Transmission Modeling with Smartphone-based Sensing

A dissertation submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Weicheng Qian

Permission to Use

In presenting this dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my dissertation work was done. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my dissertation.

Disclaimer

Reference in this dissertation to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this dissertation in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Infectious disease spread is difficult to accurately measure and model. Even for well-studied pathogens, uncertainties remain regarding the dynamics of mixing behavior and how to balance simulation-generated estimates with empirical data. Smartphone-based sensing data promises the availability of inferred proximate contacts, with which we can improve transmission models. This dissertation addresses the problem of informing transmission models with proximity contact data by breaking it down into three sub-questions.

Firstly, can proximity contact data inform transmission models? To this question, an extended-Kalman-filter enhanced System Dynamics Susceptible-Infectious-Removed (EKF-SD-SIR) model demonstrated the filtering approach, as a framework, for informing Systems Dynamics models with proximity contact data. This combination results in recurrently-regrounded system status as empirical data arrive throughout disease transmission simulations—simultaneously considering empirical data accuracy, growing simulation error between measurements, and supporting estimation of changing model parameters. However, as revealed by this investigation, this filtering approach is limited by the quality and reliability of sensing-informed proximate contacts, which leads to the dissertation's second and third questions—investigating the impact of temporal and spatial resolution on sensing inferred proximity contact data for transmission models.

GPS co-location and Bluetooth beaconing are two of those common measurement modalities to sense proximity contacts with different underlying technologies and tradeoffs. However, both measurement modalities have shortcomings and are prone to false positives or negatives when used to detect proximate contacts because unmeasured environmental influences bias the data. Will differences in sensing modalities impact transmission models informed by proximity contact data? The second part of this dissertation compares GPS- and Bluetooth-inferred proximate contacts by accessing their impact on simulated attack rates in corresponding proximate-contact-informed agent-based Susceptible-Exposed-Infectious-Recovered (ABM-SEIR) models of four distinct contagious diseases. Results show that the inferred proximate contacts resulting from these two measurement modalities are different and give rise to significantly different attack rates across multiple data collections and pathogens.

While the advent of commodity mobile devices has eased the collection of proximity contact data, battery capacity and associated costs impose tradeoffs between the frequency and scanning duration used for proximate-contact detection. The choice of a balanced sensing regime involves specifying temporal resolutions and interpreting sensing data—depending on circumstances such as the characteristics of a particular pathogen, accompanying disease, and underlying population. How will the temporal resolution of sensing impact transmission models informed by proximity contact data? Furthermore, how will circumstances alter the impact of temporal resolution? The third part of this dissertation investigates the impacts of sensing regimes on findings from two sampling methods of sensing at widely varying inter-observation intervals by synthetically downsampling proximity contact data from five contact network studies—with each of these five studies measuring participant-participant contact every 5 minutes for durations of four or more weeks. The impact

of downsampling is evaluated through ABM-SEIR simulations from both population- and individual-level for 12 distinct contagious diseases and associated variants of concern. Studies in this part find that for epidemiological models employing proximity contact data, both the observation paradigms and the inter-observation interval configured to collect proximity contact data exert impacts on the simulation results. Moreover, the impact is subject to the population characteristics and pathogen infectiousness reflective (such as the basic reproduction number, R_0). By comparing the performance of two sampling methods of sensing, we found that in most cases, periodically observing for a certain duration can collect proximity contact data that allows agent-based models to produce a reasonable estimation of the attack rate. However, higher-resolution data are preferred for modeling individual infection risk. Findings from this part of the dissertation represent a step towards providing the empirical basis for guidelines to inform data collection that is at once efficient and effective.

This dissertation addresses the problem of informing transmission models with proximity contact data in three steps. Firstly, the demonstration of an EKF-SD-SIR model suggests that the filtering approach could improve System Dynamics transmission models by leveraging proximity contact data. In addition, experiments with the EKF-SD-SIR model also revealed that the filtering approach is constrained by the limited quality and reliability of sensing-data-inferred proximate contacts. The following two parts of this dissertation investigate spatial-temporal factors that could impact the quality and reliability of sensor-collected proximity contact data. In the second step, the impact of spatial resolution is illustrated by differences between two typical sensing modalities—Bluetooth beaconing versus GPS co-location. Experiments show that, in general, proximity contact data collected with Bluetooth beaconing lead to transmission models with results different from those driven by proximity contact data collected with GPS co-location. Awareness of the differences between sensing modalities can aid researchers in incorporating proximity contact data into transmission models. Finally, in the third step, the impact of temporal resolution is elucidated by investigating the differences between results of transmission models led by proximity contact data collected with varying observation frequencies. These differences led by varying observation frequencies are evaluated under circumstances with alternative assumptions regarding sampling method, disease/pathogen type, and the underlying population. Experiments show that the impact of sensing regimes is influenced by the type of diseases/pathogens and underlying population, while sampling once in a while can be a decent choice across all situations. This dissertation demonstrated the value of a filtering approach to enhance transmission models with sensor-collected proximity contact data, as well as explored spatial-temporal factors that will impact the accuracy and reliability of sensor-collected proximity contact data. Furthermore, this dissertation suggested guidance for future sensor-based proximity contact data collection and highlighted needs and opportunities for further research on sensing-inferred proximity contact data for transmission models.

Acknowledgements

My deepest gratitude goes to my supervisors—Dr. Kevin Gordon Stanley and Dr. Nathaniel David Osgood—for their erudite supervision, contributions, and support throughout my doctoral degree program. Dr. Osgood is a transcendent beacon of intelligence and knowledge tirelessly guiding me to realize my mistakes. Dr. Stanley's wisdom and patience encouraged me to tread through shadows and despair. With their helps, I gradually found my way from mistakes and learned to live as a mortal. I was so fortunate to be enlightened by such great mentors.

I am deeply indebted to my committee members. Dr. Juxin Liu provided me with precious opportunities to study maths and statistics and spent numerous hours in her spare time answering my questions. My encounter with the arts of problem-solving started with Dr. Christopher Dutchyn in an afternoon a decade ago, when he lent me a hand with my misconfigured laptop. Dr. Dutchyn then taught me to appreciate the beauty of algorithms and programming languages as just another form of problem-solving. Dr. Derek Eager lent me invaluable insights since my first class in 2010; the theory he taught in my first semester benefited me throughout my doctoral study. Dr. Dwight Makaroff gave me unwavering guidance throughout my studies, and I benefited greatly from the DISCUS reading group he led.

One cannot make (data) bricks without straw (of computing resources)—my experiments cannot be conducted without the support and nurturing of our lovely and formidable wizards of the technical team. It is a great joy to chat with and learn from our technical wizards, pardon me, but they are the most precious computing power and firewalls ever. Thank you—Greg Oster, Merlin Hasen, Cary Bernath, Raof Ajami, and endless emerging magicians.

Special thanks to Christine Hills, the "granny" of the CEPHIL, for renovating the CEPHIL physically and metaphorically. Special thanks to Sophie Findlay for her guardian to all graduate students—her kind reminder prevents us from the late-registration fee or other troubles. Special thanks to all preceding graduate program assistants—Gwen Lancaster, we used to pass by on our ways back/forth to the gym; and Jan Thompson, you helped me land the department of Computer Science to start my decade-long journey.

Thanks should also go to Heather Webb, Maurine Powell, and Linda Gesy—you taught me to use the scanner/printer in the main office and helped me receive parcels when my accommodation cannot.

I cannot begin to express my thanks to Dr. Kurt Kreuger—his positive and goodwill illuminate everyone such that imaginarily the joyful time goes faster around him. I am so grateful to Allen McLean and Wade McDonald—anywhere shines when they are nearby.

I would also like to extend my sincere to Dr. Geoff McDonnell, the "Santa Claus" of the System Dynamics Society, and Mohammad Hashemian, who never wavered in their support. I also wish to thank Jill McMillan for her patience and insightful advice on academic writing.

Many thanks to CEPHIL and DISCUS labmates, collaborators, and everyone who came into my life.

I acknowledge the Natural Sciences and Engineering Research Council of Canada for providing funding.

To all those who helped me throughout the decade

Contents

List of Tables

List of Figures

List of Abbreviations

ABM	Agent-based Model
AFP	Apple Filing Protocol
AP	Access Point
AGPS	Assisted Global Positioning System
API	Application Programming Interface
BER	Bit Error Rate
BLE	Bluetooth Low Energy
BT	Bluetooth
CCDF	Complementary Cumulative Distribution Functions
CRC	Cyclic Redundant Check
DES	Discrete Event Simulation
ECCDF	Empirical Complementary Cumulative Distribution Functions
ECDF	Empirical Cumulative Distribution
EEG	Electroencephalogram
EKF	Extended Kalman Filtering
FSPL	Free-space Path Loss
GDOP	Geometric Dilution of Precision
GRPR	Golden Receive Power Range
HMM	Hidden Markov Chain
IQR	Interquartile Range
KDE	Kernel Density Estimation
LMA	Locationally Mandatory Activities
LQ	Link Quality
LSA	Locationally Stochastic Activities
LoS	Line of Sight
MCMC	Markov Chain Monte Carlo
MERS	Middle East Respiratory Syndrome
PF	Particle Filtering
RF	Radio Frequency
RFID	Radio Frequency Unification
RSSI	Received Signal Strength Indicator
SARS	Severe Acute Respiratory Syndrome
SD	System Dynamics

SEIR	Susceptible-Exposed-Infectious-Removed
SIR	Susceptible-Infectious-Removed
SIS	Sequential Importance Sampling
SMC	Sequential Monte Carlo
SSID	Service Set Identifier
TDoA	Time Difference of Arrival
TF-IDF	Term Frequency–Inverse Document Frequency
TPL	Transmit Power Level
TTFF	Time-to-first Fix
UEE	User Equipment Error
UERE	User Equivalent Range Error
UTM	Universal Transverse Mercator
VZV	Varicella-Zoster Virus
WGS	World Geodetic System
WSN	Wireless Sensor Networks

1 Introduction

Infectious diseases have strained civilization since its dawn. A century after the 1918 influenza pandemic, which infected nearly a third[1] of the world's population [1], we still appear underprepared for the ongoing coronavirus 2019 (COVID-19) pandemic—an epidemic estimated to have infected over 180 million individuals [2] and precipitated a $16 trillion economic loss [3] as of July 11, 2021. Escalating impacts of infectious diseases, exploiting increased human mobility [4–6] and booming social media [7, 8], seep into every corner of our lives, from compulsive hoarding [9, 10] to misinformation spread [11], threatening everyone's physical health [12, 13] and mental health [9–11, 14]. The extended impact of infectious diseases affects wild animals and water reservoirs [15], threatening us through ecological systems [16].

Public health efforts have mitigated and even eliminated [17] infectious diseases by reducing infectious contacts, such as by implementing quarantines [18, 19], and actively increasing the immunized portion of the population, such as via promotion of vaccination [17]. Behind the scenes, mathematical modeling driven by empirical data guide us in evaluating [20–22] and improving [23, 24] the efficiency of both measures. In the ongoing fight against the COVID-19 pandemic, societies have introduced social distancing [25–27], lockdown [28–31], and quarantine and isolation measures [32–34] to buy us time [25, 28, 30, 33] to develop and roll out vaccines [35–37]. Epidemiological models [38–40] with machine learning [41–44] and various data [44] are employed to guide public health measures [45–47] and, at a micro-level, to send individuals protective notifications based on device-supported proximity detection schemes [48]. While lockdown and vaccination continue to protect us against COVID-19, the extended lockdowns have also burdened many [49–51], and vaccine hesitancy and hostility have slowed vaccination campaigns [52–54], exposing us to emerging variants, some of which render the vaccine less effective [55–59].

The availability of high-resolution proximity contact data enables models to evaluate the individual risk of infection [60–63] based on personal and group behavior patterns [61, 63], capturing heterogeneities among the population. This individual-level granularity is likely to offer strong advantages for modeling scenarios when the number of cases in the region of interest is low—such as in the early stage of a potential outbreak—or when contemplating relaxing local lockdowns. Such resolution also appears likely to improve model resolution when there are smaller numbers of susceptibles, such as with pertussis outbreaks after prolonged quiescence, which are increasingly common in North America [21, 64, 65]. Models making use of high-resolution proximity contact data can provide insights that support designs of more effective, lower-burden lockdowns, as well as

[1]The 1918 influenza pandemic infected about 500 million people in four waves from February 1918 to April 1920, and the world population at that time was about 1.8 billion.

optimize overall protection despite residual vaccine hesitancy or hostility [66, 67].

The prevalence of smartphones and personal wearable sensors facilitates scanning for proximity contacts with a high frequency and storing longitudinal historical contact data, providing higher resolution and fidelity than traditional self-reporting-based contact tracing methods [68–70]. However, our understanding of constraints of battery capacity and impacts of sensor measurement biases are limited, and additional studies are required to enable models to utilize high-resolution proximity contacts to better support public health planning and decision-making regarding measures such as precision lockdown [71–73] and robust reopenings following lockdowns [54, 74–76].

1.1 Motivation

Epidemiological modeling [63, 77] and related machine learning methods [78–81] bring advantages to in mitigating the impact of infectious diseases [82, 83]. Such models support many uses, from guiding effective regional interventions [84, 85] to multi-regional collaborations [86, 87], to confining outbreaks [88] and flattening the epidemic curve [89, 90]. Epidemiological models have three primary purposes: estimating parameters, simulating (or predicting) dynamics [91], and exploring "what-if" (or counterfactual) scenarios [92, 93]. Modern epidemiological modeling for simulating dynamics primarily use three approaches: System Dynamics/compartmental modeling (SD), agent-based/microsimulation/network-based modeling (ABM), and discrete-event simulation (DES) [63, 77, 94, 95]. For infectious diseases like COVID-19, both SD and ABM often model the infection status of an individual into states including *susceptible*, *exposed* (representing those in a latent state of infection), and one or more *infectious* states. Depending on the hypotheses and scope of the model, after the *infectious* state, there could be states such as *removed*, *dead*, *recovered*. A *vaccinated* state is also commonly used. Within such models, the transition from *susceptible* to *exposed* is a particularly key one, and is governed by the force of infection, which is generally governed in part by both the transmissibility of the pathogen and the contact rate of the parties involved. When considered over the time an individual spends infectious, these factors also govern a key epidemiological parameter known as the reproductive number[1], that is, the expected number of new infections (transition from *susceptible* to *exposed*) caused by an infectious individual throughout the course of their infectious period. Keeling and Eames [61] noted that the reproductive number provides an approximate summary of the emergent dynamics of contact networks, and advocated studying this connection in order to achieve deeper understanding of epidemiological system and more insightful epidemiological models [60, 62, 96].

The prevalence of smartphones supports the gathering of richer data concerning the dynamics of contact networks, and also offers the prospect of deploying personalized protections for those at risk. For example,

[1]Here we use *reproductive number* to refer to, more specifically, the *basic reproductive number* (which assumes an otherwise susceptible population) or *effective reproductive number* (considering whatever epidemiological situation is currently in place) under corresponding contexts.

modern contact tracing employing smartphone-based sensing has been initiated simultaneously by diverse governments [97–100] and tech giants [48], triggering massive research on the ethical and privacy impact and efficiency of these contact tracing apps [101–107]. However, more empirical studies and theories of smartphone-based contact tracing are required to enhance the efficiency and utility of ongoing data collection. The higher efficiency means derived contact networks for a targeted population from fewer sampled people and with lower sensing frequency; by contrast, while better utility refers to more innovative usages of smartphone-based contact data will be sufficient—not only predicting the attack rates, but also spotting critical paths of spreading, minimize targets of lockdowns if possible, or advising measures to compensate non-compliance with public health orders. Kreuger and Osgood have evaluated the potential of particle-filtered agent-based models [108] while limited by the quality and reliability of the contact data, as well as the representation of contact networks beyond theoretical network models such as small-world and ring lattice networks. Current theoretical network models, despite being able to capture some heterogeneity, still offer poor fidelity in representing the occurrence of empirical proximate contacts. This limited resolution impairs the accuracy of dynamic model outcomes. For example, these theoretical network models tend to systematically overestimate attack rates [109].

Contact data collection and interpretation for the purposes of epidemic modeling and simulation purposes have unique characteristics and require further studying [110, 111]. When compared with other time-series, contact data has more characteristics, because if we encoding occurrence of contacts between a pair of objects (i, j) as dichotomous dummy variable $x_{ij}(t)$, then $x_{ij}(t) \in \{0, 1\}$ is governed by a stochastic process where $W_{ij}(t) = \int_0^t x_{ij}(t)$ is cumulative contact time between i and j during time period $(0, t]$. The challenge lies in the fact that we cannot safely adopt the naïve assumption and treat $\{x_{ij}\}$ $i, j \in V, i \neq j$ as mutually independent, due to simple counterexamples such as schedules of group meetings during which all pairs of group members are expected to be connected, or conflicting contemporaneous activities making contacts between pairs mutually exclusive. Furthermore, failure modes, such as overlooking one contact or recording a contact that did not happen, will have an impact when estimating the spread of disease [63]. Further studies specifically evaluating the dynamics of the contact data collection and interpretation for epidemic modeling and simulation are required, particularly with the high-resolution contact data available through smartphone-based data collecting methods, when compared with traditional less-effective and less efficient manual tracing based on self-reported contacts.

In order to take advantage of high-resolution contact data, improvements in modeling methodology are required [61]. Traditional System Dynamics and compartmental modeling approaches work on aggregate contact matrices. The employment of the standard random mixing assumption for both contacts within a compartment and contacts between a pair of compartments can be appropriate when modeling at larger scales. However, models that employ the random mixing assumption do not incorporate individual-level data—such as contact history or individual-level of heterogeneity of social activity—or provide updatable estimates of individual risk during the early stage of an outbreak or post-outbreak reemergence.

Many machine learning techniques leveraging Bayesian statistics at this stage are ready to predict outbreaks and evaluate interventions. For example, Markov chain Monte Carlo [112], the sequential Monte Carlo method of particle filtering [108, 113–118], and particle Markov Chain Monte Carlo methods [119] developed by Osgood and Liu are well suited for drawing insight from sophisticated combinations of empirical data and dynamics drawn from an epidemiological model.

This thesis argues that combining resolution-aware sensor data collection and models capable of taking into account high-resolution data with a hybrid modeling and machine learning approach can be of help analyze and refine interventions, enabling epidemiological modeling to better support decision making.

1.2 Problem Statement

It is important to study high-resolution contact records because they provide opportunities for existing epidemiological models to simulate and study individual risks of infection and reveal individual-level granularity on how the infections spread over a community. Understanding both individual risks of infections and individual-level granularity on the empirical spread of infection seem likely to be of strong importance for informing precision lockdown, slowing outbreak spread, and safeguarding the process of reopening. Broader availability of proximity contact data (smartphone-based sensing) and increased computing power (machine-learning and supporting hardware) have brought us access to individual proximity contact data. However, traditional group-level compartmental transmission models cannot readily exploit empirical data reflecting the individual-level heterogeneity within a group [120]. Another challenge concerns the impact of sensing resolutions on individual-level estimations when fused into models.

Traditional compartmental transmission models usually reflect group-level heterogeneities on proximity contacts as preferential contact matrices and cannot benefit from individual-level proximity contact records directly. On the other hand, although we can fuse agent-based models with individual-level proximity contact records, their progress is increasingly burdened as the population grows. It is important to investigate the connection between periodically regrounded cohort-level compartmental models and individual-level proximity contact data-driven agent-based models. In the first part of this thesis, I present a solution employing a Kalman filter, which subsequently led other researchers to explore how other filtering techniques, such as the Sequential Monte Carlo method of particle filtering, can be used to enable compartmental models [113] or even agent-based models [108]. The Kalman filtering study revealed that filtering techniques are now limited by the quality and reliability of the sensing data-inferred proximity contact networks. While other researchers in the group investigated other fusion algorithms based on my initial work, I focused on understanding the impact of the input data.

Sensing comes with different types of noise, and improving accuracy will, in general, require increasing costs. Whether we have sensor-collected proximity contacts derived from a co-location approach (such as via GPS-inferred distances) or via a beaconing approach (such as with Bluetooth discovery), errors exist due to

complex real-world factors; for example, signal attenuation and electromagnetic wave interference can result in failures to detect proximate individuals. Anthropogenic effects, such as those related to privacy concerns or leaving wearable devices off-person, can lead to biased proximity contacts, over-representing contacts with certain subsets of individuals, and under-representing others. In the second major component of this work, the research seeks to understand the degree to which GPS co-location and Bluetooth beaconing will lead to different derived proximity contacts. Moreover, this work seeks to understand the impact of spatial resolution—reflected as the Euclidean distance derived from measures of location and via signal strength in beaconing—on estimations made with agent-based models.

The information entropy rate can be used to measure the expected amount of information conveyed by sensor-collected proximity contact data. Because the information entropy rate of sensor-collected proximity contact data grows at the cost of power consumption [121], the study of the outcomes of agent-based models at different temporal resolutions can help assess the impacts of both observation frequency and observation method. Assessment of this sort can inform us of potential opportunities to optimize battery consumption and balance privacy protection when rolling out personal-wearable-device-based proximate-contact sensing. Furthermore, observations undertaken with higher temporal resolution can capture transient contact patterns in the network structures of proximate contacts. These transient contact patterns can provide insights into assessing individual risks of infection in places such as hospitals, incarceration facilities, and care homes.

To summarize, this dissertation focuses on addressing the problem: How can we improve transmission models with sensor-collected proximity contact data? To address this general problem, this dissertation investigates three specific questions:

- Is it possible to use the Kalman filter to advance transmission modeling by incorporating high-resolution proximity contact data?
- Whether and to what degree does increased spatial resolution in proximate-contact sensing impact simulation outcomes?
- Whether and to what degree does increased temporal resolution in proximate-contact sensing impact simulation outcomes?

1.3 Methodology to Address Stated Problems

The last two lines of investigations noted above involve evaluating the impact of high-resolution proximity contact data on outcomes of agent-based models. While the proximate contacts among our modeled population consist of deterministic replays of sensor recordings, the stochastic nature of agent-based modeling governs that each realization is still probabilistic due to the following factors: (1) Poisson arrivals characterizing infection attempts (such as coughing- or sneezing-induced cascades of droplets), (2) Bernoulli distributed outcome of infection attempts between a proximate infectious and susceptible pair, and (3) the probability

distributions associated with the duration of the latent period and (separately) infectious period. To arrive at reliable conclusions in the face of such stochastic processes, we ran multiple realizations for each parameter set and developed metrics to compare the impacts arising from differences between proximity contact data when used with different transmission models associated with different parameter sets, such as those involving different temporal-spatial resolution or differing pathogens.

1.4 Contribution

In three steps, this dissertation explored the problem of improving transmission models with proximate-contact sensing. The first part of this work employed the Kalman filtered System Dynamics model as an example, and demonstrated the potential of applying filtering techniques and proximity contact data to improve transmission models. The second and third part of this work addresses limitations found through this earlier work by focusing on better understanding sensing-data-inferred proximate contacts for transmission models—from both temporal and spatial perspectives.

The work of Chapter 3 is among the first to explore applying filtering techniques to recurrently reground state estimates of epidemiological models in light of dynamically unfolding data. This work evaluated the effectiveness of extended Kalman filtering (EKF) in periodically regrounding an aggregate (System Dynamics) susceptible-infectious-removed (SIR) model. The evaluation involved comparing the results of the EFK-filtered SIR model against synthetic data generated by an individual-level agent-based SIR model parameterized by proximity contact data. This work found that the EKF solution improves outbreak peak estimation and can compensate for inaccuracies induced by model structure, aggregation, and parameter estimates. Findings of the Kalman filtering work encouraged further studies on filtering techniques to fuse individual-level data, System Dynamics, and agent-based models.

The investigation described in Chapter 4 investigates how the accuracy of proximate-contact detection varies by sensing technique. Because the distance threshold within which infection can be transmitted varies by pathogen/disease, experiments were undertaken with both GPS co-location derived- and Bluetooth beaconing derived-proximity contact data for agent-based SEIR models over four common diseases. This work found that, generally, proximate contacts derived through GPS co-location and Bluetooth beaconing will result in different simulation results, particularly from the perspective of individual infection risks. Furthermore, the degree of difference induced by these distinct measurement modalities varies by pathogen/disease and the transmission distance thresholds of proximate contacts of interest. The findings of this work emphasize the need for further disease-specific- and sensing technique-specific- studies to better understand the influence of sensing techniques on the induced outcomes of proximity-contact-data-informed transmission models.

Finally, the investigation characterized in Chapter 5 studied the impacts of temporal resolution on proximity contact sensing and findings of proximity-informed simulation models by varying the sampling method and observation frequency. As a result, this work corroborated that temporal resolution matters for

proximity contact sensing, with a better sampling method and higher observation frequency helping to mitigate overestimation and underestimation. In addition, this work found the practical and widespread `Snapshot` sampling method to be relatively reliable in estimating attack rates, but that the impact of temporal resolution varies by disease, underlying populations, and the combination of sampling method and observation frequency. This work identified the need for improved sampling methods to support the estimation of individual-level risk of infection and reveal the network structure of proximity contacts in the context of lower observation frequencies.

1.5 Dissertation Overview

The thesis is manuscript-styled [122]. There are three papers related to the body of this thesis, included here as (Chapter 3 through Chapter 5). The following is a summary of the chapter contents.

Chapter 1

Introduces the field, frames the problem, and provides an overview of the contributions. Mathematical and algorithmic background and definitions required to understand the dissertation are provided.

Chapter 2

This chapter discusses developments related to the problems characterized in the Problem Statement (Section 1.2). This chapter starts by recalling classical epidemiological models and the importance of contact networks, coverage followed by a characterization of three dynamic simulation modeling approaches and their applications to epidemiological modeling. This characterization is followed by noting the prospects of informed understanding of the dynamics of contact networks due to developments in technologies and lifestyles. The chapter then reviews recent development in mobile sensing techniques and their capability to record high-resolution proximity contacts. This chapter then dives into a discussion of two primary types of sensors—GPS and Bluetooth—revisiting their limitations. Finally, this chapter summarizes the motivation for further studies on sensing proximity contacts and integration into transmission models.

Chapter 3 (Manuscript 1)

Citation: W. Qian, N. D. Osgood, and K. G. Stanley, "Integrating Epidemiological Modeling and Surveillance Data Feeds: a Kalman Filter Based Approach," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer, 2014, pp. 145–152. DOI: `10.1007/978-3-319-05579-4_18`.

The uncertainty regarding the dynamics of mixing behavior is one of the major reasons infectious disease spread is difficult to accurately measure and model. In practice, this induces a need to balance simulation-generated estimates with empirical data. This chapter demonstrated and evaluated an Extended Kalman Filter (EKF) approach to recurrently regrounding simulations when empirical data arrives throughout outbreaks.

This approach simultaneously considers empirical data accuracy, the growing simulation error as time passes between measurements, and supports estimations of changing model parameters. The work of this chapter compared simulations between a "synthetic ground truth" SIR agent-based model (SIR-ABM) fused with high-proximity contact data and an EKF-filtered System Dynamics SIR model (EKF-SD-SIR) recurrently updated with noisy measurements reflected in the synthetic data output by the SIR-ABM. The work of this chapter finds that the EKF-SD-SIR solution improves outbreak peak estimation compared to a SIR aggregate model in the absence of such filtering, and can compensate for inaccuracies and approximations in the structure and parameter estimates of the aggregate model.

Chapter 4 (Manuscript 2)

Citation: W. Qian, A. Cooke, K. G. Stanley, and N. D. Osgood, "Comparing Contact Tracing Through Bluetooth and GPS Surveillance Data," *Submitted to the Journal of Medical Internet Research*, Apr. 2022.

Two primary sensing techniques for proximate-contact detection exhibit distinct accuracy tradeoffs: GPS co-location can have errors of at least ten meters when used outdoors; Bluetooth beaconing usually has errors on the order of meters. Both measurement modalities have shortcomings and are subject to false positives or negatives as unmeasured environmental influences bias the data. Meanwhile, communicable respiratory diseases predominantly infect through one of two mechanisms. Many spread through respiratory droplets during coughing and sneezing, which can, as clinical experiences tell, affect people within 6 feet in line of sight (LOS) when unmasked. Contrastingly, airborne diseases (such as those spreading via aerosol-based mechanisms) can spread much further. The manuscript of this chapter presents a comparison of GPS and Bluetooth-inferred contact patterns and assesses their impact on the attack rate induced in corresponding agent-based Susceptible, Exposed, Infectious, Recovered (SEIR) models of four different communicable diseases. The work of this chapter shows that the contact networks recorded by these two measurement modalities are different and give rise to significant discrepancies in the estimates of attack rates across multiple datasets and pathogens.

Chapter 5 (Manuscript 3)

Citation: W. Qian, K. G. Stanley, and N. D. Osgood, "Impacts of observation frequency on Reconstruction of Close-proximity Contact Networks and Modeled Transmission Dynamics," *Submitted to the PLOS Computational Biology*, May 2022.

Constraints and efficiencies—such as those involving battery capacity and energy consumption—limit observation frequency when detecting and collecting proximity contact data. This chapter sought to investigate how the acceptable measurement frequencies depend on the characteristics of a particular pathogen/accompanying disease, population cohesion, downsampling method (`Snapshot` and `Upperbound`), and analysis involving disease transmission modeling simulation to be performed. The investigation in this chapter downsampled data from five contact network studies, each measuring participant-participant contact every 5 minutes for

durations of four or more weeks. The studies included a total of 284 participants exhibiting different community structures. The work of this chapter found that for epidemiological models employing high-resolution proximity data, both the observation method and the observation interval configured to collect proximity data impact the simulation results. The impact is subject to the population characteristics and pathogen infectiousness (R_0). By comparing the performance of two observation methods, this investigation found that in most cases, periodically observing for a particular duration can collect proximity data that allows agent-based models to produce a decent estimation of the (cumulative incidence) attack rate. However, high-resolution data (such as with sampling intervals shorter than half an hour in the context of simulation experiments covered by this chapter) are preferred to model individual infection risk. The findings of this chapter represent a step toward establishing the empirical basis for guidelines to inform data collection that is simultaneously efficient and effective.

Chapter 6

The contributions are summarized in light of the Problem Statement. The chapter includes a discussion of potential future work and the overall conclusions from this body of work.

1.6 Summary

Studies on high-resolution proximate contacts and their integration into transmission modeling are essential to better support decision-making with regard to epidemiological controls. This work has demonstrated the Kalman filter as one promising approach to fuse high-resolution proximity contacts with epidemiological models. In a subsequent contribution, this work studied the impacts of different proximity contact detection approaches (co-location and beaconing) when fed into simulation models; finally, this work investigated the impact of temporal resolution and downsampling approaches (`Snapshot` and `Upperbound`), under the related factors of pathogen/disease type and population type. This dissertation has demonstrated the plausibility of integrating high-resolution proximity contact into simulation models, discussed factors that could impact the precision and accuracy of these individual-granularity models, and envisioned paths that lead to improvements. This dissertation has also recommended sensor data collection settings for several high-burden communicable diseases and community types.

2 Background and Literature Review

2.1 Epidemiological Modeling and Machine Learning

2.1.1 Epidemiological Modeling

Simulation modeling methods employ mathematical characterizations to represent the operation of a system or a process. Such models can be applied to the study of the behavior of the actual system, evaluating and optimizing the performance of a system, and experimenting with the interventions and their corresponding impact over a longer period in various scenarios [126, 127]. There are three primary simulation paradigms within the sphere of Health and Health Care: Those using Compartmental or System Dynamics (SD) modeling, agent-based modeling (ABM), and discrete event simulation (DES). By combining these three modeling paradigms and taking advantage of each for suitable portions of a model, hybrid models can be created.

Common Epidemiological Compartmental/System Dynamics Models

The susceptible-infectious-removed (SIR) model is a classical epidemic model describing fundamental concepts of infectious disease spread [128–130]. It classifies a population according to the stages of diseases, namely *Susceptible*, *Infectious*, and *Removed*. Using ordinary differential equations, it describes the rate of changes in the count of people in these states ($\mathrm{d}S$, $\mathrm{d}I$, $\mathrm{d}R$) in terms of the current count of people in these states (S, I, R), along with parameters describing the contact networks (c) and disease characteristics (β, τ), as shown in Equation (2.1).

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\bar{c}\frac{I}{N}\beta S \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \bar{c}\frac{I}{N}\beta S - \tau I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \tau I \end{aligned} \qquad \text{(Eq. 2.1)}$$

Despite of being simple, SIR models are widely used. For example: Towers *et al.* [131] used an SIR disease model with periodic transmission rate to access the control strategies via antiviral drug treatment during an outbreak of pandemic influenza; Huppert and Katriel [132] took a general assumption of SIR and equations to model influenza and vaccination; Cooper *et al.* [133] drew on SIR models to study the effectiveness of a

modeling approach for the COVID-19 pandemic, they also developed an SIR model that provides a theoretical framework to investigate the COVID-19 spread within a community.

The SEIR model

The mathematical model of a System Dynamics SEIR model (`SEIR-SD`) is composed of the following ordinary differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= -\bar{c}\frac{I}{N}\beta S \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \bar{c}\frac{I}{N}\beta S - \sigma E \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \sigma E - \gamma I \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I
\end{aligned} \tag{Eq. 2.2}$$

where $N = S + E + I + R$ denotes the total size of the population; $\bar{c}$ is the mean number of contacts made by each susceptible per unit time, β is the probability of transmission per contact between a susceptible and an infective; σ is the rate of exposed persons finishing the latent period and becoming infectious; and γ is the rate at which an infectious person recovers or otherwise transitions to the *Removed* state. S, E, I, and R are the current number of susceptible, exposed, infectious, and removed people.

The `SEIR-SD` model, despite being widely used, contains some strong assumptions, notably the random-mixing of contacts between those infectious and susceptible. Given that each susceptible mixes with $\bar{c}$ people per unit time, there is an assumption that $\bar{c}\frac{I}{N}$ of them will be with infectives; the fact that this quotient considers the infective fraction of the entire population reflects an assumption that any infective within the population will contribute equally to this group, and that a given contact made by a given susceptible will have an equal probability of occurring with any other person within the model population. The probability per unit time that this susceptible is infected by any of their contacts with infectives—the force of infection λ—is then approximated as $\lambda = \bar{c}\frac{I}{N}\beta$. It follows that the count of new infections per unit time is $\bar{c}\frac{I}{N}\beta S = \lambda S$. By recognizing that in an entirely susceptible population, $\frac{S}{N} \approx 1$ The term $\bar{c}\beta$ can also be seen as approximating the count of new infections per unit time transmitted by an index infective in an otherwise susceptible population.

Within this framework, the basic reproductive number R_0, denoting the expected number of secondary cases produced by a typical infective during its entire period of infectiousness in an otherwise susceptible population, is given as $R_0 = \frac{\beta\bar{c}}{\gamma}$.

The $\bar{c}$ denotes the mean contact rate across the population. Extending the mean $\bar{c}$ into individual samples $\hat{c}$, extra resolution can reflect heterogeneity between cohorts with various contact activity level. Consider the mean contact rate between two cohorts i, j is $\widehat{c^{ij}}$, then the basic reproductive number for cohorts i, j is $R_0^{ij} = \frac{\beta\widehat{c^{ij}}}{\gamma}$, denoting the expected number of secondary cases in the cohort i produced by a typical infective from cohort j during its entire period of infectiousness when everyone in cohort i is susceptible [134].

When $i = j$, $\widehat{c^{ij}}$ denoting the mean contact rate within the cohort, when everyone except the infective is susceptible. When the cohorts have a size of one, $\widehat{c^{ij}}$ denotes expected contact rate between two individuals of the susceptible i and infectious j.

Compartment models or Bayesian spatial models can provide further resolution on sub-populations with different $\hat{c}$ and even the spread between geographical regions, but they fail to resolve individual-level contact patterns and the root node (or patient zero) where epidemics originate [135].

Recent Models

While a large number of compartmental/System Dynamics models have been developed to address the ongoing COVID-19 pandemic, we will not cite those papers one by one here. Instead, we will summarize prominent spheres of focus:

Estimating R_0 and R_e One category of research focuses on estimating the basic reproduction number R_0, effective reproduction number R_e, or both. The basic reproduction number R_0 is the expected number of secondary cases produced by a typical infected individual during its entire period of infectiousness, given a situation where the entire population is susceptible at the start of an epidemic, before widespread immunity starts to develop and before any attempt has been made at immunization [136]. If, when surrounded by susceptibles, the index person develops the infection and passes it on to two others prior to recovery, the R_0 is 2. The effective reproductive number, R_e, sometimes also called R_t or R_*, similarly quantifies an expected number of infections generated by an index infective, but under a wider set of epidemiological circumstances. Specifically, it is the expected number of people in a population who would be infected by an index infective before their recovery at the present time (or time t, for R_t). The effective reproductive number changes as the epidemiology evolve—for example, as the count of susceptibles in the population decrease, either by natural immunity (immunity following infection) or by vaccination, and or as individuals are born, immigrate or emigrate, die, or are otherwise removed.

At the beginning of the COVID-19 pandemic, researchers focused on estimating the basic reproduction number R_0 from limited and highly regional dependent infection data; as the pandemic spread, health surveillance efforts and data reporting standards have supported regular reporting of new cases, active cases, hospitalization and mortality estimates for various levels of jurisdictions (such as countries, provinces, regions, municipalities and neighborhoods) over time, which allows estimation of the effective reproduction number R_e [137–140].

Separating Infectious States COVID-19 is known for its relatively long latent period and long infectious period. The variation in symptomology, contagiousness, and propensity to seek care over the course of this long course of infection, has led to researchers delineating several infectious compartments in SEIR models, with potentially multiple mutually exclusive states on each stage. These modules can be characterized as

SEIIR [141] or SEIIIR models[1] [142]. Another motivation for having these sorts of SEIIR and SEIIIR models is to characterize asymptomatic infection pathways or model the impact of quarantine and interventions, by splitting the infectious state into many parallel mutually exclusive sub-states. Such splitting allows modelers to assign different infectious rate β or different contact rate c to those parallel sub-states [143].

Compartment Models and Preferential Contact Matrices In order to capture heterogeneity in contact preferences and parameter differences, compartmental models commonly split each stock of S, E, I, R through the use of subscriptions. Different values of a subscript could represent, for example, different age groups, distinct ethnicity groups, or successive behavior groups (such as with respect to frequency of mask use). The model would then represent preferential mixing probabilities characterized in contact matrices [20, 144–149]. System Dynamics models work better on a large enough scale and with more homogeneous populations such that stochastic variability across flows is not a major issue and that sampling variability does not lead to pronounced deviations from means. However, the System Dynamics modeling approach is limited compared to the agent-based modeling approach for behavioral change models (BCMs), which focus on studying the coupling of individual behavior-changing—such as social distancing and contact precautions—and disease transmission [150–152]. Because a given person's changes in behavior over time commonly depends on the preferences, circumstances, history and context of that particular person, agent-based models much more effectively capture the factors needed to reason about realistic behavior change than do aggregate models.

From System Dynamics Models to Agent-based Models

The mutually exclusive and collectively exhaustive states characterizing the natural history of infection or diagnosis/health status within a compartmental/System Dynamics model can be captured within individual agents using one or more statecharts. Contacts between each pair of individuals within such an individual-based model can be characterized by reading either sensor-collected proximity contact data or model-generated contact schedules. Through this process, we can create a corresponding agent-based model that is able to simulate scenarios when the homogeneous assumption of the SD model no longer holds [95, 153, 154]. Agent-based modeling is particularly flexible in terms of being able to use contact schedules from a generative mobility model, which describes patterns of daily life such as traveling amongst and dwelling at home, working places, and grocery stores, with preferential mixing considered. Agent-based modeling can also include latent factors determining individuals' behaviors, such as adherence with quarantine order or the wearing masks.

Hybrid Models

Hybrid models can be quite useful in combining compartmental/System Dynamics models and agent-based models, because agent-based models are usually computationally expensive and demand more detailed data

[1] To be precise, Arenas *et al.* [142] did not name their model as SEIIIR; instead, we coined their model as SEIIIR due to the fact that they have an additional infectious state representing inpatient infectious people.

[155–158]. For example, when modeling and monitoring the impact of caloric intake and daily workouts on gestational diabetes [158], we used agent-based modeling to represent individual-level heterogeneity in dietary and kinetic habits. Within each agent representation of an individual, this model used an Ordinary Differential Equation/System Dynamics model to summarize the complex dynamics of β-cell mass and glycemia of a diabetic individual. The hybrid modeling approach enables us to easily characterize the broader context of present knowledge from the mathematical model from Hardy *et al.* [159].

2.1.2 Calibrating and Grounding an Epidemic Model

Epidemic models and filtering techniques provide two important means for characterizing and exploiting information extracted from surveillance datasets. Epidemic models specify processes posited to govern the data represented—in a scattered way—in such datasets. Filters sieve noise out of the data and support calibrating parameters of epidemic models to achieve estimates of evolving system states and parameters.

Although epidemic models vary notably across specific diseases, filtering techniques are generic and programmable. There are many applied filters in automation and robotics with the purpose of process control and artificial intelligence. Kalman filters and particle filters are two of those filters. Such filters have been implemented in many signal processing software libraries. However, it is not until recently that these filters have been formally applied in mathematical and computational epidemiology to enhance and calibrate epidemic models [108, 113–117, 134, 160].

Both the Kalman filter and particle filter can be regarded as examples of Bayesian estimation. In Bayesian estimation, unknown variables are treated as stochastic variables, with different Bayesian estimators differing on how they characterize the distribution unknown variables follow. The Kalman filter assumes that all stochastic variables in the discussed system follow Gaussian distributions, an assumption that sacrifices a degree of generality for reduced computational complexity. In accordance with the principle of importance sampling, the particle filter represents distributions with sets of weighted samples, referred to as particles. Both the Kalman filter and particle filter update the prior distribution of an instantaneous system state (treated as a vector of stochastic variables) with the likelihood function of instantaneous system states given acquired observations [161]. A particle filter commonly takes 10 to $1,000$ times the computational time than a Kalman filter, with the specific computational time required by a particle filter scaling with the number of particles it employs [162–164].

Hidden Markov Model

Hidden Markov models (HMMs) and linear dynamical systems (LDSs) are closely related, except that, traditionally HMMs are considered to have discrete state-spaces while LDSs can have continuous state-spaces [165]. The HMM is the foundation of applied estimators, including both the Kalman filter and particle filter [78, 165]. The essence of an HMM lies in its abstraction of the relationship of measurement and system-state over time.

HMMs employ the Markov assumption under which the Markov property holds, with the Markov property (also referred to as the memorylessness property) positing that the probability distribution of future states of the system is conditionally independent of historical states given the current state. Under the Markov assumption, the forward propagation of a discrete-time state-space model, that is, to estimate the marginal $p(X_k \,|\, X_0)$ of state distribution of a state $p(X_{1:k} \,|\, X_0)$ at time t_k given the initial state distribution $p(X_0)$, can be simplified by iteratively computing $p(X_k) = p(X_k \mid X_{k-1})\, p(X_{k-1})$; this iterated system helps us reduce the complexity associated with describing and computing [166].

A discrete-time hidden Markov model usually contains a Markov chain, which describes the system-state dynamics in terms of state transition probabilities. This Markov chain of system-state dynamics is hidden, which means they cannot be directly measured or verified—a given model could differ from the most natural abstraction characterizing the underlying structure governing the system in the world. For example, that model could be some missing transitions or states, or contain needless states. One benefit of the Markov chain system model is that the governing factors can be encoded into a state-transition matrix.

Another part of the hidden Markov model formalism concerns the transitions from system-states to measurement-states. Again, the memoryless characteristics of transitions within a Markov process helps to simplify the computation, because given a measurement Z_k, the k-th of a series of measurements, we could interpret the posterior probability of the hidden state X_k as follows:

$$p(X_k \mid Z_{1:k}) = \frac{p(Z_k \mid X_k)\, p(X_k \mid Z_{1:k-1})}{p(Z_k \mid Z_{1:k-1})} \tag{Eq. 2.3}$$

where

$$p(X_k = x_k \mid Z_{1:k-1} = z_{1:k-1}) = \int p(x_k \mid x_{k-1})\, p(x_{k-1} \mid z_{1:k-1})\, \mathrm{d}x_{k-1} \tag{Eq. 2.4}$$

Note that in Equation (2.4), we used the assumption of Markov process, $X_k \perp\!\!\!\perp Z_{1:k-1} \mid X_{k-1}$, to have $p(X_k \mid X_{k-1}, Z_{1:k-1}) = p(X_k \mid X_{k-1})$.

For a multi-dimensional discrete system with unbiased system noise and unbiased measurements, its state-space model is given as follows [78, 80, 81]:

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{f}_k(\boldsymbol{x}_{k-1}, \boldsymbol{v}_{k-1}) \\ \boldsymbol{z}_k &= \boldsymbol{h}_k(\boldsymbol{x}_k, \boldsymbol{u}_k) \\ \boldsymbol{v}_k &\sim \mathcal{N}(0, Q) \\ \boldsymbol{u}_k &\sim \mathcal{N}(0, R) \end{aligned} \tag{Eq. 2.5}$$

where $\{\boldsymbol{x}_k,\ k = 1, 2, \dots\}$ is the state sequence of the underlying system; $\boldsymbol{f}_k : \mathbb{R}^{n_x} \times \mathbb{R}^{n_v} \to \mathbb{R}^{n_z}$ is a potentially nonlinear function of the state $\boldsymbol{x}_{k-1}$, termed the state transition function for system state $\boldsymbol{x}_k$; $\boldsymbol{z}_k$ is the measurement at time t_k; Q and R are covariance matrices of Gaussian noises associated with the model and measurements. Similarly to the system-state transition, per measurement space, we have a mapping function from the system-space to the measurement-space $\boldsymbol{h}_k : \mathbb{R}^{n_x} \times \mathbb{R}^{n_n} \to \mathbb{R}^{n_z}$. There are also two noise sources that are treated as independent and identically distributed (i.i.d.) stochastic processes: $\boldsymbol{v}_k$ represents the

system noise—the deviation of model estimation from underlying ground truth of the system state. Beyond this, $\boldsymbol{u}_k$ represents the *measurement noise*, *i.e.*, the deviation of measurement from underlying ground truth regarding the measurand.

For systems with additive system noise and measurement noise, Equation (2.5) can be simplified as:

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{f}_k(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_{k-1} \\ \boldsymbol{z}_k &= \boldsymbol{h}_k(\boldsymbol{x}_k) + \boldsymbol{u}_k \end{aligned} \tag{Eq. 2.6}$$

Kalman Filter

According to the assumption of a Kalman filter, the system is characterized by additive i.i.d. noise (per Equation (2.6)), and both $\boldsymbol{v}_k$ and $\boldsymbol{u}_k$ follow a zero-mean normal (Gaussian) distribution.

For a discrete-time linear system, we could rewrite Equation (2.6) as:

$$\begin{aligned} \boldsymbol{x}_k &= F_k\boldsymbol{x}_{k-1} + \boldsymbol{v}_{k-1} \\ \boldsymbol{z}_k &= H_k\boldsymbol{x}_k + \boldsymbol{u}_k \end{aligned} \tag{Eq. 2.7}$$

where F_k is the state-transition matrix of shape $n_x \times n_x$, and H_k is the measurement matrix of shape $n_z \times n_x$.

It is worth noting that, given the state-space of $\boldsymbol{x}_k$, the rank of $H_k^\intercal \boldsymbol{z}_k$ is determined by both H_k and $\boldsymbol{u}_k$. In practice, the rank of H_k is usually less than or equal to the dimension of the state-space of $\boldsymbol{x}_k$—this means that the intrinsic dimension of idealized measurement (without noise) should be less than or equal to the dimension of the state vector.

To estimate the underlying state $\boldsymbol{x}_k^{\text{real}}$, The Kalman filter works when the rank of $H_k^\intercal \boldsymbol{z}_k$ is greater than or equal to the dimension of the state-space of $\boldsymbol{x}_k$. This is usually achieved by having rank of $H_k^\intercal \boldsymbol{u}_k$ greater than or equal to the rank of $H_k^\intercal H_k \boldsymbol{x}_k$, and the rank of $H_k^\intercal \boldsymbol{u}_k$ is the number of linearly independent sources of measurement noise in vector $\boldsymbol{u}_k$. To be more specific, the dimension of vector space of $\boldsymbol{z}_k$ is the number of independent measurements taken on the system-state, and the core of Kalman filter is to fuse independent noisy measurements to get a more accurate estimation of the underlying system state.

When dealing with a continuous-time non-linear system, we have Equation (2.5) further expressed as:

$$\begin{aligned} \boldsymbol{x}(t_k) &= \boldsymbol{x}(t_{k-1}) + \int_{t_{k-1}}^{t_k} \boldsymbol{f}\left(\boldsymbol{x}(\tau), \tau\right) \mathrm{d}\tau + \int_{t_{k-1}}^{t_k} \boldsymbol{v}(\tau)\mathrm{d}\tau \\ \boldsymbol{z}(t_k) &= \boldsymbol{h}_{t_k}\left(\boldsymbol{x}(t_k), \boldsymbol{u}_{t_k}\right) \end{aligned} \tag{Eq. 2.8}$$

Note that usually we will still use the discrete form for measurement rather than the continuous form. The term $\boldsymbol{v}(\tau)\mathrm{d}\tau$ can not be integrated directly unless we use Itô calculus [78]. Luckily, from the assumption of Gaussian noise, $\mathbb{E}[\boldsymbol{v}(\tau)] = 0$, removing $\int_{t_{k-1}}^{t_k} \boldsymbol{v}(\tau)\mathrm{d}\tau$ in the state-update from $\boldsymbol{x}(t_{k-1})$ to $\boldsymbol{x}(t_k)$.

The equations for the extended Kalman filter will use Taylor expansion up to the first order (shown in one-dimension, when $(x - \hat{x}) \to 0$) [78],

$$f(x) = f(\hat{x}) + \left.\frac{\partial f}{\partial x}\right|_{x=\hat{x}} (x - \hat{x}) + O(x - \hat{x})^2 \tag{Eq. 2.9}$$

following is the equation set for Extended Kalman Filter

$$\begin{aligned}
\dot{\hat{\boldsymbol{x}}}(t) &= \boldsymbol{f}\left(\hat{\boldsymbol{x}}(t), t\right) + K(t)\left[\boldsymbol{z}(t) - \boldsymbol{h}\left(\hat{\boldsymbol{x}}(t), t\right)\right] \\
\dot{P}(t) &= F\left(\hat{\boldsymbol{x}}(t), t\right) P(t) + P(t) F^{\intercal}\left(\hat{\boldsymbol{x}}(t), t\right) + Q(t) \\
&\quad - P(t) H^{\intercal}\left(\hat{\boldsymbol{x}}(t), t\right) R^{-1}(t) H\left(\hat{\boldsymbol{x}}(t), t\right) P(t) \\
K(t) &= P(t) H^{\intercal}\left(\hat{\boldsymbol{x}}(t), t\right) R^{-1}(t)
\end{aligned} \tag{Eq. 2.10}$$

where

$$\begin{aligned}
F\left(\hat{\boldsymbol{x}}(t), t\right) &= \left.\frac{\partial \boldsymbol{f}(\boldsymbol{x}(t), t)}{\partial \boldsymbol{x}(t)}\right|_{x=\hat{x}} \\
H\left(\hat{\boldsymbol{x}}(t), t\right) &= \left.\frac{\partial \boldsymbol{h}(\boldsymbol{x}(t), t)}{\partial \boldsymbol{x}(t)}\right|_{x=\hat{x}}
\end{aligned} \tag{Eq. 2.11}$$

One issue that arises when applying EKF concerns the case in which $R^{-1}(t)$ does not exist, which can be caused by $R(t)$ not being of full-rank, such as by dependent measurements in our measurement set. Under this condition, measurements should be removed to eliminate the dependency.

Importance Sampling

Both importance sampling technique and Markov chain Monte Carlo method use a proposal distribution $q(x)$ to approximate the target distribution $p(x)$. Particle filtering relies on the importance sampling technique and variants thereof to track estimates of system states with even non-Gaussian noise from the system model and measurements. Importance sampling uses weighted particles to approximate the target distribution $p(x)$, and the importance weight w_i of the i-th particle x_i, $i = 1, \cdots, n$, is defined as

$$w_i = \frac{p(x_i)}{q(x_i)} \,. \tag{Eq. 2.12}$$

The expected value of $f(X), X \sim p(x)$ can be estimated by

$$\begin{aligned}
\int f(x)p(x)dx &= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\
&= \mathbb{E}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right] \\
&\simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{p(x_i)}{q(x_i)} \\
&= \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i} \,.
\end{aligned} \tag{Eq. 2.13}$$

Particle Filter

Unlike the Kalman filter, which uses the mean value and covariance matrix of a Gaussian distribution to characterize the likelihood of the underlying system-state at an instance, a particle filter applied to a set of state equations characterizing stochastic evolution of system state represents the required posterior distribution of system-state values at a given time given a set of observations until that time by a set of importance-weighted particles. Estimates are calculated based on sequential importance sampling with such particles.

Suppose the sources of noise $\boldsymbol{v}(t)$ and $\boldsymbol{u}(t)$ in Equation (2.6) follow two distributions, with general form $\Psi_v(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$ and $\Psi_u(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$. Suppose further that at time t_k we have measurement $\boldsymbol{z}_k$ and n particles $\boldsymbol{x}_k^{(i)}, i = 1, 2, \cdots, n$, collectively to represent the distribution of system-state $\boldsymbol{x}_k$. We could rewrite Equation (2.5) as

$$\begin{aligned} \boldsymbol{x}_k &\sim \Psi_v(\boldsymbol{f}_k(\boldsymbol{x}_{k-1}) + \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v) \\ \boldsymbol{z}_k &\sim \Psi_u(\boldsymbol{h}_k(\boldsymbol{x}_k) + \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u) \end{aligned} \tag{Eq. 2.14}$$

In terms of notation, $\boldsymbol{f}_k$ will continue to represent the system-state transition function and $\boldsymbol{h}_k$ the measurement function. If we use $w_{t_k}^i$ to represent the weight of ith particle at time t_k, we will have weight-updating equation as following,

$$\begin{aligned} w_k^i &\propto \frac{p\left(\boldsymbol{x}_{0:k}^i | \boldsymbol{z}_{1:k}\right)}{q\left(\boldsymbol{x}_{0:k}^i | \boldsymbol{z}_{1:k}\right)} \\ w_k^i &\propto w_{k-1}^i \frac{p\left(\boldsymbol{z}_k | \boldsymbol{x}_k^i\right) p\left(\boldsymbol{x}_k^i | \boldsymbol{x}_{k-1}^i\right)}{q\left(\boldsymbol{x}_k^i | \boldsymbol{x}_{k-1}^i, \boldsymbol{z}_k\right)} \end{aligned} \tag{Eq. 2.15}$$

where $q(x)$ is the proposal distribution, and $p(x)$ is the target distribution. This iterated approach of updating importance weights of particles is also referred to as sequential importance sampling (SIS).

Markov Chain Monte Carlo

There are three important methods that are frequently used to generate samples following an arbitrary distribution: Metropolis, Metropolis-Hasting, and Gibbs. Within these three, the Metropolis-Hasting method could be regarded as the general case for both Metropolis and Gibbs methods [167].

Successive samples we generate as $\theta^{(t)}$, a genetically inspired algorithm for Metropolis and Metropolis-Hasting methods to sample from the target distribution $p(x)$, could be summarized as follows:

1. Start with any initial value $\theta^{(0)}$ satisfying $p(\theta^{(0)}) > 0$
2. Use proposal distribution to generate a candidate sample $\theta^\star$
3. For step t, Compute α for $\theta^\star$ (varies per algorithm; see below)
4. Compare α with a sample $\epsilon \sim \mathcal{U}[0, 1)$, if $\epsilon < \alpha$, then accept $\theta^\star$ as $\theta^{(t)}$, otherwise still use $\theta^{(t-1)}$ as $\theta^{(t)}$. It bears noting that when $\alpha = 1$, $\theta^\star$ is always accepted as $\theta^{(t)}$.

For the Metropolis sampling,

$$\alpha = \min\left(\frac{p(\theta^\star)}{p(\theta^{(t-1)})}, 1\right) \tag{Eq. 2.16}$$

where the target distribution $p(x)$ has to be symmetric.

For the Metropolis-Hasting method,

$$\alpha = \min\left(\frac{p(\theta^\star) q(\theta^\star \mid \theta^{(t-1)})}{p(\theta^{(t-1)}) q(\theta^{(t-1)} \mid \theta^\star)}, 1\right) \tag{Eq. 2.17}$$

When $q(x)$ is symmetric, then $q(\theta^\star \mid \theta^{(t-1)}) = q(\theta^{(t-1)} \mid \theta^\star)$, the equation reduces to the form used in the Metropolis method.

In contrast to Metropolis sampling and Metropolis-Hasting sampling, which have to keep throwing out rejected proposals $\theta^\star$, an advantage of Gibbs sampling is that it will make use of all the generated proposals $\theta^\star$ after the initial burn-in period. One assumption of a Gibbs sampler is that the target distribution should be multivariate and the conditional distributions of one variable conditioned on all of the other variables can be computed and sampled exactly from these distributions.

Assume we want to generate a joint sample $\boldsymbol{\theta}^{(t)} = \begin{bmatrix} \theta_1^{(t)} & \theta_2^{(t)} & \cdots & \theta_{n_\phi}^{(t)} \end{bmatrix}$ for a vector of n_ϕ variables $\boldsymbol{\phi} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{n_\phi} \end{bmatrix}$ of the target distribution p. Assume further that, for every ϕ_i, we can compute conditional distributions $p(\phi_i \mid \phi_{-i}) = p(\phi_i \mid \phi_1, \cdots, \phi_{i-1}, \phi_{i+1}, \cdots, \phi_{n_\phi})$ and we can also sample $\theta_i^{(t)} \sim p(\phi_i \mid \phi_{-i})$. Based on the joint sample $\boldsymbol{\theta}^{(t-1)}$ from step $t-1$, we generate $\boldsymbol{\theta}^{(t)}$ by

1. Sample $\theta_1^{(t)} \sim p(\phi_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_{n_\phi}^{(t-1)})$
2. Sample $\theta_2^{(t)} \sim p(\phi_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_{n_\phi}^{(t-1)})$
3. Sample $\theta_3^{(t)} \sim p(\phi_3 \mid \theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)}, \cdots, \theta_{n_\phi}^{(t-1)})$

...

$n_\phi - 1$. Sample $\theta_{n_\phi-1}^{(t)} \sim p(\phi_{n_\phi-1} \mid \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{n_\phi-2}^{(t)}, \theta_{n_\phi}^{(t-1)})$

n_ϕ. Sample $\theta_{n_\phi}^{(t)} \sim p(\phi_{n_\phi} \mid \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{n_\phi-1}^{(t)})$

Putting together the n_ϕ samples generated above, we get the joint sample $\boldsymbol{\theta}^{(t)} = \begin{bmatrix} \theta_1^{(t)} & \theta_2^{(t)} & \cdots & \theta_{n_\phi}^{(t)} \end{bmatrix}$ for step t.

Applications in Epidemiological Modeling

Osgood and Liu [113, 115, 117] have led particle filter and particle MCMC applications for latent state and parameter estimation for re-grounding epidemiological models. This use of the particle filter strongly elevated the predictive accuracy of epidemiological models, relative to their open-loop counterparts [113]. The University of Saskatchewan Computational Epidemiology and Public Health Informatics lab (CEPHIL), founded and led by Professor Osgood, who has contributed a series of applications: Kreuger *et al.* [108] has pioneered applying particle filter on agent-based models; Li *et al.* [116] has applied particle filtering in both aggregated and age-structured population compartmental models of pre-vaccination measles; Safarishahrbijari *et al.* [117] investigated three applications of particle filtering involving a SEIRV simulation model of H1N1 influenza. This work included study of the impact of particle filtering on predictive accuracy of particle filtering in dynamic models to support outbreak projections, investigation of how predictive accuracy is affected by the inter-observation interval and key parameters used in particle filtering, as well as how the use of search data affected particle filtering predictive accuracy [115]. Oraji, Safarishahrbijari *et al.* [114] combined particle filtering and transmission modeling for Tuberculosis control; Mohammadbagheri *et al.* [118] has applied mathematical modeling of the Hypothalamus-Pituitary-Adrenal gland (HPA) axis using particle filter algorithm.

2.2 Human Mobility and Contact Networks

Human mobility and contact networks are two behavior-related matters of great importance to epidemiological modeling. Human mobility models focus on describing the characteristics of both individual and collective movements, from larger scale inter-continent flights to small scale daily trips between home, working spaces, and grocery stores [168–170]. Human mobility patterns at the scale of a country or a city may be reflected in real-time travel data and statistics regarding cumulative population migration over a corresponding time frame. The co-location is an even between two persons when their physical distance is within the threshold of interests. Deduced co-location from human mobility patterns can suggest contact rates and exposure risks for epidemiological modeling [168, 169, 171] and aid in the evaluation of interventions involving mobility restrictions such as lockdowns [172, 173].

By contrast, dynamic measurement of contact networks traces contacts between people, providing information that can be directly used to feed an epidemiological model [61]. Access to information on contact networks and their representation in a model can aid evaluation of interventions that cannot be straightforwardly expressed as alteration of mobility patterns, for example, public health orders requiring social distancing [149] or personal hygiene measures such as mask use [63].

Intuitively, there is a natural role for the use of mobility patterns when modeling larger scale disease spreads between countries and cities, while information on contact networks offers ready exploitation at smaller scales of disease spread, such as within a city, a community, or an aged care home [61, 63, 168, 169]. Multiscale epidemiological models using both mobility patterns and contact networks are expected to better express the multiscale nature of dynamic behaviors driving contact within the system [169, 174].

Studies collecting reliable records or formulating accurate models and predictions of human mobility patterns can benefit intelligent routing decisions, whether based on infrastructures such as cell tower or WiFi router placement, or opportunistic peer-to-peer routing in delay tolerant networks. There has been significant research on employing user movement patterns to support transit opportunity estimation [175–178], device resource management [179, 180], and battery savings [181, 182]. Similarly, information on human mobility patterns of the sort which informs the placement of cell towers and WiFi hotspots can also support decision making on the placement of other services or amenities, such as bus stops [183, 184].

Both human mobility and contact networks are shifting with our ever-changing lifestyles, as well as emergency events. Shared (micro-)mobility services such as bike-share and scooter-share changed both trip duration and spatiotemporal signatures [185]. Online shopping and food delivery is changing human mobility patterns not only of their users and couriers [186], but also of others due to delivery service-induced changes in urban traffic patterns and routing [187, 188]. Such changes can be particularly marked during public emergencies such as COVID-19, when, due to social distancing policy, adoption or personal protective behaviors, or occurrence of lockdown, residences rely more on delivery services while staying at home [189]. The COVID-19 pandemic demonstrated a capacity of the early stage of an outbreak to alter such residents'

regular grocery purchasing schedules due to hoarding of goods such as toilet paper; such changes can eventually alter the entire supply chain [190]. Emerging services such as drone delivery may further change human mobility patterns and contact networks [191]. The cost of energy, an aging population, and "996" working systems for young people can further accelerate the demand for delivery services and shared mobility services [186, 191].

Contact patterns are patterns of interaction or mixing in close proximity amongst population members over a time frame. Although greater numbers of such interactions are occurring in the virtual sphere because of convenience with emerging social networks, this thesis focuses on contacts of proximate individuals directly impacting disease transmission models; the balance of the dissertation will commonly refer to such contacts as "proximate contacts" and data records of these contacts as "proximity contact data". Common aggregate metrics used in characterizing contact patterns include the distribution of contact duration, distribution of inter-contact duration [192, 193], and the average contact frequency in a underlying population [61]. These metrics impact parameters of critical importance in health modeling, such as the basic reproductive number. The basic reproductive number, as introduced in Section 2.1.1, is the expected number of others to whom an index infective individual transmits infection over the course of their infectious period, in an otherwise susceptible population. Additionally, contact patterns play a central role in agent-based infection transmission models, infection prevention, and shape individual decision-making [194, 195]. Human mobility patterns underlie the contact patterns between both people and places, and are suitable as metrics at the population level [171, 196]. Human mobility patterns have impacts on the transmission of contagious diseases, attitudes and norms, access to services, and contribute to exposure to environmental risks such as toxins, pedestrian-unfriendly built environments, and food environments, and form an important causal influence on both environmentally mediated diseases such as asthma and socially mediated diseases such as obesity [60, 62, 197–199].

Barbosa *et al*. systematically reviewed human mobility models and their applications [168] from four perspectives: data sources, metrics, models, and applications. Inspired by their perspective, we will review human mobility and contact network by their data sources, metrics, and models. Because human mobility patterns and contact networks are quite related in such a way that many data sources and metrics apply to both, we will only separate these two topics when discussing their models.

2.2.1 Data Source and Important Patterns Found

Unsurprisingly, findings on human mobility patterns and contact networks accompany new data sources enabled by new technologies with higher resolution and better fidelity. The following data sources and associated important findings are presented in chronological order.

Census, Surveys and Circulatable Notes

Census data started to trace migration, living place and later workplace and transportation in 1841 led by the British [200]. With census data, the "Laws of Migration" were formulated in 1885 [201], positing that "migrants only move a comparatively short distance from the place which gave them birth"; a "Law of Limited Circulation of Population" was later articulated in 1937 [202], adding that migration may exhibit "occasional radiating inequalities which indicate favored routes of migration". The improvement of census questions may have facilitated the revealing of these changes in migration patterns. However, they more plausibly could be a reflection of human mobility changing from the 1880s to 1930s—especially when considering that the American industrial revolution took place from the 1880s to 1920s, when the workforce shifted in a pronounced manner from agriculture to industry, and the number of workers increased rapidly [203].

With the availability of an online dollar bill tracking system, Brockmann *et al.* [204] in 2006 studied dollar bill trajectories as secondary data to reveal patterns of underlying human mobility. Their research brought human mobility research into a new phase by attempting to use physical models such as random walks and Lévy walks [205, 206] to explain human mobility. Within their paper, an investigation regarding the distribution of traveling path lengths at a relatively large scale led to the discovery of truncated power-law distributed trip length as one of the potential characteristics of human mobility. The physical limits of human beings are the primary driver for the truncated tails in the power-law distribution and technology [207]. The researchers also derived from lattice network assumptions a model that can reproduce patterns in human mobility with truncated power-law distributed trip length. However, trip length only focuses on memoryless or quasi-memoryless transitions, without considering the patterns of absolute location created by our days or weeks [168]. The memoryless processes have only recorded relative location aspects, which could not be used to represent the characteristic clustering of absolute locations.

Cellular Tower, GPS and WiFi Records

Jensen *et al.* [208] used both GPS and WiFi supported location information as feed-in data, using the same criteria as the Song and Barabási adopted [209]; their results support Song and Barabási's finding of a high entropy upper-bound [209]. Besides using the Lempel-Ziv estimator for entropy, they also applied a first-order Markov model with the transition probability estimated from the finite process and obtained the same result as did the Lempel-Ziv estimator.

Hashemian *et al.* [210] developed an app for the Google Android operating system-based smartphones. Multi-sensor data provides a potential opportunity for data fusion that could mitigate noise caused by a single sensor or allow for better reliability in the presence of missing data. However, an analysis of collected data in terms of human mobility patterns was not provided.

Qian [211] completed the post-processing of data collected for [210], obtaining multi-sensor data allowing the filtering of noisy data, and—in addition to previously used techniques [208, 212]—investigated the impact of and interaction between granularity and geometric representation.

RFID and Bluetooth Discovery Records

Integrating human traces from six studies, with data collected from smartphone-based GPS, WiFi, and Bluetooth sensors, Karagiannis *et al*. [212] found that the distribution of inter-contact time possesses an invariant property: a characteristic elapsed time threshold—on the order of half a day—beyond which the distribution decays exponentially. Furthermore, Karagiannis *et al*. proposed the random way point method of generating synthetic human mobility records involves randomly generating points to represent sites for possible visits, and choosing a destination from these points as a component of the subsequent trajectory. The random way point method can generate human trajectories while allowing truncated exponential decay of inter-contact time pairs in the simulated community.

Cattuto *et al*. [213] used active Radio Frequency Identification (RFID) devices to detect and collect person-to-person proximity contacts, and found super-linear behavior between the number of connections and cumulative contact duration, indicated "the possibility of defining super-connectors both in the number and intensity of connections". Hashemian *et al*. [109] found that using a putative "typical day" to represent contact networks of coworkers tended to overestimate incidence, and argued that in some circumstances, high-resolution data of contact dynamics are required to secure high fidelity in transmission models.

Auxiliary Data Sources and Their Impacts

Riding the tide of cloud computing and machine learning, emerging data sources such as social media, open data, and crowdsourced data, allow new perspectives on human mobility and contact networks. These types of data differ from previous datasets in several ways, including the fact that their collection demands fewer labor hours per record, are less likely to be affected by certain subjective biases, with the data capturing information on various aspects of context [214–216]. Such auxiliary data sources can provide extra insights, as shown in the following studies, arranged according to their auxiliary data source types.

Social Media To help describe traffic anomalies, Pan *et al*. [217] used term frequency-inverse document frequency (TF-IDF) [218] to extract keywords from posts on Weibo (a Twitter-like social media service in China). They demonstrated that when using TF-IDF to extract descriptive keywords of a traffic anomaly, it is more efficient to focus only on posts both spatially and temporally related to a detected traffic anomaly. Their finding reveals a potential to reduce the demand for posts used for information extraction from millions down to hundreds.

Grabowicz *et al*. [219] found entanglement between social ties and human mobility and interactions by building and calibrating their stylistic "travel and friendship" model based on location check-ins on Twitter. The researchers identified user interaction data on Twitter with respect to following, replying, sharing, and check-in locations for over 714,000 users during a month. Additional location check-in datasets are employed from two location-based services, namely Gowalla and Brightkite. Their "travel and friendship" model, simulating traveling to friends and making new friends, was parameterized with two important stylistic

factors. For the *travel* component, an important parameter was the probability of visiting a randomly selected friend at their current location. The *friend* component of the model included the probability of randomly connecting to an individual anywhere, in analogy to making a new friend on social media independent of the geography. They found their calibrated model fits well with the social media collected check-in locations (as human mobility data) and interactive activities on the social media (as the social network structure) in terms of the following metrics:

- Node degree distribution, where an edge represents a directed social link, such as message following in social media. Two individuals with at least one social link in either direction is considered a linked pair.

- At a given distance, the ratio of linked pairs among all distance pairs, where a distance pair is two individuals with physical distance at the given threshold. Because individuals' locations are rounded into grids of points, pairs of two individuals can be grouped by the distance between them.

- For a linked pair, that is, two individuals with at least one link in either direction, the probability of these two individuals mutually linking to each other. Two individuals are mutually linking to each other form a mutually-linked pair.

- Among distance pairs for a given distance, the ratio of mutually-linked pairs over linked pairs.

- The ratio of closed triads over all triads for a given distance, where a triad is defined as any three individuals (i, j, k) such that (i, j) and (j, k) are distance pairs for the given distance, and this triad is closed if (i, k) are also a distance pair for the given distance.

- Distribution of distance disparity among a closed triad, where the distance disparity is defined as

$$D = 6\left[\frac{d_1^2 + d_2^2 + d_3^2}{(d_1 + d_2 + d_3)^2} - \frac{1}{3}\right], \qquad \text{(Eq. 2.18)}$$

  where d_1, d_2, and d_3 are geographical distances between (i, j), (j, k), and (i, k) of a triad (i, j, k). The distance disparity ranges from 0 to 1 as the triangle passes from equilateral to isosceles.

Wu *et al*. [220] modeled intra-urban human mobility based on 15 million check-in records collected during a yearlong. They analyzed these records by dividing travel demands into locationally mandatory activities (LMA) and locationally stochastic activities (LSA). They further classified check-in locations into six categories: home, transportation, work, dining, entertainment and other, and removed random walks (as judged by when the displacement between two adjacent records is less than 100 meters), remaining in the vicinity of a single location or missing (the time interval between two adjacent records is over 12 hours), abnormal data (as judged by an average speed over 431 kilometers per hour). Their classification of travel demands summarized check-in records into a temporal transition probability matrix of activities to simulate agent-based intra-urban mobility.

Wu[1] *et al.* [221] evaluated prediction of keywords associated with a mobility record given historical spatiotemporal documents using kernel density estimation (KDE). Based on 37 million tweets across three cities (New York, Chicago, and Los Angeles) spanning over nine months, their evaluation shows that the prediction of keywords associated with mobility records can be carried out with relatively high accuracy (over 90%) for over one-third of the cases. Their finding confirmed the existence of locationally mandatory activities or locationally landmark tweet keywords as the other Wu *et al.* [220] highlighted.

Zhang *et al.* [222] found that group-level mobility records can help improve the accuracy of predicting individual mobility from a predefined set of probable locations, and that better user grouping with higher within-group consistency boosts model reliability. Zhang *et al.* drew on over 1.3 million Twitter messages during twenty days across two cities. Their model takes their check-in locations and their text to explain the activities users undertook while at that location. Text in natural language often exhibits high sparsity, which can be reflected in the use of many different words to present the same underlying activity. For example, the use of the keywords "pasta" and "pizza" from two tweets checked-in near a shopping mall during lunchtime may indicate the same underlying activity of "having lunch". Zhang *et al.* employed sampling-based text augmentation [223, 224] to handle sparsity of text, such that different expressions indicating similar underlying activities from different users across the spatiotemporal space can be grouped to infer the emission distribution of tweet-text given an underlying activity. Considering each tweet as a multi-dimensional observation, and following emission distributions conditional on the hidden-state of the individual who published the very tweets, Zhang *et al.* used a hidden Markov model (HMM), parameterized by the number of hidden states. They evaluated the trained HMM to predict the top k places at which each individual is likely located, and found that the performance of their HMM is not sensitive to the number of latent states K, as long as $K \geq 5$. Their finding supports the hypothesis that the number of "hotspots" that a group could collectively exhibit follows a power-law distribution [207, 225].

The studies characterized above benefit significantly from social media data. Without social media data, it would be difficult to conduct studies with millions of participants, not to mention linking their information to the activities that they undertake and locations that they visit for months. In addition, information gleaned from social media data can enable researchers to investigate the semantic meaning of human mobility patterns to improve human mobility models and behavior analyses. However, the social media data used by these studies are still likely to suffer from sampling bias, with two notable reasons being the fact that social media data under-represents those who do not use social media activity and cannot rule out self-reporting bias [226].

Open Data Open Data is an evolving idea currently manifested in notable initiatives of Open Government Data [227] and open science data [228]. The open data movements can be traced back to Ancient Greece [227] and substantially influenced by the Open Source and Free Software movements amid the 1990s [228]. Then the principles enabling linked open data and [229] and open government data were conceived in 2006–2007.

[1]This Wu [221] is not the same researcher as the Wu [220] above.

Until 2014, the Open Definition defined data as open if "anyone can freely access, use, modify, and share for any purpose—subject, at most, to requirements that preserve provenance and openness [230]." During the COVID-19 pandemic, Lassig *et al.* proposed the Open Data progression model for opening data for global health [231].

Open Data sectors and communities cover various aspects of our lives, such as statistics regarding education, geospatial planning, land ownership, population-level health status, reporting from telecommunication operators, and transportation statistics. Many sectors of open data sharing rely on national or regional statistics conducted by the government [232]. Meanwhile, trends of, for example, transportation methods, telecommunication, and internet user hobbies rely on third-party organizations, tech giants, and even small businesses to contribute and self-regulate. Such forms of open data can provide background context that helps understanding and forecasting shifts of human mobility and behavior patterns from city construction to online shopping and delivery [233].

Crowdsourced Data The idea of crowdsourcing was proposed in 2006, contemporary with linked open data and open government data [234]. One general pattern that distinguishes crowdsourced data and open data lies in the fact that crowdsourcing initiatives commonly target more specific topics, such as are seen for the CRAWDAD community for archiving wireless data[1] and OpenStreetMap.[2] Moreover, crowdsourcing of data can be more responsive towards emergencies, providing insight before organizations or governments kick in. For example, during the COVID-19 pandemic, software developers and communities self-motivated to crowdsource data to help mitigate epidemics before large-scale government-backed initiatives were undertaken. Crowdsourcing data helped capture and communicate statistics of infections and mortality from the disease in its early stages and aided government tracking and tracing. However, due to non-systematic collecting methods, sampling bias, and the lack of authoritative validation [235–237], conspiracy theories, misinformation, disinformation, and fake messages have corrupted the flow of information regarding the pandemic. Studies on how to effectively use crowdsourced data to help respond to epidemics are necessary [238, 239].

2.2.2 Metrics and Laws

This section summarizes commonly used metrics for analyzing human mobility and contact networks. Metrics are often employed together to effectively demonstrate patterns within data observing specific laws.

Radius of Gyration

The radius of gyration measures the degree of dispersion among a set of relative locations. The distribution of personal radii of gyration for a population reflects the rationale that humans at certain scales "tend to

[1] `https://crawdad.org/`

[2] `https://www.openstreetmap.org/`

move a characteristic distance from their starting locations". Related discovery can be traced back to the "Law of Migration" in the 1880s using census data [201].

Formally the radius of gyration, r_{g}, is defined as [168]:

$$r_{\mathrm{g}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{r}_i - \mathbf{r}_0\right)^2}, \tag{Eq. 2.19}$$

where position coordinates vector $\{\mathbf{r}_i\},\ i = 1, \ldots, N$ consists of N observations of location at different time; $\mathbf{r}_0$ denotes the center of mass of the set of points $\mathbf{r}_0 = \frac{1}{N}\sum_i \mathbf{r}_i$; and $\{\mathbf{r}_i - \mathbf{r}_0\},\ i = 1, \ldots, N$ are effectively N displacements (or relative locations) of a person relative to that center of mass.

Gonzalez *et al*. [170] found that for individual location data collected by smartphones with configurations of both "high temporal resolution shorter period (every two hours over one week)" and "low resolution longer period (every call or text data over six months)," the distribution of r_{g} among a population can be approximated with a truncated power-law

$$P\left(r_{\mathrm{g}}\right) = \left(r_{\mathrm{g}} + r_{\mathrm{g}}^0\right)^{-\beta_r} \exp\left(-\frac{r_{\mathrm{g}}}{\kappa_r}\right), \tag{Eq. 2.20}$$

where $\beta_r = 1.65 \pm 0.15$ reflects a significant degree of heterogeneity of the travel habits of the observed population (the larger the β_r, the less likely individuals will to travel further from their initial location), and $\kappa_r \approx 350$km represents an upper cutoff mostly due to the finite size of the study area or even the limit of human mobility region [168]. The fact that Equation (2.20) exhibits both a power-law body and fat-tail characteristics reflects the fact that when $r_{\mathrm{g}} \ll \kappa_r,\ \exp\left(-\frac{r_{\mathrm{g}}}{\kappa_r}\right) \to 1,\ P\left(r_{\mathrm{g}}\right) \approx \left(r_{\mathrm{g}} + r_{\mathrm{g}}^0\right)^{-\beta_r}$, and clearly $\lim_{r_{\mathrm{g}}\to\infty} \exp\left(Cr_{\mathrm{g}}\right) P\left(r_{\mathrm{g}}\right) = D\left(r_{\mathrm{g}} + r_{\mathrm{g}}^0\right)^{-\beta_r} = \infty$, indicating that this distribution also exhibits a fat-tail. Interestingly as the author noted, at extreme data collection configurations:

- When the observation frequency is low—for example, annually—the smartphone data with respect to the location time series exhibits similarities to a census data with respect to family location, where the law on radius of gyration becomes an expression of the "Laws of Migration" [201] and "Law of Limited Circulation of Population" [202].

- When the observation period is sufficiently short that it only includes a single observation, a law regarding radius of gyration becomes a law constraining trip lengths—lengths which observe a similar power-law, as we will discuss next.

Trip Length

Trip length refers to characteristics of the distance an individual travels over a period of time. In practice, it could be the measure of the displacement $l = \|\mathbf{x}_2 - \mathbf{x}_1\|$ [170, 204] (or distance $l = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbb{I}(\mathbf{x} \in C)\, \mathrm{d}\mathbf{x}$ on route C [240]) between two locations before and after a fixed interval of time [170], or consecutive opportunistic events such as phone calls [170, 204].

A general form to express the truncated power-law characteristic of trip length is [170]

$$P(l) = (l + l_0)^{-\beta_l} \exp(-\frac{l}{\kappa_l}). \quad \text{(Eq. 2.21)}$$

where l_0 can be considered as the minimum trip length, especially in the case when there are a significant amount of trip lengths accrued while remaining in the immediate vicinity of a single location ($l = 0$) in the data collection; κ_l is the cutoff value reflecting an overall upper bound of trip length; for trip between consecutive smartphone calls or messages [170], the exponent $\beta_l = 1.75 \pm 0.15$. The power-law characteristic of trip length is connected to characteristics of a Lévy walk [204]. In fact, by changing the exponent β_l, we can characterize a series of random walks [241].

The empirical complementary cumulative distribution function (ECCDF) of trip length may appear to be piecewise characterized by a power-law body and exponential decay due to cost and physical limits of extremely long trips [207]. On the other hand, the trip length distribution may resemble a single-piece power law when studying trips within a small region.

Trip length and human mobility predictability are often considered together. This reflects the fact that given a threshold as the maximum random walk radius, within the conventional model led by Song *et al.* [209], the distribution of trip length fully describes the length distribution of both random walks and flights. In the classic model, directions of both random walks and flights follow uniform distributions. As a result, the trip length distribution is all we need to describe a mobility model and its predictability [242].

Inter-Contact Time

In the ideal case featuring continuous observations, the inter-contact time between two devices (as the representatives of their holders) is defined as the length of the time interval over which two devices remain not in contact, between endpoints in which they are in contact [212]. However, in practice, due to cost and energy consumption, most devices can only detect contacts during a period in which they are mutually awake. Measured inter-contact time is often subject to a one-sided bias that leads measurements to tend to overestimate the actual intercontact time.

For devices using sensors such as Bluetooth, which has both discovery and discoverable mode in each duty cycle, discovery records for a pair of devices are expected to be symmetric. That is, if device A has a record stating it discovered device B at time t, then it is expected that B also has a record stating it discovered A between $(t - \Delta t, t + \Delta t)$. If only A or B has a record of such a contact, we call this discovery record asymmetric. Asymmetric Bluetooth discovery records may be caused by shifted schedule times, conflicting Bluetooth modes (explored in Section 2.3), or other reasons [243, 244].

Studies investigating inter-contact time often involve analyses of its ECCDF curves [212, 245–248]. Using data resulting from the SHED7 study, we undertook a preliminary investigation of the impact on the ECCDF curve induced by both the inexact scheduled discovery process and asymmetric discovery records. The results demonstrated that the power-law decay in the body of the ECCDF remained regardless of whether asymmetric

records are included, and missing records tend to fatten the tail of the distribution due to overestimating inter-contact time.

Inter-contact time is also studied from the perspective of stochastic processes and diffusion on temporal networks [96, 249], especially in the context of systems involving human activities. The dynamics of such systems, deviating strongly from that of a Poisson process, are "characterized by bursts of rapidly occurring events separated by long periods of inactivity" due to the highly scheduled activities of human working and living [250].

2.3 Measuring Human Mobility and Contact Networks

Privacy-respecting sensor data security is a need as smartphone-based sensing becomes popular. Much pioneer work has been conducted to address this need, including asymmetric encryption on `User ID`, one-time generated `User ID`, and non-centralized storage of contact data [103, 251, 252]. Homomorphic Encryption is an emerging solution to support privacy-preserving machine learning with sensitive data [253, 254].

2.3.1 Smartphone-Based Behavior Sensing

The prevalence and richness of smartphone based sensors have inspired a diverse set of studies ranging from agriculture management [255], vehicle telemetries [256], and smartphone-based applications for healthcare and well-being [257–261].

Most applications of smartphone-based sensing in health and healthcare can be classified into two categories: behavior-related and physiological-related. Behavior-related sensing (also referred to as community sensing [262]) focuses on capturing health-related behaviors, such as proximate contacts with other people [257], physical activity [263], dietary intakes [264], degree of stress and other aspects of psychological status [265–267]. Uses of smartphones for physiological sensing, on the other hand, record measurands such as pulse rateand oxygen saturation level[1] [268], and blood pressure [269, 270]. Approaches employing external sensing devices, often referred to as "biosensors" [271, 272], connected to a smartphone via Bluetooth, can perform unobtrusive ambulatory monitoring and assessment[2] such as electroencephalogram (EEG) [274], mood assessment and mood recognition [275], and symptoms of chronic diseases such as Parkinson's disease [276].

Many current applications of smartphone-based sensing in healthcare focus on comparing and improving its accuracy (or rather its consistency) with existing medical devices [277–279]. These analyses are likely accelerating adoption of their respective technologies in accordance with the history of now-ubiquitous pulse oximetry—a history which demonstrated that adoption of a novel measurement modality relies on comprehension by potential users of the diagnostic value of the measurements from that modality in practice [268, 280].

This section will focus primarily on how smartphone-based behavior sensing informs two sensing problems: the mobility pattern identification problem and the proximate-contact tracing problem. We will further investigate how to integrate solutions of learning problems with sensing problems to form feedback loops, improving solutions to both problems.

[1] Modern pulse oximeters measure pulse rate along with pulse oximetry. A pulse oximeter primarily works by analyzing changes in the amount of backscattered light from (invisible) infrared light sent into the tissue, in the forms of fingertips, bands on the chest or wrist, or stickers.

[2] Ambulatory assessment focuses on minimizing retrospective biases while gathering ecologically valid data from people in (near) real-time in their natural environment, including self-reported, observational and biological, physiological, or behavioral [273].

The Mobility Pattern Problem

Human mobility patterns are patterns of human geospatial trajectories. The sensing problem of recognizing human mobility patterns involves collecting trajectory data and pattern recognition with trajectory data.

Depending on the context, human trajectories are studied in either real space—where properties such as the distance between places and the travel time of each route vary—or a projected space, such as an abstract space structured[1] such that distances between places and travel times for traversing each route are equal [281]. In both real space and projected space, researchers [282–284] often simplify the characterization of human trajectories as graphs or graphs over temporal spaces, denoting places as vertices and routes between places as edges with attributes (such as accessibility or the cost of time). Trajectory data collection can be accomplished by periodically collecting the current location, so as to define a sequence of vertices, and then inferring trajectories as paths of edges between sequences of vertices. Remaining in the vicinity of a single location and traveling can be further distinguished within timestamped location sequences given a threshold. For behavior-related studies, researchers may study trajectories in a projected space so as to highlight features of interests [282, 283].

Even before the emergence of wireless sensor networks (WSNs), analysis of human mobility patterns formed a part of an emergent field called network science. Barabási *et al*. studied cellular-tower tagged human mobility data over time [209, 285] and found the predictability of human mobility is higher than many had previously suggested. Since then, a series of human mobility models [286–288] have been proposed based on study of human mobility patterns collected from pocket-sized sensors or smartphone-based sensors. With the advent of WSNs and growing prevalence of smartphones, human mobility models and human mobility data were further studied for applications such as low-capacity packet-switched networks [289], detecting traffic anomalies [217, 290], and modeling responses to the large-scale spread of infectious diseases [291] and further activity-based human mobility patterns [292]. More recent studies also proposed frameworks for feature extraction [283, 284].

The Proximate-Contact Tracing Problem

Proximity is another fundamental behavior that our definition can address as it bears on two subproblems: proximate-contact detection and (separately) tracing contact networks. The first subproblem focuses on classifying if two individuals are in proximity via sensor-collectible signals by optimizing parameters such as sensor types, available measures from each sensor type, and controllable configurations of each sensor. The second subproblem builds on the first by emphasizing optimization of the collection of proximate contacts amongst a population over time to capture important proximate contacts for individuals and reconstructing contact patterns for characterizing such individuals' contact network. The parameters being optimized in

[1] Formally, abstract space is a geographic space that is entirely homogeneous, such that all movements and activities would be equally easy or difficult in all directions and across all locations within this space.

the context of the second subproblem include but are not limited to those involving the sampling schedule (to whom to deploy sensors; when to activate deployed sensors), and interpretation of sampled contact data—estimating and (separately) mitigating bias of samples—subject to energy-related constraints of the sensors and the costs of the deployment.

For proximate-contact detection, a variety of sensor types have been employed, such as radio-frequency identification (RFID) tags [293], Wi-Fi MAC tags [294, 295], those collecting supporting GPS binning [296], Cellular Tower tags [159], and Bluetooth device-discovery based sensing [257]. Smartphone-based proximate-contact sensing is expected to have higher intrinsic validity because of the ubiquity of smartphone use. This longitudinal granularity offers unique opportunities for alleviating the impacts of communicable diseases [61].

All the sensors listed above use radio frequency (RF) signals, which radiate as electromagnetic waves (radio waves). RF signals weaken along their paths, with obstacles generally accelerating the attenuation. As a result, the signal strength of a sender sensed by a receiver can offer some insights into the distance between the receiver and sender. However, because signal strengths measured by an obstruction-separated sender-receiver pair could be similar to that from a considerably more distant pair within line-of-sight (LoS) of each other, there remains the issue of distinguishing these two scenarios offering similar signal strength readings. Whether this issue impacts the inferred proximate contacts depends on the transmission signal characteristics used, the criteria used to define proximate contacts, the character of the physical environment, and the behaviors of the population [297, 298].

We use generalized proximate contacts to include contacts that are in proximity but separated by obstacles (such as those generated by a wall or floor and ceiling). For generalized proximate contacts, we consider two "contactees" as being in contact as long as their absolute distance lies within a certain threshold. However, for contact tracing and many other behavior-related tracking purposes, we are attempting to sense only those proximate contacts not separated from each other by obstacles. This reflects the fact that many obstacles (e.g., floors/ceilings) may shield one of the pair in contact from transmission of infection or block a pair from physical interactions [63].

Sensors support generalized proximate-contact detection through either direct measures or indirect measures. Direct measures map sensor signals directly to distance, while indirect measures map sensor signals first to locations and then calculate distance from locations. Researchers often refer to the generalized proximate-contact detection problem as a co-location problem: the generalized proximate-contact detection problem is the dual to the locating problem with error. In its simplest form, determining whether two individuals are in proximate contact is equivalent to finding whether the Euclidean distance between those two individuals is less than a threshold. To locate an object with tolerance for error is equivalent to finding landmarks and other points or parties in contact with it, then use those features as location references and the distance threshold within which two points or parties are judged as proximate contacts as error bounds. Indirect measures may suffer from error amplification due to the extra transform from measured location to distance [299, 300]. In addition, RF signals with higher frequencies have fewer diffraction effects, thus lowering signal

strength beyond obstacles, making detection of contact with obstacles easier. In summary, proximate-contact detection and locating are two related questions but can still be quite different depending on the context.

Contemporary smartphones have the following on-board sensors that have been demonstrated in studies to contribute to the challenge of locating and proximity contact detection [301]:

- GPS module: A GPS module can directly measure latitude, longitude and accuracy as a radius of 68% confidence. Most modern modules can also provide bearing, the horizontal direction of travel of this device.

- Bluetooth module: While details will be covered in a dedicated section below, in short, the Bluetooth module can take advantage of the discovery phase to scan other devices with a Bluetooth module set to the discoverable mode.

- Wi-Fi module: A Wi-Fi module can search for discoverable routers, or serve as a router providing connections to other Wi-Fi modules.

- Battery level: In light of the reliable availability of sensor data on battery state, presence of such sensor data within recordings can be used to determine whether the device is powered on and recording. This can aid in distinguishing a situation in which a device is not recording from one in which recording is occurring but no connection was found.

- Magnetometer: The change of the surrounding magnetic field as measured by a magnetometer can be used to infer orientation. More importantly, it can assist detection of whether the sensing device is on the body. This can be helpful to avoid conflating information about sensor device surroundings with those concerning the surroundings of the owner when the sensing device is merely left unattended. Additionally, a linear correlation between two series of magnetometer readings—each from a smartphone—can aid in inferring that these two smartphones coexisted within a disease-contractible distance [302].

- Accelerometer: Similar to the magnetometer, within the context of detecting proximate contacts, readings from the accelerometer are employed primarily for the purpose of ensuring that the phone was carried on the person of the owner.

- Sound intensity: Jeong *et al*. [302] found that linear correlations of ambient sound level measurements from different study devices can be used to detect co-location of phones, thus serving a purpose similar to that of the magnetometer in their study. At a practical and ethical level, such use of sound is difficult due to audio being privacy sensitive.

- Temperature: Similar to a magnetometer, this sensor can help determining whether a phone was carried on-person. In some settings, this can also be used to distinguish the presence of the owner indoors or outdoors.

Due to the scope of this thesis, we will further confine our discussion on the data processing pipeline and error propagation to the GPS module, Wi-Fi, and Bluetooth, because these three sensors are most widely used for proximate-contact detection [303, 304].

GPS Module

GPS receivers rely on receiving timestamped satellites signals (with the commonly used L1 band at 1575.42 MHz and L2 band at 1227.60 MHz) to calculate distances to satellites. By combining estimated distances of the current point from at least four satellites and the expected position of each satellite, a set of quadratic equations can be solved to estimate the receiver's coordinates. Errors in the GPS receiver are classified into User Equivalent Range Error (UERE), and Geometric Dilution of Precision (GDOP) [305]. GDOP is a dimensionless multiplicative factor that reflects the geometric relationships of the GPS satellites. Researchers focused on methods and assistant systems reducing and analyzing the UERE [306]. UERE is classified into three segments, namely space, control, and user. UERE which is allocated into space and control segments is called user range error (URE). By contrast, UERE which is allocated into the user segment is called user equipment error (UEE), and is due to errors sourced from user equipment. Notable error sources by the amount of their error-budget are as follows, indicating for each the segment to which they correspond. [305]

- (URE) Frequency standard stability
- (URE) Space vehicle acceleration uncertainty
- (URE) Ephemeris prediction and model implementation
- (UEE) Ionospheric delay compensation
- (UEE) Tropospheric delay compensation
- (UEE) Receiver noise and resolution
- (UEE) Multipath effects

It is worth noting that the time-to-first-fix (TTFF) error is not stated in the error budget as listed above. TTFF is a measure of the elapsed time required for a GPS receiver to acquire the satellite signals' navigation data and calculate a position solution [305]. TTFF is commonly discussed in three scenarios: cold start, warm start, and hot start. Cold start means lacking a cached recently solved position, while warm and hot start means having a valid cache. A valid cache to enable a warm start (also referred to as normal operation) usually means having an estimated GPS solution from within the past 20 seconds, and the solved position

is within a few hundred kilometers of the current position.[1,2] The cold start has longer TTFF than the hot start, roughly on the scale of minutes versus seconds [305, 307]. Modern cellphones reduce the time to first fix by using assisted GPS (A-GPS) and other localization methods such as those involving Wi-Fi routers and cellular tower locations.

Signal attenuation is another common error source caused when using GPS for indoor positioning. A GPS receiver cannot receive attenuated signals behind a wall of buildings or other objects [307]. Other error sources for detecting proximity include projection error when converting GPS reading (usually refers to WGS84[3]) to UTM[4] zones to allow calculating Euclidean distances, and error caused by altitude inaccuracy (such that person on a different floor may be considered in close proximity to another below them when in fact they are not).

Kjærgaard *et al*. [307] evaluated GPS indoor positioning in terms of GPS availability and signal strength, time first to fix (TTFF), and accuracy. Such characteristics were evaluated by field testing with both dedicated receivers and smartphone onboard receivers.

Potential methods to deal with GPS error during data processing could be:

- GPS with an L1 receiver on smartphones has difficulties achieving a resolution finer than a 10-meter circle, especially when used for indoor positioning. Without losing useful positioning information, we can safely trim more than four decimal places from latitude and longitude readings. Given the fact that Geometric Dilution of Precision is not random but depends on satellite position and number of available satellites dependent, it will not be easy to perform noise canceling with extra decimal points to achieve higher accuracy.[5]

- Due to TTFF, GPS reading can take one or a few seconds to become available. It can be hard to attach accurate positions to a randomly timed event. We may mitigate this type of error by reading from the GPS receiver after the event occurrence and then estimate the actual position of the scene of the past event.[6]

[1]The GPS receiver relies on a reasonably close estimate of its current position and a reasonably fresh almanac to know which satellite should be visible and quickly acquire and track satellite signals. The last solved position of a GPS receiver is often used in estimating its current position. Therefore, presence of the receiver in a location further away from the last solved position delays the GPS receiver in acquiring and tracking satellite signals.

[2]`https://www.gpsworld.com/innovation-faster-higher-stronger/`

[3]World Geodetic System 1984 (WGS84) is the latest revision of the WGS standard, a geographic coordinate system used to assign a coordinate for a location on the surface of the Earth. It approximates the Earth's surface as the surface of an oblate spheroid, and coordinates are on a polar axis, with the unit of degree. Euclidean distance cannot be directly calculated between two WGS84 coordinates.

[4]The Universal Transverse Mercator (UTM) coordinate system projects a geographic coordinate system, usually WGS84, into sixty zones (ignoring areas having latitudes beyond 84°N and 80°S). Each zone is a planar surface using a Cartesian axis where Euclidean distance can be directly calculated between coordinates within the same zone. This projects has maximum scale error within 0.04% [308].

[5]`http://wiki.gis.com/wiki/index.php/Decimal_degrees`

[6]`https://docs.huihoo.com/android/4.4/guide/topics/location/strategies.html`

- The location manager of Android can return a GPS reading of $(0°, 0°)$—a location off the west coast of Africa—during cold start,[1] requiring filtering or outlier removal.

- To calculate Euclidean distance from latitude and longitude, it can be easier to choose an appropriate UTM zone to convert WGS84 coordinates into UTM coordinates[2] [309].

- GPS readings usually have a larger error when measuring in cold start, measuring at high velocity, measuring indoors, or reading from an area with lower coverage of assistance systems like cell towers, filtering out these readings may of help [305, 307].

Wi-Fi and Bluetooth Modules

Both Wi-Fi and Bluetooth are each wireless network protocols, rather than dedicated beaconing systems for proximity contact detection. Devices that support these protocols are intended for communicating and transmitting packets. The discovery ability of such devices is used to find transmitters/receivers nearby, and signal strength measurements can be used to infer the existence of and distance to nearby Wi-Fi or Bluetooth devices. We will review both protocols and their service devices for proximate-contact detection by focusing on the discovery mechanism and signal strength measures.

Wi-Fi Wi-Fi refers to a family of wireless network protocols, standardized in IEEE802.11. Devices supporting Wi-Fi follow the IEEE standards but are regulated by a non-profit third party organization called the Wi-Fi Alliance. The IEEE802.11 family has its base on 802.11-1997, and then evolved according to a series of "amendments" notably 802.11a (provided 5GHz in addition to existing 2.4GHz), 802.11b (data rates upgrade for 2.4GHz), 802.11g (further data rates upgrade for 2.4GHz), 802.11n (Wi-Fi 4, dual-band), 802.11ac (Wi-Fi 5), and newest 802.11ax (Wi-Fi 6).

Wi-Fi has supported ad hoc networks since 802.11-1997 [310]. Technically, modern Wi-Fi-supported devices can be controlled to stay in passive scanning mode; while communicating, these devices can still record other nearby devices, including access points (APs). However, these approaches will likely disturb the normal use of smartphones from communicating via Wi-Fi.

There are two scanning modes that a station can choose to search APs—passive scanning and active scanning—but only one mode can be chosen at any time [311]. Under the passive scanning mode, the Wi-Fi module on a smartphone gathers a service set identifier (SSID) from a beacon message sent every 100ms from each working AP nearby. The Wi-Fi module will listen to each channel for no longer than a maximum duration defined by the MaxChannelTime parameter during the passive scanning period. For IEEE802.11a with 2.4GHz and 13 usable channels (or three frequently used channels, 1, 6, 11), this regime means that the process of scanning all channels can require at least 300ms to finish. Due to privacy issues, APs may

[1] stackoverflow:37715680, stackoverflow:42192608, stackoverflow:24627745

[2] `https://learn.arcgis.com/en/projects/choose-the-right-projection/`

be configured to exclude the SSID field in beacon messages, making some private network sensing records effectively useless (because the SSID is empty, the receiver cannot distinguish to which AP the record belongs).

On the other hand, active scanning requires the smartphone to broadcast probe requests and wait for nearby APs to reply with probe responses. It usually costs less than 150ms to finish a full scan [312, 313]. While passive scanning can take a relatively long time to finish, frequent active scanning can increase the latency of all devices connected to the probed AP [314].

The Received Signal Strength Indicator (RSSI) is specified in 802.11-1997 and its following amendments. According to these standards, RSSI is intended to be interpreted on an ordinal scale, suitable as long as the RSSI is a monotonically increasing function of the received power on the scale of 0 to up to 255. In practice, each manufacturer chooses a different "`RSSI_Max`" [315], which can be converted to some corresponding quantity of power (dBm) [310].

Bluetooth Bluetooth is intended for ad hoc short-range wireless networks, enabling various close-proximity devices connected in a secure, reliable, and low power consumption manner. With the emergence of IoT, Bluetooth Low Energy (BLE) superseded ZigBee and became the de-facto standard of connectivity in low-power, low-cost IoT devices. It is expected that with the ongoing Bluetooth 5.x, Bluetooth will be an essential part of the data collection pipeline for many fields [316, 317].

Both Wi-Fi and Bluetooth use radio frequency (RF) signals as a form of communication. However, when multiple devices simultaneously send radio frequency signals across the same frequency range, their interference causes receivers to fail to receive messages. Wi-Fi resolves this interference problem by partitioning a radio frequency range into non-overlapping sub-ranges, called channels, and by confining signals to operate within a chosen channel. However, as discussed in the previous section, this segmenting approach causes problems when performing passive and active scanning—since typically a device can only send/receive messages on one specific channel at a time, the full scan of all nearby devices requires the scanner to iterate between each channel. In contrast, Bluetooth uses adaptive frequency hopping to address interference, with two connected devices periodically hopping to another frequency following a coordinated pseudo-random sequence. According to three of their early founders, this "gives a reasonable bandwidth and the best interference immunity ... [318]". In addition, Bluetooth's frequency-hopping virtual channel simplifies communication across channels and makes it easier to scan.

Since the inception of its architecture, Bluetooth has emphasized a quick connect procedure [318]. For each virtual channel (that is, a different frequency hopping sequence), there is at most one piconet, and within a piconet, there is one master node and one or more slave nodes. A Bluetooth device can simultaneously take a master role in some piconets and a slave role in other piconets. For example, for Bluetooth-based proximity contact detection, when a device (such as a smartphone) wants to scan all nearby discoverable Bluetooth devices, it can initiate a new piconet as the master node of this piconet, starting an inquiry taking an average 1.92 seconds (or 3.84 seconds if the first 1.92 seconds does not overlap with slaves' wake-up periods) to finish

the scan.

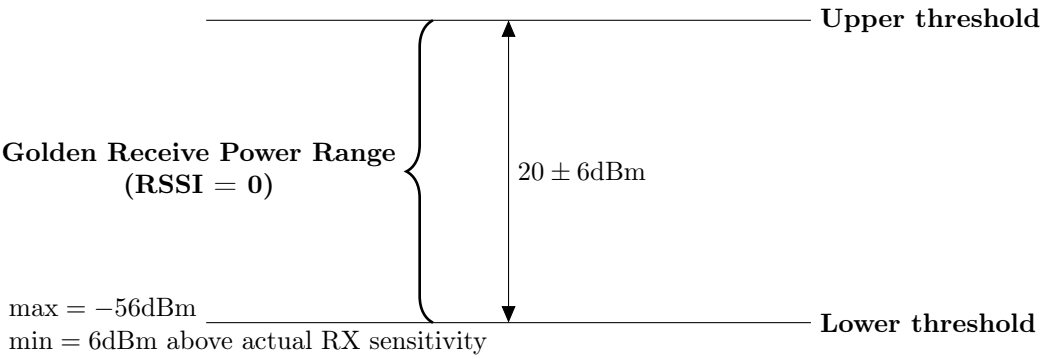


Figure 2.1: Convert Received Power to RSSI Given the GRPR

Reproduced and annotated based on [319], this graph shows the mapping of optional RSSI value to received power. The figure shows the Golden Receive Power Range and its mapping to RSSI value 0; beyond the lower/upper threshold, any positive RSSI value indicates how many dBm the RSSI is above the upper limit, any negative value indicates how many dBm the RSSI is below the lower limit.

To measure signal strength, Bluetooth also employs an RSSI. The optional RSSI fields were bound to the received power in a way that the "Golden Receive Power Range" (GRPR) was defined (see Figure 2.1). Its lower threshold level corresponds to a received power between -56dBm and 6dB above the actual sensitivity of the receiver. The upper threshold level is 20dB above the lower threshold level to an accuracy of ±6dB.

Some studies of Bluetooth-based positioning [320] make use of a Bluetooth access point (AP), yet APs are not defined in Bluetooth standards. When consulting industry specifications,[1] and patents,[2,3] we found that these so-called Bluetooth access points are one or more Bluetooth devices working as master, while devices connected to the AP are slaves.

Interference Bluetooth works on the 2.4GHz industrial scientific and medical (ISM) RF band.[4] However, many other bands overlap this band—most notably, Wi-Fi running in 2.4GHz,[5] Zigbee (standardized by IEEE 802.15.4), and other S-band microwaves such as those emitted by microwave ovens and cellular phones (the Ultra high frequency band in the definition of ITU frequency bands, running from 300MHz to 3GHz).

Bluetooth exhibits lower interference than other overlapping protocols [321–323]. Pei *et al.* [324] found Wi-Fi positioning will primarily interfere with Bluetooth even with the Apple Filing Protocol (AFP) enabled on Bluetooth. Most researchers in this area found that Wi-Fi substantially disrupts Zigbee, while Bluetooth continues to perform adequately [322, 323, 325, 326].

[1] `https://www.silabs.com/community/blog.entry.html/2019/06/28/how_to_use_bluetoothlow-energyforwi-fionboardi-CQTu`

[2] `https://patentimages.storage.googleapis.com/e5/1d/38/8523b85acbe685/US20030134596A1.pdf`

[3] `https://patentimages.storage.googleapis.com/3d/25/47/73158d3ad17c65/US7606600.pdf`

[4] Although the 2.4GHz ISM band is defined from 2.4GHz to 2.5GHz with 100MHz bandwidth, Bluetooth uses from 2400MHz to 2483.5MHz. Thus, strictly speaking, not every radio wave within the 2.4GHz ISM band will interfere with Bluetooth

[5] Wi-Fi in the 2.4GHz band defined 14 channels. Channel 1 to channel 11 are designated worldwide, from 2401MHz to 2473MHz; channel 12 to channel 14 are only designated in some regions, spanning from 2456MHz to 2495MHz.

2.3.2 Bluetooth-Based Location and Co-location

In the following section, we will focus on Bluetooth-based measures as the primary sensor data source to address the localization and co-location (LCL) problem in order to answer the following questions:

- What are those measures from Bluetooth devices that can be of help to the LCL problem?
- What is the theoretical relationship between such measures and the LCL problem?
- In practice, how are those measures used in studies, and how well have these studies addressed the LCL problem?

Bluetooth Signal Parameters

Hossain and Soh [327] provided a comprehensive overview of Bluetooth signal parameters, that is, an overview of all the status parameters of a Bluetooth connection together with any other signal values made available in the Bluetooth Core Specification, which we summarize as follow:

- Link Quality (LQ): an 8-bit unsigned integer from 0 to 255 that evaluates the perceived link quality at the receiver, derived from the average bit error rate (BER) . However, the exact mapping from BER to LQ is device-specific. Based on the authors' experiments, CSR® chips report LQ with BER resolution reduced as BER value increases, for example, with mappings like $BER \propto \log(LQ)$. In general, the authors summarize the following LQ characteristics:
  - BER-to-LQ mapping is sensitive to Bluetooth class and device specific.
  - LQ readings do not vary much at close-range distance.
- RSSI: an 8 bit signed integer from -128 to 127, logarithmically scaled with a ratio of received signal power. The authors' summary of RSSI is as follows:
  - RSSI readings tend to change significantly at close-range.
  - Combining both LQ and RSSI may be a viable option.
  - RSSI has a poor correlation with distance.
- Transmit Power Level (TPL): an 8 bit signed integer from -128 to 127, which specifies the Bluetooth module's transmit power level (in dBm). The transmitter will set TPL either to its device-specific default power setting or vary it during a connection in accordance with possible power control processes.
- Inquiry Result with RSSI: an approach for collecting the RSSI value by sending an inquiry. This approach it requires no active connection, and the radio layer of a nearby device monitors the RX power level of the current inquiry and infers the corresponding RSSI. The following characterize authors' concerns regarding inquiry results with RSSI:

– The retrieval of inquiry-based signal parameters tends to induce latency of about 9 seconds to reach the potential of discoverability (and 4 seconds to reach above 50% of its potential).

According to Hossain and Soh [327], RSSI, along with "inquiry results with RSSI"—a special inquiry procedure that perceives RSSI from the responses sent by its nearby devices—is suitable for use as the primary measurable parameters for Bluetooth signal based distance estimation; other parameters, such as LQ, may be used for data fusion.

RSSI to Distance

Path loss models based on the free-space path loss (FSPL) formula, which is derived from the Friis transmission equation, have been widely used [320, 328–333] to convert RSSI to distance. It is often cited as:

$$RSSI = -10n \log_{10}\left(\frac{d}{d_0}\right) + RSSI_0, \tag{Eq. 2.22}$$

where RSSI is the observed RSSI in dB, n is the path loss exponent that corresponds to the environment, d is the distance between the beacon and the user, d_0 is the reference distance and RSSI_0 is the average RSSI value in dB at the reference distance. In practice, we often let $d_0 = 1$ meter and consider $d \geq d_0$, rewriting Equation (2.22) into

$$d = 10^{\frac{RSSI - RSSI_0}{10n}}. \tag{Eq. 2.23}$$

Critics of this formula have noted discrepancies with empirically measured $(d, RSSI)$ pairs [320, 329, 330, 332, 334]. This has led to mitigating the model error with methods such as sensor fusion [335], filtering [330, 332], or neural networks [336].

The path loss exponent n is required to map between RSSI and distance. However, the literature exhibits very limited attempts to estimate down the realistic value(s) of n. David Young from the Android Beacon Library estimated a fully parameterized function based on experiments [337]:

$$d = 0.89976 \times \left(\frac{RSSI}{RSSI_0}\right)^{7.7095} + 0.111, \tag{Eq. 2.24}$$

where $RSSI_0$ is the RSSI value at a distance of one meter and d has the unit of meter.

We analyzed both Young's RSSI model (Equation (2.24)) and the simple Path Loss model (Equation (2.23)) with the assumption $RSSI_0 = -60$, based on our analyses from SHED series data collected by Ethica Data from smartphone-base sensors [257]. In short, both Young's RSSI model and the simple Path Loss model resembles each other in practice, given our assumption of $RSSI_0 = -60$ and for close proximity distances, which we informally defined to involve proximity within 1 to 20 meters.

Note that the simple path loss model derived from the Friis transmission equation did not check Friis's assumption that d in Equation (2.22) needs to be sufficiently large to assume a planar wavefront at a distance d to the emitter, and free space is needed to assume that the receiver power does not have further decay, such as absorption in the transmission medium, whose impact is accumulated over distance [299, 300]. Moreover,

taking the non-bijective mapping between receiver power and RSSI (due to GRPR Figure 2.1), it would be hard to calibrate data from various scenarios into a single constant n in Equation (2.23).

Beyond the simple path loss model (Equation (2.23)), a working group of the National Institute of Standards and Technology (NIST) has come up with a series of models for channel propagation or path loss [338]. The fundamental idea of their modeling of the path loss by distance is to build specific models for each environment featured by characteristics of noises from RF reflection, RF attenuation, and background noise. For different ranges of distance, piecewise forms assuming different path loss exponents were employed. The generic path loss model without considering noise as referred to in [338], when converted to a dBm scale from the original dB scale, has:

$$PL_{\mathrm{d,dBm}}(d) = PL_{0,\mathrm{dBm}} + \begin{cases} -10n_0 \log_{10}(d/d_0), & d \le d_1 \\ -10n_0 \log_{10}(d_1/d_0) - 10n_1 \log_{10}(d/d_1), & d > d_1 \end{cases}, \qquad \text{(Eq. 2.25)}$$

where d, d_0, d_1 are all in meters, and d_1 is the breakpoint where the path loss exponent increased from n_0 to n_1. The $n_1 \ge n_0$ is in accordance with the accelerated decay of power at a greater distance. $PL_{0,\mathrm{dBm}}$ is the reference path loss at $d_0 = 1$ meter, modeled with:

$$PL_{0,\mathrm{dBm}} = -20 \log_{10}(2\pi d_0/\lambda), \qquad \text{(Eq. 2.26)}$$

where λ is wavelength in meters. Plugging in $\lambda = v/h$ with further assuming $v = 3 \times 10^8$ as speed of light in a vacuum, and approximating Bluetooth RF as $h \approx 2.4$GHz, we have $\lambda \approx 0.125$ meters and $PL_{0,\mathrm{dBm}} \approx -34$ dBm. Note that by definition in Figure 2.1, this could result in a positive RSSI value, for example if the lower threshold were chosen at maximum -56 dBM, and GRPR range were taken as $20 - 6 = 14$ dBm, with the resulting value of $RSSI^{\max}_{0,\mathrm{ideal}} = 9$.

To further correct model error, random component of the path loss $PL_{\mathrm{r,dB}}$ is proposed [338], having complete path loss PL_{dB} modeled as:

$$PL_{\mathrm{dBm}} = PL_{\mathrm{d,dBm}} + PL_{\mathrm{r,dBm}}, \qquad \text{(Eq. 2.27)}$$

where $PL_{\mathrm{d,dB}}$ is the ideal path loss modeled in Equation (2.25). The random path loss can be further separated into two terms:

$$PL_{\mathrm{r,dBm}} = X_{\mathrm{s,dBm}} + X_{\mathrm{f,dBm}}, \qquad \text{(Eq. 2.28)}$$

where $X_{\mathrm{s,dBm}}$ is referred to as shadow fading, representing the deviation of the signal from its predicted deterministic model due to the presence of large obstructions in the wireless path. This effect is modeled as a Gaussian noise:

$$X_{\mathrm{s,dBm}} \sim \mathcal{N}(0, \sigma). \qquad \text{(Eq. 2.29)}$$

The second term $X_{\mathrm{f,dBm}}$ is referred to as small-scale or fast fading, representing the deviation of the signal due to the presence of smaller obstruction on the path which cause scattering of the signal or multipath effects, as a gamma distributed noise:

$$X_{\mathrm{f,dBm}} \sim \Gamma(\alpha, \beta), \qquad \text{(Eq. 2.30)}$$

where $\alpha = \beta = m$ and m is the Nakagami fading parameter [338].

The Location and Co-location Problem

Sensors usually acquire their location (and co-location) by communicating with other devices. The coverage limit and signal transmission cost determine the effective distance threshold. It is assumed that receiving a signal stronger than a preset RSSI indicates co-location (within a threshold of the RSSI indicated distance) of both the sender and the receiver, as in [106, 257, 302, 339]. We refer to this dichotomous measure-based co-location method as tagging: co-located devices serve as taggers to each other, and the co-location is represented as an edge between two vertices (denoting the sender and receiver) in a graph, which we refer to as a tagging graph. Two sensors are considered co-located if and only if there exists at least one path between their corresponding vertices in the tagging graph. Similarly, the tagging-based distance between two vertices can be defined as the minimum number of edges to be added to the tagging graph to make a path between their corresponding vertices. Although the tagging-based method can be used to detect co-location with places of interest if transmitters are placed at places of interest, in practice, the tagging method is rarely used except in Bluetooth-based locating studies [340].

The distance between sender and receiver can be estimated by measuring signal quality, such as time-to-receive and receiver signal strength. We refer to this weighted measure as the distancing method. The distancing graph can be viewed as a weighted tagging graph, where each edge is a weight indicating the distance between two nodes. We could rephrase the distancing method into classical graph problems, defining the path length (sum of all weights of edges on the path) as the maximum distance between two devices and the maximum flow (minimum of all weights of edges on the path) as the minimum distance between two devices. In addition to Bluetooth RSSI based distancing methods, acoustic sensor-based location methods use a time difference of arrival (TDoA) measure and yield high accuracy (mean error 30cm) [341–343]. However, the acoustic sensor-based methods suffer from the crowdsensing problem [344], in an analogy to the speech recognition problem known as the cocktail party effect [345]. By combining video of mouth movement, the problem of cocktail party effect can be tackled [346–348], but beaconing with acoustic sensors in crowded areas remains an open problem.

Given the absolute location of each tagger, with distances to a sufficient number of non- co-located taggers, multilateration [328, 332] can be undertaken to determine an absolute location. We refer to this kind of approach as trilateration as a minimum of three sources are received [328]. Co-location can be estimated by either directly measuring the distance to the sender or calculating distances between devices given their absolute location [332]. Note that the latter approach does not require communications between two devices to whom co-location status is to be determined. Trilateration methods are applied on GPS and widely re-applied with Bluetooth indoor positioning studies [328, 332, 349]. GPS measures use multilateration and results measured in ellipsoid coordinates. To calculate distance between two points with an ellipsoid coordinate requires projection into planar coordinates first. We often use Euclidean distance to calculate

distance between two coordinates of the Universal Transverse Mercator (UTM) coordinate system [309].

Machine learning methods have been applied based on distancing and trilaterating methods to improve the accuracy of measurements [350–353]. RSSI readings of signals sent from Bluetooth devices that are physically associated with a landmark (usually a cell from a gridded planar surface in experiments) along with their device MAC address as identities, are usually referred to as the Bluetooth fingerprints of landmarks, which can be stored as dictionaries,

$$\{\text{Landmark}_1, [(\text{MAC}_1, \text{RSSI}_1), \cdots]\}, \cdots .$$

Bluetooth fingerprints of landmarks are often taken as features to determine the closest landmark given a tuple of fingerprints from an unknown place. Analyses would then seek to locate the current unknown place by referring to the known landmark's location [354–356]. Using the fingerprint method to predict current location without history can be considered as a classification learning problem if one seeks to output the closest landmark [354, 356], or as a posterior point estimate (for example, the mean of the posterior of a Bayesian static estimate [355], the kernel method used in [335] is essentially the same). Meanwhile, to predict the current location based on previous locations with a state-space model incorporating the assumption of a random-walk can be addressed through an application of the sequential Monte Carlo method (SMC) [355].

3 Integrating Epidemiological Modeling and Surveillance Data Feeds: A Kalman Filter Based Approach

Citation: W. Qian, N. D. Osgood, and K. G. Stanley, "Integrating Epidemiological Modeling and Surveillance Data Feeds: a Kalman Filter Based Approach," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer, 2014, pp. 145–152. DOI: 10.1007/978-3-319-05579-4_18

Abstract Infectious disease spread is difficult to accurately measure and model. Even for well-studied pathogens, uncertainties remain regarding dynamics of mixing behavior and how to balance simulation-generated estimates with empirical data. While Markov chain Monte Carlo approaches sample posteriors given empirical data, health applications of such methods have not considered dynamics associated with model error. We present here an extended Kalman filter (EKF) approach for recurrent simulation regrounding as empirical data arrives throughout outbreaks. The approach simultaneously considers empirical data accuracy, growing simulation error between measurements, and supports estimation of changing model parameters. We evaluate our approach using a two-level system, with "ground truth" generated by an agent-based model simulating an outbreak over empirical micro-contact networks, and noisy measurements fed into an EKF corrected aggregate model. We find that the EKF solution improves outbreak peak estimation and can compensate for inaccuracies in model structure and parameter estimates.

Relationship to This Thesis To improve transmission modeling with sensing data, one fundamental question is whether there exists an approach to combine sensing data informed contact network with existing transmission models and resulting better estimation. The manuscript of this chapter demonstrated EKF as an approach to bridge a System Dynamics SIR model with an agent-based SIR model, provided high-resolution proximity contact data. The improvement of the EKF empowered System Dynamics SIR model on outbreak peak estimations indicates the potential of SD models to benefit from high-resolution proximity contact data and overcome inaccuracies in its model structure and parameter estimates. At the same time, the EKF improved SD models lead to the potentials for ABM to delegate to filtering-informed aggregate models some expensive simulation tasks for quick responsive "what-if" questions. Finally, this demonstration revealed that our understanding of sensing inferred proximity contact networks limits the reliability of the model. We need better understanding the quality and reliability of the sensing data inferred proximity contact network. This paper demonstrates that integrating sensed contact data to epidemiological models during an outbreak

is possible and beneficial, but replaces the errors on modeling infectious contacts due to random-mixing assumptions with unknown errors due to inferences of proximity contacts from sensing data. This paper then serves as a motivation for the work of the next two papers which attempt to analyze the impact of spatial and temporal errors on simulation results.

Author's Note In this chapter, we applied minor updates to the published manuscript to fix typos and improve clarity.

3.1 Introduction

Infectious diseases are notoriously difficult to manage, because they can exhibit great instability, with periods of quiescence interspersed by sudden outbreaks. Anticipating the future behavior of the outbreak and how interventions will affect the disease spread is important for policy makers who must marshal prophylactic and treatment campaigns. However, in the case of emerging pathogens such as SARS or H1N1, the disease dynamics and appropriate treatment regime are unknown [199].

System Dynamics (SD) models can project possible epidemiological dynamics, and aid assessment of trade-offs between interventions. Models are traditionally parameterized and calibrated when they are constructed, but frequently the underlying parameters are dependent on hard-to-predict dynamic factors such as human contact patterns, diagnosing practices, or even the weather [62]. In particular, dynamics of human contact patterns have been shown to play a significant role in the spread of disease [109, 357], and are poorly captured by even the best open-loop models.

Filtering techniques leveraging statistical inferences for dynamic models, such as the sequential Monte Carlo (SMC) method and Markov chain Monte Carlo (MCMC) methods, can be used to estimate model parameters as information becomes available [358]. The typical formulation identifies posterior distributions for parameters or outputs conditional on the model, and do not explicitly recognize the growth in model projection error as time elapses since an observation. In this paper, we demonstrate that the extended Kalman filter (EKF) can be used to adapt System Dynamics[1] models to better estimate an epidemic outbreak even when the parameters in question are not empirically or logically observable. In particular, we provide the first demonstration of an EKF-enhanced System Dynamics model evaluated against empirically observed and evolving proximity contact data. We demonstrate that the EKF-enhanced system provides obviously better estimates of the number of infectious individuals and the peak timing of an outbreak than a calibrated but open-loop SD-SIR model, particularly when contact rate in the model is repeatedly regrounded with incoming data.

3.2 Related Work

System Dynamics models have been used in a variety of epidemiological studies, and have made contributions to all areas of epidemiology [112, 359, 360]. Recently, research has highlighted the importance of population heterogeneity and network structure in shaping outbreak emergence and progression [357]. While this work has promoted the recognition of agent-based models (ABMs), such models exhibit diverse and textured tradeoffs with aggregate models [153], including the use of ABMs as synthetic ground truth whose dynamics the aggregate model is seeking to adequately characterize [95].

[1] In this manuscript, we referred to System Dynamics models as aggregate models in contrast to agent-based/individual-based models.

Some practitioners have sought to address aspects of parameter uncertainty in dynamic models using Monte Carlo methods [112, 359]. While such approaches can offer great insight into parameter-related uncertainty, using them to understand model-related uncertainty is more involved [153]. Similarly, SMC methods are applied in the context of stochastic processes [361]. The Kalman filter has been employed to address shortcomings in disease models or calibrations including temporal-spatial integration [362], dynamics of the HIV/AIDS epidemic [363], and time-varying effects of the covariates in accessing short-term pollutant exposure effects on health [364].

Recently, researchers have collected high-resolution empirical data of proximate contacts [293, 357, 365, 366] and have reported strong heterogeneity in contact patterns [293, 357, 366]. Moreover, researchers found that dynamics of proximate contacts impact the spread of disease [357, 358]. However, methods for incorporating such empirical contact data tend to work better with ABM and not aggregate dynamic models.

3.3 Model Description

For simplicity and to demonstrate that even stylized models benefit from the closed-loop design, we employed a classic System Dynamics susceptible-infectious-removed (SD-SIR) model. To test the effectiveness of the Kalman filtering approach in improving the projective accuracy of this aggregate model, we used, as a comparison, an agent-based model that is more textured but still adheres to the SD-SIR characterization of health status.

3.3.1 Agent-Based Model

We used an agent-based model (ABM) to provide synthetic ground truth data against which we could compare aggregate model estimates. The ABM employed a classic dynamic-network-based SIR formulation whose details as given in [366]. We refer to this agent-based SIR model as ABM-SIR. The ABM-SIR has 36 agents, each denoting a specific (real-world) participant. Proximate contacts among these 36 agents were replayed according to empirical proximity contact data collected among 36 participants in [366] over 92 days. Each infectious agent generated exposure events of a Poisson process with $\lambda = 0.003$; for each such event, a connected susceptible agent experienced a 25% likelihood of infection. Within simulations, no infectious individuals remained beyond the first 30 days. The time span of our experiments was therefore refined to 30 days.

3.3.2 Population Models

The SIR model is widely known and well-studied in mathematical epidemiology. The model contains three state variables: the number of *Susceptible* (S), *Infectious* (I), and *Removed* (R) individuals. Individuals are assumed to mix randomly and transitions between states are governed by memoryless processes. Equations

depicting dynamics of system states are as follows:

$$\begin{aligned}\dot{S} &= -\frac{c \cdot \beta \cdot I \cdot S}{S+I+R} \\ \dot{I} &= \frac{c \cdot \beta \cdot I \cdot S}{S+I+R} - \frac{I}{\tau}\,, \\ \dot{R} &= \frac{I}{\tau}\end{aligned} \tag{Eq. 3.1}$$

where c is the mean number of contacts made by each susceptible per unit time, β is the probability of transmission per contact between a susceptible and an infective, and τ is the mean time to recovery.

The SIR model assumes a constant mixing parameter, c, which implies emergent behavior unlikely to obtain in many empirical dynamic contact networks. Reflecting the observation that the mean contact rate can change substantially over the course of an outbreak due to both behavioral changes [367] and network heterogeneity [109, 357], we generalize to an SIRc model in which c is changed from a fixed parameter to a state variable of the model, allowing it to change over time. Rather than seeking to impose a flaw-prone behavioral model, the model initially estimates no change in c, forcing updates to c to be entirely driven by the EKF, we refer to this EFK-enhanced System Dynamics SIRc model as the EKF-SIR model.

$$\begin{aligned}\dot{S} &= -\frac{c \cdot \beta \cdot I \cdot S}{S+I+R} \\ \dot{I} &= \frac{c \cdot \beta \cdot I \cdot S}{S+I+R} - \frac{I}{\tau} \\ \dot{R} &= \frac{I}{\tau} \\ \dot{c} &= 0\end{aligned} \tag{Eq. 3.2}$$

3.3.3 Extended Kalman Filter Model

The EKF provides estimates by combining information from both model estimates and measurements. To evaluate EKF effectiveness, we used the ABM to generate noisy measurements of the count of agents in the *Infectious* state and the *Removed* state at the (discrete) time step k. Then, at each time step k, the EKF updates state estimates based on a weighted average of the model output at time k (as shaped by previous EKF updates) and the incoming measurement at time k.

Because the System Dynamics of our SIR model is nonlinear, we employed an EKF variant known as the continuous-discrete extended Kalman filter [78], referred to herein simply as the EKF. In iterated steps, the EKF updates state estimates and the covariance matrix of process noise. The covariance matrix of process noise specifies the accuracy of model estimates. The EKF consists of two processes for each iterated step: the continuous-time update, which governs the behavior of the System Dynamics SIR model and estimates covariance between measurement points; and the discrete-time measurement update, which modifies (regrounds) system state and updates covariance estimates with a consensus estimate derived from both measurements and model estimates. Interested readers are referred to [78] for details. A standard EKF formulation posits a state vector $\boldsymbol{x}(t)$ governed by a set of state equations with a nonlinear right-hand

side $\boldsymbol{f}_k(\boldsymbol{x}(k), k)$, plus process noise following a zero-mean Gaussian distribution; and a measurement-state mapping function $\boldsymbol{h}_k(x(k), k)$, mapping system states to measurements. Typically, $\boldsymbol{f}$ and $\boldsymbol{h}$ are nonlinear. The EKF approximates each function with its first-order Taylor expansion.

Initial values for the system noise covariance matrix and the measurement noise covariance matrix are usually generated based on expert knowledge and heuristics reflecting the uncertainty of dynamic model estimated system states and uncertainty of measurements. In our model, the covariance matrix of measurement noise was initialized to expect contagious contacts from zero to three individuals, based on empirical data [366] and observations. Our experiments use results from ABM simulations as the ground truth. We generate synthetic noisy measurements of I and R by artificially adding noises to the ABM simulated I and R at time k. Our EKF-SIR model uses I, R, and c as state variables. Starting from the initial conditions or the "consensus" results of the previous measurement update (whichever is later), the ordinary differential equations (ODEs) for both the process state equations and the covariance matrix P are numerically integrated from the previous state until the next measurement update. The covariance process noise, in general, grows in iterated updates, reflecting the accumulation of ongoing process noise. At the next measurement update process, the gain matrix K is calculated, and the EKF-SIR generates a new consensus state estimate based on the noisy measurements, model state estimates, and covariance matrices. In the event of physically impossible situations—such as where the numbers in the S, I, and R bins exceed their upper-/lower- bounds or move in a direction that is unrealistic (for example, the number of recovered decreasing)—we reset the Kalman filter to entirely weight towards the noisy measurements.

3.4 Experimental Setup

For each experiment, we assume an initial contact rate $c = 1$. Disease-specific parameters β and τ are assumed to hold the same value in both the aggregate (Open-Loop-SIR and EKF-SIR) models and the ABM-SIR model ($\beta = 0.25,\ \tau = 7,\ N = 36$). In the Open-Loop-SIR model and continuous-time update processes of the EKF-SIR, we used MATLAB's function `ode15s` to numerically integrate I and R over time. Since we have a closed population, for time t, the number of susceptible is simply calculated by using $S(t) = N - I(t) - R(t)$. We used similar settings to [109, 366] to simulate the ABM-SIR model. We added zero-mean Gaussian noise to the output of the ABM-SIR simulation to generate synthetic noise measurements.

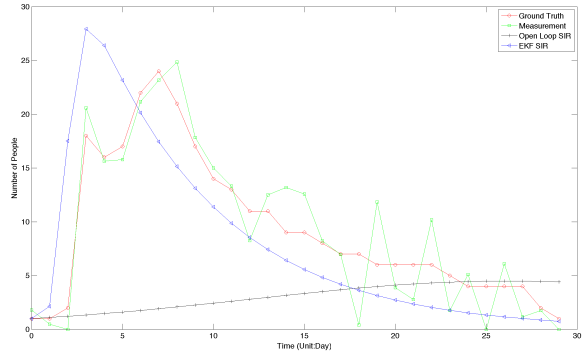
3.4.1 Kalman Filter Configuration

Within the EKF-SIR models, we used the aggregate SIRc model to update the system state estimate, and assumed that β, τ, and the noise distribution were well estimated. Two rounds of experiments were performed: firstly, based on the SD-SIR model (which has a calibrated constant c), we configured an EKF-enhanced model to investigate the filter's performance compensating for measurement noise; secondly, we configured the EKF-SIR model by applying EKF on the System Dynamics SIRc model (which includes c as a dimension

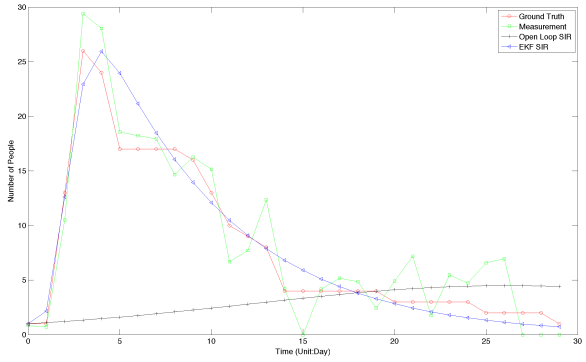
of the system states) to determine the ability of the EKF enhanced system to compensate for varying, but unmeasured parameter. For each round of experiments, the initial contact rate $c = 1$ was used based on the results of the ABM-SIR simulation to provide a highly plausible starting point for the aggregate models.

3.5 Results

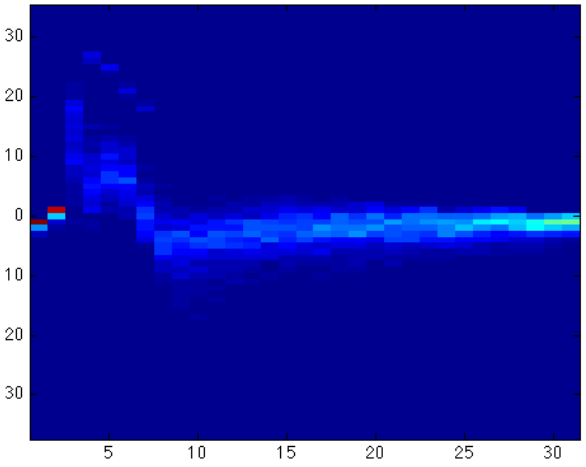
We compared the performance of our EKF-SIR model against the synthetic ground truth, the measurements corrupted by zero-mean Gaussian noise, and the Open-Loop-SIR model. It is important to note that in these comparisons, the only difference is the addition of the EKF-mediated feedback. We chose to focus our evaluation on the ability of EKF-SIR to correctly estimate the number of infectious people at each time point. Figures 3.1a and 3.1b show two example infection trajectories.



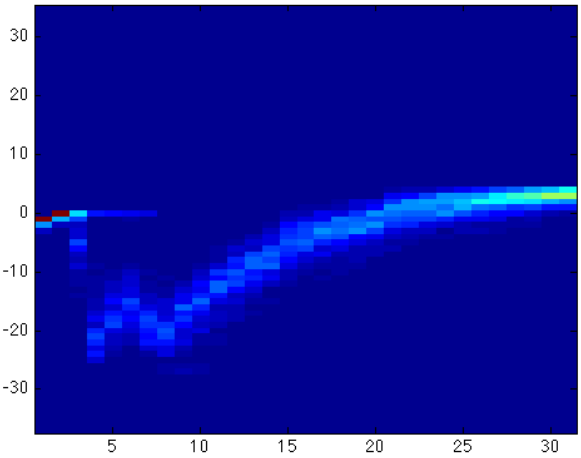
(a) Trackings of Infections I (A Typical Example)



(b) Trackings of Infections I (An Excellent Example)



(c) Histogram of Error on I with EKF



(d) Histogram of Error on I without EKF (Open-loop)

Figure 3.1: EKF Infection Trackings and Error Histograms

Infection trajectories for the results; Figure 3.1a an example of a typical tracking performance for the Kalman filter for a single realization; Figure 3.1b an example of an exceptionally good tracking performance for a single realization; Figure 3.1c histogram of error trajectories for EKF-SIR model; Figure 3.1d histogram of error trajectories for the Open-Loop-SIR model. All x-axis are time (days) and y-axis are counts of infectious people.

Figure 3.1a is an example of typical tracking performance, while Figure 3.1b is an example of exceptionally

good tracking performance. As shown in Figures 3.1a and 3.1b, obviously, the EKF-SIR does a substantially better job of tracking the infection than the Open-Loop-SIR, which diverges from the ground truth rapidly. The noisy measurement as an estimate of the ground truth is still better than either the EKF-SIR model or the Open-Loop-SIR model, perhaps due to the conservative error model we employed for measurement noise; however, the noised measurement cannot predict the system state in a future time, and cannot be used as a forecast.

Figure 3.1c and Figure 3.1d show two-dimensional histograms of the difference between the synthetic ground truth versus the Open-Loop-SIR model and the synthetic ground truth versus the EKF-SIR model. at the onset of the outbreak, both the EKF-SIR model and the Open-Loop-SIR model have errors of similar magnitude but opposite in sign. But the EKF-SIR model converges more quickly to the synthetic ground truth, reaching near-zero error after 10 days rather than 20 days. Because the Open-Loop-SIR model is running in an open loop, it makes exactly the same prediction regardless of the dynamics resulting from varying the initial infectious person for the outbreak. As such, variability in the case of the Open-Loop-SIR model is entirely due to the variation of the outbreak dynamics with a different initial infectious person and disease dynamics. The EKF-SIR model, on the other hand, attempts to re-ground the simulation at measurement steps. The performance of the EKF-SIR model is mainly due to EKF's capacity to track the epidemic given the model and the measurement data quality. The fact that the variance of the two cases is similar indicates that any additional variance introduced by the EKF-SIR estimates is of the same order of magnitude as normal variation due to disease dynamics.

3.6 Discussion

We have proposed a method for integrating epidemiological surveillance data with population-based mathematical epidemiology models. This approach creates consensus estimates of the underlying epidemiological situation from measured data and model estimates in a way that reflects the confidence modelers place in each. Compared to MCMC implementations, the EKF is computationally parsimonious, allowing for ready incorporation in sensitivity analyses and large-scale scenario exploration.

However, the current approach is accompanied by limitations. The EKF's requirement for the process noise and measurement covariance matrices may be difficult to obtain in practice, but bounding distributions can be readily derived. The assumption that the system is corrupted by white Gaussian noise could render the model temporarily obsolete if presented with non-Gaussian disturbances. The EKF will attempt to correct for the deviations due to model inaccuracies, distorting the result. However, these risks are no worse than in existing open-loop systems. We believe that there is the potential for such systems to broaden the contributions of models to practical decision-making. The capacity to keep models up to date raises the potential for much greater trust being placed in them. More fundamentally, closed-loop models lower the divisions between epidemiological data collection and modeling, encouraging decision-makers to consider an

integrated process. Finally, modelers have traditionally sought model precision through additional model complexity [153] or ABMs [368]. Closed-loop models offer a third way of enhancing prediction reliability.

3.7 Summary

We have presented a technique for creating closed-loop epidemiological models, compensating for dynamic human contact patterns. The technique described here is based on well-established stochastic optimization techniques and permits flexibly incorporating ongoing data streams data directly into the model. We have established the efficacy using cross-validation across different model types in light of empirical data on observed human contact patterns. The approach is easily integrated with population-level models, provides better than open-loop estimates with noisy measurements, and is more computationally efficient than MCMC approaches. This work has the potential to serve as a valuable tool for policymakers and researchers alike when attempting to compensate for changing parameters and flawed models during epidemic outbreaks.

4 Comparing Contact Tracing Through Bluetooth and Gps Surveillance Data

Citation: W. Qian, A. Cooke, K. G. Stanley, and N. D. Osgood, "Comparing Contact Tracing Through Bluetooth and GPS Surveillance Data," *Submitted to the Journal of Medical Internet Research*, Apr. 2022

Abstract The COVID-19 pandemic has highlighted the need for accurate and responsive transmission modeling of epidemic outbreaks. These simulations must be grounded in quantities derived from the measurement, for example, of the period over which a person is infectious or a disease's mortality rate. Challenging parameters to estimate are those associated with contacts between individuals. Digital contact tracing can provide more precise measures of proximate contacts than traditional methods based on direct observation or self-reporting. Bluetooth beaconing and GPS co-locating are two sensing modalities to collect proximity contact data; both have shortcomings and are prone to false positives or negatives as unmeasured environmental influences bias collected data. In this paper, we present a comparison of GPS and Bluetooth-inferred contact patterns and assess their impact on the attack rate induced in corresponding agent-based Susceptible, Exposed, Infectious, Recovered (SEIR) models of four different contagious diseases. We show that the contact networks generated from these two measurement modalities are different and generate significantly different attack rates across multiple datasets and pathogens. While both modalities offer higher resolution portraits of contact behavior than is possible with most traditional contact measures, the differential impact of measurement modality on simulation outcome cannot be ignored.

Relationship to This Thesis GPS co-locating and Bluetooth beaconing are two common approaches to sensing proximate contacts, but are associated with different underlying geometries. It is essential to understand whether collocating- and beaconing-collected proximity contact data will lead to fundamental differences in the inferred proximity contact networks. The investigation described in this chapter compared simulation results from an ABM model parameterized with time series of proximity contacts derived from two different sensing techniques—GPS co-locating and Bluetooth beaconing. It answers whether the sensing techniques and the spatial resolution matter to the networks inferred. Furthermore, we demonstrated to what degree and under what conditions these two sensing techniques—and, by extension, their spatial resolution—impact simulation outcomes. It is impossible to evaluate which modality was more accurate without independent ground truth. However, the substantial differences indicate that simulations of population spread of the same disease/pathogen with contact data collected with different modalities cannot be directly

compared, even for the same underlying population. This difference has implications for study design and meta-analysis.

4.1 Introduction

Infectious disease has imposed a heavy burden across the span of human civilization [369, 370]. Prior to the COVID-19 epidemic, annual influenza alone accounted for $87.1 billion worth of lost economic activity in the United States [371]. Even preventable diseases such as measles, tuberculosis, and polio have had substantial impacts on indigenous or otherwise marginalized societies [372–374]. The COVID-19 epidemic has brought the threat of contagious disease into sharp focus. With 3.44 million dead, 166 million infected [375] and over $16 trillion in lost economic activity [3] as of July 11, 2021, COVID-19 has been one of the defining global crises of the 21st century [376]. While the rapid development and deployment of vaccines have blunted the spread of COVID-19 in some parts of the world, it appears likely that the disease will become endemic, and that society will face a prolonged battle characterized by continuous monitoring, vaccine boosters reflecting the evolving variant ecology, and intermittent outbreaks [377, 378].

Transmission models have served as a key tool in the fight against contagious diseases. These mathematical models date back over a century [128–130], but have become more useful through leveraging sophisticated algorithms [113, 379] and increased computing power [380, 381]. Transmission models to predict, plan and respond to periodic COVID-19 outbreaks will be required for as long as COVID-19 remains endemic. Both compartmental [381] and agent-based [379] transmission models require well-grounded parameters characterizing not only the biology of the pathogen and host, but also the host's behavior to provide reasonable estimates of disease spread [61]. Among the most difficult of these parameters to reliably estimate are the links between population spatial behavior and infectious events—that is, how a given population of interest's movement through space aids or inhibits the spread of disease. Many public health interventions are based on altering spatial behavior to slow disease spreading, with quarantine and lockdown protocols being amongst the most direct, whereas mask use and handwashing are meant to reduce the probability of infection given exposure to pathogen.

For airborne contagious diseases like measles [373] or COVID-19 [376], the key enabler for disease spread is collocation. Spatiotemporal proximity of an infectious person to a susceptible person dramatically increases the probability of disease transmission. Spatially, collocation is being in the same volume of space at the same time, where effective spatial volume for COVID-19 is determined by aerosol dynamics, and often approximated as two meters [382]. Measuring collocation can be conducted by self-reporting, as is commonly used in classic contact tracing [383], direct observation and counts [383], or more recently by electronic means [105, 257, 384].

Two primary modalities for determining collocation using electronic devices exist: Measurements based on estimating the distance from one person to another directly (beaconing), and measurements based on estimating the location of each person of interest within a coordinate system, and calculating distances (localizing). Beaconing measurements are typically made by detecting an electromagnetic ping from one device on another device. Devices can be bespoke such as the sociometric badge [354–356], or can leverage

existing technologies like Bluetooth phones, beacons or dongles [341–343]. Localization techniques use GPS or local localization systems to place every user at a specific location at a specific time [305, 307], and can conveniently be piggybacked on existing smartphones, or mined from some social media platforms [385–387].

Unlike many systems sponsored by governments, and supported by technology from Apple and Google, determining collocation for parameter estimation of transmission models requires complete records of all interactions, not only interactions that result in infection. To estimate probabilities of disease transmission, the total number of interactions, proportion of times spent in the vicinity of frequently visited locations (also referred to as s, and spatial proximity must be measured to properly baseline the parameter estimates. Techniques from companies like Ethica Data [388] and other companies made possible by a Google-Apple partnership [389] can be used to obtain this data for target populations under transparent and ethical data acquisition practices (preferably overseen by some type of institutional ethics board). However, the underlying physical processes and mathematical treatment of beaconing and collocation data are substantially different, and have different failure modes. Previous research had not elucidated the degree to which the use of such techniques would yield disparities in the estimated contact patterns for the same population. It is simple to hypothesize that co-locating and beaconing will yield different contact patterns, but it is less apparent how the differences will interact with diseases dynamics and impact the overall simulation outcomes.

In this paper, we examine the contact patterns derived from three previously collected datasets employing both Bluetooth beaconing and GPS localization on smartphones running the Ethica Data app. We demonstrate that while the underlying contact patterns generated from co-locating and beaconing are broadly similar, they contain salient differences. For each of four pathogens marked by different dynamics, we compare the results of an agent-based simulation of a communicable disease outbreak for that pathogen parameterized with beaconing and localization derived contact patterns. The results demonstrate that the method used to estimate contact patterns gives rise to significant differences between estimates of key outbreak parameters. In particular, we show that GPS-based contact patterns estimate significantly fewer and less severe outbreaks than Bluetooth-derived contact patterns for the same participant and device. This result is mostly insensitive to disease and contact distance threshold, and for the most part holds across datasets, with the magnitude of the effect changing, but not the direction.

4.2 Literature Review

Transmission models for communicable diseases are based on the characterization of the natural history of a condition and contact networks [61]. Beyond traditional population-based non-spatial approaches, agent-based epidemiological models can take individual-level contact records and behaviors to identify emergent patterns in a bottom-up approach [379].

Real-world proximity tracking has applications in contact tracing, location-based risk assessment, mobility tracking, and outbreak detection [390]. Deriving real-world proximity contact mainly falls into two categories:

calculating the delta of measured absolute positions—with, for example, GPS and Wi-Fi network-assisted locationing [294, 296]—and directly measuring the relative distance with, for example, Bluetooth [257, 391], or RFID [293].

Exemplars of each of these two approaches—GPS and Bluetooth—have been studied for digital contact tracing and epidemiological simulation [382, 390]. Recent comparisons between GPS- and Bluetooth-inferred proximity contact collection approaches focus on privacy-preservation, adoption, and compliance rates [382]. By contrast, the accuracy of simulations with GPS- and Bluetooth- derived proximity contacts have yet to be quantified across different underlying populations and pathogens [391].

Advances in digital contact tracing have also contributed to disease parameter estimation. At the beginning of the COVID-19 pandemic, researchers focused on estimating the basic reproduction number R_0 from limited and highly regional dependent infection data; as the pandemic spread, data collection and reporting standards have enabled daily reporting of incident cases, active cases and mortality for various geographic scales over time, allowing estimation of the effective reproduction number R_e [137–140].

4.3 Background

4.3.1 Bluetooth Proximity

Bluetooth is a short-range communications protocol incorporated into most smartphones, where it is commonly used to pair with devices such as wireless headsets. It is a low-power protocol designed to operate over short ranges to facilitate local connections. By default, Bluetooth is configured to be in a quiescent state, not advertising its presence and only communicating with devices that have been paired. Prior to 2020, it was possible to lock an Android phone into a more active discovery mode, where the device would beacon approximately every eight seconds, advertising its presence to other devices. While this functionality was intended to provide ease of initial device pairing, it could be repurposed to detect the proximity of two devices by registering when one device received a discovery ping from another.

Several studies [244, 330] have investigated the use of Bluetooth to estimate the spatial proximity between devices representing people. The simplest methodology will be to create a proximity event between two devices if one device detects a discovery ping from the other or vice versa. However, the distance between devices is a relevant parameter for determining a valid proximity event or contact in many applications. Researchers have typically used the Received Signal Strength Indicator as a proxy for distance [350–352], assuming an exponential falloff of signal strength with distance [392, 393]. This approximation is confounded by reflections or transmissions off or through objects, meaning that RSSI cannot be strictly interpreted as distance except in all but the most controlled conditions. RSSI values can plausibly be used to filter out contacts that are either far away or on the other side of a barrier, such as a wall.

The Received Signal Strength Indicator (RSSI) measures signal strength in decibel-milliwatts (dBm), where $RSSI = 0$ is defined by a "Golden Receive Power Range" (GRPR), whose lower threshold level

corresponds to a received power between -56dBm and 6dB above the actual sensitivity of the receiver, and whose upper threshold level is 20dB above the lower threshold level to an accuracy of ±6dB. Beyond the lower/upper threshold, any positive RSSI value indicates how many dBm the RSSI is above the upper limit, any negative value indicates how many dBm the RSSI is below the lower limit. Usually, the stronger the signal strength (higher RSSI) indicates closer distances between two Bluetooth devices, but orientation, barriers, and interference can attenuate the signal strength beyond what the distance would suggest [331]. David Young [337], and Android Beacon Library [394] contributed an RSSI to distance function based on Nexus 4 and Apple's iBeacon performance which is often used as a first approximation for similar location awareness services on modern smartphones

$$d = 0.89976 \times \left(\frac{RSSI}{RSSI_0}\right)^{7.7095} + 0.111, \qquad \text{(Eq. 4.1)}$$

where $RSSI_0$ is the RSSI value at a one-meter distance.

4.3.2 GPS and Location Proximity

Global Positioning System receivers are standard on smartphones, enabling location-based services and route finding. Consumer-grade GPS receivers typically have a nominal accuracy of 10 meters, but can be subject to substantially larger error due to environmental factors. Neither iOS nor Android employs pure GPS localization in their location estimation services. Both additionally employ initial estimates from cell tower locations (AGPS) as well as fingerprinting-based localization employing databases of detected Wi-Fi routers. Because GPS receivers often take several seconds to obtain a position lock, even with APGS assistance, smartphone localization services tend to default to Wi-Fi-based localization initially, then switch to GPS as better location estimates become available. For simplicity of presentation, the balance of this paper uses the term GPS to refer to location estimation regardless of whether it was obtained through GPS, AGPS, Wi-Fi fingerprinting, or some combination thereof.

Given records of locations, a dichotomous notion of proximity can be defined in which two agents are considered proximate if they are in the same place at the same time. Precision and accuracy of the measurements and the context of the definition of proximity determine how close in time and space agents must be to be considered proximate or in contact. When using commodity smartphone localization hardware and services, accuracy below 5 meters is rare [307], so spatial proximity has a strong lower resolution limit. Temporal resolution is substantially better—on the order of seconds—and is more likely to be limited by the measurement regime or application requirements. Elevation estimates are even less reliable than spatial estimates, so commodity GPS receivers are often projected onto a two-dimensional plane, introducing the potential for erroneous connections between people at the same location but on different floors of a building, for example.

While both GPS and Bluetooth can provide higher fidelity estimates of proximity and contact than traditional surveys or diaries, both are prone to false positives and negatives. Given two devices separated

by a mutually proximate wall or ceiling/floor, Bluetooth can still report contacts because the attenuation of RSSI will be such that they appear in contact but farther away. GPS is prone to false positives for detecting proximity for communicable pathogens because the distance over which transmission can occur is smaller than the accuracy threshold for commodity devices. GPS proximity can only be interpreted as close enough that contact was possible, given the error in measurement, not that contact actually occurred. Bluetooth can produce false negatives if the beaconing and listening cycles of the devices are misaligned, such that one device is beaconing while the other is asleep. GPS can lose signal or accuracy when indoors, causing false negative contacts by either having no location reported for an agent, or exhibit position inaccuracies which render inaccurate co-location calculations. While the underlying true contact dynamics for the same devices are identical, the differing failure modes of GPS and Bluetooth means that data drawn from those data collection modalities may generate different contact networks, and thereby suggest different contact dynamics, and ultimately, different outbreak dynamics.

4.3.3 Agent-Based SEIR Models

The SEIR disease state model is a classic model to characterize, the pathogen transmission and the natural history of infection across a range of communicable diseases. Disease state transitions are unidirectional in the order of susceptible, exposed, infectious and removed. The initial state of the model specifies the amount of population in each disease state, and the rate of transition between disease states are subject to both disease-characteristic parameters (such as latent period and infectious period) and contact network (such as preferential mixing and average contact rate). It is common for a specific disease, given surveillance data, to have more detailed models. For example, there are models of COVID-19 splitting SEIR into more states and re-route transitions in states [141, 142]. Because our goal is to probe the impact of contact measurement modality, in accordance with Occam's razor, we choose the simplest SEIR model.

Agent-based models incorporate individual interactions and track the state and state transitions through which each individual progress. Unlike a stock and flow model, which uses differential equations to model the flow of individuals from one state to another in aggregate, an agent-based model knows the state of every agent individually at any time step of the simulation, and aggregate statistics, for example, on infections, are queried and computed during post-processing. An agent-based SEIR model captures both individual disease state transitions based on disease-specific parameters such as latent period, infectious period, R_0, as well as some abstraction of the contact behavior of the population. Because the simulation of an infectious disease can capture emerging patterns in a bottom-up manner [379], and more faithfully reflect dynamics due to the proximity contact network than compartmental models, agent-based models provide higher fidelity at the cost of computation when compared to stock and flow models. Because an agent-based model can directly employ a contact pattern as part of the simulation, it is the logical choice for examining the sensitivity of simulations to the contact detection methodology.

4.4 Methods

4.4.1 Dataset Description

For this work, we employed three previously collected datasets, all of which were collected from the city of Saskatoon, a city in the mid-western Canadian province of Saskatchewan. In all these datasets, readings from additional sensor modalities (for example, accelerometers, gyroscopes, and Wi-Fi traces) were also collected, but only the Bluetooth, GPS traces, and battery data were used in this study. Battery data were used to identify gaps in data collection. If the phone is on, and Ethica is running, then battery data will be recorded, providing a more reliable way to assess the continuity of data collection than is possible with GPS, where signals can be obscured by the built environment, but where the phone is still actively recording. The Saskatchewan Human Ethology Datasets (SHEDs) are a collection of pilot projects and technical trials taking place during the iEpi project—the academic precursor for the Ethica Data system—and associated post-processing and methodological outcomes [395, 396]. The SHED datasets were exclusively collected from populations at the University of Saskatchewan in Saskatoon. The SHED7 dataset was collected between July 11, 2016 and August 8, 2016, and included 61 students. The SHED8 dataset was collected between September 25, 2016 and October 25, 2016, and included 74 students. The SHED9 dataset was collected between October 28, 2016 and December 9, 2016, and included 88 students. These participants were part of a social science student study pool that included both undergraduate and graduate students, weighted towards undergraduates. All data were collected with the informed consent of the participants and under the oversight of the institutional research ethics review board.

4.4.2 Sensor Data Processing

To evaluate the performance of each sensor under real-world scenarios, we needed to account for the impact of participant compliance. We defined the active period of a study with the start day as the first day when we have no less than 80% of participants' battery reading, and the end day as the first day with all following days having less than 80% of participants' battery reading. We retained participants who had at least 50% of daily battery data. Descriptive statistics are shown in Table 4.1.

Ethica's multi-sensor sensing mechanisms request that the Android operating system perform sensor scanning and reading periodically. We call our requested period length—that is, each of the repeated 5-minute time windows—a duty cycle. For location and Bluetooth contact data used by this paper, Ethica records for 1 minute starting every 5 minutes.

The Bluetooth discovery record from the Android API includes the MAC-address of the discovered Bluetooth device and the associated RSSI. After linking such discovery records to the participant-id via the smartphone Bluetooth MAC-address table collected after consent and before the experiment started, we created a table of Bluetooth discovery records of eligible participants. Those RSSI values were filtered to

Table 4.1: Sensor Data Table

	shed7	**shed8**	**shed9**
Total number of participants	61	74	88
Number of retained participants	61	74	78
Total number of days in studies	35	31	41
Number of active days in studies	28	30	38
Number of Bluetooth-inferred contacts (dist. threshold 8m)	37,804	34,400	20,597
Number GPS-inferred contacts (dist. threshold 10m, accuracy 10m)	4338	6784	5064

include records associated with an RSSI stronger than the RSSI values associated with the desired distance thresholds, and then aggregated maximum RSSI for unique tuples of discovered-participant and duty cycle (data collection epoch), resulting in the final BT contact record table. Although Bluetooth discovery records are directional, our usage of unique tuples will consider a pair of participants potentially in contact if at least one's Bluetooth device discovers that of the other.

Starting with raw GPS readings, for each participant, we first discarded GPS readings having an accuracy radius larger than 10 meters as being too inaccurate to allocate even approximate co-location estimates. For each participant, we used the median of their GPS readings within a duty cycle as the estimated geolocation of that participant. We then mapped the estimated GPS coordinates of latitude and longitude onto UTM coordinates as Northing and Easting with units of meters. For the sake of estimating inter-participant proximity, we used the Euclidean distance between the estimated geolocation for all pairs of participants within the same duty cycle as the estimated distance between pairs of participants. For each duty cycle, participants who lacked GPS readings within that duty cycle were considered as not being in contact with any of the other participants for the duration of that cycle.

ABM-SEIR model

An agent-based susceptible-exposed-infectious-removed (ABM-SEIR) simulation model was employed to characterize pathogen transmission and describe the natural history of infection. The model assumed the following:

- There is no reinfection during the simulation period.
- The population is closed, and no birth, death, and migration occur during the simulation time horizon.
- The latent periods for diseases under consideration are similar to the incubation periods.
- During the infectious period, an infectious patient will have a constant hazard rate of transmission to

every one of their currently contacted persons, normalizing passive shedding from active spread (e.g. sneezing) over a contact period.

- There are no behavior changes for participants during the simulation period, conditional on the contact patterns measured. For small outbreaks, this is reasonable, but COVID-19 has demonstrated the importance and magnitude of changes that can occur to hygienic personal protective behavior (e.g., mask use) over the course of a pandemic.

We made use of a four-fold duplication and concatenation of both GPS- and Bluetooth-inferred proximity contact data—like successively replaying a movie—to allow the outbreak to run its course, without running out of contact data.

All participants connected to at least one other participant after filtering were included in the simulation. Each simulation starts with one initially exposed participant. We conducted multiple simulations with different random seeds to account for stochastics. Every simulation began with a single infectious participant. All active participants were the initially exposed participant in turn, for 50 realizations each.

Table 4.2: Number of Participants With at Least One Contact Within the Study

	shed7	**shed8**	**shed9**
Bt8	58	71	76
Bt20	58	71	76
GPS8	49	63	66
GPS20	58	71	74

During the initialization of each simulation realization, for each participant, the latent period and infectious period were drawn uniformly from the minimum-maximum range of corresponding parameters as shown in Table 4.3.

Table 4.3: Disease Parameter Table

	R0	**Incubation Period (min–max)**	**Infectious Period (min–max)**
covid19 (non-variant)	†2.2 [144, 397]	5.6 – 7.7 [144]	3 – 7 [144]
flu	‡3 [398]	1 – 4 [399]	3 – 5 [399]
norovirus	1.75 [400]	0.5 – 2 [401]	2 – 3 [401]
measles	‡15 [402]	10 – 12 [403]	8 – 11 [403]

†Derived as midpoint of reported range. ‡Derived range from different reports.

Simulation Configuration

For each SHED study, after pre-processing, we obtained GPS- and Bluetooth-inferred proximity contact data with distance-equivalent RSSI thresholds of -80 dBm (corresponding to about 8 meters) and -90 dBm

(about 20 meters). A group of simulations for each of the four diseases—namely flu, COVID-19, measles, and norovirus—was run. Within each group of simulations sharing the same derived proximity contact data and diseases parameters, we iterated each of the active participants as the initially exposed patient with 50 realizations, where each realization has a different predetermined random seed, resulting in 170400 realizations, in total, across all datasets and conditions.

In our agent-based SEIR model, at any given time during the simulation, each agent resides in one of the four disease states (susceptible, exposed, infectious, removed). At the start of the simulation, all agents except the designated initially exposed agent are susceptible. The transition from susceptible to exposed has a probability p when exposed by proximity to an infectious agent. Such occurrences of exposure are characterized by a Poisson process with a mean inter-arrival time of 5-minutes. The value assumed for p is derived from the disease-specific R_0 and the average empirically observed frequency of population contacts. The timing, duration and the pair of agents involved in each proximity contact are given by to the proximity contact data fed to the simulation. The transitions of exposed-to-infectious, and infectious-to-removed are timeouts with timers set as the corresponding latent period and infectious period as initialized to each individual.

Simulations were run on two servers, each with Intel® Xeon® CPU E5-2690 v2 and 503GB memory. Models were created in AnyLogic 8.1.0 and exported to a standalone Java application with OpenJDK 1.8.0_252 as the runtime environment. Analysis was conducted in R 4.0.2 with major packages including tidyverse 1.3.0, ggprah 2.0.5, igraph 1.2.6, and in Python 3.8.0 with major packages including pandas 1.2.0, numpy 1.20.2 and scipy 1.6.1.

Evaluate Impacts on Transmission Models

We use attack ratio as the metric to evaluate the impact of proximity contact data on transmission models. The attack ratio is the proportion of the total population who ever get infected throughout the simulation. Although the ABM-SEIR model can produce many estimates for different disease parameters given proximity contact data, the attack rate and individual risk of infection was chosen for simplicity, accessibility and on account of serving as single summary statistics [404, 405].

Welch's t-Test Assuming disease M and underlying population V, the choice of an initial infectious individual ν is independent of the data collection configuration (sensor type ω and proximate distance threshold ϵ). We are interested in the marginal $P(\Theta \mid \omega, \epsilon, M, V) = \sum_{\nu \in V} P(\Theta \mid \omega, \epsilon, M, V) P(\nu \mid M, V)$. Limited by our knowledge of $P(\nu \mid M, V)$, we assume the initial infectious individual ν is chosen with uniform probability from the underlying population V, that is $P(\nu \mid M, V) = \frac{1}{\|V\|}$. Consider $\mathcal{X} = \bar{X}$ as the sample mean from a sample $X_i \sim P(\Theta \mid \nu = v_i, \omega, \epsilon, \mathcal{M}, V),\ i = 1, \cdots, \|V\|$, and we sample $\bar{X}$ by repeating $\|V\|$ simulations iterating every individual of the population V as the initial infectious individual. By the central limit theorem, samples of $\mathcal{X} \sim P(\mathcal{X} \mid \omega, \epsilon, M, V)$ tends toward normally distributed to suffice the assumption

of Welch's t-test.

Pairwise t-Test Without assuming the initial infectious individual ν is homogeneous among the underlying population V (that is $P(\nu \,|\, M, V) = \frac{1}{\|V\|}$), we could construct a pairwise t-test by pairing the samples of attack rate having the same initial infectious individual μ, given sensor type ω, distance threshold ϵ, for each pair of disease M and underlying population V. In this case, we assume the pairwise differences of attack rate, such as for $\Theta_i^{\text{BT8-GPS20}} = \Theta^{\text{BT8}} - \Theta^{\text{GPS20}}$, are normally distributed, where $\Theta^{\text{BT8}} \sim P(\Theta | \nu = v_i,\ \omega = \text{BT}, \epsilon = 8, M, V)$, $\Theta^{\text{GPS20}} \sim P(\Theta \,|\, \nu = v_i,\ \omega = \text{GPS}, \epsilon = 20, M, V)$.

Kullback-Leibler Divergence of Individual Infection Risks Given sensor type, proximate distance threshold, disease and underlying-population, we estimated individual infection risk based on the Laplacian-smoothed rate of being infected across realizations, denoted by $\rho_{v \in V}(\omega, \epsilon, \mathcal{M}, \mathcal{V})$. The likelihood of being the most likely infected individual for an individual $v \in V$ follows $P(v \,|\, \omega, \epsilon, \mathcal{M}, \mathcal{V})$, which can be estimated by normalizing vector $\boldsymbol{\rho} = \{\rho_v \,|\, v \in V\}$. The Kullback-Leibler divergence was used to summarize differences between pairs of sensor type and proximate distance threshold (ω, ϵ) within blocks by disease and underlying-population. For disease M and underlying population V, we have $D_{\text{KL}}(\phi_{w_1,e_1} \parallel \phi_{w_2,e_2})$ between sensing configurations (w_1, e_1) and (w_2, e_2), where $\phi_{w_i,e_i} = P(v \,|\, \omega = w_i, \epsilon = e_i, \mathcal{M} = M, \mathcal{V} = V),\ w_i \in \{\text{BT}, \text{GPS}\},\ e_i \in \{8, 20\}$.

4.5 Results

While the agent-based simulation utilizes dynamic contacts, some insight can be gained by examining the aggregate contact network for participants in each study. Figure 4.1 shows the aggregate contact networks for SHED7, 8 and 9 using Bluetooth (BT), and GPS at 8m and 20m thresholds. If a connection ever occurred between two nodes given the protocol, a corresponding edge is drawn in the network, with the color of the edge proportional to the total contact duration over the course of the experiment between those nodes. Reflecting the Pareto-like distribution of contact duration, colors move from blue (weakly connected) to red (strongly connected) on a logarithmic scale, consistent with other human network observations [207, 406]. As expected, most nodes appear to have weak connections compared to highly connected dyads and triads in the network. The Bluetooth networks are denser and more highly connected than their GPS counterparts, implying a greater potential for disease spread. There is a greater preponderance of weak edges in the Bluetooth datasets than amongst their corresponding GPS counterparts. There is a modest increase in the number of edges between the 8m and 20m thresholds for each dataset.

Contact frequency (the rate at which contacts occur) and inter-contact time (the time between contacts) are common aggregate distributions used to characterize contact datasets. Like many other datasets, both the BT and GPS demonstrate power law decay for the probability of a contact duration and inter-contact time (Figure 4.2). GPS-based contact detection tends to infer more and shorter duration contacts, but exhibits truncated tails. In SHED7 and SHED9, the tail truncation leads to fewer long duration contacts (more than

600 minutes) than BT. Inter-contact times are similar for all datasets, but BT distributions are skewed more heavily towards longer inter-contact times than is the case for GPS. By contrast, for SHED8 and SHED9, BT tracking detects notably fewer moderately long contacts (those in the range of 50-600 minutes). This may be due to localization noise-induced false positives in the GPS dataset skewing the apparent contact durations higher.

After filtering connections for the appropriate distance threshold (8 meters and 20 meters), the agent-based simulation was run according to the protocol described in Section 4.4.2. Many runs do not produce an outbreak, with the initially exogenously infected individual the only member of the network infected. This results in a zero-heavy bimodal distribution of cumulative infection counts per realization, with a Poisson spike at zero cumulative endogeneous infections (one exongenous infection) and a second distribution describing the probability of an outbreak of a given size, conditional on outbreak occurrence (i.e., the probability of at least one endogenous infection). A stacked bar plot showing the ratio of runs in which further incidences beyond the initial infectious individual did or did not occur is shown in Figure 4.3. The figure clearly shows a higher likelihood of an outbreak occurring with the Bluetooth data, as expected from the aggregate network diagrams and aggregate contact duration and frequency plots. The consistent difference in the probability of outbreak occurrence between the two conditions is our first substantial indication that the two means of measuring contact are not equivalent. To determine the impact of each dynamic contact pattern on the outbreaks themselves, the trials in which no endogenous infection occurred were removed, and statistical analysis conducted on the distribution of outbreak severity conditional on outbreak occurrence.

The core research question of this paper was whether and to what extent the differences in GPS and Bluetooth based proximity detection would alter the contact network and therefore the implied attack rate. We consider the attack rate $\theta = \frac{\int_{t=0}^{T} I(t)\mathrm{d}t}{N} \in [0,1]$ (the proportion of the population that is infected) as the response variable (denote as Θ) to controlled variables of the disease/pathogen $\mathcal{M}$, the initial infectious individual $\nu \in V = \{v_1, v_2, \cdots, v_n\}, n = \|V\|$, and collected proximity contact data $\mathcal{D}(\omega, \epsilon, V)$. For that proximity contact data $\mathcal{D}(\omega, \epsilon, V)$, $\omega \in \{\mathrm{BT}, \mathrm{GPS}\}$ is the sensor type, $\epsilon \in \{8, 20\}$ is the distance threshold of proximate contacts, and V is the underlying population. So with the ABM-SEIR model as $P(\cdot)$ for specific disease $\mathcal{M}$ and underlying population V, we can sample $\Theta \sim P(\Theta = \theta \mid \mathcal{V} = \nu, \omega, \epsilon, \mathcal{M}, V)$ with simulation realizations. While the initial infectious individual ν has been known to impact attack rate Θ, investigation of that impact lies outside the scope of this manuscript.

The Bonferroni-corrected Shapiro-Wilk test checking normality is passed for each 50 samples of $\mathcal{X}$ for every pair of disease $\mathcal{M}$ and the underlying population V, except for COVID-19 with contact records collected via GPS using distance threshold 20 over underlying population SHED8. The result of a Bonferroni-corrected Welch's t-test are shown in Table 4.4.

The core research question of this paper was whether and to what extent the differences in GPS and Bluetooth based proximity detection would alter the inferred contact network and therefore the induced attack rate. The results of Bonferroni-corrected pairwise t-tests [407] between observed attack rates (having filtered

Table 4.4: Welch's T-Test Table

Pairwise t-test with Bonferroni correction		number of incidences ~ proximity contact data source with distance threshold					
p-value		Bt8-Bt20	Bt8-GPS8	Bt8-GPS20	Bt20-GPS8	Bt20-GPS20	GPS8-GPS20
norovirus	shed7	$p < 0.001$ ***	0.052	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
	shed8	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
	shed9	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	0.001 **	$p < 0.001$ ***
flu	shed7	$p < 0.001$ ***	0.117	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
	shed8	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	0.001 **
	shed9	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
covid19	shed7	$p < 0.001$ ***	0.845	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
	shed8	0.039 *	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	0.038 *
	shed9	0.001 **	$p < 0.001$ ***	1	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
measles	shed7	$p < 0.001$ ***	0.003 **	$p < 0.001$ ***	$p < 0.001$ ***	1	$p < 0.001$ ***
	shed8	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***
	shed9	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***	$p < 0.001$ ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

out scenarios with zero endogenous infections) across all simulation runs for a condition are shown in Table 4.5. These results confirm our hypothesis that Bluetooth and GPS-based contact histories induce significantly different estimates of total disease burden across multiple simulated realizations. The primary comparisons are the BT8-GPS8 and BT20-GPS20, with others included for completeness. For SHED7 BT8-GPS8, the results are not significant. For all other diseases and datasets, the results are statistically significantly different. In the case of BT20-GPS20, all results are significantly different with the exception of SHED7 measles. While we suspected that the infectiousness of the disease would impact simulated outcomes, the results seem to be dominated by differences in dataset and contact measurement modality. Looking at the impact of resolution, some combinations of dataset and disease are not significantly different, but for the most part, increasing the threshold increases the number of contacts, driving differences in simulated outcomes. The exception to this general rule seems to be SHED8 GPS8-GPS20, where increasing the threshold did not significantly alter the outcomes for most diseases, and only marginally for measles.

Figure 4.4 shows violin plots for the attack rates over each realization across all simulated conditions, and provides insight into the statistical results from Table 4.5. In Figure 4.4, the width of each violin indicates the empirical probability density of the attack rate at the corresponding y-coordinate. Each violin represents an empirical probability density distribution of attack rate resulting from the multiple realizations across different initially exposed agents and random seeds. The y-coordinates of each red dot and corresponding bar indicates the mean and mean plus or minus the sample standard deviation. As shown in Figure 4.4, SHED7 consistently exhibits smaller attack rates for all diseases, with smaller variance and mean than other data sources. The limited attack rate likely drives the similarity between Bluetooth and GPS. The denser SHED8 and SHED9 networks have substantially larger variance, leading to significant differences between measurement modality conditions. In particular, the highly contagious measles virus exhibits marked differences within SHED8 and SHED9 datasets. In general, Bluetooth contact patterns have longer tails, indicating a greater possibility for

Table 4.5: Pairwise T-Test Table

Pairwise t-test with Bonferroni correction		Cumulative incidence count ~ proximity contact data source with distance threshold					
		Bt8-Bt20	**Bt8-GPS8**	**Bt8-GPS20**	**Bt20-GPS8**	**Bt20-GPS20**	**GPS8-GPS20**
norovirus	**shed7**	p < 1e-10 ***	0.2505	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***
	shed8	0.4436	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	1
	shed9	p < 1e-10 ***	0.0114 *	0.0046 **	p < 1e-10 ***	0.0075 **	p < 1e-10 ***
flu	**shed7**	p < 1e-10 ***	1	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***
	shed8	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	1
	shed9	p < 1e-10 ***	0.0036 **	0.0189 *	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***
covid19	**shed7**	1e-04 ***	0.729	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***
	shed8	0.0689	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	1
	shed9	p < 1e-10 ***	p < 1e-10 ***	1	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***
measles	**shed7**	p < 1e-10 ***	1	p < 1e-10 ***	p < 1e-10 ***	0.4877	p < 1e-10 ***
	shed8	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	0.002 **
	shed9	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***	p < 1e-10 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05							

larger outbreaks throughout the population. In cases where a substantial probability mass is contained in the tail, the median is also drawn higher, as in SHED8 with BT20 for measles.

Figure 4.5 shows Kullback-Leibler divergence (KL-divergence) on individual infection risks within blocks of disease and underlying population, where bars indicate corresponding KL-divergence $D_{\mathrm{KL}}(p \parallel q)$ between pairs of sensor configurations (p, q) annotated as ticks p-q on the x-axis. The individual infection risks are reflected by the likelihood of being the most likely infected individual between different sensing configurations, where a sensing configuration is a pair of selected sensor types and proximate distance thresholds. Whether it is considered to be 8 meters or 20 meters, the distance threshold of proximate contact does not appear to impact GPS-collocated inferred proximity contacts in terms of individual infection risks, regardless of the underlying population. This invariance to distance thresholds suggests that the primary bottleneck lies in the GPS-collocation method's inability to identify exact proximity contacts from a group of collocated individuals. Meanwhile, the BT-beaconing method may capture proximity contacts at certain distance thresholds (such as for SHED7 and SHED8), which can be important when considering droplet-based pathogen transmission. The underlying population in SHED9 is known to behave in a less spatially clustered manner than for SHED7, SHED8, which might contribute to a lower magnitude of KL-divergence for BT8-BT20 and BT20-BT8. The KL-divergence among pairs of different sensor types is similar regardless of distance thresholds of proximate contact, suggesting that BT-beaconing and GPS-collocating collect different proximity contacts regardless of the resolution on distance thresholds of proximate contacts. The magnitude of asymmetric $|D_{\mathrm{KL}}(p \parallel q) - D_{\mathrm{KL}}(q \parallel p)|$ shown in red lines is lower than either $D_{\mathrm{KL}}(p \parallel q)$ or $D_{\mathrm{KL}}(q \parallel p)$, indicating the asymmetry of KL-divergence is not impairing our analyses above.

4.6 Discussion

Our results clearly indicate that GPS and Bluetooth-based contact tracking yield disparate results for the same cohort under measurement. The ground truth contact network, while unknown, was the same for each dataset—it was the same set of participants carrying a single phone measuring both quantities. Both Bluetooth- and GPS-derived contact measurements are estimates of the underlying contact pattern, admitting false positives (for example, Bluetooth contacts through a wall) and negatives (for example, a missed GPS contact because it occurred in an area of poor satellite reception). GPS-based contact tracking identifies fewer shorter contacts, leading to a significant decrease in expected outbreak intensity and number of outbreaks, potentially because both participants need to have a sufficiently good location fix to estimate collocation. The denser contacts reported by Bluetooth-based contact tracking lead to a higher probability of an outbreak and larger outbreaks, resulting in significantly different attack rates for most datasets and diseases. While there were conditions under which no significant differences were observed across the data collection modalities (particularly SHED7 BT8-GPS8), differences were significant often enough to encourage caution in uptake and interpretation of these sensed contact networks. GPS8 tends to underestimate attack rate relative to the others (BT8, BT20, GPS20), indicating the general inability of GPS-collocating to capture proximate contacts within a short distance. Sensing configurations tend to estimate similar attack rates for infectious diseases without a comparatively high R_0 in a more distant underlying population, except for SHED9-measles. Our study cannot conclusively determine if the higher outbreak frequency and size in Bluetooth derived networks is due to false positives in Bluetooth, or false negatives in GPS, but based on the precision of commodity GPS receivers, and their propensity to lose signal in large buildings, we suspect that the observed disparities are predominantly driven by GPS false negatives. If this suspicion is warranted, GPS location-based proximity measurement should be employed in epidemiological simulations with caution, and in a fashion that anticipates and accounts for the fact that the data collection modality employed may be systemically underestimating contact. This is particularly true for the short contacts outside of normal contact networks that drive mixing.

The significance results were relatively insensitive to differences in simulated disease impacting differences in GPS and BT, but the data collection modality induced fewer differences in the results for less contagious diseases such as seasonal flu than for more contagious diseases like measles. It is possible that weakly contagious diseases might not demonstrate differences, as outbreaks would be rare and limited in both GPS and BT networks. These findings hold for both a nominal 8m and 20m threshold for determining if contact has occurred. The thresholds chosen are already judicious, and indicate participants being close enough during a measured portion to have come into close contact during a sensor sleep period, rather than explicitly detecting a close contact. Comparing within sensor outcomes, the contact threshold impacted simulated attack rate for most cases, with the exception of SHED8, which was generally consistent across resolution.

We employed a stylized agent-based SEIR model to determine the attack rate using both Bluetooth- and GPS-inferred temporal contact patterns. The stylized nature of the simulation implies that the results should

be generally correct, but that more detailed models may diverge in the magnitude of the differences observed. SHED7, 8, and 9 are interesting datasets due to the multiple sensor modalities, but are also highly biased, being drawn from a university social science participant pool comprised primarily of undergraduate students in the social and physical sciences. GPS or Bluetooth data from other demographics will almost certainly have different contact patterns, leading to different outcomes. At one extreme, institutionalized individuals (for example, in incarceration facilities or care homes) have limited mobility and would be expected to have much more convergent GPS and Bluetooth contact patterns. Perhaps not surprisingly, some of the worst COVID-19 outbreaks happened in these institutional settings. Similarly, we analyzed four relatively contagious diseases, and ignored diseases where a specific type of contact initiates infection, such as sexually transmitted or blood-borne diseases, or where disease propagation is slow or exhibits prolonged latent periods, such as with tuberculosis. Because the definition of contact for such excluded diseases is substantially different from those analyzed here, the difference between GPS and Bluetooth contact patterns may be more or less pronounced. The process we have used to evaluate the differences should generalize to any contagious disease or measured contact pattern and can be used to evaluate the impact of novel contact detection algorithms or other or novel diseases such as COVID-19 variants of concern.

While this study has made several meaningful contributions to the literature, particularly in highlighting divergent attack rates for GPS and BT measurements of the same underlying contact network, it is subject to notable limitations. We employed three datasets drawn from a social sciences participant pool at our institution. These datasets included individuals who were often unknown to each other, and likely produced more diffuse datasets than would have been expected had we used snowball or respondent-driven sampling or other socially connected recruiting techniques. Running a similar analysis on other datasets could provide more broadly generalizable or representative results. However, for reasonable privacy reasons, public datasets containing both GPS and Bluetooth records are not available, requiring additional measurement effort to extend this analysis. We employed an agent-based SEIR model as it provided the most direct link between the data and the simulated diseases. We chose the stylized SEIR model to emphasize the role of evolving contact networks over other disease dynamics. These results could be extended to include more sophisticated disease models and compare those results against compartmental transmission models grounded in aggregate representations of the underlying contact network. The COVID-19 epidemic has driven innovation in contact tracing, and new measurement techniques based on dongles, beacons or badges are now readily available. A similar analysis including these data sources could be valuable. Finally, we constrained our analysis to four canonical contagious diseases with relatively well parameterized behaviors. However, novel diseases will have novel disease parameters. An exploratory simulation study which outlined how diseases might be expected to behave over these contact networks using, for example, a random-walk through parameter space, might be valuable in predicting new variants or existing diseases, or new diseases emerging from animal reservoirs.

4.7 Conclusions

Epidemiological models of disease propagation are an important tool in controlling and containing epidemic outbreaks. These models rely on accurate measurement of key biological and behavioral parameters to ground the simulation results. Quantifying the characteristics dynamic contact networks is a particularly challenging aspect of grounding these simulations. The significant differences in predicted outcomes for contact networks demonstrated here between GPS and Bluetooth-based contact tracking highlight the difficulty of grounding these simulations. Because of the nature of our data, we know that the contact networks being sought via measurement by BT and GPS should have been identical, as they corresponded to the same device held by the same high-adherence individual as they went about their lives. That the resulting contact networks and predicted attack rates were different indicates that these modalities are not interchangable and that caution should be exercised by modelers employing these measures. While BT and GPS data provides more precise measurement than traditional surveys, they are still prone to error and to disparate estimates of underlying network structure and dynamics.

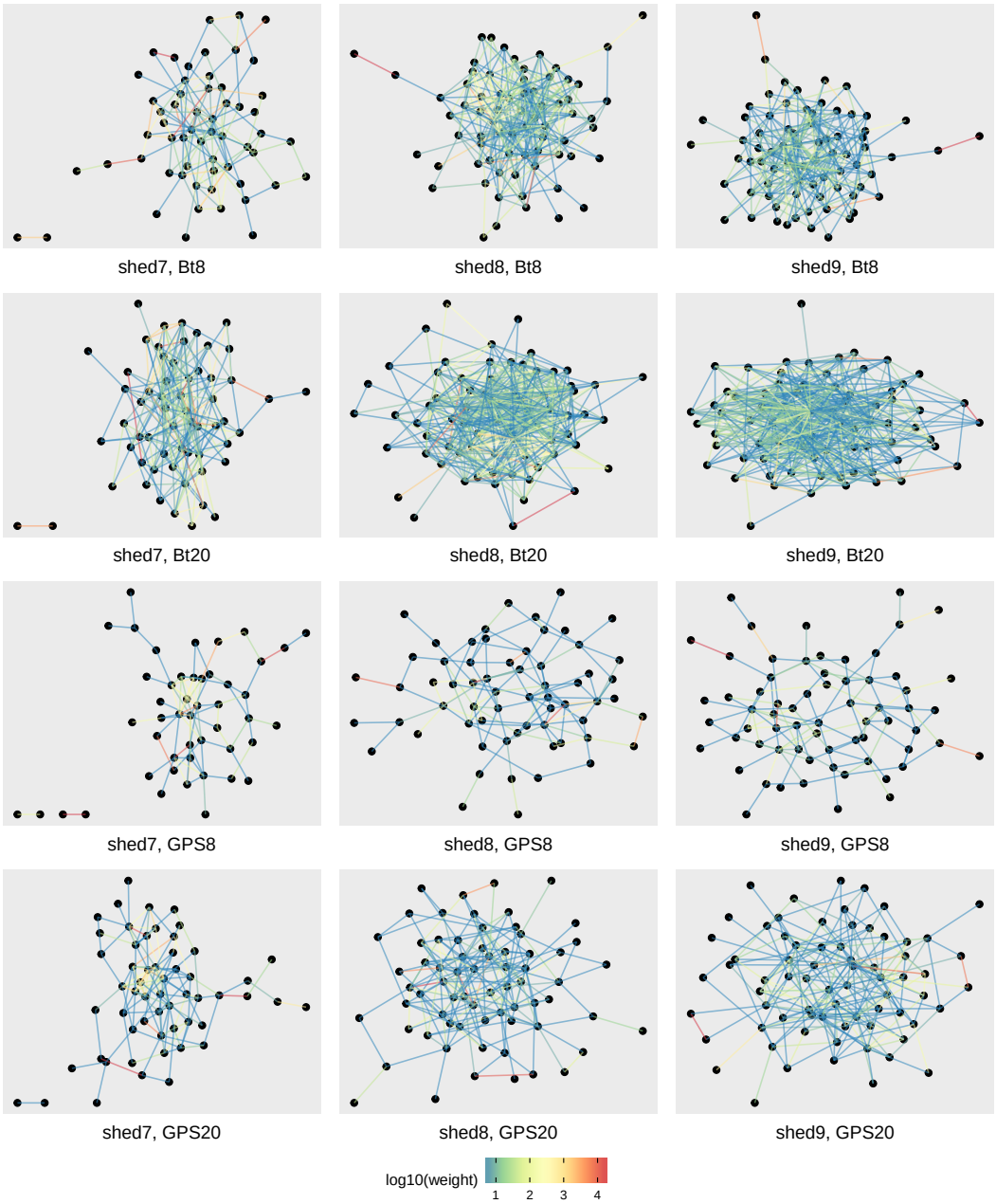


Figure 4.1: Stress-layout of Aggregated Weighted Contact Network by Underlying Population and Data Source, with Edges Colored in Log-Scale by Weights

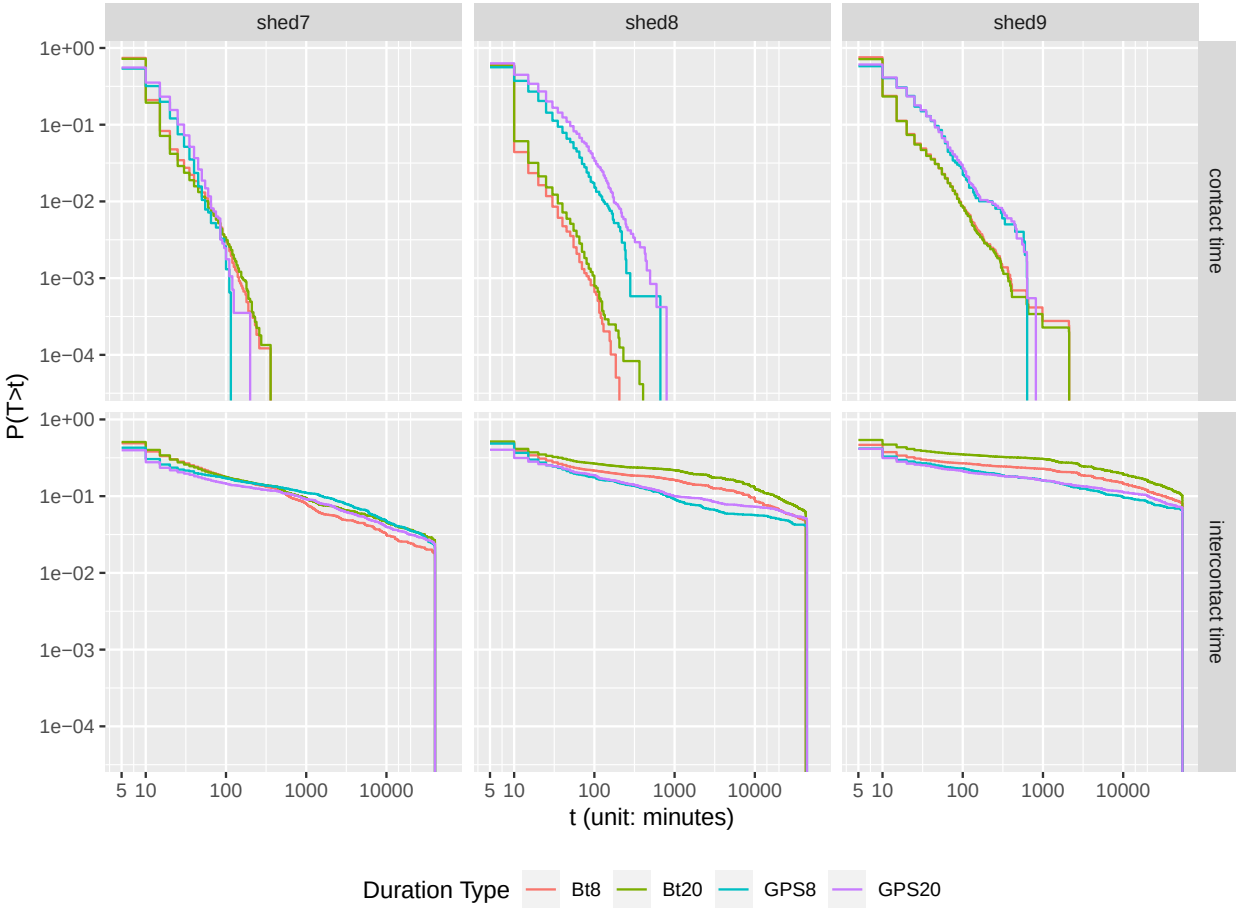


Figure 4.2: Empirical Complementary Cumulative Distribution Function of Contact Duration and Inter-contact Time with Different Sources and Distance Thresholds

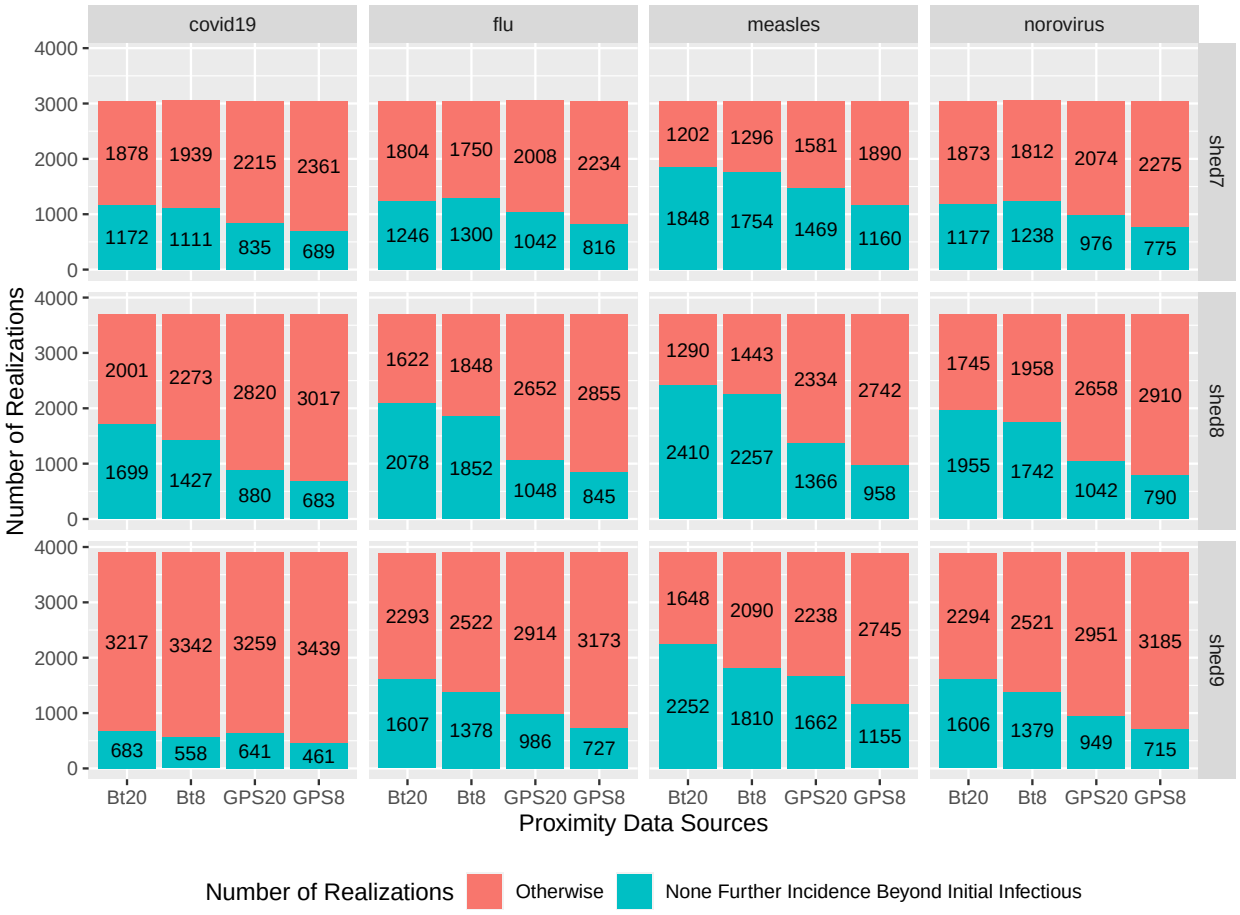


Figure 4.3: Number of Realizations with or without Further Infections Beyond Initial Infectious

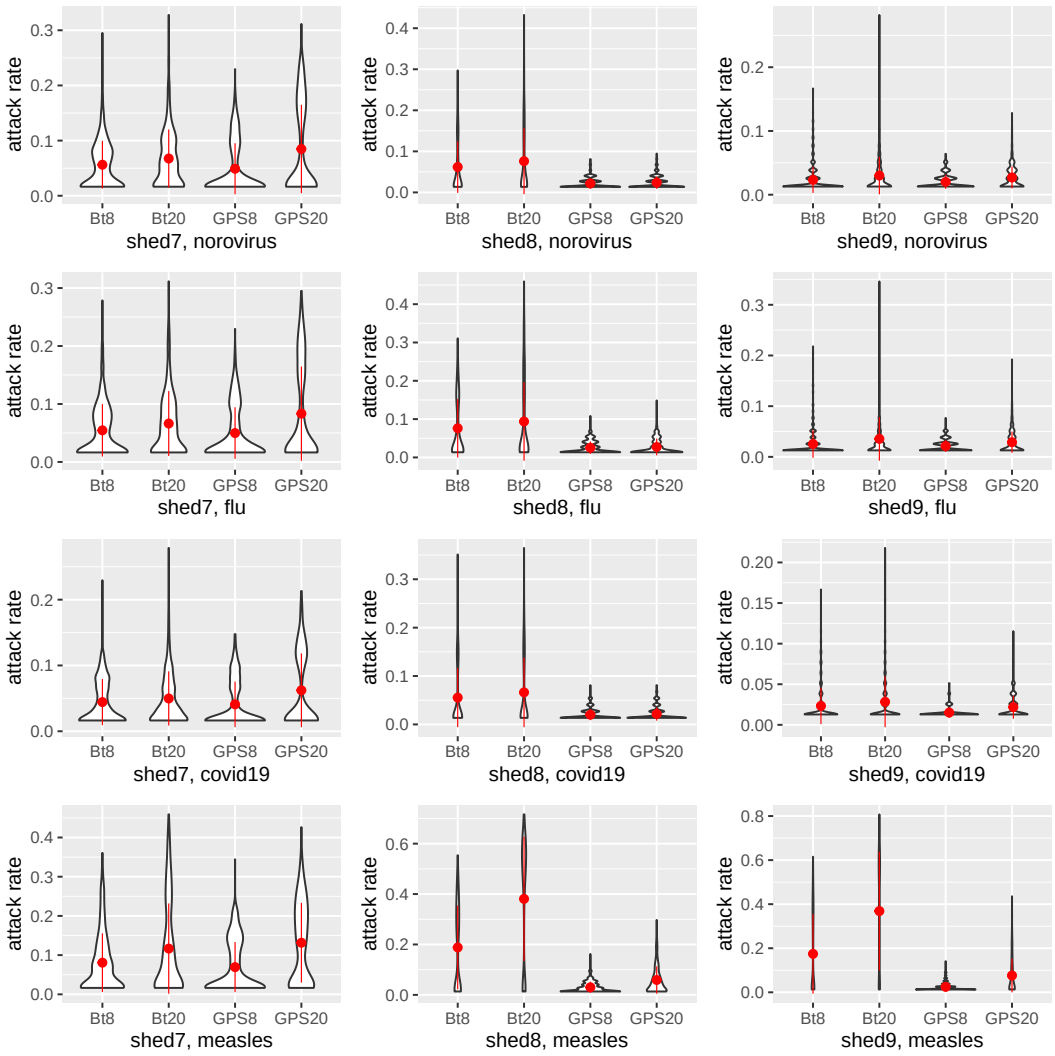


Figure 4.4: Distribution of the Attack Rate (filtered out zero) for Data Collections and Diseases

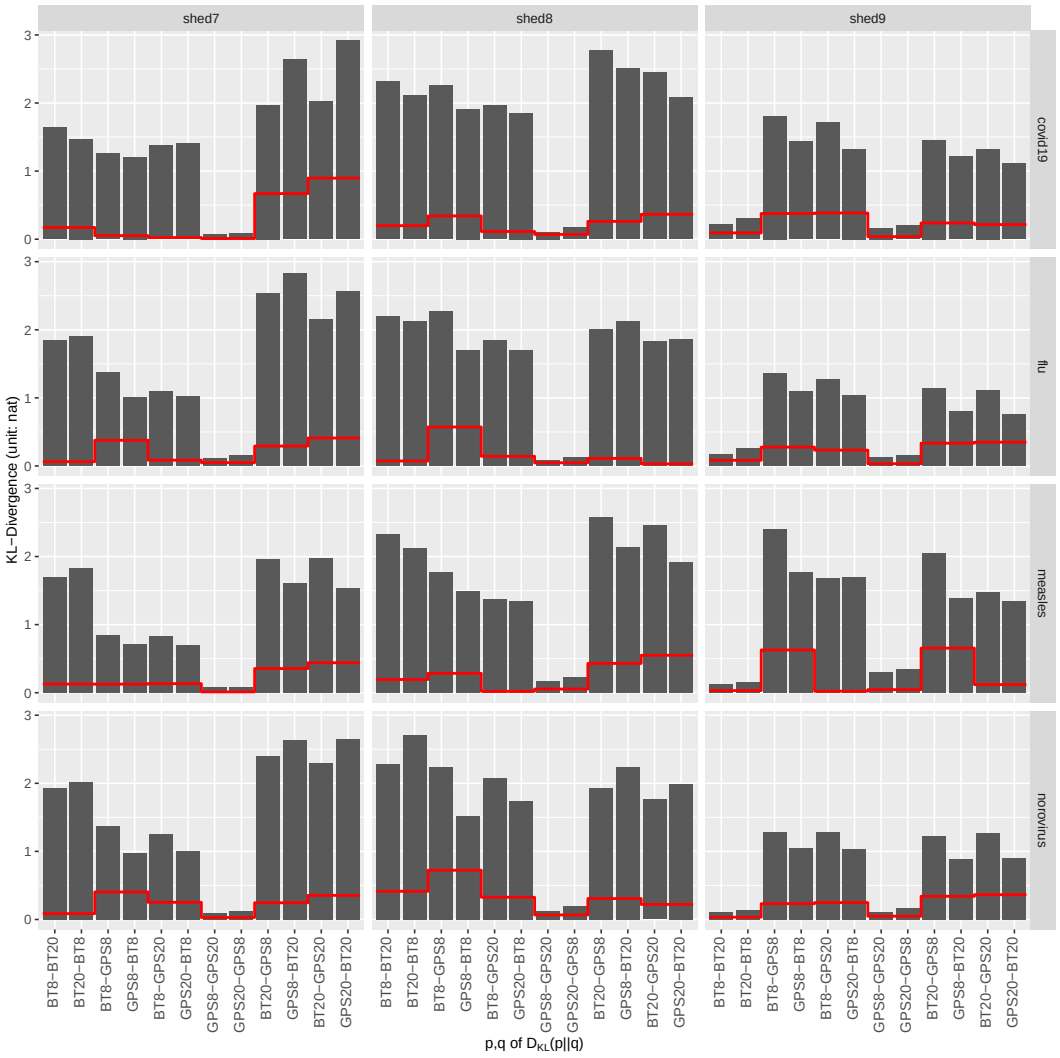


Figure 4.5: Kullback-Leibler Divergence of Individual Infection Risks

5 Impacts of Observation Frequency on Proximity Contact Data and Modeled Transmission Dynamics

Citation: W. Qian, K. G. Stanley, and N. D. Osgood, "Impacts of observation frequency on Reconstruction of Close-proximity Contact Networks and Modeled Transmission Dynamics," *Submitted to the PLOS Computational Biology*, May 2022

Abstract Transmission of many communicable diseases depends on proximity contacts among humans. Modeling the dynamics of proximity contacts can help determine whether an outbreak is likely to trigger an epidemic. While the advent of commodity mobile devices has eased the collection of proximity contact data, battery capacity and associated costs impose tradeoffs between the observation frequency and scanning duration used for contact detection. The choice of observation frequency should depend on the characteristics of a particular pathogen and accompanying disease. We downsampled data from five contact network studies, each measuring participant-participant contact every 5 minutes for durations of four or more weeks. These studies included a total of 284 participants and exhibited different community structures. We found that for epidemiological models employing high-resolution proximity data, both the observation method and observation frequency configured to collect proximity data impact the simulation results. This impact is subject to the population characteristics as well as pathogen infectiousness. By comparing the performance of two observation methods, we found that in most cases, half-hourly Bluetooth discovery for one minute can collect proximity data that allows agent-based transmission models to produce a reasonable estimation of the attack rate, but more frequent Bluetooth discovery is preferred to model individual infection risks or for highly transmissible pathogens. Our findings inform the empirical basis for guidelines to inform data collection that is both efficient and effective.

Relationship to This Thesis The sensing regime of Bluetooth discovery matters when used for gathering proximity contact data because the energy constraints and network bandwidth costs must be balanced against the temporal resolution of the measurement. Increasing the observation frequency of Bluetooth discovery aids in assessing under-reported proximity contacts at the cost of increased battery consumption and network bandwidth required for gathering sensed data. This chapter investigated the impacts of changing temporal resolution by varying the observation frequency under scenarios considering sampling methods, diseases, and underlying populations. Our findings confirmed that temporal resolution of proximity contact data collection impacts outcomes of transmission modeling simulation. In general, higher temporal resolution lessens the

risks of overestimation or underestimation on metrics drawn from simulation results, as judged relative to finding at the baseline frequency (every 5 minutes) of sampling. We have found that downsampling by simply increasing the periodic duty cycle interval is practically straightforward and reliable. We found that reducing temporal resolution can distort the structure of reconstructed contact networks, and higher temporal resolution diminishes this distortion. Findings from this part of the thesis suggest that a tailored sensing regime can improve collections of proximity contact data to reduce battery footprint while ensuring data quality on derived proximate contacts for transmission model simulation. The tailoring of a sensing regime depends on the combination of disease/pathogen and the underlying population.

5.1 Introduction

Despite a century of advances, the burden of contagious diseases remains troublingly high. In the context of growing rates of drug resistance and virus mutations, development patterns which elevate human contact with vectors and animal disease reservoirs, and the capacity of infections to be disseminated via historically growing rates of global travel, the potential burden of infectious disease is historically high. From the shocking worldwide death toll from SARS-CoV-2 [408–411], to Middle East respiratory syndrome coronavirus (MERS-CoV), to Ebola in central Africa [412, 413], to the burden of endemic tuberculosis worldwide and in indigenous communities [414, 415], to the lost productivity due to seasonal flu [416–418] and the common cold [419, 420], and the resurgent patterns of childhood communicable diseases [421–423], contagious disease continues to impose a heavy adverse impact on society. This impact has driven substantial and ongoing research into the transmission, population spread, treatment and prevention of common viral and bacterial pathogens [421, 424, 425]. For the past century, dynamic models of communicable diseases have served as a key tool in the understanding, prevention and control of communicable disease. A central element of such models is a representation of contact patterns between hosts, transmission, and the natural history of infection within a host [128, 426].

Close-proximity human contact networks constitute a key mechanism in the spread of communicable diseases [61, 427, 428]. Together with pathogen-specific parameters, high-fidelity representations of such contact networks within transmission models [61] can enable a much higher resolution view of the process of a disease spreading than is possible with the random mixing assumptions required in compartmental models within the traditional susceptible-infectious-recovered (SIR) family [128, 426, 429]. Such a view can support real-time identification early of outbreaks and an estimation of the attack rate, as well as retrospective evaluation and assessment of improved effectiveness of altered vaccine schedules, aid in planning of interventions such as outbreak response immunization [21], public health orders and quarantine, and support assessment of the impact of the scope, speed, and breadth of contact tracing [430]. Transmission models structured with a detailed contact network aid inferencing of population-scale effects from individual-level behavior of infections by enabling characterization of the transmission of contagious diseases over the close-proximity contacts shaping outbreak dynamics [61, 368].

The ubiquity of smartphones with their rich complement of sensors, and emergence of wearable proximity-detection device have enriched data collection systems [213, 257, 293, 396, 431, 432]. Automatic contact tracing apps using Bluetooth low-energy [391] have allowed researchers to collect contact information whose self-reporting would be burdensome [68, 433], and likely infeasible due to limited awareness of contacts [434]. As envisioned by some observers [61], the growing availability of proximity contact data in high-resolution has further encouraged analytics taking empirical data of proximate contacts into transmission modeling [109, 119, 134, 293, 435]. Salathé *et al.* [293] pioneered collecting high-resolution proximity contact data with mote sensors, and taking such high-resolution data into a transmission model to analyze influenza outbreaks.

Despite the increasing scale of computing power in the form of expanding storage capacity and accessible high-performance computing, we still struggle to collect, store, and process individual-level contact data sufficient to parameterize a longitudinal transmission model with even a municipal-scale population. When configuring smartphones to collect proximity contact data, a sensing regime with sampling frequencies on the scale of minutes notably elevates power consumption, risking adverse impacts on study recruitment and adherence. Such impacts are of particular concern among low-socioeconomic status populations who are subject to elevated risks of communicable disease transmission due to crowding and other risk factors [436–438].

In light of such technology constraints, past contributions [213, 428, 439] have argued that a clear understanding of the sensing regime is required—a sensing regime schedules short periods to turn sensors on for scanning throughout an experiment. The proximity contact data in our study are derived from Bluetooth discovery records, and the Bluetooth discovery is performed at the first minute of each duty cycle, where duty cycles are consecutive periods of identical length. The reciprocal of the duty cycle interval is referred to as the observation frequency, and the observation frequency is in inverse relationship with the inter-observation interval. This study investigated how varying sensing regimes affects captured proximity contact data and impacts the results of an empirical contact empowered transmission model (ECTM). Specifically, we sought to investigate the following three questions:

- How does the structure of the inferred contact network skew as the observation frequency of Bluetooth discovery reduces?
- How do the results of a transmission model when taking proximity contact data collected at a reduced observation frequency deviate from taking proximity contact data collected at a baseline frequency (the highest frequency among our scenarios)?
- Under which disease/pathogen and community structure contexts may observation frequency be reduced, and to what extent, without undermining confidence in conclusions?

We addressed these questions by analyzing proximity contacts derived from downsampled contact data collected from participant smartphones in five high-resolution human contact network studies. Each study has an effective duration of four or more weeks, and includes at least 30 participants, yielding a total of 284 participants across all studies. Close-proximity contact data were collected approximately every 5 minutes by smartphone-based Bluetooth handshakes. We analyzed how network structure changed as observation frequency is reduced.

To study the impact of downsampling on the model-estimated attack rate and individual infection risks, we provided downsampled contact data to an SEIR agent-based simulation model for 12 different transmissible diseases/pathogens. Using findings at the baseline resolution (involving sampling every 5 minutes) as the reference, we found that the bias-variability of the attack rate shifted as observation frequency was reduced. Our findings further demonstrate that in terms of both variability and bias, the magnitude of the impact of

reducing observation frequency is both disease and community specific. Specifically, for diseases with low basic reproduction number, such as Middle East Respiratory Syndrome (MERS), simulation results with respect to both attack rate and individual infection risk were relatively insensitive to observation frequency. On the other hand, pathogens such as *Bordetella pertussis* showed a marked dependence on sampling frequency. Maintaining a higher observation frequency notably turns to be more important in denser communities. Finally, we found that individual infection risk varied according to which edges of contact network served as parts of transmission chains within a given simulation.

5.2 Data Sources

This study drew contact data from five high-velocity micro-contact data sets each with a month or longer duration, employing the Saskatchewan Human Ethology Datasets (SHED) 1, 2, 7, 8, and 9 [210, 257, 395]. These SHED data sets employed the iEpi system and its successor Ethica Data [257] to collect longitudinal data via smartphone-based sensors, including with respect to the battery level, charging state, Bluetooth, Wi-Fi, GPS, accelerometer, magnetometer, in addition to pre- and post-surveys. Only the Bluetooth discovering records and battery level records were used in this research. It is important to emphasize that the SHED datasets, though sharing high acquisition velocity and a duration of a month or greater, exhibit notable heterogeneity in the characteristics of the participant population and—by extension—the network structures. SHED1 and SHED2, represent "closer" communities, composed of graduate students and staff from the Department of Computer Science from University of Saskatchewan, with SHED1 having the majority of its participants coming from two research laboratories. In contrast, SHED7, SHED8, and SHED9 recruited undergraduate students from across the University of Saskatchewan through a social sciences study pool, representing a more diverse and diffuse community.

All SHED studies' participants were volunteers. No experimental manipulations were conducted during data collection. The studies did not undertake stratified sampling as to ethnicity, grade, or gender. The study did not proscribe participation by those connected with the department or research laboratories involved, and the study team informed colleagues in labs and the Department of Computer Science first. Awareness of the potential study involvement can be assumed to have spread across social networks. For SHED1 and SHED2, participants were provided with a pre-configured Android phone that they carried in conjunction with any other personal mobile device. By contrast, participants used their own phones for SHED7, SHED8, and SHED9. Although for these three studies, both Android and iPhone users were welcome, because Bluetooth beaconing did not work reliably on iPhone due to security settings, iPhone users were removed from the analysis and all participants reported here were Android users.

5.2.1 Contact Data Collection Method

Data collection for Bluetooth contacts and battery levels on both iEpi and Ethica Data apps equipped smartphones occurs within discontinuous epochs. Study periods (consecutive days spanning at least one month) were divided into 5-minute (exactly for SHED1 and SHED2, and approximate intervals for SHED7, 8, 9) duty cycles. Within each duty cycle, battery levels were recorded as long as the apps were running, and Bluetooth scan was enabled during the first minute of each duty cycle. Phones were discoverable while scanning for nearby discoverable devices.

5.3 Methods

We synthesized collections of proximity contact data with varying sensing regimes by downsampling from a baseline. The impact of varying sensing regimes are measured on two types of findings: those regarding network structure, and those involving population-wide disease spread. For the network analyses, we compared network structure with successive levels of downsampling and interpreted the results in terms of classical network models [207, 225, 406]. For the simulation analyses, we used an individual-level Susceptible-Exposed-Infectious-Recovered (SEIR) model [135], with reconstructed contact networks using 12 distinct common communicable diseases/pathogens (flu, SARS, fifth, pertussis, measles, chickenpox, MERS, diphtheria, COVID-19, COVID-19 Alpha variant, COVID-19 Beta variant, COVID-19 Delta variant). We investigated how downsampling (decrements in observation frequency) impacts findings regarding the attack rates, individual infection risks, and outbreak timing from simulation outputs, by employing two distinct downsampling methods named `Snapshot` and `Upperbound`. For every combination of choices from downsampling methods, sampling rates, communicable diseases, and studies, the contact network for that study induced by that downsampling rate was derived and analyzed, and simulations conducted using those networks were analyzed.

5.3.1 Downsampling Approach

We assume that the behavior of close-proximity contacts is time-varying and denoted by an undirected graph $G_t = (V_t, E_t)$, with vertices representing participants and edges denoting pairs of participants that exhibit close-proximity contact at time t. We assume that, given a sufficiently small temporal quantum ξ_0 (for example, one second), the state of our close-proximity contacts can be considered constant across each such time quantum without significant loss of precision, meaning our analysis only considers dynamics over a unidimensional lattice with spacing ξ_0. This leads to proximity contacts evolving over time as a series of undirected graphs $G_{t_0},\ t_0 \in \xi_0\mathbb{N} \subset \mathbb{R}$, where t_0 projects the discrete-time index onto a real-world clock. We denote proximity contacts among participants V at time t as $G_t,\ t \in \mathbb{N}$. Downsampling according to a heuristic is essentially aggregating $\{G_t\},\ t \in [t_i, t_j)$, which can be considered as a coding problem [440, 441].

Because the baseline frequency of longitudinal data obtained is approximately every 5 minutes, the original

sampling of close-proximity contact network is a series of $\{G_t\},\ t \in [0, T)$, where t has the unit of minute and T is the effective length of a study in minutes. After post-processing, t represents an integer index representing the minute associated with the observation, where minute 0 corresponds to the first minute of the first day of the study. For convenience, we rephrase the sample time as a period rather than a specific point, $\{G_{t_i}\},\ t_i \in [\xi i, \xi i + \xi),\ i = 0, 1, 2, \cdots$, where $\xi = 5$ is the (expected) duty cycle interval and $1/\xi$ is the observation frequency for our original data, and is referred to as the baseline frequency.

A further consideration relates to data availability. Such availability is affected by many factors, including—but not limited to—participants opting to "snooze" the sensor data recording during a private period, cases where the operating system temporarily evicts the data collection app from memory due to resource shortages, or—especially for the case of SHED7-8—due to misaligned duty cycles reflecting system scheduling. After aggregation, each sample $G_t = (V_{t_i}, E_{t_i})$ is an unweighted unidirected simple graph which can be represented as a $(0, 1)$-adjacency matrix. This adjacency matrix is symmetrical and each of its element $a_{ij} \in \{0, 1\}$ indicates individual v_i and v_j have a contact ($a_{ij} = 1$) or no contact ($a_{ij} = 0$).

We considered two downsampling strategies: A physically realizable sampling strategy (named `Snapshot`), and a theoretical upper-bound (named `Upperbound`). Snapshot periodically samples a snapshot of the current contacts in place at that time, thereby providing a simulated answer to the question "what if we sampled less frequently?". The `Upperbound` downsampling strategy instead records all contacts throughout the downsampling interval, and reports those as applying at the sampling time. It instead answers the question "What would be the impact of these same contacts, if they were to change less frequently?" `Upperbound` provides an oracle which maintains all contacts during the period regardless of whether the downsampled schedule would have measured them.

Snapshot

The `Snapshot` downsampling method is conceptually straightforward: For each downsampling period $[\xi' i, \xi' i + \xi'),\ i = 0, 1, 2, \cdots$, we choose the first available sample index $G_{\tilde{t}_i},\ \tilde{t}_i \in [\xi' i, \xi' i + \xi')$. This results in subsampling $\left\{G_{\tilde{t}_i}\right\},\ \tilde{t}_i \in [\xi' i, \xi' i + \xi')$. If a contact occurred during the specific duty cycle captured by that index, it will be reflected within the sampled record. `Snapshot` simulates the effect of selecting a longer duty cycle for measurement, including the loss of contacts due to undersampling.

Upperbound

In contrast to `Snapshot`, we sought to investigate the impact of a theoretical downsampling method, which could provide a sample summary that included information drawn from throughout that interval. Specifically, we considered the union $G_{\check{t}_i}$ for $\{G_{t_i}\},\ t_i \in [\xi i, \xi i + \xi),\ i = 0, 1, 2, \cdots$, where the union, in general for any discrete set $j \in \mathbb{N}$, is defined as $\bigcup_{j \in J} G_j = \bigcup_{j \in J} (V_j, E_j) = \left(\bigcup_{j \in J} V_j, \bigcup_{j \in J} E_j\right)$. This downsampling mechanism serves to conserve all pairwise contacts which are observed at any time during a downsampling interval. `Upperbound` cannot practically be deployed in data collection using the most common sensors used

for proximity detection, but could be used during post-processing to reduce the number of time steps realized during ABM-based analyses, increasing simulation speed. As ξ approaches the study period, the `Upperbound` downsampling results in a more homogeneously weighted random mixing graph of contacts, resembling compartmental models with less heterogeneous preferential mixing among compartments. `Upperbound` maintains the density of the contact graph during downsampling.

While the investigation of the effects of `Upperbound` was motivated predominantly by its theoretical properties, it bears noting that some technologies—such as privacy-preserving or battery-sensitive contact tracking and reporting systems—do perform similar temporal aggregation of contact information over a period of time [442]. `Snapshot` performs temporal quantization in a sampling context. `Upperbound` performs both temporal quantization and aggregation via accumulation across that interval.

5.3.2 Network Structure Analyses

Past research suggests that contact networks are types of small-world networks [225], and may be associated with scale-free properties [406]. The scale-free property is characterized by a power-law decay in degree distribution. Amaral *et al*. further classified small-world networks into three classes, and advanced the hypotheses that faster decay in the degree distribution with a rising degree is due to aging vertices and constraints on adding new links to highly connected vertices [207]. We measured cumulative contact time as a measure proportional to contact time $\bar{c}$, which specifies the average time that an individual is in close-proximity contact with another, and has been used in epidemiological models [357, 443].

5.3.3 SEIR Simulation

We built an agent-based SEIR transmission model (`SEIR-ABM`) with the proximate contacts derived from synthetic proximity contact data collected with different observation frequencies. These synthetic proximity contact data are generated by downsampling with both `Snapshot` and `Upperbound` methods across datasets and diseases/pathogens.

Agent-Based SEIR model

The agent-based model treats each person in the study population as an actor with one of four possible states with respect to a natural history of infection: *Susceptible*, latently infected (*Exposed*), *Infective*, and in a *Removed* state conferring persistent immunity to future infection. At any one time quantum, a given agent is further parameterized by a vector of active contacts, as specified by the proximity contact data for that agent for the current study, at the current level and type of aggregation.

Our SEIR agent-based model (`SEIR-ABM`) takes proximity contact data $\mathcal{D} = \left(\{G_{t_i}\}, \xi \right)$, $t_i \in [\xi i, \xi i + \xi)$, $i = 0, 1, \cdots, n-1$, an initial infected agent $\mathcal{V} \in \bigcup_i V_{t_i}$, $i = 0, 1, \cdots, n-1$, and a disease $\mathcal{M}$ from the set {flu, SARS, fifth, pertussis, measles, chickenpox, MERS, diphtheria, COVID-19, COVID-19 Alpha variant, COVID-19 Beta variant, COVID-19 Delta variant}. The proximity contact observations were repeated four

times, ensuring at least four months of proximity contacts time series for transmission simulation, to avoid underestimation of attack rate induced by right-censored data—particularly considering diseases/pathogens whose *Exposed* and *Infectious* periods add up to more than twenty days.

For each disease, we gathered the base reproductive number R_0, and range estimates of the latent period and infectious periods. Each agent in the `SEIR-ABM` was associated with a latent period and personal infectious period drawn uniformly from corresponding ranges. Although in practice R_0 varies along with the rate of human-human or human-vector interactions spatially and temporally [138], we assumed identical R_0 for scenarios with different SHED datasets, because participants spend a considerable amount of their time on campus. This paper focuses on analyzing the impact of temporal resolution of Bluetooth discovering sensed proximate contact. Even if the R_0 is not calibrated separately for the specific population represented in each SHED dataset, it will not block us from interpreting how R_0 changes with the temporal resolution.

Table 5.1: Disease Parameter Table

	Basic Reproduction Number	Incubation Period	Infectious Period
chickenpox	†15 [444]	10 – 12 [403]	⋆8 – 11 [445]
COVID-19 Wild Type	†2.5 [144]	5.6 – 7.7 [144]	3 – 7 [144]
COVID-19 Alpha Variant	⋆3.23 [446]	5.6 – 7.7 [144]	3 – 7 [144]
COVID-19 Beta Variant	⋆3.13 [446]	5.6 – 7.7 [144]	3 – 7 [144]
COVID-19 Delta Variant	⋆4.93 [446]	5.6 – 7.7 [144]	3 – 7 [144]
diphtheria	†6.5 [402]	‡2 – 5 [403]	‡14 – 28 [447]
fifth	1.8 [448]	⋆6 – 11 [449]	⋆4 – 9 [449]
flu	‡1.31 [199]	2.28 – 3.12 [199]	‡2.06 – 4.69 [199]
measles	†15 [402]	⋆5 – 10 [450]	⋆4 – 6 [451]
MERS	0.69 [452]	2 – 14 [453]	⋆1 – 5 [454]
pertussis	†14.5 [402]	7 – 10 [403]	‡14 – 21 [403, 455]
SARS	3.6 [456–458]	2 – 10 [457]	4 – 14 [459]

†Derived as midpoint of reported range.
‡Derived range from different reports.
⋆Derived from starting range plus average duration.
⋆Derived from other disease or comparative estimations.

A simplified model was employed because we were primarily interested in the impact of measurement frequency. That model supports a stylized notion of the characteristics of the diseases explored, under a variety of epidemiological contexts:

- **Closed-population** Despite the fact that some for the 12 communicable diseases examined here are potentially lethal, we assume a closed population with no mortality or care-seeking that would cause an infected individual to be removed from circulation prior to recovery.

- **No intervention** Occurrence of infection within an individual or public health messaging regarding an identified outbreak can lead to the adoption of personal protective behavior such as elevated hygienic adherence and social distancing by population members; outbreaks can also lead to triggering of public health interventions, such as outbreak response immunization campaigns, quarantine efforts, contact tracing or increased vaccination. Within our simulation, we assume that infection status does not change agent behavior.

- **Consistent stages of infection** While the different communicable diseases considered in this paper differ considerably in the features of their natural history of infection (*e.g.*, the presence of both symptomatic and alternative oligo-/asymptomatic pathways, lack of permanent immunity) and routes of transmission (*e.g.*, airborne, droplet, fecal-oral), to focus on the effects of temporal quantization, we treated them as all being characterized by a 4-stage natural history of infection and proximity-based transmission, and as differing merely in terms of a disease-specific residence time within each state. This structure proceeds from *Susceptible* to *Exposed*, *Infectious*, and *Removed* states. In light of the 4-month time horizon of the model, we assumed that no re-infection is possible for each of the 12 communicable diseases.

- **Homogeneous infectious rate** We assume that for every discordant pair of individuals engaged in contact, the probability that the pathogen will be transmitted is governed by a constant hazard rate and the duration of the contacting period. This hazard rate is determined by a rate of potentially infecting exposures β (for example, sneezing, aerosol production or hand-shaking), and a transmission probability per such exposure.

A System Dynamics/compartmental SEIR model (`SEIR-SD`) typically has $R_0 = \lambda \cdot \gamma^{-1} = \beta \cdot \bar{c} \cdot \gamma^{-1}$, where λ is the force of infection, β is the probability of transmission per contact between a susceptible and an infective, $\bar{c}$ is the average number of contacts made by each susceptible per unit time, and γ is the rate at which an infectious person recovers or otherwise transitions to the *Removed* state. In the `SEIR-ABM`, because of the no intervention assumption, we estimate $\bar{c} = \dfrac{1}{T\|V\|} \sum_{i,j,k} \mathbb{I}(e_{jk} \in E_{t_i})$ given observed temporal graphs $\{G_{t_i} = (V_{t_i}, E_{t_i})\}$, $t_i \in [\xi i, \xi i + \xi)$, $i = 0, 1, \cdots, n-1$, where $T = \xi n$ is the effective study period, $\mathbb{I}$ is the indicator function, $\|V\|$ is the number of participants whose contact networks are recorded and $\|V\|$ does not change within the model because of the closed-population assumption.

All agents in the SEIR-ABM start as susceptible, with the exception of one initial infective. To address the potential impact on an outbreak outcome of the index infective individual, we iterate the initially infected person over the entire population. For each initial infection setting, we simulated the model across 30 distinct realizations, each associated with a distinct random number seed. At a high level, the algorithm of our SEIR-ABM can be summarized as follows:

Algorithm 1: Outline of SEIR-ABM

input : contact data $\mathcal{D}$,
disease $\mathcal{M}$,
initially infected person $\mathcal{V}$

output : list of infectious events $\mathcal{R}$

$(\mathcal{G}, \xi) \leftarrow \mathcal{D}$;
it $\leftarrow$ `iterator(`$\mathcal{G}$`)`;
Vs $\leftarrow$ `init_population`;
`set_health_state(`$\mathcal{V}$`,` Infectious`)`;
$t \leftarrow 0$;
G $\leftarrow$ `next(`it`)`;
while $t < T$ **do**
  `map(update_health_state,` Vs`)`;
  `map(`λv. `append(`$\mathcal{R}$`,` `expose_all_connected(`v`,` G`))`, `filter(`λv. `at_health_state(`v`,` Infectious`)`, Vs`))`;
  if `get_timestamp(`G`)` $+ \xi < t$ **then**
    G $\leftarrow$ `next(`it`)`;
  end
  $t \leftarrow$ `tick_tock(`t`)`;
end

Parameter Variation Grid

For each of the two downsampling methods, our simulation considers scenarios involving all combinations over three parameter classes: underlying populations, diseases/pathogens, and downsampling intervals (mimicking observation frequencies). This paper specifically investigates the impact of the downsampling method and downsampling intervals given a specific underlying study population and pathogen. We consider all combinations of the following:

- Two downsampling methods: `Snapshot` and `Upperbound`
- Five datasets (SHED1-2, SHED7-9) with populations $\{39, 32, 61, 74, 78\}$, considering each possible

exogenously infected index within each population

- Twelve pathogens and their accompanying communicable diseases: influenza type A, SARS-CoV, parvovirus B19, *Bordetella pertussis*, *Measles morbillivirus* (MeV), varicella-zoster virus (VZV), MERS-CoV, *Corynebacterium diphtheriae*, SARS-CoV-2, SARS-CoV-2 (B.1.1.7), SARS-CoV-2 (B.351), SARS-CoV-2 (B.1.617.2)
- Seven sampling intervals: 5 minutes (baseline), and 6 downsampling intervals: 10, 30, 60, 90, 180, 360 minutes
- An ensemble of 30 Monte Carlo realizations per parameterization

Considering all combinations of the above, we planned 1312080 realizations of the `SEIR-ABM` model.

Realizations were evaluated on a server with an Intel® Xeon® CPU E5-2690 v2 and 503GB memory. Models were created in AnyLogic 8.1.0 and exported to a standalone Java application with OpenJDK 1.8.0_252, resulting in 85GB of output data.

5.3.4 Impact Metrics

The transmission dynamics that emerged from an SEIR agent-based model have various usages, typified by evaluating interventions [460, 461], understanding transmission paths [462, 463], estimating disease parameters [464, 465], and forwarning outbreaks [466, 467]. These usages often have their basis on model simulated results, such as attack rates, transmission pathways in contact networks, and individual infection risks. We employed corresponding metrics to summarize changes in these simulation results across the parameter variation grid, bearing variations due to stochastics of Monte Carlo realizations and rotations of the index infective within each population.

Cumulative Cases The cumulative cases of a realization are the number of endogenous infections throughout that realization, starting from an infectious due to exogenous infection (not counted) until no one at the states of *Exposed* or *Infectious*. Because of assumptions as to the closed-population and acyclic stages of infection and persistent immunity, the cumulative cases are capped at one less than the size of underlying population. Without imposing assumptions on the distribution of cumulative cases over thirty Monte Carlo realizations, we employed median and inter-quantile range (IQR) as statistics to summarize cumulative cases by groups. In an agent-based model, the results of disease spread can be strongly influenced by the contact network of the initially infected individual. We explored two approaches to grouping the cumulative cases by constructing blocks with/without the index infectives.

Attack Rates The attack rate of a realization is the ratio of cumulative cases to the size of its underlying population. The attack rate reflects the proportion of people who become infected started with an exogenously infected index in an otherwise susceptible population under our assumptions. Considering the attack rate as

corresponding cumulative cases normalized by the size of its underlying population, we can compare attack rates among different underlying populations for a given disease/pathogen but with different downsampling methods and observation frequencies. These comparisons may shed light on whether an underlying population will alter the importance of the temporal resolution to transmission simulation results of our interests.

Transmission Pathways in Contact Networks Studies on transmission pathways in contact networks investigate how a disease/pathogen may spread on routes of transmission available between infectious and susceptible individuals, given the structure of contact networks where they reside [468, 469]. Sensor-data-derived proximate contacts reveal contact networks to study transmission pathways for diseases/pathogens relying on routes of aerosol transmission and potentially direct contact transmission [257, 293, 302]. An infection pair of a realization, denoted by an ordered tuple of $(V_{\text{susceptible}}, V_{\text{infectious}})$, states the infection of $V_{\text{susceptible}}$ by $V_{\text{infectious}}$ during the infectious period of $V_{\text{infectious}}$ in the realization. Because infection pairs are elemental results reflecting transmission pathways from a realization, we sought to measure impacts of temporal resolution on transmission pathways in contact networks by comparing statistics of infections pairs from corresponding realizations.

Given a set of realizations from the `ABM-SEIR` model with a size $\|V\|$ population, if we put all possible tuples of individuals $\mathcal{T} = \{(i,j) \mid i,j \in V \wedge i \neq j\}$ into a canonical form with a rule $(i,j) \prec (k,l) \iff i \prec j \vee (i = k \wedge j \prec l)$, then we can express the frequencies of infection pairs in the set of realizations as a vector $\mathbf{\Omega} \in {\mathbb{N}_0}^{\|\mathcal{T}\| \times 1}$, where $\|\mathcal{T}\| = \|V\|^2 - \|V\|$. Assuming $\mathbf{\Omega} \neq 0$ and an uniform prior of an individual becoming the exgenously infected index, the L_1-normalized vector of infectious pairs' frequencies, denoted by $\frac{\mathbf{\Omega}}{\|\mathbf{\Omega}\|_1}$, is the relative risk of infection pairs occur in a realization.

Now we consider two parameter sets $\mathcal{P}_p$ and $\mathcal{P}_q$ sharing a disease/pathogen, an underlying population, and a sampling method, but with different duty cycle intervals ξ_p and ξ_q. The differences of realizations resulted by $\mathcal{P}_q$ from $\mathcal{P}_p$, in terms of frequencies of infection pairs, can be reflected by a weighted-Minkowski distance of order one, denoted by $D_{\text{M}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right) = \left(\frac{\mathbf{\Omega}_p}{\|\mathbf{\Omega}_p\|_1}\right)^{\intercal} \cdot \left(\mathbf{\Omega}_p - \mathbf{\Omega}_q\right)^{\text{abs}}$, where $\frac{\mathbf{\Omega}}{\|\mathbf{\Omega}\|_1}$ is the weight and $(\cdot)^{\text{abs}}$ is the element-wise absolute value operator for a vector, such that $\mathbf{\Omega}^{\text{abs}} = \begin{bmatrix}\text{abs}(\Omega_1) & \cdots & \text{abs}(\Omega_{\|\mathcal{T}\|})\end{bmatrix}^{\intercal}$. When $\xi_p < \xi_q$, $\mathcal{P}_q$ is a parameter set with a larger duty cycle interval than $\mathcal{P}_q$, this weighted-Minkowski distance between frequencies of infection pairs $\mathbf{\Omega}_p$ and $\mathbf{\Omega}_q$ can be interpreted as the risk-weighted L_1-distance resulted by a downsampled proximate data of duty cycle interval ξ_q from a reference proximate data of duty cycle interval ξ_p. This expected L_1-distance handles variations due to stochastics of Monte Carlo realizations and rotations of the index infective within each population, allowing us to infer impacts of observation frequencies on simulated results of infection pairs sharing an underlying population. Notice that $0 \leq D_{\text{M}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right) \leq 30\|V\|$, where 30 is due to the number of Monte Carlo realizations per parameterization, and $\|V\|$ is due to the rotation of index infectives. We have $D_{\text{M}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right) = 0$ when $\mathbf{\Omega}_p = \mathbf{\Omega}_q$, and $D_{\text{M}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right) = 30\|V\|$ when $\nexists\, i,j \in V,\ \Omega_p^{(i,j)} > 0 \wedge \Omega_q^{(i,j)} > 0 \wedge i \neq j$. To unify the scale of $D_{\text{M}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right)$ for underlying populations with different sizes, we employed $D_{\text{NM}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right) = \frac{D_{\text{M}}(\mathbf{\Omega}_p, \mathbf{\Omega}_q)}{\|V\|}$ and $0 \leq D_{\text{NM}}\left(\mathbf{\Omega}_p, \mathbf{\Omega}_q\right) \leq 30$.

Individual Infection Risks The infection risk of an individual is estimated by the fraction of realizations where the individual got infected. For a population V, its individual infection risks under a circumstance of a disease, a sampling method, and an duty cycle interval can be presented by a vector of infection risks for everyone in the population, denoted by $\mathbf{\Psi} \in \mathbb{N}_0^{\|V\|}$. The L_1-normalized vector of individual infection risks, denoted by $\rho = \frac{\mathbf{\Psi}}{\|\mathbf{\Psi}\|_1}$ can be considered as the likelihood of an individual to be the most likely infected.

5.4 Results

We evaluated the impact of downsampling methods and frequencies from two perspectives: the resultant distortions of network structure, and deviation in transmission model outcomes. Each such evaluation employed the results of the baseline fidelity network representation as the reference for assessing such distortions/deviations. The network structure analyses show that, as observation frequency reduces, the `Snapshot` method and the `Upperbound` method distort network structure in different ways—the `Snapshot` keeps the average cumulative contact time at the cost of underestimating node degrees. In contrast, the `Upperbound` method results in inflated average cumulative contact time but retains the node degree distribution. The evaluated the deviation in transmission model outcomes at both the population and individual levels are analyzed in the following sections.

5.4.1 Impacts on Population-Level Simulation Results

The impacts of the observation frequency on simulation results from the `ABM-SEIR` model can be considered at the population and/or individual level. Cumulative cases and attack rates were used to measure the impacts of observation frequency on simulation results at the population level—population-level results of a transmission model are often used to evaluate the size of the outbreak or the overall severeness of an upcoming wave. We performed Welch's t-test on cumulative cases with different ξ to draw quantitative conclusions as to the impact of observation frequency on the mean of cumulative cases. We used the Prentice modified Friedman tests on cumulative cases with different ξ to test the impact of observation frequency on the distribution of cumulative cases.

Cumulative Cases

Figures 5.1 and 5.2 show grids of violin plots visualizing the empirical distributions of the cumulative cases in realizations of the agent-based SEIR model taking downsampled contact data at different duty cycle intervals, with one grid for each of the `Snapshot` and the `Upperbound` downsampling methods. Each grid of the violin plots characterizes how cumulative cases varies by diseases (row) and underlying populations (column). Each cell of a grid is a violin plot consists of violins arranged by increasing duty cycle interval, with 5 minutes being the left-most and 360 minutes being the right-most. Each violin in a violin plot illustrates the distribution of cumulative cases for realizations given the duty cycle interval (x-axis value), the disease/pathogen (row-label),

and the underlying population (column-label)—aggregated over random seeds and index infectives.

Violin plots of cumulative cases illustrate the risk of outbreak occurrence. In general cases, the `Snapshot` method preserves the distribution of cumulative regardless of increasing duty cycle interval. Meanwhile, the `Upperbound` method suffers from systematically overestimating the plausibility of having an outbreak, except for diseases with low R_0 (MERS). For diseases having relatively high R_0 (e.g., chickenpox, measles, pertussis) and close population (SHED1-2), the `Snapshot` method risks underestimating plausible outbreaks with sparse observations—those sampled an hour or more apart—while the `Upperbound` method retains the risk of corresponding outbreak occurrence at the cost of results varying between either universal infection or no further infection after the initial infection.

Welch's t-test We validated our interpretation of Figures 5.1 and 5.2 with the Bonferroni-corrected Welch's t-test, provided by R package *stats*, version 4.0.2. For each sampling method of the `Snapshot` and `Upperbound`, we tested cumulative cases with different duty cycle intervals blocked by diseases and underlying populations. Resulting 60 blocks, with each group (observation frequencies) having at least 960 samples (cumulative cases of realizations), sufficiently large to consider the robustness of t-test given the distribution of cumulative cases' departure from normality [470, 471], as shown in Figures 5.1 and 5.2. Setting the α-value as 5%, our null hypothesis is that given a disease/pathogen other than high R_0 diseases/pathogens (chickenpox, measles, pertussis) and a underlying population, the mean of cumulative cases resulted by proximate contacts collected with different observation frequencies of at least once per half-hour (equivalent to duty cycle intervals of 5, 10, and 30 minutes) are equal. For each block, pairwise by duty cycle intervals resulting three comparisons per block and $\alpha_{\text{altered}} = 0.05/3 = 0.0167$. It turned out for the `Snapshot` method null hypotheses are failed to reject except for SHED8-diphtheria between pairs of duty cycle intervals 30–5 $\big(t(4360.8) = 2.72,\, p = 0.0065\big)$, 30–10 $\big(t(4368.3) = 3.10,\, p = 0.0019\big)$; SHED8-SARS 30–5 $\big(t(4294.5) = 3.24,\, p = 0.0012)\big)$, 30–10 $\big(t(4338.9) = 2.78,\, p = 0.0006\big)$; and SHED9-diphtheria 30–5 $\big(t(4399.6) = 3.61,\, p = 0.0003\big)$, 30–10 $\big(t(4524.3) = 2.87,\, p = 0.0040\big)$. For the `Upperbound` method hypotheses are rejected, except for SHED2-fifth 10–5 $\big(t(1910.7) = 1.84,\, p = 0.0667\big)$, SHED2-MERS 10–5 $\big(t(1902.2) = 1.04,\, p = 0.2998\big)$; SHED7-fifth 10–5 $\big(t(3623.1) = 2.36,\, p = 0.018\big)$.

Prentice-Modified Friedman Test We further validated our interpretation of Figures 5.1 and 5.2 with the Prentice-modified Friedman test, provided by R package *muStat*, version 1.7.0. We tested cumulative cases grouped by sampling interval and blocked by data collection, sampling method, population (dataset), disease, and initial infection node. Resulting $\chi^2 = 222081$, with 6 degrees of freedom (reflecting the fact that the sampling interval $\xi \in \{5, 10, 30, 60, 90, 180, 360\}$ has 7 choices in total), and $p < 2.2\text{e}{-16}$, with the null hypothesis being that the sampling interval does not differentiate the distribution of cumulative cases, for the same data collection, sampling method, dataset, disease, and initial infection node.

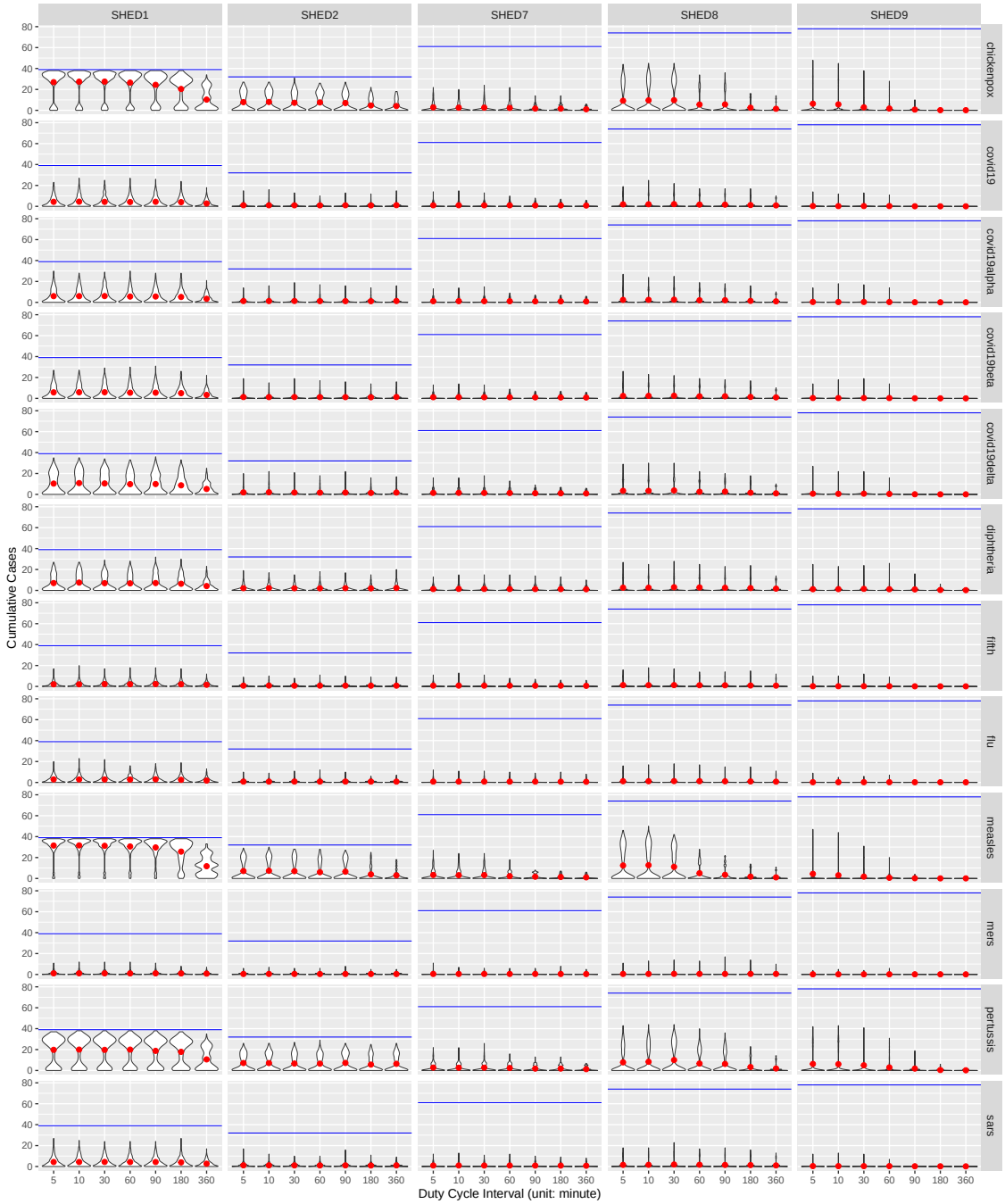


Figure 5.1: Grids of Violin Plots of Cumulative Cases for the `Snapshot` Method

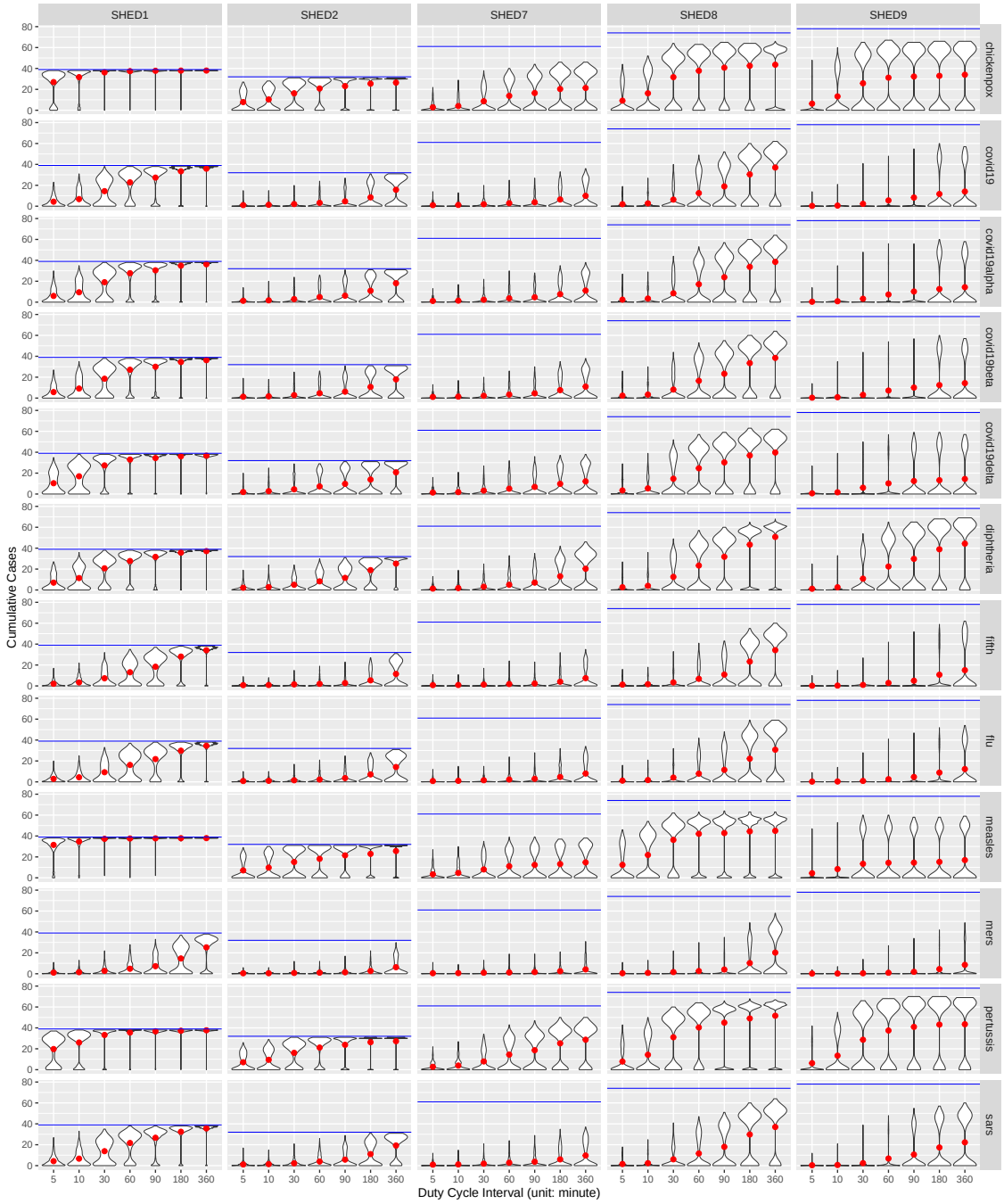


Figure 5.2: Grids of Violin Plots of Cumulative Cases for the `Upperbound` Method

Attack Rate

The accuracy-precision view measures the deviation with respect to the attack rate of simulations parameterized by the downsampled contact observations $\mathcal{D}_{\xi'},\ \xi' \in \{10, 30, 60, 90, 180, 360\}$ from the baseline $\mathcal{D}_{\xi_0},\ \xi_0 = 5$. Subplots are arranged as grids according to the combinations of underlying population V and disease $\mathcal{M}$. Within each subplot specific for a given combination of $\{\mathcal{D}, \mathcal{V}, \mathcal{M}\}$, deviation of the median attack rate is shown on the horizontal axis (reflecting accuracy), and deviation of the inter-quartile range (IQR) for attack rate is depicted on the vertical axis (negatively correlated with precision). Each datapoint within such a subplot is associated with a specific sample interval ξ of $\mathcal{D}$, whose value is denoted by both color and shape for visual clarity. In both Figures 5.3 and 5.4, points with the same color and shape tend to cluster instead of mixing with other colors, indicating that sample interval impacts govern both the accuracy and precision of the attack rate more than the initial infection node. In Figure 5.3, when downsampling with the `Snapshot` method, points are closer to the origin for communicable diseases/pathogens with low R_0 and for "diffuse" population such as $\{\text{SHED7}, \text{SHED8}, \text{SHED9}\}$, indicating the advantage of `Snapshot` at maintaining an estimate of attack rate as downsampling interval increases. For diseases with high R_0 (chickenpox, measles, pertussis) and "closer" communities $\{\text{SHED1}, \text{SHED2}\}$, `Snapshot` underestimates the attack rate as ξ increases, whereas `Upperbound` slightly overestimates. `Upperbound` reduces IQR deviation of estimated attack rate while the `Snapshot` increases interquartile range (IQR) deviation.

In Figures 5.3 and 5.4, we summarized two statistics: median and IQR, across the values of the attack rate drawn from an ensemble of 30 realizations for each scenario defined by observations of contact network $\mathcal{D}$, initial infectious individual $\mathcal{V}$, and a type of communicable disease $\mathcal{M}$. The accuracy-precision deviation of simulation results in terms of attack rate depends on the underlying population structure ("closer" or "diffuse"), the type of communicable disease, the sampling method (`Snapshot` or `Upperbound`), and the sample interval. The sampling interval is denoted with color. Figure 5.3 depicts median and IQR specifically for the `Snapshot` sampling method, while Figure 5.4 depicts for the `Upperbound` sampling method. Casual inspection of the skewed nature of the distributions towards higher values of the horizontal axis within each subplot (indicating increasing median deviation in incidence) confirms that increasing the sample interval results in over-estimation of the attack rate, as is suggested in [109]. By contrast, the clustering of the points by color in each subplot suggests that the initial infection node exerts a smaller impact on the two statistics we have chosen to reflect the accuracy-precision tradeoffs.

Comparing within subplots column-wise, the `Snapshot` method performs well with diffuse communities, resulting in both low deviation of median and low deviation of IQR. When used with close networks, `Snapshot` tends to overestimate the attack rate but underestimate the IQR. `Upperbound` exhibits greater deviation than `Snapshot`, and is more consistent as sampling interval increases given other factors—from left to right. When sampling interval is brief and sampling rate high, attack rate exhibits low median and IQR deviation from the ground truth, because the reconstructed contact network is less distorted. As the sampling interval increases and sampling rate decreases, `Upperbound` tends to become both less accurate and less precise. As

the sample interval increases further, the overestimation of the attack rate reaches a limit as people directly or indirectly connected to the initially infectious person are reliably infected for high R_0 pathogens, or people remain uninfected for pathogens with low R_0.

Comparing within subplots row-wise, disease-specific patterns are also visible: estimates for the attack rate of diseases with low R_0, such as MERS, seem relatively insensitive to the sampling interval. Pathogens/communicable diseases with sufficiently high R_0 tend to behave similarly as sampling interval increases, regardless of their differences in R_0 value.

Outbreaks and Outbreak Timing

Outbreak timing and behavior are commonly studied characteristics of communicable diseases, yet the quantifiable definition of an outbreak varies due to challenges regarding data collection and characterization of the appropriate cohort to be counted. Instead of imposing a quantitative definition, this work employs cumulative cases over time as a measure to reflect outbreak dynamics of disease in simulations for a given underlying population V and observed contact data $\mathcal{D}$.

We computed the empirical cumulative distribution (ECDF) of incidence occurrence time given disease $\mathcal{M}$ and observed contact data $\mathcal{D}$ by marginalizing the time of incidence across initial infectious person $\mathcal{V}$ and realizations. In Figure 5.5, the empirical cumulative distribution (ECDF) of incidence occurrence time across alternative assumptions regarding the initial infectious person $\mathcal{V}$ are arranged as grids by columns of underlying population V and rows given by downsampling interval ξ. Lines in color denote ECDF of incidence time given observed population/dataset (columns) samples with the `Snapshot` method and `Upperbound` method at corresponding duty cycle intervals (row), with other factors remaining unchanged. Within each grid, a baseline ECDF of incidence time given observed contact data without downsampling $\mathcal{D}_{\xi_0}$ is provided as a reference. Grids with three colored ECDF curves close to each other indicate that under the scenario of underlying population V and downsampling interval ξ, the timing and existence of outbreaks are insensitive to the downsampling interval ξ, regardless of the sampling method (`Snapshot` vs. `Upperbound`).

We selected four representatives. Similar diseases appear to have similar dynamics, as shown, for example, in Figure 5.5a and Figure 5.5b; diseases with extremely low and high R_0 tend to behave quite differently regardless of sampling method, interval, and dataset, as can be seen by contrasting Figure 5.5c and Figure 5.5d. The steeper ECDF curves of the `Snapshot` method exhibiting similar shape but ending earlier on the horizontal axis as the downsampling interval increases the pathogen spread is halting by disconnections among infectious and susceptible due to missed contacts. Our measure automatically normalized outbreak size to lie in the interval $[0, 1]$. This analysis demonstrates that:

- As would be expected, given an initially susceptible population, a pathogen with a tendency to catalyze an outbreak will often exhibit an apparent, sharp increases in infections during the outbreak period. Weakly spreading pathogens have an initial ascent followed by a long tail. More notable is that this tendency holds largely invariant of sampling method, population, and sampling interval.

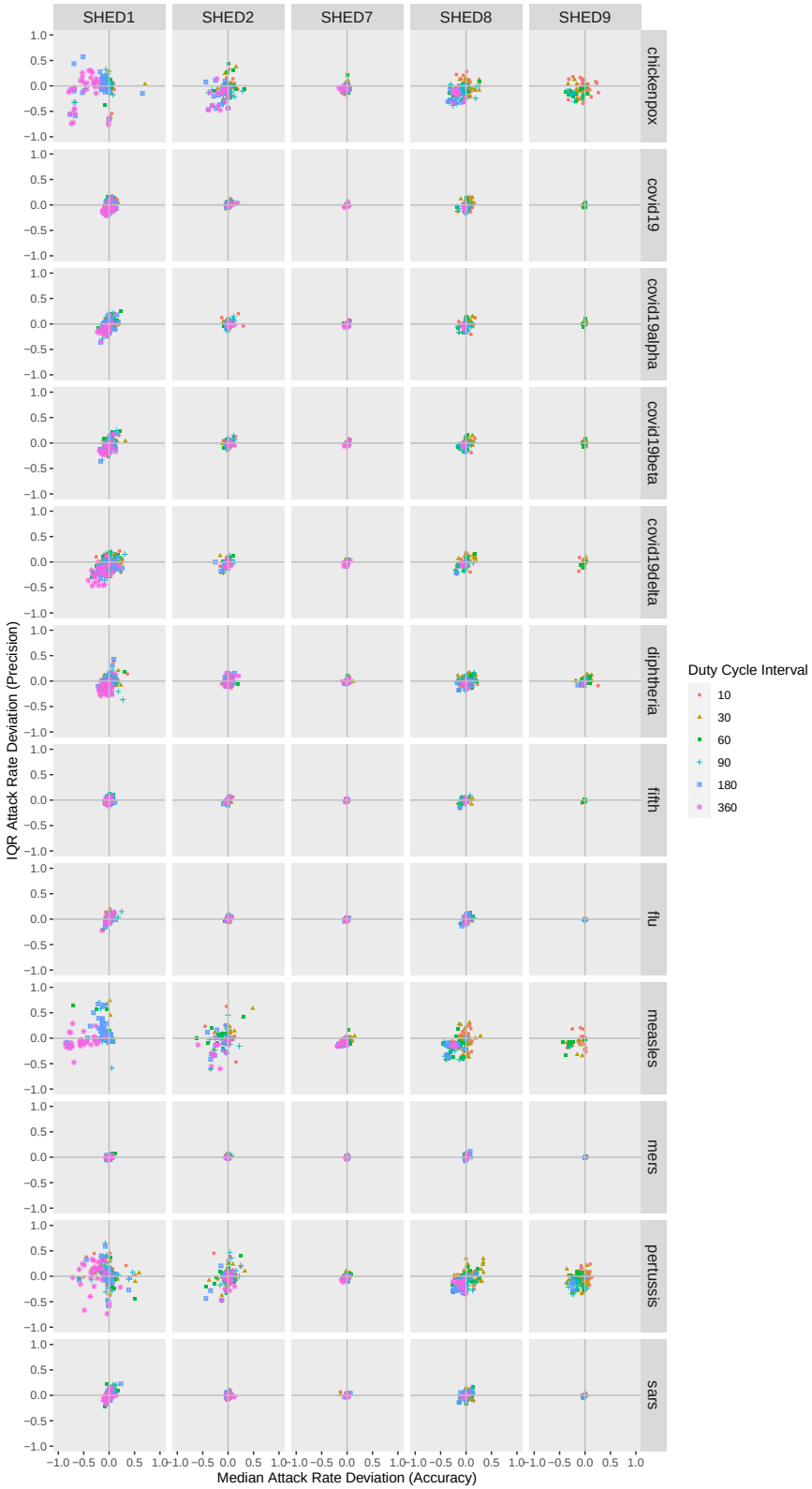


Figure 5.3: Attack Rate Given Initial Infection Node for the `Snapshot` Method

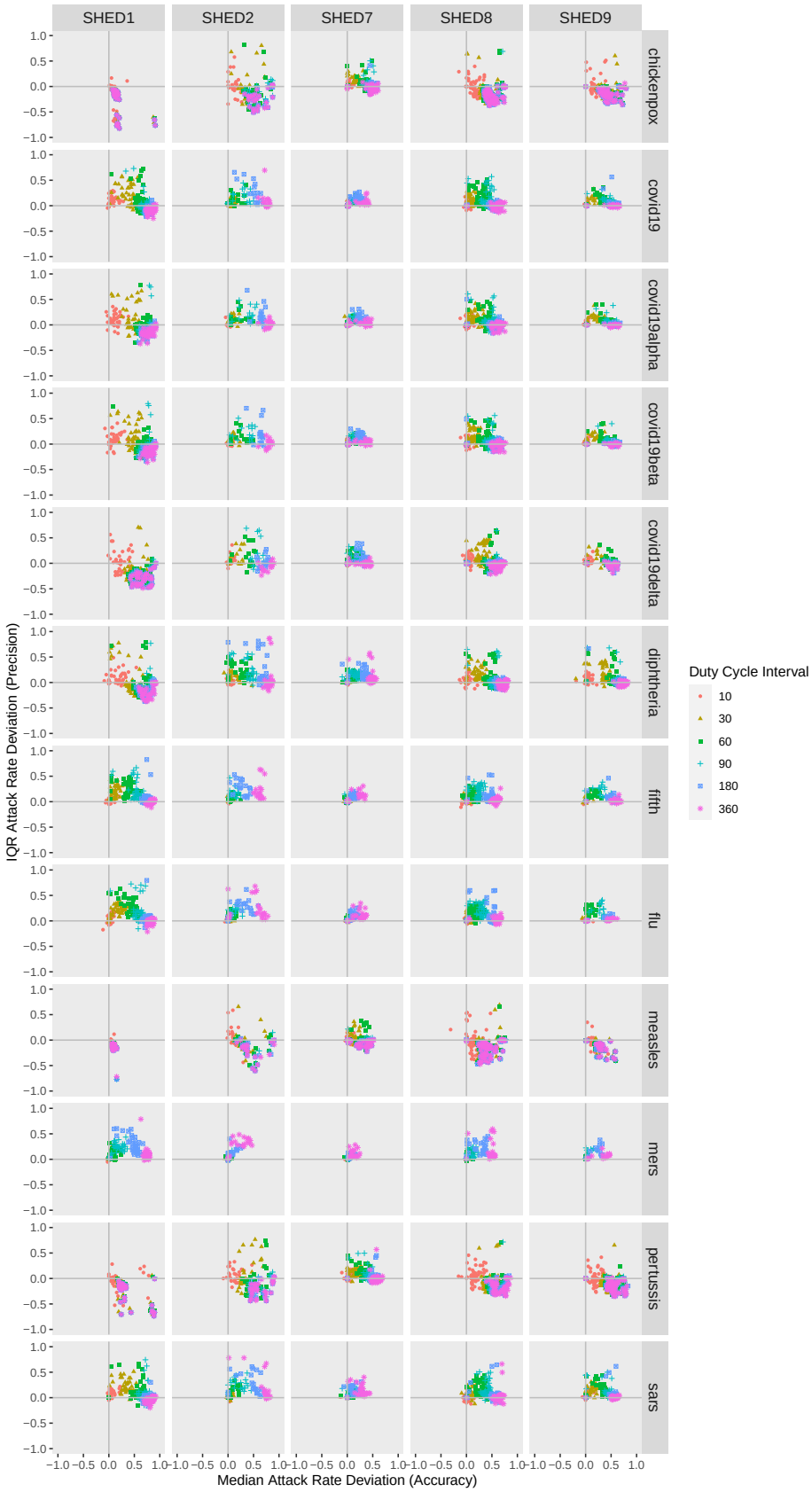


Figure 5.4: Attack Rate Given Initial Infection Node for the `Upperbound` Method

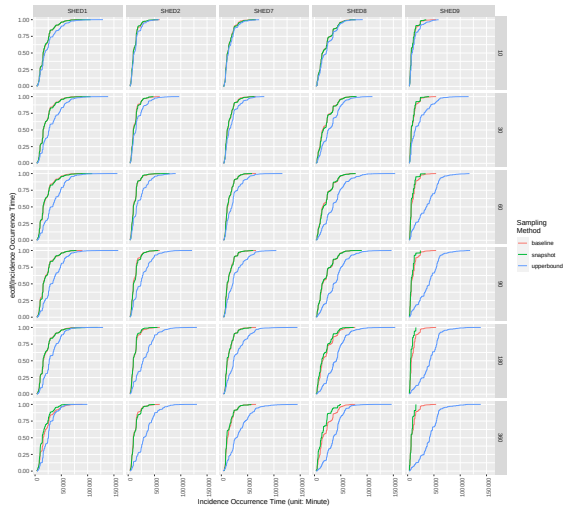
- In a pattern that is maintained—*mutatis mutandis*—across populations, sampling method, and interval, similar diseases exhibit clinically similar curves: SARS (Figure 5.5b) is known to have similar characteristics to flu (Figure 5.5a), and they exhibit similar patterns for our measure. The discrepancy is small for close populations.
- In a pattern that again holds independent of sampling method and interval as well as population (dataset), diseases with different R_0 behave differently. Pertussis (Figure 5.5c) has the highest R_0 amongst the diseases we simulated, while MERS (Figure 5.5d) has the lowest. Their pattern is distinct—pertussis tends to have a clearer outbreak. By contrast, SARS exhibits a steep curve in the beginning and a long tail, indicating limited disease spread.
- Discrepancies between `Snapshot` and `Upperbound` from the baseline increased with the sampling interval ξ. Discrepancies induced by the sampling interval exert less impact than the characteristics of the study population, with "diffuse" communities (like SHED9) exhibiting substantial discrepancies for both `Snapshot` and `Upperbound`.
- For a "diffuse" community, `Snapshot` outperforms `Upperbound` when the sampling interval is extensive, with more minor differences in median and IQR deviation.

5.4.2 Impacts on Individual-Level Simulation Results

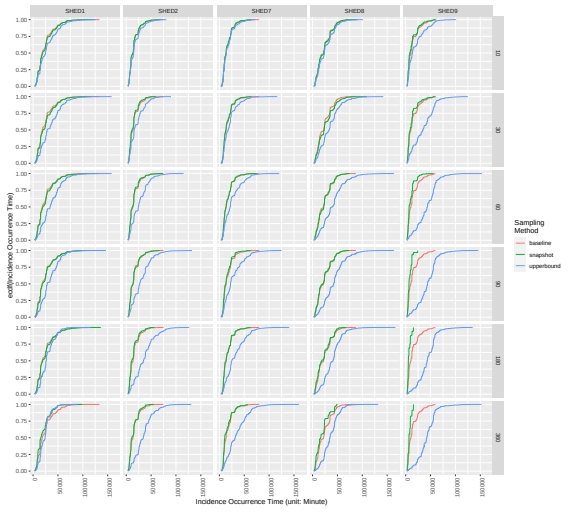
We measure the impacts of observation frequencies on the simulation results at the individual level with transmission pathways and individual infection risks. Individual risk of infection can suggest vulnerable group to prioritize resource allocation and ensure health equity [472]. Individual risk of infection is asymptotically approached by the fraction of realizations in which an individual is infected. The difference of individual risk of infection can be compared pairwise in terms of the weighted-Minkowski distance among scenarios with different datasets $\mathcal{D}$ for the same underlying population V and disease $\mathcal{M}$. We calculated the Kullback-Leibler divergence on individual infection probabilities with different ξ to draw quantitative conclusions on the impact of downsampling frequency on simulation results at the population level. Higher KL-divergence values from the `Snapshot` method for SHED9 were observed for chickenpox, COVID-19, diphtheria, measles, and pertussis, and are indicated by reddish colors of the corresponding column on Figure 5.6. Lower KL-divergence values associated with MERS, regardless of dataset and downsampling frequency, induces its greenish color in the corresponding column in that figure. We find that the KL-divergence can effectively summarize the information shown on Figure 5.6 and therefore can serve as an efficient metric to measure differences in individual risk.

Distances Matrices of Infection Pairs

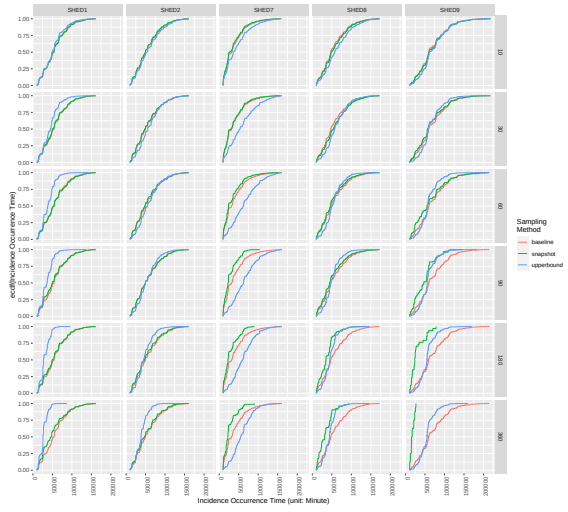
Figure 5.6 shows matrices of pairwise weighted-Minkowski distances of frequencies of infections pairs given downsampling methods, disease, and sampling frequencies for underlying populations, with the color shifting



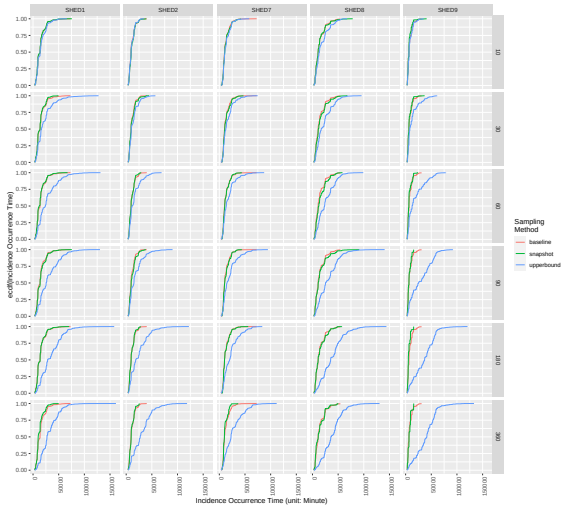
(a) flu



(b) SARS

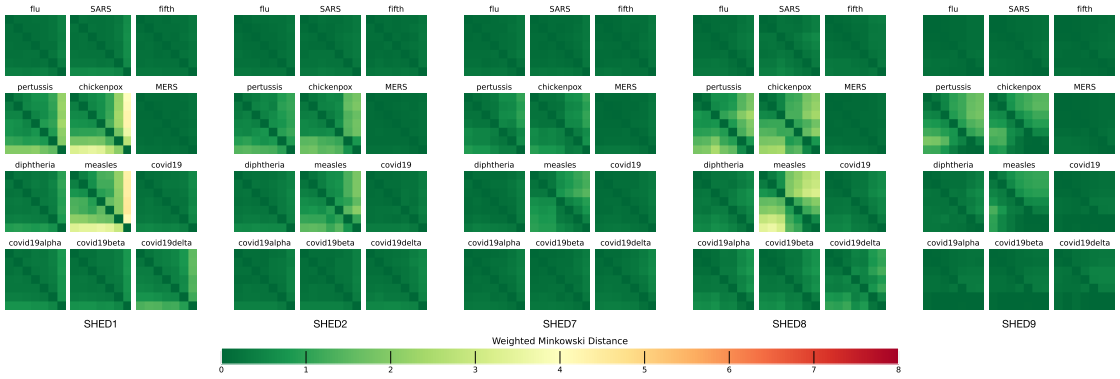


(c) pertussis

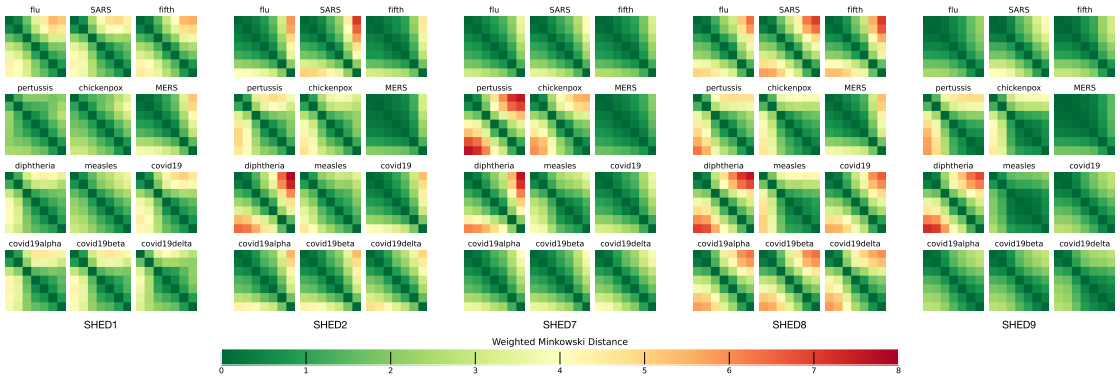


(d) MERS

Figure 5.5: Comparison of Outbreak Timing



(a) `Snapshot`

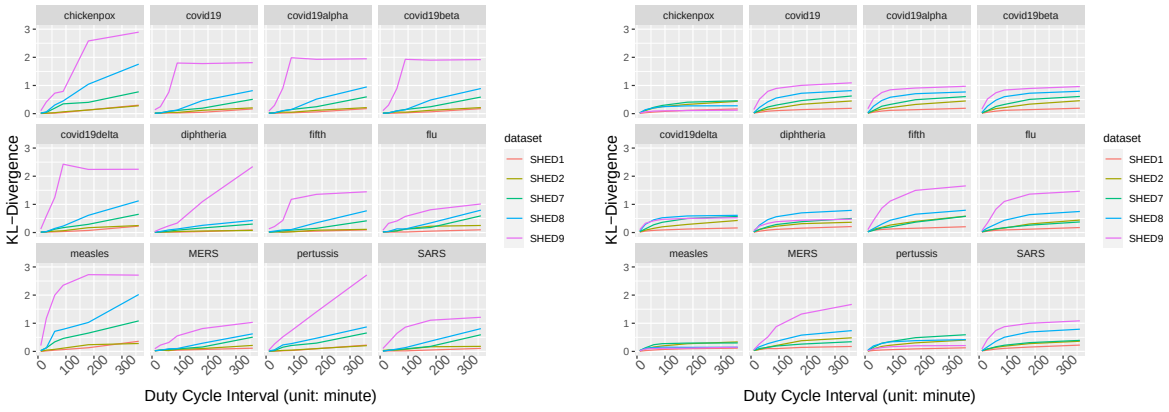


(b) `Upperbound`

Figure 5.6: Distance Matrix of Infection Pairs

from greenish to reddish with the increasing degree of dissimilarity. For each matrix, starting from its top left corner, inter-observation intervals are arranged in ascending order—$\xi = 5, 10, 30, 60, 90, 180, 360$—horizontally from left to right and vertically from top to bottom. We found `Snapshot` is better at preserving consistent frequencies of infection pairs, particularly with an observation frequency higher than once per half-hour, except for higher R_0 diseases in "diffuse" communities, such as chickenpox and measles in SHED9. For lower R_0 diseases in general, particularly in "closer" communities like SHED1, the `Snapshot` method have weighted-Minkowski distance less than 1 even between the observation frequencies of 5-minute and 360-minute.

We found `Upperbound` is better at preserving likely paths than `Snapshot`, and the limits of the sampling interval needed to preserve likely paths of disease spreading lies amongst $\xi \in \{10, 30, 60\}$. Under `Upperbound`, diseases with similar R_0 resemble each other, and MERS with a low $R_0 = 0.69$, has its likely paths varying notably over sampling intervals for a less diffuse population, while other diseases—despite exhibiting a wide range of $R_0 \in [0.69, 15]$—maintain a similar pattern of those likely paths with rising sampling interval, for a given population.



(a) `Snapshot` **(b)** `Upperbound`

Figure 5.7: Kullback-Leibler Divergence of Individual Infection Risk

Kullback-Leibler Divergence on Individual Infection Risk

For each individual given each combination of disease and datasets, we calculated the Laplacian-smoothed individual infection probability based on infection counts from simulations fed with ξ-sampled contact data using downsampling method μ, where $\mu \in \{\texttt{Upperbound}, \texttt{Snapshot}\}$; then we assembled the individual infection probability into a vector of the individual infection risk, denoted $\rho_{(\mathcal{M},\mathcal{D}_{\xi,\mu})}$. Laplacian-smoothing was employed to ensure that those who were not infected in simulation outcomes are still assigned a small probability of being infected.

To characterize $\rho_{(\mathcal{M},\mathcal{D}_{\xi,\mu})}$ as shown in Figure 5.7, we arranged the presentation top-down, characterizing the distinct downsampling methods (μ) using two sub-figures, with disease ($\mathcal{M}$) within each sub-figure as wrapped facets. Finally, within each facet, we plotted a line for each underlying population, with the x-value as the duty cycle interval ξ, and the y-value as $\delta_{\xi_+} = D_{\mathrm{KL}}(\rho_{(\mathcal{M},\mathcal{D}_{\xi_0,\mu})} \parallel \rho_{(\mathcal{M},\mathcal{D}_{\xi_+,\mu})})$, where $\xi_0 = 5$, and $\xi_+ \in \{10, 30, 60, 90, 180, 360\}$.

As shown in Figure 5.7, the `Snapshot` method in general will exhibit higher divergence than the `Upperbound` method, except for diseases with low R_0, such as influenza type A (1.31) and MERS (0.69). In general, the higher the δ_{ξ_+}, the higher the divergence of individual infection risk from estimations with ξ_+-downsampled contact data when compared to ξ_0-sampled contact data.

5.5 Discussion

The `Snapshot` method generally resulted in faithful population-level estimates—in terms of securing estimates on the mean of cumulative cases at the cost of distort the probability distribution of cumulative cases across realizations—invariant to the duty cycle interval ξ while imposing the risk of underestimating the attack rate for high R_0 pathogens, particularly for denser communities such as SHED1. By contrast, for population-level

estimations, the `Upperbound` method can produce close estimates only for short duty cycle intervals. However, for individual-level estimations, we found that the `Upperbound` method generally outperforms the `Snapshot` method, except for low R_0 pathogens and diffuse communities.

We found that both the sampling method and sampling interval exert impacts on the simulation results; moreover, their impacts vary depending on the type of pathogen and the sparse-versus-dense attributes of the community. Such differences are evident in Figures 5.1 to 5.4 by the distinct patterns distinguishing different downsampling methods and different intervals within a sampling method.

For most subplots in Figures 5.1 and 5.2, locations and shapes of boxplots within each subplot across different sampling intervals exhibit greater similarity to each other for the `Snapshot` method than for the `Upperbound` method. Moreover, in Figures 5.3 and 5.4, dots are closer to the origin for the `Snapshot` method than the `Upperbound` method. These closer-to-the-origin dots indicates that in estimating the attack rate, as the duty cycle interval increases, the `Snapshot` method achieves lower bias and variance than does `Upperbound`. In short, when fed into the ABM model, the results of the `Snapshot` method—which only captures proximity contacts within a short duration window within each duty cycle—can produce more accurate estimates at a lower variance than results sampled from the `Upperbound` method.

Despite `Snapshot`'s effectiveness in capturing contact networks in terms of estimating the mean of cumulative cases and attack rate, when considering individual infection risk, `Upperbound` is superior. In Figure 5.6, the weighted-Minkowski distances for the `Snapshot` method is, in general, larger than that for the `Upperbound` method, particularly for sparse populations like that SHED9, and with diseases having higher R_0, such as pertussis, and measles.

For simulations to yield reliable conclusions with respect to infection transmission across sparsely connected communities, there needs to be compensation for reductions in observation frequency through elevation of the ensemble sizes. Reducing observation frequency can alter simulation results, and comparisons on network structure changes induced by downsampling suggest the need to maintain information on the ordering of contacts to better simulate contact networks reconstructed from high-resolution contact data. This also indicates that apps that have already employed the `Snapshot` method, such as EthicaData [388], can be an excellent tool for efficiently and effectively capturing proximity contact data.

In terms of network structure presented by downsamplings with the `Snapshot` and `Upperbound` method, the `Snapshot` method ensures accuracies of cumulative contact time at the cost of distorting the node-degree distribution. The `Snapshot` method seems to retain better population-level estimates than the `Upperbound` method for results of an empirical contact data empowered transmission model. There might be associations between metrics of contact network (such as cumulative contact time) and transmission model results (such as the mean of cumulative cases and attack rate).

Our findings are subject to a number of important limitations:

- **Limited Population Size**—Given the confined population size, it is possible that the observed behavior of `Snapshot` and `Upperbound` here is materially altered by quantization effects exhibited by agent-based

simulation when population size is small [95, 153, 473]. Because the attack rate represents the quotient of two integers, when both nominator and denominator are small, the possible values of their ratio can be sparse. A partial result is that realizations in which the entire population is infected can happen more frequently than when the population size is large.

- **Limited Diversity of Participants**—As university students, most participants of our experiments share similar lifestyles and activity spaces for their working and studying hours. The findings resulting from applying such methods to larger and more societally representative communities may vary notably from the results shown here.

- **Simple Modeling Methodologies**—While the method of feeding high-resolution proximity contact data into the epidemiological model used in our experiment is mathematically and practically straightforward, there is much opportunity to apply more versatile methods to combine data into the model as modeling methodologies advance.

5.6 Conclusion

This work has investigated the impacts of observation frequency and sampling methods for proximity contact records in capturing proximity contact networks for epidemiological simulations. We evaluated the impacts induced by the temporal granularity of sampling networks in terms of distortion of measured network structure and of population-level and individual-level simulation outcome metrics in light of combinations of specific diseases and underlying types of communities. These results emphasize classes of pathogens and population structures in which the design of new studies should prioritize frequent sampling of contact networks. Our findings also provide guidance as to how network density and lower sampling rates might distort measures such as attack rate and individual risk.

6 Conclusions

6.1 Summary

The incorporation of high-resolution proximity contact data offers an important avenue towards advancing transmission models. Effective public interventions and measurements to control infections—such as contact tracing, precision lockdowns, safeguarded reopening, and vaccine planning—can benefit from a better understanding of proximate contacts. High-resolution proximity contact data and methods to fuse these data with transmission models help secure insights into the significance of proximate-contact patterns.

The first contribution of this dissertation is the demonstration of the utility of Extended Kalman Filtering (EKF) as an approach to enable a System Dynamics SIR (SD-SIR) model to benefit directly from high-resolution proximity contact data. The experiment designed in Chapter 3 reveals connections between a filtered SD model and an agent-based SIR (ABM-SIR) model, with both taking advantage of recurrent incoming data. The filtered SD model takes recurrent noisy surveillance data such as a time series of new cases, and the agent-based model takes high-resolution proximity contact data. The demonstrated EKF empowered SD model has improved estimations better matching outbreak peaks. This improvement indicates the potential to train the SD model with surveillance data to better match the high-resolution proximity contact data assisted ABM model, overcoming inaccuracies in the SD model structure and its parameter estimates. Meanwhile, it is plausible to have agent-based models hand over some expensive simulation tasks to a filtered SD model for rapid response to interactive what-if questions. The work of EKF also revealed a need to study the impact of unknown sensing noise on simulation results of transmission models, particularly considering the dynamics of human behavior and disease spreading.

The second contribution of this dissertation demonstrated that both spatial and temporal resolution have a fundamental impact on high-resolution proximity contact data. Experiments designed in Chapter 4 and Chapter 5 substantiated the importance of such data in general, and pointed to the need for further studies with various populations, diseases, sensing techniques, and resolutions. For simulation results exhibiting invariance under changes in resolution, general patterns and exceptions suggest opportunities to optimize sensing regimens and generalize findings from simulations to similar scenarios. For example, identifying a minimum sufficient observation frequency can help slow battery drain. Doing so may support the viability of using varying pedigrees of smartphones, promoting smartphone-assisted contact tracing globally to populations at various degrees of development and socioeconomic status. This research lays the groundwork for further studies that could support identifying characteristics of the contact network giving rise to such invariance.

Such work could, for instance, shed light on contexts under which potential outbreaks are oblivious to certain types of variation in contact patterns, helping spot safe opportunities for lockdown relaxation and reopening.

Beyond allowing us to measure and quantify contact patterns, the advent of sensing technologies, such as GPS and Bluetooth, is also shaping such patterns. For example, GPS-equipped navigation systems may route a vast number of users to the same small set of optimized routes, thereby inadvertently boosting co-location of individuals. On the other hand, Bluetooth lessens dependence on wired communications, enabling users to move away from their electronic devices and mingle in proximity to others. There are circumstances involving underlying populations and pathogens under which the results of transmission simulation exhibit invariance to changes in resolution of proximity contact data. Studies characterizing such circumstances can aid under-parameterized models in emerging or poorly studied diseases, helping us better prepare for features of the next pandemic.

6.2 Future Work

This dissertation raises the opportunity for and highlights the importance of some interesting future work. The groundwork of Chapter 4 and Chapter 5 hypothesized and evaluated the character of impacts of sensors' temporal-spatial resolution on transmission models informed by sensor-collected proximity contact data. However, the findings of Chapter 4 and Chapter 5 are limited to the scope of proximity contact data used in experiments. Proximity contact data collected from a larger-sized underlying population with more diversified occupations can enrich sensitivity analyses of impacts of temporal-spatial resolution.

In Chapter 5, the `Snapshot` sampling method with at least half-hourly observation seems to be a decent default sensing regime. Simulations with parameters of other communicable diseases may lead to better estimations on configuring the default sensing regime. Particularly, communicable diseases can be classified into subgroups based on their sensitivities to spatial resolution. Identifying an effective classifier for a communicable disease's sensitivity to the sensing regime can suggest sensor configuration for a newly emerged disease whose parameters remain unknown.

Personal privacy-protected proximity contact data collection is crucial to acquire data to inform transmission models. Sensing modalities with lower spatial resolution can naturally provide better protection of personal privacy. Following the experiment in Chapter 5, comparisons with other sensing modalities, such as co-location based on cellular tower trilateration, can support finding a better solution for personal privacy-protected proximity contact data collection.

6.3 Conclusions

This dissertation studied the fusion of high-resolution proximity contact data with transmission models by answering the following three questions:

- Is it possible to use the Kalman filter to advance transmission modeling by incorporating high-resolution proximity contact data?
- Whether and to what degree does increased spatial resolution in proximate-contact sensing impact simulation outcomes?
- Whether and to what degree does increased temporal resolution in proximate-contact sensing impact simulation outcomes?

We addressed these problems with simulation experiments taking and designed around empirical longitudinal datasets. Simulation results suggest transmission models can benefit from high-resolution proximity contact data: With recurrent EKF updates and incoming estimates from an ABM-SIR model, an SD-SIR model can mitigate structural and parametric inaccuracies at the cost of uncertainties in observed proximate contacts due to the temporal-spatial resolution. An agent-based model can benefit from proximity contact data directly, but its estimates' sensitivity to the temporal-spatial resolution of proximity contact data varies under circumstances of diseases/pathogens, underlying population, and the stochastic of disease spreading. Characterization how this sensitivity changes under the circumstances reveals opportunities to tailor sensing regimes and modalities for proximity contact data collection, ensuring model estimates' accuracy and data collections' energy efficiency. To summarize, the contributions of this dissertation to the literature are as follow:

- The extended Kalman filter (EKF) can improve a System Dynamics/compartmental aggregate susceptible-infectious-removed (SD-SIR) model by periodically regrounding that model with even noisy surveillance data. The EKF filtered SD model has its outbreak peak estimations in better accordance with synthetic data outcoming from an agent-based SIR (ABM-SIR) model taking individual-level proximity contact data. This better accordance reflects that the EKF solution compensates for an SD-SIR model's structural and parametric inaccuracies. For example, an SD-SIR model usually assumes random mixing within its compartments, aggregating individual-level preferential contacts into group-level preferential mixing rates. This aggregation replaces contact rates between two individuals with the norm of contact rates between two groups to which these two individuals belong. EKF updates taking recurrent incoming data can reground model states, mitigating inaccuracies incurred from replacing individual contact rates with group norm and estimating the norm of contact rates between two groups. Because an EKF improved aggregate model has improved accuracies without a surging burden of computational complexities, it may support offloading expensive computing from the agent-based model.
- For transmission modeling using derived proximate contacts from proximity contact data, the impact of sensing modality and the distance threshold of two individuals having proximity contact can substantially impact modeling outcomes. GPS co-location and Bluetooth beaconing are two typical sensing modalities employed to collect proximity contact data. Even with an identical agent-based susceptible-exposed-infectious-removed (ABM-SEIR) model, GPS co-location derived- vs. Bluetooth beaconing derived-

proximate contacts can still result in different predictions regarding disease trajectories. The magnitude of differences in predictions varies by disease type and underlying population but is usually notable from individual risks of infection. Proximity contact data measured using different spatial resolutions cannot be treated as interchangeable, despite combinations of disease/pathogen types and underlying populations under which the results of an ABM-SIR can exhibit insensitivity to the spatial resolution. Studies on conditions governing low sensitivity to spatial resolution may lead to improvements in data compatibility and reusability.

- The observation frequency and the sampling method are referred to as the sensing regime. The sensing regime impacts the temporal resolution of proximity contact data collection, hence the derived proximate contacts for transmission modeling. For example, when collecting proximity contact data for derived proximate contacts taken by an ABM-SEIR model, with the sampling method fixed, a sensing regime with higher observation frequency usually results in low deviance of the ABM-SEIR model estimated attack rate and individual risks of infection from the baseline—the sensing regime with the highest possible observation frequency. The `Snapshot` sampling method can moderate the lowering observation frequency caused diminishing accuracy in simulation-based outcomes, thus is worthy of further consideration. Although disease type and underlying population can considerably alter the model's sensitivity to temporal resolution, results suggest, with the `Snapshot` sampling method, it is plausible to lower observation frequency to once per 10 minutes or even once per 30 minutes while retaining simulation results in bearable accordance with the baseline of once per 5 minutes.

This dissertation demonstrated a practical approach to improving transmission models. The first part demonstrated that filtering techniques and proximity-sensing-informed contact data could lessen the distortions caused by System Dynamics and compartmental transmission models. The distortions due to assumptions of random-mixing are lessened and replaced with distortions characteristic of the sensing regime and modality involved. The second component of this thesis studied the impacts on simulated results of potential outbreaks of spatial and temporal resolution and two key configurable factors of sensing modalities for proximate contacts, shedding light on managing distortions due to sensing data and modality. When synthesized together, further research related to either part will bring new insights into improving transmission models and our approaches.

Our approach also serves as a way to present, store and retrieve proximity contact networks. Proximity contact networks have been presented by stylized models, such as with Poisson random, small-world and scale free networks, generative algorithms such as Erdős–Rényi, and Watts-Strogatz and Barabási–Albert algorithms. These approaches focus on characterizing network features with key parameters, and relying on calibrated key parameters to parametrically approximate the structure of particular contact networks. While such methods are recommended by conciseness, such parametric characterizations of networks are static—thereby lacking temporal dependency—and serve as a notably lossy approximation to proximity contact networks. Our approach of storing proximate contacts as event records and downsampling and filtering

techniques to retrieve stored proximity contact networks can preserve temporal dependency and population characteristics, particularly when used with transmission models. Keeping simulation results comparable while adjusting the downsampling method and parameters suggests compressed or feature extracted proximity contact networks.

We found that systematic approaches are required to study transmission simulation results, which, notwithstanding exceptions, cannot be generally independently analyzed with respect to sensing modalities, underlying population, and disease parameters in isolation. Insensitivity of simulation results to sensing modalities, underlying population, and disease parameters was only noticed in certain combinations, such as with a diffuse population with the GPS co-location method given a less transmissible disease. Our findings suggest that changing parameters, such as via implementation of public health orders, will give rise to different strengths of impact given different underlying populations or different diseases parameters, such as given for variants of concern of SARS-CoV-2.

References

[1] J. K. Taubenberger and D. M. Morens, "1918 influenza: The mother of all pandemics," *Emerging Infectious Diseases*, vol. 12, no. 1, pp. 15–22, 2006. DOI: 10.3201/eid1201.050979.

[2] World Health Organization, "Weekly operational update on COVID-19—12 July 2021," *Emergency Situational Updates*, no. 63, 2021.

[3] D. M. Cutler and L. H. Summers, "The COVID-19 pandemic and the $16 trillion virus," *JAMA*, vol. 324, no. 15, pp. 1495–1496, 2020. DOI: 10.1001/jama.2020.19759.

[4] R. Zheng, Y. Xu, W. Wang, G. Ning, and Y. Bi, "Spatial transmission of COVID-19 via public and private transportation in China," *Travel Medicine and Infectious Disease*, vol. 34, p. 101 626, 2020. DOI: 10.1016/j.tmaid.2020.101626.

[5] H. Lau *et al.*, "The association between international and domestic air traffic and the coronavirus (COVID-19) outbreak," *Journal of Microbiology, Immunology and Infection*, vol. 53, no. 3, pp. 467–472, 2020. DOI: 10.1016/j.jmii.2020.03.026.

[6] A. Findlater and I. I. Bogoch, "Human mobility and the global spread of infectious diseases: A focus on air travel," *Trends in Parasitology*, vol. 34, no. 9, pp. 772–783, 2018. DOI: 10.1016/j.pt.2018.07.004.

[7] M. Cinelli *et al.*, "The COVID-19 social media infodemic," *Scientific Reports*, vol. 10, no. 1, p. 16 598, 2020. DOI: 10.1038/s41598-020-73510-5.

[8] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychological Science*, vol. 31, no. 7, pp. 770–780, 2020. DOI: 10.1177/0956797620939054.

[9] D. Greenberg, "Compulsive hoarding," *American Journal of Psychotherapy*, vol. 41, no. 3, pp. 409–416, 1987. DOI: 10.1176/appi.psychotherapy.1987.41.3.409.

[10] D. Banerjee, "The other side of COVID-19: Impact on obsessive compulsive disorder (OCD) and hoarding," *Psychiatry Research*, vol. 288, 2020. [Online]. Available: https://doi.org/10.1016/j.psychres.2020.112953.

[11] Z. Barua, S. Barua, S. Aktar, N. Kabir, and M. Li, "Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation," *Progress in Disaster Science*, vol. 8, 2020. DOI: 10.1016/j.pdisas.2020.100119.

[12] F. K. Kommoss *et al.*, "The pathology of severe COVID-19-related lung damage: Mechanistic and therapeutic implications," *Deutsches Ärzteblatt International*, vol. 117, no. 29-30, pp. 500–506, 2020. DOI: 10.3238/arztebl.2020.0500.

[13] P. Chen, L. Mao, G. P. Nassis, P. Harmer, B. E. Ainsworth, and F. Li, "Coronavirus disease (COVID-19): The need to maintain regular physical activity while taking precautions," *Journal of Sport and Health Science*, vol. 9, no. 2, pp. 103–104, 2020. DOI: https://doi.org/10.1016/j.jshs.2020.02.001.

[14] W. Cullen, G. Gulati, and B. Kelly, "Mental health in the COVID-19 pandemic," *QJM: An International Journal of Medicine*, vol. 113, no. 5, pp. 311–312, 2020. DOI: 10.1093/qjmed/hcaa110.

[15] S. Mallapaty, "The search for animals harbouring coronavirus-and why it matters," *Nature*, vol. 591, no. 7848, pp. 26–28, 2021. DOI: 10.1038/d41586-021-00531-z.

[16] M. E. El Zowalaty, S. G. Young, and J. D. Järhult, "Environmental impact of the COVID-19 pandemic–a lesson for the future," *Infection Ecology & Epidemiology*, vol. 10, no. 1, p. 1 768 023, Jan. 2020. DOI: 10.1080/20008686.2020.1768023.

[17] S. A. Plotkin, *History of Vaccine Development*. Springer Science & Business Media, 2011.

[18] E. Tognotti, "Lessons from the history of quarantine, from plague to influenza A," *Emerging Infectious Diseases*, vol. 19, no. 2, pp. 254–259, 2013. DOI: 10.3201/eid1902.120312.

[19] K. Drews, "A brief history of quarantine," *The Virginia Tech Undergraduate Historical Review*, vol. 2, 2013. DOI: 10.21061/vtuhr.v2i0.16.

[20] E. Rafferty, W. McDonald, W. Qian, N. D. Osgood, and A. Doroshenko, "Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling," *PeerJ*, vol. 6, e5012, 2018. DOI: 10.7717/peerj.5012.

[21] A. Doroshenko, W. Qian, and N. D. Osgood, "Evaluation of outbreak response immunization in the control of pertussis using agent-based modeling," *PeerJ*, vol. 4, e2337, 2016. DOI: 10.7717/peerj.2337.

[22] S. T. Goldstein, F. Zhou, S. C. Hadler, B. P. Bell, E. E. Mast, and H. S. Margolis, "A mathematical model to estimate global hepatitis B disease burden and vaccination impact," *International Journal of Epidemiology*, vol. 34, no. 6, pp. 1329–1339, 2005. DOI: 10.1093/ije/dyi206.

[23] L. Ribassin-Majed, R. Lounes, and S. Clémençon, "Efficacy of vaccination against HPV infections to prevent cervical cancer in France: Present assessment and pathways to improve vaccination policies," *PLOS ONE*, vol. 7, no. 3, e32251, 2012. DOI: 10.1371/journal.pone.0032251.

[24] L. Willem, S. Stijven, E. Vladislavleva, J. Broeckhove, P. Beutels, and N. Hens, "Active learning to understand infectious disease models and improve policy making," *PLOS Computational Biology*, vol. 10, no. 4, e1003563, 2014. DOI: 10.1371/journal.pcbi.1003563.

[25] J. A. Lewnard and N. C. Lo, "Scientific and ethical basis for social-distancing interventions against COVID-19," *The Lancet Infectious Diseases*, vol. 20, no. 6, pp. 631–633, 2020. DOI: 10.1016/S1473-3099(20)30190-0.

[26] S. M. Kissler, C. Tedijanto, M. Lipsitch, and Y. Grad, *Social Distancing Strategies for Curbing the COVID-19 Epidemic*, Mar. 2020. [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42638988.

[27] M. Qian and J. Jiang, "COVID-19 and social distancing," *Journal of Public Health*, vol. 30, pp. 259–261, 2022. DOI: 10.1007/s10389-020-01321-z.

[28] F. Schlosser, B. F. Maier, O. Jack, D. Hinrichs, A. Zachariae, and D. Brockmann, "COVID-19 lockdown induces disease-mitigating structural changes in mobility networks," *Proceedings of the National Academy of Sciences*, vol. 117, no. 52, pp. 32 883–32 890, 2020. DOI: 10.1073/pnas.2012326117.

[29] H. Lau *et al.*, "The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China," *Journal of Travel Medicine*, vol. 27, no. 3, taaa037, 2020. DOI: 10.1093/jtm/taaa037.

[30] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?" *The Lancet*, vol. 395, no. 10228, pp. 931–934, 2020. DOI: 10.1016/S0140-6736(20)30567-5.

[31] J. Caulkins *et al.*, "How long should the COVID-19 lockdown continue?" *PLOS ONE*, vol. 15, no. 12, e0243413, 2020. DOI: 10.1371/journal.pone.0243413.

[32] S. H. Shahidi, J. S. Williams, and F. Hassani, "Physical activity during COVID-19 quarantine," *Acta Paediatrica (Oslo, Norway: 1992)*, vol. 109, no. 10, pp. 2147–2148, Oct. 2020. DOI: 10.1111/apa.15420.

[33] C. R. Wells *et al.*, "Optimal COVID-19 quarantine and testing strategies," *Nature Communications*, vol. 12, no. 1, p. 356, 2021. DOI: 10.1038/s41467-020-20742-8.

[34] F. Piguillem and L. Shi, "Optimal COVID-19 quarantine and testing policies," Einaudi Institute for Economics and Finance (EIEF), EIEF Working Papers Series 2004, Apr. 2020. [Online]. Available: https://ideas.repec.org/p/eie/wpaper/2004.html.

[35] A. Scala, "The mathematics of multiple lockdowns," *Scientific Reports*, vol. 11, no. 1, p. 8078, 2021. DOI: 10.1038/s41598-021-87556-6.

[36] M. Shen *et al.*, "Projected COVID-19 epidemic in the United States in the context of the effectiveness of a potential vaccine and implications for social distancing and face mask use," *Vaccine*, vol. 39, no. 16, pp. 2295–2302, 2021. DOI: 10.1016/j.vaccine.2021.02.056.

[37] J. H. Buckner, G. Chowell, and M. R. Springborn, "Dynamic prioritization of COVID-19 vaccines when social distancing is limited for essential workers," *Proceedings of the National Academy of Sciences*, vol. 39, no. 16, pp. 2295–2302, Apr. 2021. DOI: 10.1016/j.vaccine.2021.02.056.

[38] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal," *BMJ*, vol. 369, no. m1328, Apr. 2020. DOI: 10.1136/bmj.m1328.

[39] Y. Xiang, Y. Jia, L. Chen, L. Guo, B. Shu, and E. Long, "COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models," *Infectious Disease Modelling*, vol. 6, pp. 324–342, 2021. DOI: 10.1016/j.idm.2021.01.001.

[40] I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," *Neural Computing and Applications*, pp. 1–11, 2021. DOI: 10.1007/s00521-020-05626-8.

[41] W. T. Li *et al.*, "Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 247, 2020. DOI: 10.1186/s12911-020-01266-z.

[42] H. B. Syeda *et al.*, "Role of machine learning techniques to tackle the COVID-19 crisis: Systematic review," *JMIR Medical Informatics*, vol. 9, no. 1, e23811, 2021. DOI: 10.2196/23811.

[43] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review," *Chaos, Solitons & Fractals*, vol. 139, p. 110 059, 2020. DOI: 10.1016/j.chaos.2020.110059.

[44] I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham, "Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5330, 2020. DOI: 10.3390/ijerph17155330.

[45] A.-E. Hassanien, N. Dey, and S. Elghamrawy, Eds., *Big data analytics and artificial intelligence against COVID-19: innovation vision and approach*. Springer Nature, 2020, vol. 78. DOI: 10.1007/978-3-030-55258-9.

[46] J. Wu, J. Wang, S. Nicholas, E. Maitland, and Q. Fan, "Application of big data technology for COVID-19 prevention and control in China: Lessons and recommendations," *Journal of Medical Internet Research*, vol. 22, no. 10, e21980, 2020. DOI: 10.2196/21980.

[47] J. Blumenstock, "Machine learning can help get COVID-19 aid to those who need it most," *Nature*, 2020. DOI: 10.1038/d41586-020-01393-7.

[48] Apple Inc. and Google Inc., *Privacy-preserving contact tracing*. [Online]. Available: https://covid19.apple.com/contacttracing.

[49] C. L. Niedzwiedz *et al.*, "Mental health and health behaviours before and during the initial phase of the COVID-19 lockdown: Longitudinal analyses of the UK household longitudinal study," *Journal of epidemiology and community health*, vol. 75, no. 3, pp. 224–231, 2021. DOI: 10.1136/jech-2020-215060.

[50] C. Pieh, S. Budimir, J. Delgadillo, M. Barkham, J. R. Fontaine, and T. Probst, "Mental health during COVID-19 lockdown in the United Kingdom," *Psychosomatic Medicine*, vol. 83, no. 4, pp. 328–337, 2021. DOI: 10.1097/PSY.0000000000000871.

[51] J. Du *et al.*, "Mental health burden in different professions during the final stage of the COVID-19 lockdown in China: Cross-sectional survey study," *Journal of Medical Internet Research*, vol. 23, no. 1, e24240, 2021. DOI: 10.2196/24983.

[52] J. Murphy *et al.*, "Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom," *Nature Communications*, vol. 12, no. 1, p. 29, 2021. DOI: 10.1038/s41467-020-20226-9.

[53] S. Wood and K. Schulman, "Beyond politics—promoting COVID-19 vaccination in the United States," *New England Journal of Medicine*, vol. 384, no. 7, e23, 2021. DOI: 10.1056/NEJMms2033790.

[54] W.-Y. S. Chou and A. Budenz, "Considering emotion in COVID-19 vaccine communication: Addressing vaccine hesitancy and fostering vaccine confidence," *Health Communication*, vol. 35, no. 14, pp. 1718–1722, 2020. DOI: 10.1080/10410236.2020.1838096.

[55] J. Lopez Bernal *et al.*, "Effectiveness of COVID-19 vaccines against the B.1.617.2 (Delta) variant," *New England Journal of Medicine*, vol. 385, no. 7, pp. 585–594, 2021. DOI: 10.1056/NEJMoa2108891.

[56] K. Kupferschmidt and M. Wadman, "Delta variant triggers new phase in the pandemic," *Science*, vol. 372, no. 6549, pp. 1375–1376, 2021. DOI: 10.1126/science.372.6549.1375.

[57] T. Farinholt *et al.*, "Transmission event of SARS-CoV-2 Delta variant reveals multiple vaccine breakthrough infections," *BMC Medicine*, vol. 19, no. 1, p. 255, 2021. DOI: 10.1186/s12916-021-02103-4.

[58] C. Davis *et al.*, "Reduced neutralisation of the Delta (B.1.617.2) SARS-CoV-2 variant of concern following vaccination," *PLOS Pathogens*, vol. 17, no. 12, e1010022, 2021. DOI: 10.1371/journal.ppat.1010022.

[59] J. M. Musser *et al.*, "Delta variants of SARS-CoV-2 cause significantly increased vaccine breakthrough COVID-19 cases in Houston, Texas," *The American journal of pathology*, vol. 192, no. 2, pp. 320–331, 2022. DOI: 10.1016/j.ajpath.2021.10.019.

[60] M. Keeling, "The effects of local spatial structure on epidemiological invasions," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 266, no. 1421, pp. 859–867, 1999. DOI: 10.1098/rspb.1999.0716.

[61] M. J. Keeling and K. T. Eames, "Networks and epidemic models," *Journal of the Royal Society Interface*, vol. 2, no. 4, pp. 295–307, 2005. DOI: 10.1098/rsif.2005.0051.

[62] M. Keeling, "The implications of network structure for epidemic dynamics," *Theoretical Population Biology*, vol. 67, no. 1, pp. 1–8, 2005. DOI: 10.1016/j.tpb.2004.08.002.

[63] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008. DOI: 10.1515/9781400841035.

[64] M. Gambhir, T. A. Clark, S. Cauchemez, S. Y. Tartof, D. L. Swerdlow, and N. M. Ferguson, "A change in vaccine efficacy and duration of protection explains recent rises in pertussis incidence in the United States," *PLOS Computational Biology*, vol. 11, no. 4, e1004138, 2015. DOI: 10.1371/journal.pcbi.1004138.

[65] S. L. Sheridan, K. Frith, T. L. Snelling, K. Grimwood, P. B. McIntyre, and S. B. Lambert, "Waning vaccine immunity in teenagers primed with whole cell and acellular pertussis vaccine: Recent epidemiology," *Expert Review of Vaccines*, vol. 13, no. 9, pp. 1081–1106, 2014. DOI: 10.1586/14760584.2014.944167.

[66] Q. Cassam, "Misunderstanding vaccine hesitancy: A case study in epistemic injustice," *Educational Philosophy and Theory*, pp. 1–15, 2021. DOI: 10.1080/00131857.2021.2006055.

[67] A. L. Hsu and R. Trotman, "Overcoming vaccine hesitancy, skepticism, and hostility in Missouri," *Missouri Medicine*, vol. 118, no. 5, pp. 396–400, 2021.

[68] K. Farrahi, R. Emonet, and M. Cebrian, "Epidemic contact tracing via communication traces," *PLOS ONE*, vol. 9, no. 5, 2014. DOI: 10.1371/journal.pone.0095133.

[69] S. Kojaku, L. Hébert-Dufresne, E. Mones, S. Lehmann, and Y. Y. Ahn, "The effectiveness of backward contact tracing in networks," *Nature Physics*, vol. 17, no. 5, pp. 652–658, 2021. DOI: 10.1038/s41567-021-01187-2.

[70] I. Braithwaite, T. Callender, M. Bullock, and R. W. Aldridge, "Automated and partly automated contact tracing: A systematic review to inform the control of COVID-19," *The Lancet Digital Health*, vol. 2, no. 11, e607–e621, 2020. DOI: 10.1016/S2589-7500(20)30184-9.

[71] K. Leuzinger *et al.*, "Epidemiology and precision of SARS-CoV-2 detection following lockdown and relaxation measures," *Journal of Medical Virology*, vol. 93, no. 4, pp. 2374–2384, 2021. DOI: 10.1002/jmv.26731.

[72] S. Bandyopadhyay, K. Chatterjee, K. Das, and J. Roy, "Learning versus habit formation: Optimal timing of lockdown for disease containment," *Journal of Mathematical Economics*, vol. 93, p. 102 452, 2021. DOI: 10.1016/j.jmateco.2020.11.008.

[73] D. G. Bausch, "Precision physical distancing for COVID-19: An important tool in unlocking the lockdown," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 1, pp. 22–24, 2020. DOI: 10.4269/ajtmh.20-0359.

[74] J. Harrison, S. Berry, V. Mor, and D. Gifford, "'Somebody like me': Understanding COVID-19 vaccine hesitancy among staff in skilled nursing facilities," *Journal of the American Medical Directors Association*, vol. 22, no. 6, pp. 1133–1137, 2021. DOI: 10.1016/j.jamda.2021.03.012.

[75] T. Callaghan *et al.*, "Correlates and disparities of COVID-19 vaccine hesitancy," *Social Science & Medicine*, vol. 272, p. 113 638, 2021. DOI: 10.1016/j.socscimed.2020.113638.

[76] E. Robertson *et al.*, "Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study," *Brain, Behavior, and Immunity*, vol. 94, pp. 41–50, 2021. DOI: 10.1016/j.bbi.2021.03.008.

[77] E. Vynnycky and R. White, *An introduction to infectious disease modelling*. OUP Oxford, 2010. DOI: 10.1017/S0950268811000422.

[78] A. Gelb, *Applied Optimal Estimation*. MIT press, 1974.

[79] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1, pp. 5–43, 2003. DOI: 10.1023/A:1020281327116.

[80] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *The Oxford Handbook of Nonlinear Filtering*, 656-704, Oxford University Press, 2009.

[81] A. Doucet, N. De Freitas, and N. Gordon, "Sequential monte carlo methods in practice. statistics for engineering and information science," in, A. Doucet, N. De Freitas, and N. Gordon, Eds. Springer, 2001, ch. An introduction to sequential Monte Carlo methods, pp. 3–14. DOI: 10.1007/978-1-4757-3437-9_1.

[82] S. Eker, "Validity and usefulness of COVID-19 models," *Humanities and Social Sciences Communications*, vol. 7, no. 1, p. 54, 2020. DOI: 10.1057/s41599-020-00553-4.

[83] C. S. Currie *et al.*, "How simulation modelling can help reduce the impact of COVID-19," *Journal of Simulation*, vol. 14, no. 2, pp. 83–97, 2020. DOI: 10.1080/17477778.2020.1751570.

[84] F. Della Rossa *et al.*, "A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic," *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020. DOI: 10.1038/s41467-020-18827-5.

[85] P. J. Turk *et al.*, "Modeling COVID-19 latent prevalence to assess a public health intervention at a state and regional scale: Retrospective cohort study," *JMIR Public Health and Surveillance*, vol. 6, no. 2, e19353, 2020. DOI: 10.2196/19353.

[86] R. Li *et al.*, "Global COVID-19 pandemic demands joint interventions for the suppression of future waves," *Proceedings of the National Academy of Sciences*, vol. 117, no. 42, pp. 26 151–26 157, 2020. DOI: 10.1073/pnas.2012002117.

[87] R. Carli, G. Cavone, N. Epicoco, P. Scarabaggio, and M. Dotoli, "Model predictive control to mitigate the COVID-19 outbreak in a multi-region scenario," *Annual Reviews in Control*, vol. 50, pp. 373–393, 2020. DOI: 10.1016/j.arcontrol.2020.09.005.

[88] Y. Zhang, B. Jiang, J. Yuan, and Y. Tao, "The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: A data-driven SEIQR model study," *MedRxiv*, 2020. DOI: 10.1101/2020.03.04.20031187.

[89] M. Saez, A. Tobias, D. Varga, and M. A. Barceló, "Effectiveness of the measures to flatten the epidemic curve of COVID-19. the case of Spain," *The Science of the Total Environment*, vol. 727, p. 138 761, 2020. DOI: 10.1016/j.scitotenv.2020.138761.

[90] L. Thunström, S. C. Newbold, D. Finnoff, M. Ashworth, and J. F. Shogren, "The benefits and costs of using social distancing to flatten the curve for COVID-19," *Journal of Benefit-Cost Analysis*, vol. 11, no. 2, pp. 179–195, 2020. DOI: 10.1017/bca.2020.12.

[91] H. S. Hurd and J. B. Kaneene, "The application of simulation models and systems analysis in epidemiology: A review," *Preventive Veterinary Medicine*, vol. 15, no. 2-3, pp. 81–99, 1993. DOI: 10.1016/0167-5877(93)90105-3.

[92] T. Zhang, M. Lees, C. K. Kwoh, X. Fu, G. K. K. Lee, and R. S. M. Goh, "A contact-network-based simulation model for evaluating interventions under 'what-if' scenarios in epidemic," in *Proceedings of the 2012 Winter Simulation Conference (WSC)*, IEEE, IEEE, 2012, pp. 1–12. DOI: 10.1109/WSC.2012.6465056.

[93] J. Banks, "Introduction to simulation," in *Proceedings of the 31st Conference on Winter Simulation: Simulation—a Bridge to the Future - Volume 1*, ser. WSC '99, 1999, pp. 7–13. DOI: 10.1145/324138.324142.

[94] G. E. Mobus and M. C. Kalton, *Principles of Systems Science*, 1st ed., ser. Understanding Complex Systems. Springer, 2015. DOI: 10.1007/978-1-4939-1920-8.

[95] H. Rahmandad and J. Sterman, "Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models," *Management Science*, vol. 54, no. 5, pp. 998–1014, 2008. DOI: 10.1287/mnsc.1070.0787.

[96] H. J. Wearing, P. Rohani, and M. J. Keeling, "Appropriate models for the management of infectious diseases," *PLOS Medicine*, vol. 2, no. 7, e174, 2005. DOI: 10.1371/journal.pmed.0020174.

[97] Government of Singapore, *TraceTogether, safer together*. [Online]. Available: https://www.tracetogether.gov.sg/.

[98] Government of Canada, *COVID Alert*. [Online]. Available: https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html.

[99] Public Health Ontario, *COVID-19 contact tracing initiative*. [Online]. Available: https://www.publichealthontario.ca/en/diseases-and-conditions/infectious-diseases/respiratory-diseases/novel-coronavirus/contact-tracing-initiative.

[100] D. of Health and S. C. of UK, *Next phase of NHS coronavirus (COVID-19) app announced*. [Online]. Available: https://www.gov.uk/government/news/next-phase-of-nhs-coronavirus-covid-19-app-announced.

[101] M. J. Parker, C. Fraser, L. Abeler-Dörner, and D. Bonsall, "Ethics of instantaneous contact tracing using mobile phone apps in the control of the COVID-19 pandemic," *Journal of Medical Ethics*, vol. 46, no. 7, pp. 427–431, 2020. DOI: 10.1136/medethics-2020-106314.

[102] M. E. Kretzschmar, G. Rozhnova, M. C. Bootsma, M. van Boven, J. H. van de Wijgert, and M. J. Bonten, "Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study," *The Lancet Public Health*, vol. 5, no. 8, e452–e459, 2020. DOI: 10.1016/S2468-2667(20)30157-2.

[103] J. Bell, D. Butler, C. Hicks, and J. Crowcroft, "Tracesecure: Towards privacy preserving contact tracing," *arXiv*, Apr. 2020. DOI: 10.48550/arXiv.2004.04059.

[104] F. Rowe, "Contact tracing apps and values dilemmas: A privacy paradox in a neo-liberal world," *International Journal of Information Management*, vol. 55, p. 102 178, 2020. DOI: 10.1016/j.ijinfomgt.2020.102178.

[105] N. Ahmed *et al.*, "A survey of COVID-19 contact tracing apps," *IEEE Access*, vol. 8, pp. 134 577–134 601, 2020. DOI: 10.1109/ACCESS.2020.3010226..

[106] T. M. Yasaka, B. M. Lehrich, and R. Sahyouni, "Peer-to-peer contact tracing: Development of a privacy-preserving smartphone app," *JMIR mHealth and uHealth*, vol. 8, no. 4, e18936, 2020. DOI: 10.2196/18936.

[107] H. Cho, D. Ippolito, and Y. W. Yu, "Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs," *arXiv*, 2020. DOI: 10.48550/arXiv.2003.11511.

[108] K. Kreuger and N. Osgood, "Particle filtering using agent-based transmission models," in *2015 Winter Simulation Conference (WSC)*, IEEE, 2015, pp. 737–747. DOI: 10.1109/WSC.2015.7408211.

[109] M. Hashemian, W. Qian, K. G. Stanley, and N. D. Osgood, "Temporal aggregation impacts on epidemiological simulations employing microcontact data," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 132, 2012. DOI: 10.1186/1472-6947-12-132.

[110] P. Beutels, Z. Shkedy, M. Aerts, and P. Van Damme, "Social mixing patterns for transmission models of close contact infections: Exploring self-evaluation and diary-based data collection through a web-based interface," *Epidemiology & Infection*, vol. 134, no. 6, pp. 1158–1166, 2006. DOI: 10.1017/S0950268806006418.

[111] E. Yoneki, "The importance of data collection for modelling contact networks," in *2009 International Conference on Computational Science and Engineering*, IEEE, vol. 4, 2009, pp. 940–943. DOI: 10.1109/CSE.2009.332.

[112] N. Osgood and J. Liu, "Bayesian parameter estimation of System Dynamics models using Markov chain Monte Carlo methods: An informal introduction," in *Proceedings of the 30th International Conference of the System Dynamics Society*, 2013, pp. 22–26.

[113] N. Osgood and J. Liu, "Towards closed loop modeling: Evaluating the prospects for creating recurrently regrounded aggregate simulation models using particle filtering," in *Proceedings of the Winter Simulation Conference 2014*, IEEE, 2014, pp. 829–841. DOI: 10.1109/WSC.2014.7019944.

[114] R. Oraji, V. H. Hoeppner, A. Safarishahrbijari, and N. D. Osgood, "Combining particle filtering and transmission modeling for TB control," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, 2016, pp. 392–398. DOI: 10.1109/ICHI.2016.70.

[115] A. Safarishahrbijari, T. Lawrence, R. Lomotey, J. Liu, C. Waldner, and N. Osgood, "Particle filtering in a SEIRV simulation model of H1N1 influenza," in *2015 Winter Simulation Conference (WSC)*, IEEE, 2015, pp. 1240–1251. DOI: 10.1109/WSC.2015.7408249.

[116] X. Li, A. Doroshenko, and N. D. Osgood, "Applying particle filtering in both aggregated and age-structured population compartmental models of pre-vaccination measles," *PLOS ONE*, vol. 13, no. 11, e0206529, 2018. DOI: 10.1371/journal.pone.0206529.

[117] A. Safarishahrbijari, A. Teyhouee, C. Waldner, J. Liu, and N. D. Osgood, "Predictive accuracy of particle filtering in dynamic models supporting outbreak projections," *BMC Infectious Diseases*, vol. 17, no. 1, p. 648, 2017. DOI: 10.1186/s12879-017-2726-9.

[118] A. Mohammadbagheri, C. Lillas, and N. D. Osgood, "Mathematical modeling of HPA axis using particle filter algorithm," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, 2018, pp. 400–402. DOI: 10.1109/ICHI.2018.00073.

[119] X. Li *et al.*, "Illuminating the hidden elements and future evolution of opioid abuse using dynamic modeling, big data and particle Markov chain Monte Carlo," in *11th International Conference, SBP-BRiMS 2018, Washington, DC, USA*, July 10-13, 2018. [Online]. Available: http://sbp-brims.org/2018/proceedings/papers/challenge_papers/Illuminating%20the%20HiddenElements.pdf.

[120] J. Almagor and S. Picascia, "Exploring the effectiveness of a COVID-19 contact tracing app using an agent-based model," *Scientific Reports*, vol. 10, no. 1, p. 22 235, 2020. DOI: 10.1038/s41598-020-79000-y.

[121] T. Paul, "Modeling human mobility entropy as a function of spatial and temporal quantizations," Ph.D. dissertation, PhD thesis, University of Saskatchewan, 2017. [Online]. Available: http://hdl.handle.net/10388/7777.

[122] U. of Saskatchewan, *Manuscript-style theses and dissertations*, Sep. 2021. [Online]. Available: https://web.archive.org/web/20210630172436/https://students.usask.ca/graduate/manuscript-style.php.

[123] W. Qian, N. D. Osgood, and K. G. Stanley, "Integrating Epidemiological Modeling and Surveillance Data Feeds: a Kalman Filter Based Approach," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer, 2014, pp. 145–152. DOI: 10.1007/978-3-319-05579-4_18.

[124] W. Qian, A. Cooke, K. G. Stanley, and N. D. Osgood, "Comparing Contact Tracing Through Bluetooth and GPS Surveillance Data," *Submitted to the Journal of Medical Internet Research*, Apr. 2022.

[125] W. Qian, K. G. Stanley, and N. D. Osgood, "Impacts of observation frequency on Reconstruction of Close-proximity Contact Networks and Modeled Transmission Dynamics," *Submitted to the PLOS Computational Biology*, May 2022.

[126] D. A. Marshall *et al.*, "Applying dynamic simulation modeling methods in health care delivery research—the SIMULATE checklist: Report of the ISPOR simulation modeling emerging good practices task force," *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, vol. 18, no. 1, pp. 5–16, 2015. DOI: 10.1016/j.jval.2014.12.001.

[127] A. Maria, "Introduction to modeling and simulation," in *Proceedings of the 29th Conference on Winter Simulation*, 1997, pp. 7–13. DOI: 10.1145/268437.268440.

[128] 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. DOI: 10.1098/rspa.1927.0118.

[129] R. Ross and H. P. Hudson, "An application of the theory of probabilities to the study of a priori pathometry—Part II," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 93, no. 650, pp. 212–225, 1917. DOI: 10.1098/rspa.1917.0014.

[130] R. Ross, "An application of the theory of probabilities to the study of a priori pathometry—Part I," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 92, no. 638, pp. 204–230, 1916. DOI: 10.1098/rspa.1916.0007.

[131] S. Towers, K. V. Geisse, Y. Zheng, and Z. Feng, "Antiviral treatment for pandemic influenza: Assessing potential repercussions using a seasonally forced SIR model," *Journal of Theoretical Biology*, vol. 289, pp. 259–268, 2011. DOI: 10.1016/j.jtbi.2011.08.011.

[132] A. Huppert and G. Katriel, "Mathematical modelling and prediction in infectious disease epidemiology," *Clinical Microbiology and Infection*, vol. 19, no. 11, pp. 999–1005, 2013. DOI: 10.1111/1469-0691.12308.

[133] I. Cooper, A. Mondal, and C. G. Antonopoulos, "A SIR model assumption for the spread of COVID-19 in different communities," *Chaos, Solitons & Fractals*, vol. 139, p. 110 057, 2020. DOI: 10.1016/j.chaos.2020.110057.

[134] W. Qian, N. D. Osgood, and K. G. Stanley, "Integrating epidemiological modeling and surveillance data feeds: A Kalman filter based approach," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer, 2014, pp. 145–152. DOI: 10.1007/978-3-319-05579-4_18.

[135] H. W. Hethcote, "Qualitative analyses of communicable disease models," *Mathematical Biosciences*, vol. 28, no. 3-4, pp. 335–356, 1976. DOI: 10.1016/0025-5564(76)90132-2.

[136] O. Diekmann, J. A. P. Heesterbeek, and J. A. Metz, "On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations," *Journal of Mathematical Biology*, vol. 28, no. 4, pp. 365–382, 1990. DOI: 10.1007/BF00178324.

[137] D. Adam, "A guide to R—the pandemic's misunderstood metric," *Nature*, vol. 583, no. 7816, pp. 346–348, 2020. DOI: 10.1038/d41586-020-02009-w.

[138] P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, and K. H. Jacobsen, "Complexity of the basic reproduction number (R0)," *Emerging Infectious Diseases*, vol. 25, no. 1, pp. 1–4, 2019. DOI: 10.3201/eid2501.171901.

[139] H. Nishiura and G. Chowell, "The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends," in *Mathematical and Statistical Estimation Approaches in Epidemiology*, Springer, 2009, pp. 103–121. DOI: 10.1007/978-90-481-2313-1_5.

[140] K. R. M. Jeffrey K Aronson Jon Brassey, "'when will it be over?': An introduction to viral reproduction numbers, R0 and Re," Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, Tech. Rep., Apr. 2020. [Online]. Available: https://www.cebm.net/covid-19/when-will-it-be-over-an-introduction-to-viral-reproduction-numbers-r0-and-re/.

[141] M. Tomochi and M. Kono, "A mathematical model for COVID-19 pandemic—SIIR model: Effects of asymptomatic individuals," *Journal of General and Family Medicine*, vol. 22, no. 1, pp. 5–14, 2021. DOI: 10.1002/jgf2.382.

[142] A. Arenas *et al.*, "Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions," *Physical Review X*, vol. 10, no. 4, p. 041 055, 2020. DOI: 10.1103/PhysRevX.10.041055.

[143] I. Holmdahl and C. Buckee, "Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us," *New England Journal of Medicine*, vol. 383, no. 4, pp. 303–305, 2020. DOI: 10.1056/NEJMp2016822.

[144] K. Prem *et al.*, "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study," *The Lancet Public Health*, vol. 5, no. 5, e261–e270, 2020. DOI: 10.1016/S2468-2667(20)30073-6.

[145] X. Wang *et al.*, "Impact of social distancing measures on coronavirus disease healthcare demand, central Texas, USA," *Emerging Infectious Diseases*, vol. 26, no. 10, pp. 2361–2369, 2020. DOI: 10.3201/eid2610.201702.

[146] J. Zhang *et al.*, "Age profile of susceptibility, mixing, and social distancing shape the dynamics of the novel coronavirus disease 2019 outbreak in China," *medRxiv*, 2020. DOI: 10.1101/2020.03.19.20039107.

[147] J. Zhang *et al.*, "Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China," *Science*, vol. 368, no. 6498, pp. 1481–1486, 2020. DOI: 10.1126/science.abb8001.

[148] J. Hilton and M. J. Keeling, "Estimation of country-level basic reproductive ratios for novel coronavirus (SARS-CoV-2/COVID-19) using synthetic contact matrices," *PLOS Computational Biology*, vol. 16, no. 7, e1008031, 2020. DOI: 10.1371/journal.pcbi.1008031.

[149] R. Singh and R. Adhikari, "Age-structured impact of social distancing on the COVID-19 epidemic in India," *arXiv*, 2020.

[150] E. P. Fenichel *et al.*, "Adaptive human behavior in epidemiological models," *Proceedings of the National Academy of Sciences*, vol. 108, no. 15, pp. 6306–6311, 2011. DOI: 10.1073/pnas.1011250108.

[151] F. Verelst, L. Willem, and P. Beutels, "Behavioural change models for infectious disease transmission: A systematic review (2010–2015)," *Journal of The Royal Society Interface*, vol. 13, no. 125, p. 20 160 820, 2016. DOI: 10.1098/rsif.2016.0820.

[152] Z. Wang, M. A. Andrews, Z.-X. Wu, L. Wang, and C. T. Bauch, "Coupled disease–behavior dynamics on complex networks: A review," *Physics of life reviews*, vol. 15, pp. 1–29, 2015. DOI: 10.1016/j.plrev.2015.07.006.

[153] Y. Tian and N. Osgood, "Comparison between individual-based and aggregate models in the context of tuberculosis transmission," in *Proceedings, the 29th International Conference of the System Dynamics Society*, 2011, pp. 10–025.

[154] H. V. D. Parunak, R. Savit, and R. L. Riolo, "Agent-based modeling vs. equation-based modeling: A case study and users' guide," in *Multi-Agent Systems and Agent-Based Simulation*, J. S. Sichman, R. Conte, and N. Gilbert, Eds., Springer, vol. 1534, Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 10–25. DOI: 10.1007/10692956_2.

[155] N. Osgood and G. Kaufman, "A hybrid model architecture for strategic renewable resource planning," *New York: System Dynamics Organization*, 2003.

[156] L. K. Kreuger, W. Qian, N. Osgood, and K. Choi, "Agile design meets hybrid models: Using modularity to enhance hybrid model design and use," in *2016 Winter Simulation Conference (WSC)*, IEEE, 2016, pp. 1428–1438. DOI: 10.1109/WSC.2016.7822195.

[157] A. Gao, N. D. Osgood, W. An, and R. F. Dyck, "A tripartite hybrid model architecture for investigating health and cost impacts and intervention tradeoffs for diabetic end-stage renal disease," in *Proceedings of the Winter Simulation Conference 2014*, IEEE, 2014, pp. 1676–1687. DOI: 10.1109/WSC.2014.7020018.

[158] Y. Qin, L. Freebairn, J.-A. Atkinson, W. Qian, A. Safarishahrbijari, and N. D. Osgood, "Multi-scale simulation modeling for prevention and public health management of diabetes in pregnancy and sequelae," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Springer, vol. 11549, 2019, pp. 256–265. DOI: 10.1007/978-3-030-21741-9_26.

[159] T. Hardy, E. Abu-Raddad, N. Porksen, and A. De Gaetano, "Evaluation of a mathematical model of diabetes progression against observations in the diabetes prevention program," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 303, no. 2, E200–E212, 2012. DOI: 10.1152/ajpendo.00421.2011.

[160] J. B. S. Ong *et al.*, "Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore," *PLOS ONE*, vol. 5, no. 4, e10036, 2010. DOI: 10.1371/journal.pone.0010036.

[161] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002. DOI: 10.1109/78.978374.

[162] C. J. Panetta, "Network of unmanned surface vehicles: Design and application to target tracking," M.S. thesis, Michigan State University, 2021. [Online]. Available: https://www.proquest.com/docview/2486549245.

[163] P. S.-h. Won, M. Biglarbegian, and W. Melek, "Development and performance comparison of extended kalman filter and particle filter for self-reconfigurable mobile robots," in *2014 IEEE Symposium on Robotic Intelligence in Informationally Structured Space (RiiSS)*, IEEE, 2014, pp. 1–6. DOI: 10.1109/RIISS.2014.7009168.

[164] S. Koyama, L. Castellanos Pérez-Bolde, C. R. Shalizi, and R. E. Kass, "Approximate methods for state-space models," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 170–180, 2010. DOI: 10.1198/jasa.2009.tm08326.

[165] T. Minka, "From hidden Markov models to linear dynamical systems," Technical report, MIT, Tech. Rep., 1999. [Online]. Available: https://www.media.mit.edu/publications/from-hidden-markov-models-to-linear-dynamical-systems-2/.

[166] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986. DOI: 10.1109/MASSP.1986.1165342.

[167] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[168] H. Barbosa *et al.*, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018. DOI: 10.1016/j.physrep.2018.01.001.

[169] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009. DOI: 10.1073/pnas.0906910106.

[170] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008. DOI: 10.1038/nature06958.

[171] E. Hernández-Orallo and A. Armero-Martıénez, "How human mobility models can help to deal with COVID-19," *Electronics*, vol. 10, no. 1, p. 33, 2021. DOI: 10.3390/electronics10010033.

[172] M. U. Kraemer *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. DOI: 10.1126/science.abb4218.

[173] A. Scala *et al.*, "Time, space and social interactions: Exit mechanisms for the COVID-19 epidemics," *Scientific Reports*, vol. 10, no. 1, p. 13 764, 2020. DOI: 10.1038/s41598-020-70631-9.

[174] A. Vespignani, "Predicting the behavior of techno-social systems," *Science*, vol. 325, no. 5939, pp. 425–428, 2009. DOI: 10.1126/science.1171990.

[175] A. Smith, H. Balakrishnan, M. Goraczko, and N. Priyantha, "Tracking moving devices with the cricket location system," in *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '04, 2004, pp. 190–202. DOI: 10.1145/990064.990088.

[176] P. Hui, J. Crowcroft, and E. Yoneki, "BUBBLE rap: Social-based forwarding in delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, 2010. DOI: 10.1109/TMC.2010.246.

[177] S. Havlin and D. Ben-Avraham, "Diffusion in disordered media," *Advances in Physics*, vol. 36, no. 6, pp. 695–798, 1987. DOI: 10.1080/00018738700101072.

[178] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '04, vol. 34, Association for Computing Machinery, 2004, pp. 145–158. DOI: 10.1145/1015467.1015484.

[179] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed diffusion for wireless sensor networking," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 2–16, 2003. DOI: 10.1109/TNET.2002.808417.

[180] C. Konstantopoulos, A. Mpitziopoulos, D. Gavalas, and G. Pantziou, "Effective determination of mobile agent itineraries for data aggregation on sensor networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 12, pp. 1679–1693, 2010. DOI: 10.1109/TKDE.2009.203.

[181] W. Ye, J. Heidemann, and D. Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 3, pp. 493–506, 2004. DOI: 10.1109/TNET.2004.828953.

[182] J.-S. Lee, "Performance evaluation of IEEE 802.15.4 for low-rate wireless personal area networks," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 742–749, 2006. DOI: 10.1109/TCE.2006.1706465.

[183] M. W. Horner and M. E. O'Kelly, "Embedding economies of scale concepts for hub network design," *Journal of Transport Geography*, vol. 9, no. 4, pp. 255–265, 2001.

[184] R. Kitamura, C. Chen, R. M. Pendyala, and R. Narayanan, "Micro-simulation of daily activity-travel patterns for travel demand forecasting," *Transportation*, vol. 27, no. 1, pp. 25–51, 2000. DOI: 10.1023/A:1005259324588.

[185] G. McKenzie, "Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services," *Computers, Environment and Urban Systems*, vol. 79, p. 101 418, 2020. DOI: 10.1016/j.compenvurbsys.2019.101418.

[186] J. Visser, T. Nemoto, and M. Browne, "Home delivery and the impacts on urban freight transport: A review," *Procedia-social and behavioral sciences*, vol. 125, pp. 15–27, 2014. DOI: 10.1016/j.sbspro.2014.01.1452.

[187] C.-I. Hsu, S.-F. Hung, and H.-C. Li, "Vehicle routing problem with time-windows for perishable food delivery," *Journal of Food Engineering*, vol. 80, no. 2, pp. 465–475, 2007. DOI: 10.1016/j.jfoodeng.2006.05.029.

[188] L.-W. Chen, "Impact assessment of food delivery on urban traffic," in *2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, IEEE, 2019, pp. 236–241. DOI: 10.1109/SOLI48380.2019.8955108.

[189] A. Albalawi, C. Hambly, and J. R. Speakman, "The impact of the novel coronavirus movement restrictions in the UK on food outlet usage and body mass index," *Obesity Science & Practice*, vol. 7, no. 3, pp. 302–306, 2021. DOI: 10.1002/osp4.477.

[190] J. E. Hobbs, "Food supply chains during the COVID-19 pandemic," *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, vol. 68, no. 2, pp. 171–176, 2020. DOI: 10.1111/cjag.12237.

[191] S. M. Zahraei, J. H. Kurniawan, and L. Cheah, "A foresight study on urban mobility: Singapore in 2040," *Foresight*, vol. 22, no. 1, pp. 37–52, 2020. DOI: 10.1108/FS-05-2019-0044.

[192] N. Masuda and P. Holme, "Predicting and controlling infectious disease epidemics using temporal networks," *F1000Prime Reports*, vol. 5, no. 6, 2013. DOI: 10.12703/P5-6.

[193] N. Masuda, J. C. Miller, and P. Holme, "Concurrency measures in the era of temporal network epidemiology: A review," *Journal of The Royal Society Interface*, vol. 18, no. 179, p. 20 210 019, 2021. DOI: 10.1098/rsif.2021.0019.

[194] R. M. May, "Network structure and the biology of populations," *Trends in Ecology & Evolution*, vol. 21, no. 7, pp. 394–399, 2006. DOI: 10.1016/j.tree.2006.03.013.

[195] M. Newman, "Spread of epidemic disease on networks," *Physical Review E*, vol. 66, no. 1, p. 016 128, 2002. DOI: 10.1103/PhysRevE.66.016128.

[196] S. Redhu and R. M. Hegde, "Optimal relay node selection in time-varying IoT networks using apriori contact pattern information," *Ad Hoc Networks*, vol. 98, no. 1, p. 102 065, 2020. DOI: 10.1016/j.adhoc.2019.102065.

[197] J. Stehlé *et al.*, "Simulation of an seir infectious disease model on the dynamic contact network of conference attendees," *BMC Medicine*, vol. 9, no. 1741-7015, p. 87, 2011. DOI: 10.1186/1741-7015-9-87.

[198] P. Stroud, S. Sydoriak, J. Riese, J. Smith, S. Mniszewski, and P. Romero, "Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing," *Mathematical Biosciences*, vol. 203, no. 2, pp. 301–318, 2006. DOI: 10.1016/j.mbs.2006.01.007.

[199] A. Tuite *et al.*, "Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza," *Canadian Medical Association Journal*, vol. 182, no. 2, pp. 131–136, 2010. DOI: 10.1503/cmaj.091807.

[200] M. Campbell-Kelly, "Information technology and organizational change in the British census, 1801–1911," *Information Systems Research*, vol. 7, no. 1, pp. 22–36, 1996. DOI: 10.1287/isre.7.1.22.

[201] E. G. Ravenstein, "The laws of migration," *Journal of the Statistical Society of London*, vol. 48, no. 2, pp. 167–235, 1885. DOI: 10.2307/2979181.

[202] C. Lively, "Spatial mobility of the rural population with respect to local areas," *American Journal of Sociology*, vol. 43, no. 1, pp. 89–102, 1937.

[203] C. Hirschman and E. Mogford, "Immigration and the American Industrial Revolution from 1880 to 1920," *Social science research*, vol. 38, no. 4, pp. 897–920, 2009. DOI: 10.1016/j.ssresearch.2009.04.001.

[204] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–5, Jan. 2006, ISSN: 1476-4687. DOI: 10.1038/nature04292.

[205] G. Uhlenbeck and L. Ornstein, "On the theory of the Brownian motion," *Physical Review*, vol. 36, no. 5, p. 823, 1930. DOI: 10.1103/PhysRev.36.823.

[206] J. Klafter, M. Shlesinger, and G. Zumofen, "Beyond Brownian motion," *Physics Today*, vol. 49, no. 2, pp. 33–39, 1996. DOI: 10.1063/1.881487.

[207] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the National Academy of Sciences*, vol. 97, no. 21, pp. 11 149–11 152, 2000. DOI: 10.1073/pnas.200327197.

[208] B. Jensen, J. Larsen, L. Hansen, J. Larsen, and K. Jensen, "Predictability of mobile phone associations," in *21st European Conference on Machine Learning: Mining Ubiquitous and Social Environments Workshop*, 2010, pp. 91–105. [Online]. Available: https://eprints.gla.ac.uk/119589/.

[209] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. DOI: 10.1126/science.1177170.

[210] M. S. Hashemian, K. G. Stanley, D. L. Knowles, J. Calver, and N. D. Osgood, "Human network data collection in the wild: The epidemiological utility of micro-contact and location data," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 2012, pp. 255–264. DOI: 10.1145/2110363.2110394. [Online]. Available: IHI%20'12.

[211] W. Qian, K. G. Stanley, and N. D. Osgood, "The impact of spatial resolution and representation on human mobility predictability," in *Proceedings of the 12th International Conference on Web and Wireless Geographical Information Systems*, ser. W2GIS'13, Springer, 2013, pp. 25–40. DOI: 10.1007/978-3-642-37087-8_3.

[212] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnović, "Power law and exponential decay of intercontact times between mobile devices," *IEEE Transactions on Mobile Computing*, vol. 9, no. 10, pp. 1377–1390, 2010. DOI: 10.1109/TMC.2010.99.

[213] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, "Dynamics of person-to-person interactions from distributed RFID sensor networks," *PLOS ONE*, vol. 5, no. 7, 2010. DOI: 10.1371/journal.pone.0011596.

[214] M. Kassen, "A promising phenomenon of open data: A case study of the Chicago open data project," *Government Information Quarterly*, vol. 30, no. 4, pp. 508–513, 2013. DOI: 10.1016/j.giq.2013.05.012.

[215] C. W. Schmidt, "Trending now: Using social media to predict and track disease outbreaks," *Environmental Health Perspectives*, vol. 120, no. 1, a30–a33, 2012. DOI: 10.1289/ehp.120-a30.

[216] K. Siła-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demšar, and A. S. Fotheringham, "Analysis of human mobility patterns from gps trajectories and contextual information," *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 881–906, 2016. DOI: 10.1080/13658816.2015.1100731.

[217] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL'13, 2013, pp. 344–353. DOI: 10.1145/2525314.2525343.

[218] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003. DOI: 10.1016/S0306-4573(02)00021-3.

[219] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves, and V. M. Eguíluz, "Entangling mobility and interactions in social media," *PLOS ONE*, vol. 9, no. 3, e92196, 2014. DOI: 10.1371/journal.pone.0092196.

[220] L. Wu, Y. Zhi, Z. Sui, and Y. Liu, "Intra-urban human mobility and activity transition: Evidence from social media check-in data," *PLOS ONE*, vol. 9, no. 5, e97010, 2014. DOI: 10.1371/journal.pone.0097010.

[221] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang, "Semantic annotation of mobility data using social media," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15, 2015, pp. 1253–1263. DOI: 10.1145/2736277.2741675.

[222] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han, "GMove: Group-level mobility modeling using geo-tagged social media," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, 2016, pp. 1305–1314. DOI: 10.1145/2939672.2939793.

[223] X. Lu, B. Zheng, A. Velivelli, and C. Zhai, "Enhancing text categorization with semantic-enriched representation and training data augmentation," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 526–535, 2006. DOI: 10.1197/jamia.M2051.

[224] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," *arXiv*, 2018.

[225] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998. DOI: 10.1038/30918.

[226] K. Pelechrinis and P. Krishnamurthy, "Location-based social network users through a lense: Examining temporal user patterns," in *AAAI Fall Symposium - Technical Report*, vol. FS-12-, 2012, pp. 61–68. [Online]. Available: http://d-scholarship.pitt.edu/id/eprint/18809.

[227] J. Tauberer, *Open Government Data*, 2nd ed. 2014. [Online]. Available: https://opengovdata.io/.

[228] P. Murray-Rust, "Open data in science," *Nature Precedings*, pp. 1–1, 2008. DOI: 10.1038/npre.2008.1526.1.

[229] T. Berners-Lee, *Linked data*. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html.

[230] O. K. Foundation, *The open definition*, 2014. [Online]. Available: https://opendefinition.org/od/2.0/en/.

[231] M. Laessig, B. Jacob, and C. AbouZahr, "Opening data for global health," in *The Palgrave Handbook of Global Health Data Methods for Policy and Practice*, S. B. Macfarlane and C. AbouZahr, Eds. London: Palgrave Macmillan UK, 2019, pp. 451–468. DOI: 10.1057/978-1-137-54984-6_23.

[232] S. Chignard, *A brief history of open data.* [Online]. Available: http://www.paristechreview.com/2013/03/29/brief-history-open-data/.

[233] T. Davies, S. B. Walker, M. Rubinstein, and F. Perini, Eds., *The State of Open Data: Histories and Horizons.* African Minds, 2019. DOI: 10.5281/zenodo.2668475.

[234] J. Howe *et al.*, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006. [Online]. Available: https://www.wired.com/2006/06/crowds/.

[235] W. Willett, J. Heer, and M. Agrawala, "Strategies for crowdsourcing social data analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12, New York, NY, USA: Association for Computing Machinery, 2012, pp. 227–236. DOI: 10.1145/2207676.2207709.

[236] C. Heipke, "Crowdsourcing geospatial data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 550–557, 2010. DOI: 10.1016/j.isprsjprs.2010.06.005.

[237] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 901–911, 2016. DOI: 10.1109/TKDE.2016.2518669.

[238] G. M. Leung and K. Leung, "Crowdsourcing data to mitigate epidemics," *The Lancet Digital Health*, vol. 2, no. 4, e156–e157, 2020. DOI: 10.1016/S2589-7500(20)30055-8.

[239] A. Desai *et al.*, "Crowdsourcing a crisis response for COVID-19 in oncology," *Nature cancer*, vol. 1, no. 5, pp. 473–476, 2020. DOI: 10.1038/s43018-020-0065-z.

[240] S. Lämmer, B. Gehlsen, and D. Helbing, "Scaling laws in the spatial structure of urban road networks," *Physica A: Statistical Mechanics and its Applications*, vol. 363, no. 1, pp. 89–95, 2006. DOI: 10.1016/j.physa.2006.01.051.

[241] A. Clementi, F. d'Amore, G. Giakkoupis, and E. Natale, "Search via Parallel Lévy Walks on $\mathbb{Z}^2$," in *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, ser. PODC'21, New York, NY, USA: Association for Computing Machinery, 2021, pp. 81–91. DOI: 10.1145/3465084.3467921.

[242] R. Metzler and J. Klafter, "The random walk's guide to anomalous diffusion: A fractional dynamics approach," *Physics Reports*, vol. 339, no. 1, pp. 1–77, 2000. DOI: 10.1016/S0370-1573(00)00070-3.

[243] K. E. Persson, D. Manivannan, and M. Singhal, "Bluetooth scatternets: Criteria, models and classification," *Ad Hoc Networks*, vol. 3, no. 6, pp. 777–794, 2005. DOI: 10.1016/j.adhoc.2004.03.014.

[244] T. Salonidis, P. Bhagwat, L. Tassiulas, and R. LaMaire, "Proximity awareness and ad hoc network establishment in Bluetooth," University of Maryland, Tech. Rep., 2001. [Online]. Available: http://hdl.handle.net/1903/6194.

[245] V. Conan, J. Leguay, T. Friedman, *et al.*, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *Proceedings of the 1st International Conference on Autonomic Computing and Communication Systems*, ser. Autonomics '07, vol. 1, Brussels, BEL: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007, pp. 1–9. [Online]. Available: https://dl.acm.org/doi/10.5555/1365562.1365588.

[246] A. Passarella, M. Conti, C. Boldrini, and R. I. Dunbar, "Modelling inter-contact times in social pervasive networks," in *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2011, pp. 333–340. DOI: 10.1145/2068897.2068955.

[247] E. Hernández-Orallo, J. C. Cano, C. T. Calafate, and P. Manzoni, "New approaches for characterizing inter-contact times in opportunistic networks," *Ad Hoc Networks*, vol. 52, pp. 160–172, 2016. DOI: 10.1016/j.adhoc.2016.04.003.

[248] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. M. Ni, "Recognizing exponential inter-contact time in VANETs," in *2010 Proceedings IEEE INFOCOM*, IEEE, 2010, pp. 1–5. DOI: 10.1109/INFCOM.2010.5462263.

[249] M. Gueuning, J.-C. Delvenne, and R. Lambiotte, "Imperfect spreading on temporal networks," *The European Physical Journal B*, vol. 88, no. 11, p. 282, 2015. DOI: 10.1140/epjb/e2015-60596-0.

[250] A.-L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005. DOI: 10.1038/nature03459.

[251] L. Reichert, S. Brack, and B. Scheuermann, "Privacy-preserving contact tracing of COVID-19 patients," in *41st IEEE Symposium on Security and Privacy*, 2020. [Online]. Available: https://www.ieee-security.org/TC/SP2020/program-posters.html.

[252] Q. Tang, "Privacy-preserving contact tracing: Current solutions and open questions," *arXiv*, 2020. DOI: 10.48550/arXiv.2004.06818.

[253] E. Hesamifard, H. Takabi, M. Ghasemi, and R. N. Wright, "Privacy-preserving machine learning as a service," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 3, pp. 123–142, 2018. DOI: 10.1515/popets-2018-0024.

[254] E. Hesamifard, H. Takabi, M. Ghasemi, and C. Jones, "Privacy-preserving machine learning in cloud," in *Proceedings of the 2017 on Cloud Computing Security Workshop*, ser. CCSW '17, New York, NY, USA: Association for Computing Machinery, 2017, pp. 39–43. DOI: 10.1145/3140649.3140655.

[255] S. Pongnumkul, P. Chaovalit, and N. Surasvadi, "Applications of smartphone-based sensors in agriculture: A systematic review of research," *Journal of Sensors*, vol. 2015, p. 195 308, 2015. DOI: 10.1155/2015/195308.

[256] J. Wahlström, I. Skog, and P. Händel, "Smartphone-based vehicle telematics: A ten-year anniversary," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2802–2825, 2017. DOI: 10.1109/TITS.2017.2680468.

[257] M. Hashemian *et al.*, "iEpi: An end to end solution for collecting, conditioning and utilizing epidemiologically relevant data," in *Proceedings of the 2nd ACM International Workshop on Pervasive Wireless Healthcare*, ser. MobileHealth '12, New York, NY, USA: Association for Computing Machinery, 2012, pp. 3–8. DOI: 10.1145/2248341.2248345.

[258] H. Þórarinsdóttir, L. V. Kessing, and M. Faurholt-Jepsen, "Smartphone-based self-assessment of stress in healthy adult individuals: A systematic review," *Journal of Medical Internet Research*, vol. 19, no. 2, e41, 2017. DOI: 10.2196/jmir.6397.

[259] M. A. Habib, M. S. Mohktar, S. B. Kamaruzzaman, K. S. Lim, T. M. Pin, and F. Ibrahim, "Smartphone-based solutions for fall detection and prevention: Challenges and open issues," *Sensors*, vol. 14, no. 4, pp. 7181–7208, 2014. DOI: 10.3390/s140407181.

[260] V. P. Cornet and R. J. Holden, "Systematic review of smartphone-based passive sensing for health and wellbeing," *Journal of Biomedical Informatics*, vol. 77, pp. 120–132, 2018. DOI: 10.1016/j.jbi.2017.12.008.

[261] G. Rateni, P. Dario, and F. Cavallo, "Smartphone-based food diagnostic technologies: A review," *Sensors*, vol. 17, no. 6, p. 1453, 2017. DOI: 10.3390/s17061453.

[262] A. Krause, E. Horvitz, A. Kansal, and F. Zhao, "Toward community sensing," in *2008 International Conference on Information Processing in Sensor Networks (ipsn 2008)*, IEEE, 2008, pp. 481–492. DOI: 10.1109/IPSN.2008.37.

[263] J. Bort-Roig, N. D. Gilson, A. Puig-Ribera, R. S. Contreras, and S. G. Trost, "Measuring and influencing physical activity with smartphone technology: A systematic review," *Sports Medicine*, vol. 44, no. 5, pp. 671–686, 2014. DOI: 10.1007/s40279-014-0142-5.

[264] C. M. Wharton, C. S. Johnston, B. K. Cunningham, and D. Sterner, "Dietary self-monitoring, but not dietary quality, improves with use of smartphone app technology in an 8-week weight loss trial," *Journal of Nutrition Education and Behavior*, vol. 46, no. 5, pp. 440–444, 2014. DOI: 10.1016/j.jneb.2014.04.291.

[265] C. Crema *et al.*, "Smartphone-based system for the monitoring of vital parameters and stress conditions of amatorial racecar drivers," in *2015 IEEE SENSORS*, IEEE, 2015, pp. 1–4. DOI: 10.1109/ICSENS.2015.7370521.

[266] M. Samaha and N. S. Hawi, "Relationships among smartphone addiction, stress, academic performance, and satisfaction with life," *Computers in Human Behavior*, vol. 57, pp. 321–325, 2016.

[267] A. J. Van Deursen, C. L. Bolle, S. M. Hegner, and P. A. Kommers, "Modeling habitual and addictive smartphone behavior: The role of smartphone usage types, emotional intelligence, social stress, self-regulation, age, and gender," *Computers in Human Behavior*, vol. 45, pp. 411–420, 2015. DOI: 10.1016/j.chb.2014.12.039.

[268] M. W. Wukitsch, M. T. Petterson, D. R. Tobler, and J. A. Pologe, "Pulse oximetry: Analysis of theory, technology, and practice," *Journal of Clinical Monitoring*, vol. 4, no. 4, pp. 290–301, 1988. DOI: 10.1007/BF01617328.

[269] A. Chandrasekhar, K. Natarajan, M. Yavarimanesh, and R. Mukkamala, "An iphone application for blood pressure monitoring via the oscillometric finger pressing method," *Scientific Reports*, vol. 8, no. 1, p. 13 136, 2018. DOI: 10.1038/s41598-018-31632-x.

[270] A. Chandrasekhar, C.-S. Kim, M. Naji, K. Natarajan, J. O. Hahn, and R. Mukkamala, "Smartphone-based blood pressure monitoring via the oscillometric finger-pressing method," *Science Translational Medicine*, vol. 10, no. 431, 2018. DOI: 10.1126/scitranslmed.aap8674.

[271] X. Huang *et al.*, "Smartphone-based analytical biosensors," *The Analyst*, vol. 143, no. 22, pp. 5339–5351, 2018. DOI: 10.1039/c8an01269e.

[272] A. Roda, E. Michelini, M. Zangheri, M. Di Fusco, D. Calabria, and P. Simoni, "Smartphone-based biosensors: A critical review and perspectives," *TrAC Trends in Analytical Chemistry*, vol. 79, pp. 317–325, 2016. DOI: 10.1016/j.trac.2015.10.019.

[273] T. J. Trull and U. Ebner-Priemer, "Ambulatory assessment," *Annual Review of Clinical Psychology*, vol. 9, pp. 151–176, 2013. DOI: 10.1146/annurev-clinpsy-050212-185510.

[274] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Scientific Reports*, vol. 5, p. 16 743, 2015. DOI: 10.1038/srep16743.

[275] A. Bachmann *et al.*, "How to use smartphones for less obtrusive ambulatory mood assessment and mood recognition," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ser. UbiComp/ISWC'15 Adjunct, New York, NY, USA: Association for Computing Machinery, 2015, pp. 693–702. DOI: 10.1145/2800835.2804394.

[276] W.-Y. Cheng *et al.*, "Smartphone-based continuous mobility monitoring of Parkinsons disease patients reveals impacts of ambulatory bout length on gait features," in *2017 IEEE Life Sciences Conference (LSC)*, IEEE, 2017, pp. 166–169. DOI: 10.1109/LSC.2017.8268169.

[277] R. Archdeacon, R. Schneider, and Y. Jiang, "Critical flaws in the validation of the instant blood pressure smartphone app—a letter from the app developers-reply," *JAMA Internal Medicine*, vol. 176, no. 9, pp. 1410–1411, 2016. DOI: 10.1001/jamainternmed.2016.4765.

[278] M. A. Case, H. A. Burwick, K. G. Volpp, and M. S. Patel, "Accuracy of smartphone applications and wearable devices for tracking physical activity data," *JAMA*, vol. 313, no. 6, pp. 625–626, 2015. DOI: 10.1001/jama.2014.17841.

[279] J.-M. Lee, Y. Kim, and G. J. Welk, "Validity of consumer-based physical activity monitors.," *Medicine and Science in Sports and Exercise*, vol. 46, no. 9, pp. 1840–1848, 2014. DOI: 10.1249/MSS.0000000000000287.

[280] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals," *Current cardiology reviews*, vol. 8, no. 1, pp. 14–25, 2012. DOI: 10.2174/157340312801215782.

[281] N. Andrienko, G. Andrienko, N. Pelekis, and S. Spaccapietra, "Basic concepts of movement data," in *Mobility, data mining and privacy*, F. Giannotti and D. Pedreschi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 15–38. DOI: 10.1007/978-3-540-75177-9_2.

[282] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi, "Space transformation for understanding group movement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2169–2178, 2013. DOI: 10.1109/TVCG.2013.193.

[283] R. Zhang, K. G. Stanley, S. Bell, and D. Fuller, "A feature set for spatial behavior characterization," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '18, New York, NY, USA: Association for Computing Machinery, 2018, pp. 512–515. DOI: 10.1145/3274895.3274973.

[284] R. Zhang, K. G. Stanley, D. Fuller, and S. Bell, "Differentiating population spatial behavior using representative features of geospatial mobility (ReFGeM)," *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, vol. 6, no. 1, pp. 1–25, 2020. DOI: 10.1145/3362063.

[285] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. DOI: 10.1126/science.286.5439.509.

[286] S. Hong, K. Lee, and I. Rhee, "STEP: A spatio-temporal mobility model for humans walks," in *The 7th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE MASS 2010)*, IEEE, 2010, pp. 630–635. DOI: 10.1109/MASS.2010.5663776.

[287] A. Munjal, T. Camp, and W. C. Navidi, "SMOOTH: A simple way to model human mobility," in *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '11, New York, NY, USA: Association for Computing Machinery, 2011, pp. 351–360. DOI: 10.1145/2068897.2068957.

[288] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: A new mobility model for human walks," in *IEEE INFOCOM 2009*, IEEE, 2009, pp. 855–863. DOI: 10.1109/INFCOM.2009.5061995.

[289] K. Rasul, S. A. Chowdhury, D. Makaroff, and K. Stanley, "Community-based forwarding for low-capacity pocket switched networks," in *Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '14, New York, NY, USA: Association for Computing Machinery, 2014, pp. 249–257. DOI: 10.1145/2641798.2641801.

[290] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016. DOI: 10.1016/j.trc.2016.04.005.

[291] 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, p. 62, 2011. DOI: 10.1038/srep00062.

[292] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 208–219, 2017. DOI: 10.1109/TBDATA.2016.2631141.

[293] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences*, vol. 107, no. 51, pp. 22 020–22 025, 2010. DOI: 10.1073/pnas.1009094108.

[294] A. Carlotto, M. Parodi, C. Bonamico, F. Lavagetto, and M. Valla, "Proximity classification for mobile devices using Wi-Fi environment similarity," in *Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-Less Environments*, ser. MELT '08, New York, NY, USA: Association for Computing Machinery, 2008, pp. 43–48. DOI: 10.1145/1410012.1410023.

[295] V. Osmani, I. Carreras, A. Matic, and P. Saar, "An analysis of distance estimation to detect proximity in social interactions," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 3, pp. 297–306, 2014. DOI: 10.1007/s12652-012-0171-6.

[296] G. M. Vazquez-Prokopec *et al.*, "Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment," *PLOS ONE*, vol. 8, no. 4, e58802, 2013. DOI: 10.1371/journal.pone.0058802.

[297] M. Bolic, M. Rostamian, and P. M. Djuric, "Proximity detection with RFID: A step toward the Internet of Things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 70–76, 2015. DOI: 10.1109/MPRV.2015.39.

[298] X. Jiang *et al.*, "Design and evaluation of a wireless magnetic-based proximity detection platform for indoor applications," in *2012 ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN)*, 2012, pp. 221–232. DOI: 10.1109/IPSN.2012.6920959.

[299] H. T. Friis, "A note on a simple transmission formula," *Proceedings of the IRE*, vol. 34, no. 5, pp. 254–256, 1946. DOI: 10.1109/JRPROC.1946.234568.

[300] W. L. Stutzman and G. A. Thiele, *Antenna theory and design*, 3rd ed. John Wiley & Sons, 2012.

[301] S. Odenwald, *Experimenter's Guide To Smartphone Sensors*. NASA Space Science Education Consortium, 2019. [Online]. Available: https://spacemath.gsfc.nasa.gov/Sensor/SensorsBook.pdf.

[302] S. Jeong, S. Kuk, and H. Kim, "A smartphone magnetometer-based diagnostic test for automatic contact tracing in infectious disease epidemics," *IEEE Access*, vol. 7, pp. 20 734–20 747, 2019. DOI: 10.1109/ACCESS.2019.2895075.

[303] V. Shubina, S. Holcer, M. Gould, and E. S. Lohan, "Survey of decentralized solutions with mobile devices for user location tracking, proximity detection, and contact tracing in the covid-19 era," *Data*, vol. 5, no. 4, p. 87, 2020. DOI: 10.3390/data5040087.

[304] L. Cheng, C. Wu, Y. Zhang, H. Wu, M. Li, and C. Maple, "A survey of localization in wireless sensor network," *International Journal of Distributed Sensor Networks*, vol. 8, no. 12, p. 962 523, 2012. DOI: 10.1155/2012/962523.

[305] U.S. Coast Guard, *Navstar GPS user equipment introduction (public release version)*, Sep. 1996. [Online]. Available: https://www.navcen.uscg.gov/pubs/gps/gpsuser/gpsuser.pdf.

[306] N. Ashby, "Relativity in the global positioning system," *Living Reviews in Relativity*, vol. 6, no. 1, p. 1, 2003. DOI: 10.12942/lrr-2003-1.

[307] M. B. Kjærgaard, H. Blunck, T. Godsk, T. Toftkjær, D. L. Christensen, and K. Grønbæk, "Indoor positioning using GPS revisited," in *Pervasive Computing*, P. Floréen, A. Krüger, and M. Spasojevic, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 38–56. DOI: 10.1007/978-3-642-12654-3_3.

[308] A. Troy and B. Voigt, *Introduction to projections and coordinate systems (lecture notes)*, 2011. [Online]. Available: http://www.uvm.edu/rsenr/gradgis/lectures/lecture6.pptx.

[309] J. P. Snyder, "Map projections: A working manual," U.S. Government Printing Office, Washington, D.C., Tech. Rep., 1987. DOI: 10.3133/pp1395.

[310] IEEE 802.11 Working Group and others, "IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11-2007 (Revision of IEEE Std 802.11-1999)*, pp. 1–1076, 2007. DOI: 10.1109/IEEESTD.2007.373646.

[311] G. Castignani, A. Arcia, and N. Montavont, "A study of the discovery process in 802.11 networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 15, no. 1, pp. 25–36, 2011. DOI: 10.1145/1978622.1978626.

[312] F. Goovaerts, G. Acar, R. Galvez, F. Piessens, and M. Vanhoef, "Improving privacy through fast passive Wi-Fi scanning," in *Secure IT Systems*, A. Askarov, R. R. Hansen, and W. Rafnsson, Eds., Cham: Springer International Publishing, 2019, pp. 37–52. DOI: 10.1007/978-3-030-35055-0_3.

[313] D. Murray, M. Dixon, and T. Koziniec, "Scanning delays in 802.11 networks," in *The 2007 International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST 2007)*, IEEE, 2007, pp. 255–260. DOI: 10.1109/NGMAST.2007.4343430.

[314] D. Jaisinghani, V. Naik, S. K. Kaul, R. Balan, and S. Roy, "Improving the performance of WLANs by reducing unnecessary active scans," *arXiv*, 2018. DOI: 10.48550/arXiv.1807.05523.

[315] A. Vlavianos, L. K. Law, I. Broustis, S. V. Krishnamurthy, and M. Faloutsos, "Assessing link quality in IEEE 802.11 wireless networks: Which is the right metric?" In *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, IEEE, 2008, pp. 1–6. DOI: 10.1109/PIMRC.2008.4699837.

[316] E. Au, "Bluetooth 5.0 and beyond [standards]," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 119–120, 2019. DOI: 10.1109/MVT.2019.2905520.

[317] Bluetooth, SIG, "Specification of the Bluetooth system, core version 5.2," *Bluetooth SIG*, 2019. [Online]. Available: https://www.bluetooth.com/specifications/specs/core-specification-5-2/.

[318] J. Haartsen, M. Naghshineh, J. Inouye, O. J. Joeressen, and W. Allen, "Bluetooth: Vision, goals, and architecture," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 2, no. 4, pp. 38–45, 1998. DOI: 10.1145/1321400.1321402.

[319] Bluetooth, SIG, "Specification of the Bluetooth system, core version 1.1," *Bluetooth SIG*, 2001. [Online]. Available: https://people.inf.ethz.ch/hvogt/proj/btmp3/Datasheets/Bluetooth_11_Specifications_Book.pdf.

[320] J. Figueiras, H. P. Schwefel, and I. Kovacs, "Accuracy and timing aspects of location information based on signal-strength measurements in Bluetooth," in *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, IEEE, vol. 4, 2005, pp. 2685–2690. DOI: 10.1109/PIMRC.2005.1651931.

[321] S. Gollakota, F. Adib, D. Katabi, and S. Seshan, "Clearing the RF smog: Making 802.11n robust to cross-technology interference," in *Proceedings of the ACM SIGCOMM 2011 Conference*, ser. SIGCOMM '11, New York, NY, USA: Association for Computing Machinery, 2011, pp. 170–181. DOI: 10.1145/2018436.2018456.

[322] A. Hithnawi, H. Shafagh, and S. Duquennoy, "Understanding the impact of cross technology interference on IEEE 802.15.4," in *Proceedings of the 9th ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization*, ser. WiNTECH '14, New York, NY, USA: Association for Computing Machinery, 2014, pp. 49–56. [Online]. Available: 10.1145/2643230.2643235.

[323] Y. Wang and Q. Wang, "Evaluating the IEEE 802.15.6 2.4GHz WBAN proposal on medical multi-parameter monitoring under Wi-Fi/Bluetooth interference," in *International Journal of E-Health and Medical Communications (IJEHMC)*, 3, vol. 2, IGI Global, 2013, pp. 48–62. DOI: 10.4018/jehmc.2011070103.

[324] L. Pei *et al.*, "The evaluation of Wi-Fi positioning in a Bluetooth and Wi-Fi coexistence environment," in *2012 Ubiquitous Positioning, Indoor Navigation, and Location Based Service (UPINLBS)*, IEEE, 2012, pp. 1–6. DOI: 10.1109/UPINLBS.2012.6409768.

[325] F. Hermans, O. Rensfelt, T. Voigt, E. Ngai, L.-Å. Nordén, and P. Gunningberg, "SoNIC: Classifying interference in 802.15.4 sensor networks," in *2013 ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2013, pp. 55–66. DOI: 10.1145/2461381.2461392.

[326] B. d. Silva, A. Natarajan, and M. Motani, "Inter-user interference in body sensor networks: Preliminary investigation and an infrastructure-based solution," in *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, IEEE, 2009, pp. 35–40. DOI: 10.1109/BSN.2009.36.

[327] A. K. M. M. Hossain and W.-S. Soh, "A comprehensive study of Bluetooth signal parameters for localization," in *2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, IEEE, 2007, pp. 1–5. DOI: 10.1109/PIMRC.2007.4394215.

[328] Y. Wang, X. Yang, Y. Zhao, Y. Liu, and L. Cuthbert, "Bluetooth positioning using RSSI and triangulation methods," in *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, IEEE, 2013, pp. 837–842. DOI: 10.1109/CCNC.2013.6488558.

[329] J. Jung, D. Kang, and C. Bae, "Distance estimation of smart device using Bluetooth," in *ICSNC 2013 : The Eighth International Conference on Systems and Networks Communications*, 2013, pp. 13–8. [Online]. Available: http://www.thinkmind.org/download.php?articleid=icsnc_2013_1_30_20039.

[330] A. Mackey, P. Spachos, L. Song, and K. N. Plataniotis, "Improving BLE beacon proximity estimation accuracy through Bayesian filtering," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3160–3169, 2020. DOI: 10.1109/JIOT.2020.2965583.

[331] Y. Gu and F. Ren, "Energy-efficient indoor localization of smart hand-held devices using Bluetooth," *IEEE Access*, vol. 3, pp. 1450–1461, 2015. DOI: 10.1109/ACCESS.2015.2441694.

[332] Z. Jianyong, L. Haiyong, C. Zili, and L. Zhaohui, "RSSI based Bluetooth low energy indoor positioning," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, 2014, pp. 526–533. DOI: 10.1109/IPIN.2014.7275525.

[333] S. Zhou and J. K. Pollard, "Position measurement using Bluetooth," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 2, pp. 555–558, 2006. DOI: 10.1109/TCE.2006.1649679.

[334] M. S. Aman, H. Jiang, C. Quint, K. Yelamarthi, and A. Abdelgawad, "Reliability evaluation of iBeacon for micro-localization," in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, IEEE, 2016, pp. 1–5. DOI: 10.1109/UEMCON.2016.7777904.

[335] P. Mirowski, T. K. Ho, S. Yi, and M. MacDonald, "SignalSLAM: Simultaneous localization and mapping with mixed Wi-Fi, Bluetooth, LTE and magnetic signals," in *International Conference on Indoor Positioning and Indoor Navigation*, IEEE, 2013, pp. 1–10. DOI: 10.1109/IPIN.2013.6817853.

[336] M. Altini, D. Brunelli, E. Farella, and L. Benini, "Bluetooth indoor localization with multiple neural networks," in *IEEE 5th International Symposium on Wireless Pervasive Computing 2010*, IEEE, 2010, pp. 295–300. DOI: 10.1109/ISWPC.2010.5483748.

[337] D. G. Young, *How far can you go?* 2021. [Online]. Available: http://www.davidgyoungtech.com/2020/05/15/how-far-can-you-go (visited on 2021).

[338] D. Cypher and N. Golmie, *NIST SGIP priority action plan 2, guidelines for assessing wireless standards for smart grid applications*, 2014. DOI: 10.6028/NIST.IR.7761r1.

[339] J. Benavides *et al.*, "3G smartphone technologies for generating personal social network contact distributions and graphs," in *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, IEEE, 2011, pp. 182–189. DOI: 10.1109/HISB.2011.2.

[340] J. Hallberg, M. Nilsson, and K. Synnes, "Positioning with Bluetooth," in *10th International Conference on Telecommunications, 2003. ICT 2003.*, IEEE, vol. 2, 2003, pp. 954–958. DOI: 10.1109/ICTEL.2003.1191568.

[341] A. Ganguly, C. Reddy, Y. Hao, and I. Panahi, "Improving sound localization for hearing aid devices using smartphone assisted technology," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, IEEE, 2016, pp. 165–170. DOI: 10.1109/SiPS.2016.37.

[342] A. Ens *et al.*, "Acoustic self-calibrating system for indoor smart phone tracking," *International Journal of Navigation and Observation*, vol. 2015, p. 694 695, 2015. DOI: 10.1155/2015/694695.

[343] F. Höflinger *et al.*, "Acoustic self-calibrating system for indoor smartphone tracking (ASSIST)," in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, 2012, pp. 1–9. DOI: 10.1109/IPIN.2012.6418877.

[344] D. V. Le, J. W. Kamminga, H. Scholten, and P. J. Havinga, "Nondeterministic sound source localization with smartphones in crowdsensing," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, IEEE, 2016, pp. 1–7. DOI: 10.1109/PERCOMW.2016.7457115.

[345] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953. DOI: 10.1121/1.1907229.

[346] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, Jul. 2018. DOI: 10.1145/3197517.3201357.

[347] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 639–658. DOI: 10.1007/978-3-030-01231-1_39.

[348] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 587–604. DOI: 10.1007/978-3-030-01246-5_35.

[349] T. M. Fernandez, J. Rodas, C. J. Escudero, and D. I. Iglesia, "Bluetooth sensor network positioning system with dynamic calibration," in *2007 4th International Symposium on Wireless Communication Systems*, IEEE, 2007, pp. 45–49. DOI: 10.1109/ISWCS.2007.4392299.

[350] Y. Shi, W. Shi, X. Liu, and X. Xiao, "An RSSI classification and tracing algorithm to improve trilateration-based positioning," *Sensors*, vol. 20, no. 15, p. 4244, 2020. DOI: 10.3390/s20154244.

[351] L. Khalil and P. Jung, "Scaled unscented Kalman filter for RSSI-based indoor positioning and tracking," in *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies*, IEEE, 2015, pp. 132–137. DOI: 10.1109/NGMAST.2015.20.

[352] W. Xue, W. Qiu, X. Hua, and K. Yu, "Improved Wi-Fi RSSI measurement for indoor localization," *IEEE Sensors Journal*, vol. 17, no. 7, pp. 2224–2230, 2017. DOI: 10.1109/JSEN.2017.2660522.

[353] C. Zhou, J. Yuan, H. Liu, and J. Qiu, "Bluetooth indoor positioning based on RSSI and Kalman filter," *Wireless Personal Communications*, vol. 96, no. 3, pp. 4115–4130, 2017. DOI: 10.1007/s11277-017-4371-4.

[354] W. M. Yeung and J. K. Ng, "An enhanced wireless LAN positioning algorithm based on the fingerprint approach," in *TENCON 2006—2006 IEEE Region 10 Conference*, IEEE, 2006, pp. 1–4. DOI: 10.1109/TENCON.2006.343696.

[355] L. Chen, L. Pei, H. Kuusniemi, Y. Chen, T. Kröger, and R. Chen, "Bayesian fusion for indoor positioning using Bluetooth fingerprints," *Wireless Personal Communications*, vol. 70, no. 4, pp. 1735–1745, 2013. DOI: 10.1007/s11277-012-0777-1.

[356] H. J. Pérez Iglesias, V. Barral, and C. J. Escudero, "Indoor person localization system through RSSI Bluetooth fingerprinting," in *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, IEEE, 2012, pp. 40–43. [Online]. Available: https://ieeexplore.ieee.org/document/6208163.

[357] 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, p. 185, 2013. DOI: 10.1186/1471-2334-13-185.

[358] I. S. Mbalawata, S. Särkkä, and H. Haario, "Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering," *Computational Statistics*, vol. 28, no. 3, pp. 1195–1223, 2013. DOI: 10.1007/s00180-012-0352-y.

[359] I. Dorigatti, S. Cauchemez, A. Pugliese, and N. M. Ferguson, "A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: Application to the Italian 2009–2010 A/H1N1 influenza pandemic," *Epidemics*, vol. 4, no. 1, pp. 9–21, 2012. DOI: 10.1016/j.epidem.2011.11.001.

[360] F. C. Coelho, C. T. Codeço, and M. G. M. Gomes, "A Bayesian framework for parameter estimation in dynamical models," *PLOS ONE*, vol. 6, no. 5, e19616, 2011. DOI: 10.1371/journal.pone.0019616.

[361] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000. DOI: 10.1023/A:1008935410038.

[362] M. Chiogna and C. Gaetan, "Hierarchical space-time modelling of epidemic dynamics: An application to measles outbreaks," *Statistical Methods and Applications*, vol. 13, no. 1, pp. 55–71, 2004. DOI: 10.1007/s10260-004-0085-3.

[363] B. Cazelles and N. P. Chau, "Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic," *Mathematical Biosciences*, vol. 140, no. 2, pp. 131–154, 1997. DOI: 10.1016/s0025-5564(96)00155-1.

[364] Chiogna and Carlo Gaetan, Monica and C. Gaetan, "Dynamic generalized linear models with application to environmental epidemiology," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 51, no. 4, pp. 453–468, 2002. DOI: 10.1111/1467-9876.00280.

[365] N. Eagle, A. (Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15 274–15 278, 2009. DOI: 10.1073/pnas.0900282106.

[366] M. S. Hashemian, K. G. Stanley, and N. D. Osgood, "Leveraging H1N1 infection transmission modeling with proximity sensor microdata," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, pp. 35–35, 2012. DOI: 10.1186/1472-6947-12-35.

[367] S. Funk, M. Salathé, and V. A. Jansen, "Modelling the influence of human behaviour on the spread of infectious diseases: A review," *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1247–1256, 2010. DOI: 10.1098/rsif.2010.0142.

[368] N. Osgood, "Using traditional and agent based toolsets for system dynamics: Present tradeoffs and future evolution," *Proceedings of the Proceedings The 2007 International Conference of the System Dynamics Society*, vol. 5, p. 3364, 2007. [Online]. Available: http://toc.proceedings.com/02047webtoc.pdf.

[369] D. Satcher, "Emerging infections: Getting ahead of the curve," *Emerging Infectious Diseases*, vol. 1, no. 1, pp. 1–6, 1995. DOI: 10.3201/eid0101.950101.

[370] W. H. McNeill, "Disease in history," *Social Science & Medicine. Part B: Medical Anthropology*, vol. 12, pp. 79–81, 1978. DOI: 10.1016/0160-7987(78)90012-1.

[371] N.-A. M. Molinari *et al.*, "The annual impact of seasonal influenza in the US: Measuring disease burden and costs," *Vaccine*, vol. 25, no. 27, pp. 5086–5096, 2007. DOI: 10.1016/j.vaccine.2007.03.046.

[372] K. K. Tibbetts, R. A. Ottoson, and D. T. Tsukayama, "Public health response to tuberculosis outbreak among persons experiencing homelessness, Minneapolis, Minnesota, USA, 2017–2018," *Emerging Infectious Diseases*, vol. 26, no. 3, pp. 420–426, 2020. DOI: 10.3201/eid2603.190643.

[373] S. Sarkar, A. Zlojutro, K. Khan, and L. Gardner, "Measles resurgence in the USA: How international travel compounds vaccine resistance," *The Lancet Infectious Diseases*, vol. 19, no. 7, pp. 684–686, 2019. DOI: 10.1016/S1473-3099(19)30231-2.

[374] M. Patel *et al.*, "National update on measles cases and outbreaks – United States, January 1–October 1, 2019," *Morbidity and Mortality Weekly Report*, vol. 68, no. 40, pp. 893–896, 2019. DOI: 10.15585/mmwr.mm6840e2.

[375] World Health Organization, "Weekly epidemiological update on COVID-19—25 May 2021," 2021. [Online]. Available: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---25-may-2021.

[376] World Health Organization, *Coronavirus disease (COVID-19): Situation report, 200*, Online, 2020. [Online]. Available: https://apps.who.int/iris/handle/10665/333832.

[377] D. Skegg *et al.*, "Future scenarios for the COVID-19 pandemic," *The Lancet*, vol. 397, no. 10276, pp. 777–778, 2021. DOI: 10.1016/S0140-6736(21)00424-4.

[378] T. L. Microbe, "COVID-19 vaccines: The pandemic will not end overnight," *The Lancet Microbe*, vol. 2, no. 1, e1, 2021. DOI: 10.1016/S2666-5247(20)30226-3.

[379] L. Perez and S. Dragicevic, "An agent-based approach for modeling dynamics of contagious disease spread," *International Journal of Health Geographics*, vol. 8, no. 1, pp. 1–17, 2009. DOI: 10.1186/1476-072X-8-50.

[380] G. Chowell, L. Sattenspiel, S. Bansal, and C. Viboud, "Mathematical models to characterize early epidemic growth: A review," *Physics of Life Reviews*, vol. 18, pp. 66–97, 2016. DOI: 10.1016/j.plrev.2016.07.005.

[381] G. V. Bobashev, D. M. Goedecke, F. Yu, and J. M. Epstein, "A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches," in *2007 Winter Simulation Conference*, IEEE, 2007, pp. 1532–1537. DOI: 10.1109/WSC.2007.4419767.

[382] P. Rodríguez *et al.*, "A population-based controlled experiment assessing the epidemiological impact of digital contact tracing," *Nature Communications*, vol. 12, no. 1, pp. 1–6, 2021. DOI: 10.1038/s41467-020-20817-6.

[383] T. Hoang *et al.*, "A systematic review of social contact surveys to inform transmission models of close-contact infections," *Epidemiology (Cambridge, Mass.)*, vol. 30, no. 5, pp. 723–736, 2019. DOI: 10.1097/EDE.0000000000001047.

[384] K. H. Grantz *et al.*, "Age-specific social mixing of school-aged children in a US setting using proximity detecting sensors and contact surveys," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021. DOI: 10.1038/s41598-021-81673-y.

[385] B. Han, P. Cook, and T. Baldwin, "Geolocation prediction in social media data by finding location indicative words," in *Proceedings of COLING 2012*, 2012, pp. 1045–1062. [Online]. Available: https://aclanthology.org/C12-1064.

[386] S. Hasan, S. V. Ukkusuri, and X. Zhan, "Understanding social influence in activity location choice and lifestyle patterns using geolocation data from social media," *Frontiers in ICT*, vol. 3, p. 10, 2016. DOI: 10.3389/fict.2016.00010.

[387] M. Rizwan, W. Wan, and L. Gwiazdzinski, "Visualization, spatiotemporal patterns, and directional analysis of urban activities using geolocation data extracted from LBSN," *ISPRS International Journal of Geo-Information*, vol. 9, no. 2, p. 137, 2020. DOI: 10.3390/ijgi9020137.

[388] Ethica Data, *Ethica data: Empower your research with smartphones and big data*, 2021. [Online]. Available: https://ethicadata.com/.

[389] K. Michael and R. Abbas, "Behind COVID-19 contact trace apps: The Google–Apple partnership," *IEEE Consumer Electronics Magazine*, vol. 9, no. 5, pp. 71–76, 2020. DOI: 10.1109/MCE.2020.3002492.

[390] F. N. Wirth, M. Johns, T. Meurers, and F. Prasser, "Citizen-centered mobile health apps collecting individual-level spatial data for infectious disease management: Scoping review," *JMIR mHealth and uHealth*, vol. 8, no. 11, e22594, 2020. DOI: 10.2196/22594.

[391] L. Reichert, S. Brack, and B. Scheuermann, "A survey of automatic contact tracing approaches using Bluetooth low energy," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 2, pp. 1–33, 2021. DOI: 10.1145/3444847.

[392] R. Al Alawi, "RSSI based location estimation in wireless sensors networks," in *2011 17th IEEE International Conference on Networks*, IEEE, 2011, pp. 118–122. DOI: 10.1109/ICON.2011.6168517.

[393] A. Harun *et al.*, "Comparative performance analysis of wireless RSSI in wireless sensor networks motes in tropical mixed-crop precision farm," in *2012 Third International Conference on Intelligent Systems Modelling and Simulation*, IEEE, 2012, pp. 606–610. DOI: 10.1109/ISMS.2012.57.

[394] R. Networks, *Android beacon library*, 2019. [Online]. Available: https://altbeacon.github.io/android-beacon-library/distance-calculations.html (visited on 2019).

[395] K. Stanley *et al.*, "Opportunistic natural experiments using digital telemetry: A transit disruption case study," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1853–1872, 2016. DOI: 10.1080/13658816.2016.1145224.

[396] D. L. Knowles, K. G. Stanley, and N. D. Osgood, "A field-validated architecture for the collection of health-relevant behavioural data," in *2014 IEEE International Conference on Healthcare Informatics*, IEEE, 2014, pp. 79–88. DOI: 10.1109/ICHI.2014.18.

[397] World Health Organization, "Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19)," World Health Organization, Tech. Rep., 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

[398] M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, and L. Finelli, "Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: A systematic review of the literature," *BMC Infectious Diseases*, vol. 14, no. 1, p. 480, 2014. DOI: 10.1186/1471-2334-14-480.

[399] I. A. Coalition, *Influenza: Questions and answers*, Oct. 2020. [Online]. Available: https://www.immunize.org/catg.d/p4208.pdf.

[400] M. K. Steele *et al.*, "Characterizing norovirus transmission from outbreak data, United States," *Emerging Infectious Diseases*, vol. 26, no. 8, p. 1818, 2020. DOI: 10.3201/eid2608.191537.

[401] R. M. Lee *et al.*, "Incubation periods of viral gastroenteritis: A systematic review," *BMC Infectious Diseases*, vol. 13, no. 1, p. 446, 2013. DOI: 10.1186/1471-2334-13-446.

[402] P. E. Fine, "Herd immunity: History, theory, practice," *Epidemiologic Reviews*, vol. 15, no. 2, pp. 265–302, 1993. DOI: 10.1093/oxfordjournals.epirev.a036121.

[403] J. Hamborsky, A. Kroger, *et al.*, *Epidemiology and prevention of vaccine-preventable diseases, E-Book: The Pink Book*. Public Health Foundation, 2015. [Online]. Available: https://www.cdc.gov/vaccines/pubs/pinkbook/index.html.

[404] D. Gao *et al.*, "Prevention and control of zika as a mosquito-borne and sexually transmitted disease: A mathematical modeling analysis," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016. DOI: 10.1038/srep28070.

[405] V. Belik, T. Geisel, and D. Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases," *Physical Review X*, vol. 1, no. 1, p. 011 001, 2011. DOI: 10.1103/PhysRevX.1.011001.

[406] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, no. 5, pp. 60–69, 2003. DOI: 10.1038/scientificamerican0503-60.

[407] R. J. Cabin and R. J. Mitchell, "To Bonferroni or not to Bonferroni: When and how are the questions," *Bulletin of the Ecological Society of America*, vol. 81, no. 3, pp. 246–248, 2000. [Online]. Available: https://www.jstor.org/stable/20168454.

[408] World Health Organization, *Coronavirus disease 2019 (COVID-19): Situation report, 70*, Online, 2020. [Online]. Available: https://apps.who.int/iris/handle/10665/331683.

[409] World Health Organization, *Coronavirus disease 2019 (COVID-19): Situation report, 72*, Online, 2020. [Online]. Available: https://apps.who.int/iris/handle/10665/331685.

[410] World Health Organization, *Coronavirus disease 2019 (COVID-19): Situation report, 85*, Online, 2020. [Online]. Available: https://www.who.int/publications/m/item/situation-report---85.

[411] Y.-C. Wu, C.-S. Chen, and Y.-J. Chan, "The outbreak of COVID-19: An overview," *Journal of the Chinese Medical Association*, vol. 83, no. 3, p. 217, 2020. DOI: 10.1097/JCMA.0000000000000270.

[412] World Health Organization, *Latest updates on the ebola outbreak*, Jun. 2017. [Online]. Available: http://www.who.int/csr/disease/ebola/top-stories-2016/en/.

[413] T. Garske *et al.*, "Heterogeneities in the case fatality ratio in the West African Ebola outbreak 2013–2016," *Philosophical Transactions of the Royal Society B*, vol. 372, no. 1721, p. 20 160 308, 2017. DOI: 10.1098/rstb.2016.0308.

[414] R. Long and E. Ellis, "Tuberculosis elimination in Canada: Truce or victory?" *Canadian Medical Association Journal*, vol. 187, no. 16, pp. 1191–1192, 2015. DOI: 10.1503/cmaj.150317.

[415] V. Gallant, S. Ogunnaike-Cooke, and M. McGuire, "Tuberculosis in Canada: 1924–2012," *Canada Communicable Disease Report*, vol. 40, pp. 99–108, 2014. DOI: 10.14745/ccdr.v40i06a02.

[416] World Health Organization, *Influenza (seasonal) fact sheet*, Jun. 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs211/en/.

[417] G. J. Rubin, R. Amlôt, L. Page, and S. Wessely, "Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: Cross sectional telephone survey," *BMJ*, vol. 339, b2651, 2009. DOI: 10.1136/bmj.b2651.

[418] World Health Organization, *Influenza update - 290*, Jun. 2017. [Online]. Available: http://www.who.int/influenza/surveillance_monitoring/updates/latest_update_GIP_surveillance/en/.

[419] G. Worrall, "Common cold," *Canadian Family Physician*, vol. 57, no. 11, pp. 1289–1290, 2011. [Online]. Available: https://www.cfp.ca/content/53/10/1735.

[420] National Center for Immunization and Respiratory Diseases, Division of Viral Diseases, *Common colds: Protect yourself and others*, Jun. 2017. [Online]. Available: https://www.cdc.gov/features/rhinoviruses/.

[421] Y. Zhang *et al.*, "Resurgence of pertussis infections in Shandong, China: Space-time cluster and trend analysis," *The American Journal of Tropical Medicine and Hygiene*, vol. 100, no. 6, pp. 1342–1354, 2019. DOI: 10.4269/ajtmh.19-0013.

[422] F. R. Mooi, N. A. Van Der Maas, and H. E. De Melker, "Pertussis resurgence: Waning immunity and pathogen adaptation–two sides of the same coin," *Epidemiology & Infection*, vol. 142, no. 4, pp. 685–694, 2014. DOI: 10.1017/S0950268813000071.

[423] J. D. Cherry, "Epidemic pertussis in 2012—the resurgence of a vaccine-preventable disease," *New England Journal of Medicine*, vol. 367, no. 9, pp. 785–787, 2012. DOI: 10.1056/NEJMp1209051.

[424] K. C. Chong *et al.*, "Were infections in migrants associated with the resurgence of measles epidemic during 2013–2014 in southern China? a retrospective data analysis," *International Journal of Infectious Diseases*, vol. 90, pp. 77–83, 2020. DOI: 10.1016/j.ijid.2019.10.014.

[425] M. R. Weigand *et al.*, "Genomic survey of *Bordetella pertussis* diversity, United States, 2000–2013," *Emerging Infectious Diseases*, vol. 25, no. 4, pp. 780–783, 2019. DOI: 10.3201/eid2504.180812.

[426] R. M. Anderson, B. Anderson, and R. M. May, *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[427] L. Pellis *et al.*, "Eight challenges for network epidemic models," *Epidemics*, vol. 10, pp. 58–62, 2015. DOI: 10.1016/j.epidem.2014.07.003.

[428] K. Eames, S. Bansal, S. Frost, and S. Riley, "Six challenges in measuring contact networks for use in modelling," *Epidemics*, vol. 10, pp. 72–77, 2015. DOI: 10.1016/j.epidem.2014.08.006.

[429] N. T. J. Bailey, *The Mathematical Theory of Epidemics*. Charles Griffin and Co. Ltd, London, 1957.

[430] Y. Tian, N. D. Osgood, A. Al-Azem, and V. H. Hoeppner, "Evaluating the effectiveness of contact tracing on tuberculosis outcomes in Saskatchewan using individual-based modeling," *Health Education & Behavior*, vol. 40, no. 1 Suppl, 98S–110S, 2013. DOI: 10.1177/1090198113493910.

[431] E. Yoneki, "FluPhone study: Virtual disease spread using haggle," in *Proceedings of the 6th ACM Workshop on Challenged Networks*, ser. CHANTS '11, 2011, pp. 65–66. DOI: 10.1145/2030652.2030672.

[432] T. R. Katapally *et al.*, "The SMART study, a mobile health and citizen science methodological platform for active living surveillance, integrated knowledge translation, and policy interventions: Longitudinal study," *JMIR Public Health and Surveillance*, vol. 4, no. 1, e31, 2018. DOI: 10.2196/publichealth.8953.

[433] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PLOS ONE*, vol. 10, no. 9, pp. 1–26, 2015. DOI: 10.1371/journal.pone.0136497.

[434] M. S. Hashemian, K. G. Stanley, and N. Osgood, "Flunet: Automated tracking of contacts during flu season," in *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, IEEE, 2010, pp. 348–353.

[435] N. Eagle and A. S. Pentland, "Reality mining: Sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006. DOI: 10.1007/s00779-005-0046-3.

[436] B. Schwartz, P. S. Moore, and C. V. Broome, "Global epidemiology of meningococcal disease," *Clinical Microbiology Reviews*, vol. 2, no. suppl, S118–S124, 1989. DOI: 10.1128/CMR.2.Suppl.S118.

[437] M. Baker, A. McDonald, J. Zhang, and P. Howden-Chapman, *Infectious diseases attributable to household crowding in New Zealand: A systematic review and burden of disease estimate*. Wellington: He Kainga Oranga/Housing and Health Research Programme, University of Otago, 2013.

[438] P. M. Coffey, A. P. Ralph, and V. L. Krause, "The role of social determinants of health in the risk and prevention of group A streptococcal infection, acute rheumatic fever and rheumatic heart disease: A systematic review," *PLOS Neglected Tropical Diseases*, vol. 12, no. 6, e0006577, 2018. DOI: 10.1371/journal.pntd.0006577.

[439] M. Génois and A. Barrat, "Can co-location be used as a proxy for face-to-face contacts?" *EPJ Data Science*, vol. 7, no. 1, p. 11, 2018. DOI: 10.1140/epjds/s13688-018-0140-1.

[440] S. Rao and W. A. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1160–1178, 1996. DOI: 10.1109/18.508839.

[441] D. Geman and A. Koloydenko, "Invariant statistics and coding of natural microimages," in *IEEE Workshop on Statistical and Computational Theories of Vision*, Fort Collins, Colorado, 1999. [Online]. Available: http://www.stat.ucla.edu/~sczhu/Workshops/sctv99.html.

[442] Apple Inc. and Google Inc., *Privacy-preserving contact tracing*, 2021. [Online]. Available: https://covid19.apple.com/contacttracing.

[443] J. Stehlé *et al.*, "High-resolution measurements of face-to-face contact patterns in a primary school," *PLOS ONE*, vol. 6, no. 8, e23176, 2011. DOI: 10.1371/journal.pone.0023176.

[444] H. Petousis-Harris, P. Carter, N. Turner, and M. Nowlan, "2012 antigen review for the New Zealand national immunisation schedule: Meningococcal B and C," Immunisation Advisory Centre, University of Auckland, New Zealand, Tech. Rep., 2014. [Online]. Available: https://www.immune.org.nz/2012-antigen-review-new-zealand-national-immunisation-schedule-meningococcal-b-and-c.

[445] A. S. Lopez and M. Marin, "Strategies for the control and investigation of varicella outbreaks manual, 2008," National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases, Tech. Rep., 2008. [Online]. Available: https://www.cdc.gov/chickenpox/outbreaks/manual.html.

[446] F. Campbell *et al.*, "Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021," *Eurosurveillance*, vol. 26, no. 24, p. 2 100 509, 2021. DOI: 10.2807/1560-7917.ES.2021.26.24.2100509.

[447] National Center for Immunization and Respiratory Diseases, Division of Viral Diseases, *Diphtheria*, Mar. 2020. [Online]. Available: https://nt.gov.au/wellbeing/health-conditions-treatments/bacterial/diphtheria#heading1.

[448] G. Gonçalves, A. Correia, P. Palminha, H. Rebelo-Andrade, and A. Alves, "Outbreaks caused by parvovirus B19 in three Portuguese schools," *Euro Surveill*, vol. 10, no. 6, pp. 121–124, 2005. DOI: 10.2807/esm.10.06.00549-en.

[449] K. A. Kho, K. Eisinger, and K. T. Chen, "Management of an obstetric health care provider with acute parvovirus B19 infection," *American Journal of Obstetrics and Gynecology*, vol. 198, no. 5, e33–e34, 2008. DOI: 10.1016/j.ajog.2007.10.779.

[450] G. D. Murray and A. D. Cliff, "A stochastic model for measles epidemics in a multi-region setting," *Transactions of the Institute of British Geographers*, vol. 2, no. 2, pp. 158–174, Jan. 1977. DOI: 10.2307/621855.

[451] M. J. Keeling and B. T. Grenfell, "Disease extinction and community size: Modeling the persistence of measles," *Science*, vol. 275, no. 5296, pp. 65–67, 1997. DOI: 10.1126/science.275.5296.65.

[452] R. Breban, J. Riou, and A. Fontanet, "Interhuman transmissibility of Middle East respiratory syndrome coronavirus: Estimation of pandemic risk," *The Lancet*, vol. 382, no. 9893, pp. 694–699, 2013. DOI: 10.1016/S0140-6736(13)61492-0.

[453] National Center for Immunization and Respiratory Diseases, Division of Viral Diseases, *Mers clinical features*, Aug. 2020. [Online]. Available: https://www.cdc.gov/coronavirus/mers/clinical-features.html.

[454] World Health Organization, *Rubella fact sheets*, Jan. 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs367/en/.

[455] P. Rohani, C. Green, N. Mantilla-Beniers, and B. Grenfell, "Ecological interference between fatal diseases," *Nature*, vol. 422, no. 6934, pp. 885–888, 2003. DOI: 10.1038/nature01542.

[456] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *American Journal of Epidemiology*, vol. 160, no. 6, pp. 509–516, 2004. DOI: 10.1093/aje/kwh255.

[457] M. Lipsitch *et al.*, "Transmission dynamics and control of severe acute respiratory syndrome," *Science*, vol. 300, no. 5627, pp. 1966–1970, 2003. DOI: 10.1126/science.1086616.

[458] M. Ceccarelli, M. Berretta, E. V. Rullo, G. Nunnari, and B. Cacopardo, "Differences and similarities between Severe Acute Respiratory Syndrome (SARS)-CoronaVirus (CoV) and SARS-CoV-2. would a rose by another name smell as sweet?" *European Review for Medical and Pharmacological Sciences*, vol. 24, no. 5, pp. 2781–3, 2020. DOI: \verb|10.26355/eurrev_202003_20551|.

[459] R. M. Anderson *et al.*, "Epidemiology, transmission dynamics and control of SARS: The 2002–2003 epidemic," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 359, no. 1447, pp. 1091–1105, 2004. DOI: 10.1098/rstb.2004.1490.

[460] D. K. Sewell, A. Miller, and CDC MInD-Healthcare Program, "Simulation-free estimation of an individual-based SEIR model for evaluating nonpharmaceutical interventions with an application to COVID-19 in the District of Columbia," *PLOS ONE*, vol. 15, no. 11, e0241949, 2020. DOI: 10.1371/journal.pone.0241949.

[461] F.-Z. Jaouimaa *et al.*, "An age-structured SEIR model for COVID-19 incidence in Dublin, Ireland with framework for evaluating health intervention cost," *PLOS ONE*, vol. 16, no. 12, e0260632, 2021. DOI: 10.1371/journal.pone.0260632.

[462] M. Altmann, B. C. Wee, K. Willard, D. Peterson, and L. C. Gatewood, "Network analytic methods for epidemiological risk assessment," *Statistics in Medicine*, vol. 13, no. 1, pp. 53–60, 1994. DOI: 10.1002/sim.4780130107.

[463] Z. Yang, J. Song, S. Gao, H. Wang, Y. Du, and Q. Lin, "Contact network analysis of COVID-19 in tourist areas—based on 333 confirmed cases in China," *PLOS ONE*, vol. 16, no. 12, e0261335, 2021. DOI: 10.1371/journal.pone.0261335.

[464] M. Ward, D. Maftei, C. Apostu, and A. Suru, "Estimation of the basic reproductive number (R0) for epidemic, highly pathogenic avian influenza subtype H5N1 spread," *Epidemiology & Infection*, vol. 137, no. 2, pp. 219–226, 2009. DOI: 10.1017/S0950268808000885.

[465] S.-Z. Huang, "A new SEIR epidemic model with applications to the theory of eradication and control of diseases, and to the calculation of R0," *Mathematical Biosciences*, vol. 215, no. 1, pp. 84–104, 2008. DOI: 10.1016/j.mbs.2008.06.005.

[466] J.-D. Van Wees *et al.*, "Forecasting hospitalization and ICU rates of the COVID-19 outbreak: An efficient SEIR model," *Bulletin of the World Health Organization*, 2020. DOI: 10.2471/BLT.20.256743.

[467] L. López and X. Rodo, "A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics," *Results in Physics*, vol. 21, p. 103 746, 2021. DOI: 10.1016/j.rinp.2020.103746.

[468] J. M. Read and M. J. Keeling, "Disease evolution on networks: The role of contact structure," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1516, pp. 699–708, 2003. DOI: 10.1098/rspb.2002.2305.

[469] K. VanderWaal, E. A. Enns, C. Picasso, C. Packer, and M. E. Craft, "Evaluating empirical contact networks as potential transmission pathways for infectious diseases," *Journal of The Royal Society Interface*, vol. 13, no. 121, p. 20 160 166, 2016. DOI: 10.1098/rsif.2016.0166.

[470] T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The importance of the normality assumption in large public health data sets," *Annual Review of Public Health*, vol. 23, no. 1, pp. 151–169, 2002. DOI: 10.1146/annurev.publhealth.23.100901.140546.

[471] S. S. Sawilowsky and R. C. Blair, "A more realistic look at the robustness and Type II error properties of the *t*-test to departures from population normality," *Psychological Bulletin*, vol. 111, no. 2, pp. 352–360, 1992. DOI: 10.1037/0033-2909.111.2.352.

[472] P. Tugwell, D. de Savigny, G. Hawker, and V. Robinson, "Applying clinical epidemiological methods to health equity: The equity effectiveness loop," *BMJ*, vol. 332, no. 7537, pp. 358–361, 2006. DOI: 10.1136/bmj.332.7537.358.

[473] M. A. Trecker, D. J. Hogan, C. L. Waldner, J. A. R. Dillon, and N. D. Osgood, "Revised simulation model does not predict rebound in gonorrhoea prevalence where core groups are treated in the presence of antimicrobial resistance," *Sexually Transmitted Infections*, vol. 91, no. 4, pp. 300–302, 2015. DOI: 10.1136/sextrans-2014-051792.