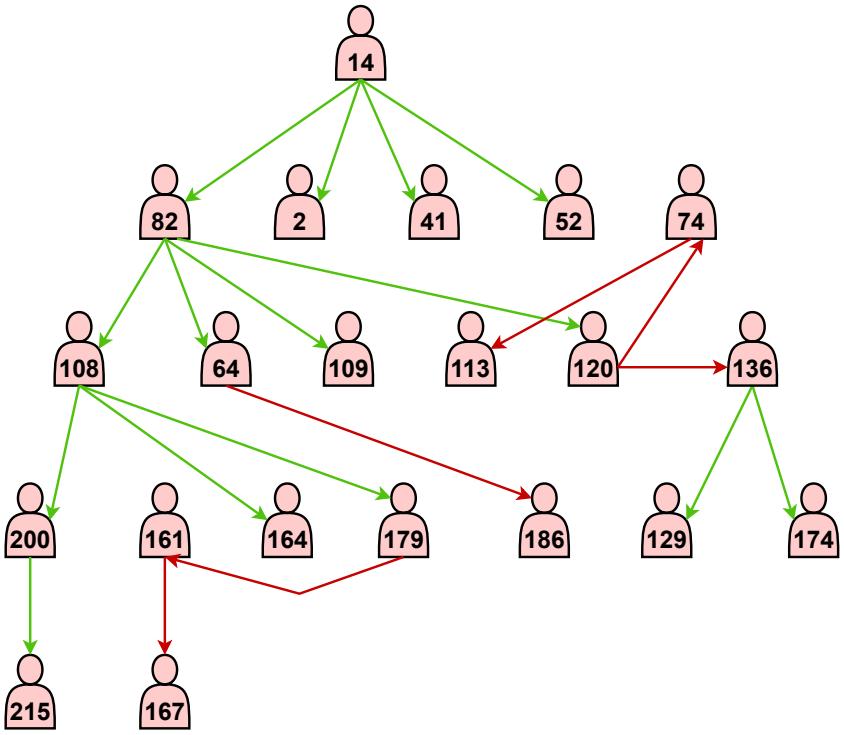
DEPARTMENT OF PHYSICS



Reconstructing Transmission Trees in Healthcare Setting using Bayesian Inference

Degree project for Master of Science (120 hec) in Complex Adaptive Systems

Minal Kumbhar

DEPARTMENT OF PHYSICS
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

Master's thesis 2024

Reconstructing Transmission Trees in Healthcare Setting using Bayesian Inference

Author: Minal Kumbhar
Supervisor: Philip Gerlee

Department of Physics
University of Gothenburg
Gothenburg, Sweden 2024

Reconstructing Transmission Trees in Healthcare Setting using Bayesian Inference
Minal Kumbhar

Supervisor: Philip Gerlee, Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden
Examiner: Kristian Gustafsson, Department of Physics, University of Gothenburg, Sweden

Master's Thesis 2024
Department of Mathematical Sciences
University of Gothenburg

Cover: Reconstructed transmission tree with the Bayesian inference

Typeset in LaTeX
Gothenburg, Sweden 2024

Reconstructing Transmission Trees in Healthcare Setting using Bayesian Inference
Minal Kumbhar
Department of Physics
University of Gothenburg

Abstract

Outbreaks of multidrug-resistant bacteria, such as *Klebsiella oxytoca*, present critical challenges to healthcare systems worldwide. Such bacteria can cause severe infections in immune-suppressed patients, spreading through contact with infected individuals, equipment, or contaminated environments. This thesis focuses on reconstructing transmission trees in healthcare systems using Bayesian modeling, focusing on the significance of data integration for effective infection control strategies. First, the study examines how patient, and contact data generated in hospitals contribute to understanding transmission trees. Second, it explores the incremental impact on inferred transmission trees' accuracy by incorporating different data sources, such as temporal, contact, diagnostic, and genetic data. Lastly, the study evaluates the effects of varying sampling on transmission inference accuracy. The results indicate that integrating temporal, contact, reporting, and genetic data enhances the accuracy of transmission tree reconstructions. Furthermore, our investigation into the impact of sampling revealed that increased sampling improves accuracy and reduces variability in transmission tree structure. Overall, this research emphasizes the importance of comprehensive data integration for effective infection control strategies and provides insights for managing outbreaks of multidrug-resistant organisms in hospital environments.

Keywords: Bayesian inference, transmission tree, MCMC, healthcare system, disease outbreak, mathematical outbreak modeling

Acknowledgements

First, I would like to thank my supervisor Philip Gerlee for our discussions, feedback sessions, and the guidance he has provided me throughout my thesis. His guidance and steady support have made this thesis experience meaningful and strengthened my ability to apply my knowledge. In addition, I would also like to thank Jon Wallér (Hygiene Doctor at Södra Älvsborgs Sjukhus, Sweden) for the helpful information he provided at the beginning, which was greatly beneficial during my thesis work.

Further, I want to express my gratitude to my husband Mahesh Kumbhar, for his constant support and encouragement throughout this process. His belief in me has been a source of strength for me. I am grateful for everything he has done to help me to achieve my goals. Finally, I would like to thank my mother Anita Kumbhar, father Madhukar Kumbhar, brother Abhijit Kumbhar and other family members, for their support. Their encouragement, love, and belief in me have been invaluable during this process, and I am thankful for their understanding and motivation.

Minal Kumbhar, Gothenburg, November 2024

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

LOS	Length of stay
P-P	Patient to patient
P-R	Patient to room
R-R	Room to room
R-P	Room to patient
SNP	Single nucleotide polymorphism

Nomenclature

Below is the nomenclature of probabilities, parameters, and variables that have been used throughout this thesis.

Probabilities

β_r	Patient to patient infection probability within room
β_w	Patient to patient infection probability within ward
γ_{rp}	Room to patient infection probability
γ_{pr}	Patient to room infection probability
P_s	Probability of sampling

Parameters

π	Proportion of cases sampled
λ_r	Non-infectious contact probability between cases within room
λ_w	Non-infectious contact probability between cases within ward

Variables

Pt_{id}	Patient Id
R_{no}	Room number
T^{adm}	Time of admission
T^{dis}	Time of discharge
T^{inf}	Time of infection
T^{samp}	Time of sampling
$InfF$	Infection status flag
$Test$	Test result

R_{con}	Connected room
T_p	Total number of patients
Inf	Infected patients list
$roomP$	Room pairs
R_i^{no}	Room number of case i
R_i^{con}	Room connected to room cases i staying
$Trans_{pairs}$	Transmission pairs list
Pt_{ConR}	patients in contaminated room
$R_{inf_{pt}}$	Room infected patient staying
R_{Cont_R}	Room connected to contaminated room

Contents

List of Figures

List of Tables

1

Introduction

Infectious disease outbreaks, particularly those caused by multidrug-resistant bacteria, pose significant challenges to healthcare facilities worldwide. *Klebsiella oxytoca* is a type of bacteria found in soil, sewage, and the digestive systems of humans and animals [1]. While usually harmless, it can cause infections, especially in people with weakened immune systems or in hospital settings [2]. Being in the hospital increases the chance of acquiring a *Klebsiella* infection, especially if someone has a cut, uses a ventilator, receives medicine intravenously or through other medical devices. Also those with underlying conditions such as diabetes, chronic lung diseases, or who have undergone invasive medical procedures have a higher risk of acquiring infection [3] [4]. This spreads through direct or indirect contact with infected people or contaminated environments. It can cause serious infections like sepsis, stomach, and lung infections. Infections may present in the urinary tract, respiratory tract, bloodstream, or wounds. The treatment might involve antibiotics, but some strains are becoming more resistant to common antibiotics like penicillin and ampicillin. It's also part of the group of bacteria that produce Extended-Spectrum Beta-Lactamase (ESBL), which makes it resistant to many antibiotics [4], including Cephalosporin and Ceftazidime. The antibiotic-resistant strains of *Klebsiella oxytoca* pose significant treatment challenges. Proper hygiene practices and infection control measures are essential for preventing the transmission of *Klebsiella oxytoca*.

Understanding the transmission dynamics of these pathogens within hospital settings is crucial for implementing effective infection control measures and preventing further spread among patients and healthcare workers. In recent years, significant research has focused on understanding the transmission dynamics of multidrug-resistant organisms, particularly in healthcare settings. A major challenge in this field is determining how pathogens spread in different environments, such as hospitals and communities, and identifying effective strategies for containment.

Mathematical modeling is crucial for understanding and managing infectious diseases by providing insights into how pathogens spread and identifying factors that drive transmission. These models allows to simulate outbreaks, predict future trends, and assess intervention impacts by incorporating biological and environmental variables. Basic models like SIR (Susceptible-Infectious-Recovered) track disease progression, while network-based models simulate contact-specific transmission [5]. With stochastic elements, these models capture real-world variability, making them essential for optimizing public health strategies, evaluating preventive measures, and enhancing disease management in both community and healthcare settings [6].

Modeling infectious diseases in hospitals involves understanding the structured and changing environment, where patients, healthcare workers, and surfaces con-

tribute to spreading infections [7]. Models designed specifically for hospitals take into account the layout of the facility, how patients move, and how well hygiene measures are followed. These models help simulate how infections spread and pinpoint high-risk areas where hygiene practices, like handwashing and equipment cleaning, are most needed. Newer models, especially those using Bayesian methods, aim to capture the complex nature of drug-resistant bacteria in hospitals, though challenges remain due to the varying patient groups, infection control practices, and environmental factors [8]. Models that combine genetic and clinical data offer more accurate pictures of infection spread and can guide targeted interventions to control hospital infections [9], [8], [10].

1.1 Aim

The overarching goal of the thesis is to develop a method for constructing transmission trees for infectious disease outbreaks in a healthcare setting, specifically focusing on *Klebsiella oxytoca*. This study aims **to develop a model for inferring transmission chains of infectious diseases in hospitals using Bayesian inference**. Understanding the dynamics of transmission within healthcare facilities is crucial for effective outbreak management and improving hygiene. Traditional epidemiological methods primarily rely on genomics and epidemiological data to reconstruct transmission chains, but the complexities of healthcare settings including interactions among patients, their admission and discharge date, healthcare workers, and environmental factors complicate this task. Ultimately, this thesis aims to provide a more accurate inference of transmission trees, facilitating improved outbreak management strategies within healthcare environments by addressing the below research questions,

Research question 1:
What type of data generated in a hospital setting can be used for inference of transmission trees?

Research question 2:
How much better can the inference of transmission trees become by adding new data (inference with only temporal data, then by adding contact data etc.) into a model?

Research question 3:
How does the number of patients sampled (tested) affect the accuracy of the transmission tree?

1.2 Limitations

Since this study does not include sampling of rooms, the inferred transmission tree cannot capture potential transmission events between patients and contaminated rooms. As a result, both patient-to-room and room-to-patient transmissions are missing from the analysis, potentially leading to an incomplete picture of transmission dynamics within the hospital setting.

The method has been tested on synthetic data, which relies on multiple assumptions that may not hold in real-life scenarios. For instance, assumptions include that

all diagnostic tests are perfectly accurate with no false positives or negatives and that all patient contacts are fully reported. Such assumptions are rarely met in real hospital data, meaning this method may yield different results or accuracy when applied directly to real-world data. Adjustments to the method may be necessary to improve accuracy with real data.

Additionally, the transmission model in this method considers only a single ward. However, real hospitals have multiple wards with fluctuating patient counts and outpatient visits, which are not accounted for in this study.

2

Theory

This section covers previous work on reconstructing transmission trees and some background on transmission models and Bayesian inference.

2.1 Previous Work

Since the late 1980s, extended-spectrum β-lactamase (ESBL)–producing Enterobacteriaceae have become a major cause of hospital infections and outbreaks [11]. As a result, healthcare systems are facing significant challenges in controlling these outbreaks leading to the need for better monitoring and action plans. Many research efforts have been started to understand the spread, causes, and genetic factors related to ESBL-producing outbreaks. Recent improvements in statistical methods and genetic analysis have made it easier to track how these infections spread, providing important information that can help improve infection control in hospitals.

Transmission trees, which trace the spread of an infectious disease from one individual to another, have been a central focus in epidemiological studies for decades. Understanding these trees is crucial for planning effective measures and predicting how outbreaks will progress. Several approaches such as simple deterministic models and complex stochastic simulations have been developed to model transmission trees [8], [12], [13], [14]. In recent years, Bayesian inference has become a powerful tool in this context because it helps account for uncertainty and integrate diverse sources of data, such as epidemiological data and genetic information. By applying Bayesian methods, probabilistic transmission trees can be generated that provide deeper insights into the dynamics of disease spread.

Jombart *et al.* (2014) [15] developed a method, implemented in the R package *outbreaker*, which combines infection timing and pathogen genetic sequences to infer key outbreak features such as transmission trees, infection dates, secondary infections, and unobserved cases. They applied this approach to the 2003 SARS outbreak in Singapore. Similarly Didelot *et al.* (2017) [13] used a Bayesian framework but focused on inferring transmission trees by addressing within-host genetic diversity and partial sampling in ongoing outbreaks. Their method, which integrates phylogenetic analysis with epidemiological models, uses a reversible jump MCMC algorithm to explore transmission scenarios, by handling incomplete data and unsampled intermediates, which improves outbreak reconstruction accuracy. While, in 2020, Cassidy *et al.* [8] proposed a discrete-time transmission model for an outbreak of Methicillin-Resistant Staphylococcus Aureus (MRSA) in a hospital setting, with a combination of epidemiological data (such as patient admission and discharge

times) and whole-genome sequencing data from pathogen isolates. This study extended the work of Worby *et al.* (2014) [14], which considers homogeneous mixing between colonized and susceptible patients.

Further, Campbell *et al.* (2019) [9] improved outbreak reconstruction methods by building on Jombart *et al.*'s work with a Bayesian model called *outbreaker2*. This model integrates contact tracing data with epidemiological and genetic information, which makes transmission tree reconstruction more accurate. It addresses the limits of using only genetic data, which can be less useful when there is low genetic diversity or when individuals carry multiple pathogen strains. The *outbreaker2* model showed that even incomplete contact tracing data can improve the accuracy of identifying transmission events. This approach was also used for the 2003 SARS outbreak in Singapore, showing that combining contact, genetic, and epidemiological data gives a better understanding of how diseases spread. Lindsey *et al.* (2020) [10] extended the *outbreaker2* model and applied Bayesian inference to investigate within-hospital SARS-CoV-2 transmission events by combining viral genetic and epidemiological data. Their approach demonstrated the strength of Bayesian methods in integrating diverse datasets and handling uncertainty by providing a robust framework for understanding transmission dynamics in hospital settings.

Many studies have used Bayesian statistical methods for reconstructing disease outbreaks (transmission trees) by using epidemiological and genetic data [15], [8], [12] with only a few incorporating contact data to reconstruct transmission trees [9], [14]. This report aims to address this gap by integrating contact data to further refine transmission chain reconstruction. The *outbreaker2* model will be extended in a hospital setting, highlighting that interactions between patients, as well as the contamination of the environment, are critical factors in healthcare transmission dynamics.

2.2 Transmission Models

Transmission models are essential tools in epidemiology, providing insights into how infectious diseases spread within populations.

2.2.1 Reed-Frost Model

The Reed-Frost model is a fundamental epidemiological framework that describes the spread of infectious diseases in a closed population. Developed in the early 20th century, it uses a probabilistic approach to illustrate how infections propagate, making it particularly relevant for reconstructing transmission trees in healthcare settings [16].

The Reed-Frost model explains how infectious diseases spread in a closed group of people. In this model, the infection is passed directly from one infected person to others through specific interactions called "adequate contact". When a healthy person has adequate contact with someone who is infected during a set time period, they will catch the infection. Once infected, they can spread the disease to others for a short time before becoming completely immune. Each person has the same chance of having adequate contact with anyone else in the group, and these chances

stay the same throughout the outbreak. Additionally, individuals do not interact with anyone outside the group, keeping the spread of the infection contained within the population [17].

The model is formulated in discrete time steps, allowing for the simulation of successive generations of infection. At each time step, a susceptible individual can become infected if they contact an infected individual. The dynamics are typically represented through the following equation:

$$I_{t+1} = S_t.(1 - (1 - p)^{I_t}) \tag{2.1}$$

where:

- I_{t+1}**:** The number of newly infected individuals in the next time step, $t + 1$.
- S_t: The number of susceptible individuals at the current time step t, who are still uninfected.
- p: Probability of infection occurring between any two individuals with "adequate contact."
- $(1 - (1 - p)^{I_t})$: The probability that a susceptible individual will become infected, given contact with I_t currently infected individuals.

Susceptible individuals are infected based on their chance of adequate contact with currently infected cases, simulating transmission across discrete time steps. This equation allows for the calculation of infection spread generation by generation.

2.2.2 SIR Model

The epidemic susceptible-infected-removed (S-I-R) model is a well-known compartmental framework that serves as a foundational model for infectious diseases. This model enhances the Reed-Frost model by introducing distinct compartments for susceptible, infected, and removed individuals, allowing for a more nuanced understanding of disease dynamics. However, the S-I-R model is primarily valid for closed populations and requires further consideration when applied to hospital data [18].

In hospital settings, the migration process such as patient admissions and discharges plays a critical role in the dynamics of infection spread and cannot be overlooked. For instance, patients already colonized with pathogens may enter the unit, influencing the epidemic process. New uncolonized patients represent a susceptible population at risk of acquiring infections. Additionally, susceptible patients may be discharged (or may die) before they acquire the pathogen, making discharge a competing event for colonization. Consequently, the transmission rate must be interpreted in conjunction with the discharge rate of susceptible patients within a competing-risk framework. To address these complexities, Wolkewitz *et al.* extend this further with the martingale-based statistical methods to make them applicable to populations with migration, specifically focusing on patients in hospitals who enter and leave the unit during their stay [18]. This approach emphasizes the time-dependent nature of the transmission rate and its interpretation within a competing-event framework. This method allows for a more accurate description

of epidemic spread at the unit level in hospitals and provides valuable insights for clinicians managing infectious disease outbreaks.

2.3 Bayes' Rule

Bayesian methods can be used to estimate the model's parameters based on observed data, providing a way to incorporate prior knowledge and uncertainty into the predictions [15], [19], [10], [8]. By applying Bayesian inference techniques, transmission trees can be reconstructed and identify sources of infection, enhancing our understanding of outbreak dynamics.

The basis for Bayesian inference is Bayes' rule which describes how to update probabilities based on new evidence or data [20]. As new data is observed, the posterior can be used as a new prior, and the process can be repeated (sequential updating). The prior refers to the initial probability distribution representing beliefs or knowledge about a parameter before observing any new evidence. It reflects what is known about the parameter based on previous information or assumptions. The posterior, on the other hand, is the updated probability distribution that results from applying Bayes' rule after incorporating new data or evidence. This posterior distribution then serves as the new prior for future updates as additional data becomes available.

Bayes' theorem is a mathematical statement of conditional probability. Given two events, A and B, Bayes' theorem states that the probability of event A occurring, given the occurrence of another event B, is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.2}$$

In Bayesian inference, we start with a formula 2.2 to help us work with our data. We use a likelihood function that includes the parameters we are trying to estimate. The main goal is to find out what the parameter values are after considering the data we have. We can write this mathematically as finding the probability of the parameters θ given the data D, which depends on how likely the data is given the parameters and our initial beliefs about those parameters.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}, \tag{2.3}$$

where:

- θ represents a set of parameters or hypotheses.
- D represents the observed data.
- $P(\theta|D)$ is the posterior probability of θ, given the data D.
- $P(D|\theta)$ is the likelihood, the probability of observing the data D given θ.
- $P(\theta)$ is the prior probability of θ, representing the prior belief before observing the data.
- $P(D)$ is the marginal likelihood or evidence, a normalizing constant ensuring that the posterior is a valid probability distribution.

$P(D)$ is the integral of the product of the likelihood and the prior with respect to θ, evaluated over the range of possible values of θ.

$$P(D) = \int_{\Theta} P(D|\theta)P(\theta)d\theta \tag{2.4}$$

Since the parameters have been removed from consideration, the value of $P(D)$ is not influenced by θ. Therefore, a modified version of Bayes' rule, shown in Equation (2.4), is frequently used as:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{2.5}$$

This implies that the posterior distribution can be obtained up to a constant of proportionality.

$$P(\theta|D) = KP(D|\theta)P(\theta), \tag{2.6}$$

where $K = 1/P(D)$. In practice, calculating K is often too complex. Because of this, the practical use of Bayesian methods for statistical inference was limited for many years. It was only feasible in certain special cases where the likelihood and prior were chosen in a way that made the calculations easier, allowing direct sampling from the posterior distribution. However, with modern Markov Chain Monte Carlo (MCMC) methods, it's now possible to sample from posterior distributions without calculating K directly [21].

2.4 Bayesian Inference

Bayesian inference is a method of statistical reasoning where beliefs or confidence about different possible outcomes are updated as new information or data becomes available. Rather than adhering to a fixed assumption, this approach allows for the continuous adjustment of how much one trusts each possibility based on the evidence at hand [22]. The outcome is affected by several factors, each with its own starting likelihood based on what we've seen before. If some factors don't have enough evidence or examples, the other factors become more important in shaping the outcome, which helps guide the inference process. When new observations are made, the chance of not seeing a certain factor goes down, and the other factors are reassessed. This updated likelihood multiplied by prior is called the posterior probability. This posterior probability then becomes the starting point for future observations, creating a continuous cycle of updating our beliefs as we gather more data [23].

2.4.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are computational techniques designed for sampling from complex probability distributions, especially useful in Bayesian inference and in contexts where direct sampling is infeasible. MCMC methods construct a Markov chain—a sequence of random variables in which each state

depends only on the previous one to approximate the desired distribution. Fundamental algorithms, such as the Metropolis-Hastings algorithm and Gibbs sampling, Random Walk Metropolis, and Hamiltonian Monte Carlo provide frameworks for generating these Markov chains. No matter the algorithm, the sampling will generate a Markov chain of samples, each of which is correlated with nearby samples. The Metropolis-Hastings algorithm, introduced by Metropolis [24]. and later expanded by Hastings [25], is foundational in MCMC, offering a way to handle asymmetrical proposals through an acceptance-rejection step. This makes it applicable in a broad range of inference tasks involving complex likelihoods. Gibbs sampler of MCMC developed by Gelfand and Smith [26], simplifies high-dimensional sampling by sequentially drawing each variable from its conditional distribution, making it especially efficient when conditional distributions are known. Burn-in and thinning are common techniques used to improve the quality of samples from a Markov chain.
Burn-in: At the start of an MCMC simulation, samples may not accurately represent the target distribution due to the chain's initial state. To mitigate this, the initial part of the chain, known as the burn-in period, is discarded. By removing these early samples, it is ensured that the retained samples are more representative of the steady-state or stationary distribution of the chain [27].
Thinning: MCMC chains often have high autocorrelation, where successive samples are highly similar. Thinning reduces this by retaining only every n^{th} sample, which lowers autocorrelation and makes the remaining samples more independent. This technique can reduce storage and computational costs when working with large data sets [27].

2.4.2 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a method in Markov Chain Monte Carlo (MCMC) for obtaining samples from complex probability distributions, especially when direct sampling is challenging [25], [24]. By iteratively proposing candidate states based on an initial distribution and accepting or rejecting these states based on the target distribution, the algorithm helps generate samples that approximate the desired distribution over time [28].
The core steps of the Metropolis-Hastings algorithm are as follows:

- **Initialize**: Start with an initial state x_0. The process begins with an initial state, often chosen based on prior knowledge or random selection.
- **Generate a Proposal**: Each step proposes a new candidate state using a proposal distribution, which could be symmetric (like Gaussian) or asymmetric, depending on the problem. At each step t, propose a new candidate state x' based on a proposal distribution $Q(x' \mid x_t)$, which depends on the current state x_t.
- **Compute Acceptance Probability**: Acceptance ratio is calculated to assess the likelihood of moving to the new state. This ratio includes the target distribution and proposal terms.

  - Calculate the acceptance ratio α:

$$\alpha = \min\left(1, \frac{P(x')Q(x_t \mid x')}{P(x_t)Q(x' \mid x_t)}\right)$$

  - This ratio α compares the probability of the new state relative to the current state, adjusting for any asymmetry in the proposal distribution Q.
- **Accept or Reject**:
  - Draw a random number u from a uniform distribution $U(0, 1)$.
  - If $u < \alpha$, accept x' as the next state; otherwise, retain x_t.
- **Iterate**: Repeat steps 2–4 for a sufficient number of steps to approximate the target distribution.

3
Methods

In this chapter, the methods and data are presented. First, the process of synthetic data generation is described and followed by the framework for Bayesian inference.

3.1 Model for Disease Transmission

The transmission model is based on the spread of the pathogen from infected patients to other patients and the environment. Since the environment can also become infected and transmit the pathogen to patients, the model includes four types of transmissions: patient-to-patient (P-P), patient-to-room (P-R), room-to-patient (R-P), and room-to-room (R-R). The P-P within room, P-P within ward, P-R, and R-P transmissions occur with the probabilities β_r, β_w, γ_{pr}, γ_{rp} respectively. If two sinks are connected by a pipe, the pathogen can spread from one sink to the other, which is the R-R transmission. Unlike the other transmission types (P-P, P-R, R-P), which involve some probability, R-R transmission occurring through a shared sewage pipe is considered a direct pathway rather than a probabilistic one. Since a connected pipe creates an uninterrupted route for pathogen movement between sinks, the model assumes that pathogen spread will occur whenever contamination is present, without probabilistic barriers.

Since transmission can be both direct and indirect, each patient has a different chance of coming into contact with other patients. If two patients share the same room, the probability of contact between them is higher and they can infect each other more easily, with a probability of β_r. If the two patients are in the same ward but in different rooms, there is still a chance of contact, such as through the hands of healthcare workers or contaminated medical devices, but the probability of transmission is slightly lower than for patients sharing the same room. In this case, the pathogen can spread from one patient to another within the ward with a probability of β_w. If transmission occurs from a patient to a room (P-R), the pathogen can be transmitted to the room with a probability of γ_{pr}. On the other hand, if the transmission is from a room to a patient (R-P), it occurs with a probability of γ_{rp}, which is lower than γ_{pr}. I assume that the following relationship holds between different transmission probabilities:

$$\beta_r > \beta_w > \gamma_{pr} > \gamma_{rp}. \tag{3.1}$$

The numerical values for these probabilities can be seen in table 3.1

β_r is greatest in all values as, there will be more frequent contact between patients within room than within ward (β_w), and these contacts could transmit the pathogen. Patients can transmit pathogen to room, but probability of transmission is less than β_r and β_w. The pathogen is also transmitted from room to patient but it's probability is smaller than P-R, so γ_{rp} is the smallest probability of transmission.

I assumed that a patient who becomes infected on day x, can transmit the pathogen from day $x + 1$ to their discharge date or until it is positive and the patient is isolated in a special room. However, if a room becomes infected on day x, it can transmit the pathogen from day $x + 1$ as long as it continues to have patients in it since the patients are only tested.

Further infection/transmission	Probability	Value
Patient to patient within room	β_r	0.05
Patient to patient within ward	β_w	0.03
Patient to the room	γ_{pr}	0.025
Room to the patient	γ_{rp}	0.02

Table 3.1: Notations and probability of further infection from an infected case to a new patient or room

3.1.1 Data Generation

This section describes the generation of data used in our study. The data we are generating for our study is epidemiological data, which contains patient ids, with room no. of patients, admission and discharge date of each patient, length of stay (LOS) of each patient, and diagnostic test or sampling date.

The data is simulated for one ward, which contains 10 rooms, and each room has two patients. Therefore, at a time 20 patients are staying in a ward. For the first 20 patients, the admission date was kept the same for convenience and the discharge date of patients is calculated based on each patient's length of stay. The distribution of LOS is considered right-skewed beta geometric distribution [29] as very few patients stay in the hospital for a long time, while most patients are discharged relatively quickly. So, most of the patients within a ward are discharged after a few days of stay while very few are discharged after a long time. As soon as a patient is discharged, a new patient is admitted to that room the next day, ensuring that the ward is always full (Appendix A in Algorithm 6).

Since *Klebsiella oxytoca* can spread through the environment, such as from one sink to another, the room number is also considered (each room has its sink). The index case is introduced randomly within the first five days of an outbreak. From the next day of infection of the first case, the pathogen begins to spread to other patients and rooms according to the probabilities β_r, β_w, γ_{pr}, γ_{rp} i.e. probability of transmission P-P within a room, P-P within a ward, P-R, and R-P respectively. This continues until the infected patient is tested positive and is moved to a separate room or discharged. The newly infected cases will then continue to spread the infection to rooms and patients. An infected patient can spread the infection to others until discharged, but if the room is infected, it continues to infect patients

staying in that room with probability γ_{rp}. As the disease is transmitted, diagnostic tests are also performed from the first day of simulation. Every day, patients are sampled randomly with sampling probability P_s. The results of diagnostic tests are considered as perfect, therefore we do not have false positive or false negative test results. So, if the patient is infected that will show a positive test result. Then that positive case is isolated in a separate room and a new patient is added into that room the next day. The data is simulated for 60 days. The pseudocode for synthetic data generation can be found in Appendix A, Algorithms 7, 8.

For all infected patients genetic distances of pathogen sequence between two patients are calculated. Genetic distance is a matrix of Hamming distance between genetic sequences of two cases. It is based on single nucleotide polymorphism (SNP). In Figure 3.1, patient 1 to 2 or patient 2 to 8 there is direct transmission from one patient to another patient. For such patients who has direct transmission genetic distance is randomly assigned between 1 to 30 number of mutations. For other patients who do not have direct transmission (genetic distance between patient 2 and 4, 7 and 12) it is calculated based on shortest path between those two cases. Therefore, the distance between patient 7 and 12 is calculated as

$$D(7,12) = D(7,2) + D(2,1) + D(1,4) + D(4,12)$$
$$D(7,12) = 12 + 7 + 3 + 4$$
$$D(7,12) = 26$$

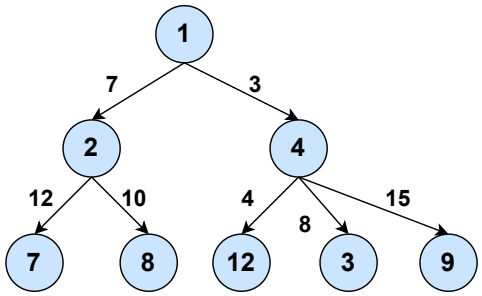


Figure 3.1: Disease transmission tree. Nodes represent patient Ids, whereas number on edges show number of mutations (genetic distance), and arrows represent the direction of transmission

In addition, the contact matrix is calculated with the help of admission and discharge dates. The contact matrix C is a $N \times N$ square binary matrix, where each row/column represents a reported case.

$$c_{ij} = \begin{cases} 1 & \text{, when i and j stays in hospital at the same time} \\ 0 & \text{, otherwise.} \end{cases}$$

Symbol	Type	Description
i	Data	Index case
s_i	Data	Genetic sequence of case i
T^{samp}	Data	Time of sampling
T^{adm}	Data	Time of admission
C	Data	Contact matrix
α_i	Augmented data	Index of the most recent sampled ancestor of case i
κ_i	Augmented data	Number of generations between cases α_i and i
T_i^{inf}	Augmented data	Time of infection of case i
w	Function	Generation time distribution
f	Function	Difference in infection and sampling time distribution
$d(s_i, s_j)$	Function	Number of mutations between s_i and s_j
$l(s_i, s_j)$	Function	Number of comparable nucleotide positions between s_i and s_j
π	Parameter	Proportion of cases sampled
ϵ	Parameter	Proportion of contacts reported
λ_r	Parameter	Non-infectious contact prob between cases within room
λ_w	Parameter	Non-infectious contact prob between cases within ward

Table 3.2: Data, parameters, and functions used in the inference method.

3.2 Framework for Bayesian Inference of Transmission Trees

The model I have developed aims to reconstruct the transmission tree with the help of generated epidemiological, genetic, and contact data. The model looks at the likelihood of different transmission scenarios for each infected case, based on when the case is sampled, admitted to the hospital, the distance between the pathogen's genetic sequence, and the contact of cases. Two key assumptions are made about the timing of transmission between individuals. One is the generation time distribution which refers to the statistical distribution that describes the time interval between two successive cases in the transmission tree. It gives us insight into how quickly an infection spreads from one person to another. The second is the distribution of the delay between the time of infection and the sampling. It gives us insight into how much delay is there between the time of infection and sampling. Both the generation time distribution w and delay period distribution f are assumed to be already known, so they are not calculated during the model's analysis. The overall structure of Bayesian inference of transmission trees represented in Figure 3.2

3.2.1 Likelihood

To reconstruct the transmission tree using Bayesian inference, the model considers four types of likelihoods - genetic, temporal, contact, and reporting likelihood, which are discussed in this section.

3.2.1.1 Genetic Likelihood

The genetic distance model is based on a single nucleotide polymorphism (SNP) analysis. It refers to a variation in a single nucleotide (A, T, C, or G) at a specific

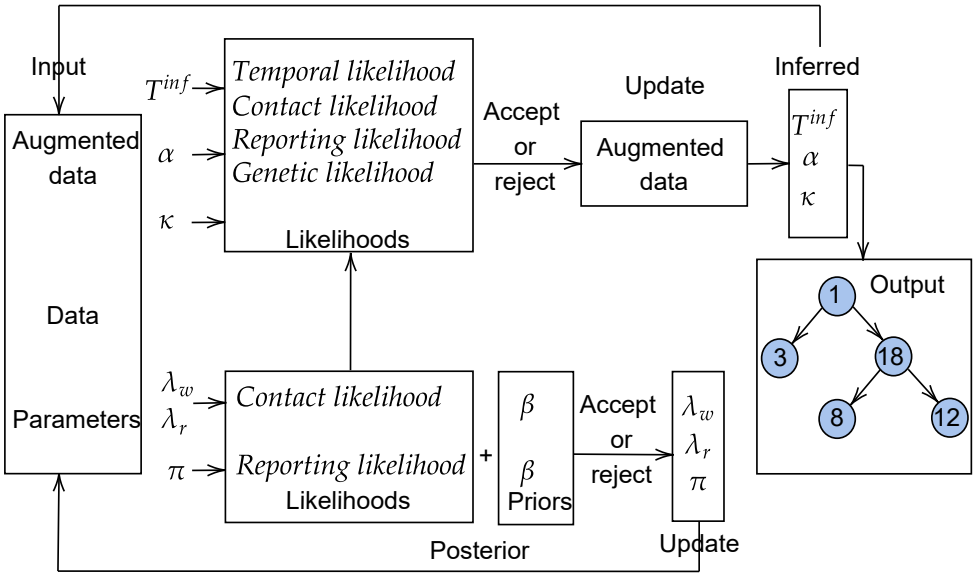


Figure 3.2: Framework for Bayesian Inference of Transmission Trees. Initially, the model starts with initialized augmented data and parameters, and the respective likelihood and posterior are calculated. For each subsequent iteration augmented data and parameters are being updated based on recent and new posterior. Finally, augmented data and parameters are sampled for the reconstruction of an inferred transmission tree.

position in the genome that occurs in a significant proportion of the population. The model assumes that each sequence can undergo mutation with SNP when the pathogen is transmitted from one case to another case. For all infected patients, genetic distance is calculated with the help of the shortest path between that pair.

The genetic likelihood measures the probability of observing the genetic distance between a case and its ancestor in an epidemiological transmission chain. It is calculated as follows: First, the transmission chain is traced to find ancestors with genetic data. Then the number of mutations between a case and its sequenced ancestor is calculated, with the help of a distance matrix. The mutation probability (μ) and the number of generations (κ) are used to compute the log-likelihood for each pair of cases and ancestors. Finally, Summing the log-likelihood across all cases in the dataset completes the genetic likelihood.

The genetic likelihood of a specific case i, the probability of observing a genetic distance $d(s_i, s_{\alpha_i})$ between the genetic sequence s_i and that of its most recent sampled ancestor s_{α_i}. The case i and it's ancestor α_i are separated by κ_i generations of infection. The number of comparable nucleotide positions between s_i and s_{α_i} is denoted as $l(s_i, s_{\alpha_i})$ [9]

$$\text{Genetic likelihood} = \Omega_i^1 = p(s_i|\alpha_i, s_\alpha, \kappa_i, \mu) \tag{3.2}$$

and it is calculated as:

$$(\kappa_i, \mu)^{d(s_i, s_{\alpha_i})}(1 - \kappa_i\mu)^{l(s_i, s_{\alpha_i}) - d(s_i, s_{\alpha_i})} \tag{3.3}$$

This calculation determines the probability of seeing some mutations (denoted as $d(s_i, s_j)$) between cases and it's ancestor at certain nucleotide positions in a DNA sequence while ensuring that no mutations happened at the other positions. The mutation rate is not updated in each model, it is kept fixed. The pseudocode for genetic likelihood can be found in Algorithm 1.

Algorithm 1 llGene

Description: This computes log-likelihood for genetic data based on the number of mutations and the generation count.

Require: n : Case index, L : length of the sequence, D : distance matrix, α : $ancestors$, κ : ancestor's generation, μ : mutation rate

if $\alpha[n] \neq 0$ & $\kappa[n] \neq 0$ **then**
  $N_{mut} \leftarrow D[n, \alpha[n]]$ ▷ Number of mutations
  $res \leftarrow N_{mut} \cdot \log(\kappa[n] \cdot \mu) + (L - N_{mut}) \cdot \log(1 - \kappa[n] \cdot \mu)$ res
else
  $res \leftarrow -\infty$
end if

Output: res

3.2.1.2 Temporal Likelihood

The temporal likelihood helps to find the temporal relationships between transmission pairs in the model. It consists of two parts, such as the likelihood of time of infection and the time of sampling. The likelihood of time of infection evaluates the likelihood of the proposed time of infection, incorporating information from the generation time distribution (w). The generation time distribution is assumed to be a negative binomial distribution. The negative binomial distribution is used to model the generation time because it captures the variation in how long infected individuals avoid transmission. It models the number of failures, i.e., days avoiding transmission, in a sequence of independent trials (each day) before reaching a specified number of successes, i.e., transmitting the infection. The likelihood of sampling describes the probability of estimating the time of infection under the distribution of delay between the time of infection and sampling (f). The distribution of delay between time of infection and sampling is assumed to be a uniform distribution as time of infection can be anywhere uniformly between admission and sampling date.

It is based on the form described in [9], but using admission and sampling time data instead of onset data, and with a delay distribution between infection time and sampling instead of incubation period distribution. The pseudocode can be found in Algorithms 2 and 3.

$$\text{Temporal likelihood} = \Omega_i^2 = p(T^{samp}|T_i^{inf}, T^{adm}, T^{samp})p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i) \quad (3.4)$$

and it is calculated as:

$$f(T^{samp} - T_i^{inf})w^{\kappa_i}(T_i^{inf} - T_\alpha^{inf}) \quad (3.5)$$

The first term in the above equation $f(T^{samp} - T_i^{inf})$ is a distribution of delay between the time of infection and sampling. This distribution is considered a uniform

distribution with the minimum value as T^{inf} and the maximum value as T^{samp}. The second term $w^{k_i}(T_i^{inf} - T_\alpha^{inf})$ describes the probability of observing the delay between the time of infection of the current case and its most recent sampled ancestor under the generation time distribution, w over the imputed number of generations, κ. $w^\kappa = w * w * w * \cdots * w$ where $*$ is the convolution operator and is applied κ times.

Algorithm 2 llSamp

Description: Calculates the log-likelihood of sampling for an individual case.

Require: i: case index, T^{inf}: time of infection, T^{samp} : time of sampling, $fDens$: delay period distribution

1: $delay \leftarrow T^{samp} - T^{inf}[i]$
2: **if** $1 \leq delay \leq \text{length}(fDens)$ **then**
3: $\quad res \leftarrow (\ln fDens)\,[delay]$
4: **else**
5: $\quad res \leftarrow -\infty$
6: **end if**

Output: res

Algorithm 3 llInf

Description: This computes log-likelihood for infection timing based on individual case.

Require: i: case index, α_i: ancestor of case i, κ_i: number of generations, T^{inf} : time of infection, $wDens$: generation time distribution

1: $delay \leftarrow T^{inf}[i] - T^{inf}[\alpha_i]$
2: **if** $1 \leq delay \leq \text{length}(wDens[0])$ & $1 \leq \kappa_i \leq \text{length}(wDens)$ **then**
3: $\quad res \leftarrow (\ln wDens)\,[\kappa_i][delay]$
4: **else**
5: $\quad res \leftarrow -\infty$
6: **end if**

Output: res

3.2.1.3 Contact Likelihood

This is a hierarchical model with two processes; the occurrence of contact and then the reporting of contact (Figure 3.3). Transmission pairs experience contact with probability (η). Sampled infected individuals that do not constitute a transmission pair experience contact with another probability (λ) i.e. non-infectious contact probability. Contacts that have occurred, whether between transmission pairs or non-transmission pairs, are then reported with probability (ϵ), which is assumed as one for all reported contacts. The contact likelihood quantifies the probability of observing the provided contact data given a proposed transmission tree and associated parameters such as ϵ and λ. As per the approach by Campbell and others [9]. we consider the non-infectious contact probability within ward, and room as λ_w, and λ_r respectively. Individuals may have frequent contact with each other, but not

all of these contacts result in transmitting the infectious pathogen. Therefore, we establish different probabilities for the likelihood of non-infectious contacts, denoted as

$$\lambda_w > \lambda_r$$

λ_r is the probability of staying together for n days without transmission within room. Therefore, it is calculated as $(1-\beta_r)^n$ where n is average LOS. Further, I assume that $\lambda_w = 1.5 \times \lambda_r$. The contacts between patients either between transmission or non-transmission pairs are then reported with probability ϵ. The ϵ is considered as one as we already have perfect contact information. We assume that each transmission is reported as contact therefore set η as one. The pseudocode for contact likelihood is given in Algorithm 4

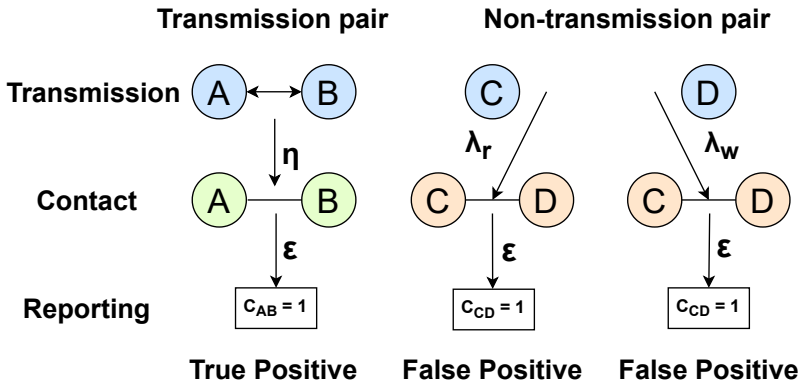


Figure 3.3: Representation of contacts between transmission and non-transmission pairs with their reporting probability. Circles represents sampled cases and $C_{ij} = 1$ indicate reported contacts (adapted from [9]).

$$\text{Contact likelihood } = \Omega_i^3 = \prod_{i=1, j \neq i}^{N} p(C_{ij} | \alpha_i, \kappa_i, \epsilon, \lambda_r, \lambda_w) \tag{3.6}$$

Using contact model above in Figure 3.3 the contact likelihood is calculated as:

$$p(C_{ij} = 1 | \alpha_i = j, \kappa_i = 1) = \eta * \epsilon = \epsilon$$
$$p(C_{ij} = 1 | \alpha_i \neq j) = p(C_{ij} = 1 | \alpha_i = j, \kappa_i > 1) = \lambda_r * \epsilon$$
$$p(C_{ij} = 1 | \alpha_i \neq j) = p(C_{ij} = 1 | \alpha_i = j, \kappa_i > 1) = \lambda_w * \epsilon$$

Since, $\epsilon = 1$ in this study, the above expressions can be simplified further:

$$p(C_{ij} = 1 | \alpha_i \neq j) = p(C_{ij} = 1 | \alpha_i = j, \kappa_i > 1) = \lambda_r$$
$$p(C_{ij} = 1 | \alpha_i \neq j) = p(C_{ij} = 1 | \alpha_i = j, \kappa_i > 1) = \lambda_w$$

3.2.1.4 Reporting Likelihood

The reporting likelihood evaluates the likelihood of not observing certain intermediate cases in the transmission chain, i.e. it considers the possibility that some cases in the chain might go unnoticed or unobserved during the outbreak investigation.

$$\text{Reporting likelihood} = \Omega_i^4 = p(\kappa_i | \pi), \tag{3.7}$$

Algorithm 4 llCont

Description: This computes log-likelihood based on contact data between cases.

Require: N : number of cases, C : contact matrix, α : ancestors, κ: ancestor's generation, ϵ : reported contacts, λ_r : Non-infectious contacts within room, λ_w : Non-infectious contacts within ward

$TP = 0.0,\ FP = 0.0$
for $j = 1$ to N **do**
  if $\alpha_j > 0$ and $\alpha_j \leq N$ **then**
    $TP \leftarrow TP + C[j, \alpha_j]$
  end if
end for
$FP \leftarrow N - TP$
$res \leftarrow TP \cdot \log(\epsilon) + FP \cdot \log(\epsilon \cdot \lambda_r) + FP \cdot \log(\epsilon \cdot \lambda_w)$

Output: res

which is calculated as

$$NB(1|\kappa_i, \pi). \tag{3.8}$$

The probability mass function of the negative binomial distribution is used to calculate the likelihood of missing intermediate cases in an outbreak, based on the proportion of cases sampled π. The negative binomial distribution is used because it models the chance of detecting one observed case while accounting for the possibility of missing several unobserved cases in the transmission chain. The dispersion parameter κ_i reflects the variability in the number of undetected cases, which makes it useful for understanding outbreaks where some infections may go unnoticed. Reference for pseudocode is in Algorithm 5

Algorithm 5 llRep

Description: This computes the log-likelihood of not observing certain intermediate cases in the transmission chain.

Require: N: number of cases, K: value from $wDens$ matrix, κ: number of generations, π: contact reporting probability

for $j = 1$ to N **do**
  if $\kappa_j < 1 || \kappa_j > K$ **then**
    $res \leftarrow -\infty$
  else
    $res \leftarrow \log(\pi) + \kappa_j * \log(1 - \pi)$
  end if
end for

Output: res

3.2.2 Prior Distribution

π, λ_r, and λ_w represent probabilities and are assigned Beta distributed priors with shape parameters $(10, 1), (1, 1), (1, 1)$ respectively.

3.2.3 Posterior Distribution

The updated joint posterior distribution is proportional to the product of the joint prior and four likelihood terms.

$$p(A, \theta|D) \propto p(\pi, \lambda_r, \lambda_w) \prod_{i=1}^{N} \Omega_i^1 \Omega_i^2 \Omega_i^3 \Omega_i^4 \tag{3.9}$$

3.2.4 Bayesian Inference

The model infers ancestors (α), number of generations (κ), and time of infection (T^{inf}) as augmented data, using MCMC sampling. The MCMC is iterated 4000 times with burn-in period 100. The initial T^{inf} is estimated for each case by uniformly sampling between the admission and sampling date. The initial ancestral relationships between cases are based on their infection dates if the model does not have genetic data. For each case, potential ancestors are identified based on whether their infection occurred before the current case and if the sampling date of the current case falls after the admission date. Out of possible ancestors, the ancestor with the closest infection time is selected as the most likely ancestor. When the model has genetic data, the initial ancestor is based on the genetic distance between cases. The only samples collected earlier than another are considered as potential ancestors. If the sampling date of the supposed ancestor is later than the target sample then those ancestors are excluded from the possible ancestors. The most likely ancestor is selected with the smallest DNA distance, which represents the most genetically similar and temporally feasible ancestor. The initial value of κ is set to one, due to the unknown number of generations between a case and it's ancestor.

The value of α is updated for each case in a Bayesian inference model as shown in Figure 3.2. For each case it first computes the current temporal-, contact-, reporting-, and genetic-log-likelihood for the existing ancestor (Figure 3.2). Then a new ancestor is proposed using an infection time (similar to the initialization of α without genetic data), and its log-likelihood is computed similarly. The acceptance of the new ancestor is determined by comparing the new and old log-likelihoods, with the random acceptance probability, which is drawn from a uniform distribution between 0 and 1. If the likelihood ratio between the new and old values is greater than this random threshold, the new ancestor is accepted and replaces the old one (Appendix A in Algorithm 10 lines $15 - 34$).

Similarly, the value of κ is also updated for each case in a Bayesian inference model which can be seen in Figure 3.2. While updating κ old log-likelihoods are calculated similarly to α. To propose a new κ random step either ± 1 is generated, and new log-likelihood is calculated, as long as κ stays within valid bounds (1 to the maximum delay). The new κ's acceptance is determined by comparing the new and old likelihoods with a random acceptance threshold similar to α. If the new κ improves the model's likelihood sufficiently, it replaces the old value (Appendix A in Algorithm 10 lines $35 - 54$).

The T^{inf} is augmented data which is iteratively processed for each case to propose adjustments in each model similar to α and κ (Figure 3.2). The temporal old

log-likelihood of the current model configuration is calculated based on the existing T^{inf} value. The new T^{inf} proposed such as, first the T^{inf} is divided into some number of groups as per the maximum T^{inf}. To propose a new T^{inf}, the range of infection times is divided into groups based on the maximum T^{inf} value, with each group covering a specific time interval (e.g. group 1: T^{inf} between 1 to 7, group 2: 8 to 15). Each group is assigned a step size $(1, 2, 3, \dots)$, determining how much the infection time can change within that group. For each case, the model identifies the group corresponding to the current T^{inf} and generates a new proposed T^{inf} by randomly increasing or decreasing the current value by the step size of the assigned group. This approach ensures controlled variability in infection times, which is useful for adjusting the likelihood in line with the temporal structure of the data. Then the temporal new log-likelihood for this new configuration is calculated. The acceptance probability for the proposed change is determined by the exponential difference between the new and old log-likelihoods. This ratio reflects how much the new proposal improves the fit compared to the old configuration. The proposed change is accepted or rejected with the accepted probability (Appendix A in Algorithm 10 lines $1 - 14$).

The values of λ_r and λ_w are updated for each model in a Bayesian inference model. A new value for the non-infectious contact rate λ_r and λ_w are proposed using a random normal distribution. The contact likelihood for the current and proposed values of λ_r and λ_w are calculated as shown in Figure 3.2. The new value of λ_r and λ_w is accepted with acceptance probability based on the difference between the contact log-likelihood of the old and new values. If the new proposal improves the fit compared to the old configuration. The proposed change is accepted with the acceptance probability. The acceptance probability for the proposed change is determined by the exponential difference between the new and old log-likelihoods. This ratio reflects how much the new proposal improves the fit compared to the old configuration. The proposed change is accepted or rejected with the accepted probability (Appendix A in Algorithm 10 lines $55 - 80$).

In the inference process, the parameter π i.e. reporting rate is updated as shown in Figure 3.2. For each iteration, a step size is sampled from $\mathcal{N}(0, \pi)$ (a normal distribution with a mean of zero and a variance of π), and a new proposed value for π is calculated by adding this step to the current value. The old log-likelihood of the reporting process is computed for both the current π and proposed new π values, and a beta prior is applied to each. The posterior log-likelihood for the old and new π is then compared, and the new value is accepted with an acceptance probability. The acceptance probability of the proposed change is determined by the exponential difference between the new and old posterior log-likelihood. If the new value improves the posterior likelihood sufficiently, it replaces the current value of π. This process iteratively refines π to better fit the model (Appendix A in Algorithm 10 lines $81 - 94$).

3.2.5 Accuracy of Transmission Tree

The ground truth validates the accuracy of an inferred transmission tree. For this, edges ('from and to') are calculated for actual simulated data (ground truth)

and the inferred data using the support. Support indicates how confident the model is in an inferred edge. The ground truth indicates the actual connections between nodes, while the inferred data represents predicted connections. The accuracy assessment involves several steps. First, the edges present in both the ground truth and the inferred data are identified, checking each pair in the inferred data against those in the ground truth. True positives (TP) are counted as the number of edges that appear in both ground truth and inferred edges, representing correctly identified connections. False positives (FP) are edges present in the inferred edges but absent in the ground truth, indicating incorrectly inferred connections. False negatives (FN) are edges present in the ground truth but not in the inferred data, showing missed connections. The overall accuracy can be calculated using the formula:

$$Accuracy = \frac{TP}{TP + FP + FN} \tag{3.10}$$

This provides a ratio of correctly inferred edges to the total edges that should have been identified, which offers a measure of how well the inferred transmission tree reflects the true connections within the disease transmission model.

4

Results

In this section, I present the results of a reconstructed transmission tree based on applying Bayesian inference to different data sources and using sampling probabilities. First, I describe the results from the disease transmission model, followed by the reconstruction of the transmission tree using different data sources and various sampling probabilities.

4.1 Disease Transmission Model

The generated data contained multiple data sources within the hospital setting, including patient admission and discharge records, diagnostic tests, and patient contact logs. Additionally, genetic sequencing of isolated pathogens provided insights into the genetic relationships between cases, aiding in the inference of transmission events. Some patients, identified as index cases, played a primary role in initiating transmission chains. These index cases contributed to further spread within the ward by infecting additional patients. Admission and discharge dates were used in tracking patient movement and overlap periods within hospital units. Patients admitted during an infectious period had a higher likelihood of becoming infected.

The example of the transmission tree (Figure 4.1) shows the transmission pathways among patients and rooms in the facility. Nodes represent individual patients or rooms, and edges indicate the transmission links based on recorded P-P and P-R interactions. Key findings from the tree include:

1. A clear pathway of spread originating from an index patient showing initiation of the outbreak.
2. A clustering of cases around rooms in the hospital, indicates the infected room infects patients staying in that room until that patient is discharged or tested positive and isolated.
3. Pink nodes which correspond to reported cases, show that only a fraction of the outbreak is detected.

Figure 4.2 shows the tree of only sampled cases from the infected patients tree. If there is no direct link between the two cases then the node is linked to the last sampled ancestor. Figure 4.3 presents the cumulative number of cases as the outbreak progresses, based on the time of infection. The results demonstrate a steady increase in cumulative cases as the outbreak goes further, reflecting the progression of the outbreak.

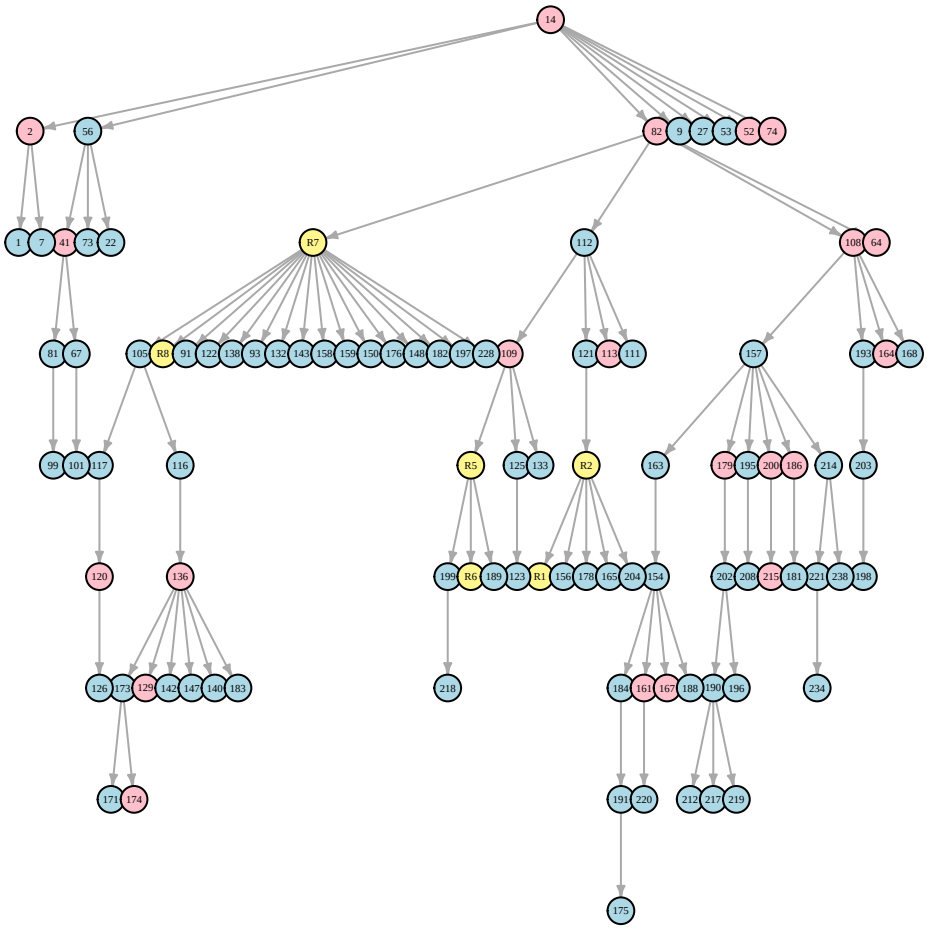


Figure 4.1: The original transmission tree of all infected patients and rooms. The pink nodes represent sampled cases, while the yellow nodes show contaminated rooms. The direction of the arrows indicates 'who infected whom'.

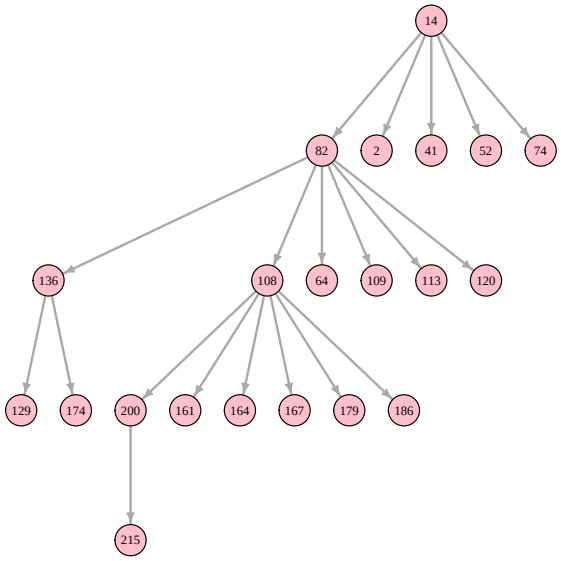


Figure 4.2: The transmission tree of sampled cases (pink nodes from Figure 4.1). Arrows shows how transmission has transferred from one case to next case

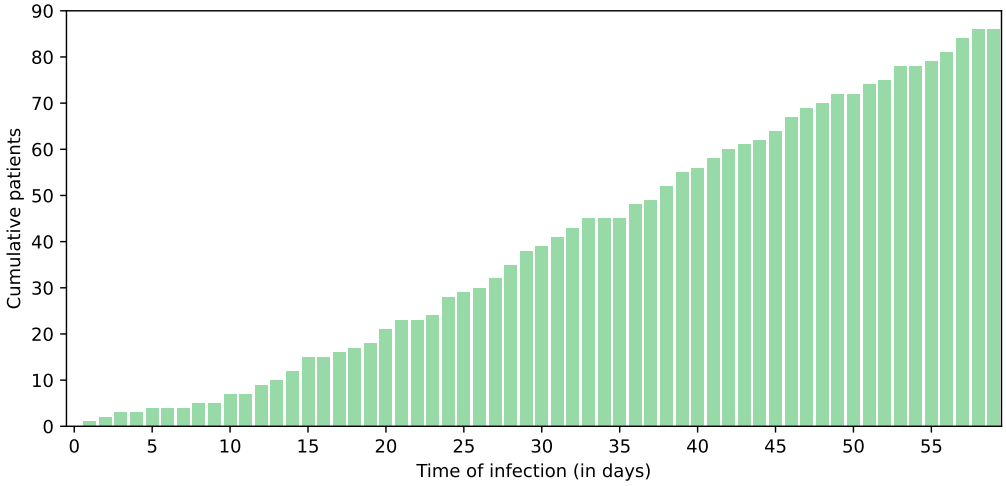
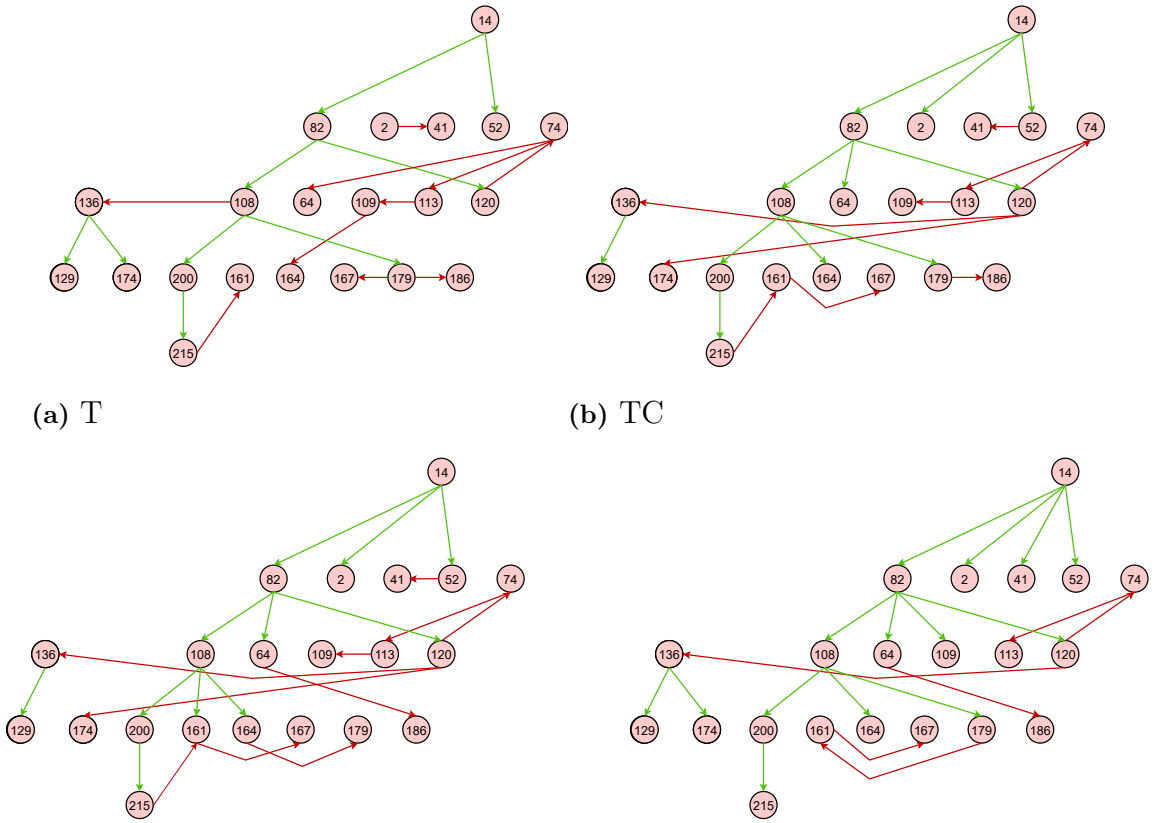


Figure 4.3: Cumulative number of cases over time of infection of a simulated data.

4.2 Reconstruction of Transmission Tree using Multiple Data Sources

The transmission trees were reconstructed from a set of sampled cases, as shown in Figure 4.2, based on multiple types of data: temporal, contact, reporting, and genetic. Four versions of the transmission tree were constructed to assess how incorporating additional data sources impacts the inferred transmission ancestry, each integrating an increasing number of data sources. These versions (Figure 4.4) show how each data source adds unique and complementary information, improving the accuracy of the inferred transmission trees. The first transmission tree (Figure 4.4a) was constructed using only temporal data, including admission, discharge, and test dates for each patient. By incorporating contact data with existing temporal data, the transmission tree in Figure 4.4b shows slightly more defined connections than the first tree. This is evident due to interactions between patients providing evidence for potential transmission events, especially among patients who overlapped in shared rooms or interacted within the same time frame within the ward. In the third reconstructed transmission tree, an additional parameter is used which is the probability of unobserved intermediate cases (π). This helps to add intermediate cases that are infected but not sampled during the outbreak (Figure 4.4c). Finally, the integration of genetic data provided the most accurate representation of the transmission tree (Figure 4.4d). Genetic sequencing of pathogens isolated from each case allowed for confirming genetic relatedness between cases, highlighting direct ancestries with higher certainty. These inferred trees T, TC, TCR, and TCRG can be compared with the tree as shown in Figure 4.2

This stepwise construction of transmission trees shows the importance of combining multiple data sources to achieve a reliable inference of transmission pathways in a hospital setting. The results emphasize that while temporal and contact data provide a foundational understanding, incorporating genetic data is essential for accurate ancestry determination. The accuracy of transmission trees reconstructed using different data can be seen in Figure 4.5. Ten experiments were carried out for each model (T, TC, TCR, TCRG) with sampling probability 0.1. The figure on the right side shows the accuracy distribution across these models, whereas the figure on the left shows how the mean of accuracy changes as new data is incorporated. The outliers in the boxplot are marked with diamonds with values greater than $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$, where $Q1$ and $Q3$ are lower and upper quartiles, and IQR is inter-quartile range. The accuracy with TC increased slightly compared to T, but accuracy is stayed stable with TCR. The TCRG model achieves the highest accuracy with the lowest standard deviation, indicating both high reliability and stability in transmission tree inference, which incorporates all data sources.



(a) T **(b)** TC

(c) TCR **(d)** TCRG

Figure 4.4: Four transmission trees reconstructed, for the tree Figure 4.2, using different data sources. Green edges represent the correctly assigned ancestor for a case while red edges represent the incorrectly assigned ancestor. Arrows show the direction of transmission from one case to another. (a) Transmission tree reconstructed with temporal data (T) (b) Transmission tree reconstructed with temporal and contact data (TC) (c) Transmission tree reconstructed with temporal, contact, and reporting data (TCR) (d) Transmission tree reconstructed with temporal, contact, reporting, and genetic data (TCRG)

4.3 Reconstruction of Transmission Tree using various Sampling Probabilities

The sampling probabilities ranging from 0.05 to 0.2 were tested to assess their impact on the reconstruction of the transmission tree. Increasing sampling probabilities means increasing the number of cases sampled. Ten experiments were conducted for each probability to observe how changes in sampling probability influenced the accuracy and structure of the transmission tree reconstruction. These experiments conducted for model TCRG, these are different experiments from the previous experiments which shown in Figure 4.5. The right side of Figure 4.6 shows the proportion of cases sampled at each sampling probability. As the sampling probability increases, a larger proportion of patients become recorded cases. This increase in

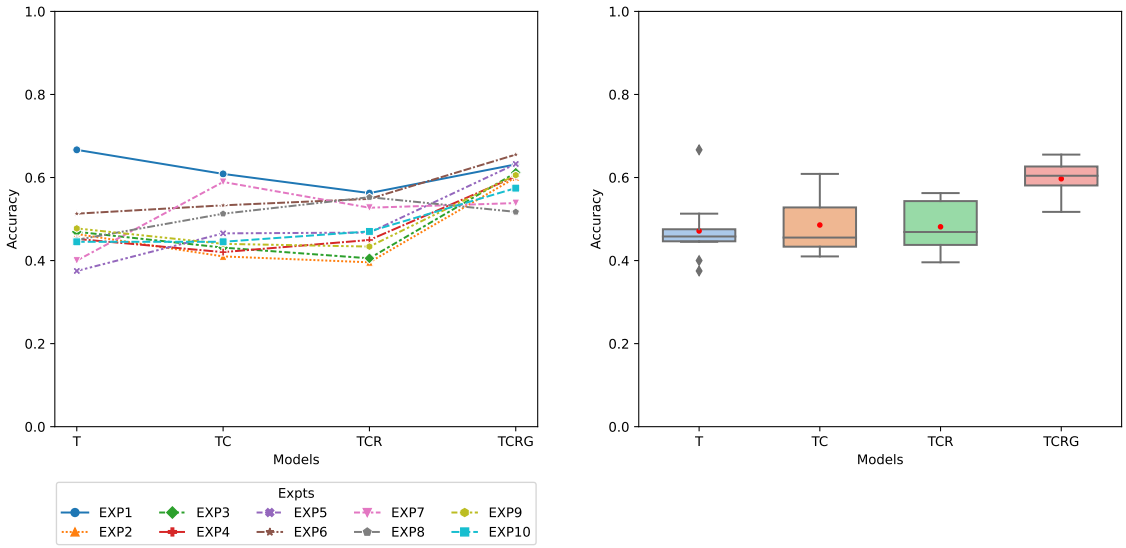


Figure 4.5: Left figure shows lineplot of accuracy over various data sources. The different lines correspond to ten different experiments. Ten experiments for each model conducted with sampling probability 0.1. Right figure shows a boxplot of increased accuracy by incremental new data sources. The red dots indicate mean accuracy for ten experiments.

sampled patients led to a reduction in intermediate cases, resulting in fewer indirect connections between cases. Therefore, indirect links between two cases were reduced, which influences intermediate cases. This affects the accuracy of reconstructing the transmission tree, as shown on the right side of the figure. At a sampling probability of 0.05, accuracy is lower, with more variability, while as the sampling probability increases, accuracy improves, and variability decreases. The red dot in each box shows the mean accuracy for each sampling probability. The mean accuracy is the highest at 0.2 sampling probability compared to the other with lesser variation. The outliers in the boxplot are marked with diamonds with values greater than $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$, where $Q1$ and $Q3$ are lower and upper quartiles, and IQR is inter-quartile range.

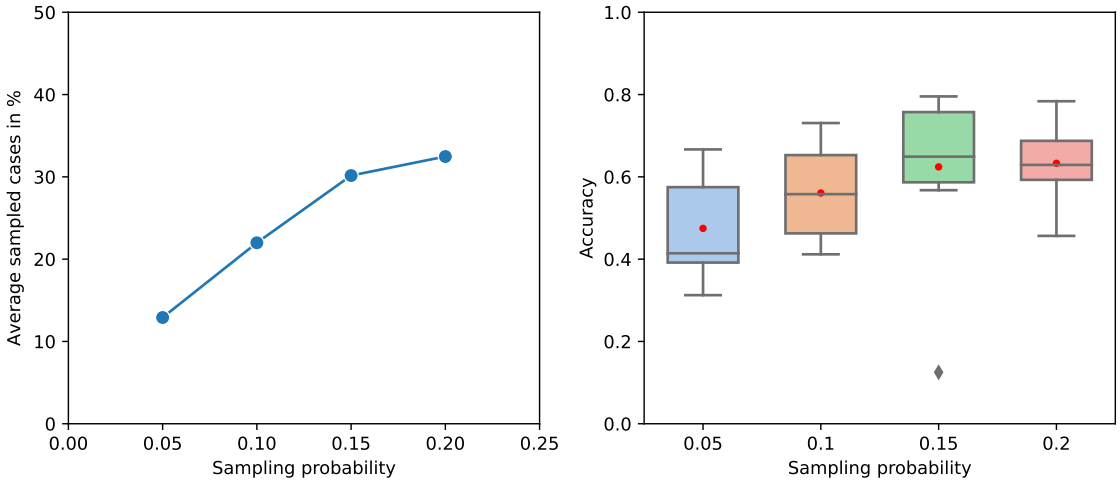


Figure 4.6: Left figure shows the average proportion of cases sampled for each sampling probability and right figure shows a boxplot of accuracy over different data sources. Ten experiments are conducted for each probability with TCRG model. (red dots indicate mean accuracy for each probability).

5

Discussion

This chapter provides an analysis of the materials, methods, and results utilized in this research, alongside a comparative evaluation of the *outbreaker2* framework and other research that has extended its application. Additionally, this chapter considers the broader implications of these findings for future research. The flexible methodology demonstrated here provides opportunities for further refinement, such as incorporating real-time data updates or adapting the model to multi-ward or inter-facility outbreak scenarios. Since hospital infections continue to change, extending this work could help develop stronger, more proactive strategies for controlling infections in a range of healthcare settings.

This work extends the methodology developed by Campbell *et al.* (*Outbreaker2*)[9] but focuses specifically on hospital settings. A distinction lies in the approach to assigning initial ancestries. While Campbell *et al.* used only genetic data for this purpose, our approach initializes ancestries bases on their admission and discharge date to check if the patients stay in hospital at the same time. We chose to add genetic data later in the process because getting genetic sequences can be difficult and time-consuming in a lab. To initialize the transmission trees referred to as T, TC, and TCR we used information about patients who were in the hospital at the same time, which helped us identify possible links between them. Additionally, our method for initializing the time of infection, is different from what Campbell *et al.* used. This flexibility in our approach is important for accurately reflecting how infections spread in healthcare environments, where many factors come into play.

Lindsey *et al.* [9] also extended *Outbreaker2* by modeling ward-based genetic transmission in a hospital outbreak. While Lindsey *et al.* leveraged genetic data and ward occupancy to model transmission events within wards, our inclusion of contact data offers an additional dimension for reconstructing within-ward transmission pathways. This approach potentially improves inference by identifying more precise infection links in a structured setting, where interactions between patients are integral to understanding spread dynamics. Furthermore, getting genetic sequences can be difficult and time-consuming so incorporating other easily available data like contact, temporal can be useful to to learn transmission dynamics of disease spread.

5.1 Transmission Model

The transmission model in this study simplifies how infections spread in hospitals by focusing on four main types of spread (patient-to-patient, patient-to-room, room-to-patient, and room-to-room) based on where patients are located. To make the

model more realistic, future research could include the role of healthcare workers, who often carry infections between patients. It would also help to adjust infection chances based on how infectious a patient or room is, which depends on factors like how long a patient stays or how often they interact with others. Adding room cleaning schedules to the model would show how cleaning can reduce infection spread over time, while considering how long a pathogen can survive on surfaces would make the model even more accurate. Lastly, separating different indirect transmission methods, like surfaces, or equipment, could make the model a stronger tool for managing real hospital outbreaks.

5.1.1 Reconstruction of Transmission Tree Using various Data Sources

The transmission tree was reconstructed using a specific sequence of data T, TC, TCR, and TCRG to understand how each data type affects transmission accuracy. However, other combinations, such as TRG or TCG, were not explored. Testing these alternatives could provide valuable insights into which data combinations most effectively enhance transmission accuracy and reduce uncertainty. For example, pairing genetic data with only temporal or reporting data might yield comparable accuracy, or it may be possible to achieve effective transmission mapping with contact and genetic data.

The reported contact assumes all contacts are reported, but in real case, there could be some missing contacts that should also be considered while reconstructing the model. The inference algorithm used to reconstruct trees can be improved by priors on parameters such as non-infectious contact probability, and intermediate cases reporting probability.

5.1.2 Reconstruction of Transmission Tree Using various Sampling probabilities

Sampling was kept simple for model simplicity, with patients sampled randomly each day based on a fixed probability. Prioritizing the sampling of patients who have had contact with positive cases can make it a more realistic model. For instance, if a patient tests positive, other patients sharing the same room would have a higher likelihood of being infected and should be prioritized for sampling. If patients have wounds or cuts, using a ventilator, or people with weakened immune systems, have more probability of getting infected easily [3] [4]. These patients should be prioritized for sampling if they are within ward at the infectious period. Additionally, as the environment plays a role in *Klebsiella oxytoca* transmission, such as through sink-to-sink spread via sewage pathways [30], [2], [1] future models could incorporate environmental sampling. This approach of sampling would add realism to the model.

The model's accuracy can be further increased by adding more data on room sampling (sewage pipe attached to room), as once room gets contaminated it infects patients staying in it. This could increase model accuracy as it decreases intermediate cases between two sampled cases, as seen in Figure 4.1.

6
Conclusions

This thesis aimed to develop a model for inferring transmission chains of infectious diseases, specifically *Klebsiella oxytoca*, using Bayesian inference. This enhances outbreak management strategies by integrating data from multiple sources including patient interactions and genomics. This study has provided insights into the reconstruction of disease transmission tree in complex healthcare environments.

Three research questions guided this work. First, the study explored data generation within hospital settings, identifying how patient interactions and environmental factors contribute to data complexity and accuracy in transmission inference. Second, the findings demonstrated that adding new data sources (e.g., contact, reporting, and genetic data) incrementally improved the accuracy of reconstructed transmission trees, highlighting the value of integrating diverse data sources. Finally, the analysis showed that sampling probability significantly impacts the accuracy of inferred transmission trees, with higher sampling probabilities leading to greater accuracy and reduced variability.

This thesis highlights the importance of combining different data sources to create clear insights into how infections spread in healthcare facilities. By tackling issues related to complex data and sampling probabilities, this research aims to help manage outbreaks better and improve how we model diseases in hospitals. Future studies could enhance these models by incorporating other data sources or different sampling methods. It will lead to more effective responses to outbreaks that may occur in healthcare settings.

Bibliography

1. Bagley, S. T. Habitat Association of Klebsiella Species. *Infection Control* **6,** 52–58 (1985).
2. Podschun, R. *et al.* "Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors" (1998).
3. Peleg, A. Y. & Hooper, D. C. Hospital-acquired infections due to gram-negative bacteria. *New England Journal of Medicine* **362,** 1804–1813 (2010).
4. Podschun, R. & Ullmann, U. Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clinical microbiology reviews* **11,** 589–603 (1998).
5. Keeling, M. J. & Rohani, P. *Modeling infectious diseases in humans and animals* (Princeton university press, 2008).
6. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology* **6,** 477–487 (2008).
7. Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-a., Comte, B. & Voirin, N. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one* **8,** e73970 (2013).
8. Cassidy, R., Kypraios, T. & O'Neill, P. D. Modelling, Bayesian inference, and model assessment for nosocomial pathogens using whole-genome-sequence data. *Statistics in Medicine* **39,** 1746–1765 (2020).
9. Campbell, F., Cori, A., Ferguson, N. & Jombart, T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology* **15,** e1006930 (2019).
10. Lindsey, B. B., Villabona-Arenas, C. J., Campbell, F., Keeley, A. J., Parker, M. D., Shah, D. R., Parsons, H., Zhang, P., Kakkar, N., Gallis, M., *et al.* Characterising within-hospital SARS-CoV-2 transmission events using epidemiological and viral genomic data across two pandemic waves. *Nature communications* **13,** 671 (2022).
11. Hilty, M., Betsch, B. Y., Bögli-Stuber, K., Heiniger, N., Stadler, M., Küffer, M., Kronenberg, A., Rohrer, C., Aebi, S., Endimiani, A., *et al.* Transmission dynamics of extended-spectrum β-lactamase–producing Enterobacteriaceae in the tertiary care hospital and the household setting. *Clinical infectious diseases* **55,** 967–975 (2012).

12. Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T. & Soubeyrand, S. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS computational biology* **8,** e1002768 (2012).

13. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution* **34,** 997–1007 (2017).

14. Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., De Angelis, D., Cartwright, E. J., Peacock, S. J. & Cooper, B. S. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics* **10,** 395 (2016).

15. Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. & Ferguson, N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology* **10,** e1003457 (2014).

16. Abbey, H. An examination of the Reed-Frost theory of epidemics. *Human biology* **24,** 201 (1952).

17. Fine, P. E. A commentary on the mechanical analogue to the Reed-Frost epidemic model. *American journal of epidemiology* **106,** 87–100 (1977).

18. Wolkewitz, M., Dettenkofer, M., Bertz, H., Schumacher, M. & Huebner, J. Statistical epidemic modeling with hospital outbreak data. *Statistics in medicine* **27,** 6522–6531 (2008).

19. Forrester, M., Pettitt, A. & Gibson, G. Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. *Biostatistics* **8,** 383–401 (2007).

20. Berry, D. A. *Statistics: A Bayesian Perspective* (Duxbury Press, 1995).

21. Jones, B. G. *Bayesian Methods for the Design and Analysis of Cluster Randomised Controlled Trials* PhD thesis (University of Plymouth, 2022).

22. Kruschke, J. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Academic Press, 2014).

23. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian data analysis* (Chapman and Hall/CRC, 1995).

24. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21,** 1087–1092 (1953).

25. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications (1970).

26. Gelfand, A. E. & Smith, A. F. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* **85,** 398–409 (1990).

27. Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. *Handbook of markov chain monte carlo* (CRC press, 2011).

28. Robert, C. *Monte Carlo Statistical Methods* 1999.

29. Dehouche, N., Viravan, S., Santawat, U., Torsuwan, N., Intharakosum, A. & Sirivatanauksorn, Y. Hospital length of stay: A cross-specialty analysis and Beta-geometric model. *Plos one* **18,** e0288239 (2023).

30. Lowe, C., Willey, B., O'Shaughnessy, A., Lee, W., Lum, M., Pike, K., Larocque, C., Dedier, H., Dales, L., Moore, C., *et al.* Outbreak of extended-spectrum β-lactamase–producing Klebsiella oxytoca infections associated with contaminated handwashing sinks. *Emerging infectious diseases* **18,** 1242 (2012).

A

Appendix

A.1 Pseudocode for Data Generation

Algorithm 6 An algorithm for generating synthetic patient data

Require: N_{rooms}: No. of rooms, P_{room} : Patients per room, LOS : Length of stay

$T_p \leftarrow N_{room} \times P_{room}$, $R \leftarrow [R_1, R_2, \ldots, R_n]$ ▷ T_p : Total patients

$Pt_{id} \leftarrow [Pt_1, Pt_2, \ldots]$, $R_{no} \leftarrow \{Pt_i \rightarrow R_j \mid Pt_i \in R_j\}$

$T^{adm} \leftarrow [1, 1, 1, \ldots]$ ▷ T^{adm}: Time of admission

$\text{LOS} \leftarrow \sum_{k=\min(\text{LOS})}^{\max(\text{LOS})} \frac{\beta}{\alpha+\beta} \prod_{i=1}^{k-1} \frac{\alpha}{\alpha+\beta+i}$

$T^{dis} \leftarrow [\,]$ ▷ List to store discharge date

$InfF \leftarrow [FALSE, FALSE, FALSE, \ldots]$

$Inf \leftarrow$ [[0,0],[0,0],....] ▷ List to store infected cases and infection date

roomP $\leftarrow \begin{bmatrix} (R_1, R_2) \\ \vdots \\ (R_{N_{room-1}}, R_{N_{room}}) \end{bmatrix}$

/* —— **Assign connected rooms** ——*/

for $i = 1, \ldots, T_p$ **do**

  $R_i^{no} \leftarrow \text{Extract}(R, \text{Pt}_{id})$

  $R_i^{con} \leftarrow \{roomP[i, 2] \mid roomP[i, 1] = R_{no}\}$

  $\text{Pt}_{id}.\text{R}_{con} \leftarrow R_i^{con}$

end for

for $i = 1, \ldots, T_p$ **do**

  $T_i^{dis} \leftarrow T_i^{adm} + LOS_i$ ▷ Simulate patients discharge date

end for

/* —— **Randomly generate index case patient** ——*/

$Inf_{pt} \leftarrow$ `sample(1:`T_p`, 1)` ▷ Index cases

$Max_{infDay} \leftarrow 5$

$Inf_{day} \leftarrow$ `sample(1:`Max_{infDay}`, 1)`

`InfF` $(Inf_{pt}) \leftarrow$ `TRUE`

$Inf \leftarrow$ `list(c(`$Inf_{\texttt{pt}}, Inf_{\texttt{day}}$`))`

Algorithm 7 Transmission of pathogen and sampling

Description: This algorithm transmits pathogen from P-P, P-R, R-P, R-R and also sampling of patients goes on every day. Transmission and sampling occurs from day 1 to total number of days.

Initialize:
$Trans_{\texttt{pairs}} \leftarrow [[0,0],[0,0],...]$, $\texttt{Days} \leftarrow 60$
for day = 1 to Days **do**
  for each id in Inf **do**
    $id \leftarrow id[1]$, $T^{inf} \leftarrow id[2]$, $Inf_{id} \leftarrow Inf[1]$
    if id is not numeric **then** ▷ If id is character, then it's room id (ex. R1)
/* —— **Room-to-Patient Transmission**—*/
      for each p in $Trans_{pairs}$ **do**
        if p[2] == id **then**
          $Pt_{ContR} \leftarrow$ Find Patients in contaminated room
          for each pt in Pt_{ContR} **do**
            if $\texttt{pt} \notin Inf_{id}$ **then** ▷ Patient not already infected
              Generate $x_{RP} \sim \mathcal{N}(0,1)$
              if $x_{RP} < \gamma_{rp}$ **then**
                $Trans_{pairs} \leftarrow Trans_{pairs} \cup (id, pt)$
                $Inf \leftarrow Inf \cup (pt, day)$
                $InfF \leftarrow TRUE$ ▷ Infected patient flag
              end if
            end if
          end for
        end if
      end for
    else ▷ Id is numerical then it is patient id
      if day $\geq T^{inf}$ **and** day $\leq T^{dis}$ **then** ▷ If day is today
        $P_{ward} \leftarrow P \in Pt \mid (T_{ad}(p) \leq$ day and $T_{dis}(p) \geq$ day)
/* ——**Patient-to-Room Transmission**—*/
        Find $R_{inf_{pt}}$ Room infected patient staying
        if $R_{inf_{pt}}$ is not in Inf_{id} **then**
          $x_{PR} \leftarrow\sim \mathcal{N}(0,1)$
          if $x_{PR} < \gamma_{pr}$ **then**
            $Trans_{pairs} \leftarrow Trans_{pairs} \cup (\texttt{id},\ R_{inf_{pt}})$
            $Inf \leftarrow Inf \cup (R_{inf_{pt}},\ \texttt{day})$
            $InfF[R_{inf_{pt}}] \leftarrow$ TRUE
            Find room connected to contaminated room R_{Cont_R} with the help of $roomP$
/* ——**Room-to-Room Transmission**—*/
            if R_{Cont_R} is not in Inf_{id} **then**
              $Trans_{pairs} \leftarrow Trans_{pairs} \cup (R_{inf_{pt}},\ R_{Cont_R})$
              $Inf \leftarrow Inf \cup (R_{Cont_R},\ \texttt{day})$
            end if
          end if
        end if

Algorithm 8 Transmission of pathogen and sampling (Continued)

/* ——**Patient-to-Patient Transmission**—*/
for each O_{id} in P_{ward} **do**
  $x_{PP} \leftarrow\sim \mathcal{N}(0,1)$
  if O_{id} is not in Inf_{id} **then**
    if id and O_{id} common room **then** ▷ Within room transmission
      if $x_{PP} < \beta_r$ **then**
        $Trans_{pairs} \leftarrow Trans_{pairs} \cup$ (id, O_{id})
        $Inf \leftarrow Inf \cup$ (O_{id}, day) , $InfF[O_{id}] \leftarrow$ TRUE
      end if
    else ▷ Within ward transmission
      if $x_{PP} < \beta_w$ **then**
        $Trans_{pairs} \leftarrow Trans_{pairs} \cup$ (id, O_{id})
        $Inf \leftarrow Inf \cup$ (O_{id}, day) , $InfF[O_{id}] \leftarrow$ TRUE
      end if
    end if
  end if
end for
end if
end if
end for
/* ——**Testing and isolate**—*/
$Pt_{today} \leftarrow$ Filter patients where $T^{adm} \leq$ day **and** $T^{dis} \geq$ day
$Test_{Pt} \leftarrow$ sample Pt_{today} with probability P_s
$Test_R = [0, 0, 0, ...]$
for i = 1 to nrow($Test_{Pt}$) **do**
  if $InfF[i] ==$ TRUE **then**
    $Test_R[i] \leftarrow 1$
    T^{samp} <- day
    add new patient into room $R_{No}[i]$ on next day $day + 1$
  else
    $Test_R[i] \leftarrow 0$
  end if
end for
/* ——**Process Other Patients which are not tested**—*/
$other \leftarrow P_{t\_today} \notin$ Test_pt
for j = 1 to nrow($other$) **do**
  if $T^{dis}[j] = day$ **then**
    add new patient into room $R_{No}[j]$ on next day $day + 1$
  end if
end for
end for ▷ end For loop: for each day until last day

A.2 Pseudocode for Bayesian Inference

Algorithm 9 searchAnces

Description: This function finds possible ancestors for a given case based on admission and sampling times.

Require: T^{adm}: time of admission, T^{samp} time of sampling, T^{inf}: time of infection, idx : case id, N : Number of cases

$initAnces \leftarrow \left[\mathbf{0}\right]_{1\times|\mathcal{I}_{\mathbb{N}}|}$
$counter \leftarrow 0$
for $j = 1$ to N **do**
  if $j \neq idx$ & $T^{inf}[j] < T^{inf}[idx]$ **then**
    $adjAdm_i \leftarrow \max(T^{adm}[idx] - 5, T^{adm}[idx] - 1)$
    if $adjAdm_i < T^{samp}[j]$ **then**
      $counter \leftarrow counter + 1$
      $initAnces[counter] \leftarrow j$
    end if
  end if
end for
if $counter = 0$ **then**
end if
$recAnces \leftarrow 0$
$recAncesT \leftarrow 0$
for $k = 1$ to $counter$ **do**
  if $t_{\text{inf}}[initAnces[k]] > recAncesT$ **then**
    $recAnces \leftarrow initAnces[k]$
    $recAncesT \leftarrow t_{\text{inf}}[initAnces[k]]$
  end if
end for

Output: $recAnces$

Algorithm 10 bayesianInference

Require: N: No. of cases, T^{inf}: Time of infection, α : Indices of sampled ancestor, κ: Indices of sampled ancestor generations, μ: Mutation rate, λ_r : non-infectious contact within room, λ_w : non-infectious contact within ward, ϵ : contact reporting probability, π : reporting probability, $wDens$: generation time distribution, $fDens$: delay time distribution

/* —— **Move** T^{inf} ——*/

for $i = 1, \ldots, N$ **do**
  $llInf^{old} \leftarrow \texttt{llInf}(i, \alpha_i, \kappa_i, T^{inf}, wDens)$
  $llSamp^{old} \leftarrow \texttt{llSamp}(i, T^{inf}, T^{samp}, fDens)$
  $like^{old}_{agg} \leftarrow llInf^{old} + llSamp^{old}$
  $T^{inf}_{prop} \leftarrow T^{inf}_i + \mathcal{N}(0,1)$ ▷ Proposing new T^{inf}_i
  $\tilde{T}^{Inf} \leftarrow T^{inf},\ \tilde{T}^{Inf}[i] \leftarrow T^{inf}_{prop}$
  $llInf^{new} \leftarrow \texttt{llInf}(i, \alpha_i, \kappa_i, \tilde{T}^{inf}, wDens)$
  $llSamp^{new} \leftarrow \texttt{llSamp}(i, \tilde{T}^{inf}, T^{samp}, fDens)$
  $like^{new}_{agg} \leftarrow llInf^{new} + llSamp^{new}$
  if $\exp(like^{new}_{agg} - like^{old}_{agg}) \geq \mathcal{N}(0,1)$ **then**
    $T^{inf}[i] \leftarrow T^{inf}_{prop}$ ▷ Accepting proposed T^{inf}_{prop}
  else
  end if
end for

/* —— **Move** α ——*/

for $i = 1, \ldots, N$ **do**
  $llInf^{old} \leftarrow \texttt{llInf}(i, \alpha_i, \kappa_i, T^{inf}, wDens)$
  $llSamp^{old} \leftarrow \texttt{llSamp}(i, T^{inf}, T^{samp}, fDens)$
  $llGene^{old} \leftarrow \texttt{llGene}(i, L, D, \alpha, \kappa, \mu)$
  $llCont^{old} \leftarrow \texttt{llCont}(i, C, \alpha, \kappa, \epsilon, \lambda_r, \lambda_w)$
  $llRep^{old} \leftarrow \texttt{llRep}(i, K, \kappa, \pi)$
  $like^{old}_{agg} \leftarrow llInf^{old} + llSamp^{old} + llGene^{old} + llCont^{old} + llRep^{old}$
  $\alpha_{prop} \leftarrow \texttt{searchAnces}(T^{\text{adm}}, T^{\text{sampl}}, T^{\text{inf}}, i, n)$ ▷ proposing new α_i reference Algorithm 9
  $\tilde{\alpha} \leftarrow \alpha,\ \tilde{\alpha}[i] \leftarrow \alpha_{prop}$
  $llInf^{new} \leftarrow \texttt{llInf}(i, \alpha_{prop}, \kappa_i, T^{inf}, wDens)$
  $llSamp^{new} \leftarrow \texttt{llSamp}(i, T^{inf}, T^{samp}, fDens)$
  $llGene^{new} \leftarrow \texttt{llGene}(i, L, D, \tilde{\alpha}, \kappa, \mu)$
  $llCont^{new} \leftarrow \texttt{llCont}(i, C, \tilde{\alpha}, \kappa, \epsilon, \lambda_r, \lambda_w)$
  $llRep^{new} \leftarrow \texttt{llRep}(i, K, \kappa, \pi)$
  $like^{new}_{agg} \leftarrow llInf^{new} + llSamp^{new} + llGene^{new} + llCont^{new} + llRep^{new}$
  if $\exp(like^{new}_{agg} - like^{new}_{agg}) \geq \mathcal{N}(0,1)$ **then**
    $\alpha[i] \leftarrow \alpha_{prop}$ ▷ Accepting proposed α_{prop}
  else
  end if
end for

Algorithm 11 bayesianInference (Continued)

/* —— **Move** κ ——*/
for $i = 1, \ldots, N$ **do**
  $llInf^{old} \leftarrow \texttt{llInf}(i, \alpha_i, \kappa_i, T^{inf}, wDens)$
  $llSamp^{old} \leftarrow \texttt{llSamp}(i, T^{inf}, T^{samp}, fDens)$
  $llGene^{old} \leftarrow \texttt{llGene}(i, L, D, \alpha, \kappa, \mu$
  $llCont^{old} \leftarrow \texttt{llCont}(i, C, \alpha, \kappa, \epsilon, \lambda_r, \lambda_w)$
  $llRep^{old} \leftarrow \texttt{llRep}(i, K, \kappa, \pi)$
  $like_{agg}^{old} \leftarrow llInf^{old} + llSamp^{old} + llGene^{old} + llCont^{old} + llRep^{old}$
  $\kappa_{prop} \leftarrow \kappa_i + \mathcal{N}(0, 1)$ ▷ Proposing new κ_i
  $\tilde{\kappa} \leftarrow \kappa,\ \tilde{\kappa}[i] \leftarrow \kappa_{prop}$
  $llInf^{new} \leftarrow \texttt{llInf}(i, \alpha_i, \kappa_{prop}, T^{inf}, wDens)$
  $llSamp^{new} \leftarrow \texttt{llSamp}(i, T^{inf}, T^{samp}, fDens)$
  $llGene^{new} \leftarrow \texttt{llGene}(i, L, D, \alpha, \tilde{\kappa}, \mu$
  $llCont^{new} \leftarrow \texttt{llCont}(i, C, \alpha, \tilde{\kappa}, \epsilon, \lambda_r, \lambda_w)$
  $llRep^{new} \leftarrow \texttt{llRep}(i, K, \tilde{\kappa}, \pi)$
  $like_{agg}^{new} \leftarrow llInf^{new} + llSamp^{new} + llGene^{new} + llCont^{new} + llRep^{new}$
  if $\exp(like_{agg}^{new} - like_{agg}^{new}) \geq \mathcal{N}(0, 1)$ **then**
    $\kappa[i] \leftarrow \kappa_{prop}$ ▷ Accepting proposed κ_{prop}
  else
  end if
end for
/* —— **Move** λ_r ——*/
for $i = 1, \ldots, N$ **do**
  $llCont^{old} \leftarrow \texttt{llCont}(i, C, \alpha, \kappa, \epsilon, \lambda_r, \lambda_w)$
  $priorContact^{old} \leftarrow \beta(\lambda_r, 1, 1)$
  $like_{agg}^{old} \leftarrow llCont^{old} + priorContact^{old}$
  $\lambda_r^{prop} \leftarrow \lambda_r + \mathcal{N}(0, 1)$ ▷ Proposing new λ_r
  $priorContact^{new} \leftarrow \beta(\lambda_r^{prop}, 1, 1)$
  $llCont^{new} \leftarrow \texttt{llCont}(i, C, \alpha, \kappa, \epsilon, \lambda_r^{prop}, \lambda_w)$
  $like_{agg}^{new} \leftarrow llCont^{new} + priorContact^{new}$
  if $\exp(like_{agg}^{new} - like_{agg}^{new}) \geq \mathcal{N}(0, 1)$ **then**
    $\lambda_r \leftarrow \lambda_r^{prop}$ ▷ Accepting proposed λ_r^{prop}
  else
  end if
end for

Algorithm 12 bayesianInference (Continued)

/* —— Move λ_w ——*/
for $i = 1, \ldots, N$ **do**
  $llCont^{old} \leftarrow \texttt{llCont}(i, C, \alpha, \kappa, \epsilon, \lambda_r, \lambda_w)$
  $priorContact^{old} \leftarrow \beta(\lambda_w, 1, 1)$
  $like_{agg}^{old} \leftarrow llCont^{old} + priorContact^{old}$
  $\lambda_w^{prop} \leftarrow \lambda_w + \mathcal{N}(0, 1)$ ▷ Proposing new λ_w
  $priorContact^{new} \leftarrow \beta(\lambda_w^{prop}, 1, 1)$
  $llCont^{new} \leftarrow \texttt{llCont}(i, C, \alpha, \kappa, \epsilon, \lambda_r, \lambda_w^{prop})$
  $like_{agg}^{new} \leftarrow llCont^{new} + priorContact^{new}$
  if $\exp(like_{agg}^{new} - like_{agg}^{new}) \geq \mathcal{N}(0, 1)$ **then**
    $\lambda_w \leftarrow \lambda_w^{prop}$ ▷ Accepting proposed λ_w^{prop}
  else
  end if
end for
/* —— Move π ——*/
for $i = 1, \ldots, N$ **do**
  $likeReport^{old} \leftarrow \texttt{llRep}(i, K, \kappa, \pi)$
  $likeReport^{old} \leftarrow \beta(\pi, 10, 1)$
  $like_{agg}^{old} \leftarrow likeReport^{old} + likeReport^{old}$
  $\pi^{prop} \leftarrow \pi + \mathcal{N}(0, 1)$ ▷ Proposing new π
  $likeReport^{new} \leftarrow \beta(\pi^{prop}, 10, 1)$
  $likeReport^{new} \leftarrow \texttt{llRep}(i, K, \kappa, \pi^{prop})$
  $likeReport^{new} \leftarrow \beta(\pi, 10, 1)$
  $like_{agg}^{new} \leftarrow likeReport^{new} + likeReport^{new}$
  if $\exp(like_{agg}^{new} - like_{agg}^{new}) \geq \mathcal{N}(0, 1)$ **then**
    $\pi \leftarrow \pi^{prop}$ ▷ Accepting proposed π^{prop}
  else
  end if
end for

DEPARTMENT OF PHYSICS
University of Gothenburg
Gothenburg, Sweden
www.gu.se