWES) detect outbreaks weeks in advance to help public health decision-makers make more efficient allocations of public resources to avoid or minimize an overflood of contagious in the healthcare system. EWES are undergoing significant investments and changes due to the COVID-19 disruption. However, most of them harvest vast amounts of data and do not exploit the explanatory and predictive power of the network heterogeneity where a disease-informational epidemic is spreading.

In this study, we demonstrated that social media traces, like Twitter, could be used as a source of social-behavioral data to monitor disease-informational epidemics that mirror offline biological contagious disease epidemics, like ILI, by exploiting the network heterogeneity whenever social centrality measures of the network are available. By having a simple centrality metric, such as the out-degree, we can define suitable sensors for the disease-informational epidemic in the network. When aggregated correctly, we can use sensors to feed autoregressive models that could yield signals of an outbreak up to four weeks in advance. Although previous studies showed the advantage of using social network metrics to detect, monitor, and forecast contagious outbreaks [188, 189]. The usage of sensors in a network to detect early warnings of an outbreak in a biological disease contagious epidemic [140, 194], or informational epidemics [141], our study is the first to combine the use of sensors

in social media to anticipate epidemics in real life. Our results are based on the hypothesis that social media networks are related to offline contact networks, which has been validated directly in other works [142–144]. Our empirical and theoretical results show that instead of harvesting large amounts of data and metrics from social networks [94], we can track and anticipate early outbreaks of a disease-informational epidemic by inexpensively looking at a small set of specific users (sensors).

We also demonstrated that sensors could be profiled and detected automatically from social media raw data by using their topological network properties and based on the content posted by individuals and their mobility patterns. Explicitly, we found that sensors talk more about some topics like National, Politics, and Government and less about Sports and Entertainment. The fact that those topics could also be related to their income, educational attainment [206], but also to other traits like more extroversion personality traits [208] opens the possibility to investigate the potential overlapping reasons why sensors not only are more prone to get infected earlier but also that they would like to post about it on social media. For instance, Music topic requires further investigation, previous literature suggests individual differences in personality in the way we use and experience music [209], possible having a social component.

Finally, our method uses the out-degree in the social media platform as a proxy for centrality. Better knowledge of the network structure could yield more optimized methods to detect highly-central users. Our approach also has other limitations. For example, our data corresponded only to a given epidemic in a given country and were not tested against more global epidemics like the COVID-19 pandemic. However, given that our findings rely on the collective behavior of people in social media and the observed relationship between offline and online networks [210,211], we think that our findings could be extrapolated to other epidemiological situations. We hope our research can help study the role of sensors in other pandemics, specially COVID-19, where more information about real-world offline contact networks exists due to better mobility data [3] or contact tracing applications.

In summary, this study proposes a feasible approach to exploit the network heterogeneity underneath social media sites, like Twitter, to detect more efficiently and earlier outbreaks from a disease-informational epidemic that mirror a biological disease contagious epidemic, like ILI. Furthermore, the sensors approach we used to detect early outbreaks within informational epidemics and biological contagious

disease epidemics, but this is the first time in a disease-informational epidemic as we have done in this study. Finally, novel epidemiological systems have been developed for other pathogens such as Zika, SAR, or COVID-19, among others, in addition to influenza, using conventional and non-conventional data sources such as the official public cases, online searches, or health forums. For instance, for the COVID-19 pandemic, some studies used social media traces to try to predict the dynamics of the pandemic [212, 213]. Such approaches, along with our findings about the power of the network structure, could improve the results of their predictions.

Also, health systems and health organizations initiatives, like the Global Outbreak Alert and Response Network (GOARN) [68] from WHO that is composed of 250 technical institutions and networks globally and projects like the Integrated Outbreak Analytics (IOA) [69], Epidemic Intelligence from Open Sources (EIOS) [70], and Epi-Brain [71] that respond to acute public health events. This network is already moving in a double direction of incorporating early warnings from Big Data, social sciences techniques and behavioral data into epidemic response systems [72] to control outbreaks and public health emergencies across the globe. Also, syndromic surveillance platforms like InfluenzaNet could ask for twitter profiles or number of people an individual interacted with in the last week. Our innovative approach might help detect early outbreaks without having to monitor and harvest data from a whole population, making EWES more accurate in time prediction of an outbreak, more efficient in resources and more respectful regarding citizens' data privacy. Additionally, in this chapter, we introduce a novel approach for monitoring network sensors derived from social media traces and their mirror power, enabling indirect observations of human health-related behaviors for the purpose of modeling and quantifying social epidemics, contributing to more effective and respectful approach with users data.

3

Data-Driven Contact Networks: Modeling and Quantifying Infectious Epidemics through Real Human Mobility Data

"The observer, when seems to himself to be observing a stone, is really, if physics is to be believed, observing the effects of the stone upon himself."

– Bertrand Russel[1]

3.1 Introduction

HUMAN behaviors can now be observed more effectively than ever before, thanks to the availability of novel data streams. One such stream, behavioral mobility data, has emerged as a reliable proxy for understanding human behavior in geographical spaces [96, 123]. To gain a more accurate and insightful assessment of disease transmission dynamics during an epidemic, it is imperative to possess a comprehensive understanding of people's mobility behaviors, focusing on their locations and durations of stay, rather than solely relying on homogeneous mixing models. Homogeneous mixing models, while useful in certain contexts, oversimplify

[1]Bertrand Russell. British Mathematician and Philosopher. Quote extracted from the book 'An Inquiry Into Meaning and Truth'

the intricate complexities of human interactions and movement patterns. In contrast, by integrating data on individuals' locations and duration of stays, we can capture the true heterogeneity of their behaviors and socialcontacts. This approach acknowledges that people exhibit diverse travel habits, interact with varying frequencies, and cluster in specific geographic areas, all of which significantly influence the spread of infectious diseases.

Effectively processing and analyzing behavioral mobility data can yield valuable insights for modeling and quantifying viral epidemics. Identifying high-traffic regions, gathering places, and travel hotspots provide essential information for pinpointing potential transmission hubs. Additionally, it enables us to evaluate the effectiveness of interventions and forecast the course of an epidemic with greater accuracy. For instance, the COVID-19 pandemic highlighted the effectiveness of stringent social distancing measures in slowing down the virus's spread. However, as restrictions were gradually eased, changes in human mobility posed the potential for second-wave scenarios to emerge. Only by understanding how humans change and adapt their mobility we can anticipate and manage potential new epidemics or new waves within them.

To address this concern, we integrated anonymized, geo-localized mobility data with census and demographic data to feed a detailed agent-based model (ABM) of SARS-CoV-2 transmission, specifically in the Boston metropolitan area. This utilization of data allowed for obtaining granular and detailed information on the virus's spread across real contact matrices, without relying on compartmentalizing the population in broad groups or assuming homogeneous mixing contact matrices as traditional SIR differential equation models or ABM do.

Furthermore, our approach enables the exploration of various what-if and counterfactual scenarios, providing crucial insights to health decision-makers regarding the potential effects of different social distancing policies on the healthcare system. Our research findings indicated that a period of strict social distancing, followed by a comprehensive combination of testing, contact tracing, and household quarantine measures, could effectively manage the disease within the healthcare system's capacity while allowing for the gradual reopening of economic activities. Notably, our results emphasized the significant role that an enhanced testing and contact tracing response system could play in easing social distancing interventions,

particularly in the absence of herd immunity against SARS-CoV-2.

By harnessing these novel mobility data streams and employing an ABM framework, we significantly enhanced our understanding of viral agent spread in the environment. Our approach not only improved modeling accuracy but also empowered health decision-makers with valuable information for guiding policy choices related to social distancing measures and healthcare planning. Therefore, in this chapter, we present a comprehensive demonstration of how to process mobility data to construct social contact matrices, which can be incorporated into a SIR ABM using real-world data. As far as we know, we were the first research group to pioneer the use of real mobility data to build social contact matrices to study epidemics. Our research was done during the daunting times of the first wave of COVID-19 and contributed to the discussion of what measures could be implemented after the first wave to prevent subsequent waves.

In the following sections, an updated version of the article *Modelling the impact of testing, contact tracing, and household quarantine on second waves of COVID-19* [3] is provided, offering further insights into our methodology and findings.

3.2 Background & Hypotheses

The first report of a new infectious disease, later coined COVID-19, appeared on 31 December 2019 [214]. As of 15 July 2020, when we finished writing this research, the virus spread to 188 countries with more than 13.3 million confirmed cases worldwide, and killed more than 579,500 people [215]. As the number of confirmed COVID-19 cases increased and the expansion of the disease entered into a global exponential growth phase, a large number of affected countries were forced to adopt non-pharmaceutical interventions at an unprecedented scale. Given the absence of specific antiviral prophylaxis, therapeutics, or a vaccine, non-pharmaceutical interventions ranging from case isolation and quarantine of contacts to the lock-down of entire populations were implemented with the aim of suppressing/mitigating the epidemic before it could overwhelm the healthcare system. Although these aggressive measures appeared to be successful in reducing the number of deaths and hospitalizations [131, 216], and in reducing the transmission of the SARS-CoV-2 virus, the absence of herd immunity after the first wave of the epidemic pointed to a large

risk of a resurgence when interventions were relaxed and societies went back to a "business as usual" lifestyle [217–219]. It was therefore of paramount importance to analyze different mitigation and containment strategies aimed at minimizing the risk of potential additional waves of the COVID-19 epidemic while providing an acceptable trade-off between economic and public health objectives.

In the present work, through the integration of anonymized and privacy-enhanced data from mobile devices and census data, we built a detailed sample of the synthetic population of the Boston metropolitan area in the United States (see Figure 3.1). This synthetic population (Figure 3.1.a) was used to define a data-driven agent-based model of SARS-CoV-2 transmission and to provide a quantitative analysis of the evolution of the epidemic and the effectiveness of social distancing interventions. The model allowed us to explore strategies concerning the lifting of social distancing interventions in conjunction with testing and isolation of cases and tracing and quarantine of exposed contacts. Our results indicated that after the abatement of the epidemic through the "stay at home" orders and halt to all nonessential activities, a proactive policy of testing, contact tracing, and contacts' household quarantine could allow the gradual reopening of economic activities and workplaces, with a low COVID-19 incidence in the population and a manageable impact on the health care system.

Modeling COVID-19 using real human mobility data and ABM models

To provide a quantitative estimate of the contact patterns for the population of agents and to build the synthetic population of the Boston Metropolitan Area (BMA), we used detailed mobility and socio-demographic data and generated a network that encodes the contact patterns of about 85,000 agents in the area during a period of six months (see *section 3.3* for more details). Agents were chosen to be representative of the different census areas in the Boston area following the methodology used in Ref. [220]. This defines a weighted multilayer network consisting of three layers representing the network of social interactions at (1) workplace/community level (W+C), (2) households, and (3) schools, as shown in Figure 3.1.a. Connections between two agents in the W+C layer were estimated from the data by the probability of both being present in a specific place (e.g. restaurant, workplace, shopping) weighted according to the time they spent in the same place. A second

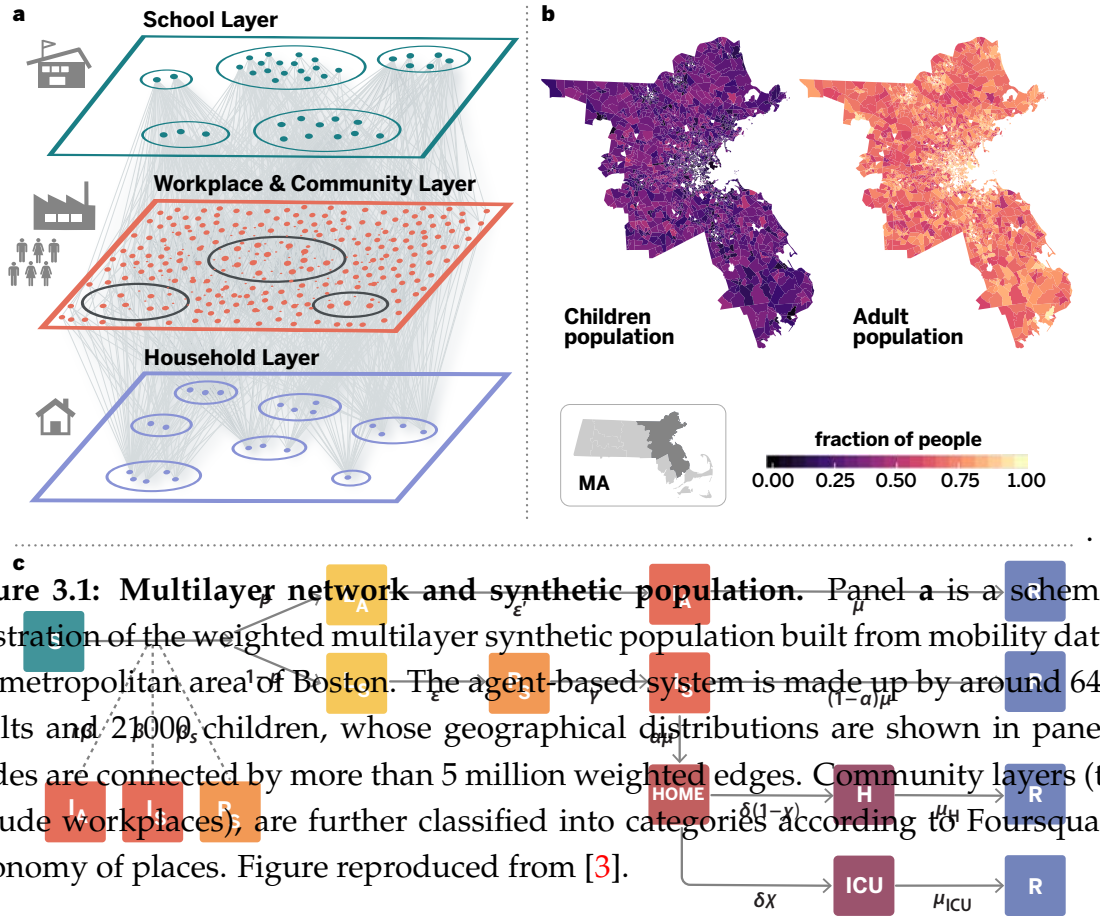

Figure 3.1: Multilayer network and synthetic population. Panel **a** is a schematic illustration of the weighted multilayer synthetic population built from mobility data in the metropolitan area of Boston. The agent-based system is made up by around 64000 adults and 21000 children, whose geographical distributions are shown in panel **b**. Nodes are connected by more than 5 million weighted edges. Community layers (that include workplaces), are further classified into categories according to Foursquare's taxonomy of places. Figure reproduced from [3].

layer represented the households of each anonymous individual. Using the home census block group of each anonymous user we associated each individual to a specific household profile based on socio-demographic data at US census block group level [221]. Families were generated by randomly mixing nodes from the community living in the same census block group, following the statistical features of family types and sizes. Finally, a third layer represented the contacts in the schools (i.e., every node represents one synthetic student and has contacts only with other individuals attending the same school).

To study the evolving dynamics of the infection, we implemented a stochastic, discrete-time compartmental model in which individuals transition from one state to the other according to key time-to-event intervals (e.g., incubation period, serial interval, and time from symptom onset to hospital admission) as from available data on SARS-CoV-2 transmission. The natural history of the disease was captured by the epidemiological model represented in Figure 3.4, where we also showed the

transition rates among compartments [187, 220, 222, 223]. The model considered that susceptible individuals (S) become infected through contact with any of the infectious categories (infectious symptomatic (I_S), infectious asymptomatic (I_A) and pre-symptomatic (P_S)), transitioning to latent compartments (L_S) and (L_A), where they were infected but not infectious yet. Latent individuals branched out in two paths according to whether the infection was symptomatic or not. We also considered that symptomatic individuals experience a pre-symptomatic phase and that once they developed symptoms, they could experience diverse degrees of illness severity, from mild symptoms to being hospitalized (H) or in need of an intensive care unit (ICU) [224]. Finally, individuals transitioned in the removed compartment (identifying recovered or dead individuals). The model assumed a basic reproductive number $R_0 = 2.5$, which is the number of cases directly generated by one case, which together with the rest of the parameters yields a generation time $T_g = 6.6$ days, which is the time interval between the infections of the infector and infectee in a transmission chain. We considered a 25% fraction of asymptomatic individuals. We report the full set of parameters used in the model in *Appendix B*, Figure B.1. For an extensive sensitivity analysis of the ABM model, which is out of the scope of this thesis, see the original *Supplementary Materials of the article* [3]. The model was not calibrated to account for the specific evolution of the COVID-19 epidemic in Boston as it was aimed at showing the effect of different NPIs rather than providing a forensic analysis of the outbreak in the BMA.

3.3 Data & Methods

In this section we detail how we constructed the contact patterns from mobility data, points of interest (POIs) and socio-demographic data from the Boston metropolitan area. We also explain the epidemic ABM model used, and how we implemented the social distancing strategies using the contact matrices defined from the mobility data.

Mobility data

The mobility data was obtained from Cuebiq, a location intelligence and measurement company. The dataset consists of anonymized records of GPS locations from users that opted-in to share the data anonymously in the Boston metropolitan

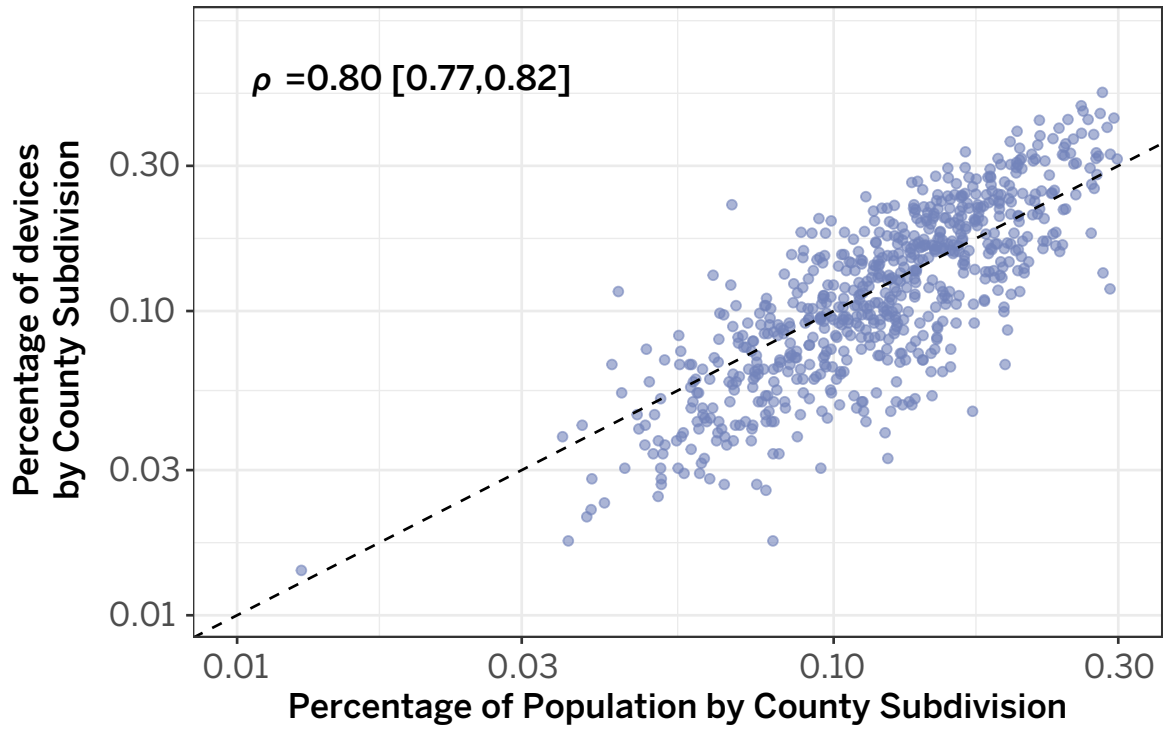

Figure 3.2: Mobility data and population representativeness. The correlation between the population for each county subdivision and the number of devices in our dataset. Figure reproduced from [3].

area over a period of 6 months, from October 2016 to March 2017. Data was shared in 2017 under a strict contract with Cuebiq through their Data for Good program where they provide access to de-identified and privacy-enhanced mobility data for academic research and humanitarian initiatives only. All researchers were contractually obligated to not share data further or to attempt to de-identify data. Mobility data was derived from users who opted in to share their data anonymously through a General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) compliant framework. Our sample from devices was very representative of the population in the Boston area. As we can see in Figure 3.2, population and number of anonymous devices detected in the real data by census area are highly correlated: $\rho = 0.80$ (Pearson correlation) with a CI between 0.77 and 0.82 for county subdivisions.

Points of Interest

We used a dataset of 86k POIs in the BMA collected using the Foursquare API. Those POIs were categorized using the Foursquare taxonomy of places which has ten main categories. We used the following eight principal categories: Art & Entertainment (4.4%), Colleges & Universities (4.8%), Food (16.7%), Nightlife Spots (3.9%), Outdoors & Recreation (10.6%), Workplaces (23.7%), Shops & Services (29.1%) and Travel & Transport (6.4%) (Table 3.1). There are also 638 subcategories, see [225] for a complete list of them.

Stays

From the combination of the mobility data and the POIs we extracted the "stays", as the unique places where anonymous users stayed (stopped) for at least 5 minutes. Each device frequently broadcasts its location to a central server by sending its latitude, longitude, device ID, and the exact date and time of the event. When a person spent significant time at a single location, measurement uncertainty caused a number of events to be scattered around the actual location. To map these events to a single stay with an accurate time and location, we used the Infostop algorithm [226]. First, to extract the locations of stays, the algorithm clusters consecutive events together if the locations are less than 25 meters apart. The location of this cluster is computed by taking the median of the latitudes and longitudes. Moreover, to better estimate the location of places that were visited frequently by the same user, the algorithm also checks whether different clusters appear within 25 meters of each other and assigns a single consistent location to all connected clusters by recomputing the median latitude and longitude. Finally, a stay is registered whenever at least two subsequent events were registered at one of these locations where the first and last event, respectively, mark the start and end time of the stay. The minimum duration of a stay was set to 5 minutes to make sure we were only including actual contact between people instead of people that, for example, pass each other at an intersection.

For privacy reasons, our data was obfuscated around home and workplaces to the level of Census Block Groups (CBGs). Thus the attribution between the mobility data and home and workplaces happened at the level of CBGs and not specific POIs. We estimated the home CBG of the anonymous users as the one in which they were more likely located during nighttime. This resulted in a dataset of the places people stayed

including the POIs in the community layer, the CBG of their workplaces that anonymous users visited, and the most likely CBG of where the device owner lived.

Network structure

Agents. Our population consisted of two different sub-populations, adults and children. Adults were sampled from anonymous individuals in the mobility data collected by Cuebiq, each adult was associated with a home location assigned to a US CBG which was provided by our location data provider. We used those anonymous individuals to construct synthetic populations by assigning them different socio-demographics using highly detail macro (census) and micro (survey) data. We used this procedure to create synthetic representative households and demographic traits as documented in [227]. With this data we designed a population building pipeline that consisted of three steps.

- First step, we built synthetically a set of households, their size and the presence of children based on our adult population and the US Census [221] tables B11016 (Household Type by Household Size) and B11003 (Family Type by Presence and Age of Own Children)
- Second step, we assigned adults to households and in case of presence of children we generated them up to reach the size of the household assigned in the first step.
- And final step, we assigned ages to nodes using table B01001 (Ages by Sex) of age distribution within the CBG.

This process generated our synthetic population consisting of 85k agents (2% of the population in the Boston Metropolitan Area), 64k (75%) of them are adults and 21k (25%) are children. Age groups are distributed as follows: 6,027 (7%) agents for the age group between zero and five years old, 16,250 (18.9%) agents for the age group between six and eighteen years old, 36,207 (42,2%) agents for the age group between nineteen and fifty years old, 13,176 (15,4%) agents for the age group between fifty one and sixty five years old, and final group, 13,945 (16,2%) agents for the age group between sixty six and older. All of agents together formed 43,167 households distributed as follows: 23,293 (53.9%) households with only one agent, 7,886 (18.2%)

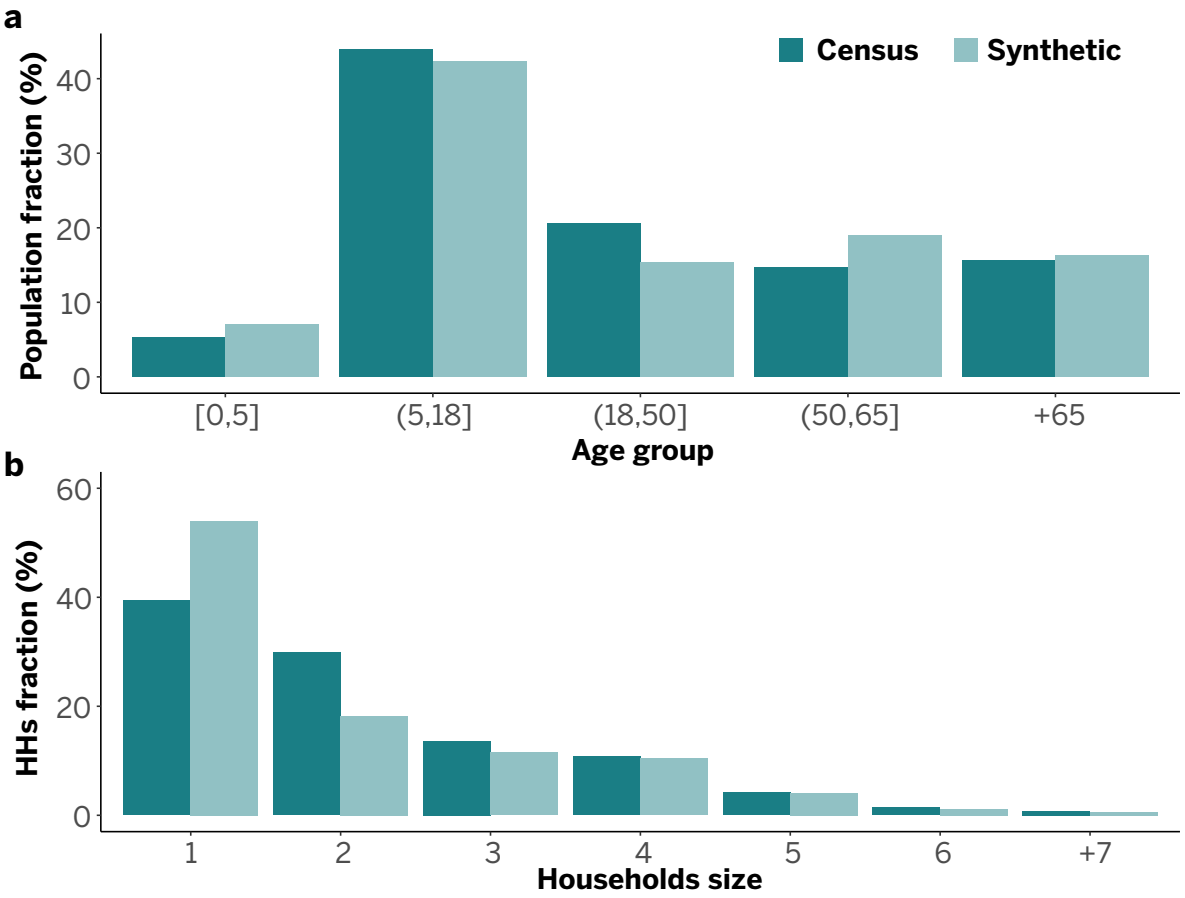

Figure 3.3: Synthetic population representativeness. Age groups and households demographics compared against US Census data. (a) Age groups distribution. (b) Households size distribution. Figure reproduced from [3].

with two agents, 4,959 (11.4%) with three agents, 4,486 (10.4%) with four agents, 1,784 (4.1%) with five agents, 514 (1.2%) with six agents, and finally, 245 (0.5%) with seven agents. In Figure 3.3 we can see the comparison of our synthetic population against census data.

Contacts. Visits to different POIs were used to estimate the contacts between anonymous users. Although the mobility dataset we used was large, co-location events between individuals were quite sparse. Because of this sparsity, and to protect individual privacy in our analysis, we adopted a probabilistic approach to measure co-presence (and probability of transmission) in all locations mapped in the dataset.

Our objective was to build the contact matrix ω_{ij} between individuals i and j using those estimations of co-presence in the different layers where those contacts were possible, Home, Schools, Workplace, and Community.

In order to explain better our approach let us consider the homogeneous mixing approach in a contact network perspective. We assume to have N individuals who are homogeneously mixed. This implies that each individual is potentially in contact with anybody else. Thus, we have a connection $\omega_{ij} = 1$, among each pair of nodes that belong to the same group, neighborhood or city. This implies that the rate of contacts c_i for the individual i is $c_i = \sum_j m\,\omega_{ij} = m(N-1)$, where m is an appropriate factor ensuring that the number of average effective contacts per individual unit time in the system is equal to κ. This implies that

$$\kappa = N^{-1}\sum_i c_i = N^{-1}\sum_{i,j} m\,\omega_{i,j} \tag{3.1}$$

yielding

$$m = \frac{\kappa}{N^{-1}\sum_{ij}\omega_{ij}} = \frac{\kappa}{N-1} \tag{3.2}$$

This finally provides the usual expression for the rate of contact $\omega_{ij} = \kappa/(N-1)$, that is multiplied by the transmissibility per contact α to give the rate (or probability) of infection per contact. This finally leads to the force of infection of a susceptible as

$$P_{S\to I} = 1-(1-\frac{\alpha\kappa}{N-1})^I = 1-(1-\frac{\beta}{N-1})^I \simeq \frac{\beta I}{N}, \tag{3.3}$$

where $\beta = \alpha\kappa$ is the transmissibility used in homogeneous model and the last approximations is valid for very large N. This expression is the traditional homogeneous-mixing result that appear in simple SIR traditional models.

In order to go beyond the homogeneous assumption, from our data we can consider that individuals who were never visiting the same places were never in contact. This is additional information of which we were certain. So for each individual we can list each of the places p that they visit and assume that we can have a link between two individuals if they have the same place in their list $\omega_{ij}^p = \delta_{i,p}\delta_{j,p}$, where $\delta_{i,p} = 1$ if the place p is on the list of visited places of individual i and zero otherwise. This step improved on the homogeneous assumption as it ruled out possible contacts among individuals that could never meet. Furthermore, we can

consider that the potential contacts among individuals are larger for individuals that can meet in more than one place. We can then define $\omega_{i,j} = \sum_p \omega_{i,j}^p$, thus considering that some individuals have more potential contacts. It is worth remarking that we are still considering that each potential contact has the same weight as in the homogeneous assumption. In order to define properly the contact rate/probability per unit time we need to use Eq. (3.1) thus defining

$$m = \frac{\kappa}{N^{-1}\sum_{i,j}\omega_{ij}} = \frac{\kappa}{\langle \omega_{ij} \rangle} \tag{3.4}$$

where we defined $\langle \omega_{i,j} \rangle$ as the average weighted contacts among individuals. This yields the effective rate of contact among individuals i and j as

$$\omega'_{ij} = \frac{\kappa \sum_p \delta_{i,p}\delta_{j,p}}{\langle \omega_{ij} \rangle} \tag{3.5}$$

In order to improve further on this approach we can consider that places are not visited in a probabilistic way. This implies that each individual has a probability to visit a specific place that is $1/n_{i,p}$, where $n_{i,p}$ is the number of places visited by the individual i in a given period. We can therefore define

$$\omega_{ij} = \sum_p \frac{1}{n_{i,p}} \frac{1}{n_{j,p}}. \tag{3.6}$$

This approach still considers potential contacts only among individuals however with a weight that depends on the variability of places of each individual. As before the rate/probability of contact would be:

$$\omega'_{ij} = \frac{\kappa \sum_p n_{i,p}^{-1} n_{j,p}^{-1}}{\langle \omega_{ij} \rangle} \tag{3.7}$$

So far we did not consider at all the time spent in each location. We can therefore improve on the probability to be in a place by weighting the number of places $n_{i,p}$ by the time spent on average in each place. This finally leads to the expression:

$$\omega_{ij} = \sum_p \frac{T_{i,p}}{T_i} \frac{T_{j,p}}{T_j} \tag{3.8}$$

where $T_{i,p}$ is the time spent by individual i at location p and T_i is equal to the sum of all time spent in places in the community by individual i. In this case the rate of interaction will be:

$$\omega'_{ij} = \frac{\kappa \sum_p \frac{T_{i,p}}{T_i} \frac{T_{j,p}}{T_j}}{\langle \omega_{ij} \rangle}. \tag{3.9}$$

This is the expression we used in our work. It is important to stress that this expression improves on the homogeneous assumption as it considers that effective contacts can occur only in places visited by both individuals, and considers that each contact is weighted by the probability for each individual to be in that place. The approach however did not account for concurrency of visits. In this respect it is still adopting an homogeneous perspective in that all places visited at any time corresponds in a potential contact. For this reason we decided to work with the approach of Eq. (3.9), for which all the assumptions can be clearly stated and provided an obvious improvement with respect to the fully homogeneous assumption.

The next steps to improve on this approach would be indeed to consider concurrency of visits. It is thus tempting to consider that each contact is weighted by $T_{i,p}/T$, where T would be the specific amount of time of the day. One could assume the 8 hours of the working time or the 24 hours cycle of the day. This is a tempting solution but introduces a number of issues. For instance the time that should be considered in the normalization depends on the places and restaurants have specific bracket of times during the day, and concurrency should be evaluated on specific hours of the day and specific days (for instance the week-end). The same was for places like movie theatres, museums etc. Furthermore, during the lockdown the concurrency normalization was re-evaluated to be consistent in their definition as the number of hours in the community of the population drastically changed. In other words, we were not sure if the simple normalization by a fixed number of hours although trying to capture the concurrency of contacts was actually introducing unwanted and uncontrolled biases. For this reason we decided to work with the approach of Eq. (3.9), for which all the assumptions can be clearly stated and provided an obvious improvement with respect to the fully homogeneous assumption.

Using our probabilistic approach to detect contacts, we built our contact network in each of the layers:

1) Community weighted contact network. The community network was based on estimation of co-presence of two devices in POIs visited by the anonymous users. We had approximately 6 months of data observation in the Boston area from anonymized users. In this layer each agent in our synthetic population represented an anonymous individual of the real population. The data allowed us to understand how infection could propagate in each layer by estimating co-location of two individuals in the same setting. Specifically, the weight , $\omega_{C_{ij}}$, of a link between individuals i and j within the workplace plus community layer was computed according to the expression:

$$\omega_{ij}^{C} = \sum_{p}^{n} \frac{T_{ip}}{T_i} \frac{T_{jp}}{T_j}, \qquad \forall i, j \tag{3.10}$$

where T_{ip} is the total time that individual i was observed at place p and T_i is the total time that individual i was observed at any place set within the workplace plus community layer. Since agents were representative of the different census areas and groups of the Boston area, our probabilistic approach was a good proxy for the real probability of co-presence between those groups/areas when networks were scaled up to the total population of the Boston area, that was approximately 4,628,910 inhabitants. Finally, for robustness and computational reasons, we included only links for which $\omega_{ij}^{C} > 0.01$.

2) Household weighted contact network. We first identified individuals' approximate home place as their most likely visited census block group at night. Then we assigned a synthetic representative household and demographic traits as documented before in Section 3.3. To assign weights, we assumed that the probability of interaction within a household was proportional to the number of people living in the same household (well-mixing). Therefore, the weight, ω_{ij}^{H}, of a link between individuals i and j within the same household is given by:

$$\omega_{ij}^{H} = \frac{1}{(n_h - 1)} \tag{3.11}$$

where n_h is the number of household members. This fraction was assumed to be the same for all individuals in the population.

3) School weighted contact network. To calculate the weights of the links at the school layer, we mixed together all children that lived in the same census tract.

Interactions were considered well-mixed, hence, the probability of interaction at a school is proportional to the number of children at the same school. Therefore, the weight, ω_{ij}^{S}, of a link between children i and j within the same school is given by:

$$\omega_{ij}^{S} = \frac{1}{(n_s - 1)} \tag{3.12}$$

where n_s is the number of school members.

This process yielded a network with a total number of 5,029,888 unique daily contacts, 3,924,694 (78%) of them in the community layer obtained using the mobility data, 160,748 (3%) and 944,446 (19%) are synthetically built for the household and school layers, respectively.

Calibration of intra-layer links. Within each connected component of the network in each layer (e.g., a household, a school), the links between nodes were weighted to account for the effective daily number of contacts. For example, if we consider a school, while a student can potentially contact all her/his schoolmates, she/he only meet a relatively small fraction of them on a daily basis as estimated in empirical studies on mixing patterns [228, 229]. To account for this, we calibrated the weight of the links in each layer of the synthetic network [230] so that the mean number of daily contacts matches the estimation provided in Mistry et al. [231] (see *Appendix B, section B.1* for more details). Based on the analysis of contact survey data from 9 countries [228, 229, 232, 233], these studies estimated the mean number of daily contacts at 10.86, 4.11 and 11.41 in the community+workplace, household and school layers, respectively.

Stochastic simulations of the COVID-19 dynamics. Using the contact network, we simulated an ABM epidemic model for COVID-19. We described the SARS-CoV-2 transmission process using a discrete-time Susceptible-Latent-Infected-Removed (SLIR) stochastic model, with some extra compartments to incorporate the special characteristics of SARS-CoV-2 infection, Figure 3.4. In particular, at each time-step t (1 day), the infectious asymptomatic (I_A), infectious symptomatic (I_S) and pre-symptomatic (P_S) individuals could transmit the disease to susceptible (S) subjects with probability $r\beta$, β and β_S, respectively. If the transmission is successful, the susceptible node will move to the latent asymptomatic state (L_A) with probability

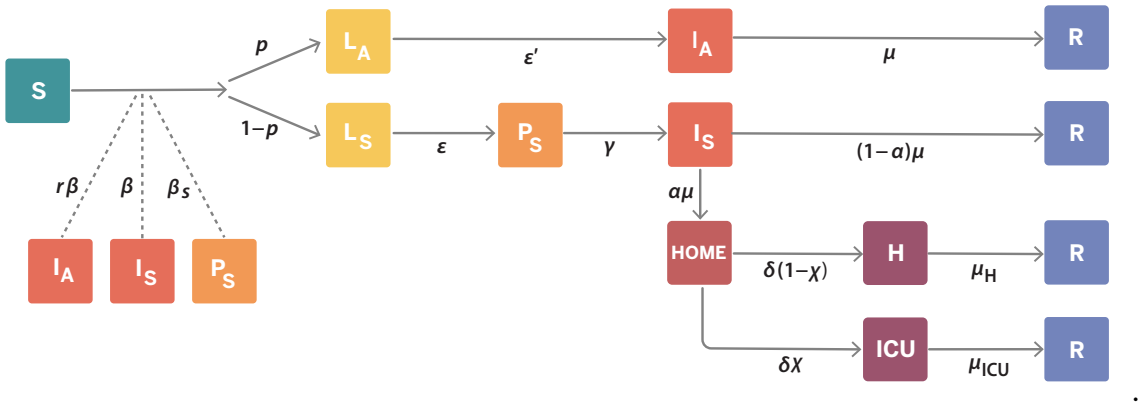

Figure 3.4: COVID-19 compartmental model. Panel displays the compartmental model used to describe the natural history of the disease as well as the transition rates between the different states. Specifically, we consider Susceptible (S), Latent asymptomatic (L_A), Latent symptomatic (L_S), Pre-symptomatic (P_S), Infectious asymptomatic (I_A), Infectious symptomatic (I_S), Hospitalised (H), Hospitalized in intensive care (ICU) and Recovered (R) individuals. More details of the model and the transitions between compartments are provided in *Appendix B*. Figure reproduced [3]

p or to the latent symptomatic state (L_S) with probability $(1-p)$. A latent asymptomatic individual becomes infectious asymptomatic after a period $(\epsilon^{'})^{-1}$, whereas latent symptomatic subjects transition, after a period ϵ^{-1}, to the pre-symptomatic (P_S) compartment. The average period to develop the disease and move to the infectious symptomatic state is γ^{-1}. Infectious asymptomatic nodes will be removed (R) after an average of μ steps. Conversely, infectious symptomatic nodes can either recover after that period with probability $(1-\alpha)$ or, with probability α, these nodes will need hospitalization. It is considered that due to their symptoms they will self-isolate at home after an average period of μ^{-1}. Then, depending on the severity of the symptoms, after a period δ^{-1} the individual will end in hospitalization with probability $(1-\chi)$ or require hospitalization and ICU care with probability χ. Finally, individuals that are either hospitalized or at ICU become removed with probability μ_H or μ_{ICU}, respectively. We initialized the model in the city of Boston by selecting an attack rate on the 17th of March of 1.5% (see *Appendix B, section B.2* for more details about the model parameters).

	Baseline		Medium closure			Non-essential closure		
Layer	Contacts	%.	Contacts	%	% Diff.	Contacts	%	% Diff.
Community	3,924,694	78	1,378,054	27.4	-72.6	357,144	7.1	-92.9
Households	160,748	3.2	160,748	3.2	0	160,748	3.2	0
Schools	944,446	18.8	0	0	-100	0	0	-100
Total	5,029,888	100	1,538,802	30.6	-69.4	517,892	10.3	-89.7

Table 3.1: Social distancing network structure. Number of daily contacts by layer and social distancing policy. Figure reproduced from [3]

Social distancing strategies. To simulate social distancing measures, we modified the synthetic population such that:

- School closures were simulated by removing all the schools from the system simultaneously.
- Partial "stay at home" assumed that all places were open except from restaurants, nightlife and cultural places. Closures of these places were simulated by removing the interactions that occurred in any place that falled into that category according to Foursquare's taxonomy of places. This was the situation after the first reopening.
- Full lock-down and confinement assumed that schools and all non-essential workplaces were closed. Here we closed all workplaces except from essential ones and removed interactions that occured at them. Essential workplaces were: Hospitals, Salons, Barbershops, Grocery Stores, Dispensaries, Supermarkets, Pet Stores, Pharmacies, Urgent Care Centers, Dry Cleaners, Drugstores, Maternity Clinics, Medical Supplies and Gas Stations.

We simulated two different scenarios for social distancing policies. This produced three contact networks: i) *baseline*, ii) *medium closure*, and iii) *non-essential closure*, as we see it in more detail in Table 3.1. Schools were closed in the medium and non-essential closure, but both policies differ in the number of places kept open in the community layer. In Table 3.2 we can see the distribution of POIs, by main Foursquare category, that remained open during each social distancing policy. In the baseline scenario, we kept all the categories and thus the average number of contacts

POIs categories	Baseline Open	%.	Medium closure Open	%	% Diff.	Non-essential closure Open	%	% Diff.
Arts & Entmt.	3,692	4.44	0	0	-100	0	0	-100
Colleges & Univs.	4,016	4.83	4,016	4.83	0	0	0	-100
Restaurants	13,860	16.7	0	0	-100	0	0	-100
Nightlife Spots	3,288	3.95	0	0	-100	0	0	-100
Outdoors & Recr.	8,840	10.64	8,840	10.64	0	229	0.27	-97.4
Workplaces	19,692	23.71	19,692	23.71	0	415	0.5	-97.8
Shops & Services	24,310	29.27	24,310	29.27	0	5,139	6.19	-78.8
Travel & Transp.	5,370	6.46	5,370	6.46	0	0	0	-100
Total	83,608	100	62,228	74.91	-25.6	5,783	6.96	-93.1

Table 3.2: POIs closures by social distancing. Number of POIs open in the Community layer by the different social distancing measures and full non-essential closure. Percentages are calculated with respect to the total number of POIs in the baseline. Table reproduced from [3].

in the community layer was 63 (median 47, [15-150] 90% confidence interval), with few anonymous individuals having a large number of contacts (that could eventually lead to super-spreading events). In the medium closure scenario, POIs in the Art & Entertainment, Restaurants and Nightlife categories were closed; this drastically reduced the average number of contacts to 27 (median 15, [0-92] 90%CI). Lastly, when all non-essential places were closed, we only kept open the following subcategories: Hospital, Salon / Barbershop, Grocery Store, Dispensary, Supermarket, Pet Store, Pharmacy, Urgent Care Center, Dry Cleaner, Drugstore, Maternity Clinic, Medical Supply, and Gas Station. In this situation, the average number of contacts was reduced to 6 (median 0, [0-29] 90%CI). The distribution for the number of contacts in the community layer in these three scenarios is shown in Figure 3.5.

3.4 Results

To provide a baseline of the COVID-19 impact in the BMA, we first investigated an unmitigated scenario in which no interventions were implemented. Figure 3.6.a shows the evolution of the estimated number of new severely affected patients who require hospitalization and admission into ICUs. At the peak of the unmitigated

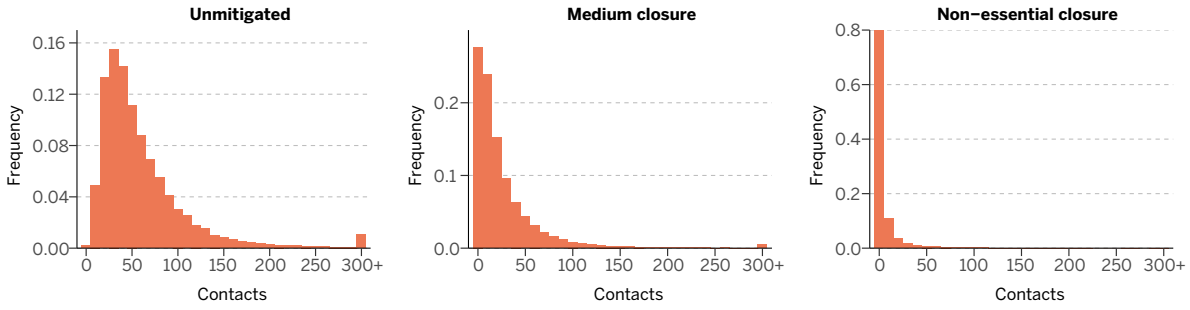

Figure 3.5: Degree distribution by social distancing policy. Degree distribution in the community layer under normal conditions, soft social distancing measures and full non-essential closure. Figure reproduced from [3].

epidemic, the number of ICU beds needed exceeds by far the available capacity (dashed horizontal line in Figure 3.6.a) by more than a factor of 10, thus indicating that the health care system would suffer large service disruptions, resulting in additional deaths due to hospitals overcrowded with patients with COVID-19 [234]. It is worth noting that estimated fatality rates at the beginning of the pandemic considered the general availability of ICU beds and critical care capacity. If this would not be possible, the fatality rate may increase dramatically. We did not report fatality estimates as it went beyond the scope of our analysis and should have considered specific data on the BMA, as well as changing medical treatment and therapeutics the over course of the pandemic.

To avoid the harmful effects of an unmitigated COVID-19 epidemic, governments and policy makers across the world relied on the introduction of aggressive social distancing measures. In the United States, as of April 15 2020, it was estimated that more than 95% of the population was under a "stay at home" or "shelter in place" order [235, 236]. To model the social distancing policies implemented in the whole BMA, we considered March 17, 2020 as the average starting date of social distancing policies that included school closures, the shut down of all non-essential work activities as well as mobility restrictions. This scenario mimicked the social distancing intervention implemented in most of the high income countries, in Europe and across states in the US. Such extreme social distancing policies came with very large economic costs and social disruption effects [237], thus prompting the question of what exit strategy could be devised to restart economic activities and normal

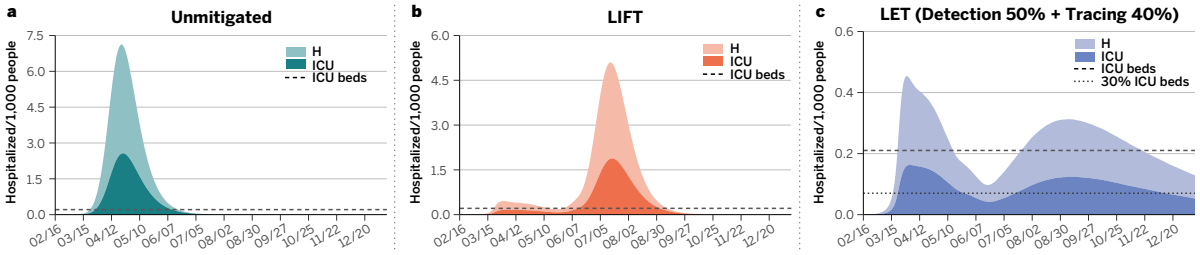

Figure 3.6: Impact on the Boston health care system. Estimated number of individuals per 1,000 inhabitants that would need hospitalization (H), and intensive care (ICU) for each of the three scenarios considered in Figure B.2. Panel **a** corresponds to the unmitigated situation, whereas results for the LIFT and LET strategies are shown in panels **b** and **c**, respectively. The horizontal dotted-dashed lines represent the ICU basal capacity of the Boston health care system. The dotted line in panel **c** indicates 30% of the ICU basal capacity. Figure reproduced from [3].

societal functions [238]. For this reason, we explored two different scenarios for lifting social distancing interventions:

- **Lift scenario (LIFT):** the "stay at home" order was lifted after 8 weeks by re-opening all work and community places, except for mass-gathering locations such as restaurants, theaters, and similar locations (see *Table 3.2*). The latter partial re-opening was enforced for another 4 weeks, which was followed by a full lifting of all the restrictions that remained. We considered that schools will remain closed given the impending summer break in July and August, 2020. In fact, some school systems, like the Boston Public schools remained closed through the 2019-2020 school year.

- **Lift and enhanced tracing (LET) scenario:** The "stay at home" order was lifted as in the previous scenario. Once partial reopening was implemented, we assumed that 50% of symptomatic COVID-19 cases could be identified for SARS-CoV-2 infection, on average, within 2 days after the onset of symptoms and that they were isolated at home and their household members were quarantined successfully for 2 weeks (a sensitivity analysis for lower rate of isolation and quarantine was presented in the original *Supplementary Materials [3]*). Although COVID-19 tests were highly specific, 50% detection accounted also for imperfect testing. We also assumed that a fraction of the

non-household contacts (we show results for 20% and 40%) of the symptomatic infections could be traced and quarantined along with their household as well – note that we considered that the contacts were identified with a rate proportional to the time duration of the interaction with the symptomatic individual.

The above scenarios were mechanistically simulated on the multilayer network of Figure 3.1.a, by allowing different interactions (between effective contacts) according to the simulated strategy. As a result, the average number of interactions in the W+C layer went from 10.86 (95% C.I.:1.51-42.39) under the unmitigated scenario, to 4.10 (95% C.I.:0-23.79) for the partial lock-down and only 0.89 (95% C.I.: 0-8.39) contacts for the stay at home policy (see *Figure 3.5*). This result were in agreement with previously published work [239] and reports in the New York City area [240]. It is worth remarking that the fluctuations in the number of contacts in the stay at home order were due to a large extent to contacts that take place in grocery stores and other public venues.

The numerical results showed that the LIFT scenario, while able to temporally abate the epidemic incidence, did not prevent the resurgence of the epidemic and a second COVID-19 wave when the social distancing measures were going to be relaxed. Indeed, at the time of lifting the social distancing intervention the population had not achieved the level of herd immunity that would protect it from the resurgence of the epidemic. It is important to stress that here we did not consider additional mitigation measures such as behavioral changes in the population, mask wearing, etc (see *Appendix B SARS-CoV-2 transmission model* section B.2). We also estimated that a second wave of the epidemic still had the potential to infect a large fraction of the population and to overwhelm the health care systems, as shown in Figure 3.6.b. The number of ICU beds needed, although half the unmitigated scenario, was still exceeding by far the estimated availability, as pointed out in similar scenario analysis [217–219, 241]. This suggested that lifting social distancing without the support of additional containment strategies was not a viable option.

In the case of the LET scenario, the lifting of the social distancing intervention wanted along with a significant amount of contact tracing and precautionary quarantine of potentially exposed individuals. The quarantine was not limited to the contacts of the identified symptomatic COVID-19 case, but extended to their

Scenario		Hospitalization	ICU	People traced
Unmitigated		4.57 (4.10-5.03)	2.56 (2.21-2.91)	-
LIFT		3.22 (2.80-3.67)	1.87 (1.55-2.20)	-
3***LET** 3*Detect 30%	No Tracing	2.70 (2.29-3.12)	1.58 (1.27-1.88)	-
	Tracing 20%	0.86 (0.65-1.10)	0.55 (0.39-0.72)	0.52 (0.36-0.69)
	Tracing 60%	0.35 (0.21-0.50)	0.22 (0.12-0.34)	0.17 (0.08-0.27)
3***LET** 3*Detect 50%	No Tracing	2.35 (1.97-2.75)	1.39 (1.11-1.68)	-
	Tracing 20%	0.44 (0.28-0.62)	0.28 (0.16-0.42)	0.39 (0.23-0.55)
	Tracing 40%	0.29 (0.18-0.43)	0.15 (0.08-0.26)	0.14 (0.05-0.23)

Table 3.3: Social distancing strategies effectiveness. Mean and 95% C.I. of the number of normal hospitalizations, ICU hospitalizations and symptomatic individuals identified/traced (when applicable) at the peak of the epidemic per 1000 people. The estimated availability of ICU beds is 0.21 beds per 1000 people. Table reproduced from [3].

households. This strategy amounted to a simplified tracing of contacts of contacts, that would not require extensive investigations. In other words, this strategy did not require the tracking of a large number of single contacts but leverage on the contacts' households as the basic unit [242]. Households could be monitored though, with daily calls or messages to ascertain the onset of symptomatic infections, and provide medical support as needed.

When 40% or more of the contacts of the detected symptomatic infections were traced and they and their households quarantined, the ensuing reduction in transmission leaded to a noticeable flattening of the epidemic curve and appeared to limit the possible resurgence of a second epidemic wave effectively. It is also worth noticing that we assumed the absence of other additional and minimally disruptive social distancing policies such as crowd control, smart working, wearing of masks, etc., that could lead to a further reduction of the transmissibility of the virus with respect to our estimates. It is important to stress that the contact tracing proposed here worked at the level of household unit, simplifying also the monitoring and follow up process, by contacting only one member of the household to monitor the onset of symptoms among all members. Figure 3.6.c and Table 3.3 show the burden in hospitalization and ICU demand in the unmitigated situation and the two mitigation scenarios. The LET scenario allowed relaxation of the social distancing interventions while maintaining the hospital and ICU demand at levels close to the health-care

availability and surge capacity.

3.5 Discussion

The efforts in the suppression and mitigation of COVID-19 were pursuing the objectives of preserving the health care system from disruptive failures due to overwhelming stress imposed by the large number of severe cases, and of minimizing the morbidity and mortality related to the epidemic. The aggressive social distancing interventions implemented by many countries in response to the COVID-19 pandemic appear to have achieved the interruption of transmission and the abatement of the epidemic, although at the price of huge societal disruption and economic costs. In such a context, the identification of “exit strategies" that allowed restarting economic and social activities while still protecting the healthcare systems and minimizing the burden of the epidemic is of primary importance. Several modeling studies already pointed out that resuming economic activities and social life was likely to lead to a resurgence of the COVID-19 epidemic, and combined social distancing interventions of different degrees and intensity had been proposed to substantially delay and mitigate the epidemic [237, 241]. These interventions generate economic loss and widespread disruption to social life. In this chapter we show how testing, contact tracing strategies at scale, based on home isolation of symptomatic COVID-19 cases and the quarantine of a fraction of their households' contacts, had the potential to provide a viable course of action to manage and mitigate the epidemic when social distancing interventions were progressive lifted [243, 244]. These strategies presented us with logistic challenges that included large-scale and rapid diagnostic capacity, and a large surge in the number of contact tracers. We investigated what fraction of the population would be isolated/quarantined under the proposed contact tracing and isolation strategy. Figure 3.7.a shows the fraction of households that needs to be quarantined. Assuming the identification of 50% of the symptomatic infections, and tracing of 40% of their contacts and households, only about 9% of the population would be quarantined at any time. While this is certainly a relevant fraction of the population, it was a much better option if compared with massive social distancing policies affecting the entire population that last for months.

In Table 3.3, we reported the number of symptomatic infections for which the

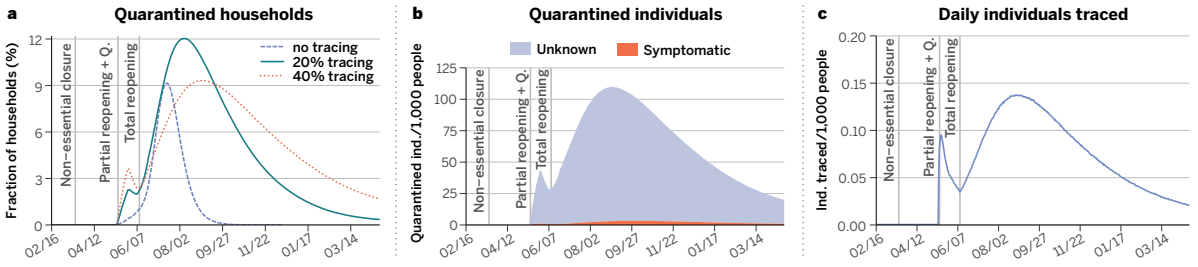

Figure 3.7: Affordability of the best way-out scenario. LET strategy with 50% detection and 40% tracing. (**a**) Fraction of the population that needs to be put under quarantine as a function of time and percentage of contact tracing. (**b**) Health status of the individuals that are quarantined for a contact tracing level of 40%. Note that only symptomatic individuals are tested, which implies that a large fraction of the quarantined population is of unknown status. This fraction of individuals quarantined with unknown health condition could be reduced if the capacity to do more tests increases. As it is shown, the pandemic might span over several months depending on the level of contact tracing. (**c**) Number of individuals whose contacts are traced each day per 1,000 persons. Relevant intervention actions are signaled by the vertical dashed lines in all panels. Figure reproduced from [3].

contact tracing investigation should have been performed in the basic scenarios. This number provided an estimate of the contact tracers per 1,000 individuals. It is important to note that the more effective the contact tracing starting from each individual, the smaller the number of generally traced households because the epidemic had lower incidence rates. In addition, as illustrated in Figure 3.7.b, the health status of the vast majority of quarantined individuals was unknown as contact tracing did not imply testing. The curves in Figure 3.7.a constitute the upper bounds for each simulated case. If we assumed that the capacity to do massive testing would have likely ramped up, then it was expected that the actual number of people in quarantine could be significantly lowered by testing the quarantined household. This would have also alleviated the burden on household members that could not go to work and increase compliance with quarantine for positive cases. It is also worth remarking that many of the logistic challenges faced with massive contact tracing could possibly be eased by digital technologies that were investigated across the world following the examples of COVID-19 response in Asian countries [244]. Also, it was difficult to quarantine the entire household of individuals who were

potentially exposed, since this is a hardship suffered with great uncertainty about their risk of infection. Offering other logistic quarantine solutions (quarantine centers, hotel rooms) could significantly raise the rate of compliance.

These results were obtained under several assumptions at the beginning of the pandemic. There were very large uncertainties around the transmission of SARS-CoV-2, in particular, the fraction of sub-clinical and asymptomatic cases and their transmission. Estimates of age-specific severity are informed from the analysis on individual-level data from China and other countries, and were subject to change as more US data become available. We also did not include specific co-morbidities or pre-existing conditions of the specific BMA population. For this reason, in the *Supplementary Materials* of the article [3] we performed an extensive sensitivity analysis showing that the modeling results discussed here are robust to the plausible range of parameter values for the key time-to-event intervals of COVID-19 (e.g., incubation period, serial interval, and time from symptom onset to hospital admission, etc.) as well as the fraction of presymtpomatic and asymptomatic transmission. We were also not considering here potential changes to the virus transmissibility due to environmental factors, in particular, seasonal drivers such as temperature and humidity. The modeling does not consider possible reintroduction of SARS-CoV-2 in the population from infected travelers. Strategies based on testing, isolation and contact tracing might be hampered by the importation through travel of a large number of infections, thus travel restrictions and screening may need to be introduced to/from places that show sustained local transmission. Finally, we also reported in the original *Supplementary Materials* of the article [3] the effect of the widespread use in the population of masks or other personal protective equipment that lead to a reduction of the transmissibility of SARS-Cov-2 during 2020. These active protection measures improved the effectiveness of the exit strategies modeled here.

The modeling of the impact of testing, contact tracing, and isolation on second-wave scenarios of the COVID-19 epidemic played a crucial role in shaping public health response planning for national and international agencies. Our research demonstrated the effectiveness of contact tracing and household quarantine at scale, even under the assumption of a complete lifting of social distancing measures. However, decisions regarding the timing and duration of policy relaxations were

marred by chaos and inadequate information for several reasons. Firstly, the absence, antiquity or delayed development of large-scale epidemiological monitoring systems during the epidemic impeded the timely collection and analysis of critical data, leading to decision-making challenges [245, 246]. Secondly, public resistance to adhering to policies due to perceived threats to individual freedoms complicated efforts to implement effective containment measures [247, 248]. Moreover, resources constraints and political interests further complicated the relaxation of policies, creating inconsistencies and inefficiencies in the public health response [249]. On the one hand, encouraging smart working from home for individuals capable of adhering to it without significant disruptions proved successful and was adopted by millions of people [250, 251]. On the other hand, contact tracing apps faced limitations in adoption and effectiveness due to privacy concerns, limited interoperability, technology limitations, and low adherence by individuals [252]. Based on our results, we found that with just 40% adoption of contact tracing apps, they could have effectively contained second-wave epidemic peaks. Lastly, in this chapter, we introduce a novel probabilistic approach for constructing contact networks using mobility data. This method enables direct observations of human health-related behaviors, elevating the modeling and quantification of viral epidemics. By leveraging this approach, we contribute to a deeper understanding of epidemic dynamics and their broader impacts. This research is a significant step forward in advancing our knowledge of infectious disease spread, informing evidence-based public health policies, and empowering agencies to make informed decisions in future public health emergencies.

4

Temporal Contact Networks: Unveiling the Spread of Viral Agents and Detecting Super-Spreading Events through Mobility Data

"Time crumbles things; everything grows old under the power of Time and is forgotten through the lapse of Time."

– Aristotle[1]

4.1 Introduction

SOCIAL interactions are dynamic over time and exert a significant influence on the transmission dynamics of infectious diseases among human populations. In recent years, epidemiological contact network models have emerged as valuable tools for comprehending and predicting pathogen transmission, including that of SARS-CoV-2. However, given the rapid feedback between the evolution of the epidemic and the behavior of people, we need to have better temporal, almost real-time models of how those contact networks evolve in time. By leveraging longitudinal mobility data from real-world sources, we can construct temporal

[1] Aristotle. Greek philosopher and polymath. Quote extracted from the book 'Physics, IV, 12'

contact networks that enable more realistic simulations of disease spread. Additionally, these temporal contact networks are important in an viral epidemic, like COVID-19, because they offer valuable insights into how a population responds and change their behaviours to Non-Pharmaceutical Interventions (NPIs) and the flow of news about the epidemic, while also capturing realistic social interactions, assessing transmission dynamics, identifying super-spreading events, and exploring the impact of social distancing policies on bending the epidemic curve over time.

In this chapter, we build upon the methodology proposed in chapter 3, where we knew how long individuals spent in different locations. Now, we introduce an innovative approach to construct temporal social contact matrices at individual and daily level. By utilizing temporal mobility data, we integrated changes in population behavior into an agent-based model (ABM) that effectively simulated the spread of infectious diseases within a given geographical area. We leveraged real-time, privacy-enhanced mobility data from the New York City and Seattle metropolitan areas, enabling us to develop a detailed ABM of SARS-CoV-2 infection and estimate the location, timing, and magnitude of transmission events during the initial wave of the COVID-19 pandemic.

Through our comprehensive analysis, we uncovered insightful findings. We discovered that a mere 18% of individuals were responsible for a significant majority (80%) of infections, with approximately 10% of events classified as super-spreading events (SSEs). While large gatherings pose a considerable risk for SSEs, our research demonstrated that the bulk of transmission occurs in smaller events within settings such as workplaces, grocery stores, or food venues. Remarkably, the specific locations driving transmission patterns evolved throughout the course of the pandemic and exhibit variations between different cities, underscoring the profound influence of behavioral factors.

This innovative approach grants us a granular understanding of viral spread dynamics and provides real-time insights into the actual effectiveness of non-pharmaceutical interventions (NPIs) in curbing transmission. By continuously informing us about the efficacy of these interventions, we gained valuable knowledge for effectively controlling the spread of viral agents. Our modeling approach, coupled with case studies and epidemiological data, suggested that real-time tracking of transmission events could inform the evaluation and formulation of

targeted mitigation policies.

In the subsequent sections, we present an updated version of the article *Building temporal human contact networks from mobility data* [4], where we delve into the details of our methodology and present our key findings in greater depth.

4.2 Background & Hypotheses

Without effective pharmaceutical interventions, the COVID-19 pandemic triggered the implementation of severe mobility restrictions and social distancing measures worldwide aimed at slowing down the transmission of SARS-CoV-2. From shelter in place orders to closing restaurants/shops or restricting travel, the rationale of those measures is to reduce the number of social contacts, thus breaking transmission chains. Though individuals may remain highly connected to household members or close contacts, these measures reduced the connections in the general community that allowed the virus to move through the network. Some venues may attract more individuals from otherwise unconnected networks, or may attract individuals who are more active and thus have greater exposure. Understanding how interventions targeted at particular venues could impact transmission of SARS-CoV-2 could help us devise better NPIs that pursue public health objectives while minimizing disruption to the economy, the education system, and other facets of everyday life.

Although it is by now clear that NPIs have helped to mitigate the COVID-19 pandemic [131], most of the evidence is based on measuring the subsequent reduction in the case growth rate or secondary reproductive number. For example, econometric models were used to estimate the effect of the introduction of NPIs on the secondary reproductive number [253, 254]. Other studies showed directly (through correlations or statistical models [255]) or indirectly (through epidemic simulations [256, 257]) the relationship between mobility or individuals' activity and number of cases. Unfortunately, most of the data used so far did not have the granularity required to assess how social contacts and SARS-CoV-2 transmission events were modified by NPIs [258].

Analysing the Impact of NPIS and Detecting Super-Spreading Events using Human Mobility Data and ABM models

This is especially important given the heterogeneous spreading of SARS-CoV-2. Overdispersion in the number of secondary infections produced by a single individual was an important characteristic of the 2003 SARS pandemic [259] and has been similarly observed for SARS-CoV-2 [260]. Several drivers of super-spreading events (SSEs) were proposed: biological, due to differences in individuals' infectiousness; behavioral, caused by unusually large gatherings of contacts; and environmental, in places where the surrounding conditions facilitate spread [261]. Transmissibility depends critically on the characteristics of the place where contacts happen, with many SSEs documented in crowded, indoor events with poor ventilation. A characteristic of this overdispersion is that most infections (around 80%) are due to a small number of people or places (20%), suggesting that better targeted NPIs or cluster-based contact tracing strategies could be devised to control the pandemic [262]. Although several studies provided insights on SSEs [258, 263], given their outsized importance for SARS-CoV-2, we needed better information about where, when, and to what extent these SSEs happen and how they could be mitigated or amplified by NPIs.

We used individual-level mobility data of over half a million individuals distributed in the New York and Seattle metropolitan areas during the months of February 2020 to June 2020 to estimate the day and type of venues where people interact. To do that we extracted from the mobility data the stays (stops) of people in a large collection of around 440k settings. With this information we built two synthetic populations, one for each metropolitan area, in which agents can interact in different settings: workplaces, households, schools, and the community (points of interest). We then explorec the transmission of SARS-CoV-2 using a compartmental and stochastic epidemic model applied on top of this population, which allowed us to track infections at the individual level.

The behavioral changes induced in the population by the introduction of several NPIs were naturally encoded in this high resolution mobility data, allowing us to characterize the effect of these interventions. We ran counterfactual simulations of our stochastic epidemic model to understand that effect. Furthermore, the resolution of this data allowed us to characterize the spreading through different types of

venues at different stages of the epidemic, depicting a complex picture in which the combination of both the characteristics of the place/setting and of the behavior of individuals who visited it determine its risk.

Lastly, the information at the individual level allowed us to study the frequency and characteristics of behavior-related super-spreading events (SSE). We studied the likelihood of finding a SSE per setting as a a function of time by looking at the number of infections produced by each individual in each location. A full description of the materials and methods is provided in the following section.

4.3 Data & Methods

While this chapter's methodology shares similarities with that of chapter 3, it also incorporates several significant differences. Firstly, the mobility data collected was specifically obtained from two different geographical regions, the New York and Seattle metropolitan areas during the period of analysis, in contrast to section 3.3, which was only for the Boston metropolitan area prior to the period of analysis. Moreover, the taxonomy of POIs mentioned in section 3.3 was manually curated to a more suitable version for this study. Additionally, the contact matrices in this chapter possess a temporal component, making them dynamic and longitudinal in nature, embedding social real distancing strategies in the mobility behaviors of observed individuals, and there is no need for simulating social distancing strategies. Furthermore, these contact matrices exhibit significantly larger sizes in terms of nodes and edges, setting them apart from the static and smaller contact matrices discussed in section 3.3. Finally, the Workplace & Community layer from chapter 3 was split into two more layers, the Workplace layer and the Community layer, and a different normalization of the data was built to approximate better the synthetic populations to the US census households size statistics. Consequently, in this chapter, we provide a comprehensive overview of the methodology once again, placing particular emphasis on the differences and crucial modifications made to accommodate the inclusion of temporal data.

Here we use a longitudinal database of detailed mobility and socio-demographic data to generate the daily contacts of 565k individuals in the New York metropolitan area and 106k individuals in the Seattle metropolitan area, during the period from

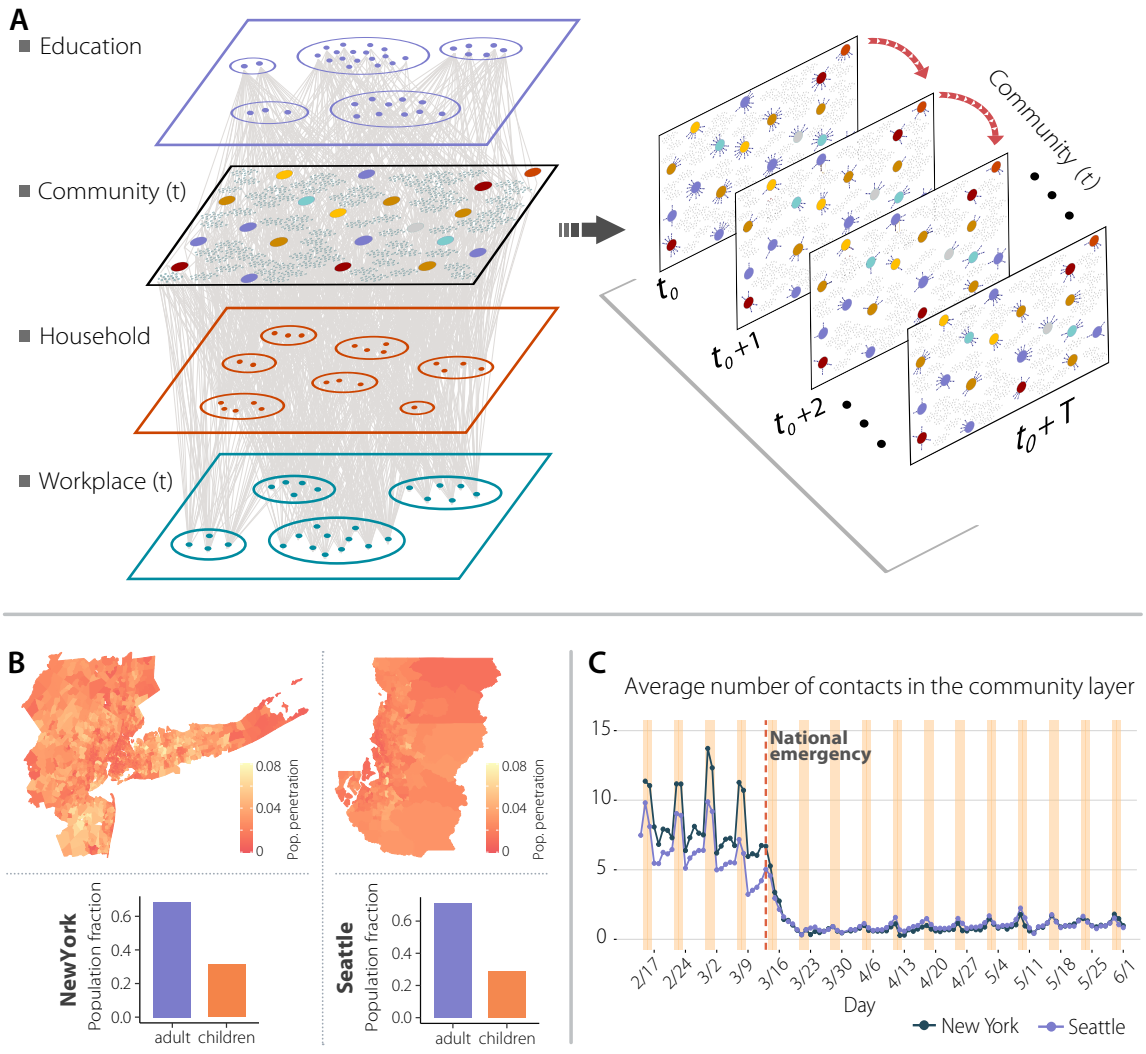

Figure 4.1: Network components, New York and Seattle metropolitan areas population and social contacts dynamics at the community layer over time. Panel **a** is a schematic illustration of the weighted multilayer and temporal network for our synthetic population built from mobility data. There are four different layers; the school and household layers are static over time, and the combined workplace and community layers have a daily temporal component. Panel **b** shows the geographic penetration of mobile devices from our mobility data compared to the total population for the New York and Seattle metropolitan areas. Panel **c** represents the average daily number of contacts in the community layer for both metropolitan areas. Figure reproduced from [4].

February 17 to June 1 of 2020. Note that the metropolitan areas considered extend beyond the city limits for both locations. We selected these areas because of their large differences in COVID-19 epidemiology, population size and density. The NY metro area has a population of 20 million people, while the Seattle metro area has 3.8 million inhabitants. Moreover, the NY metro area has a higher density (5,438 people per km^2, median by census tract) than Seattle (1,576 people per km^2). Finally the number of reported COVID-19 cases/deaths during the study period in the NY area was very large (223 per 100,000) compared to that in the Seattle area (24 per 100,000). Individuals were chosen to be representative of the different census areas (Census Block Groups (CBGs), see Figure 4.1.b). Contacts between individuals were weighted according to the likelihood of exposure between them in the different places around the metro areas. This defined a weighted temporal network consisting of four layers representing the physical/social interactions occurring in (1) the community, (2) workplaces, (3) households, and (4) schools, see Figure 4.1.a. The community and workplaces layers were generated using 4 months of data observed in the New York and Seattle metropolitan areas from anonymized users who opted-in to provide access to their location data, through a GDPR-compliant framework provided by Cuebiq, the same we mentioned in section 3.3. In these layers, each individual in our synthetic population represented an anonymous individual of the real population.

The data allowed us to understand how infection propagated in each layer by estimating co-location of individuals in the same setting at any given time. Settings were obtained from a large database of 375k locations in the New York and 70k in the Seattle from the Foursquare public API, the same as in chapter 3. By measuring the amount of time people were co-located in the different layers, we constructed the time-varying network of interactions ω_{ijt} between individuals i and j on the same day t in the education, community, work and household layers. Estimation of co-location in the community layer was done by extracting stays of users to the settings using different time and distance. Our results were independent of the particular choice of minimal time (5 minutes or 15 minutes) and maximum distance to the setting (10 meters or 50 meters), see Figure 4.1 and the following sections. Our model covered all possible interactions in urban areas and not just foot traffic to commercial locations that people visit [258], something especially important given the relevant role of households, schools or workplaces in the transmission of the

SARS-CoV-2. On the other hand, it is important to note that the underlying data did not provide a direct measurement of contacts between individuals and the nature of these contacts (masked/unmasked, with conversation). Rather, data were used to extrapolate the locations visited by each subject and the amount of time they spent there, in order to relax the homogeneous mixing assumption commonly used in mathematical modeling approaches, as we stated in chapter 3.

Geographic areas

Our sample dataset achieved broad geographic representation for our two populations, in the New York and Seattle metropolitan areas, defined as the Core Based Statistical Areas (CBSA) by the US Census [264]. This provided a self-contained metropolitan area in which people move for work, leisure or other activities. Some of the CBSAs we considered span several states, as opposed from chapter 3 where we only used data from the same state. For instance, the New York CBSA contains areas of the state of Connecticut, New Jersey, Philadelphia, and New York. We filtered all anonymous devices which were not observed each month, in order to make sure we had a stable population with enough granularity and representativeness of agents over the whole period. The population and number of anonymous devices detected in the real data by census area were highly correlated for both census county subdivision regions, with a $\rho = 0.796$ (Pearson correlation) with a CI between 0.783 and 0.807 for the New York region, and a $\rho = 0.948$ (Pearson correlation) with a CI between 0.937 and 0.957 for the Seattle region. We built such correlations between the population for each county subdivision and the number of devices in our dataset. Despite these large correlations our mobility dataset had a small income bias towards areas of higher income, specially in the NY metro area. However, as shown in *Appendix C.3*, our results did not depend on that bias.

Points of Interests

We used a dataset of 375k Points of Interest (POI) in the New York metropolitan area and 70k Points of Interest in Seattle metropolitan area collected using the public Foursquare API. In section 3.3 we used the eight main categories from Foursquare taxonomy, however, in this chapter we manually curated every subcategory in the taxonomy to be reassigned to twelve new principal categories: Arts & Museums,

College, Entertainment, Exercise, Food & Beverages, Grocery, Health, Other, Outdoors, School Service, Shopping and Transportation. In our database the New York metropolitan they were distributed as follow: Art & Museum (2.1%), College (2.9%), Entertainment (7.6%), Exercise (2.8%), Food & Beverage (17.7%), Grocery (2.6%), Health (7.5%), Other Places (13.1%), Outdoors (8.2%), School (2.3%), Service (16.6%), Shopping (8.3%), Sport & Events (0.6%) and Transportation (6.9%). For the Seattle metropolitan area POIs were distributed as follows: Art & Museum (2.7%), College (2.3%), Entertainment (7.1%), Exercise (2.7%), Food & Beverage (14.5%), Grocery (2.1%), Health (8.1%), Other Places (15.1%), Outdoors (7.8%), School (1.6%), Service (18.2%), Shopping (8.3%), Sport & Events (0.8%) and Transportation (7.8%). Despite our dataset contained many venues and places which were companies or businesses, some evidence that our dataset covered most of the public places came by comparing them to official statistics: for example, we had 2,155 art galleries in the NY metro area compared to the 1,500 estimation for NY City only. On the other hand we had 9,810 groceries in the NY metro area in our POI database which compares quite well with the 11,791 grocery business reported by the U.S. Bureau of Labor Statistics in their Quarterly Census of Employment and Wages in the NY Metro area [265].

Stays

For a detailed explanation of how we calculated stays and obfuscated data around home and workplaces, please refer to the *Data & Methods from chapter 3, section 3.3*. In this chapter, we expand upon our approach by conducting sensitivity analysis on stays distance thresholds, considering that some stays occur within or in close proximity to POIs. We attributed a stay to the closest POI up to a distance of 50 meters, otherwise that stay is discarded. We did not make this attribution if the closest place is further than 50 meters (see *Appendix C.3* for a sensitivity analysis with other maximum distance to POIs). Although we used 50 meters as an upper bound, in reality the average distance to the attributed POI is much smaller, 19.43 meters on average in the metro areas of NY and Seattle, which is smaller than the average distance between nearest POIs. In areas with large numbers of POIs like Manhattan, the distance to the attributed closest POI is even smaller. Note that we attributed each stay to a single POI and in turn, to a single category of place. We also checked that our results did not depend significantly on the 50 meters threshold for the attribution

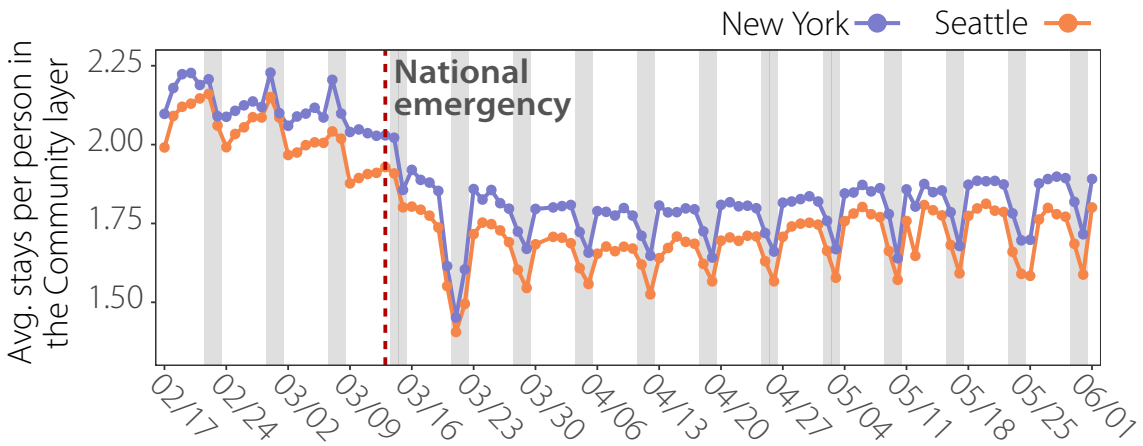

Figure 4.2: Average number of stays. Evolution of the average number of stays in the Community layer per observed person for New York and Seattle metropolitan areas. Vertical red dashed line indicates when National Emergency (N.E.) is established. Figure reproduced from [4].

of the stays (see *Appendix C.3*). Stays are then aggregated at POI level.

In Figure 4.2 we can see the daily evolution of the average number of stays per observed person for New York and Seattle only in the community layer. Also in Figure 4.3 we can see the total observed number of stays in our datasets. Two weeks before we can see that Seattle started to see a small change in the mobility behaviour, however, for New York City we can start to see that pattern one week before the national emergency. The average number of daily stays per agent for New York before the N.E. is 2.14 with a 95% CI [2.12, 2.17]. On the other hand, for Seattle is 2.05 with a 95% CI [2.02, 2.08]. After the national emergency there is an abrupt decrease for both cities in the number of stays (see Figure 4.3). Two weeks after the national emergency the average number of stays per person stabilized and starts to an slightly and steady increase. Eleven weeks after the national emergency, the average number of stays per person has recovered slightly, but it did not recover its basal state for both cities. The average number of daily stays per observed agent for New York after the N.E. is 1.83 with a 95% CI [1.81, 1.84]. On the other hand, for Seattle is 1.72 with a 95% CI [1.71, 1.74].

We can see in Figure 4.3 the daily evolution of the total number of stays to each category and their fraction distribution. Figure 4.3 (a) for New York and (c) for Seattle represent the total number of stays at the community layer, we can see a similar pattern

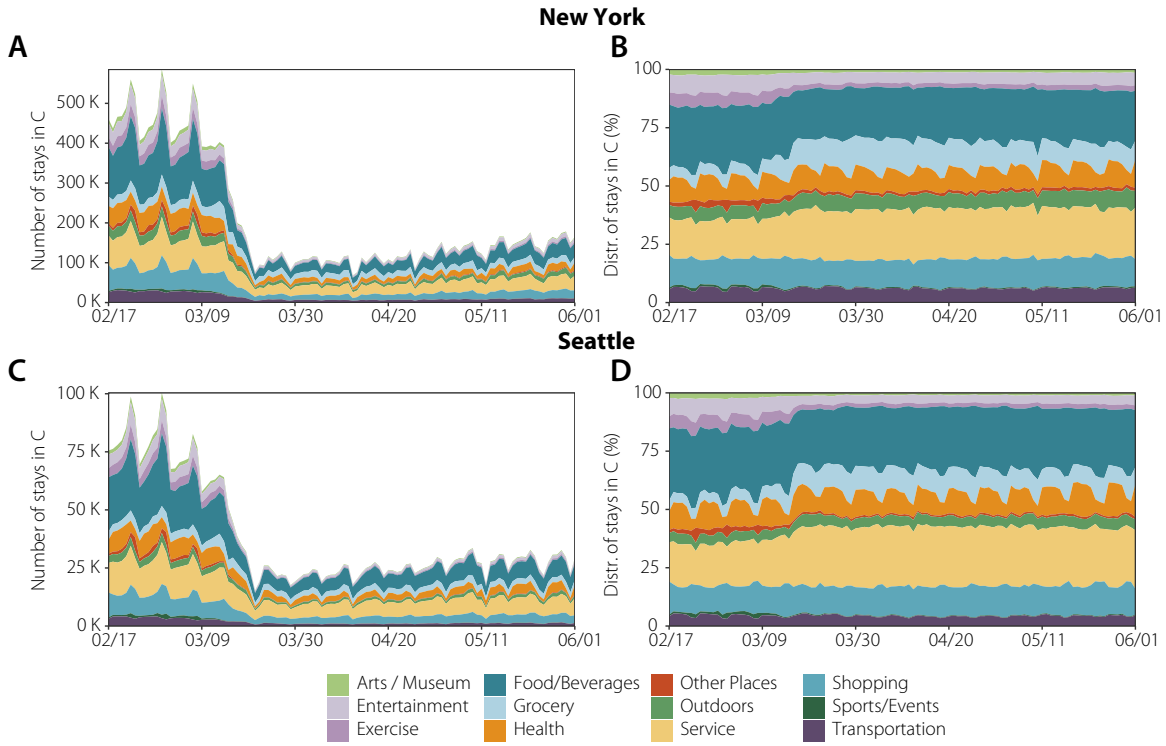

Figure 4.3: Total number of stays per POI category. The comparative evolution of the number of stays (left) and distribution (right) of stays in the Community layer for the different metropolitan areas, New York (top) and Seattle (right). Figure reproduced from [4].

as in Figure 4.2 (a) before and after the national emergency. Figure 4.3 (b) for New York and (d) for Seattle show normalized number of stays. We can see a reduction of non-essential places after the national emergency due to the social distancing policies.

Finally, in Figure 4.4, we can see the comparison of the average time per stay for each city and category before and after the national emergency. There is a significant decrease in time spent per stay for nearly each category in both cities. However, the grocery and the transportation categories are those with the smallest change in the average time for both cities. Moreover, the shopping category does not barely change in New York, but it does in Seattle. On the other hand the Food & Beverages category decrease in New York, but it does not in Seattle.

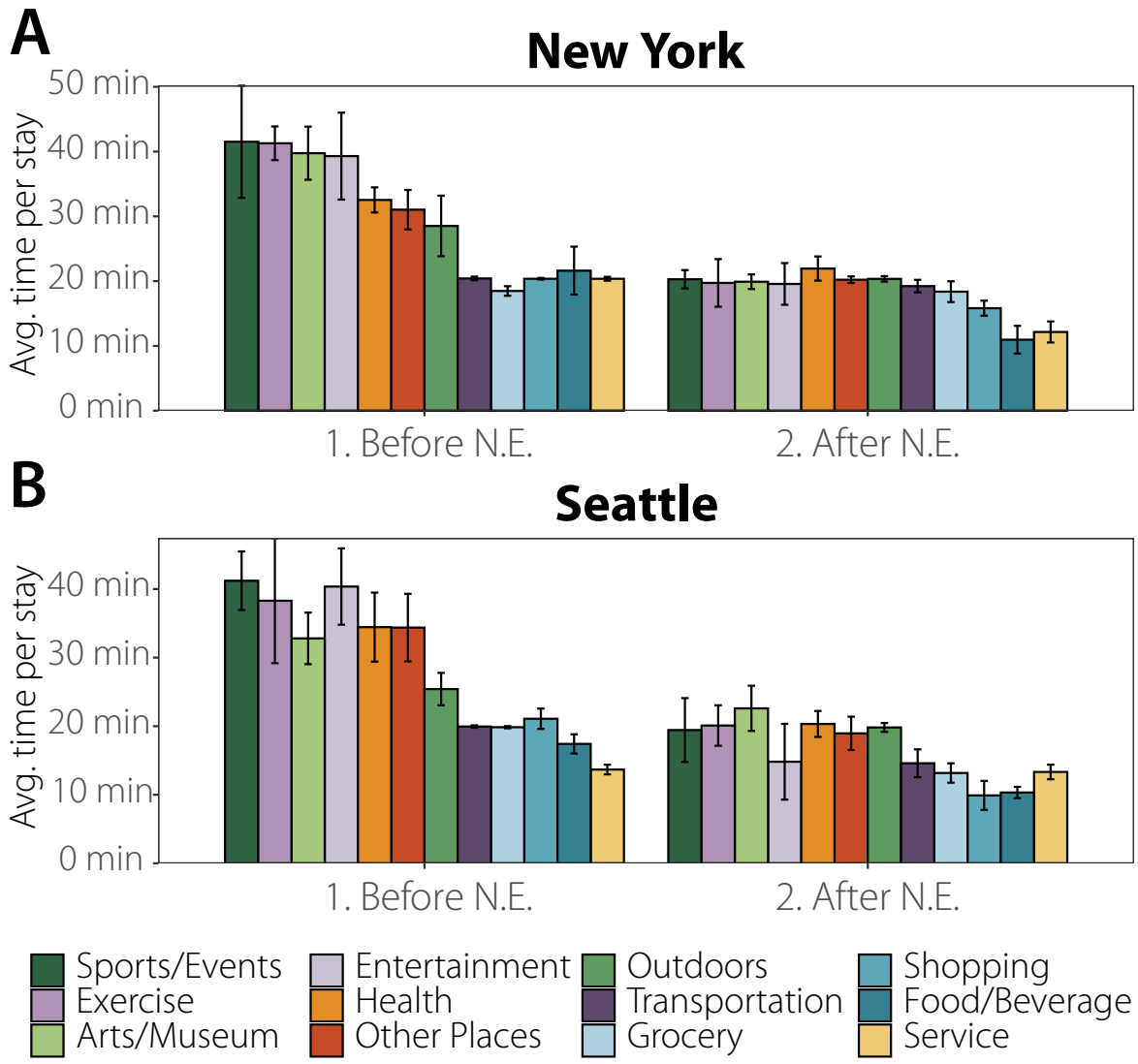

Figure 4.4: Average time per stay per category before and after National Emergency before and after the National Emergency (N.E.). for (a) the New York Metropolitan Area and for (b) the Seattle Metropolitan Area. Figure reproduced from [4].

Network structure

Agents. In the same way as in section 3.3, our population consisted of two different sub-populations, adults and children, assigned to US CBGs to create synthetic representative households and demographic traits as documented in [227].

Following this process we generated two synthetic populations, one for the New York metropolitan area and the other one for the Seattle metropolitan area. The New

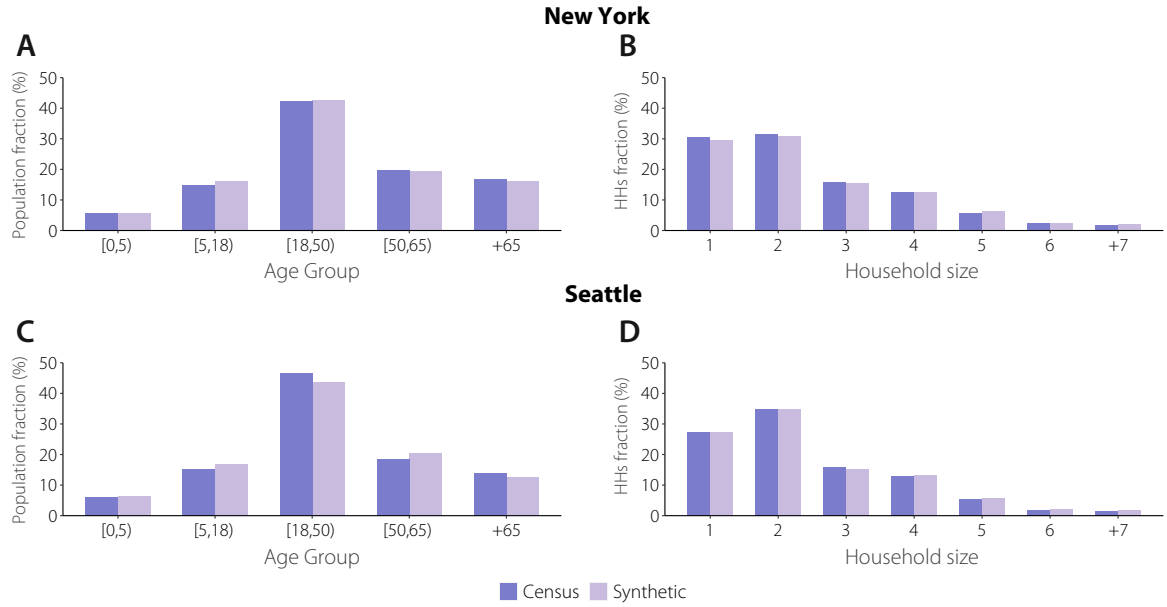

Figure 4.5: Synthetic and census population demographics. Age groups and households demographics compared against the US Census data. (a) Age groups distribution and (b) households size distribution for the New York Metropolitan Area. (c) Age groups distribution and (d) households size distribution for the Seattle Metropolitan Area. Figure reproduced from [4].

York synthetic population consisted of 565k agents (3.0% of the population in the New York metropolitan area), 78.02% of them are adults and 21.98% are children. Distribution of age groups are shown in Figure 4.5.a where we can see the that our synthetic population age distribution compares well against the US census data. The same happens for the household size distribution, where 31% of the households are of size two, 29.5% of size one and the rest are of size three or bigger, see Figure 4.5.b. The Seattle synthetic population consists of 106k agents (2.9% of the population in the Seattle metropolitan area) with 76.7% of them adults and 23.3% are children. Age groups are distributions can be found in Figure 4.5.c where we can see that they compare well with the demographic distribution. Household size distribution is very similar to the NY metro area, with 27.2% of size one, 34.8% of size one and the rest of size three or bigger. In Figure 4.5.d we can see the comparison of our synthetic households population distribution against the US census data.

Contacts. We used our probabilistic approach to detect contacts as described in section 3.3, adding the temporal dimension to it. By revisiting Eq. (3.9), we built our contact network in each of the layers:

- **Community weighted contact network**. In the community layer Contacts were built by estimating co-location of two individuals in the same POI. Eq. (4.1) is an extended version of Eq. (3.10) for day t from chapter 3. Specifically, the weight, ω_{ijt}^{C}, of a link between individuals i and j within the community layer at day t is computed according to the expression:

$$\omega_{ijt}^{C} = \sum_{p}^{n} \frac{T_{ipt}}{T_{it}} \frac{T_{jpt}}{T_{jt}}, \qquad \forall i, j, t \tag{4.1}$$

  where T_{ipt} is the total time that individual i was observed at place p in day t and T_{it} is the total time that individual i has been observed at any place set within the community layer that day t. The distribution of values of ω_{ijt} is very broad. For example in NY ω_{ijt} as a mean of 0.395, a median of 0.279 and 25% and 75% quantiles of 0.095 and 0.653, respectivelly.

  Finally, for robustness and computational reasons, we included only links for which $\omega_{ijt}^{C} > 0.01$, removing 2.88% of the original links. For other values of the threshold like $\omega_{ijt}^{C} > 0.005$ and $\omega_{ijt}^{C} > 0.02$ we would remove 1.19% and 6.19% of the links respectively. Note however that since those links have very small weights, our results for the epidemic spreading did not depend significantly of the threshold chosen provided that it is small.

- **Workplace weighted contact network**. As we mentioned, for privacy reasons, our data was obfuscated around home and workplaces to the level of CBGs. To get a proxy of contacts at the workplace, we assume that all workers in the same CBGs have a probability to interact. To account for the potential number of working places in that area, we weighted that probability by the number of POIs at the same CBG. Therefore, the contact weight, ω_{ijt}^{W}, of a link between individuals i and j within the same workplace at day t is given by:

$$\omega_{ijt}^{W} = \sum_{\alpha \in \text{CBG}} \sum_{\beta \in POI(\alpha)} \frac{\delta_{i\alpha t}}{N_{POI}(\alpha)} \frac{\delta_{j\alpha t}}{N_{POI}(\alpha)} = \sum_{\alpha \in \text{CBG}} \frac{\delta_{i\alpha t}\delta_{j\alpha t}}{N_{POI}(\alpha)}, \qquad \forall i, j, t \tag{4.2}$$

where $POI(\alpha)$ is the set of POIs in the CBG α, $N_{POI}(\alpha)$ is the number of POIs in α, $\delta_{i\alpha t}$ is the binary variable of observing or not an individual at her workplace within CBG α at day t. As before, we included only links for which $\omega_{ijt}^{W} > 0.01$.

- **Household weighted contact network**. To calculate the weights of the links at the household layer, we use Eq. (3.11) from chapter 3. We assumed this layer is static throughout our period.

- **School weighted contact network**. To calculate the weights of the links at the school layer, we use Eq. (3.12) from chapter 3. This layer is removed on March 16th 2020 in both metropolitan areas to account for the imposed school closure.

Calibration of intra-layer links. To calibrate the relative importance of each layer in the spreading process we further multiply the weights by their corresponding κ. In particular, with $\kappa = 4.11$ in the household layer, $\kappa = 11.41$ in the education layer, $\kappa = 8.07$ in the workplace layer and $\kappa = 2.79$ in the community layer [227], see Eq. (3.9)

SARS-CoV-2 transmission model

To model the natural history of the SARS-CoV-2 infection, we implemented a stochastic discrete-time compartmental model on top of the contact network ω_{ijt} in which individuals transition from one state to the other according to the distributions of key time-to-event intervals (e.g., incubation period, serial interval, time from symptom onset to hospital admission) as per available data on SARS-CoV-2 transmission model. In the infection transmission model, susceptible individuals (S) become infected through contact with any of the infectious categories (infectious symptomatic (IS), infectious asymptomatic (IA) and pre-symptomatic (PS)), transitioning to the latent compartment (L), where they are infected but not infectious yet. Latent individuals branch out in two paths according to whether the infection will be symptomatic or not. We also consider that symptomatic individuals experience a pre-symptomatic phase and that once they develop symptoms, they can experience diverse degrees of illness severity, leading to recovery (R) or death (D). The value of the basic reproduction number is calibrated to the weekly number of deaths. For further details on the model, please refer to section 3.3 and for more

detailed information on model parameters, the calibration process, model specifications, and sensitivity analysis of our results, please refer to *Appendix C*.

4.4 Results

Impact of NPIs. Our data clearly shows how the contact networks in the two metro areas changed due to the introduction of NPIs during the week of March 15th to March 22nd 2020, see Figure 4.1.c, A National Emergency was declared on March 13th 2020, and the NY City School System announced the closure of schools in March 16th [266]. NY City Mayor issued a "shelter in place" order in the city on March 17 [267], and non-essential business were ordered to close or suspend all in-person functions in New York, New Jersey and Connecticut by March 22nd. As we can see in Figure 4.1.c the individuals' total number of contacts decreased dramatically from around 7 (in our community layer) to below 2. In Seattle, the reduction of contacts started one week earlier than in NY City, coinciding with earlier closing of some schools [268], and the Seattle mayor issuing a proclamation of civil emergency on March 3rd [269].

In Figure 4.6 we report numerical simulations of the epidemic curve that accurately reproduce the evolution of the incidence of new COVID-19-related deaths in both NY and Seattle metro areas, even though both cities were affected very differently by the epidemic in the first wave. The analysis identifies the impact of the reduction in the number of contacts due to the implemented NPIs: both in the NY and Seattle metro areas, R_t dropped below 1 one week after NPIs were introduced. To estimate the importance of timely implementations of NPIs in metropolitan areas, we generated counterfactual scenarios in which the NPIs and the ensuing reduction in the number of contacts could have happened one week earlier or later than the actual timeline [272]. The comparison between NY and Seattle is relevant, because we observed that the reduction in contacts in Seattle started to happen exactly one week before that in NY. To this end, we shifted in time the contact patterns around the week where NPIs were introduced in both cities. The results for these scenarios are reported in Figure 4.6.d, where we see that a one-week delay in introducing NPIs could have yielded a peak in the number of deaths two times larger than the observed one (0.7 deaths per 1,000 people compared to the 0.35 per 1,000). This doubling in peak deaths following a one-week delay is also observed in the Seattle metro area and in the cumulative infection

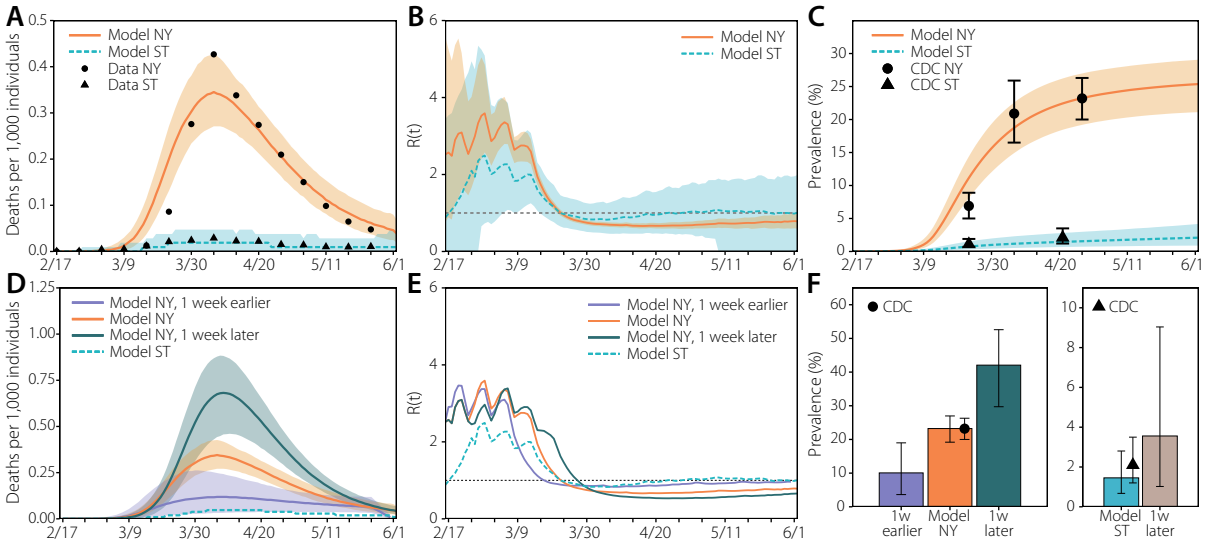

Figure 4.6: Evolution of the first wave. (a) Weekly number of deaths in New York (NY) and Seattle (ST) metro areas. The dots/triangles represent the reported surveillance data used in the calibration of the models. The lines represent the median of the model ensemble for each location and the shaded areas the 95% C.I. of the calibrated model [270]. (b) Evolution of the effective reproduction number according to the output of the simulation. The solid (dashed) line represents the median of the model ensemble and the shaded areas the 95% C.I. of the model. (c) Estimated prevalence in our model (median represented with solid/dashed lines and 95% C.I with the shaded area) and values reported by the CDC (dots/triangles represent New York and Seattle data respectively) [271]. (d) Estimated number of deaths if the NPIs had been applied in New York one week earlier/later. Solid (dashed) lines represent the median of the model ensemble and the shaded areas the 95% C.I. (e) Estimated evolution of the effective reproduction number if the measures had been applied in New York one week earlier/later. Solid (dashed) lines represent the median of the model ensemble. (f) Estimated prevalence in New York (left) and Seattle (right) if the NPIs had been applied in New York one week earlier/later and in Seattle one week later. The height of the bars represents the median of the model ensemble, while the vertical error bars represent the 95% C.I. The dot/triangle shows the value reported by the CDC. Figure reproduced from [4].

prevalence in the metro area. Conversely, a one-week earlier implementation of the NPIs timeline in the NY area could have reduced the death peak by more than a factor of three, a result similar to that found using county-level simulations [272].

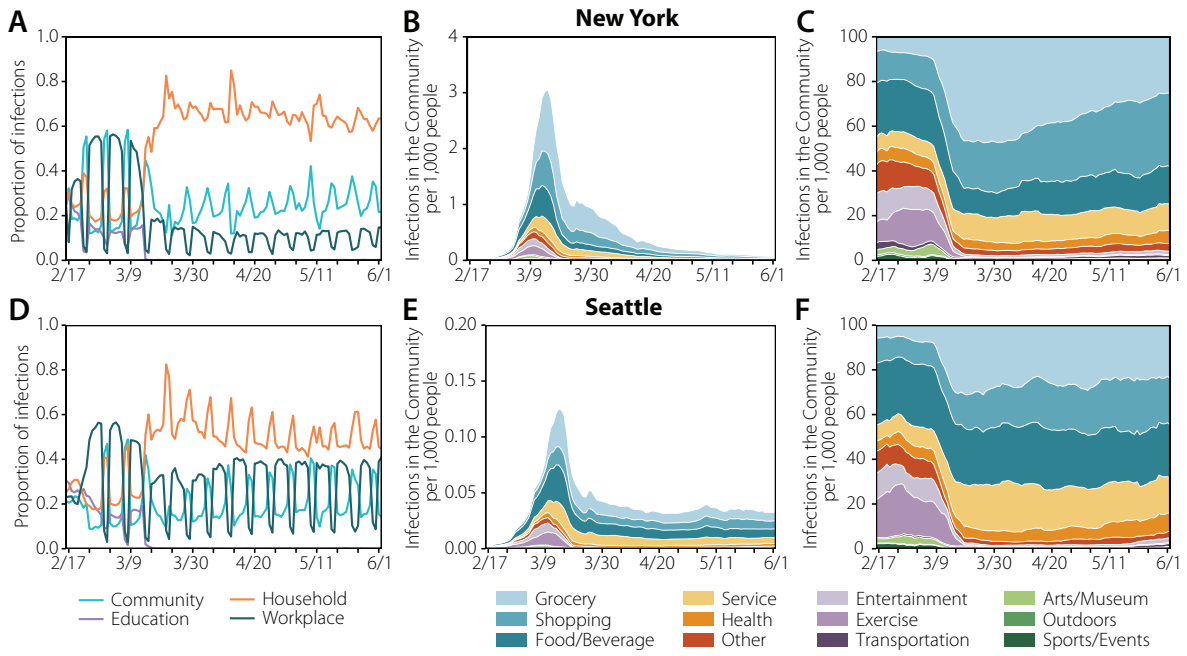

Figure 4.7: Spatial spreading of the disease. The plots in the left column represent the share of infections across layers in New York (a) and Seattle (d). In the middle column, the estimated location where the infections took place for New York (b) and Seattle (e) in the community layer. Note that the y-axis is 20 times smaller in Seattle. The evolution has been smoothed using a rolling average of 7 days. In the right column, the distributions are normalized over the total number of daily infections, showing how infections were shared across categories in the community layer. The evolution has been smoothed using a rolling average of 7 days. Figure reproduced from [4].

Taxonomy of transmission events. The high resolution of our dataset allowed us to estimate the relevance of different settings and the effects of NPIs on the transmission dynamic of SARS-CoV-2. People spent different time in each layer and place before and after the introduction of NPIs (see Figure 4.4). As a result, the number of infections varied significantly during the observed period. As we can see in Figure 4.7, before NPIs were introduced most infections took place in the community and workplace layers. Once restrictions were implemented in both cities on March 16th 2020, as expected, the proportion of infections in the household layer greatly increased, especially in the NY area. In Seattle, the number of infections in the workplace and household layers were comparable, probably because the number of cases overall was lower than in NY. We can further stratify data by venue type in the

community layer as in Figure 4.7, by looking at the estimated top categories (see section 4.3, Points of Interest for their definition) in terms of the number of total infections throughout the whole period. Before the NPIs were introduced, our model estimates that most of the infections in the community layer happened in Food/Beverage, Shopping, and Exercise venues. Also a significant number of infections happened in Art/Museums and Sport/Events venues. After the introduction of NPIs, the number of infections in Exercise, Sport/Events or Art/Museums venues decreases as expected. However, Food, Groceries and Shopping venues became the main community setting for transmission in both cities.

Super-spreading events. Our agent-based simulations also allowed us to follow specifically each individual and how many secondary infections she generated. In Figure 4.8 we reported the distribution of the number of secondary infections produced by each individual in the community layer only. As our model integrated co-location data, this was driven by individual-level differences in activity and those individuals they interacted with. The distribution is highly skewed and can be modeled by a negative binomial distribution with dispersion parameters (k) of 0.16 (NY) and 0.23 (Seattle), in agreement with the evidence accumulated from SARS-CoV-2 transmission data [260, 261, 273, 274]. As a result, super-spreading events (SSEs) are likely to be observed. We define a transmission event as a SSE if the individual infects in a specific location category more than the 99-th percentile of a Poisson distribution with average equal to R (see [259] and *Appendix C* for further details), here corresponding to an infected individual infecting 8 or more others. Interestingly, if we compare the distribution of secondary infections produced before and after the introduction of NPIs, even though we see a clear reduction of SSEs, we still find a heterogeneous distribution of secondary infections. Thus, the NPIs did not prevent the formation of SSEs, but only significantly lowered their frequency.

Consistent with this pattern of over-dispersion in the number of transmission events, we found that the majority of infections is produced by a minority of infected people: $\sim$ 20% of infected people were responsible for more than $\sim$ 85% of the infections in both metro areas (see Figure C.1 in *Appendix C*). However, note that a critical driver here of this phenomenon is that a large majority of infected people (85% in the community layer) did not infect any others in our simulations. Only a

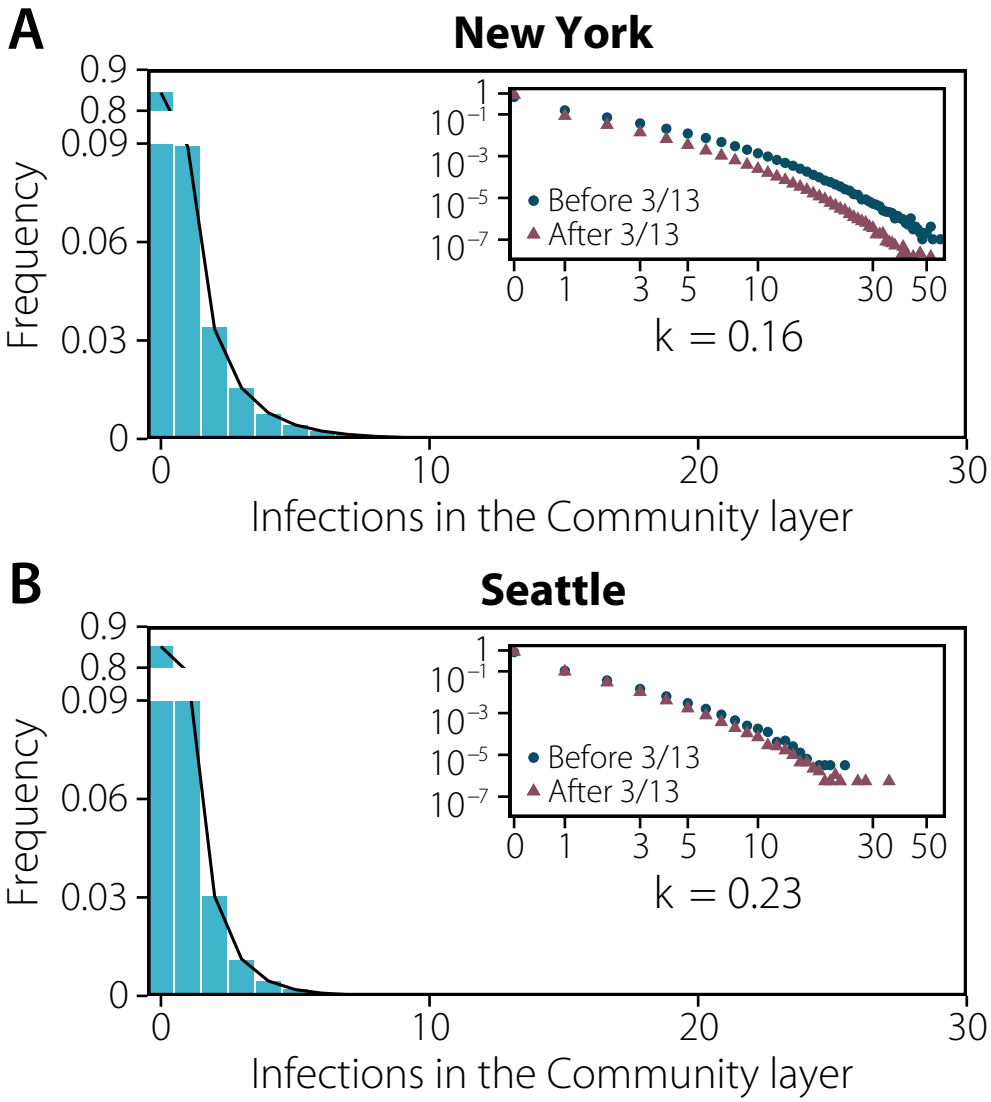

Figure 4.8: Behavioral super-spreading events. Distribution of the number of infections produced by each individual in New York (a) and Seattle (b) up to the declaration of National Emergency. The distribution is fitted to a negative binomial distribution yielding a dispersion parameter of $k = 0.163$ $[0.159 - 0.168]$ 95%CI and $k = 0.232$ $[0.224 - 0.241]$ 95%CI, respectively. In both plots the inset represents the same distribution on the log-scale and distinguishing infections that took place before the declaration of National Emergency on 03/13 and after that date. Figure reproduced from [4].

small fraction of infection events (0.08%) were made of 8 (or more) secondary infections.

Transmission events and SSEs did not happen equally in different settings or along time or geography. In Figure 4.9 we show the results of our simulations for the total number of infections produced in each category and the share of those infections that can be related to SSEs (see also Table C.2 in *Appendix C*). The combination of those two features define a continuous risk map in which places can be at different types of risk: (i) low contribution from SSEs and low contribution to the overall infections, such as Outdoor places; (ii) larger contribution from SSEs but low contribution to the overall infections, such as Sports/Events, Arts/Museum or Entertainment before the introduction of NPIs; (iii) large contribution to the overall infections but with low contribution from SSEs, such as Shopping or Food/Beverage after the introduction of NPIs; and (iv) large number of infections and with large contribution from SSEs, such as Grocery. This classification has important implications from a public health perspective. For instance, venues in (ii) do not have a major contribution to the overall infections but might represent a challenge for contact tracing. Conversely, for categories in (iii) it might be easier to trace chains of transmission but their total contribution is large. Note that this definition is not static, but could change over time due to the NPIs imposed by authorities. Indeed, looking at the weekly pattern of infections (see Figure 4.9) we observed how some categories move to a different quadrant due to the behavior of individuals. Although we estimated that SSEs and infections were more likely in Arts/Museum, Sport/Events in NY, and Entertainment and Grocery in both cities, our simulations showed that Grocery category still greatly contributes to the total number of infections, but did not have as many SSEs after March 16 2020. On the other hand, we estimated that SSEs were rare before March 9 2020 in Seattle, but their contribution doubled in the week of March 9-15 - when many individuals probably went for supplies amid preparation for the future introduction of NPIs. This observation included implicitly a very important message: a place may not be inherently dangerous; rather, the risk is a combination of both the characteristics of the place/setting and of the behavior of individuals who visited it. This suggests revisiting studies that found that settings could play always the same role in the evolution of the pandemic [258].

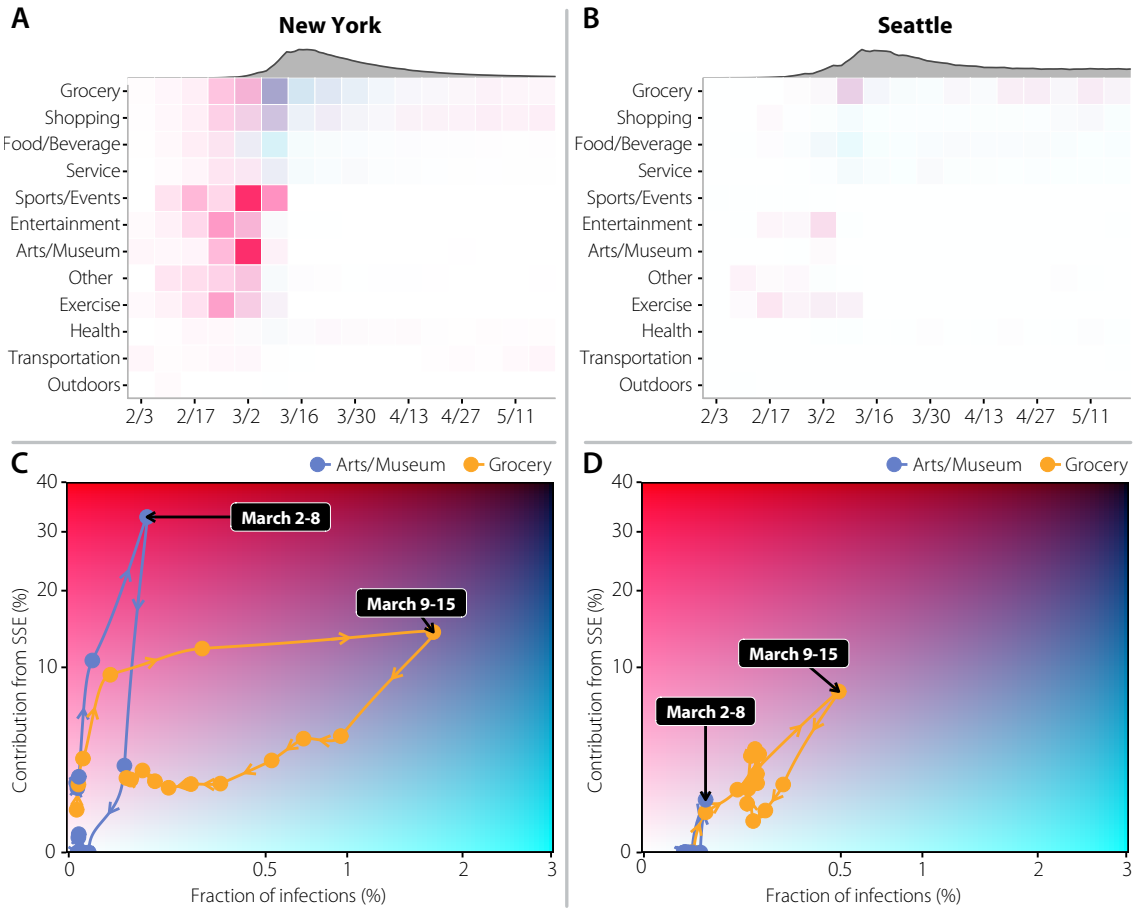

Figure 4.9: Dynamics of super-spreading events (SSE). Risk evolves with time as a function of the behavior of the population and policies in place. A) and B) : risk posed by each category per week, defined using the corresponding map below. As a reference, the gray area on top shows the estimated weekly incidence. C) and D) : the x axis represents the fraction of total infections that are associated with each category, while the y axis accounts for the share of those infections that can be attributed to SSEs in each category. This defines a continuous risk map in which places with few infections and low contribution from SSEs will be situated on the left bottom corner. Places where the number of infections is high but the contribution from SSEs is low are situated in the bottom right corner. Conversely, places with large contribution from SSEs but a low amount of infections are situated on the top left corner. Lastly, places with both large number of infections and an important contribution from SSEs are situated in the top right corner. The color associated to each tile in the top row is extracted from the position of the point in the plane defined in the bottom figure. The points in the bottom row show the evolution of the position of the categories Arts/Museum and Grocery for each week, with the arrows indicating the time evolution. Figure reproduced from [4].

4.5 Discussion

Our results emphasize the intertwined nature of human behavior, NPIs, and the evolution of the COVID-19 pandemic in two major metropolitan areas. Specifically, our results suggest that heterogeneous connectivity and behavioral patterns among individuals lead naturally to differences in risk across settings and the generation of SSEs. In particular, the implemented partial or full closures of different settings (e.g., sport venues, museums, workplaces) had a dramatic effect in shaping the mixing patterns of the individuals outside the household [239, 275]. As a consequence, the settings responsible for the majority of transmission events and SSEs varied over time. In absolute terms, the food and beverage setting is estimated to have played a key role both in determining the number of transmission events and SSEs in the early epidemic phase; however, this setting was among the first targets of interventions and thus its contribution become zero over time because of the introduced NPIs. On the other hand, settings such as grocery stores, which consistently provided a low absolute contribution to the overall transmission and SSEs, became, in relative terms, a source of SSEs during the lockdown when most of other activities were simply not available. These findings suggest that there is room for optimizing targeted measures such as extending working time to dilute the number of contacts or the use of smart working aimed at reducing the chance of SSEs. That could be especially relevant to avoid local flare ups of cases when the reproduction number is slightly above or below the epidemic threshold.

Although the overall picture emerging from studying Seattle and New York was consistent, it is important to stress that each urban area might have specific peculiarities due to local transportation, tourism, or other economic drivers differentiating the cities' life cycle. Our results suggest that a one-size-fits-all solution to minimize the spread of SARS-CoV-2 had very different impact across cities. Furthermore, the results presented may not be generalized to rural areas. Though large parts of the Seattle metro area could be considered as rural, individual connectivity patterns may be differently constrained by the generally lower population density in some other parts of the country.

Our modeling analysis did not have the ambition to substitute field investigations, which remained the primary source of evidence. Some of the reported findings (e.g., the role of food and beverage venues or groceries) appear to be in

agreement with epidemiological investigations [258, 276–279]. Future empirical analyses could provide further validation of our findings. Our modeling investigation was based on real-time data on human mobility/activity that provides an indirect proxy for infection transmission. One of the strengths of this approach is that, differently from epidemiological investigations, the data can be retrieved in real time and longitudinally, thus allowing to quickly capture possible changes in the most relevant settings for transmission. Furthermore, our approach could help minimize the noisy and biased data collection related to massive transmission events [280]. Yet, the approach used here is far from capturing all the finest details of human social contacts and thus the estimates on the contribution of different settings to SARS-CoV-2 transmission entail an unavoidable uncertainty.

To properly interpret our results, it is important to acknowledge the limitations of the assumptions included in our modeling exercise. First, we have considered a decrease of the transmission probability in outdoors as compared to indoors settings of 1/20 [281]. Although this choice is guided by empirical evidence and our results were robust to this choice, further studies better quantifying the relative risk of indoor vs. outdoor transmission were warranted. Second, our model neglected to consider differences in the behavior that people follow when in contact with each other. It is indeed possible that contacts between relatives and friends have a larger chance of resulting in a transmission event as compared with interactions with strangers [282]. Third, we did not model nursing homes, which were severely hit by the COVID-19 pandemic across the globe. However, although they represented a key setting to determine COVID-19 burden in terms of deaths and patients admitted to hospitals and ICUs, they were possibly not central to capture the transmission dynamics of SARS-CoV-2 at the population level, which is the aim of this section. Although there was some co-location information from hospitals,but we did not model them. Nonetheless, contact tracing studies from several countries revealed that transmission within hospitals wasrelatively low, and hospital staff were more at risk from interactions with their coworkers (e.g. in the breakroom) or out in their communities [283, 284].

In conclusion, the majority of NPIs introduced in large urban areas in March 2020 were effective in dramatically slowing down the first wave of COVID-19 by greatly reducing the number of effective contacts in the population. Closing down schools,

businesses, workplaces, and social venues, however, took (and still does) an enormous toll on our economy and society. Our results and methodology allow for a real-time data-driven analysis that connects NPIs, human behavior and the transmission dynamic of SARS-CoV-2 to provide quantitative information that can aid in defining more targeted and less disruptive interventions not only at a local level, but also to assess whether local restrictions could trigger undesired effects at nearby locations not subject to the same limitations. Furthermore, we extend our previously proposed probabilistic approach from chapter 3 to construct temporal contact networks. This extension enables us to make direct observations of temporal health-related behaviors, providing valuable insights for modeling and quantifying social epidemics.

5

Conclusions

"Scientific research is based on the idea that everything that takes place is determined by laws of nature, and therefore this holds for the action of people."

– Albert Einstein[1]

THE primary objective of this research was to explore the field of computational and digital epidemiology and its application in modeling mathematically human health-related behaviors to explain and predict biological viral epidemics. Specifically, this research aimed to achieve the following objectives:

- **Contextualize the importance of viral epidemics and human behaviours**: In chapter 1, we provided a comprehensive background on why viral epidemics are a crucial and compelling issue to model mathematically. We discussed the significance of social network approaches in epidemiology, the role of novel data streams from social media and mobility in modeling epidemics, and the latest advancements in data-driven epidemiological systems that include human behavioral data.

- **Understand the relationship between human behavior, viral agents, and human health**: We explored in chapter 1 the mechanisms that drive human behavior and their implications for epidemic spread. By examining the impact of changes in human behaviors on global infectious pandemics, we aimed to

[1] Albert Einstein. German-born theoretical physicist. Quote extracted from the book "Albert Einstein, The Human Side: Glimpses from His Archives".

uncover the underlying factors that make populations vulnerable to infectious diseases.

- **Utilize novel data streams for modeling epidemics**: One of the main objectives of this research was to explored the use of social media, chapter 2, and mobility digital traces, chapter 3 and chapter 4, as proxies of complex human social systems. By incorporating these data sources, we aimed to develop data-driven epidemiological systems that could simulate, explain, and predict the interaction between the environment, human behaviors, and the spread of viruses in real-time.

- **Develop methods for EWES**: A key focus of this research was to develop innovative approaches for early warning epidemiological systems. We proposed advanced social network approaches to detect super-sensors in informational epidemics on social media, enabling the prediction of seasonal biological epidemics like ILI, chapter 2. Additionally, we focused on building human contact matrices from real-world mobility data to map COVID-19 transmission, chapter 3 and chapter 4, at different spatiotemporal scales. These methodologies aimed to enhance the ability to detect outbreaks earlier and provide timely interventions.

- **Contribute to the field and inspire further research**: Throughout this thesis, we aimed to make significant contributions to the field of computational and digital epidemiology. By developing novel methodologies, exploring data-driven approaches, and investigating the implications of my findings, we sought to inspire further research in this area. Our goal was to contribute to the development of new data-driven epidemiological systems that leverage digital traces to improve health outcomes at scale.

Our work in this thesis has built upon new data streams encoding human behaviours, utilizing traditional epidemiological and social network methods alongside the latest advancements in computation and machine learning. This has enabled us to develop real-time epidemiological systems that can simulate, explain, and predict the interaction between the environment, human behaviors, and the spread of viruses, enhancing our ability to tackle the most pressing health challenges of our time. Let's see them in more detail.

In chapter 2, we presented a novel approach for building early warning epidemiological systems when direct observational data about health-related behavior is not available. We proposed an advanced social network approach to detect super-sensors in an informational epidemic on social media to predict a seasonal biological epidemic, such as ILI in Spain. By mirroring the informational epidemic on social media with the biological epidemic and identifying super-sensors based on content and centrality metrics, we can predict biological outbreaks in advance of official and social media sources. Our approach utilizes machine learning models to detect super-sensors, which can be used to build time series of the informational epidemic that replicates the biological epidemic. This approach is cost-efficient, sensitive, and respectful of citizens' data because it does not require monitoring and collecting data from the entire population, only a small subset of individuals deemed super-sensors. Furthermore, this method enables public health decision-makers to detect outbreaks earlier than traditional approaches, allowing them to mobilize resources promptly.

In chapter 3 and chapter 4, we proposed methods for building human contact matrices from real-world mobility data when direct observations of human interactions are possible. In chapter 3, we detailed a novel model that integrated anonymized real-time mobility data with census and demographic data to map COVID-19 transmission in the Boston, Massachusetts area, by building static spatiotemporal human contact matrices for the Boston metropolitan area. This study provided insights into possible pitfalls and solutions as cities lifted restrictions that were in place during the firsts COVID-19 lockdowns worldwide. In chapter 4, we focused on expanding the methodology proposed in chapter 3 for building dynamic spatiotemporal human contact matrices at a daily level. We used real-time mobility data with census and demographic data to map COVID-19 transmission from the New York metropolitan area, which included mobility data from New York, New Jersey, and Connecticut states, and Seattle metropolitan area. The goal of this study was to examine the forensic potential of such granular human contact matrices at daily and POI (points of interest) levels of viral transmissions, to quantify super-spreading events and their locations, and compare the impact of NPIs from the two metropolitan areas. Finally, articles [2,3] upon which chapter 3 is based, played a crucial role in examining the second wave and attracting significant scientific and

media attention internationally and locally, as seen in chapter .

To summarize, the computational and digital epidemiology field is vast, and we have only touched upon a small fraction of it in this thesis. Our contribution has been to explore the use of novel data streams and new methodological methods to model mathematically biological epidemics and human health-related behaviors. We developed data processing pipelines to construct human contact interaction matrices and epidemiological time series, which were used to train agent-based, regression, and machine learning models. Our hope is that these short studies will inspire further exploration in the fields of computational and digital epidemiology and lead to the development of new data-driven epidemiological systems that can leverage digital traces to improve the health outcomes of millions of people worldwide.

5.1 Future work

The results presented in this thesis open up several potential paths for future research. Our studies can be expanded upon to gain a deeper understanding of specific issues or to utilize the proposed methodologies to push the boundaries of knowledge and improve the effectiveness and accuracy of epidemiological systems. Additionally, these findings could be applied to better anticipate and respond to new and emerging epidemics.

In chapter 2, we utilized digital traces to indirectly observe human behaviors and model viral biological processes like ILI, employing a behavioral and social network approach. While our study focused on a specific epidemic in a particular country, we believe that our results, based on collective behavior, have the potential to be generalized to other epidemics, regions, and social platforms. Furthermore, our findings can encourage further investigations into the personality and behavioral traits of super-sensors.

In chapter 3 and chapter 4, we demonstrated how to make direct observations of human behavior, model and quantify epidemics, and utilize real-world mobility and census data to infer realistic synthetic populations and their social interactions, represented by social contact matrices. The proposed methodology for feeding agent-based models and simulating the dynamics of COVID-19 is robust and granular, making our synthetic populations suitable for use in a macroeconomic

model. This model could utilize the epidemiological model output from the contact matrices to measure the economic impacts of different health and behavioral interventions on both the epidemic and the economy.

In fact, we have already collaborated on such work which published in a pre-print [5]. Future pandemics require an understanding of the complex interplay between health and the economy. The synthetic populations and social interactions we developed can also be used to explore disease spread between social groups and their impact, shedding light on the correlation between infectious diseases and economic income.

6

Epilogue: Personal learning reflections

"The real scientist is ready to bear privation and, if need be, starvation rather than let anyone dictate to him which direction his work must take."

– Albert Szent-Gyorgyi[1]

THIS thesis answered many inner questions of the writer, but it has left more doors open than closed about the nature of the universe and its complexity. Hence, further study is needed and I will keep exploring other avenues of the knowledge hyper-spaces. Embarking on this research journey and undertaking the process of conducting my study has been a transformative experience that has shaped my growth and development as a researcher after many sleepless nights before and during times of war against COVID-19. Throughout this thesis, I encountered numerous challenges, unexpected discoveries, and valuable lessons that have greatly enriched my understanding of how science works as a tool to foster innovation and human evolution.

One of the key lessons I learned during this research journey was the importance of interdisciplinary collaboration at the highest international level. The field of computational and digital epidemiology bridges the domains of epidemiology, data science, and social sciences. As I delved into the complexities of mathematically modeling biological epidemics and human health-related behaviors, I realized the immense value of collaborating with experts from different disciplines. Engaging in interdisciplinary discussions and incorporating diverse perspectives enriched my

[1] Albert Szent-Györgyi. Hungarian biochemist. Quote extracted from the article 'Science needs freedom'.

research approach and allowed me to tackle complex problems from multiple angles.

Methodological insights gained during my research were invaluable in shaping the direction of my study. I discovered that the integration of anonymized real-time mobility data with census and demographic data offers a powerful means of constructing human contact matrices and mapping COVID-19 transmissions. This granular approach provided unique opportunities to explore viral transmission patterns, super-spreading events, and the impact of non-pharmaceutical interventions. The utilization of machine learning models to build time series of informational epidemics proved to be an efficient and effective strategy for predicting biological outbreaks, and detect super-sensors at scale. These methodological insights opened up new avenues for future research and contributed to the advancement of computational and digital epidemiology methodologies.

In a more theoretical framework, the application of complex systems and network theory has been transformative, allowing me to perceive the universe as an interconnected whole. Embracing these theoretical perspectives has provided me with a profound understanding of the intricate relationships and interdependencies that exist across various systems and entities. This holistic view has not only enriched my academic pursuits but has also influenced the way I perceive and engage with the world around me. Embracing the concepts of mathematical modeling, complex systems and network theory has been a powerful lens through which I appreciate the underlying unity and complexity inherent in the universe.

Along this research journey, I also encountered unexpected discoveries that further expanded my understanding of the field. The exploration of novel data streams, such as social media and mobility digital traces, revealed rich sources of information for capturing human behaviors and modeling epidemics mathematically. I was astonished by the depth and breadth of insights that could be gleaned from these unconventional data sources. These unexpected discoveries highlighted the potential of leveraging digital traces for early warning systems and the real-time monitoring of infectious diseases. The realization that non-traditional data streams could provide valuable insights and complement traditional epidemiological approaches was a significant breakthrough.

Reflecting on my journey as a researcher, I have come to appreciate the importance of adaptability and resilience in the face of challenges. Throughout the

research process, I encountered various obstacles ranging from data collection and analysis to methodological complexities. These challenges were further compounded by unforeseen obstacles and the immense stress caused by the COVID-19 pandemic. Despite the uncertainty and high levels of stress before and during the pandemic, I felt a moral obligation to contribute my technical knowledge to the world, deeply moved by my passion to humanity. However, navigating the rigorous and demanding peer-review process of top journals served as a humbling experience and taught me valuable lessons. Each challenge I encountered presented an opportunity for personal and academic growth. Overcoming these obstacles required unwavering persistence, problem-solving skills, and a willingness to explore alternative paths. The iterative nature of the research process nurtured my adaptability and taught me the significance of embracing uncertainty and adjusting my approach as needed.

For instance, one specific hypothesis we pursued in during this PhD thesis was the potential to measure and nowcast the prevalence of headaches using social media data from Twitter and environmental data, such as pollution, weather and even Schumann resonance data. After conducting extensive analyses spanning several years in collaboration with Nick Obradovich, expert in climate change and human behaviour. We recognized the need to validate our findings against a reliable ground truth. Consequently, I sought access to a public health database through the Spanish Ministry of Health, but despite our efforts, we were unable to obtain the desired results. Regrettably, we had to abandon this line of research due to the lack of conclusive findings and we decided not to include the results in this thesis. While the potential for groundbreaking discoveries was significant, the absence of results prevented us from conducting what could have been the largest epidemiological study on headaches worldwide.

Moreover, when the COVID-19 outbreak emerged, our research focus shifted from studying the spread of influenza among various social statuses with the contact matrices presented in chapter 3 and a first agent-based models coded by myself in python that was very slow. I never coded high performance computing models in python. As you have read in chapter 3, we promptly adapted our plans, modified our data for different social distancing strategies, and refined our models to simulate the spread of COVID-19 along with Alberto Aleta and Yamir Moreno. We had the expertise to build contact matrices from real human mobility data and they had the

expertise on building very efficient agent-based models to run millions of simulations in minutes. Our initial collaboration began on March 13th, 2020. Remarkably, within just nine sleepless days, we published our first preliminary report [2], and our efforts culminated in a significant achievement when our work was published [3] in Nature Human Behaviour journal on August 5th, 2022. Sustaining this fruitful collaboration over several years, we had the privilege of working alongside esteemed researchers in the field, including Alex Vespignani and his team.

Furthermore, this research journey has reinforced the importance of the scientific community and the power of collaboration. Engaging with fellow researchers, attending conferences, and participating in discussions have provided valuable insights, feedback, and encouragement. The constructive criticism and support received from my advisor and colleagues have been instrumental in shaping the trajectory of my research and fostering an environment of continuous learning.

I had the invaluable opportunity to visit the Human Dynamics Labs at the MIT Media Lab, where my thesis advisor, Esteban Moro, conducts groundbreaking research. The lab, led by the esteemed Alex Pentland, is at the forefront of leveraging big data to unravel the complexities of human behavior. Meeting Alex Pentland was a remarkable experience, and I was privileged to collaborate with him as a coauthor on a project. Visiting the MIT Media Lab was nothing short of heavenly for an innovator like me, as it provided an environment that fostered creativity and exploration at the intersection of data science and human behaviour. I have also published in Nature Human Behaviour or PNAS journals, something that I would have never imagined when I started this thesis as a part-time student at my early 30s. I am confident that the lessons learned and the knowledge gained during this research journey will continue to shape my future contributions to the field of science and my professional career, and I am excited about the potential impact of data-driven epidemiological systems powered by AI in improving global health outcomes and what the universe will bring me.

In conclusion, my PhD research in mathematical engineering has been a deep transformative and rewarding experience, personally and professionally. The magnitude of my experiences has transformed me to such an extent that I am no longer the same person, and I really mean it, my entire belief system has undergone a profound and complete shift. The unexpected challenges, discoveries,

methodological insights, and experiences gained have greatly influenced my growth as a researcher. Lastly, I want to reiterate my heartfelt gratitude to my thesis advisor, Esteban Moro, for his unwavering support, kind encouragement, and invaluable constructive feedback throughout this journey. Two students learning together.

7

Bibliography

[1] D. Martín-Corral, M. García-Herranz, M. Cebrian, and E. Moro, "Social Media Sensors to Detect Early Warnings of Influenza at Scale," *medRxiv*, p. 2022.11.15.22282355, Nov 2022.

[2] D. Martín-Calvo, A. Aleta, A. Pentland, Y. Moreno, and E. Moro, "Effectiveness of social distancing strategies for protecting a community from a pandemic with a data driven contact network based on census and real-world mobility data," *Complex. Dig*, 2020.

[3] A. Aleta, D. Martin-Corral, A. P. y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini Jr, S. Merler, *et al.*, "Modelling the impact of testing, contact tracing and household quarantine on second waves of covid-19," *Nature Human Behaviour*, vol. 4, no. 9, pp. 964–971, 2020.

[4] A. Aleta, D. Martín-Corral, M. A. Bakker, A. Pastore y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini Jr, *et al.*, "Quantifying the importance and location of sars-cov-2 transmission events in large metropolitan areas," *Proceedings of the National Academy of Sciences*, vol. 119, no. 26, p. e2112182119, 2022.

[5] M. Pangallo, A. Aleta, R. Chanona, A. Pichler, D. Martín-Corral, M. Chinazzi, F. Lafond, M. Ajelli, E. Moro, Y. Moreno, *et al.*, "The unequal effects of the health-economy tradeoff during the covid-19 pandemic," *arXiv preprint arXiv:2212.03567*, 2022.

[6] D. Martín-Corral, M. Cebrián, and E. Moro, "More people, more fun: The scaling of events in cities." https://medium.com/snsores-lab/more-people-more-fun-the-scaling-of-events-in-cities-f82d3072eb63, 2015.

[7] D. Martín-Corral, *Viviendo en una sociedad enferma*. libros.com, 2022.

[8] G. J. Armelagos, P. J. Brown, and B. Turner, "Evolutionary, historical and political economic perspectives on health and disease," *Social Science & Medicine*, vol. 61, no. 4, pp. 755–765, 2005.

[9] E. D. Kilbourne, "Influenza pandemics of the 20th century," *Emerging infectious diseases*, vol. 12, no. 1, p. 9, 2006.

[10] F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, and I. D. Ladnyi, "The history of smallpox and its spread around the world," *Smallpox and its Eradication*, pp. 209–244, 1988.

[11] D. R. Hopkins, *The greatest killer: smallpox in history*. University of Chicago press, 2002.

[12] X. Wei, S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, *et al.*, "Viral dynamics in human immunodeficiency virus type 1 infection," *Nature*, vol. 373, no. 6510, pp. 117–122, 1995.

[13] M. Chan-Yeung and R.-H. Xu, "Sars: epidemiology," *Respirology*, vol. 8, pp. S9–S14, 2003.

[14] A. C. Steere, J. Coburn, L. Glickstein, *et al.*, "The emergence of lyme disease," *The Journal of clinical investigation*, vol. 113, no. 8, pp. 1093–1101, 2004.

[15] J. P. Nataro and J. B. Kaper, "Diarrheagenic escherichia coli," *Clinical microbiology reviews*, vol. 11, no. 1, pp. 142–201, 1998.

[16] C. B. Jonsson, L. T. M. Figueiredo, and O. Vapalahti, "A global perspective on hantavirus ecology, epidemiology, and disease," *Clinical microbiology reviews*, vol. 23, no. 2, pp. 412–441, 2010.

[17] D. Guha-Sapir and B. Schimmer, "Dengue fever: new paradigms for a changing epidemiology," *Emerging themes in epidemiology*, vol. 2, no. 1, pp. 1–10, 2005.

[18] G. L. Campbell, A. A. Marfin, R. S. Lanciotti, and D. J. Gubler, "West nile virus," *The Lancet infectious diseases*, vol. 2, no. 9, pp. 519–529, 2002.

[19] T. E. Morrison, "Reemergence of chikungunya virus," *Journal of virology*, vol. 88, no. 20, pp. 11644–11647, 2014.

[20] H. Feldmann, S. Jones, H.-D. Klenk, and H.-J. Schnittler, "Ebola virus: from discovery to vaccine," *Nature Reviews Immunology*, vol. 3, no. 8, pp. 677–685, 2003.

[21] L. R. Petersen, D. J. Jamieson, A. M. Powers, and M. A. Honein, "Zika virus," *New England Journal of Medicine*, vol. 374, no. 16, pp. 1552–1563, 2016.

[22] X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, *et al.*, "Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study," *The lancet respiratory medicine*, vol. 8, no. 5, pp. 475–481, 2020.

[23] C. A. Dimala, B. M. Kadia, M. A. M. Nji, and N. N. Bechem, "Factors associated with measles resurgence in the united states in the post-elimination era," *Scientific Reports*, vol. 11, no. 1, p. 51, 2021.

[24] M. G. Reynolds and I. K. Damon, "Outbreaks of human monkeypox after cessation of smallpox vaccination," *Trends in microbiology*, vol. 20, no. 2, pp. 80–87, 2012.

[25] C. T. Darimont, S. M. Carlson, M. T. Kinnison, P. C. Paquet, T. E. Reimchen, and C. C. Wilmers, "Human predators outpace other agents of trait change in the wild," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 952–954, 2009.

[26] V. Ramaswamy, M. Schwarzkopf, W. Randel, B. Santer, B. J. Soden, and G. Stenchikov, "Anthropogenic and natural influences in the evolution of lower stratospheric cooling," *Science*, vol. 311, no. 5764, pp. 1138–1141, 2006.

[27] B. D. Santer, M. F. Wehner, T. Wigley, R. Sausen, G. Meehl, K. Taylor, C. Ammann, J. Arblaster, W. Washington, J. Boyle, *et al.*, "Contributions of anthropogenic and natural forcing to recent tropopause height changes," *science*, vol. 301, no. 5632, pp. 479–483, 2003.

[28] D. J. Wuebbles, D. W. Fahey, K. A. Hibbard, J. R. Arnold, B. DeAngelo, S. Doherty, D. R. Easterling, J. Edmonds, T. Edmonds, T. Hall, *et al.*, "Climate science special report: Fourth national climate assessment (nca4), volume i," 2017.

[29] C. Brown, "Emerging zoonoses and pathogens of public health significance–an overview," *Revue scientifique et technique-office international des epizooties*, vol. 23, no. 2, pp. 435–442, 2004.

[30] T. Allen, K. A. Murray, C. Zambrana-Torrelio, S. S. Morse, C. Rondinini, M. Di Marco, N. Breit, K. J. Olival, and P. Daszak, "Global hotspots and correlates of emerging zoonotic diseases," *Nature communications*, vol. 8, no. 1, pp. 1–10, 2017.

[31] R. Gibb, D. W. Redding, K. Q. Chin, C. A. Donnelly, T. M. Blackburn, T. Newbold, and K. E. Jones, "Zoonotic host diversity increases in human-dominated ecosystems," *Nature*, vol. 584, no. 7821, pp. 398–402, 2020.

[32] A. J. MacDonald and E. A. Mordecai, "Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22212–22218, 2019.

[33] J. Olivero, J. E. Fa, R. Real, A. L. Márquez, M. A. Farfán, J. M. Vargas, D. Gaveau, M. A. Salim, D. Park, J. Suter, *et al.*, "Recent loss of closed forests is associated with ebola virus disease outbreaks," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[34] K. D. Lafferty, "The ecology of climate change and infectious diseases," *Ecology*, vol. 90, no. 4, pp. 888–900, 2009.

[35] E. K. Shuman, "Global climate change and infectious diseases," *New England Journal of Medicine*, vol. 362, no. 12, pp. 1061–1063, 2010.

[36] S. Towers, G. Chowell, R. Hameed, M. Jastrebski, M. Khan, J. Meeks, A. Mubayi, and G. Harris, "Climate change and influenza: the likelihood of early and severe influenza seasons following warmer than average winters," *PLoS currents*, vol. 5, 2013.

[37] Q. Liu, Z.-M. Tan, J. Sun, Y. Hou, C. Fu, and Z. Wu, "Changing rapid weather variability increases influenza epidemic risk in a warming climate," *Environmental Research Letters*, vol. 15, no. 4, p. 044004, 2020.

[38] D. R. Murray, M. Schaller, and P. Suedfeld, "Pathogens and politics: Further evidence that parasite prevalence predicts authoritarianism," *PloS One*, vol. 8, no. 5, p. e62275, 2013.

[39] M. Marmot and R. Wilkinson, *Social determinants of health*. Oup Oxford, 2005.

[40] P. Braveman and L. Gottlieb, "The social determinants of health: it's time to consider the causes of the causes," *Public health reports*, vol. 129, no. 1_suppl2, pp. 19–31, 2014.

[41] A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. Kraemer, J. Wu, *et al.*, "Modelling covid-19," *Nature Reviews Physics*, vol. 2, no. 6, pp. 279–281, 2020.

[42] D. Mistry, M. Litvinova, A. Pastore y Piontti, M. Chinazzi, L. Fumanelli, M. F. Gomes, S. A. Haque, Q.-H. Liu, K. Mu, X. Xiong, *et al.*, "Inferring high-resolution human mixing patterns for disease modeling," *Nature communications*, vol. 12, no. 1, p. 323, 2021.

[43] J. M. Hyman and E. A. Stanley, "The effect of social mixing patterns on the spread of aids," in *Mathematical Approaches to Problems in Resource Management and Epidemiology: Proceedings of a Conference held at Ithaca, NY, Oct. 28–30, 1987*, pp. 190–219, Springer, 1989.

[44] L. Fumanelli, M. Ajelli, P. Manfredi, A. Vespignani, and S. Merler, "Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread," *Proceedings of the National Academy of Sciences*, 2012.

[45] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim, *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[46] C.-Y. Fan, J. C.-Y. Fann, M.-C. Yang, T.-Y. Lin, H.-H. Chen, J.-T. Liu, and K.-C. Yang, "Estimating global burden of covid-19 with disability-adjusted life years and value of statistical life metrics," *Journal of the Formosan Medical Association*, vol. 120, pp. S106–S117, 2021.

[47] W. Bank, "2014–2015 west africa ebola crisis: impact update," 2016.

[48] D. Rassy and R. D. Smith, "The economic impact of h1n1 on mexico's tourist and pork sectors," *Health economics*, vol. 22, no. 7, pp. 824–834, 2013.

[49] V. Cagnolati, S. Tempia, and A. Abdi, "Economic impact of rift valley fever on the somali livestock industry and a novel surveillance approach in nomadic pastoral systems," in *Proceedings of the 11th Symposium of the International Society for Veterinary Epidemiology and Economics*, pp. 6–11, 2006.

[50] H. Field, P. Daniels, B. Lee, A. Jamaludin, M. Bunning, *et al.*, "Manual on the diagnosis of nipah virus infection in animals," 2002.

[51] S. Begley, "Flu-conomics: The next pandemic could trigger global recession," *REUTERS, https://www. reuters. com/article/us-reutersmagazine-davos-flu-economy/flu-conomics-the-next-pandemic-could-trigger-global-recession-idUSBRE90K0F820130121*, 2013.

[52] G. Wearden, "Ftse 100 suffers worst year since 2008 financial crisis," *The Guardian, https://www.theguardian.com/business/2020/dec/31/ftse-100-suffers-worst-year-since-2008-financial-crisis*, 2021.

[53] I. M. Fund, "World economic outlook update, january 2021," *International Monterary Fund World Economic Outlooks, https://www.imf.org/en/Publications/WEO/Issues/2021/01/26/2021-world-economic-outlook-update*, 2021.

[54] "Can in-store digital experience transform the future of shopping?," 2020.

[55] M. John, *A dictionary of epidemiology*. Oxford university press, 2001.

[56] A. Morabia, *A history of epidemiologic methods and concepts*. Springer Science & Business Media, 2005.

[57] R. M. Merrill, *Introduction to epidemiology*. Jones & Bartlett Publishers, 2015.

[58] C. Buck and A. Llopis, *The challenge of epidemiology: issues and selected readings*, vol. 505. Pan American Health Org, 1988.

[59] C. Walford, "Early bills of mortality," *Transactions of the Royal Historical Society*, vol. 7, pp. 212–248, 1878.

[60] K. Dietz and J. Heesterbeek, "Bernoulli was ahead of modern epidemiology," *Nature*, vol. 408, no. 6812, pp. 513–514, 2000.

[61] S. M. Teutsch, R. E. Churchill, *et al.*, *Principles and practice of public health surveillance*. Oxford University Press, USA, 2000.

[62] F. E. Cox, "History of the discovery of the malaria parasites and their vectors," *Parasites & vectors*, vol. 3, no. 1, pp. 1–9, 2010.

[63] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[64] R. Doll and A. B. Hill, "Smoking and carcinoma of the lung," *British medical journal*, vol. 2, no. 4682, p. 739, 1950.

[65] W. B. Kannel, "The framingham study: Its 50-year legacy and future promise," *Journal of atherosclerosis and thrombosis*, vol. 6, no. 2, pp. 60–66, 2000.

[66] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England journal of medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[67] F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, I. D. Ladnyi, *et al.*, *Smallpox and its eradication*, vol. 6. World Health Organization Geneva, 1988.

[68] World Health Organization, "Global outbreak alert and response network (goarn)." https://extranet.who.int/goarn/.

[69] World Health Organization, "Integrated outbreak analytics (ioa)." https://extranet.who.int/goarn/content/integrated-outbreak-analytics-delivers-holistic-understanding-outbreak-dynamics.

[70] World Health Organization, "Epidemic intelligence from open sources (eios)." https://www.who.int/initiatives/eios.

[71] World Health Organization, "Epi-brain." https://www.epi-brain.com/.

[72] S. E. Carter, N. Gobat, J. P. Zambruni, J. Bedford, E. Van Kleef, T. Jombart, M. Mossoko, D. B. Nkakirande, C. N. Colorado, and S. Ahuka-Mundeke, "What questions we should be asking about covid-19 in humanitarian settings: perspectives from the social sciences analysis cell in the democratic republic of the congo," *BMJ global health*, vol. 5, no. 9, p. e003607, 2020.

[73] F. Tönnies, "Gemeinschaft und gesellschaft," in *Studien zu Gemeinschaft und Gesellschaft*, pp. 27–58, Springer, 2012.

[74] É. Durkheim, *De la division du travail social*. F. Alcan, 1922.

[75] G. Simmel, *Soziologie: Untersuchungen über die formen der vergesellschaftung*, vol. 2. Duncker & Humblot, 1923.

[76] L. Freeman, "The development of social network analysis," *A Study in the Sociology of Science*, vol. 1, no. 687, pp. 159–167, 2004.

[77] K. A. Fredericks and M. M. Durland, "The historical evolution and basic concepts of social network analysis," *New directions for evaluation*, vol. 2005, no. 107, pp. 15–23, 2005.

[78] J. L. Moreno, "Sociometry in relation to other social sciences," *Sociometry*, vol. 1, no. 1/2, pp. 206–219, 1937.

[79] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[80] S. H. Strogatz, "Exploring complex networks," *nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[81] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[82] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Social science. computational social science.," *Science (New York, NY)*, vol. 323, no. 5915, pp. 721–723, 2009.

[83] D. M. Lazer, A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, *et al.*, "Computational social science: Obstacles and opportunities," *Science*, vol. 369, no. 6507, pp. 1060–1062, 2020.

[84] S. Richey, "The autoregressive influence of social network political knowledge on voting behaviour," *British Journal of Political Science*, vol. 38, no. 3, pp. 527–542, 2008.

[85] S. Abrams, T. Iversen, and D. Soskice, "Informal social networks and rational voting," *British Journal of Political Science*, vol. 41, no. 2, pp. 229–257, 2011.

[86] S. Aral, "Poked to vote," *Nature*, vol. 489, no. 7415, pp. 212–214, 2012.

[87] T. Ryan and S. Xenos, "Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage," *Computers in human behavior*, vol. 27, no. 5, pp. 1658–1664, 2011.

[88] T. C. Marshall, K. Lefringhausen, and N. Ferenczi, "The big five, self-esteem, and narcissism as predictors of the topics people write about in facebook status updates," *Personality and Individual Differences*, vol. 85, pp. 35–40, 2015.

[89] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, pp. 5–es, 2007.

[90] J. L. Iribarren and E. Moro, "Impact of human activity patterns on the dynamics of information diffusion," *Physical review letters*, vol. 103, no. 3, p. 038702, 2009.

[91] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[92] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*, pp. 115–122, 2010.

[93] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PLoS Comput Biol*, vol. 11, no. 10, p. e1004513, 2015.

[94] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models," in *2014 IEEE international conference on data mining*, pp. 755–760, IEEE, 2014.

[95] R. Chunara, S. Aman, M. Smolinski, and J. S. Brownstein, "Flu near you: an online self-reported influenza surveillance system in the usa," *Online Journal of Public Health Informatics*, vol. 5, no. 1, 2013.

[96] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[97] M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, *et al.*, "Digital epidemiology," *PLoS Comput Biol*, vol. 8, no. 7, p. e1002616, 2012.

[98] M. Marathe and A. K. S. Vullikanti, "Computational epidemiology," *Communications of the ACM*, vol. 56, no. 7, pp. 88–96, 2013.

[99] J. J. V. Bavel, K. Baicker, P. S. Boggio, V. Capraro, A. Cichocka, M. Cikara, M. J. Crockett, A. J. Crum, K. M. Douglas, J. N. Druckman, *et al.*, "Using social and behavioural science to support covid-19 pandemic response," *Nature human behaviour*, vol. 4, no. 5, pp. 460–471, 2020.

[100] S. Brailsford and B. Schmidt, "Towards incorporating human behaviour in models of health care systems: An approach using discrete event simulation," *European journal of operational research*, vol. 150, no. 1, pp. 19–31, 2003.

[101] N. Ferguson, "Capturing human behaviour," *Nature*, vol. 446, no. 7137, pp. 733–733, 2007.

[102] K. Konsolakis, H. Hermens, C. Villalonga, M. Vollenbroek-Hutten, and O. Banos, "Human behaviour analysis through smartphones," *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 2, no. 19, p. 1243, 2018.

[103] C. N. Macpherson, "Human behaviour and the epidemiology of parasitic zoonoses," *International journal for parasitology*, vol. 35, no. 11-12, pp. 1319–1331, 2005.

[104] L.-Q. Fang, X.-J. Wang, S. Liang, Y.-L. Li, S.-X. Song, W.-Y. Zhang, Q. Qian, Y.-P. Li, L. Wei, Z.-Q. Wang, *et al.*, "Spatiotemporal trends and climatic factors of hemorrhagic fever with renal syndrome epidemic in shandong province, china," *PLoS neglected tropical diseases*, vol. 4, no. 8, p. e789, 2010.

[105] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection—harnessing the web for public health surveillance," *The New England journal of medicine*, vol. 360, no. 21, p. 2153, 2009.

[106] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," *PLoS computational biology*, vol. 7, no. 10, p. e1002199, 2011.

[107] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, "Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic," *PloS one*, vol. 6, no. 8, p. e23610, 2011.

[108] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, "Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales," *PLoS Comput Biol*, vol. 9, no. 10, p. e1003256, 2013.

[109] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[110] D. Butler, "When google got flu wrong: Us outbreak foxes a leading web-based method for tracking seasonal flu," *Nature*, vol. 494, no. 7436, pp. 155–157, 2013.

[111] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in china with search query from baidu," *PloS one*, vol. 8, no. 5, p. e64323, 2013.

[112] D. J. McIver and J. S. Brownstein, "Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003581, 2014.

[113] N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky, "Global disease monitoring and forecasting with wikipedia," *PLoS Comput Biol*, vol. 10, no. 11, p. e1003892, 2014.

[114] V. Lampos, M. S. Majumder, E. Yom-Tov, M. Edelstein, S. Moura, Y. Hamada, M. X. Rangaka, R. A. McKendry, and I. J. Cox, "Tracking covid-19 using online search," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–11, 2021.

[115] R. P. Soebiyanto, F. Adimi, and R. K. Kiang, "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters," *PloS one*, vol. 5, no. 3, p. e9450, 2010.

[116] N. E. Kogan, L. Clemente, P. Liautaud, J. Kaashoek, N. B. Link, A. T. Nguyen, F. S. Lu, P. Huybers, B. Resch, C. Havas, *et al.*, "An early warning approach to monitor covid-19 activity with multiple digital traces in near real time," *Science Advances*, vol. 7, no. 10, p. eabd6989, 2021.

[117] V. Lampos, T. D. Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 599–602, Springer, 2010.

[118] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pp. 702–707, IEEE, 2011.

[119] C. Dalton, D. Durrheim, J. Fejsa, L. Francis, S. Carlson, E. T. d'Espaignet, and F. Tuyl, "Flutracking: a weekly australian community online survey of influenza-like illness in 2006, 2007 and 2008," *Communicable diseases intelligence quarterly report*, vol. 33, no. 3, pp. 316–322, 2009.

[120] D. Paolotti, C. Gioannini, V. Colizza, and A. Vespignani, "Internet-based monitoring system for influenza-like illness: H1n1 surveillance in italy," in *3rd International ICST Conference on Electronic Healthcare for the 21st century*, vol. 13, pp. 2010–15, Casablanca Morocco, 2010.

[121] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin, *et al.*, "Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience," *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 17–21, 2014.

[122] A. Richard, L. Müller, A. Wisniak, A. Thiabaud, T. Merle, D. Dietrich, D. Paolotti, E. Jeannot, and A. Flahault, "Grippenet: a new tool for the monitoring, risk-factor and vaccination coverage analysis of influenza-like illness in switzerland," *Vaccines*, vol. 8, no. 3, p. 343, 2020.

[123] E. Moro, D. Calacci, X. Dong, and A. Pentland, "Mobility patterns are associated with experienced income segregation in large us cities," *Nature communications*, vol. 12, no. 1, p. 4633, 2021.

[124] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti," *PLoS medicine*, vol. 8, no. 8, p. e1001083, 2011.

[125] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, "Modeling human mobility responses to the large-scale spreading of infectious diseases," *Scientific reports*, vol. 1, no. 1, pp. 1–7, 2011.

[126] S. Merler, M. Ajelli, A. Pugliese, and N. M. Ferguson, "Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: implications for real-time modelling," *PLoS computational biology*, vol. 7, no. 9, p. e1002205, 2011.

[127] M. E. Halloran, A. Vespignani, N. Bharti, L. R. Feldstein, K. Alexander, M. Ferrari, J. Shaman, J. M. Drake, T. Porco, J. N. Eisenberg, *et al.*, "Ebola: mobility data," *Science*, vol. 346, no. 6208, pp. 433–433, 2014.

[128] L. Alessandretti, "What human mobility data tell us about covid-19 spread," *Nature Reviews Physics*, vol. 4, no. 1, pp. 12–13, 2022.

[129] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun, *et al.*, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, 2020.

[130] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of covid-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[131] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, *et al.*, "The effect of human mobility and control measures on the COVID-19 epidemic in China," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[132] S. R. Friedman and S. Aral, "Social networks, risk-potential networks, health, and disease," *Journal of Urban Health*, vol. 78, no. 3, pp. 411–418, 2001.

[133] K. P. Smith and N. A. Christakis, "Social networks and health," *Annual review of sociology*, vol. 34, no. 1, pp. 405–429, 2008.

[134] N. A. Christakis and J. H. Fowler, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in medicine*, vol. 32, no. 4, pp. 556–577, 2013.

[135] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[136] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[137] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature communications*, vol. 5, no. 1, pp. 1–10, 2014.

[138] S. L. Feld, "Why your friends have more friends than you do," *American journal of sociology*, vol. 96, no. 6, pp. 1464–1477, 1991.

[139] N. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, pp. 225–233, 2013.

[140] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PloS one*, vol. 5, no. 9, p. e12948, 2010.

[141] M. Garcia-Herranz, E. Moro, M. Cebrian, N. A. Christakis, and J. H. Fowler, "Using friends as sensors to detect global-scale contagious outbreaks," *PloS one*, vol. 9, no. 4, p. e92413, 2014.

[142] M. Galesic, W. Bruine de Bruin, J. Dalege, S. L. Feld, F. Kreuter, H. Olsson, D. Prelec, D. L. Stein, and T. van Der Does, "Human social sensing is an untapped resource for computational social science," *Nature*, vol. 595, no. 7866, pp. 214–222, 2021.

[143] R. I. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella, "The structure of online social networks mirrors those in the offline world," *Social networks*, vol. 43, pp. 39–47, 2015.

[144] J. Zhang and D. Centola, "Social networks and health: New developments in diffusion, online and offline," *Annual Review of Sociology*, vol. 45, pp. 91–109, 2019.

[145] M. Morris, "Epidemiology and social networks: Modeling structured diffusion," *Sociological methods & research*, vol. 22, no. 1, pp. 99–126, 1993.

[146] N. M. Ferguson and G. P. Garnett, "More realistic models of sexually transmitted disease transmission dynamics: sexual partnership networks, pair models, and moment closure," *Sexually transmitted diseases*, pp. 600–609, 2000.

[147] S. Eubank, H. Guclu, V. Anil Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.

[148] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe, "Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks," in *SC'08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, pp. 1–12, IEEE, 2008.

[149] S. Merler and M. Ajelli, "The role of population heterogeneity and human mobility in the spread of pandemic influenza," *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1681, pp. 557–565, 2010.

[150] A. Machens, F. Gesualdo, C. Rizzo, A. E. Tozzi, A. Barrat, and C. Cattuto, "An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices," *BMC infectious diseases*, vol. 13, no. 1, pp. 1–15, 2013.

[151] H. W. Hethcote and J. A. Yorke, *Gonorrhea transmission dynamics and control*, vol. 56. Springer, 2014.

[152] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, *et al.*, "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS medicine*, vol. 5, no. 3, p. e74, 2008.

[153] A. Aleta, G. Ferraz de Arruda, and Y. Moreno, "Data-driven contact structures: from homogeneous mixing to multilayer networks," *PLoS computational biology*, vol. 16, no. 7, p. e1008035, 2020.

[154] Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani, "Measurability of the epidemic reproduction number in data-driven contact networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 50, pp. 12680–12685, 2018.

[155] E. M. McCulley, P. H. Mullachery, A. F. Ortigoza, D. A. Rodríguez, A. V. Diez Roux, and U. Bilal, "Urban scaling of health outcomes: a scoping review," *Journal of Urban Health*, pp. 1–18, 2022.

[156] K. Sankaran and S. P. Holmes, "Generative models: An interdisciplinary perspective," *arXiv preprint arXiv:2208.06011*, 2022.

[157] J. G. Cárcamo, R. G. Vogel, A. M. Terwilliger, J. P. Leidig, and G. Wolffe, "Generative models for synthetic populations.," in *SummerSim*, pp. 7–1, 2017.

[158] B. D. Marshall and S. Galea, "Formalizing the role of agent-based modeling in causal inference and epidemiology," *American journal of epidemiology*, vol. 181, no. 2, pp. 92–99, 2015.

[159] N. Meade and T. Islam, "Modelling and forecasting the diffusion of innovation–a 25-year review," *International Journal of forecasting*, vol. 22, no. 3, pp. 519–545, 2006.

[160] R. N. Thompson and E. Brooks-Pollock, "Detection, forecasting and control of infectious disease epidemics: modelling outbreaks in humans, animals and plants," 2019.

[161] M. Jit and M. Brisson, "Modelling the epidemiology of infectious diseases for decision analysis," *Pharmacoeconomics*, vol. 29, no. 5, pp. 371–386, 2011.

[162] R. Deardon, S. P. Brooks, B. T. Grenfell, M. J. Keeling, M. J. Tildesley, N. J. Savill, D. J. Shaw, and M. E. Woolhouse, "Inference for individual-level models of infectious diseases in large populations," *Statistica Sinica*, vol. 20, no. 1, p. 239, 2010.

[163] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, p. 3200, 2001.

[164] J. A. Jacquez, "A note on chain-binomial models of epidemic spread: what is wrong with the reed-frost formulation?," *Mathematical Biosciences*, vol. 87, no. 1, pp. 73–82, 1987.

[165] H. Abbey, "An examination of the reed-frost theory of epidemics," *Human biology*, vol. 24, no. 3, p. 201, 1952.

[166] R. M. May and A. L. Lloyd, "Infection dynamics on scale-free networks," *Physical Review E*, vol. 64, no. 6, p. 066112, 2001.

[167] E. Teweldemedhin, T. Marwala, and C. Mueller, "Agent-based modelling: a case study in hiv epidemic," in *Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, pp. 154–159, IEEE, 2004.

[168] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.

[169] J. Wiens and E. S. Shenoy, "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology," *Clinical Infectious Diseases*, vol. 66, no. 1, pp. 149–153, 2018.

[170] T. L. Wiemken and R. R. Kelley, "Machine learning in epidemiology and health outcomes research.," *Annual review of public health*, vol. 41, pp. 21–36, 2019.

[171] M. J. Paul, A. Sarker, J. S. Brownstein, A. Nikfarjam, M. Scotch, K. L. Smith, and G. Gonzalez, "Social media mining for public health monitoring and surveillance," in *Biocomputing 2016: Proceedings of the Pacific symposium*, pp. 468–479, World Scientific, 2016.

[172] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez, "Utilizing social media data for pharmacovigilance: a review," *Journal of biomedical informatics*, vol. 54, pp. 202–212, 2015.

[173] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems and Signal Processing*, vol. 107, pp. 241–265, 2018.

[174] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[175] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[176] J. A. Roth, M. Battegay, F. Juchler, J. E. Vogt, and A. F. Widmer, "Introduction to machine learning in digital healthcare epidemiology," *Infection Control & Hospital Epidemiology*, vol. 39, no. 12, pp. 1457–1462, 2018.

[177] S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning: a review," *International Journal of Engineering & Technology*, vol. 7, no. 3, pp. 1019–1023, 2018.

[178] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," *The New England journal of medicine*, vol. 376, no. 26, p. 2507, 2017.

[179] S. B. Thacker, R. G. Parrish, and F. L. Trowbridge, "A method for evaluating systems of epidemiological surveillance," *World Health Statistics Quarterly 1988; 41 (1): 11-18*, 1988.

[180] H. V. Fineberg and M. E. Wilson, "Epidemic science in real time," 2009.

[181] J. Parry, "H7n9 avian flu infects humans for the first time," 2013.

[182] K. Stadler, V. Masignani, M. Eickmann, S. Becker, S. Abrignani, H.-D. Klenk, and R. Rappuoli, "Sars—beginning to understand a new virus," *Nature Reviews Microbiology*, vol. 1, no. 3, pp. 209–218, 2003.

[183] R. A. Fouchier, T. Kuiken, M. Schutten, G. Van Amerongen, G. J. Van Doornum, B. G. Van Den Hoogen, M. Peiris, W. Lim, K. Stöhr, and A. D. Osterhaus, "Koch's postulates fulfilled for sars virus," *Nature*, vol. 423, no. 6937, pp. 240–240, 2003.

[184] H. Feldmann and T. W. Geisbert, "Ebola haemorrhagic fever," *The Lancet*, vol. 377, no. 9768, pp. 849–862, 2011.

[185] S. Briand, E. Bertherat, P. Cox, P. Formenty, M.-P. Kieny, J. K. Myhre, C. Roth, N. Shindo, and C. Dye, "The international ebola emergency," *New England Journal of Medicine*, vol. 371, no. 13, pp. 1180–1183, 2014.

[186] J. Riou and C. L. Althaus, "Pattern of early human-to-human transmission of wuhan 2019 novel coronavirus (2019-ncov), december 2019 to january 2020," *Eurosurveillance*, vol. 25, no. 4, p. 2000058, 2020.

[187] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, *et al.*, "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1199–1207, 2020.

[188] L.-l. Ma, C. Ma, H.-F. Zhang, and B.-H. Wang, "Identifying influential spreaders in complex networks based on gravity formula," *Physica A: Statistical Mechanics and its Applications*, vol. 451, pp. 205–212, 2016.

[189] R. M. Christley, G. Pinchbeck, R. Bowers, D. Clancy, N. French, R. Bennett, and J. Turner, "Infection in social networks: using network analysis to identify high-risk individuals," *American journal of epidemiology*, vol. 162, no. 10, pp. 1024–1031, 2005.

[190] M. Alexander, L. Forastiere, S. Gupta, and N. A. Christakis, "Algorithms for seeding social networks can enhance the adoption of a public health intervention in urban india," *Proceedings of the National Academy of Sciences*, vol. 119, no. 30, p. e2120742119, 2022.

[191] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, "Statistical physics of vaccination," *Physics Reports*, vol. 664, pp. 1–113, 2016.

[192] R. Ghosh, J. Mareček, W. M. Griggs, M. Souza, and R. N. Shorten, "Predictability and fairness in social sensing," *IEEE Internet of Things Journal*, 2021.

[193] M. T. Rashid and D. Wang, "Covidsens: a vision on reliable social sensing for covid-19," *Artificial intelligence review*, vol. 54, no. 1, pp. 1–25, 2021.

[194] K. Farrahi, R. Emonet, and M. Cebrian, "Predicting a community's flu dynamics with mobile phone data," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 1214–1221, ACM, 2015.

[195] H. Shao, K. Hossain, H. Wu, M. Khan, A. Vullikanti, B. A. Prakash, M. Marathe, and N. Ramakrishnan, "Forecasting the flu: designing social network sensors for epidemics," *arXiv preprint arXiv:1602.06866*, 2016.

[196] S. Kianersi, Y.-Y. Ahn, and M. Rosenberg, "Association between sampling method and covid-19 test positivity among undergraduate students: Testing friendship paradox in covid-19 network of transmission," *medRxiv*, 2020.

[197] Twitter, "Twitter developer documentation."

[198] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

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

[200] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006.

[201] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[202] R. Troncy, "Bringing the iptc news architecture into the semantic web," in *International Semantic Web Conference*, pp. 483–498, Springer, 2008.

[203] G. de Vigilancia de Gripe del Centro Nacional de Epidemiología. ISCIII, "Sistema de vigilancia de la gripe en españa."

[204] E. Commission, "Commission implementing decision (eu) 2018/945 of 22 june 2018 on the communicable diseases and related special health issues to be covered by epidemiological surveillance as well as relevant case definitions," *Off J Eur Union*, vol. 61, pp. 1–74, 2018.

[205] N. Kishore, A. R. Taylor, P. E. Jacob, N. Vembar, T. Cohen, C. O. Buckee, and N. A. Menzies, "Evaluating the reliability of mobility metrics from aggregated mobile phone data as proxies for SARS-CoV-2 transmission in the USA: a population-based study," *The Lancet Digital Health*, vol. 4, pp. e27–e36, 1 2022.

[206] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras, "Studying User Income through Language, Behaviour and Affect in Social Media," *PLoS ONE*, vol. 10, p. e0138717, 9 2015.

[207] K. N. Nelson, A. J. Siegler, P. S. Sullivan, H. Bradley, E. Hall, N. Luisi, P. Hipp-Ramsey, T. Sanchez, K. Shioda, and B. A. Lopman, "Nationally representative social contact patterns among U.S. adults, August 2020-April 2021," *Epidemics*, vol. 40, p. 100605, 2022.

[208] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 149–156, IEEE, 2011.

[209] T. Chamorro-Premuzic and A. Furnham, "Personality and music: Can traits explain how people use music in everyday life?," *British journal of psychology*, vol. 98, no. 2, pp. 175–185, 2007.

[210] S. M. Reich, K. Subrahmanyam, and G. Espinoza, "Friending, iming, and hanging out face-to-face: overlap in adolescents' online and offline social networks.," *Developmental psychology*, vol. 48, no. 2, p. 356, 2012.

[211] G. C. Huang, J. B. Unger, D. Soto, K. Fujimoto, M. A. Pentz, M. Jordan-Marsh, and T. W. Valente, "Peer influences: the impact of online and offline friendship networks on adolescent smoking and alcohol use," *Journal of Adolescent Health*, vol. 54, no. 5, pp. 508–514, 2014.

[212] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen, "Retrospective analysis of the possibility of predicting the covid-19 outbreak from internet searches and social media data, china, 2020," *Eurosurveillance*, vol. 25, no. 10, p. 2000199, 2020.

[213] L. Qin, Q. Sun, Y. Wang, K.-F. Wu, M. Chen, B.-C. Shia, and S.-Y. Wu, "Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index," *International journal of environmental research and public health*, vol. 17, no. 7, p. 2365, 2020.

[214] "World Health Organization, "Novel Coronavirus – China" ," 2020.

[215] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[216] S. Flaxman, S. Mishra, A. Gandy, H. Unwin, H. Coupland, T. Mellan, H. Zhu, T. Berah, J. Eaton, P. Perez Guzman, *et al.*, "Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: technical description update." Preprint at arXiv: https://arxiv.org/abs/2004.11342 (2020).

[217] P. G. T. Walker, C. Whittaker, O. J. Watson, M. Baguelin, P. Winskill, A. Hamlet, B. A. Djafaara, Z. Cucunubá, D. Olivera Mesa, W. Green, H. Thompson, S. Nayagam, K. E. C. Ainslie, S. Bhatia, S. Bhatt, A. Boonyasiri, O. Boyd, N. F. Brazeau, L. Cattarino, G. Cuomo-Dannenburg, A. Dighe, C. A. Donnelly, I. Dorigatti, S. L. van Elsland, R. FitzJohn, H. Fu, K. A. M. Gaythorpe, L. Geidelberg, N. Grassly, D. Haw, S. Hayes, W. Hinsley, N. Imai, D. Jorgensen, E. Knock, D. Laydon, S. Mishra, G. Nedjati-Gilani, L. C. Okell, H. J. Unwin, R. Verity, M. Vollmer, C. E. Walters, H. Wang, Y. Wang, X. Xi, D. G. Lalloo, N. M. Ferguson, and A. C. Ghani, "The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries," *Science*, p. eabc0035, Jun 2020.

[218] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch, "Projecting the transmission dynamics of sars-cov-2 through the postpandemic period," *Science*, vol. 368, no. 6493, pp. 860–868, 2020.

[219] Laura Di Domenico , Giulia Pullano, Chiara E. Sabbatini , Pierre-Yves Boëlle , Vittoria Colizza, "Expected impact of lockdown in Île-de-France and possible exit strategies." Preprint at medRxiv: https://www.medrxiv.org/content/10.1101/2020.04.13.20063933v1 (2020).

[220] Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani, "Measurability of the epidemic reproduction number in data-driven contact networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 50, pp. 12680–12685, 2018.

[221] U.S. Census Bureau, "2018 American Community Survey 5-Year Data," 2019.

[222] C. Poletto, S. Meloni, V. Colizza, Y. Moreno, and A. Vespignani, "Host mobility drives pathogen competition in spatially structured populations," *PLoS computational biology*, vol. 9, no. 8, 2013.

[223] J. Zhang, M. Litvinova, W. Wang, Y. Wang, X. Deng, X. Chen, M. Li, W. Zheng, L. Yi, X. Chen, *et al.*, "Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 7, pp. 793–802, 2020.

[224] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, *et al.*, "Estimates of the severity of coronavirus disease 2019: a model-based analysis," *The Lancet Infectious Diseases*, vol. 20, no. 6, pp. 669–677, 2020.

[225] "Foursquare Venue Category Hierarchy." https://developer.foursquare.com/docs/build-with-foursquare/categories/. Accessed: 09-12-2020.

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

[227] D. Mistry, M. Litvinova, A. P. y. Piontti, M. Chinazzi, L. Fumanelli, M. F. C. Gomes, S. A. Haque, Q.-H. Liu, K. Mu, X. Xiong, M. E. Halloran, M. Longini Ira, Jr., S. Merler, M. Ajelli, and A. Vespignani, "Inferring high-resolution human mixing patterns for disease modeling," *arXiv*, Feb 2020.

[228] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska11, and W. Edmunds, "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS medicine*, vol. 5, no. 3, 2008.

[229] M. Ajelli and M. Litvinova, "Estimating contact patterns relevant to the spread of infectious diseases in russia," *Journal of Theoretical Biology*, vol. 419, pp. 1–7, 2017.

[230] M. Litvinova, Q.-H. Liu, E. S. Kulikov, and M. Ajelli, "Reactive school closure weakens the network of social interactions and reduces the spread of influenza," *Proc Natl Acad Sci USA*, vol. 116, no. 27, pp. 13174–13181, 2019.

[231] D. Mistry, M. Litvinova, M. Chinazzi, L. Fumanelli, M. F. Gomes, S. A. Haque, Q.-H. Liu, K. Mu, X. Xiong, M. E. Halloran, I. J. Longini, S. Merler, M. Ajelli, and A. Vespignani, "Inferring high-resolution human mixing patterns for disease modeling." Preprint at arXiv: https://arxiv.org/abs/2003.01214 (2020).

[232] G. Béraud, S. Kazmercziak, P. Beutels, D. Levy-Bruhl, X. Lenne, N. Mielcarek, Y. Yazdanpanah, P.-Y. Boëlle, N. Hens, and B. Dervaux, "The french connection: the first large population-based contact survey in france relevant for the spread of infectious diseases," *PloS one*, vol. 10, no. 7, 2015.

[233] J. Zhang, P. Klepac, J. M. Read, A. Rosello, X. Wang, S. Lai, M. Li, Y. Song, Q. Wei, H. Jiang, *et al.*, "patterns of human social contact and contact with animals in shanghai, china," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[234] American Hospital Directory and are based on public records obtained from the US Centers for Medicare and Medicaid Services, " ," 2020.

[235] The White House, " 15 days to slow the spread.," 2020.

[236] CNN, "These states have implemented stay-at-home orders. Here's what that means for you," 2020.

[237] IMF Blog, "Global Uncertainty Related to Coronavirus at Record High," 2020.

[238] S. Gottlieb, C. Rivers, and M. B. McClellan, "National coronavirus response: A road map to reopening," 2020.

[239] J. Zhang, M. Litvinova, Y. Liang, Y. Wang, W. Wang, S. Zhao, Q. Wu, S. Merler, C. Viboud, A. Vespignani, M. Ajelli, and H. Yu, "Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China," *Science*, p. eaba8001, 2020.

[240] Bakker, Michiel and Berke, Alex and Groh, Matt and Pentland, Alex and Moro, Esteban, "Effect of social distancing measures in the New York City metropolitan area," 2020.

[241] K. Leung, J. T. Wu, D. Liu, and G. M. Leung, "First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment," *The Lancet*, vol. 395, no. 10233, pp. 1382–1393, 2020.

[242] A. J. Kucharski, P. Klepac, A. J. K. Conlan, S. M. Kissler, M. L. Tang, H. Fry, J. R. Gog, W. J. Edmunds, J. C. Emery, G. Medley, J. D. Munday, T. W. Russell, Q. J. Leclerc, C. Diamond, S. R. Procter, A. Gimma, F. Y. Sun, H. P. Gibbs, A. Rosello, K. van Zandvoort, S. Hué, S. R. Meakin, A. K. Deol, G. Knight, T. Jombart, A. M. Foss, N. I. Bosse, K. E. Atkins, B. J. Quilty, R. Lowe, K. Prem, S. Flasche, C. A. B. Pearson, R. M. G. J. Houben, E. S. Nightingale, A. Endo, D. C. Tully, Y. Liu, J. Villabona-Arenas, K. O'Reilly, S. Funk, R. M. Eggo, M. Jit, E. M. Rees, J. Hellewell, S. Clifford, C. I. Jarvis, S. Abbott, M. Auzenbergs, N. G. Davies, and D. Simons, "Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study," *Lancet Infect. Dis.*, Jun 2020.

[243] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun, *et al.*, "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *The Lancet Global Health*, vol. 8, no. 4, pp. e488–e496, 2020.

[244] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser, "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 368, no. 6491, p. eabb6936, 2020.

[245] N. K. Ibrahim, "Epidemiologic surveillance for controlling covid-19 pandemic: types, challenges and implications," *Journal of infection and public health*, vol. 13, no. 11, pp. 1630–1638, 2020.

[246] W. H. Organization *et al.*, "Who consultation to adapt influenza sentinel surveillance systems to include covid-19 virological surveillance: virtual meeting, 6–8 october 2020," tech. rep., World Health Organization, 2022.

[247] S. Engler, P. Brunner, R. Loviat, T. Abou-Chadi, L. Leemann, A. Glaser, and D. Kübler, "Democracy in times of the pandemic: explaining the variation of covid-19 policies across european democracies," *West European Politics*, vol. 44, no. 5-6, pp. 1077–1102, 2021.

[248] N. J. Saam, C. Friedrich, and H. Engelhardt, "The value conflict between freedom and security: Explaining the variation of covid-19 policies in democracies and autocracies," *Plos one*, vol. 17, no. 9, p. e0274270, 2022.

[249] D. Bol, M. Giani, A. Blais, and P. J. Loewen, "The effect of covid-19 lockdowns on political support: Some good news for democracy?," *European journal of political research*, vol. 60, no. 2, pp. 497–505, 2021.

[250] A. Bick, A. Blandin, K. Mertens, *et al.*, "Work from home after the covid-19 outbreak," 2020.

[251] N. Barbour, N. Menon, and F. Mannering, "A statistical assessment of work-from-home participation during different stages of the covid-19 pandemic," *Transportation Research Interdisciplinary Perspectives*, vol. 11, p. 100441, 2021.

[252] A. Akinbi, M. Forshaw, and V. Blinkhorn, "Contact tracing apps for the covid-19 pandemic: a systematic literature review of challenges and future directions for neo-liberal societies," *Health Information Science and Systems*, vol. 9, pp. 1–15, 2021.

[253] H. Badr, H. Du, M. Marshall, E. Dong, M. Squire, and L. M. Gardner, "Social Distancing is Effective at Mitigating COVID-19 Transmission in the United States," *medRxiv*, p. 2020.05.07.20092353, 2020.

[254] J. Y. Wu, B. D. Killeen, P. Nikutta, M. Thies, A. Zapaishchykova, S. Chakraborty, and M. Unberath, "Changes in Reproductive Rate of SARS-CoV-2 Due to Non-pharmaceutical Interventions in 1,417 U.S. Counties," *medRxiv*, p. 2020.05.31.20118687, 2020.

[255] P. Cintia, D. Fadda, F. Giannotti, L. Pappalardo, G. Rossetti, D. Pedreschi, S. Rinzivillo, P. Bonato, F. Fabbri, F. Penone, *et al.*, "The relationship between human mobility and viral transmissibility during the COVID-19 epidemics in Italy," *arXiv preprint arXiv:2006.03141*, 2020.

[256] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann, "Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions," *Science*, vol. 15, p. eabb9789, May 2020.

[257] A. Aleta and Y. Moreno, "Evaluation of the potential incidence of COVID-19 and effectiveness of containment measures in Spain: a data-driven approach," *BMC Med.*, vol. 18, pp. 1–12, May 2020.

[258] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of COVID-19 explain inequities and inform reopening," *Nature*, pp. 1–6, 2020.

[259] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, "Superspreading and the effect of individual variation on disease emergence," *Nature*, vol. 438, pp. 355–359, Nov 2005.

[260] D. C. Adam, P. Wu, J. Y. Wong, E. H. Y. Lau, T. K. Tsang, S. Cauchemez, G. M. Leung, and B. J. Cowling, "Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong," *Nature Medicine*, 2020.

[261] B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu, "Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2," *arXiv*, May 2020.

[262] A. Chande, S. Lee, M. Harris, Q. Nguyen, S. J. Beckett, T. Hilley, C. Andris, and J. S. Weitz, "Real-time, interactive website for US-county-level COVID-19 event risk assessment," *Nature Human Behaviour*, pp. 1–7, 2020.

[263] R. Laxminarayan, B. Wahl, S. R. Dudala, K. Gopal, C. M. B, S. Neelima, K. S. J. Reddy, J. Radhakrishnan, and J. A. Lewnard, "Epidemiology and transmission dynamics of COVID-19 in two Indian states.," *Science (New York, N.Y.)*, vol. 370, no. 6517, pp. 691–697, 2020.

[264] U. S. C. Bureau, "Core-Based Statistical Areas." https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html, 2019.

[265] U.S. Bureau of Labor Statistics, "Quarterly Census of Employment and Wages." https://www.bls.gov/cew/data.htm, 2020. Accessed 16-02-2021.

[266] "New York City Public Schools to Close to Slow Spread of Coronavirus," 2020. [Online; accessed 03. Dec. 2020].

[267] "New York City Mayor de Blasio Considering Shelter in Place," 2020. [Online; accessed 03. Dec. 2020].

[268] "Schools Shut in Seattle Area as Coronavirus Spreads," 2020. [Online; accessed 03. Dec. 2020].

[269] "Mayoral proclamation of civil emergency," 2020. [Online; accessed 03. Dec. 2020].

[270] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect. Dis.*, vol. 20, pp. 533–534, May 2020.

[271] "Commercial Laboratory Seroprevalence Survey Data," 2020. [Online; accessed 11. Sep. 2020].

[272] S. Pei, S. Kandula, and J. Shaman, "Differential effects of intervention timing on COVID-19 spread in the United States," *Science Advances*, vol. 6, no. 49, p. eabd6370, 2020.

[273] A. Endo, n. null, S. Abbott, A. Kucharski, and S. Funk, "Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China," *Wellcome Open Research*, vol. 5, no. 67, 2020.

[274] K. Sun, W. Wang, L. Gao, Y. Wang, K. Luo, L. Ren, Z. Zhan, X. Chen, S. Zhao, Y. Huang, Q. Sun, Z. Liu, M. Litvinova, A. Vespignani, M. Ajelli, C. Viboud, and H. Yu, "Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2," *medRxiv*, p. 2020.08.09.20171132, Nov 2020.

[275] C. I. Jarvis, K. V. Zandvoort, A. Gimma, K. Prem, M. Auzenbergs, K. O'Reilly, G. Medley, J. C. Emery, R. M. G. J. Houben, N. Davies, E. S. Nightingale, S. Flasche, T. Jombart, J. Hellewell, S. Abbott, J. D. Munday, N. I. Bosse, S. Funk, F. Sun, A. Endo, A. Rosello, S. R. Procter, A. J. Kucharski, T. W. Russell, G. Knight, H. Gibbs, Q. Leclerc, B. J. Quilty, C. Diamond, Y. Liu, M. Jit, S. Clifford, C. A. B. Pearson, R. M. Eggo, A. K. Deol, P. Klepac, G. J. Rubin, and W. J. Edmunds,

"Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK," *BMC Medicine*, vol. 18, no. 1, p. 124, 2020.

[276] J. Lu, J. Gu, K. Li, C. Xu, W. Su, Z. Lai, D. Zhou, C. Yu, B. Xu, and Z. Yang, "COVID-19 Outbreak Associated with Air Conditioning in Restaurant, Guangzhou, China, 2020," *Emerging Infectious Diseases*, vol. 26, no. 7, pp. 1628–1631, 2020.

[277] K. A. Fisher, M. W. Tenforde, L. R. Feldstein, C. J. Lindsell, N. I. Shapiro, D. C. Files, K. W. Gibbs, H. L. Erickson, M. E. Prekker, J. S. Steingrub, M. C. Exline, D. J. Henning, J. G. Wilson, S. M. Brown, I. D. Peltan, T. W. Rice, D. N. Hager, A. A. Ginde, H. K. Talbot, J. D. Casey, C. G. Grijalva, B. Flannery, M. M. Patel, W. H. Self, I. N. Investigators, C. C.-. R. Team, K. W. Hart, R. McClellan, H.-n. Tan, A. Baughman, N. A. Hennesy, B. Grear, M. Wu, K. Mlynarczyk, L. Marzano, Z. Plata, A. Caplan, S. M. Olson, C. E. Ogokeh, E. R. Smith, S. S. Kim, E. P. Griggs, B. Richards, S. Robinson, K. Kim, A. M. Kassem, C. N. Sciarratta, and P. L. Marcet, "Community and Close Contact Exposures Associated with COVID-19 Among Symptomatic Adults $\geq$ 18 Years in 11 Outpatient Health Care Facilities - United States, July 2020," *Morbidity and Mortality Weekly Report*, vol. 69, no. 36, pp. 1258–1264, 2020.

[278] F.-Y. Lan, C. Suharlim, S. N. Kales, and J. Yang, "Association between sars-cov-2 infection, exposure risk and mental health among a cohort of essential retail workers in the usa," *Occupational and environmental medicine*, vol. 78, no. 4, pp. 237–243, 2021.

[279] R. A. Shumsky, L. Debo, R. M. Lebeaux, Q. P. Nguyen, and A. G. Hoen, "Retail store customer flow and covid-19 transmission," *Proceedings of the National Academy of Sciences*, vol. 118, no. 11, 2021.

[280] Z. Susswein and S. Bansal, "Characterizing superspreading of sars-cov-2: from mechanism to measurement," *medRxiv*, 2020.

[281] M. Weed and A. Foad, "Rapid Scoping Review of Evidence of Outdoor Transmission of COVID-19," *medRxiv*, p. 2020.09.04.20188417, Sep 2020.

[282] S. Hu, W. Wang, Y. Wang, M. Litvinova, K. Luo, L. Ren, Q. Sun, X. Chen, G. Zeng, J. Li, L. Liang, Z. Deng, W. Zheng, M. Li, H. Yang, J. Guo, K. Wang, X. Chen, Z. Liu, H. Yan, H. Shi, Z. Chen, Y. Zhou, K. Sun, A. Vespignani, C. Viboud, L. Gao, M. Ajelli, and H. Yu, "Infectivity, susceptibility, and risk factors associated with SARS-CoV-2 transmission under intensive contact tracing in Hunan, China," *medRxiv*, p. 2020.07.23.20160317, Nov 2020.

[283] C. Rhee, M. Baker, V. Vaidya, R. Tucker, A. Resnick, C. A. Morris, M. Klompas, C. P. E. Program, *et al.*, "Incidence of nosocomial covid-19 in patients hospitalized at a large us academic medical center," *JAMA network open*, vol. 3, no. 9, pp. e2020498–e2020498, 2020.

[284] A. Richterman, E. A. Meyerowitz, and M. Cevik, "Hospital-acquired sars-cov-2 infection: lessons for public health," *Jama*, vol. 324, no. 21, pp. 2155–2156, 2020.

[285] J. I. Tokars, S. J. Olsen, and C. Reed, "Seasonal incidence of symptomatic influenza in the united states," *Clinical Infectious Diseases*, vol. 66, no. 10, pp. 1511–1518, 2018.

[286] J. A. Backer, D. Klinkenberg, and J. Wallinga, "Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020," *Eurosurveillance*, vol. 25, no. 5, p. 2000062, 2020.

[287] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-m. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, and N. M. Linton, "Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)," *International Journal of Infectious Diseases*, vol. 94, pp. 154–155, May 2020.

[288] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, *et al.*, "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," 2020.

[289] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch, "Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period," *Science*, vol. 368, no. 6493, pp. 860–868, 2020.

[290] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers, "Serial Interval of COVID-19 among Publicly Reported Confirmed Cases," *Emerging Infectious Diseases journal*, vol. 26, no. 6, 2020.

[291] "Coronavirus Disease 2019 (COVID-19) planning scenarios," Dec 2020. [Online; accessed 15. Dec. 2020].

Appendix A

Supplementary Materials: Social Epidemic Sensors

A.1 Data processing

In Figure A.1 we can see our data processing pipeline is composed of three different stages. First, data scrapping and storage of raw data both from the Twitter API and the surveillance system for influenza in Spain (ScVGE) [203]. Second, data cleaning. This stage has a ILI keyword-based filtering step, where we look for ILI-related spanish words, such as "gripe", "gripazo", "trancazo", "catarro" and "constipado", and a second step with a first person ILI-related posts text classifier. Then our pipeline is divided in two branches. One for building autoregressive linear models for explaining and predicting official weekly ILI cases, where is required a previous step for grouping by weeks the transactional data, before feeding our linear models. The second branch builds several logistic linear models at individual level for explaining and predicting potential sensors from the network. Before feeding the models, we create some behavioural variables, based on social activity, mobility and content posting.

A.2 Centrality sensitivity analysis

Figure A.2 shows the time series of different centrality metrics of the users making ILI-related posts compared with the total number of them. We show the average over

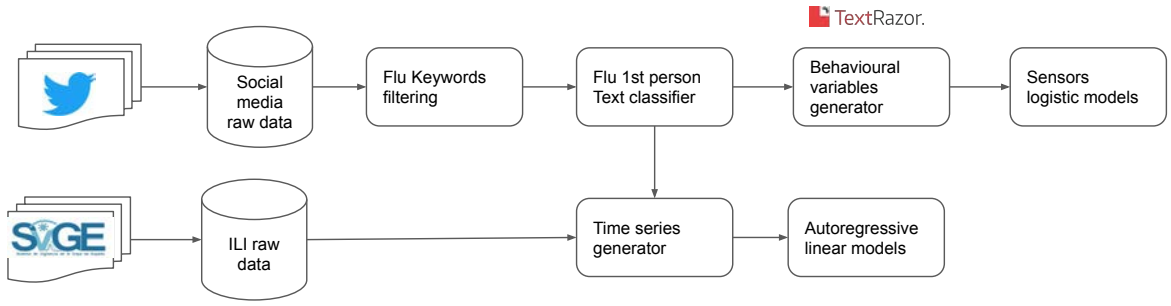

Figure A.1: Data processing pipeline schematic view. From left to right, raw data scrappers and storage, data filtering, data enrichment, time series generator and models. Figure reproduced from [1].

during the 2013, 2014, and 2015 seasons of ILI in Spain. Time series are centered around their maximum peak within the season. For the centrality metrics, we show the weekly total out-degree ($D_{N,t}$ in the main text) compared with the weekly average out-degree ($D_{N,t}$ divided over the number of users making ILI-related posts) and the weekly median of the out-degree of users making ILI-related posts. As we can see the total out-degree clearly follows the number of ILI-related mentions and shows a large spike weeks before the peak. That peak is also observed for the average degree. However, it is not present in the median. These results show that high-connected people have ILI-related posts in the social network at different times than the rest of the people. And it only appears in the total or average degree, since those estimators are more susceptible to large out-degree users than the median out-degree.

To make that difference more quantitative in Figure A.3 we show the difference on the average of those centrality metrics 18 weeks before and after the peak. We applied the statistical test t-test to see if there are differences between the groups. Figure A.3. As we can see, the total out-degree is the one that shows clearly more difference before and after the peak. For that reason, we choose it in the main paper.

A.3 Sensors selection analysis

To define sensors, Figure A.4 shows the out-degree distribution of all the users having ILI mentions. Again, it follows a power-law distribution with an exponent of 2.56 (CI

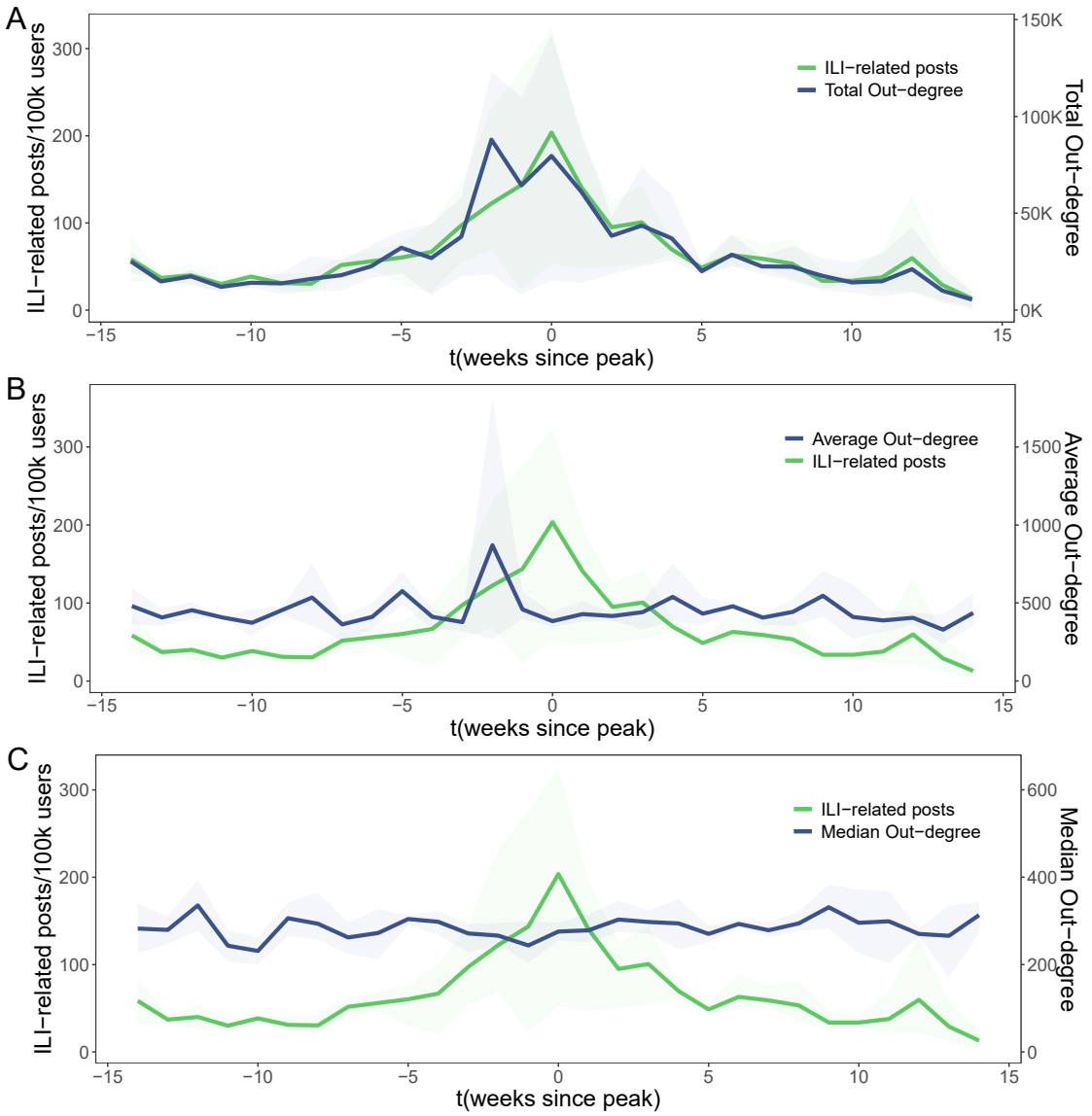

Figure A.2: Generalized ILI-related posts against weekly centrality statistics. Horizontal axis measures weeks from the peak. Green solid lines show the average incidence across seasons of weekly ILI-related posts (left Y-axis). Blue lines represent different weekly centrality metrics from individuals posting a first-person ILI-related post (right Y-axis). (A) is the weekly total out-degree, (B) is their average out-degree, and (C) is the median out-degree of those individuals. Shaded area are the confidence intervals over the different seasons. Figure reproduced from [1].

$[2.51, 2.62]$). Based on this distribution, we defined four out-degree thresholds to test different groups of sensors. Out of the users making ILI-related mentions, we defined

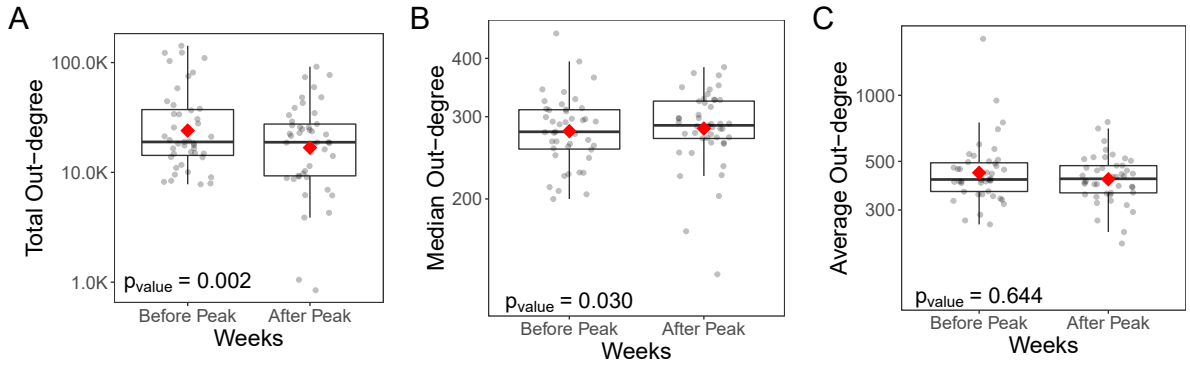

Figure A.3: Comparison of degree metrics before and after peak groups. Points correspond to different weeks grouped by before and after the peak. We also show the box-plot for each group, including their median (horizontal thick line) and mean (red diamonds). Vertical axis are centrality metrics: (A) Total out-degree, (B) Median out-degree and (C) Average out-degree. The p-value shows the t-test statistic comparing the means for the groups before and after the peak. Figure reproduced from [1].

a sensor as a user with an out-degree greater than 100, 500, 1000, or 2000 (vertical dashed lines in Figure A.4), and as control otherwise. Figures A.4B shows the results for the cumulative incidence of the ILI-related mentions for each of the out-degree sensor thresholds. As we can see, for an out-degree threshold greater than 1000, the cumulative ILI-related mentions incidence for the sensor group is ahead and starts to grow one or two weeks before the control group. Therefore, we selected the out-degree threshold to be 1000 from now onward in our study.

A.4 Agent-based model of ILI disease and information diffusion

The values of all the parameters used in our Agent-Based Model (ABM) simulating the Susceptible-Infected-Recovery epidemic spreading on a complex network are given in Table A.1. The synthetic network is generated by the Barabasi-Albert model using the `igraph` R package [200]. We simulated different realizations of the epidemic model, see Figure A.5, emulating different ILI seasons. To compare the temporal dynamics of our ABM with the real ILI epidemics, we rescale the time in our model to mimic the epidemic dynamics in the empirical data. As we can see in Figure A.5 equating four

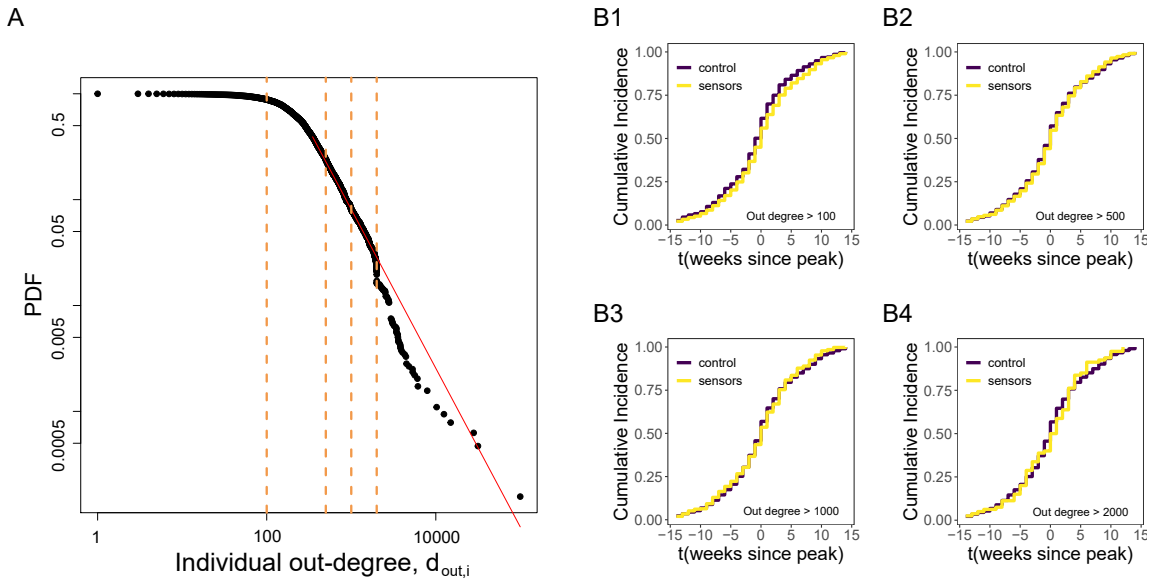

Figure A.4: Out-degree sensitivity analysis.(A) Users out-degree $d_{out,i}$ frequency distribution. Vertical orange dashed lines define the thresholds selected with out-degrees 100, 500, 1000, and 2000. (B) Empirical cumulative distribution differences in ILI-related mentions on Twitter between the sensor and randomly chosen individuals for each of the degree thresholds selected. The purple line correspond to the group of randomly chosen individuals and the yellow line is sensor group selected with an out-degree bigger than 100 (B1), 500 (B2), 1000 (B3) and 2000 (B4). Figure reproduced from [1].

time units in our simulations to one week, the epidemic curves have the same shape as the real ILI-related cases.

We also assume that each agent posts on a social media platform and that those tweets are ILI-related when he gets infected. To incorporate our hypothesis that offline and online networks degrees are correlated, we assume that

$$d_i^{\text{Twitter}} = d_i^{\text{Offline}}(1 + \nu_i) \tag{A.1}$$

where ν_i is a random number uniformly distributed between 0 and 1. This way we account for potential variability between offline and online degrees, while still getting a moderate correlation between them.

In this ABM model we assume that sensors in the social media platform are those with $d_i^{Twitter} \geq \chi$, see Table A.1. Figure A.5.C shows the average total out degree for all

Parameters	Description	Value
N	Number of nodes in graph	150k
n	Initial seeds	2
β	Infection probability ($S \rightarrow I$)	10% [285]
ϵ	Latent period	3 days [285]
α	Recovery probability ($I \rightarrow R$)	$1 \ / \ \epsilon$
χ	Sensors degree threshold	12

Table A.1: SIR Agent-based model parameters. Baseline set of parameters. Table reproduced from [1].

users and those in the sensor group compared with the number of infected agents. As expected, users with larger degree get infected earlier and those in the sensor group even a little bit earlier, as we saw in the real data.

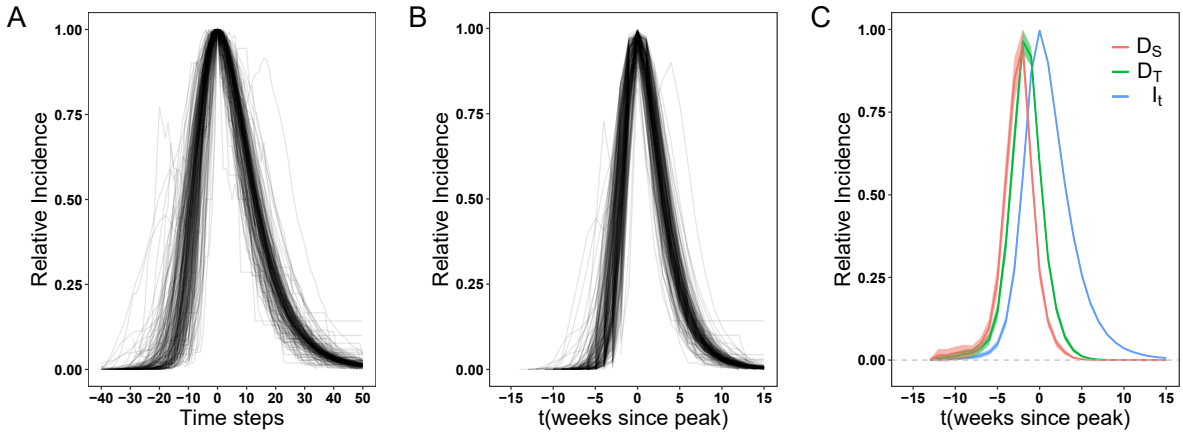

Figure A.5: Agent-based simulations of ILI disease and information diffusion. (A) Shows the incidence curves for the spreading of the diseases for each simulation in our original time scale. Time is centered around the peak for each simulation and we show the relative incidence to its maximum (peak). (B) Same as in A), but with time steps converted into weeks to compare with real ILI-cases. (C) Average total out-degree in the social network for all agents (D_T) and those in the sensor group (D_S) compared with the incidence of the disease. Figure reproduced from [1].

	Dependent variable:			
	Sensor			
	Content	Network	Mobility	All
	(1)	(2)	(3)	(4)
Association	−0.263*** (0.099)			
Basketball				−0.237** (0.098)
Christianity	0.116** (0.053)			0.082 (0.054)
Christmas	0.134** (0.060)			0.097 (0.061)
Easter	−0.144** (0.066)			−0.133** (0.067)
Entertainment	−0.154* (0.093)			−0.164* (0.094)
Folk	−0.237*** (0.080)			−0.225*** (0.081)
Government	0.125 (0.081)			0.189*** (0.058)
Human	0.118** (0.058)			0.103* (0.058)
Language	0.218*** (0.057)			0.156*** (0.058)
Music	0.499*** (0.127)			0.461*** (0.129)
National	0.292*** (0.113)			0.168* (0.093)
Organisations	0.156** (0.074)			0.109 (0.072)
Philosophical	0.090 (0.057)			
Politics	0.133* (0.081)			
Popular	−0.160** (0.076)			−0.154** (0.077)
Soccer	−0.138 (0.087)			
Out-degree		0.549** (0.233)		0.511** (0.228)
Number of posts		0.461*** (0.073)		0.315*** (0.074)
Radius of gyration			0.113** (0.053)	0.104* (0.056)
Constant	−0.492*** (0.056)	−0.444*** (0.055)	−0.463*** (0.054)	−0.470*** (0.057)
Observations	1,460	1,460	1,460	1,460
Accuracy	0.605	0.607	0.615	0.636
Accuracy CI	(0.568, 0.632)	(0.567, 0.645)	(0.575, 0.653)	(0.596, 0.673)
Log Likelihood	−926.504	−941.322	−971.694	−913.783
Akaike Inf. Crit.	1,887.008	1,888.645	1,947.388	1,861.566
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A.2: Sensor models. Logistic regression models for sensors characterization based on content, network and mobility features. Table reproduced from [1].

A.5 Sensors logistic regression model

In table A.2 we can see the coefficients of the logistic regression models to explain and identify a single node as a sensor. We used three group of variables, a categorization of the content published, user's mobility and user's network features including their out-degree and number of posts.

A.6 Data and materials availability

All data needed to evaluate the conclusions in this study are present in chapter 2, this appendix and the following repository. Access to the github repository.

Appendix B

Supplementary Materials: Data-Driven Contact Networks

B.1 Calibration of intra-layer links

As described in the main text, we define ω_{ij} as the weight associated to the link between node i and j.

In the community+workplace layer, we estimated the mean number of daily effective contacts (η_C) by using $\omega_{C_{ij}}$, which is based on the co-presence probability estimation. η_C can thus be calculated as

$$\eta_C = 1/n \sum_{i \in \{1,\dots,n\}} \sum_{j | j \neq i \wedge j \in \{1,\dots,n\}} \omega_{C_{ij}}. \tag{B.1}$$

By construction, the weights for the household layer were assigned as $\omega_{H_{ij}} = 1/(h-1)$, where h is the number of household members so that $\eta_H = 1$. Analogously, by construction, for schools, we have $\eta_S = 1$.

It is important to note that $\eta_{C,H,S}$ refer to the mean number of daily effective contacts in the synthetic (non-calibrated) network. Based on the analysis of contact survey data from 9 countries [228, 229, 231–233], the estimated number of daily effective contacts by social setting is 10.86 in community+workplace, 4.11 in household, and 11.41 in school. To calibrate the weights of intra-layer links ($\hat{\omega}_l$), we associate to each layer a single rescaling factor w_l such that the mean number of daily effective contacts in that layer matches mean number of daily effective contacts in the

Param.	Description	Age	Value	Ref.
r	relative infectiousness of asymptomatic people	-	50%	†
ϵ^{-1}	latent period	-	3 days	[286]
ϵ'^{-1}	latent period	-	5 days	[286]
p	proportion of asymptomatic	-	25%	[287]
γ^{-1}	pre-symptomatic period	-	2 days	[286]
μ^{-1}	time to removed/home stay	-	2.5 days	$*$
α	symptomatic case hospitalization ratio (%)	0-4	0.0	[224]
		5-17	0.025	
		18-49	2.672	
		50-64	9.334	
		65+	15.465	
χ	ICU % among hospitalized	0-4	5.0	[288]
		5-17	5.0	
		18-49	5.38	
		50-64	17.10	
		65+	44.71	
δ^{-1}	days from home stay to hospital admission	-	2	[289]
μ_H^{-1}	days in hospital	-	8	[224]
μ_{ICU}^{-1}	days in ICU	-	13	[224]
k	proportion of presymptomatic transmission	-	15%	[290]
R_0	basic reproduction number	-	2.5	†
β	transmission for symptomatic and asymptomatic people	-	$\frac{R_0\mu}{pr+(1-p)/(1-k)}$	
β_S	transmission for pre-symptomatic people	-	$\frac{\beta\gamma k}{\mu(1-k)}$	

Table B.1: SARS-CoV-2 transmission model baseline set of parameters. †: assumed;$*$: calibrated to the generation time T_g. Table reproduced from [3].

corresponding social setting. Therefore, the calibrated mean number of daily effective contacts in the community+workplace layer $\hat{\eta}_C$ is

$$\hat{\eta}_C = 1/n \sum_{i\in\{1,\dots,n\}} \sum_{j|j\neq i\wedge j\in\{1,\dots,n\}} \omega_{C_{ij}}\, w_{C+W} = 1/n \sum_{i\in\{1,\dots,n\}} \sum_{j|j\neq i\wedge j\in\{1,\dots,n\}} \hat{\omega}_{C_{ij}} \tag{B.2}$$

where $w_{C+W} = 10.86/\eta_C$. Analogously for household and school layers, we obtain $w_H = 4.11/\eta_H$ and $w_S = 11.41/\eta_S$.

B.2 SARS-CoV-2 transmission model

The values of all the disease parameters used for simulating the transmission dynamics are given in table B.1. Figure B.1 shows the numerical distributions of these parameters as resulting from simulations of the model.

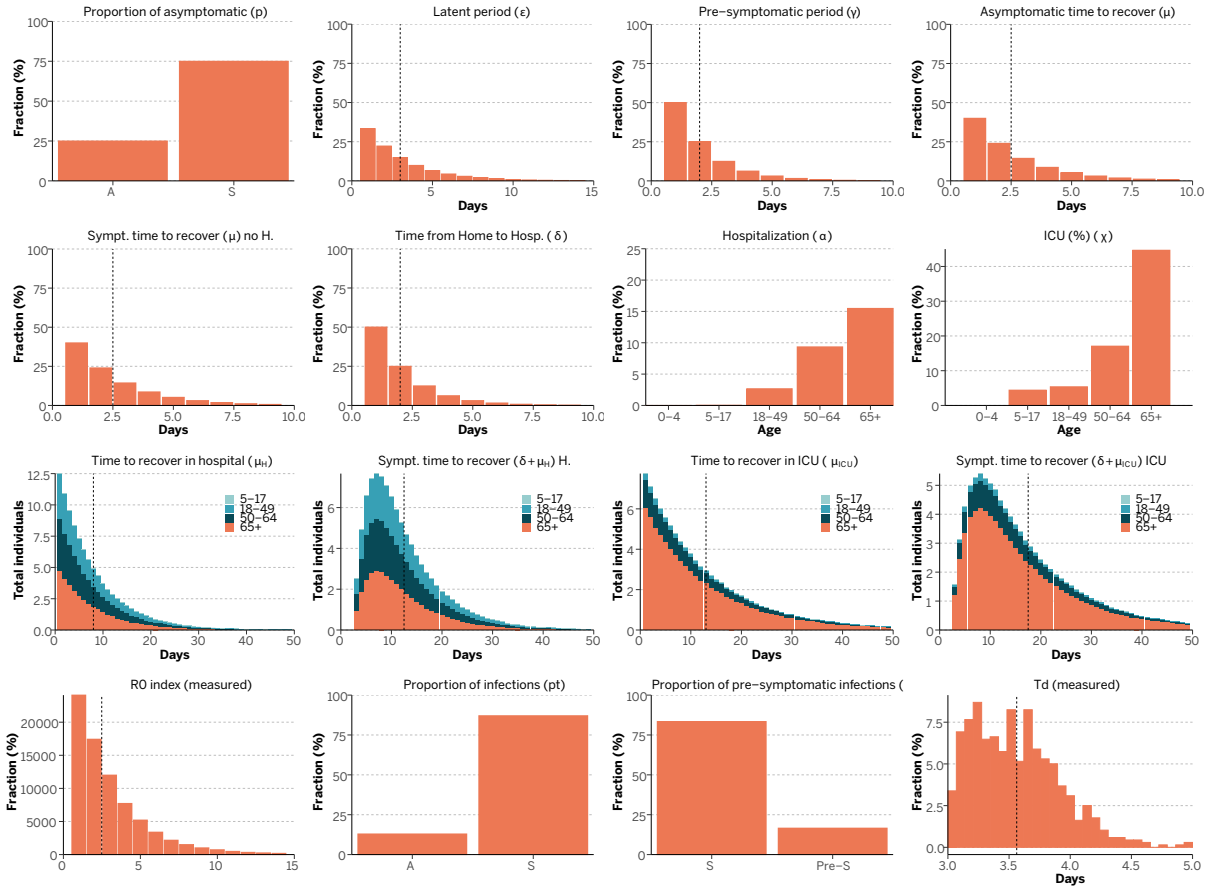

Figure B.1: Model's parameters and distributions summary. Figure reproduced from [3].

B.3 Epidemiological results of the COVID-19 What-if scenarios

Results for the unmitigated scenario are shown in Figure B.2, panels a-c (see the last page of this Appendix). A COVID-19 unmitigated epidemic would have a peak of daily incidence of 25.2 (95% C.I: 23.8-26.4) newly infected individuals per 1,000 people. The epidemic follows a typical trajectory, namely, when the effective reproduction number R_t as a function of time (panel c) becomes smaller than 1, the transmission dynamics slow down and eventually vanish after having infected about 75% of the population (Figure B.2.b In Figure B.2.d, we show that following the lifting of social distancing the infection incidence starts to increase again, and the effective reproductive number, that dropped by circa 75% and reached values below 1 with the

intervention, increases to values up to 2.05 (95%CI: 1.73-2.47) (see Figure B.2.f). Figure B.2.g shows results obtained for different levels of tracing (no tracing, 20% and 40%) of the contacts of the symptomatic isolated COVID-19 cases. By comparing Figure B.2.d with Figure B.2.g (for no tracing), we find that quarantining households of symptomatic subjects alone is not enough to significantly change the course of the epidemic and the conclusions reached for the first of these scenarios.

B.4 Data and materials availability

The data that support the findings of chapter 3 are available from Cuebiq through their Data for Good program, but restrictions apply to the availability of these data, which were used under licenses for the current study, and so are not publicly available. Aggregated data used in the models are however available from the authors upon reasonable request and permission of Cuebiq. Other data used comes from the American Community Survey (5-year) from the Census, which is publicly available.

The epidemiological model is out of the scope of this thesis, if interested in more detailed information on the calibration process, model specifications, and sensitivity analysis of our results, please refer to the original article [3] and its Supplementary Materials. These resources provide in-depth insights that go beyond the scope of this thesis.

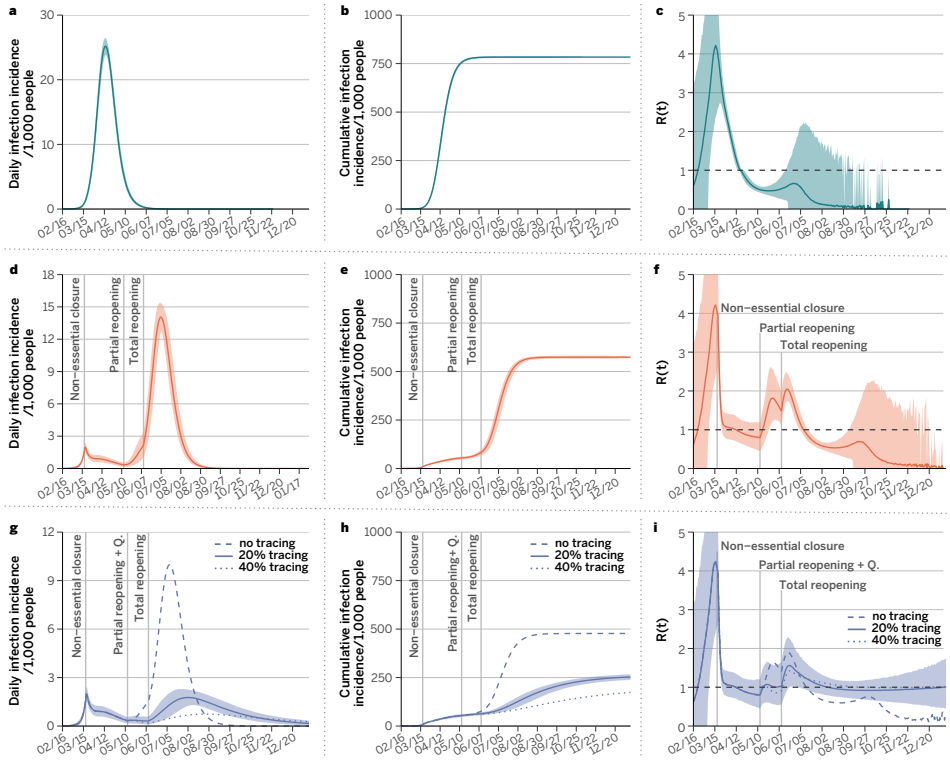

Figure B.2: Impact of COVID-19 under different scenarios. Evolution of the number of new cases (**a, d, g**), the outbreak size (**b, e, h**), and the effective reproductive number (**c, f, i**) as a function of time in each situation studied. Results of the SARS-CoV-2 transmission dynamics are shown for the unmitigated scenario (top panels **a-c**), and the two social distancing interventions considered, LIFT (**d-f**) and LET scenarios (**g-h**). In both cases, we considered the closure of schools and non-essential places for 8 weeks. This is the strictest lock-down period, which is followed by a partial lifting of the stay-at-home policy whose duration is set to 4 weeks. During the partial lifting, all places in the community layer are open except mass-gathering locations. Finally, a full reopening takes place after the period of partial lifting ends (relevant events are marked with vertical lines). Panels **d-f** consider that no other measures are adopted concurrently to the lifting of the restrictions, whereas the results in panels **g-i** have been obtained when the reopening is accompanied by an active policy consisting of testing the symptomatic individuals, home isolating them, and quarantining their household and the households of a fraction of their contacts, as indicated in the legend of the bottom panels. Note that the vertical scales of panels **a**, **d**, and **g** are not the same and that both the number of new cases and total cases are per 1,000 inhabitants. In all panels the solid line represents the average over 10,000 simulations and the shaded region the 95% C.I. Figure reproduced from [3].

Appendix C

Supplementary Materials: Temporal Contact Networks

C.1 SARS-CoV-2 transmission model

The values of all the disease parameters used for simulating the transmission dynamics are given in table C.1.

C.2 Superspreading events

In heterogeneous populations it is possible for an infected individual to produce an usually large number of secondary cases. This is known as a super-spreading event (SSE). To define a SSE we follow Lloyd-Smith et al [259]:

1. Estimate the effective reproduction number, R
2. Compute a Poisson distribution with mean R
3. Define a SSE as any infected individual who infects more than the 99-th percentile of the Poisson distribution within a certain category of place.

In C.1 we test the hypothesis of the 20/80 rule according to which 20% of the infected individuals produce 80% of the infections. Note that this does not imply that said 20% of individuals are super-spreaders. In fact, the large majority of them do not

Param.	Description	Age	Value	Ref.
r	relative infectiousness of asymptomatic people	-	50%	†
k	proportion of pre-symptomatic transmission	-	50%	[291]
ϵ^{-1}	incubation period (gamma distributed)	-	shape = 2.08 rate = 0.33	[282]
p	proportion of asymptomatic	-	40%	[291]
γ^{-1}	pre-symptomatic period	-	2 days	[286]
μ^{-1}	time to isolation	-	2.5 days	
δ^{-1}	days from isolation to death	-	12.5	[291]
IFR	infection fatality ratio	0-9	0.00161%	[224]‡
		10-19	0.00695%	
		20-29	0.0309%	
		30-39	0.0844%	
		40-49	0.161%	
		50-59	0.595%	
		60-69	1.93%	
		70-79	4.28%	
		≥ 80	7.80%	
T_n	Notification of death	-	7 days	[291]
θ	outdoor transmissibility	-	0.05	[281]

Table C.1: SARS-CoV-2 transmission model baseline set of parameters. †: assumed ;∗: calibrated to the generation time T_g; ‡ Only applied to symptomatic individuals. As such, a correction factor of 1/(1-p) is applied to all age groups. Table reproduced from [4].

produce any secondary infections, inline with what has been observed in highly detailed empirical studies [274].

In Table C.2 we report the probability of having a SSE within each category before and after the declaration of the National Emergency. We observe a drastic reduction of the probability after 03/13.

C.3 Behavioural sensitivity analysis

Distance to POIs

While constructing the network, we attributed a stay to a given POI if it was no further than 50 meters from the POI center. In this section we test more strict conditions for that attribution, i.e. a threshold of just 10 meters. Note that this more strict condition

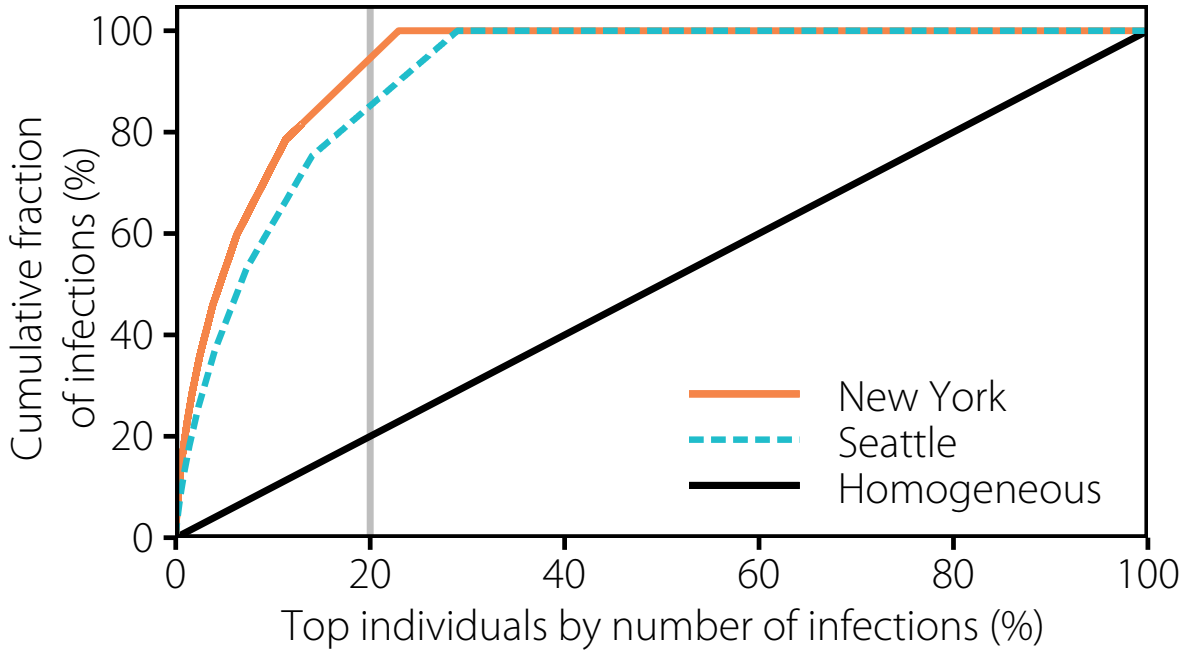

Figure C.1: Super-spread individuals distribution. Individuals are ranked according to the number of infections they produce. The cumulative fraction of infections found in both cities is compared with the one that would be obtained in a completely homogeneous system. Figure reproduced from [4].

for attribution lowers the number of potential visitors to the POI but also lowers the distance between people in the venue, making physical contact more likely. In Figure C.2 we show the results for this scenario.

A more restrictive definition of stay yield a much sparser network in the community layer, while it does not affect the rest of the layers. We can see that to obtain the observed number of deaths under these conditions, the fraction of infections attributed to the workplace layer is increased. Nevertheless, the distribution of infections across settings is fairly similar, signaling that the results are robust to this definition.

Behavioral changes

The aggregated change in behavior due to the evolution of the epidemic as well as the introduction of non-pharmaceutical interventions is already contained in the mobility data. This leads to the sudden drop in the number of contacts following the declaration

Probability of a super-spreading event (%)				
Category	New York		Seattle	
	Before 03/13	After 03/13	Before 03/13	After 03/13
Arts/Museum	7.30 [7.01-7.61]	0.52 [0.48-0.57]	0.31 [0.08-0.59]	0.00 [0.00-0.00]
Entertainment	2.42 [2.35-2.49]	0.14 [0.13-0.15]	2.16 [1.72-2.63]	0.21 [0.06-0.39]
Exercise	1.96 [1.91-2.03]	0.34 [0.32-0.36]	1.14 [0.88-1.43]	0.77 [0.51-1.06]
Food/Beverage	0.53 [0.51-0.55]	0.17 [0.17-0.18]	0.17 [0.11-0.23]	0.13 [0.10-0.17]
Grocery	2.18 [2.12-2.24]	1.31 [1.30-1.33]	0.58 [0.37-0.81]	0.93 [0.85-1.02]
Health	0.14 [0.12-0.16]	0.11 [0.11-0.12]	0.00 [0.00-0.00]	0.06 [0.02-0.10]
Other	1.61 [1.54-1.67]	0.10 [0.09-0.10]	0.40 [0.21-0.62]	0.04 [0.00-0.12]
Outdoors	0.03 [0.01-0.06]	0.00 [0.00-0.01]	0.00 [0.00-0.00]	0.00 [0.00-0.00]
Service	0.59 [0.56-0.62]	0.18 [0.17-0.18]	0.01 [0.00-0.02]	0.10 [0.07-0.13]
Shopping	1.43 [1.39-1.47]	0.84 [0.83-0.85]	0.14 [0.05-0.27]	0.09 [0.06-0.11]
Sports/Events	8.73 [8.32-9.14]	4.27 [3.90-4.66]	0.22 [0.00-0.56]	0.00 [0.00-0.00]
Transportation	0.26 [0.21-0.31]	0.04 [0.03-0.05]	0.00 [0.00-0.00]	0.00 [0.00-0.00]
All	1.73 [1.71-1.75]	0.71 [0.70-0.71]	0.93 [0.84-1.02]	0.34 [0.32-0.37]

Table C.2: Super-spreading events by POIs categories. Probability that an individual will cause a super-spreading event as defined in [259]. We aggregate all the infections produced by each individual within each category for the given period of time, and compute the fraction of individuals who produce a super-spreading event out of the total number of individuals infecting someone in that category. In brackets the 95% C.I. computed using a bootstrap percentile method is shown. Table reproduced from [4].

of the National Emergency. However, at the individual level, it might be possible that some individuals in the dataset lowered their contacts due to having developed symptoms, even if in our simulations they do not get infected at all and vice versa. But for anonymity reasons, it is not possible to relate the medical history of individuals and our agents and, thus, we cannot know the reason why an individual might have changed her behavior. From the point of view of the individual this observation is important, but since we are working on aggregated metrics this observation does not affect the results.

To demonstrate this, in Figure C.3, we show the results in which we completely remove symptomatic transmission. This extreme scenario would represent a situation in which every time an individual develops symptoms, she gets completely isolated. As we can see, the overall results are close to the ones we have presented so far. The reason is that our model is fitted to the number of deaths and, thus, the total number of infections is fixed (as a function of IFR). If we remove one type of transmission, then

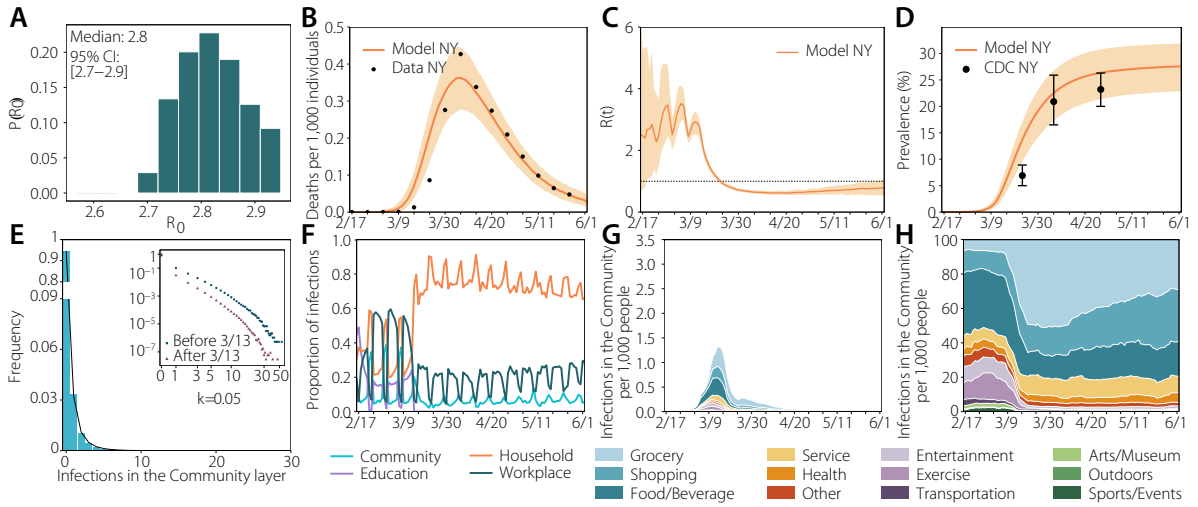

Figure C.2: Infections by distance to POIs (10 meters). Results with a more restricted definition of stay for the case of New York (10 meters): (a) estimated R_0; (b) number of deaths (fit); (c) estimated R_t; (d) prevalence; (e) distribution of infections; (f) proportion of infections per layer; (g) infections per setting; (h) normalized infections per setting. Figure reproduced from [4].

the transmissibility of the other types has to be increased to achieve the same number of deaths, yielding similar results.

Economic and age bias

The complete sample of users is slightly biased towards higher income individuals. Specifically, the penetration ratio (number of mobile phone users to adult population) in each census tract is correlated with the median household income, $\rho = 0.28 \pm 0.02$ in NY and $\rho = 0.18 \pm 0.02$ in Seattle metro areas. However the correlation of the penetration ratio with the number of people above 64 years old in each census tracts is small $\rho = 0.17 \pm 0.04$ in the NY area or not significant $\rho = -0.06 \pm 0.11$ in the Seattle area. To analyze the impact of this bias, we have investigated the dynamics of our model in a different set of users obtained by downsampling each economic groups (median income quartiles in each metro area) to have a better representation of them. In C.4 we report the results obtained using this new sample of users. As we can see, the results remain largely unaltered, signaling that the distribution of contacts per type of venue is not affected by this bias.

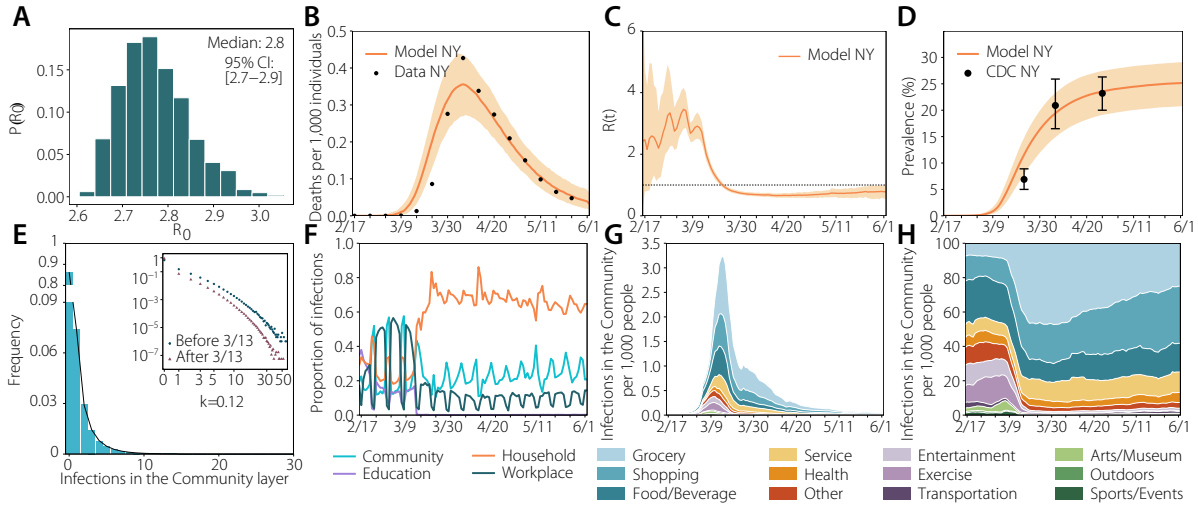

Figure C.3: Infections without symptomatic transmission. Main results in New York without symptomatic transmission: (a) estimated R_0; (b) number of deaths (fit); (c) estimated R_t; (d) prevalence; (e) distribution of infections; (f) proportion of infections per layer; (g) infections per setting; (h) normalized infections per setting. Figure reproduced from [4].

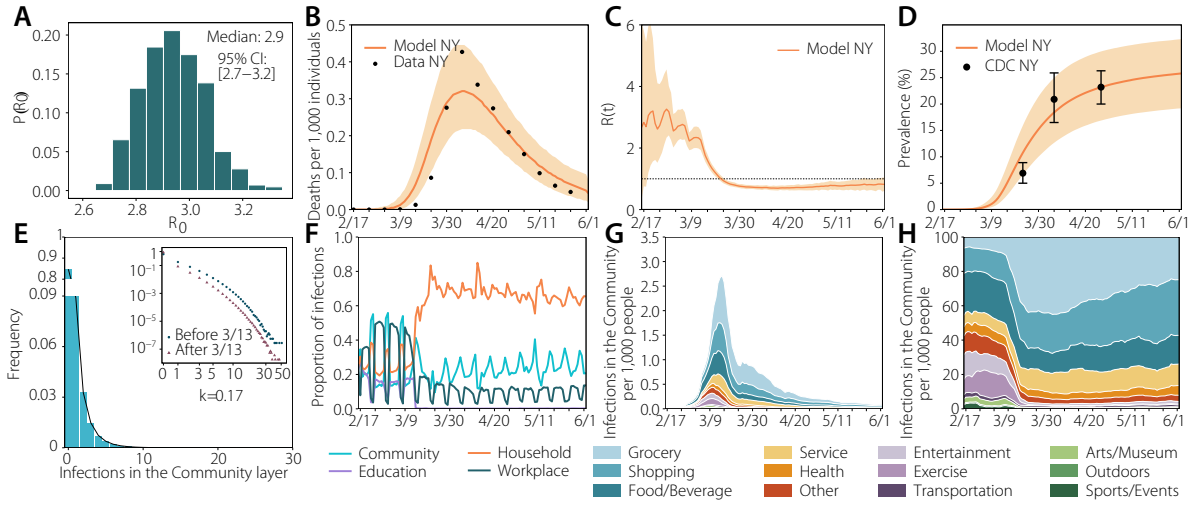

Figure C.4: Infections without economic and age bias. Results with a resampled population to remove economic bias in New York: (a) estimated R_0; (b) number of deaths (fit); (c) estimated R_t; (d) prevalence; (e) distribution of infections; (f) proportion of infections per layer; (g) infections per setting; (h) normalized infections per setting. Figure reproduced from [4].

Longer stays

We have tested the sensitivity of the results with a more strict definition of stay (minimum 15 minutes instead of 5 minutes), Figure C.5. We observe a slight increase in the Arts & Museums category before the declaration of the National Emergency, and one in the Grocery category after the declaration. This indicates that individuals tended to stay for longer in groceries in this period, but the rest of the results remain largely unaffected.

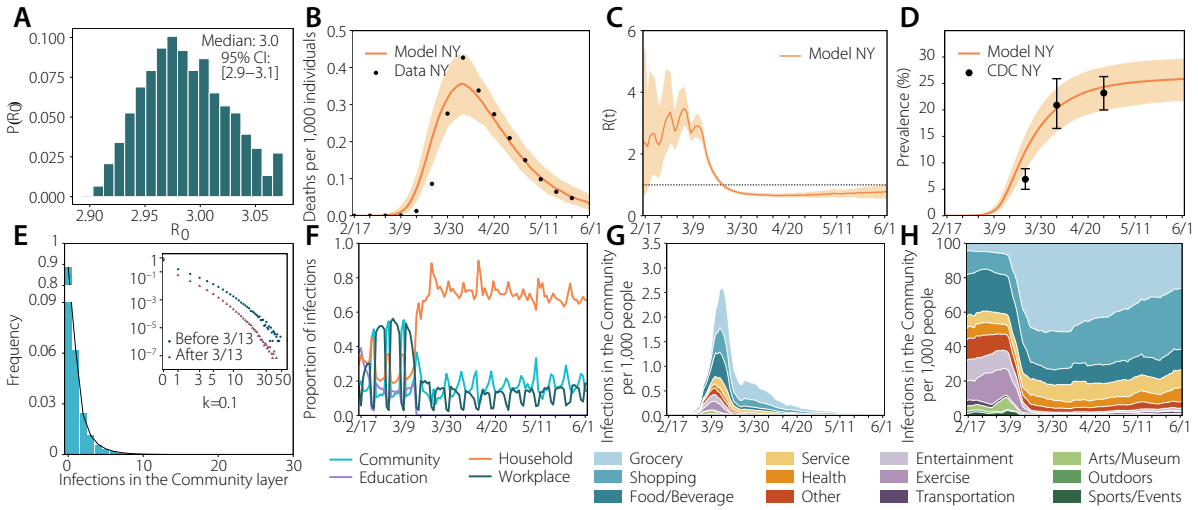

Figure C.5: Infections with longer stays. Results with stricter definition of stay in New York (minimum 15 minutes): (a) estimated R_0; (b) number of deaths (fit); (c) estimated R_t; (d) prevalence; (e) distribution of infections; (f) proportion of infections per layer; (g) infections per setting; (h) normalized infections per setting. Figure reproduced from [4].

C.4 Data and materials availability

The original mobility database is not publicly available due to license restrictions, but can be obtained from Cuebiq through their COVID-19 Data Collaborative. The anonymized temporal contact matrices in each layer for each city and the code to reproduce our results are publicly available on github. Access to the github repository.

The epidemiological model is out of the scope of this thesis, if interested in more

detailed information on the calibration process, model specifications, and sensitivity analysis of our results, please refer to the original article [4] and its Supplementary Materials. These resources provide in-depth insights that go beyond the scope of this thesis.