Unravelling transmission dynamics of infectious diseases with Bayesian and Network analysis

by

Jianing Xu

(Under the Direction of Liang Liu and Pengsheng Ji)

Abstract

Reconstructing transmission networks is critical for identifying epidemiological factors such as superspreaders and high-risk locations, informing targeted strategies for pandemic prevention and control. This dissertation introduces two Bayesian frameworks designed to reconstruct infectious disease transmission networks by integrating genomic, temporal, and social network data. The proposed models accommodate within-host genetic diversity, unobserved infection times, incomplete sampling, latent periods, and symptom onset, significantly enhancing the precision of inferred transmission dynamics. Simulation studies demonstrate the robustness of the Bayesian transmission model without network data, achieving high accuracy in identifying direct transmission pairs—93% at a genome length of 1×10^6 and 100% at 4.4×10^6. Hypothesis testing reliably identifies direct transmission events, maintaining an average false positive proportion of approximately 1%. Meanwhile, sensitivity declines with decreasing sample sizes due to increased misclassification of indirect transmissions. Implementing Nelder–Mead optimization improved sensitivity by approximately 30%, although it concurrently raised false positives by around 10%, highlighting an inherent trade-off. Furthermore, an Exponential Random Graph Model (ERGM) fitted to the inferred transmission tree demonstrated the robust effect of social distance on transmission dynamics, revealing that each unit increase in social distance decreased transmission likelihood. Perturbation analyses with 5%, 10%, and 20% noise confirmed that ERGM reliably captured the social-distance effect and remains robust to network uncertainty. Real-world analysis using a Bayesian model on genomic and temporal data from 93 tuberculosis cases identified 28 direct transmission pairs, highlighting limited within-neighborhood transmission. ERGM analysis further suggested a trend

toward increased transmission likelihood with greater social distance, implying that contacts outside immediate neighborhoods potentially drive transmission, though this association was statistically insignificant and weakened with increasing network uncertainty. Notably, this study represents the first network investigation of tuberculosis transmission in an endemic region. In future work, we will combine GPS and cell-phone trajectory data with traditional social network data using machine learning to derive personal network information, thereby refining contact probability estimation. Additionally, adopting advanced substitution models and relaxing assumptions about uniform effective population size may further enhance model accuracy. Leveraging parallel computing will improve computational efficiency, increasing the practicality and scalability of Bayesian methods in epidemiological research.

INDEX WORDS: [Bayesian Inference, Network Analysis, Phylogeny, Transmission Network, Infectious Disease]

Unravelling transmission dynamics of infectious diseases with Bayesian and Network analysis

by

Jianing Xu

B.S., Beloit College, 2014
M.S., University of Georgia, 2019

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Doctor of Philosophy

Athens, Georgia

2025

©2025
Jianing Xu
All Rights Reserved

Unravelling transmission dynamics of infectious diseases with Bayesian and Network analysis

by

Jianing Xu

Major Professor:	Liang Liu, Pengsheng Ji
Committee:	Christopher C. Whalen Paul Schliekelman

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Acknowledgments

I wish to express my sincere gratitude to my advisors, Dr. Liang Liu and Dr. Pengsheng Ji, for their steadfast guidance, insightful feedback, and unwavering support throughout this dissertation. Their mentorship has been instrumental in shaping both my research approach and analytical rigor.

I am also deeply thankful to Dr. Christopher C. Whalen, whose extensive expertise in epidemiology and network applications has broadened my understanding of the field and inspired me to explore new research directions. His valuable insights have enriched the theoretical framework of this study.

Additionally, I appreciate the contributions of Dr. Paul Schliekelman, whose stimulating discussions and critical feedback challenged me to address the limitations of my work and deepen my comprehension of complex research issues. His innovative perspectives have significantly enhanced the overall quality of this study.

Collectively, the guidance and support of these distinguished scholars have been invaluable to my academic development and the successful completion of this dissertation.

Finally, I would like to express my heartfelt gratitude to my family, especially my parents—whose unwavering love, sacrifices, and encouragement have formed the foundation of my academic journey by enabling me to overcome challenges and pursue excellence—and to my friends, whose camaraderie, shared laughter, and steadfast support have enriched my life and sustained me through both triumphs and setbacks, rendering this journey not only more manageable but also profoundly rewarding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Introduction

Infectious diseases continue to be a major global health concern, significantly contributing to morbidity and mortality worldwide, particularly in resource-limited regions (Fonkwo, 2008). A key strategy for controlling and preventing the outbreak and spread of infectious diseases is the construction and analysis of transmission networks, which trace the movement of pathogens through populations and reveal transmission networks between individuals (Luke and Harris, 2007). By analyzing transmission networks, health professionals can identify super-spreader events, pinpoint potential outbreak sources, and gain insights into the dynamics of pathogen transmission (Haydon et al., 2003;Lloyd-Smith et al., 2005). This information is essential for implementing timely interventions and developing targeted public health strategies to effectively slow or halt the progression of an epidemic (Ferguson et al., 2001).

1.2 Review of Infectious Disease: Tuberculosis

Tuberculosis (TB) has been a persistent threat to human health for centuries, with evidence of its existence traced back to ancient civilizations. Known historically as "consumption" due to its wasting effects on the body, TB was once regarded as a death sentence, claiming countless lives before the advent of modern medicine. In the 19th and early 20th centuries, TB was one of the leading causes of mortality worldwide, particularly in overcrowded and impoverished urban areas, where poor sanitation and malnutrition fueled its spread (Doege, 1965).

The discovery of *Mycobacterium tuberculosis* by Robert Koch in 1882 marked a turning point in understanding the disease (KOCH, 1882). This breakthrough spurred research into its pathogenesis, transmission, and treatment. The mid-20th century saw significant advancements, with the development of antibiotics such as streptomycin and isoniazid, which transformed TB from a terminal illness into a treatable condition (Schatz et al., 1944;Takayama et al., 1972). The introduction of the Bacillus Calmette-Guérin (BCG) vaccine further bolstered efforts to prevent TB, especially in high-risk populations (Calmette, 1922). These innovations led to dramatic declines in TB incidence and mortality, prompting optimism that the disease could be eradicated.

Despite these successes, TB remains a major global health challenge. Its ability to persist in latent form within hosts and its association with social determinants of health, such as poverty and malnutrition, have hindered its elimination. The disease disproportionately affects low- and middle-income countries, where access to healthcare and diagnostic services is limited (Rodrigues and Smith, 1990). Moreover, the rise of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB has compounded the difficulty of treatment, necessitating prolonged and costly regimens with lower success rates (Seung et al., 2015).

The importance of maintaining TB as a mainstream focus cannot be overstated. Unlike emerging infectious diseases that often garner significant media and research attention, TB represents a long-standing epidemic that silently impacts millions each year. In 2021 alone, over 10 million new TB cases and 1.6 million deaths were reported, with the majority occurring in resource-limited settings (Bagcchi, 2023). Its intersection with other global health challenges, such as HIV/AIDS, further exacerbates its burden, as co-infected individuals face higher mortality risks and require complex management (Sharma et al., 2005).

Efforts to eliminate TB require sustained investment in research, innovative diagnostic tools, and effective vaccines. While the BCG vaccine has played a critical role in reducing severe forms of TB in children, its efficacy against pulmonary TB in adults—the most transmissible form—remains limited (Andersen and Doherty, 2005). Advancing vaccine development, alongside new drugs and shorter treatment regimens, is essential to overcoming these barriers.

TB also highlights the critical role of public health infrastructure and social interventions. Addressing factors such as overcrowded living conditions, malnutrition, and stigma is key to reducing transmission and ensuring equitable access to care. Furthermore, as the COVID-19 pandemic has demonstrated,

health system disruptions can lead to TB control setbacks, underscoring the need for resilient and adaptable healthcare systems (Martin-Hughes et al., 2022).

In conclusion, tuberculosis is not merely a disease of the past but a pressing modern-day challenge. Its history, from a feared plague to a treatable condition, reflects the remarkable progress of medical science. However, the ongoing global burden of TB serves as a reminder that progress must be sustained through vigilance, innovation, and a steadfast commitment to eliminating this ancient disease once and for all.

1.3 Established frameworks and constraints

Substantial efforts have been devoted to developing statistical and computational tools for constructing transmission networks using epidemiological and genetic data (Gilbertson et al., 2018; Campbell et al., 2019). Reconstruction of transmission networks through contact tracing involves systematically identifying and monitoring individuals who have been in close contact with confirmed cases of an infectious disease (Almutiry and Deardon, 2021). While this survey-based approach is useful for tracking and controlling disease spread, it can be labor-intensive and prone to errors due to the complexity of accurately tracing interpersonal interactions (Gardy et al., 2011).

While contact tracing relies on direct observation, advances in genomic technology offer complementary insights through pathogen sequencing. The increasing availability of genomic data has revolutionized the field of epidemiology, providing a powerful tool for inferring transmission networks of infectious diseases (Klinkenberg et al., 2017). By analyzing the genetic variations in pathogen genomes from infected individuals, researchers can infer transmission events and identify risk factors (Van der Roest et al., 2023). Whole Genome Sequencing (WGS) is particularly effective in environments with high mutation rates, as it can distinguish closely related transmission chains by comparing genetic sequences (Campbell et al., 2018; Bandoy and Weimer, 2021). When combined with epidemiological data, WGS significantly enhances the precision and effectiveness of infectious disease surveillance and control efforts (Duault et al., 2022).

1.3.1 Genome-Driven Inference Framework

Phylogenetic methods provide crucial insights into the evolution, transmission, and management of infectious diseases (Parker et al., 2008;Kendall and Colijn, 2016; Skums et al., 2022). By analyzing phylogenetic variation in pathogen

genomes, researchers can trace the evolutionary history of infectious agents, map transmission routes, and determine the geographic and temporal origins of outbreaks (Ratmann et al., 2017; Didelot et al., 2018; Zhang et al., 2020). If mutation and transmission occur on similar timescales, combining genetic sequence data with epidemiological observations can provide valuable insights into the transmission dynamics of infectious diseases (Morelli et al., 2012; Ypma et al., 2012;Lau et al., 2015; Duault et al., 2022).

1.3.1.1 SNP Thresholds as Tools for Transmission Clustering

A straightforward computational approach for analyzing whole genome sequencing (WGS) data reconstructs transmission events by grouping cases that share fewer single-nucleotide polymorphisms (SNPs) than a preset threshold (Martin et al., 2018; Yang et al., 2018; Coll et al., 2020). While computationally efficient, this method has notable limitations. The selection of the SNP threshold is often arbitrary and uninformed by epidemiological context, which can result in inaccurate or inconsistent groupings. Moreover, grouping cases based solely on genetic similarity fails to provide explicit information about the directionality of transmission or identify "who infected whom." Additionally, the resulting tree represents a phylogeny—depicting genetic relationships—rather than a true transmission tree that maps the actual infection pathways. These limitations underscore the need for more sophisticated frameworks that integrate genomic, epidemiological, and temporal data to improve the accuracy and reliability of transmission inference.

A probabilistic alternative refines this threshold by integrating SNP differences with temporal data and transmission dynamics (Stimson et al., 2019). By incorporating case timing, molecular clock rates, and transmission processes, this method provides a more accurate estimation of the number of transmissions separating cases. It adapts to the inherent variability of mutation rates across the genome and leverages additional contextual information, such as spatial distributions and antibiotic resistance patterns, to improve the precision of transmission clustering. This approach is particularly beneficial in cases where clock rates are inconsistent or sample collection times are widely dispersed, offering a more reliable framework for identifying direct transmission pathways.

1.3.1.2 Bayesian Frameworks for Genomic-Driven Epidemiological Inference

Although probabilistic refinements of SNP threshold-based methods improve clustering accuracy by incorporating temporal data and accounting for muta-

tion rate variability, these approaches remain limited. They primarily focus on grouping cases by genetic similarity, offering little insight into transmission directionality or the full complexity of infection pathways. Additionally, they often overlook critical factors such as within-host genetic diversity, latent periods, and transmission bottlenecks. These limitations necessitate more advanced frameworks that jointly capture epidemiological and evolutionary dynamics.

Within-host Evolution Within-host evolution plays a critical role in understanding pathogen transmission dynamics, particularly for pathogens with long latent periods or high mutation rates. During prolonged incubation within a host, pathogens have sufficient time to accumulate genetic variation through mutation, selection, and drift. The rate of evolution within hosts can significantly affect the genetic divergence between transmitted strains, complicating the inference of direct transmission pathways. These factors highlight the necessity of accounting for within-host evolution in transmission models.

The case of Mycobacterium tuberculosis illustrates this complexity, despite its relatively low mutation rate. With a genome of approximately 4.41 million base pairs (Cole et al., 1998) and a mutation rate of 0.3–0.5 mutations per genome per year (Ford et al., 2013), the prolonged latent period of tuberculosis (TB), often lasting years, allows significant genetic diversity to emerge (Behr et al., 2018). This diversity, shaped by immune responses, antibiotic treatment, and genetic drift, complicates aligning genetic relationships with transmission events. These characteristics underscore the need to account for within-host evolution when modeling TB transmission.

Effective population size is a crucial parameter for quantifying genetic diversity and its influence on transmission dynamics. Incorporating this parameter into transmission models improves the understanding of pathogen diversity and bottleneck effects, enhancing the reconstruction of transmission networks. For example, as illustrated in Figure 1.1, within-host evolution (Panel B) generates genetic variation within a host, while transmission bottlenecks pass only a subset of this diversity to subsequent hosts. This process results in phylogenetic trees (Panel C) that reflect genetic relationships among sampled variants but may not align with the actual transmission tree (Panel A). Such discrepancies highlight the impact of within-host evolution and bottlenecks on the correspondence between phylogenetic and transmission dynamics.

Bayesian Framework To fully capture the epidemiological and evolutionary processes, novel Bayesian frameworks have been developed to simultaneously infer both phylogenetic and transmission trees from epidemiological and genetic

data (Ypma et al., 2013; Klinkenberg et al., 2017; Skums et al., 2022). These frameworks account for genetic variation arising from factors such as latent periods, within-host diversity, and mutation rate heterogeneity, enabling joint inference of both epidemiological and evolutionary dynamics (Didelot et al., 2014; Hall et al., 2015; De Maio et al., 2016). A critical aspect of these Bayesian frameworks is their reliance on biologically informed temporal constraints, which help exclude implausible transmission scenarios and ensure alignment with the natural history of the pathogen.

Temporal data provides these biologically informed constraints, offering parameters such as latent periods, infectious periods, and recovery times that are fundamental for evaluating transmission feasibility. These parameters enable the exclusion of implausible transmission scenarios, ensuring that inferred events align with the pathogen's natural history. For instance, cases with latent periods or infection intervals exceeding biologically plausible limits can be ruled out. By anchoring transmission inference in biological reality, temporal data enhances the accuracy of reconstructing epidemiological and evolutionary dynamics, making it an indispensable component of outbreak analysis. Building on these temporal constraints, recent advancements in transmission inference methods have incorporated genetic distances and other data types, streamlining the reconstruction of outbreak dynamics.

Moreover, a class of transmission network reconstruction methods directly infer transmission routes from observed genetic distances, bypassing the intermediate step of phylogenetic tree inference (Jombart et al., 2014; Ke and Vikalo, 2023). These methods employ functions that quantify the relationship between genetic distance and transmission likelihood, providing a more streamlined approach to reconstructing outbreak dynamics (Worby et al., 2016). De Miao et al. (2018) presented a Bayesian method for inferring host-to-host transmission in the presence of sequencing errors. Recent advancements in computational methods for transmission inference have integrated contact, temporal, and genetic data, leading to a substantial improvement in our capacity to reconstruct transmission trees (Campbell et al., 2019; FFujikura et al., 2019; Dawson et al., 2021). By incorporating additional data types, these approaches have greatly enhanced the accuracy of reconstructing transmission pathways, effectively capturing both the dynamic and geographical aspects of disease spread (Montazeri et al., 2020).

Despite the potential of genomic data to reveal transmission chains, some computational methods fail to differentiate between phylogenetic and transmission trees, leading to inaccurate inferences about disease spread (Nübel et al., 2010; Mutreja et al., 2011; Carson et al., 2024). These methods frequently

assume that the pathogen genomes evolve along the same path as the infection process follows, neglecting the genetic diversity that exists within each infected host. This is particularly problematic for pathogens with high mutation rates and long incubation periods, as these factors can lead to significant genetic variation within a single host (Alizon et al., 2011).

The case of Mycobacterium tuberculosis illustrates this complexity, despite its relatively low mutation rate. With a genome of approximately 4.41 million base pairs (Cole et al., 1998) and a mutation rate of 0.3–0.5 mutations per genome per year (Ford et al., 2013), the prolonged latent period of tuberculosis (TB), often lasting years, allows significant genetic diversity to emerge (Behr et al., 2018). This diversity, shaped by immune responses, antibiotic treatment, and genetic drift, complicates aligning genetic relationships with transmission events. These characteristics underscore the need to account for within-host evolution when modeling TB transmission.

Effective population size is a key parameter for quantifying genetic diversity and its role in transmission dynamics. Incorporating this parameter into transmission models improves the understanding of pathogen diversity and bottleneck effects, thereby enhancing transmission network reconstruction. As shown in Figure 1.1, within-host evolution (Panel B) generates genetic variation, while transmission bottlenecks pass only a subset of this diversity to subsequent hosts. This process results in phylogenetic trees (Panel C) that represent genetic relationships among sampled variants but may not align with the actual transmission tree (Panel A). These discrepancies highlight the need to account for within-host evolution and bottlenecks when modeling transmission dynamics.

Furthermore, certain transmission network construction methods rely on symptom onset time as a proxy for infection time, ignoring the latent period and introducing further inaccuracies (Lau et al., 2015; Ayabina et al., 2018).

To overcome these challenges, we propose a novel Bayesian hierarchical model that incorporates transmission dynamics, mutation processes, within-host diversity, uncertain infection times, and unobserved cases. By acknowledging within-host genetic diversity, our model recognizes that transmitted lineages may differ from those sampled at infection. Additionally, by incorporating the latent period and distinguishing between symptom onset and actual infection time, the model enhances the accuracy of transmission dynamics and epidemiological modeling. The ability to account for unobserved cases further reflects the complexity of real-world transmission events, leading to more realistic models of pathogen spread and improved epidemiological insights for effective public health interventions.

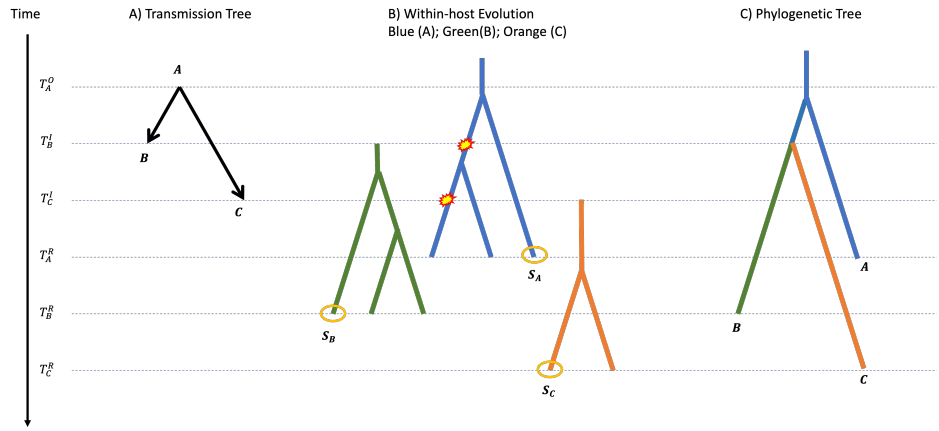


Figure 1.1: Illustration of the relationship between transmission tree, within-host evolution, and phylogenetic tree. Panel A represents the transmission tree, showing the infection pathways between hosts A, B, and C. Panel B illustrates within-host evolution, where genetic diversity (blue for A, green for B, orange for C) accumulates within hosts, with bottlenecks during transmission passing only a subset of variants. Panel C shows the phylogenetic tree, which reflects genetic relationships based on sampled variants.

1.4 Review of Statistical Methodologies

1.4.1 Bayesian Framework

Bayesian inference has its roots in the work of Reverend Thomas Bayes (Bayes, 1991), an 18th-century mathematician who first introduced the concept of updating probabilities based on new evidence. This approach, later formalized and expanded by Pierre-Simon Laplace (Laplace, 1774), provides a systematic framework for reasoning under uncertainty.

At its core, Bayesian inference combines prior knowledge or beliefs about a parameter with observed data to produce a posterior probability distribution. This process is grounded in Bayes' theorem, which mathematically expresses the relationship between the prior probability, the likelihood of the data, and the posterior probability. Unlike frequentist methods, which treat parameters as fixed values, Bayesian inference interprets probability as a measure of belief or certainty, allowing for more flexibility in decision-making and effectively incorporating prior information.

1.4.2 Markov Chain Monte Carlo (MCMC) with Metropolis-Hastings sampling

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms used to approximate complex probability distributions through sampling. These methods are particularly valuable in Bayesian inference, where calculating posterior distributions often involves high-dimensional integrals that are analytically intractable or computationally expensive to solve directly. By constructing a Markov chain that converges to the target distribution, MCMC enables efficient sampling from this distribution, even when the dimensionality of the parameter space is high. The motivation for MCMC lies in its ability to approximate probabilities and expectations with a high degree of accuracy by generating representative samples, making it an indispensable tool for modern statistical methods, particularly in scenarios involving hierarchical models, latent variables, or large datasets.

Metropolis-Hastings (MH) and Gibbs Sampling (GS) are two key methods within the Markov Chain Monte Carlo (MCMC) framework for sampling from complex probability distributions. MH generates a Markov chain by proposing new states from a chosen distribution and accepting them based on an acceptance ratio, offering flexibility for various target distributions. In contrast, GS simplifies sampling by iteratively drawing from the conditional distributions of each parameter, making it particularly effective for structured models like hierarchical Bayesian frameworks when such conditionals are easy to compute. While GS is efficient in these cases, its practicality diminishes for complex models with intractable conditionals. MH, with its broader applicability, is therefore selected for this analysis due to its adaptability and robustness in handling challenging target distributions.

1.4.3 The Coalescent Process

The coalescent process is a fundamental framework in population genetics used to trace the ancestry of gene sequences and infer evolutionary history. By modeling the lineage of sampled alleles, it provides insights into population size, mutation rates, and selection pressures. A central concept is the "coalescence time," representing the time in the past when two or more alleles shared a common ancestor.

In a haploid population with an effective population size N_e, the probability that two sampled gene sequences coalesce in the previous generation is $1/N_e$. The probability of coalescence at a specific generation t is determined by the product of non-coalescence probabilities over $t-1$ generations and coalescence

at generation t. This results in a geometric distribution for coalescence time, with a mean of N_e, expressed as:

$$P(t^* = t) = \left(1 - \frac{1}{N_e}\right)^{t-1} \frac{1}{N_e}.$$

When scaled to continuous time using the generation time g, the mean coalescence time t^* becomes $\theta = N_e g$, the within-host effective population size. For large N_e, the geometric distribution approximates an exponential distribution with a mean of θ, simplifying calculations in continuous time models.

The coalescent process is a robust tool for analyzing genetic diversity and interpreting population dynamics, precisely linking genealogical patterns to demographic and evolutionary events.

1.4.4 Nucleotide Substitution Model

The Jukes-Cantor (JC69) model, introduced by Jukes and Cantor in 1969, is a foundational nucleotide substitution model widely used to reconstruct the evolutionary history of DNA sequences. This model assumes a simple and uniform mutation process, providing a tractable framework for analyzing sequence evolution.

Key Assumptions of the Jukes-Cantor Model

1. **Site Independence**: Mutations occur independently at all sites in the sequence.
2. **Equal Base Frequencies**: All four nucleotides (A, C, G, and T) are equally frequent, with $\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$.
3. **Equal Mutation Probability**: Each nucleotide is equally likely to mutate into any of the other three, with no preference for specific substitutions.

Transition Matrix

The mutation process is described by the continuous-time Markov chain with a transition matrix Q, parameterized by the mutation rate μ. The transition matrix is defined as:

$$Q = \begin{bmatrix} -\frac{3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & -\frac{3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3\mu}{4} \end{bmatrix}.$$

Here, diagonal elements represent the rate of leaving a nucleotide state, while off-diagonal elements represent the rate of substitution between states.

Transition Probability

The probability of nucleotide change from state i to j after time t is derived using the matrix exponential $P(t) = e^{Qt}$. Each entry $P_{ij}(t)$ in the matrix denotes the transition probability, calculated as follows:

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t}, & i = j, \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t}, & i \neq j. \end{cases}$$

For a given evolutionary time t, mutation rate μ, and sequence divergence d (the number of differing sites between two sequences), the probability of observing the aligned sequences is calculated as follows:

$$P(d_{j_1}, d_{j_2} \mid t, \mu, N) = \left(P_{i \neq j}(t)\right)^{N-d} \left(P_{i \neq j}(t)\right)^{d},$$

where N is the total sequence length and d is the SNP number.

Applications and Limitations

The Jukes-Cantor model provides a foundational understanding of sequence evolution and is commonly used as a baseline for more complex models. However, its simplifying assumptions, such as equal base frequencies and uniform mutation rates, limit its accuracy for more realistic scenarios. Nonetheless, its mathematical tractability makes it an essential tool for inferring evolutionary distances and substitution rates.

1.5 Research Phases and Scope

The dissertation comprises two phases, each integrating increasingly detailed data sources and analytical goals.

- **Phase I: Temporal and Genomic Data.**
  The groundwork for this phase was established by Hu, 2023, whose dissertation laid the foundation for combining temporal data (e.g., onset

times) with genomic sequences to infer transmission events. Building on that foundation, I enhanced the Bayesian model by incorporating additional parameters to increase its robustness and conducted a more extensive series of simulation studies under diverse scenarios. In parallel, I refined the data processing pipeline and carried out an in-depth real-world data analysis.

- **Phase II: Temporal, Genomic, and Network Data.**
  With the methods from Phase I in place, I am now extending the model to incorporate network data. This step involves designing additional simulation studies, evaluating the influence of network structures on inference, and applying the enhanced model to real-world datasets. The goal is to determine whether network information can improve the accuracy and interpretability of inferred transmission pathways, while also identifying any practical challenges related to data availability and model complexity.

Through these two phases, the project evolves from a foundational framework combining temporal and genomic data to a sophisticated model that integrates multiple data streams. The efforts in Phase I, bolstered by the foundational work of Hu, 2023, have enabled a seamless transition into Phase II, where network data is further expanding our understanding of transmission processes.

Chapter 2

Materials and Methods

As described in the introduction, we adopt Bayesian methods for two primary reasons: (1) their ability to incorporate prior knowledge into the analysis, enabling more robust and informed inference about the transmission tree of infectious diseases, and (2) their flexibility in handling complex, hierarchical models. These strengths make Bayesian methods particularly well-suited for integrating heterogeneous data sources and accounting for uncertainties inherent in epidemiological studies.

To achieve our objectives, we developed two complementary frameworks. The first framework, based on genomic and temporal data, aligns with established methodologies by utilizing genetic distances and the timing of transmission events to reconstruct transmission trees effectively. The second framework extends this approach by incorporating network information, which provides additional insights into the heterogeneity of infection probabilities through contact likelihoods. In particular, this framework constitutes the first network-based investigation of tuberculosis transmission, further enhancing our nuanced understanding of the dynamics of disease spread.

In this chapter, we first outline the rationale and components of the Bayesian frameworks employed in this study, beginning with a detailed discussion of the input data and parameters that form the foundation of the models. Following this, we provide a rigorous specification of the likelihood and prior distributions underpinning each framework. A comprehensive account of the Markov Chain Monte Carlo (MCMC) algorithm is presented, illustrating its role in deriving posterior estimates and ensuring convergence for reliable results. Additionally, hypothesis testing procedures are described to identify direct transmission events within the inferred transmission network. The chapter also delves into the methodologies used for integrating genomic, temporal, and network information, with the latter providing heterogeneous infection probabilities

based on contact likelihoods. Finally, we explore network analysis using Exponential Random Graph Models (ERGMs) to assess the structural characteristics of the social network, further enriching the understanding of transmission dynamics.

2.1 Input Data and Parameters

2.1.1 Input Data

The Bayesian framework relies on three primary data sources—temporal data, genomic data, and network data. Together, these data sources provide complementary information to reconstruct the transmission network and capture the heterogeneity of infection probabilities.

Temporal Data: Temporal data consist of paired onset and removal times, where onset times

$$T^O = \{T_1^O, T_2^O, \ldots, T_n^O\}$$

represent the first recorded symptoms for n infected individuals, and removal times

$$T^R = \{T_1^R, T_2^R, \ldots, T_n^R\}$$

indicate the time at which individuals recover or are quarantined. These data are sorted by onset times, with T_1^O corresponding to the earliest onset and subsequent times following in ascending order. Patient IDs are assigned based on this order, such that ID_1 represents the individual with the earliest onset time.

The infectious period for each individual is given by

$$t^{inf} = \{T_i^R - T_i^O \mid i = 1, \ldots, n\},$$

representing the duration between their symptom onset time (T_i^O) and removal time (T_i^R). Temporal data provide critical information for reconstructing the transmission network by defining the chronological sequence of infections and constraining potential transmission events to those that respect the order of onset and removal times.

Genomic Data: Genomic data are represented by genetic distances (D), stored as an $n \times n$ matrix, where n is the number of infected individuals. Each entry d_{ij} in the matrix represents the pairwise single nucleotide polymorphism (SNP) distance between the genomes of individuals i and j. SNP distances measure the

number of genetic differences between pathogen genomes, providing a quantitative assessment of their genetic similarity. Smaller d_{ij} values indicate higher genetic similarity and a greater likelihood of direct transmission. This matrix allows for the systematic integration of genomic data into the reconstruction of transmission networks by refining possible transmission pathways based on genetic relationships.

Network Data and Contact Likelihoods: Network data represent the social or contact network structure of the population, where nodes correspond to individuals and edges indicate potential contact or interactions. This information introduces additional insights into the heterogeneity of infection probabilities, as the likelihood of transmission is influenced by the strength or frequency of contact between individuals. Contact likelihoods (w_{ij}) are assigned to each edge, reflecting the probability of interaction and potential transmission between individuals i and j. This data type ensures that the model considers variations in contact patterns, adding realism to the reconstructed transmission network.

2.1.2 Model Parameters

The Bayesian model incorporates several key parameters, which describe the dynamics of transmission and the evolutionary processes of the pathogen. Each parameter is informed by the input data described above.

Transmission Network (Φ): The transmission network (Φ) models the host-to-host spread of pathogens as a directed graph, where nodes represent infected individuals and edges denote direct transmission events. The validity of the transmission direction is governed by temporal data and latent periods, under the key assumption that individuals are infectious (capable of transmitting the pathogen) from the onset of symptoms (T_i^O) until their removal time (T_i^R). Transmission from individual i to j is only feasible if:

$$T_i^O < T_j^O - t_{j,i}^L,$$

where $t_{j,i}^L$ is the latent period, defined as the time from when an individual is infected (receives the pathogen) to the onset of symptoms. In addition to temporal constraints, network data incorporating contact likelihoods (w_{ij}) refine the edges further by quantifying the probability of transmission based on the frequency and intensity of social or physical interactions. This integration ensures a comprehensive representation of pathogen transmission dynamics.

Latent Periods (t^L): Latent periods are defined as

$$t^L = \{t^L_{i,J_i} \mid i = 2, \ldots, n, J_i \in \{1, \ldots, i-1\}\},$$

representing the time delay between the infection of the transmitter (J_i) and the individual they infect (i). These periods are derived from temporal data and contribute to the calculation of infection times.

Infection Times (T^I): Infection times are calculated as

$$T^I = \{T^I_i \mid i = 2, \ldots, n\},$$

where

$$T^I_i = T^O_i - t^L_{i,J_i}.$$

This parameter provides temporal resolution for reconstructing the transmission network.

Infection Rate (α): The infection rate (α) represents the average number of individuals that an infected person transmits the pathogen to within a year. It is informed by the structure of the transmission network (Φ) and temporal data.

Removal Rate (β): The removal rate (β) quantifies the average time it takes for an infected individual to recover from the onset of symptoms. This parameter captures recovery dynamics and is informed by temporal data, specifically the recovery times (T^R) and symptom onset times (T^O). The recovery time for each individual is calculated as the difference between their removal time (T^R_i) and symptom onset time (T^O_i). This information provides critical insight into the infectious period and overall disease progression.

Mutation Rate (μ): The mutation rate (μ) is the number of mutations per site per year, representing the rate of pathogen evolution. It links genomic data (D) to the transmission network by quantifying genetic differences between pathogens.

Effective Population Size Parameter (θ): The effective population size parameter (θ) quantifies the average number of mutations introduced per generation per site within the host population of pathogens. It is calculated as:

$$\theta = \mu N_e g,$$

where μ is the mutation rate per site per year, N_e is the effective population size of pathogens, and g is the generation time of pathogens.

2.2 Fundamental Assumptions

2.2.1 No co-infection

It is assumed that co-infection doesn't exist. Each infected patient could be traced back to the most likely transmitters.

2.2.2 Uninfectious during the latent period

It is assumed that each individual has three time points: time of infection, symptom onset time, and removal time. Each individual is only infectious/contagious during the period from symptom onset to removal. In other words, the individual stay uninfectious during the latent period.

2.2.3 Relaxed bottleneck assumption

Here, we relaxed the complete bottleneck assumption that only a single pathogen could pass to the infected at the time of infection.

2.3 The Bayesian Framework with Temporal and Genomic Data

With the input data and parameters well-defined, this framework integrates temporal and genomic data to infer the transmission network. Temporal data, comprising symptom onset (T^O) and removal times (T^R), provide critical information about the order and duration of infectious periods, ensuring the inferred transmission tree respects the observed chronology. Genomic data, represented as pairwise genetic distances (D), quantify the evolutionary relationships between pathogens, enabling the refinement of potential transmission pathways based on genetic similarity. Together, these complementary data sources provide a robust foundation for reconstructing the transmission tree.

The framework is parameterized by the transmission network Φ, which identifies who infected whom, and the latent periods t^L, representing the time between infection and symptom onset for individuals. Additional parameters include the infection rate α, which determines how quickly infections occur, the removal rate β, which captures the average duration of infectiousness, the

mutation rate μ, which describes the rate at which genetic changes occur, and the effective population size parameter θ, which shapes the genetic diversity of the pathogen. Together, these parameters govern the dynamics of transmission and pathogen evolution within the population.

Using these data and parameters, the posterior probability distribution integrates the likelihood of the temporal and genomic data with prior distributions over the parameters and transmission network. The posterior is expressed as:

$$\begin{aligned} P(\Phi, t^L, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}}) \propto\ & P(D, T^O, t^{\text{inf}} \mid \Phi, t^L, \theta, \alpha, \beta, \mu) \\ & \times P(\Phi, t^L, \theta, \alpha, \beta, \mu) \end{aligned} \tag{2.1}$$

2.3.1 Likelihood

The likelihood function $P(D, T^O, t^{\text{inf}} \mid \Phi, t^L, \theta, \alpha, \beta, \mu)$ can be factored into a product of three conditional probabilities, i.e.,

$$\begin{aligned} P(D, T^O, t^{\text{inf}} \mid \Phi, t^L, \theta, \alpha, \beta, \mu) &= P(D \mid T^O, t^{\text{inf}}, \Phi, t^L, \theta, \alpha, \beta, \mu) \\ &\quad \times P(T^O \mid t^{\text{inf}}, \Phi, t^L, \theta, \alpha, \beta, \mu) \\ &\quad \times P(t^{\text{inf}} \mid \Phi, t^L, \theta, \alpha, \beta, \mu) \\ &= P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu) \\ &\quad \times P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha) \\ &\quad \times P(t^{\text{inf}} \mid \beta) \end{aligned} \tag{2.2}$$

2.3.1.1 $P(t^{\text{inf}} \mid \Phi, t^L, \beta)$

Assuming that infectious periods t^{inf} follow an exponential distribution with rate β, the conditional probability can be factorized as a product of individual exponential probabilities for each infectious period:

$$P(t^{\text{inf}} \mid \beta) = \prod_{i=1}^{n} P(t_i^{\text{inf}} \mid \beta) = \prod_{i=1}^{n} \beta e^{-\beta t_i^{\text{inf}}} \tag{2.3}$$

This compact representation simplifies the likelihood calculation by expressing it as a product of exponential terms.

2.3.1.2 $P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha)$

The conditional probability of onset times $P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha)$ can be factorized as the product of joint probabilities for the onset times of patients who share the same infector. Denoting m as the number of individuals who have

transmitted the infection to at least one other person, each of these individuals, labeled as $j_1, \ldots, j_m$, has infected a certain number of individuals, denoted k_{j_i}, who are represented as $\{l_1, \ldots, l_{k_{j_i}}\}$. Given the infection rate α, the waiting time until the next infection follows an exponential distribution with density $\alpha e^{-\alpha t}$. For each transmitter j_i, the probability of observing the infections k_{j_i} is expressed as the product of exponential densities of these waiting times. This yields:

$$P\left(T^O_{l_1}, \ldots, T^O_{l_{k_{j_i}}} \mid T^O_{j_i}, t^{\text{inf}}_{j_i}, t^L, \alpha, \Phi\right) = \prod_{k=1}^{k_{j_i}} \alpha e^{-\alpha\left(T^O_{l_k} - T^O_{l_{k-1}}\right)} + e^{-\alpha\left(T^O_{j_i} + t^{\text{inf}}_{j_i} - T^O_{l_{k_{j_i}}}\right)}$$
$$= \alpha^{k_{j_i}} e^{-\alpha t^{\text{inf}}_{j_i}} \tag{2.4}$$

This product of joint probabilities, which accounts for each individual's latent period to recover their true infection time, provides a comprehensive view of the transmission dynamics for each transmitter. Aggregating the probabilities across all transmitters, we obtain:

$$P(T^O \mid \Phi, t^{\text{inf}}, t^L, \alpha) = \prod_{i=1}^{m} \left(\alpha^{k_{j_i}} e^{-\alpha t^{\text{inf}}_{j_i}}\right) = \alpha^{n-m} e^{-\alpha \sum_{i=1}^{m} t^{\text{inf}}_{j_i}} \tag{2.5}$$

This factorized representation enables us to depict the infection and onset times for each transmitter, thus illustrating the transmission patterns throughout the network.

2.3.1.3 $P(D \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu)$

The probability $P(D \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu)$ of the sequence data D is obtained by integrating the joint density $P(D, \Psi \mid T^0, t^{inf}, \Phi, t^L, \theta, \mu)$ over all possible phylogenetic trees Ψ, i.e.,

$$P(D \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu) = \int_{\Psi} P(D \mid \Psi) P(\Psi \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu) d\Psi$$

This means that $P(D \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu)$ represents the probability of observing the sequence data D given the transmission tree Φ and other parameters (t^L, θ, α, μ) without considering a specific phylogenetic tree. Given that the $(n-1)$ transmissions $\{\phi_2, \ldots, \phi_n\}$ in the network Φ occur independently, the probability $P(D \mid T^0, t^{inf}, \Phi, t^L, \theta, \mu)$ can be expressed as the product of the probabilities of the two sequences associated with each of the $(n-1)$ transmissions $\{\phi_2, \ldots, \phi_n\}$, i.e.,

$$P(D \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu) = \prod_{i=2}^{n} P(D \mid \phi_i, T^O, t^{inf}, t^L, \theta, \mu)$$
$$= \prod_{i=2}^{n} P(d_{J_i}, d_i \mid \phi_i, T^O_{J_i}, T^O_i, t^{inf}_{J_i}, t^{inf}_i, t^L_{i,J_i}, \theta, \mu) \tag{2.6}$$

In (6), d_{J_i} and d_i are the sequences of the individuals J_i and i associated with the transmission $\phi_i : J_i \to i$. Let x_{ij} and p_{ij} be the frequency and probability of the nucleotide doublet ij for $i, j \in \{A, C, G, T\}$. The term $P(d_{J_i}, d_i \mid \phi_i, T^O_{J_i}, T^O_i, t^{inf}_{J_i}, t^{inf}_i, t^L_{i,J_i}, \theta, \mu)$ in (6) is a multinomial probability given by

$$P(d_{J_i}, d_i \mid \phi_i, T^O_{J_i}, T^O_i, t^{inf}_{J_i}, t^{inf}_i, t^L_{i,J_i}, \theta, \mu)$$
$$= \frac{N!}{\prod_{i\in\{A,C,G,T\}} \prod_{j\in\{A,C,G,T\}} x_{ij}!} \prod_{i\in\{A,C,G,T\}} \prod_{j\in\{A,C,G,T\}} (p_{ij})^{x_{ij}} \tag{2.7}$$

The probability p_{ij} of observing the nucleotide doublet ij at a particular site in the sequences d_{Ji} and d_i can be calculated using a substitution model and a coalescence process that describes the evolution of nucleotides within a host. To simplify the calculations, we assume the Jukes-Cantor model for nucleotide substitutions (Jukes and Cantor 1969), but any substitution model can be used. Under the Jukes-Cantor model, the 16 possible doublet patterns can be reduced to two, and the multinomial probability in (8) simplifies to a binomial probability:

$$P(d_{Ji}, d_i \mid \phi_i, T^O_{J_i}, T^O_i, t^{inf}_{J_i}, t^{inf}_i, t^L_{i,J_i}, \theta, \mu) = \frac{N!}{x_i!(N - x_i)!} p_i^{x_i} (1 - p_i)^{N - x_i} \tag{2.8}$$

where N is the length (i.e., the number of sites) of the sequence alignment D, p_i is the probability of a mutation in a single site, and x_i is the frequency of mutations occurring between the sequences d_{Ji} and d_i.

The probability p_i of a mutation in a single locus occurring between the sequences d_{Ji} and d_i is calculated using the Jukes-Cantor model and a coalescence process that describes the evolution of nucleotides within a host, as detailed in Section 1.4:

$$p_i = \frac{3}{4} - \frac{3}{4}e^{-\mu \cdot t} \tag{2.9}$$

Here, t denotes the evolutionary time, which corresponds to the branch length in the phylogeny tree. Reconstructing the phylogeny tree of the two sequences is essential to determine t, as it captures the evolutionary relationships between the sequences.

t: Evolutionary Time (Branch Duration) The evolutionary time t is defined as the total length of the branch that connects the pathogens sampled d_{Ji} and d_i to their most recent common ancestor (MRCA). Specifically, t represents the sum of the time intervals from the MRCA to the removal times of J_i and i. Mathematically, t is expressed as:

$$\begin{aligned} t &= (T_{J_i}^R - T^{CA}) + (T_i^R - T^{CA}) \\ &= [(T_{J_i}^R - T_i^I) + (T_i^I - T^{CA})] + [(T_i^R - T_i^I) + (T_i^I - T^{CA})] \\ &= (t_1 + t^*) + (t_2 + t^*) = t_1 + t_2 + 2t^*, \end{aligned} \tag{2.10}$$

Here, T^{CA} represents the coalescence time of two sequences d_{Ji} and d_i. The term $t_1 = T_{Ji}^R - T_i^I$ denotes the duration from i's infection to J_i's removal, while $t_2 = T_i^R - T_i^I$ represents the duration from i's infection to i's removal. The term t^* captures the time to coalescence, tracing back to the most recent common ancestor (MRCA) for both lineages. Together, t_1, t_2, and $2t^*$ provide a complete decomposition of the total time. The relationships between t_1, t_2, t^*, and their connection to the coalescence time T^{CA} are visually illustrated in Figure 2.1, which depicts the within-host evolution of Patients A and B, the transmission event, and the phylogenetic tree showing the coalescence process.

In the absence of a complete bottleneck, the MRCA may predate the infection time of the transmitter J_i. In such cases, the time to the MRCA, t^*, can become unbounded. This emphasizes the complexity of tracing lineage relationships when transmission dynamics do not enforce strict genetic bottlenecks.

Combining (7), (9), and (10), the probability $P(D \mid T^O, t^{inf}, \Phi, t^L, \theta, \mu)$ of the sequence data D is given by:

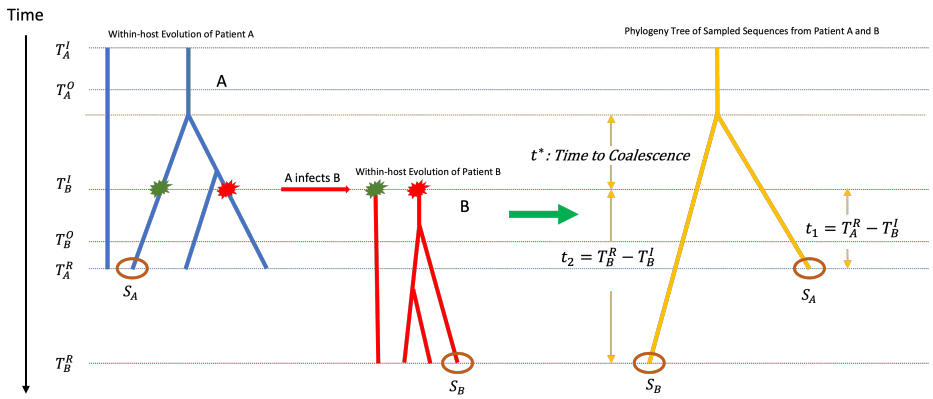


Figure 2.1: Timeline of within-host evolution and coalescence between Patient A (blue) and Patient B (red). The green star marks the transmission from A to B at T_B^I. The phylogeny (right) shows the coalescence time t^* tracing back to the MRCA. Key intervals include $t_1 = T_A^R - T_B^I$ (A's removal to B's infection) and $t_2 = T_B^R - T_B^I$ (B's infection to removal).

$$P(D \mid T^O, t^{\text{inf}}, \Phi, t^L, \theta, \mu) = \prod_{i=2}^{n} \left(\frac{3}{4} - \frac{3}{8\theta + 4} e^{-\mu(t_{i,1}+t_{i,2})} \right)^{x_i} \times \left(\frac{1}{4} + \frac{3}{8\theta + 4} e^{-\mu(t_{i,1}+t_{i,2})} \right)^{N-x_i} \tag{2.11}$$

Under the assumption of a homogeneous effective population size parameter θ across all hosts, the coalescent model for within-host evolution remains valid and integrates the roles of intermediators in the transmission chain. Consequently, the Bayesian model can effectively account for missing samples, indicating that the $n-1$ transmissions in the transmission network may be indirect.

2.3.2 Prior

Given that the transmission network Φ, the latent periods t^L, the population size parameter θ, the mutation rate μ, and the infection rate α are independent of each other, the joint prior $P(\Phi, t^L, \theta, \alpha, \mu)$ is equal to the multiplication of five independent priors $P(\Phi)$, $P(t^L)$, $P(\theta)$, $P(\alpha)$, and $P(\mu)$, i.e.,

$$P(\Phi, t^L, \theta, \alpha, \mu) = P(\Phi) \times P(t^L) \times P(\theta) \times P(\alpha) \times P(\mu) \tag{2.12}$$

Given that the $(n-1)$ transmissions in the transmission network $\Phi = \{\phi_2, \ldots, \phi_n\}$ occur independently, the prior probability $P(\Phi)$ of the transmission network $\Phi = \{\phi_2, \ldots, \phi_n\}$ is the product of the prior probabilities of $(n-1)$ transmissions, i.e.,

$$P(\Phi) = \prod_{i=2}^{n} P(\phi_i) \tag{2.13}$$

The prior probability of a transmission ϕ_i is equal to the probability that J_i is the infector associated with the transmission ϕ_i, i.e., $P(\phi_i) = P(J_i)$. When no contact information is available, the prior probability $P(J_i)$ is assumed to follow a discrete uniform distribution, i.e.,

$$P(J_i) = \frac{1}{i-1} \tag{2.14}$$

The prior probability $P(t^L)$ of latent periods t^L is the product of individual latent period probabilities, i.e.,

$$P(t^L) = \prod_{i=2}^{n} P(t^L_{J_i,i}) \tag{2.15}$$

where each latent period $t^L_{J_i,i}$ follows a truncated scaled χ^2 distribution with the upper bound $T^O_i - T^O_{J_i}$ and the lower bound $\max(0, T^O_i - T^R_{J_i})$.

The priors for the effective population size parameter θ, mutation rate μ, and infection rate α are assumed to be exponential distributions with respective rates λ_θ, λ_μ, and λ_α.

2.4 The Bayesian Framework with Temporal, Genomic and Network Data

Our Bayesian framework integrates temporal, genomic, and network information to infer infection dynamics by building upon prior methods developed for temporal and genomic data. The key innovation is the incorporation of network data G as a hyperparameter, which updates the transmission tree prior from $P(\Phi)$ to $P(\Phi \mid G)$. This modification shifts the model from a uniform infection probability to one that leverages network-derived contact likelihood based on social or spatial proximity. The posterior is iteratively updated to reflect complex interdependencies across temporal, genomic, and network domains, yielding a comprehensive and precise representation of infection dynamics and enabling robust inference of transmission pathways.

Below, we outline the posterior probability as below:

$$
\begin{aligned}
&P(\Phi, t^L, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}}, G) \\
\propto &P(D, T^O, t^{\text{inf}}, G \mid \Phi, t^L, \theta, \alpha, \beta, \mu) \times P(\Phi, t^L, \theta, \alpha, \beta, \mu) \\
\propto &P(D \mid T^O, t^{\text{inf}}, \Phi, t^L, \theta, \alpha, \beta, \mu) \times P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha) \\
&\times P(t^{\text{inf}} \mid \beta) \times P(\Phi \mid G) \times P(G) \\
&\times P(t^L) \times P(\theta) \times P(\alpha) \times P(\beta) \times P(\mu). \qquad (2.16)
\end{aligned}
$$

2.4.1 Prior of the transmission tree $P(\Phi \mid G)$

Assuming independence among transmission events, the prior probability of the transmission tree, $P(\Phi \mid G)$, can be expressed as the product of $n-1$ independent transmission events:

$$P(\Phi \mid G) = \prod_{i=2}^{n} P(\phi_i \mid G).$$

Each term $P(\phi_i \mid G)$ depends on the proximity between individual i and its corresponding transmitter J_i within the network, as quantified by the distance g_{i,J_i}. Recognizing that contact probability typically decays rapidly with increasing distance, we model this relationship using an exponential function, assuming that the transmission probability is inversely proportional to the exponential of the distance:

$$P(\phi_i \mid G) \propto \frac{1}{e^{g_{i,J_i}}}.$$

To convert these unnormalized weights into valid probabilities, we assign an arbitrary scalar k to each transmission event. For a given individual i with transmitter J_i, the unnormalized weight is

$$\frac{k}{e^{g_{i,J_i}}},$$

where g_{i,J_i} denotes the distance between individual i and its transmitter J_i. We then sum the weights over all possible transmitters (i.e., for all $j = 1, \ldots, i-1$) and normalize by dividing the individual weight by this sum. Since the scalar k appears in both the numerator and denominator, it cancels out, yielding

$$P(\phi_i \mid G) = \frac{\frac{1}{e^{g_{i,J_i}}}}{\sum_{j=1}^{i-1} \frac{1}{e^{g_{i,j}}}}.$$

We denote the resulting normalization constant by k_i, so that k and k_i effectively represent the same scaling factor, ensuring that the probabilities sum to one across all possible transmitters.

2.4.2 $P(G)$

The probability $P(G)$ represents the likelihood of observing a particular network G among n nodes. Given that there are $\binom{n}{2}$ possible pairs of nodes and each pair can either be connected or not, there exist $2^{\binom{n}{2}}$ possible network configurations. However, because G is treated as fixed and known, and $P(G)$ does not depend on the model parameters, it is absorbed into the normalizing constant of the model. Consequently, there is no need to explicitly specify $P(G)$ during model evaluation.

2.5 Markov Chain Monte Carlo (MCMC) with Metropolis-Hastings sampling

Due to the intractability of the posterior probability distributions, we approximate it by generating a sample of model parameters using a Metropolis-Hastings-based MCMC approach. In our framework, the parameter set comprises μ (the mutation rate), α (the infection rate), θ (the effective population size), and Φ, where Φ further includes t_b (latent period) and J (Infection ID). Notably, both Bayesian models—whether incorporating a contact network or not—utilize this identical set of parameters. Consequently, the initialization process and the proposal mechanisms for new parameter values can be described uniformly across models, with differences arising only in the computation of the Hastings ratio, which is tailored to reflect the distinct structural elements of the two models.

2.5.1 Initialization of Model Parameters.

In the first stage of the MCMC initialization, we identify the most probable transmitter by assigning each individual i an infector J_i from the set $\{1, 2, \ldots, i-1\}$. The assignment is chosen so that the observed genetic differences (SNPs) between the candidate infector's sequence d_j and the newly infected individual's sequence d_i are minimized. In other words, for each i, we compare d_i to all possible d_j where $j < i$, compute the SNP differences, and pick the j that yields the fewest discrepancies. By relying on this minimal-SNP criterion, we construct an initial transmission network that is most genetically plausible,

providing a solid starting point for subsequent updates within the MCMC procedure.

Based on the assumption that an individual's infectious period spans from the symptom onset time T^O to the removal time T^R, we obtain explicit bounds for the latent period associated with each transmission event. For a transmission event in which individual j infects individual i (with $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$), we assume that the latent period $t^L_{i,j}$ follows a scaled, truncated chi-squared distribution. This distribution is defined over the interval $\left(t^{L,\text{lb}}_{i,j},\, t^{L,\text{ub}}_{i,j}\right)$, where the lower bound is given by

$$t^{L,\text{lb}}_{i,j} = \max\left(0,\, T^O_i - T^R_j\right)$$

and the upper bound by

$$t^{L,\text{ub}}_{i,j} = T^O_i - T^O_j.$$

Using these bounds, the initial value of $t^L_{i,j}$ is sampled from a uniform distribution over the interval $\left(t^{L,\text{lb}}_{i,j},\, t^{L,\text{ub}}_{i,j}\right)$ for each transmission event, and these values are stored in an $n \times n$ matrix.

In the previous section, we assumed that the mutation rate μ, effective population size θ, and infection rate α have exponential priors with rate parameters λ_μ, λ_θ, and λ_α, respectively. Their initial values are drawn from uniform distributions over the following intervals:

$$\begin{aligned}
(\theta_{lb}, \theta_{ub}) &= \left(1 \times 10^{-6},\, 5 \times 10^{-6}\right), \\
(\mu_{lb}, \mu_{ub}) &= \left(5 \times 10^{-7},\, 1 \times 10^{-6}\right), \\
(\alpha_{lb}, \alpha_{ub}) &= (3,\, 5).
\end{aligned}$$

This approach provides a broad yet plausible initialization for the subsequent Markov chain Monte Carlo (MCMC) sampling process.

Based on the observed input values for symptom onset time (T^O) and removal time (T^R), the infectious period for each individual is calculated as $T^R - T^O$. Under the assumption that the removal process is determined solely by this interval, the removal rate β is estimated as the average infectious period across all n individuals:

$$\beta = \frac{1}{n} \sum_{i=1}^{n} \left(T^R_i - T^O_i\right).$$

2.5.2 Proposal of New Values

In our MCMC procedure, parameter updating relies on three components. First, iteration allocation is tuned to each parameter's convergence rate: parameters that converge rapidly are updated less frequently, while those that require more iterations are updated more often. Second, candidate proposals are generated using a random walk. For continuous parameters, a new value is obtained by perturbing the current value with a random step drawn from a uniform distribution defined by preset bounds; for discrete parameters, a new candidate is randomly chosen from the set of permissible alternatives. Third, the Metropolis-Hastings acceptance rule determines whether to adopt the proposed update by comparing the posterior density of the proposed state with that of the current state. This strategy ensures efficient exploration of the parameter space and lays the foundation for further detailed discussion of each component.

2.5.2.1 Iteration Allocation

To tailor computational effort, we assign fixed iteration proportions based on each parameter's convergence rate. Specifically, the mutation rate μ is updated in 10% of the iterations, while the effective population size θ, the infection rate α, and the transmission tree parameters (i.e., branch times t_b and infection identifiers J) are each updated in 30% of the iterations. At each iteration, a random number r is drawn from a uniform distribution on $[0, 1)$ to determine which parameter to update: if $r < 0.1$, update μ; if $0.1 \leq r < 0.4$, update θ; if $0.4 \leq r < 0.7$, update α; and if $0.7 \leq r < 1$, update the transmission tree parameters. This strategy efficiently allocates computational resources, ensuring effective exploration of the parameter space and convergence toward the target posterior distribution.

2.5.2.2 Proposal of new values

A random walk proposal mechanism is used to generate candidate values for the continuous parameters α, θ, μ, and the latent period t^L. For each parameter, the current value is perturbed by adding a random step drawn from a uniform distribution. To illustrate, consider the parameter θ. Its candidate value is proposed from a uniform distribution with bounds defined as follows:

$$U_{\text{lb}} = \max\left(\theta_{\textbf{lb}},\ \theta_{\textbf{current}} - \text{Step Size}\right),$$

$$U_{\text{ub}} = \min\left(\theta_{\textbf{ub}},\ \theta_{\textbf{current}} + \text{Step Size}\right).$$

Here, $\theta_{\mathbf{lb}}$ and $\theta_{\mathbf{ub}}$ denote the predefined lower and upper bounds for θ. For the parameters α, θ, and μ, the lower bound is set to 0, and an upper bound of 1 is imposed for θ and μ. The latent period t^L is updated similarly using its own prior-defined bounds. The step size determines the magnitude of the perturbation, ensuring efficient exploration of the parameter space during the MCMC process.

For the discrete infection identifier J_i corresponding to individual i, we first compile the list of all potential infectors, that is, all $j \in \{1, 2, \ldots, i-1\}$. Then, we explicitly remove the current infection identifier J_i from this list. Next, we randomly shuffle the remaining candidate IDs. Finally, we sequentially consider each candidate from this shuffled list as a new proposed value for J_i. This procedure ensures that the proposed update represents a genuine alternative transmission pathway, thereby facilitating a thorough exploration of the transmission network.

2.5.2.3 Update Schemes

In the Metropolis-Hastings algorithm, the acceptance ratio, often called the Hastings ratio, is computed as the product of two components: the ratio of the target (posterior) densities of the proposed and current states, and the ratio of the proposal probabilities for the reverse and forward moves. Mathematically, this is expressed as

$$r = \frac{\pi(x')}{\pi(x)} \times \frac{q(x \mid x')}{q(x' \mid x)},$$

where $\pi(x)$ is the posterior density and $q(x' \mid x)$ is the proposal probability. In many cases, when the proposal distribution is symmetric (i.e., $q(x' \mid x) = q(x \mid x')$), the proposal terms cancel out, and the ratio simplifies to the ratio of the posterior densities. In our setting, fixed lower and upper bounds and a specified step size in the random walk proposal create an asymmetric proposal distribution. Consequently, the acceptance ratio must explicitly incorporate the proposal probability ratio.

For computational stability and efficiency, we work with the logarithm of the acceptance ratio:

$$\log r = \log\left(\frac{\pi(x')}{\pi(x)}\right) + \log\left(\frac{q(x \mid x')}{q(x' \mid x)}\right),$$

which simplifies to

$$\log r = \big[\log \pi(x') - \log \pi(x)\big] + \big[\log q(x \mid x') - \log q(x' \mid x)\big].$$

Using the logarithmic form transforms multiplications into additions and helps prevent numerical underflow during computation.

The new state is accepted with probability $\alpha = \min(1, r)$, ensuring that moves toward higher posterior density are favored while still allowing exploration of lower-density regions. At each iteration, the Metropolis-Hastings algorithm determines whether to accept or reject the proposed parameter value based on the Hastings ratio. A random number k is drawn from a uniform distribution between 0 and 1, and the acceptance probability A is calculated using the formula

$$A = \min(1, \exp(H)),$$

where H is the log Hastings ratio. If $A > k$, the proposed value is accepted; otherwise, it is rejected. This algorithm is implemented in Julia, a high-level programming language known for its efficient performance and concise syntax.

θ The effective population size parameter θ is updated using a random walk with a step size of 5×10^{-6}. Since θ is constrained between 0 and 1, its proposal interval is given by

$$U_{\text{lb}} = \max\left(0,\, \theta_{\text{current}} - 5 \times 10^{-6}\right), \quad U_{\text{ub}} = \min\left(1,\, \theta_{\text{current}} + 5 \times 10^{-6}\right).$$

A candidate value θ' is drawn uniformly from this interval, yielding a proposal density of

$$q(\theta' \mid \theta) = \frac{1}{U_{\text{ub}} - U_{\text{lb}}}.$$

The reverse density $q(\theta \mid \theta')$ is defined analogously.

With the proposal distribution clearly defined, we now derive the log Hastings ratio. First, we consider the Bayesian framework without network information. The log Hastings ratio is given by

$$\begin{aligned}
H_{1,A} &= \log\left(\frac{P(\Phi, t^L, \theta', \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})} \times \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right) \\
&= \log\left(\frac{P(D \mid T^O, T^R, \Phi, t^L, \theta', \mu)\, P(\theta')}{P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu)\, P(\theta)} \times \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right) \\
&= \log P(D \mid T^O, T^R, \Phi, t^L, \theta', \mu) + \log P(\theta') + \log q(\theta \mid \theta') \\
&\quad - \log P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu) - \log P(\theta) - \log q(\theta' \mid \theta).
\end{aligned}$$

Next, we extend this formulation to the Bayesian framework that incorporates network information.

$$
\begin{aligned}
H_{1,B} &= \log\left(\frac{P(\Phi, t^L, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}}, G)} \times \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right) \\
&= \log\left(\frac{P(D \mid T^O, T^R, \Phi, t^L, \theta', \mu)\, P(\theta')}{P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu)\, P(\theta)} \times \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right) \\
&= \log P(D \mid T^O, T^R, \Phi, t^L, \theta', \mu) + \log P(\theta') + \log q(\theta \mid \theta') \\
&\quad - \log P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu) - \log P(\theta) - \log q(\theta' \mid \theta).
\end{aligned}
$$

$\boldsymbol{\mu}$ For the mutation rate μ, a finer step size of 1×10^{-7} is chosen. With μ restricted to the interval $[0, 1]$, the proposal interval is defined as

$$
U_{\text{lb}} = \max\left(0,\ \mu_{\text{current}} - 1 \times 10^{-7}\right), \quad U_{\text{ub}} = \min\left(1,\ \mu_{\text{current}} + 1 \times 10^{-7}\right).
$$

A candidate μ' is sampled uniformly from this interval, resulting in a proposal density of

$$
q(\mu' \mid \mu) = \frac{1}{U_{\text{ub}} - U_{\text{lb}}}.
$$

The reverse move is defined similarly to properly capture any asymmetry.

For the Bayesian framework without network data, the log Hastings ratio for updating μ is defined as

$$
\begin{aligned}
H_{1,A} &= \log \frac{P(\Phi, t^L, \theta, \alpha, \beta, \mu' \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})} + \log \frac{q(\mu \mid \mu')}{q(\mu' \mid \mu)} \\
&= \log \frac{P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu')\, P(\mu')}{P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu)\, P(\mu)} + \log \frac{q(\mu \mid \mu')}{q(\mu' \mid \mu)} \\
&= \Big[\log P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu') + \log P(\mu') + \log q(\mu \mid \mu')\Big] \\
&\quad - \Big[\log P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu) + \log P(\mu) + \log q(\mu' \mid \mu)\Big].
\end{aligned}
$$

When network data G is incorporated, the log Hastings ratio becomes

$$
\begin{aligned}
H_{1,B} &= \log \frac{P(\Phi, t^L, \theta, \alpha, \beta, \mu', G \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})} + \log \frac{q(\mu \mid \mu')}{q(\mu' \mid \mu)} \\
&= \log \frac{P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu')\, P(\mu')}{P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu)\, P(\mu)} + \log \frac{q(\mu \mid \mu')}{q(\mu' \mid \mu)} \\
&= \log P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu') + \log P(\mu') + \log q(\mu \mid \mu') \\
&\quad - \log P(D \mid T^O, T^R, \Phi, t^L, \theta, \mu) + \log P(\mu) + \log q(\mu' \mid \mu).
\end{aligned}
$$

In both cases, the log Hastings ratio is composed of the difference between the log posterior (or likelihood-prior product) terms and the log proposal density terms. The inclusion of network data G in the second case does not alter the form of the ratio but simply adds the additional conditioning information in the joint distribution.

α The infection rate α is updated with a step size of 0.3, reflecting its expected scale. Unlike θ and μ, α has only a lower bound of 0. Thus, the proposal interval is defined by

$$U_{\text{lb}} = \max\left(0,\ \alpha_{\text{current}} - 0.3\right), \quad U_{\text{ub}} = \alpha_{\text{current}} + 0.3.$$

A candidate α' is drawn from a uniform distribution over this interval, resulting in a proposal density of

$$q(\alpha' \mid \alpha) = \frac{1}{U_{\text{ub}} - U_{\text{lb}}}.$$

The reverse proposal density is defined in the same way to ensure that the asymmetry in the proposal mechanism is properly accounted for in the acceptance ratio.

For the Bayesian framework without network data, the log Hastings ratio for updating the infection rate α is defined as

$$\begin{aligned}
H_{1,A} &= \log\left(\frac{P(\Phi, t^L, \theta, \alpha', \beta, \mu \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})} \times \frac{q(\alpha \mid \alpha')}{q(\alpha' \mid \alpha)}\right) \\
&= \log\left(\frac{P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha')\, P(\alpha')}{P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha)\, P(\alpha)} \times \frac{q(\alpha \mid \alpha')}{q(\alpha' \mid \alpha)}\right) \\
&= \Big[\log P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha') + \log P(\alpha') + \log q(\alpha \mid \alpha')\Big] \\
&\quad - \Big[\log P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha) + \log P(\alpha) + \log q(\alpha' \mid \alpha)\Big].
\end{aligned}$$

For the Bayesian framework incorporating network data, the log Hastings ratio for updating α becomes

$$\begin{aligned}
H_{1,B} &= \log\left(\frac{P(\Phi, t^L, \theta, \alpha', \beta, \mu, G \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})} \times \frac{q(\alpha \mid \alpha')}{q(\alpha' \mid \alpha)}\right) \\
&= \log\left(\frac{P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha')\, P(\alpha')}{P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha)\, P(\alpha)} \times \frac{q(\alpha \mid \alpha')}{q(\alpha' \mid \alpha)}\right) \\
&= \log P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha') + \log P(\alpha') + \log q(\alpha \mid \alpha') \\
&\quad - \log P(T^O \mid t^{\text{inf}}, \Phi, t^L, \alpha) + \log P(\alpha) + \log q(\alpha' \mid \alpha).
\end{aligned}$$

Φ During each iteration of updating the transmission tree Φ, we select an individual $i \in \{2, \ldots, n\}$ and update both its infection identifier J_i and its latent period t^L.

To explore alternative transmission pathways, we first identify all candidate transmitters $j \in \{1, \ldots, i-1\}$ (excluding the current J_i) and randomly shuffle this candidate list to avoid ordering bias. For a candidate transmission event $j \to i$, the latent period is updated—subject to its upper and lower bounds—so that it remains consistent with the transmitter's infectious period. The update employs an adaptive step size computed as

$$\epsilon_{i,j} = \frac{T_j^O - T_i^I}{k},$$

with k defaulting to 5, which scales the available interval appropriately while ensuring that the revised latent period meets all necessary temporal constraints. Thus, the proposal interval is defined as

$$\begin{aligned} U_{\text{lb}} &= \max\left(\max\left(0,\, T_i^O - T_j^R\right),\, t_{i,j}^L - \epsilon_{i,j}\right) \\ U_{\text{ub}} &= \min\left(T_i^O - T_j^O,\, t_{i,j}^L + \epsilon_{i,j}\right) \end{aligned}$$

We then draw the candidate $t_{i,j}^{L'}$ from a uniform distribution over the interval $\left[U_{\text{lb}},\, U_{\text{ub}}\right]$. This yields a proposal density

$$q\left(t_{i,j}^{L'} \mid t_{i,j}^L\right) \;=\; \frac{1}{U_{\text{ub}} \;-\; U_{\text{lb}}}.$$

After drawing $t_{i,j}^{L'}$, we similarly compute the reverse proposal density $q\left(t_{i,j}^L \mid t_{i,j}^{L'}\right)$ by determining the corresponding interval around $t_{i,j}^{L'}$. These proposal densities are then used in the Metropolis-Hastings acceptance probability.

First, we consider the Bayesian framework without network information. The log Hastings ratio is given by

$$\begin{aligned} H_{1,A} &= \log\left(\frac{P(\Phi', t^{L'}, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})}{P(\Phi', t^L, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})} \times \frac{q(t^L \mid t^{L'})}{q(t^{L'} \mid t^L)}\right) \\ &= \log\left(\frac{P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t_{i,j}^{L'}, \theta, \mu)\, P(t_{i,j}^{L'})}{P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t_{i,j}^{L}, \theta, \mu)\, P(t_{i,j}^{L})} \times \frac{q(t_{i,j}^L \mid t_{i,j}^{L'})}{q(t_{i,j}^{L'} \mid t_{i,j}^{L})}\right) \\ &= \log P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t_{i,j}^{L'}, \theta, \mu) + \log P(t_{i,j}^{L'}) + \log q(t_{i,j}^L \mid t_{i,j}^{L'}) \\ &\quad - \log P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t_{i,j}^{L}, \theta, \mu) - \log P(t_{i,j}^{L}) - \log q(t_{i,j}^{L'} \mid t_{i,j}^{L}). \end{aligned}$$

For the Bayesian framework with network information, the log Hastings ratio is given by

$$
\begin{aligned}
H_{1,B} &= \log\left(\frac{P(\Phi', t^{L'}, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})}{P(\Phi', t^{L}, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})} \times \frac{q(t^{L} \mid t^{L'})}{q(t^{L'} \mid t^{L})}\right) \\
&= \log\left(\frac{P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^{L'}_{i,j}, \theta, \mu)\, P(t^{L'}_{i,j})}{P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^{L}_{i,j}, \theta, \mu)\, P(t^{L}_{i,j})} \times \frac{q(t^{L}_{i,j} \mid t^{L'}_{i,j})}{q(t^{L'}_{i,j} \mid t^{L}_{i,j})}\right) \\
&= \log P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^{L'}_{i,j}, \theta, \mu) + \log P(t^{L'}_{i,j}) + \log q(t^{L}_{i,j} \mid t^{L'}_{i,j}) \\
&\quad - \log P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^{L}_{i,j}, \theta, \mu) - \log P(t^{L}_{i,j}) - \log q(t^{L'}_{i,j} \mid t^{L}_{i,j}).
\end{aligned}
$$

After finishing the updating procedure for the latent period for transmission event $j \to i$, we will now compare the likelihood of $j \to i$ versus the $J_i \to i$ where J_i is the inferred infection ID (transmitter) for the individual i.

Under the Bayesian framework without network data, the log Hastings ratio is given by

$$
\begin{aligned}
H_{1,A} &= \log\left(\frac{P(\Phi', t^{L'}, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^{L}, \theta, \alpha, \beta, \mu \mid D, T^O, t^{\text{inf}})} \times \frac{q(\Phi \mid \Phi')}{q(\Phi' \mid \Phi)}\right) \\
&= \log\left(\frac{P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^{L}_{i,j}, \theta, \mu)\, P(t^{L}_{i,j})}{P(d_i, d_{J_i} \mid T^O, t^{\textbf{inf}}, \Phi, t^{L}_{i,J_i}, \theta, \mu)\, P(t^{L}_{i,J_i})} \times \frac{q(\phi_i \mid \phi'_i)}{q(\phi'_i \mid \phi_i)}\right) \\
&= \log P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^{L}_{i,j}, \theta, \mu) + \log q(\phi_i \mid \phi'_i) + \log q(t^{L}_{i,j} \mid t^{L'}_{i,j}) \\
&\quad - \log P(d_i, d_{J_i} \mid T^O, t^{\textbf{inf}}, \Phi, t^{L}_{i,J_i}, \theta, \mu) - \log P(t^{L}_{i,J_i}) - \log q(\phi'_i \mid \phi_i).
\end{aligned}
$$

Under the Bayesian framework with network data, the log Hastings ratio becomes

$$
\begin{aligned}
H_{1,B} = \log\Bigg(& \frac{P(\Phi', t^{L'}, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})}{P(\Phi, t^L, \theta, \alpha, \beta, \mu, G \mid D, T^O, t^{\text{inf}})} \times \frac{q(\Phi \mid \Phi')}{q(\Phi' \mid \Phi)} \Bigg) \\
& \log\Bigg(\frac{P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^L_{i,j}, \theta, \mu)\, P(\phi_i \mid G)\, P(t^L_{i,j})}{P(d_i, d_{J_i} \mid T^O, t^{\textbf{inf}}, \Phi, t^L_{i,J_i}, \theta, \mu)\, P(\phi'_i \mid G)\, P(t^L_{i,J_i})} \\
& \quad \times \frac{q(\phi_i \mid \phi'_i)}{q(\phi'_i \mid \phi_i)} \Bigg) \\
& = \log P(d_i, d_j \mid T^O, t^{\textbf{inf}}, \Phi', t^L_{i,j}, \theta, \mu) + P(\phi_i \mid G) \\
& \quad + \log q(\phi_i \mid \phi'_i) + \log q(t^L_{i,j} \mid t^{L'}_{i,j}) \\
& \quad - \log P(d_i, d_{J_i} \mid T^O, t^{\textbf{inf}}, \Phi, t^L_{i,J_i}, \theta, \mu) + P(\phi_i \mid G) \\
& \quad - \log P(t^L_{i,J_i}) - \log q(\phi'_i \mid \phi_i).
\end{aligned}
$$

In summary, while both $H_{1,A}$ and $H_{1,B}$ share the same structure, the inclusion of network data in $H_{1,B}$ introduces additional terms—specifically, the likelihood components $P(\phi_i \mid G)$ and $P(\phi'_i \mid G)$. These extra terms incorporate the network connectivity information, thereby providing a more detailed account of the transmission dynamics when network data are available. Consequently, $H_{1,B}$ offers a refined evaluation of the proposed transmission event compared to $H_{1,A}$, which does not leverage network information.

2.5.3 MCMC Simulation Setup

In the Markov Chain Monte Carlo (MCMC) setting, we perform a simulation with a default of $N = 100,000$ iterations. The first $20,000$ iterations are discarded as burn-in to allow the chain to converge to its stationary distribution, ensuring that the initial samples do not influence the final results. After this burn-in period, we collect every 100th sample, resulting in 800 measurements for each parameter of interest. This sampling strategy helps mitigate autocorrelation between consecutive samples, ensuring the parameter estimates are based on independent and representative draws from the posterior distribution.

2.6 Hypothesis Testing on Direct Transmissions

Capturing every individual in a full transmission network for infectious diseases is challenging due to limited healthcare access, undetected asymptomatic carriers, and logistical hurdles in testing and data collection. Typically, only a small subset of patients have their pathogen genomes collected for genetic analysis.

Consequently, many transmissions inferred from genetic data do not represent direct transmissions. To address this, we have developed a statistical tool specifically designed to identify direct transmissions. Let $S_{i,\hat{J}_i}$ be the number of SNPs between two pathogen genomes d_i and $d_{\hat{J}_i}$, where $\hat{J}_i$ is the Bayesian estimate of the transmitter J_i.

The probability distribution of $S_{i,\hat{J}_i}$ is Binomial $(N, p_{i,\hat{J}_i})$, where N is the total number of nucleotides in the genome and $p_{i,\hat{J}_i}$ is the probability of a mutation occurring at a site between genomes d_i and $d_{\hat{J}_i}$. The probability $p_{i,\hat{J}_i}$ can be estimated by

$$\hat{p}_{i,\hat{J}_i} = \frac{3}{4} - \frac{3}{8\hat{\theta} + 4} e^{-\hat{\mu}(t_{i,1}+t_{i,2})} \tag{2.17}$$

where $\hat{\theta}$ and $\hat{\mu}$ are the Bayesian estimates of θ and μ. The mean and standard deviation of $S_{i,\hat{J}_i}$ are $N\hat{p}_{i,\hat{J}_i}$ and $\sqrt{N\hat{p}_{i,\hat{J}_i}(1-\hat{p}_{i,\hat{J}_i})}$, respectively. The 95% confidence interval for the number of SNPs can be used to determine whether the observed SNP count is significantly higher than the expected count and subsequently identify direct transmissions.

2.6.1 Impact of Hypothesis Performance

The performance of the hypothesis testing method in identifying direct transmissions depends on two key factors: (1) SNP sparsity and (2) threshold estimation. These factors influence the ability to distinguish direct from indirect transmission pairs, as shown in Figure 2.2, where Panel B represents SNP sparsity and Panel C illustrates threshold estimation.

2.6.2 SNP Sparsity (Figure 2.2, Panel B)

Classification is more reliable when SNP differences between direct and indirect pairs are well-separated (Panel A) but becomes challenging when their distributions overlap (Panel B). To evaluate this effect, we conduct simulations under various parameter settings, varying infection rate (α), removal rate (β), effective population size (θ), mutation rate (μ), and sample size. This approach quantifies the influence of each parameter on SNP distributions and assesses its impact on the power of the hypothesis test, specifically its ability to correctly identify indirect pairs as indirect.

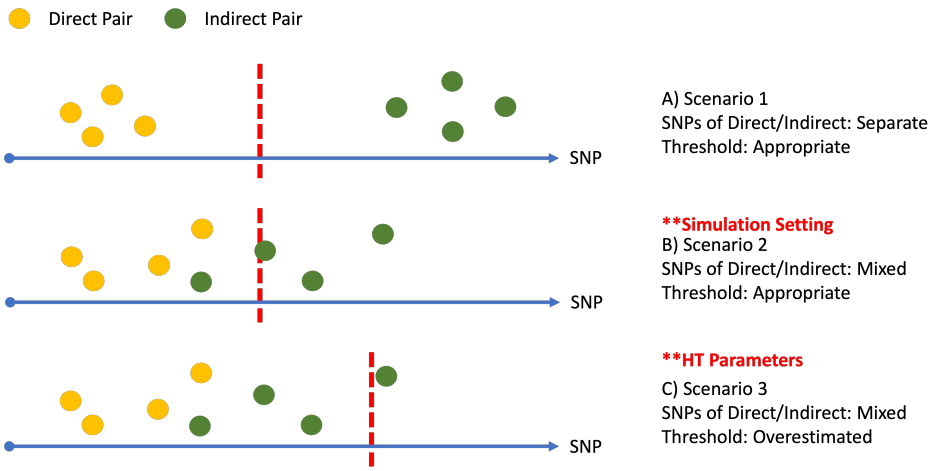


Figure 2.2: Impact of SNP distribution and threshold estimation on classification performance. This figure illustrates three scenarios of SNP distribution between direct and indirect transmission pairs and their effect on classification thresholds:
(A) Scenario 1: Direct and indirect pairs have well-separated SNP distributions, allowing for an appropriate threshold to distinguish between them.**(B) Scenario 2:** Direct and indirect pairs have overlapping SNP distributions, but the threshold remains appropriately placed for classification.
(C) Scenario 3: SNP distributions are mixed, but the threshold is overestimated, leading to potential misclassification of direct and indirect pairs.
The red dashed line represents the classification threshold in each scenario, highlighting the challenges of SNP sparsity and threshold estimation in direct transmission inference.

2.6.3 Threshold Estimation (Figure 2.2, Panel C)

The classification threshold is determined by Bayesian estimates of the mutation rate (μ) and effective population size ((θ)). Panel C illustrates how different estimation strategies influence classification outcomes. The standard Bayesian inference of the transmission tree assumes that all inferred transmission pairs are direct. When the sample size is small, more indirect pairs are included, leading to more observed SNP differences over short evolutionary timescales. This inflates the estimates of θ and μ, causing an overestimation of the mutation rate and effective population size.

To address this bias, we propose an optimization-based approach to iteratively refine the threshold. By explicitly incorporating SNP differences from inferred direct transmission pairs only and optimizing θ and μ, this method cor-

rects for the overestimation caused by indirect transmissions. The optimization process follows the algorithm:

Algorithm 1 Optimization of θ and μ for Transmission Pair Classification

Use Bayesian estimates of θ and μ to establish an initial threshold.
repeat
Optimize θ and μ by minimizing the negative log-likelihood of SNP differences given time for direct pairs.
Define the objective function as:

$$-\log[P(\text{SNP}_1|\text{time}_1, \theta, \mu) \cdot P(\text{SNP}_2|\text{time}_2, \theta, \mu) \cdots]$$

Apply the Nelder-Mead optimization to estimate optimal values θ_2, μ_2.
Update the threshold and reclassify transmission pairs.
until convergence of the classification of pairs

The Nelder-Mead method is a derivative-free optimization algorithm that updates parameter estimates by evaluating function values at simplex vertices. It is particularly effective for likelihood-based optimization when gradients are difficult to compute, making it well-suited for refining model parameters in this study.

By analyzing SNP sparsity (Panel B) and refining threshold estimation (Panel C), we assess the robustness of the hypothesis testing method under different conditions, improving its power to correctly identify indirect transmission pairs.

2.7 Network Analysis: Exponential Random Graph Models

Exponential Random Graph Models (ERGMs) are a statistical framework for modeling complex network structures by estimating the probability of network configurations based on specified characteristics and dependencies. ERGMs are essential for understanding the formation and structure of networks, as they allow researchers to account for both node-level attributes and relational dependencies, such as clustering and reciprocity. The model's probability distribution is expressed as

$$P(Y = y \mid \theta) = \frac{\exp(\theta^T s(y))}{c(\theta)}, \tag{2.18}$$

where θ is a vector of parameters, and $s(y)$ represents sufficient statistics capturing key network features, including dependencies like mutual ties and triadic closures. The term $c(\theta)$ is a normalizing constant. In our context, we can incorporate social distance by including an edge covariate term $\sum_{i,j} X_{ij} y_{ij}$, where X_{ij} denotes the social path length between nodes i and j. This allows the model to evaluate how social distance influences tie formation, with shorter paths potentially indicating stronger social proximity, thus enhancing the probability of a tie. By modeling dependencies and covariates, ERGMs provide a rigorous framework for assessing the role of social distance in network cohesion.

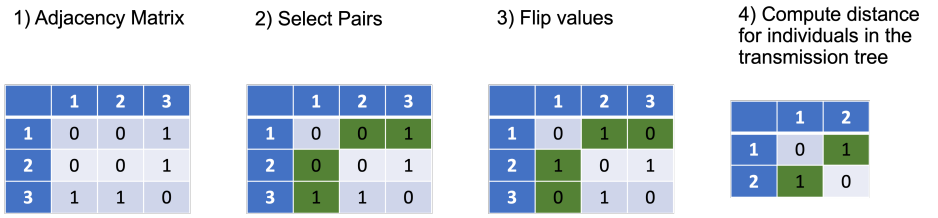
In this study, we evaluate the impact of network distance on the probability of a tie, $P(\text{tie})$, which represents a transmission event in the transmission tree. To achieve this, we fit an Exponential Random Graph Model (ERGM) using a Bayesian-inferred transmission tree and its corresponding network distance matrix.

To assess the robustness of network-based transmission modeling, we introduce noise by modifying a predefined percentage of network connections and analyzing its effect on the relationship between network distance and transmission probability. Specifically, we follow the algorithm:

Algorithm 2 Network Perturbation and ERGM Fitting

Retrieve the adjacency matrix from the previously simulated network of 10,000 nodes.
for each noise level **do**
  for iteration = 1 to 3 **do**
    Select a predefined percentage of node pairs (e.g., 10%).
    for each selected node pair (i, j) **do**
      if adjacency matrix entry $A[i, j] = 1$ **then**
        Set $A[i, j] \leftarrow 0$ (remove tie).
      else if $A[i, j] = 0$ **then**
        Set $A[i, j] \leftarrow 1$ (add tie).
      end if
    end for
    Recompute network distances for individuals in the transmission network (a subset of the full 10,000-node network) based on the modified adjacency matrix.
    Fit the ERGM using the transmission tree, incorporating the updated network distance matrix.
  end for
end for

This process quantifies the sensitivity of ERGM-based transmission modeling to structural noise, providing insights into the stability of network-based dependencies and their influence on transmission dynamics.



	1	2	3
1	0	0	1
2	0	0	1
3	1	1	0

	1	2	3
1	0	0	1
2	0	0	1
3	1	1	0

	1	2	3
1	0	1	0
2	1	0	1
3	0	1	0

	1	2
1	0	1
2	1	0

Figure 2.3: Procedures for introducing noise into the network distance matrix.

Chapter 3

Simulation Study

This chapter presents the Simulation Framework, providing an overview of the methodology used in both network-based and non-network-based simulations. The network-based simulation models transmission dynamics within a structured contact network, while the non-network-based approach simulates transmission independently of network structure. Following this high-level overview, the chapter details the simulation procedures for generating different types of data, including temporal, network, and genetic data, describing the key assumptions, parameter choices, and computational steps involved. The chapter concludes with the data formatting process for Bayesian framework analysis, which includes reordering data by symptom onset time, updating patient IDs, and aligning the network distance and pairwise SNP matrices to ensure proper integration into the Bayesian model for accurate inference and analysis.

3.1 Simulation Framework: the high-level overview

We conducted two types of simulations: network-based and non-network-based, as illustrated in Figure 3.1.

For the network-based simulation, we first generated a network of 10,000 nodes, which provided the underlying structure for disease spread. From this, we constructed a transmission network of 500 individuals, where infection spread followed the infection rate (α) and removal rate (β), as specified in Table 3.1. Rather than being restricted to directly connected individuals, infection probability was determined based on network distance, allowing transmission to occur with varying likelihood depending on proximity within the network. We then simulated pairwise SNP distance matrices based on mutation rate (μ) and effective population size (θ), as listed in Table 3.1, with a genome length of 4.4×10^6 base pairs.

For the non-network-based simulation, although a network was still generated, the transmission network of 500 individuals was modeled independently of the network structure, using only the infection rate (α) and removal rate (β) from Table 3.1. The SNP simulation remained the same as in the network-based approach, using θ and μ from Table 3.1.

We subsequently sampled 100 (20%), 200 (40%), and 400 (80%) infected individuals to compute pairwise SNP distances. Each simulation was repeated three times.

Table 3.1: Parameter combinations for (α, β) and (θ, μ) with corresponding values used in the analysis.

Parameter Combination	**Values**
(α, β)	$(3, 3), (2, 2), (1.5, 1.5)$
(θ, μ)	$(1 \times 10^{-6}, 5 \times 10^{-7})$ $(1 \times 10^{-6}, 1 \times 10^{-6})$ $(1 \times 10^{-6}, 2 \times 10^{-6})$ $(2 \times 10^{-6}, 5 \times 10^{-7})$ $(5 \times 10^{-6}, 5 \times 10^{-7})$ $(5 \times 10^{-6}, 1 \times 10^{-6})$

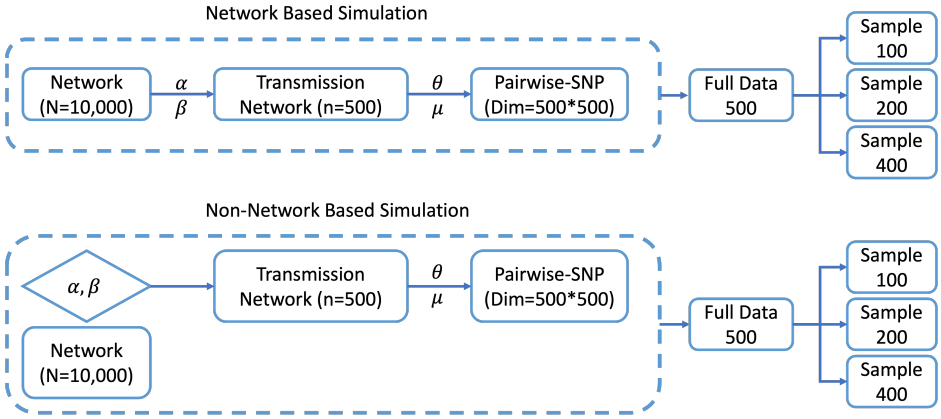


Figure 3.1: Workflow of network-based (top) and non-network-based simulations (bottom).

3.2 Simulation Implementation: the detailed breakdown

3.2.1 Network Data

3.2.1.1 Empirical Data Sources

To gain insights into simulating a network that accurately reflects real-world structures, we will rely on the social network data of $11,840$ individuals, which includes 93 tuberculosis (TB) patients within the population(Miller et al., 2021). This network contains $14,307$ edges, meaning that only 0.02% of all possible pairs have a direct connection, a key parameter that highlights its sparsity and serves as a reference for simulation. Preliminary analysis shows that 29% of the pairs in this network lack a social path, indicating no direct or indirect connections between them. Among pairs with finite social paths, distances are normally distributed, ranging from 1 to 37. The network itself consists of 47 distinct components, with the largest component dominating the structure by encompassing 84% of the total population. These characteristics will guide the design of our simulations, ensuring they realistically capture the observed connectivity patterns, distances, and low tie probability characteristic of this sparse social network.

3.2.1.2 Simulation Scheme

To replicate the observed structural properties in our simulation, we construct a network using a hybrid approach that combines an Erdős-Rényi (ER) random graph as an initial structure with a Barabási-Albert (BA) preferential attachment model. The Erdős-Rényi model generates an initial sparse network of 1,000 nodes, where edges are assigned independently with a low probability $p = 0.002$, ensuring limited initial connectivity. This approach aligns with the sparsity observed in the real-world dataset.

Subsequently, we apply the Barabási-Albert model to expand the network to 10,000 nodes, adding one new edge per incoming node ($m = 1$). The preferential attachment mechanism in the BA model means that new nodes are more likely to connect to highly connected nodes, leading to a degree distribution characterized by a few central hubs with high connectivity and many nodes with relatively few connections. This feature captures the heterogeneous connectivity patterns seen in empirical networks, particularly in social and epidemiological contexts. These models are widely used in network science to

approximate real-world structures (Erdős and Rényi, 1959;Barabási and Albert, 1999).

This approach ensures that the simulated network preserves key properties such as sparsity, path distribution, and component structure, making it a reasonable approximation of the observed social network dynamics.

3.2.2 Temporal Data and Transmission Network

3.2.2.1 Without Network Data

1. **Initial Infection:**
    - Begin with a single infected individual ($i = 1$) at infection time $T_1^I = 0$, serving as the source of infection.
    - Simulate the onset time T_1^O from a scaled chi-squared (χ^2) distribution and the removal time T_1^R from an exponential distribution with rate β.
2. **Secondary Infections:**
    - Determine the number of secondary infections n_1 caused by the initial individual during the infectious period $[T_1^O, T_1^R]$, drawn from a Poisson distribution with mean $\alpha(T_1^R - T_1^O)$.
    - Assign IDs to the new infected individuals: $2, \ldots, 1 + n_1$.
3. **Infection Timing for New Individuals:**
    - For each newly infected individual i, draw the infection time T_i^I uniformly from $[T_1^O, T_1^R]$.
    - Generate the corresponding onset time T_i^O and removal time T_i^R using the same distributions as before.
4. **Iteration:**
    - Set $i = 2$ as the next infection source and repeat from step 2 for all newly infected individuals until all infections surpass a predefined temporal threshold.

Figure 3.2 illustrates this infection process, showing the transition from the latent to the infectious period and the sequential spread of infection. The figure highlights how each infected individual transmits the disease within their infectious period, with red markers indicating transmission events.

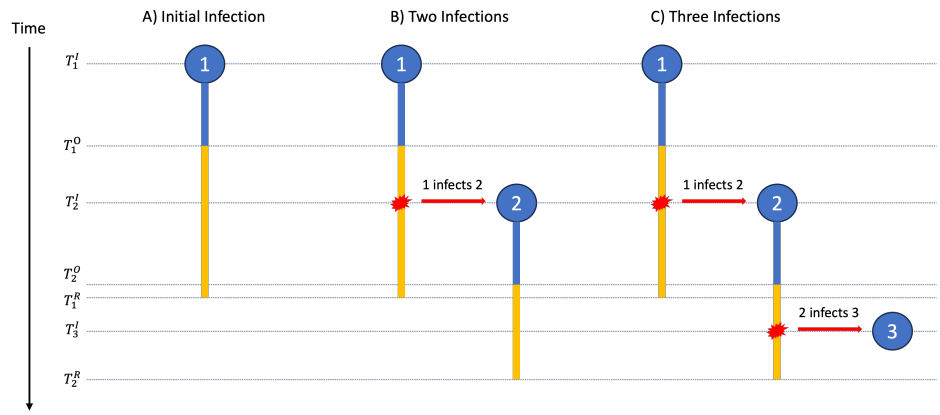


Figure 3.2: Infection Progression Over Time This figure illustrates infection dynamics, with the latent period (blue) followed by the infectious period (yellow). (A) Initial Infection: Patient 1 transitions from latent to infectious. (B) Two Infections: Patient 1 transmits the infection to Patient 2 during their infectious period. (C) Three Infections: Patient 2, now infectious, transmits to Patient 3, continuing the chain. Red markers indicate transmission events occurring within the infectious period of the transmitter.

3.2.2.2 With Network Data

We consider a network Ψ where transmissions occur as follows:

1. **Independent Transmissions:** Infections from an infectious individual a to others are independent events.

2. **Single Infection:** An individual can only be infected once.

3. **Bernoulli Trials:** A transmission event $x_{a,b}$ from a to a susceptible individual b is a Bernoulli trial with a probability mass function:

$$P(x_{a,b}) = p_{a,b}^{x_{a,b}}(1 - p_{a,b})^{1-x_{a,b}}$$

   where:

   - $x_{a,b} = 1$ if the transmission occurs, 0 otherwise.
   - $p_{a,b} = c_{a,b} \cdot k_a$ is the transmission probability.
   - $c_{a,b} = e^{-d_{a,b}}$ is the contact probability, inversely proportional to the exponential of distance $d_{a,b}$ between a and b in the network Ψ.

- $k_a = 1 - e^{-\alpha(T_a^R - T_a^O)}$ is the transmission probability of individual a given contact, where:
  - α is the infection rate (the probability of an infection per time given contact).
  - T_a^O and T_a^R are the onset and removal times of individual a, respectively.

4. **Time to New Infection:** The infection time of individual b, denoted as $t_b^I = T_b^I - T_a^O$, follows an exponential distribution:

$$f(t_b^I = t) = \lambda e^{-\lambda t}.$$

Transmission from a to b is only possible when t_b^I is within a's infectious period $(T_a^R - T_a^O)$. Thus, the time to infection t_b^I given $a \to b$ follows a **truncated exponential distribution** with an upper bound of $T_a^R - T_a^O$.

Given this, the probability of transmission per contact, k_a, is:

$$k_a = P(t_b^I \leq T_a^R - T_a^O) = F_{t_b^I}(T_a^R - T_a^O) = 1 - e^{-\alpha(T_a^R - T_a^O)}.$$

5. **Time to Symptom Removal:** The time to Symptom Removal $t_a^R = T_a^R - T_a^O$ of individual a follows an exponential distribution with a removal rate β.

1. **Initial Infection:**

   - Set the onset time of the first infectious individual $T_1^I = 0$.
   - Simulate the onset time T_1^O from a scaled chi-squared (χ^2) distribution and the removal time T_1^R from an exponential distribution with rate β.
   - The transmission event set is initialized as empty.

2. **Infection Propagation:**

   - Identify the newly infected individual a as the transmission source and determine all infectious-susceptible pairs ($a \to b_1, \ldots, a \to b_{m_1}$) associated with that source. The initial infection serves as the first transmission source.

- Simulate infection events using a Bernoulli trial as described earlier. For each successful transmission (infection event $= 1$), generate the corresponding infection time and removal time based on the defined distributions. Add these pairs to the potential transmission event set.
- Select the transmission event with the earliest infection time from the potential events and add the infected individual to the infected population. The newly infected individual becomes the next transmission source. Repeat this process from Step 1.
- Repeat from step 1 until the simulation reaches a predefined end time or a specific criterion is met (e.g., a certain number of infections).

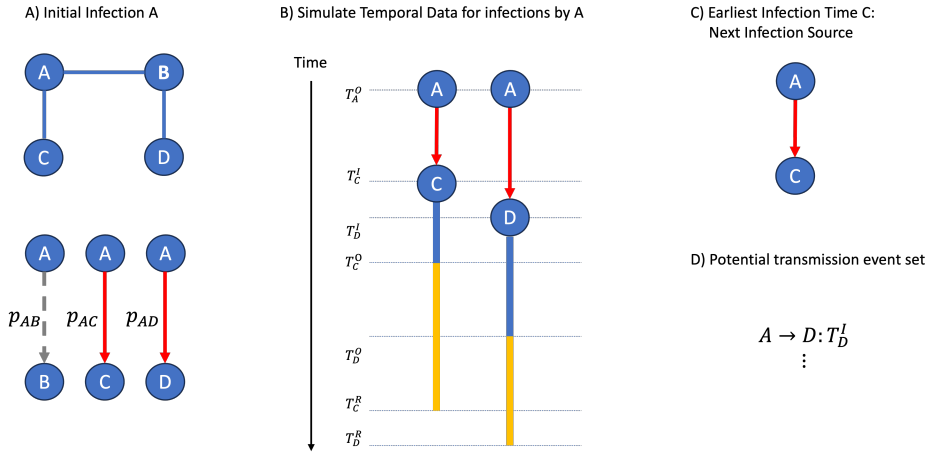


Figure 3.3: Infection Progression Under Network Over Time. (A) Possible secondary infections caused by A within the network (top). Each edge is labeled with the corresponding transmission probability. Successful transmissions are marked in **red**, while unsuccessful attempts are shown in **gray**. (B) Temporal simulation of infections initiated by A, illustrating the latent (blue) and infectious (yellow) periods. (C) The individual with the earliest infection time is selected as the next infection source. (D) Successfully transmitted infections are added to the potential transmission event set.

3.2.3 SNP Data

To simulate the genomic sequences within the transmission network, we first generate a **full genome sequence** for the initial infected patient ($i = 1$) with a length of N, assuming **equal nucleotide base frequencies** ($\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$). This serves as the **reference genome** for subsequent mutations.

For **direct transmission pairs**, we determine the number of **single nucleotide polymorphisms (SNPs)** between two genomes from Binomial$(N, p_{i,j})$, where N is the genome length and $p_{i,j}$ is the probability of a mutation happen on a single site between two pathogen genomes d_i and d_j, where j is the Bayesian estimate of the transmitter of i.

The probability $p_{i,j}$ for the pairs (i, j) involved in the $(n-1)$ transmissions in the transmission network can be calculated using the formula:

$$p_{i,j} = \frac{3}{4} - \frac{3}{8\hat{\theta} + 4} e^{-\hat{\mu}(t_{i,1}+t_{i,2})}$$

where $\hat{\theta}$ and $\hat{\mu}$ are the Bayesian estimates of θ and μ.Each SNP location is randomly assigned, and the mutated nucleotide is selected uniformly from the three possible substitutions. Instead of storing the full genome sequence for every individual, we simulate the complete genome only for the initial infection ($i = 1$). For all subsequent individuals, we record only the loci where mutations occur relative to the reference genome, ensuring efficient storage while preserving genomic variation.

Specifically, an individual's genome sequence is represented by inheriting mutations from its transmitter while accumulating new mutations. As illustrated in Figure 3.4, SNPs are identified by comparing the sampled genome sequences of two patients. For example, if the genome sequence of patient 2, when compared to the reference genome of patient 1, exhibits a mutation at locus 2 (A $\rightarrow$ C), the mutation is recorded as (Locus 2, C). If patient 3 is directly infected by patient 2, its genome sequence is compared to patient 2's sequence. In this case, patient 3 inherits the mutation at locus 2 (C) and accumulates an additional mutation at locus 4 (T $\rightarrow$ A), resulting in a recorded sequence of (Locus 2, C), (Locus 4, A).

After all direct transmission events, the pathogen sequences for all patients are fully recorded based on their inherited and newly acquired mutations. This enables comparisons between individuals who do not share a direct transmission link.

For indirect transmission pairs, the SNP count is determined by identifying all mutation loci present in either sequence. This includes both inherited mutations from intermediate hosts and independently acquired mutations. For example, when comparing individual 3 to reference 1, individual 3 inherits a mutation at locus 2 from individual 2 and acquires an additional mutation at locus 4. Consequently, the total SNP count between individual 3 and reference 1 is 2. By recording only mutation loci rather than full genome sequences, this approach efficiently captures genomic variation while optimizing storage.

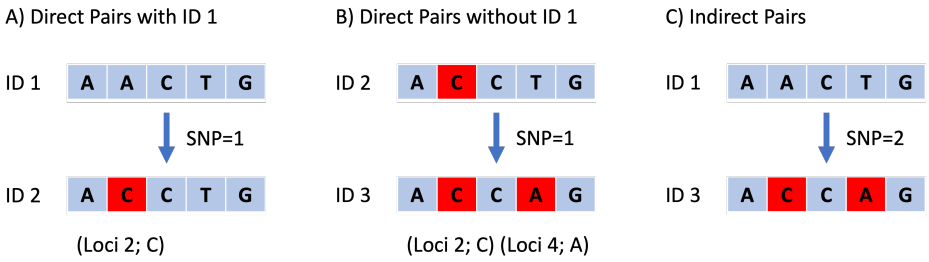


Figure 3.4: SNP Simulation in the Transmission Network. This figure illustrates how single nucleotide polymorphisms (SNPs) are recorded relative to the reference genome. (A) Direct pairs with reference ID 1: Mutations are recorded relative to the full genome of ID 1. Here, ID 2 acquires a mutation at locus 2 (A → C), resulting in SNP = 1. (B) Direct pairs without reference ID 1: ID 3, directly infected by ID 2, inherits the mutation at locus 2 (C) and accumulates an additional mutation at locus 4 (T → A). (C) Indirect pairs: ID 3 is compared to ID 1, considering both inherited and new mutations, resulting in SNP = 2.

3.3 Data Preparation

3.3.1 Format Temporal Data

To ensure consistency in the analysis, we will preprocess the temporal data by retaining only the observable information of Symptom Onset Time (T^O) and Removal Time (T^R). Additionally, we will normalize all symptom onset times by setting the initial Symptom Onset Time (T_1^O) as the reference point (0), meaning all subsequent times will be adjusted by subtracting T_1^O. This normalization ensures a consistent timeline for comparative analysis.

3.3.2 Computing Network Distances

For individuals within the transmission network, we will calculate the network distances based on their connectivity. These distances represent the relative proximity between individuals in the network and play a crucial role in understanding infection spread patterns.

3.3.3 Reordering Patients by Symptom Onset Time

Next, we will reorder the patients in ascending order of Symptom Onset Time (T^O). This ensures that individuals are arranged chronologically based on when

symptoms first appear. After reordering, we will update patient IDs accordingly to maintain consistency in data indexing.

3.3.4 Aligning Distance and SNP Matrices

After updating the patient IDs, we will reorder both the distance matrix and the pairwise SNP matrix to reflect the new patient order. This step ensures that all datasets remain synchronized, allowing for a seamless integration of temporal, genetic, and network-based information in the Bayesian model.

By following these steps, we create a well-structured dataset that preserves key temporal and network relationships, ensures a consistent reference timeline, and maintains data integrity for further analysis.

3.4 Basic Reproduction Number (R_0)

The basic reproduction number, denoted as R_0, is a fundamental epidemiological metric that quantifies the transmission potential of an infectious disease. R_0 is defined as the average number of secondary infections generated by one infected individual in a wholly susceptible population. Essentially, it represents the inherent transmissibility of a pathogen in the absence of any interventions or acquired immunity.

In network-based models of disease or information spread, each infected individual transmits the pathogen (or idea) to each neighbor with probability r. Consequently, the basic reproduction number, R_0, which represents the average number of new infections generated by one infected individual in a fully susceptible population, can be expressed as:

$$R_0 \;=\; r\,\frac{\langle m^2\rangle - \langle m\rangle}{\langle m\rangle},$$

where $\langle m\rangle$ is the average degree (i.e., the mean number of connections per node) and $\langle m^2\rangle$ is the mean-square degree. If $R_0 > 1$, on average the infection or information will propagate through the network; if $R_0 < 1$, it will likely die out. This formulation highlights how both the infection probability r and the heterogeneous structure of the network (as captured by $\langle m\rangle$ and $\langle m^2\rangle$) together govern the potential for large-scale transmission(Newman, 2008).

3.4.1 Non-network based Simulation

In our simulation framework without considering network effects, the average number of infections caused by an individual is determined by both the infectious period and the infection rate. A higher infection rate implies that an individual can transmit the infection more rapidly, while a longer infectious period increases the opportunity for transmission. Accordingly, the expected number of infections generated by an individual is given by:

$$\begin{aligned} E(\#\text{ infections}) &= E(\text{infectious period in years}) \times E(\#\text{ infections per year}) \\ &= \frac{1}{\beta} \times \alpha \\ &= \frac{\alpha}{\beta}, \end{aligned} \tag{3.1}$$

where α represents the infection rate (infections per year) and $\frac{1}{\beta}$ is the expected duration of the infectious period (in years).

3.4.2 Network based Simulation

In our setting, transmission probabilities are heterogeneous across all pairs. Instead, we explicitly define the probability of transmission from individual A to individual B as $p_{A,B}$, which depends on both A's infectious period and the distance between A and B (denoted as $g_{A,B}$).

Traditionally, the expected number of infections caused by an individual i is computed as

$$r \cdot (m_i - 1),$$

where m_i is the number of edges (contacts) associated with i, and the term $m_i - 1$ reflects the exclusion of the edge corresponding to the incoming source of infection.

Extending this idea, we generalize the computation by summing the transmission probabilities from A to all other individuals following the details in subsection 3.2.2.2:

$$\sum_{\substack{i=1 \\ i \neq A}}^{n} p_{A,i} = \sum_{\substack{i=1 \\ i \neq A}}^{n} c_{A,i} \cdot k_A = k_A \cdot \sum_{\substack{i=1 \\ i \neq A}}^{n} c_{A,i}.$$

where $C_{A,i}$ is the contact probability between A and i, K_A is the transmission probability of individual A to the other given a contact.

However, since one of these $n-1$ connections represents the incoming source and should not contribute to further transmissions, we scale the sum to account for the fact that only $n-2$ out of the $n-1$ possible edges are available for transmitting the infection. Therefore, the effective expected number of transmissions is given by:

$$\frac{n-2}{n-1} \times \sum_{\substack{i=1 \\ i \neq A}}^{n} p_{A,i}.$$

This scaling factor $\frac{n-2}{n-1}$ ensures that our calculation properly excludes the incoming source of infection while averaging over the remaining potential transmission pathways.

Lastly, in the conventional method, it needs to weigh over individual A's degree m_A. Here we may use the expected number of contacts (degree in contact network) $\sum_{\substack{i=1 \\ i \neq A}}^{n} c_{A,i}$ to approximate, denoted as c_A

$$R_0 = \frac{\sum_{i=1}^{n} c_i \cdot \frac{n-2}{n-1} \cdot k_i \cdot c_i}{\sum_{i=1}^{n} c_i} = \frac{n-2}{n-1} \cdot k_i \cdot \frac{\sum_{i=1}^{n} c_i^2 \cdot k_i}{\sum_{i=1}^{n} c_i}.$$

3.5 Simulation Data Summary

3.5.1 Temporal Data

We examined the temporal distribution of infections in simulations conducted to reach a sample size of 500, comparing scenarios with and without network-based transmission. In the network-based scenario, infections occurred within a relatively brief interval (Min: 3.467, Max: 8.733), with a median of 4.554 and a mean of 5.009. This rapid and concurrent spread, driven by social interactions, led to clusters of closely timed infections.

Conversely, the scenario without network-based transmission exhibited a significantly prolonged infection period (Min: 14.65, Max: 24.99), characterized by a median of 19.93 and a mean of 20.52. This longer duration reflects a sequential pattern of transmission, where each infection event triggered the next. These results underscore the critical influence of network structure on infection dynamics, demonstrating that social networks accelerate disease spread through clustered outbreaks, while their absence results in a slower, incremental progression.

3.5.2 SNP Data

In Figure 3.5, we present violin plots of SNP distributions for direct and indirect pairs under eight distinct parameter configurations involving θ, μ, α, and β. Violin plots provide a comprehensive visualization of each distribution's shape, range, and interquartile range (IQR), thereby facilitating direct comparisons between SNP values for different pair types.

Several key observations emerge from these plots. First, across all parameter settings, indirect pairs display a wider distribution with a higher mean and larger IQR than direct pairs. Second, as θ increases, the range of SNP values expands considerably. For instance, from subfigure A to C to E 3.5, the maximum SNP value among direct pairs grows from 25 to over 60, while the maximum SNP among indirect pairs increases from approximately 150 to 300 and eventually exceeds 500. Finally, there is no single universal threshold that cleanly separates direct from indirect pairs in every scenario, as evidenced by overlapping distributions across all configurations. This finding underscores the limitations of conventional methods that rely on a fixed SNP cutoff and highlights the importance of a hypothesis-testing framework that can adapt to diverse data and complex underlying assumptions.

In Figure 3.6, we compare violin plots of SNP distributions for sample sizes of 500, 400, 200, and 100 under the parameters $\theta = 1 \times 10^{-6}$, $\mu = 5 \times 10^{-7}$, and $(\alpha, \beta) = (3, 3)$. The overall distribution patterns remain consistent across all sample sizes, indicating that the sampling procedure successfully preserves the essential characteristics of the full dataset.

3.6 Data Analysis

3.6.1 Convergence

To evaluate the convergence of the MCMC sampler, we monitored the trace plots of the log posterior probability over the course of the runs. Convergence is indicated by a chain that stabilizes around a relatively constant mean, with only minor fluctuations thereafter. In Figure 3.7, we illustrate four scenarios under different parameter settings, each showing that after the burn-in period, the log posterior values remain stable and oscillate within a narrow band. This pattern confirms that the sampler has adequately converged, allowing us to proceed with reliable parameter inference.

3.6.2 Accuracy of Transmission Trees Φ

Overall Accuracy Individuals can be categorized based on two criteria: the availability of a direct transmitter in the sample and the correctness of the inferred transmission. As shown in Table 3.2, the columns indicate whether the direct transmitter is present in the sample, and the rows indicate whether the inferred transmission correctly identifies the true transmitter. Based on this categorization, we define the overall accuracy as:

$$\frac{n_1 + n_2}{n - 1}$$

where $n - 1$ excludes the initial infection. This accuracy measure accounts for correctly identifying the direct transmitter when available and the most recent ancestor when the direct transmitter is missing.

Table 3.2: Summary of Transmission Inference Outcomes.

Transmission Inference	**Direct Transmitter Availability**	
	Yes	**No**
Correct	n_1	n_2
Incorrect	n_3	n_4

Table 3.8 presents the average overall accuracy under various parameter settings when genome length is set at 1×10^6, grouped by both the simulation type (with or without network data) and the Bayesian model (with or without network data). When network data are included in the simulation, the Bayesian model that also utilizes network information achieves a 5% higher accuracy than its counterpart without network data, aligning with the expectation that sufficient information improves performance. In contrast, for simulations without network data, relying on a network-based model leads to a 15% decrease in accuracy, reflecting the adverse impact of misleading network information that does not match the underlying simulation assumptions.

When the genome length increases to 4.4×10^6 (Table 3.9), both Bayesian models converge to the same average overall accuracy in simulations with network data. This result indicates that sufficiently large genomic information can, by itself, reliably identify the true transmitter. Nevertheless, in simulations lacking network data, the model incorporating network information remains 16% less accurate, underscoring the persistent negative impact of relying on irrelevant or misleading network assumptions.

In conclusion, when ample genetic information is available (genome length 4.4×10^6), the network-free model demonstrates robust performance across

diverse simulation scenarios, regardless of the inclusion of social network data. Moreover, overall accuracy tends to decline as the sample size decreases. This decline arises because smaller samples contain relatively fewer direct transmission pairs; for individuals without a direct transmitter, correctness is determined by identifying the most recent ancestor. In such cases, the existence of multiple potential transmitters with similar SNP distances and closer temporal proximity increases the risk of misidentification.

Accuracy over Direct Pairs More importantly, we focus on the accuracy among direct transmission pairs. As shown in Table 3.2, this accuracy is defined as the ratio of correct direct transmission inferences (n_1) to all cases where a direct transmitter is available (both correct, n_1, and incorrect, n_3):

$$\frac{n_1}{n_1 + n_3}.$$

This metric quantifies the performance of our inference method in correctly identifying the true direct transmitter when it is present.

Table 3.10 presents the average accuracy over direct pairs for a genome length of 1×10^6, while Table 3.11 shows the corresponding results for a genome length of 4.4×10^6. The observed pattern is consistent with the overall accuracy trends. When genetic information is abundant, both Bayesian models yield comparable performance, achieving 100% accuracy. In contrast, the network-based Bayesian model consistently underperforms in scenarios that do not incorporate network data, thereby reinforcing the robust performance of the network-free model. Given that real-world data typically feature a genome length of 4.4×10^6, our results favor the use of the network-free model.

Furthermore, we do not observe a decline in accuracy as the sample size decreases. This suggests that the adverse impact of smaller sample sizes is primarily limited to cases without a direct transmitter—where correctness depends on identifying the most recent ancestor—since multiple potential transmitters with similar SNP distances and temporal proximity may increase the risk of misidentification.

3.6.3 Posterior Estimates θ, μ, α

We will examine the Bayesian parameters, including θ, μ, and α, by merging posterior samples obtained after discarding the burn-in phase across three independent runs to derive the 95% credible intervals. Our goal is to understand how parameter estimates vary across different sample sizes, simulation schemes, and Bayesian model structures. Intuitively, we anticipate that smaller sample

sizes will yield reduced parameter precision, characterized by increased variance and potential bias. This bias arises partly because smaller samples often contain a greater proportion of indirectly linked transmission pairs, resulting in extreme SNP distances that can lead to the overestimation of parameters such as θ and μ. Here, we aim to provide a high-level summary to guide future hypothesis testing that critically depends on accurate posterior estimation of these parameters.

We observe consistent trends in the posterior distributions for parameters θ (Figure 3.12) and μ (Figure 3.13) across simulation scenarios, highlighting how estimation accuracy is influenced by sample size and network structure. As sample size decreases, both parameters show increased overestimation accompanied by broader credible intervals, reflecting higher uncertainty. Specifically, network-based simulation scenarios (panels A and B in Figures 3.12 and 3.13) yield relatively robust and unimodal distributions down to a sample size of 200 (40%), whereas scenarios without network structure (panels C and D) exhibit pronounced bimodality or multimodality at the same reduced sample sizes. At the smallest sample size (100; 20%), all scenarios consistently demonstrate multimodal distributions, emphasizing substantial variability due to indirect transmission pairs and reduced direct linkage data.

In summary, these findings highlight the critical importance of accounting for potential biases and increased variability in Bayesian parameter estimation at smaller sample sizes. The observed overestimation and multimodal posterior distributions for both θ and μ emphasize the need for correction strategies or methodological adjustments when conducting hypothesis testing under limited data conditions. Such adjustments are essential to ensure robust inference, particularly in contexts where indirect transmission pathways and limited direct-linkage data significantly influence parameter estimates.

The posterior distributions for parameter α (Figure 3.14) show relatively consistent estimates across various sample sizes compared to the distributions of θ and μ. Unlike the substantial instability observed for θ and μ, α maintains greater stability, suggesting that the estimation of α is less sensitive to reductions in sample size. Notably, differences in the posterior means of α across scenarios (panels A–D) are likely attributable to the variability inherent in the simulation procedures themselves. Specifically, stochastic variation in the simulated transmission networks can lead to fluctuations in estimated α, resulting in values that are higher or lower relative to the true parameter setting ($\alpha = 3$). This randomness highlights how the realization of particular transmission chains, rather than structural factors like the presence or absence of network data, may predominantly drive variation in the posterior estimates for α.

In summary, these results underscore the necessity of implementing correction strategies or methodological refinements when using posterior estimates of θ and μ in hypothesis testing, especially under smaller sample sizes. While α estimates remain relatively stable across scenarios, highlighting their robustness to network structure and sample-size limitations, estimates for θ and μ are susceptible to overestimation and substantial variability. Thus, to ensure robust inference, hypothesis-testing procedures must explicitly address biases and multimodality in parameter estimates resulting from indirect transmissions and limited direct linkage data.

3.6.4 Hypothesis Testing

In our study, we test the following hypotheses regarding the inferred transmission pair:

Null Hypothesis (H_0): The inferred pair is direct, meaning the SNP follows a Binomial distribution with genome length N and success probability p (estimated from the posterior).

Alternative Hypothesis (H_1): The inferred pair is indirect, so the SNP does not follow the Binomial distribution with parameters N and p.

We assess performance using two key metrics: False Positive Proportion (FPP) and Sensitivity. FPP quantifies the proportion of direct transmission pairs incorrectly classified as indirect, while Sensitivity measures the proportion of indirect transmission pairs correctly identified as such. Together, these metrics evaluate the hypothesis test's accuracy in distinguishing between direct and indirect transmission events.

Furthermore, we categorize each individual based on (i) the true status of the null hypothesis and (ii) the outcome of the hypothesis test. Table 3.3 summarizes these categories. When H_0 is true, the individual has a direct transmitter in the sample, and Bayesian inference correctly identifies it; thus, the total number of cases where H_0 is true is n_1, consistent with Table **??**.

Based on this categorization, the performance metrics are defined as:

$$\text{False Positive Proportion (FPP)} = \frac{n_{1,R}}{n_1}, \quad \text{Sensitivity} = \frac{n_{1B,R}}{n-1-n_1}.$$

This formulation ensures that FPP correctly represents the proportion of truly direct transmission pairs incorrectly classified as indirect, while Sensitivity accurately captures the test's ability to detect indirect transmissions when present.

Table 3.3: Summary of Hypothesis Testing Outcomes.

Hypothesis Testing	**Null Hypothesis**	
	True	**False**
Accept H_0	$n_{1,A}$	$n_{1B,A}$
Reject H_0	$n_{1,R}$	$n_{1B,R}$
	n_1	$n-1-n_1$

Sensitivity and False Positive Proportion We examine the average false positive proportion (FPP) and sensitivity across various parameter settings. Overall, FPP is well controlled, ranging from 0% to 2% across both simulation schemes and Bayesian models (see Figure 3.16). In contrast, sensitivity decreases as sample size diminishes—falling from 68% to 48% for network-based simulations and from 77% to 65% for network-free simulations. This trend is expected because smaller samples contain a higher proportion of indirect transmission pairs, which leads to an overestimation of parameters (θ, μ). The resulting higher hypothesis test threshold reduces sensitivity.

To mitigate this issue, we applied a correction to the hypothesis testing procedure using Nelder-Mead optimization of the parameters, as described in the Subsection **??**. As shown in Figure **??**, this correction yields an average increase of 30% in sensitivity across scenarios; for smaller sample sizes, the improvement is even more pronounced (e.g., an increase from 48% to 88% for the network-based model). However, this enhancement in sensitivity comes at the cost of a modest increase in FPP—approximately 9% on average—likely due to a more conservative threshold that results in the rejection of more null hypotheses. These findings provide valuable insight into the trade-offs involved in parameter correction and underscore the need for cautious implementation.

Impacting Factors of Sensitivity In this section, we investigate how key parameters affect statistical sensitivity in our Bayesian framework. Specifically, we focus on effective population size θ, mutation rate μ, and the infection and removal rates α and β. By varying one parameter at a time while holding the others constant, we can isolate each parameter's contribution to inference accuracy.

First, we fix $\theta = 1 \times 10^{-6}$ and $\mu = 5 \times 10^{-7}$. Under these conditions, we examine three paired (α, β) settings—$(3, 3)$, $(2, 2)$, and $(1.5, 1.5)$—which are shown in red, green, and blue bars, respectively. The results are grouped by sample sizes of 100, 200, and 400. As illustrated in Figure 3.17, sensitivity

generally increases with both sample size and infection/removal rates, reaching its peak for $(\alpha, \beta) = (1.5, 1.5)$ at a sample size of 400. In contrast, higher α and β values lead to consistently reduced sensitivity, emphasizing the influence of transmission dynamics on inference accuracy.

Next, we fix $\theta = 1 \times 10^{-6}$ and set $\alpha = \beta = 3$. Under these conditions, we vary μ among three values—5×10^{-7}, 1×10^{-6}, and 2×10^{-6}—depicted by red, green, and blue bars, respectively. Again, the results are grouped by sample sizes of 100, 200, and 400. As shown in Figure 3.18, higher mutation rates yield higher sensitivity, while increasing the sample size further improves performance across all μ values. These patterns highlight the interplay between mutation rate and sample size in shaping the model's effectiveness.

Finally, we fix $\mu = 5 \times 10^{-7}$ and set $\alpha = \beta = 3$. Under these conditions, we vary θ among three values—1×10^{-6}, 2×10^{-6}, and 5×10^{-6}—shown in red, green, and blue bars, respectively. The results, grouped by sample sizes of 100, 200, and 400, are displayed in Figure 3.19. Larger values of θ and greater sample sizes both contribute to higher sensitivity, underscoring the importance of effective population size for accurate inference.

Overall, these experiments demonstrate that each of the parameters α, β, μ, and θ plays a critical role in determining statistical sensitivity. High infection and removal rates can sometimes reduce sensitivity by increasing the complexity of transmission dynamics, whereas a sufficiently large sample size or effective population size (θ) tends to enhance inference accuracy. Mutation rate (μ) also exerts a clear influence, with higher values generally yielding better performance. Taken together, these findings emphasize the importance of carefully tuning epidemiological and genetic parameters to achieve robust Bayesian inference in transmission modeling.

ERGM Analysis with Network Perturbation

We follow the network perturbation algorithm, starting with fitting an Exponential Random Graph Model (ERGM) to the inferred transmission tree under the simulation setting with $\alpha = \beta = 3$, $\theta = 1 \times 10^{-6}$, and $\mu = 5 \times 10^{-7}$. The ERGM includes social distance as an edge covariate to capture its influence on transmission dynamics. The negative edges coefficient (-4.87930) indicates a low baseline probability of forming a tie, meaning transmission links are rare.

Additionally, the negative effect of social distance (-0.06836) implies that for each unit increase in social distance, the probability of a tie **decreases by approximately 6.6% relative to its previous value**. This follows from the relationship:

$$\frac{P(\text{tie}|\text{Distance}+1)}{P(\text{tie}|\text{Distance})} = e^{\beta_1} = e^{-0.06836} \approx 0.934 \tag{3.2}$$

which shows that each unit increase in social distance scales the probability of a tie by 0.934, resulting in a 6.6% reduction per unit increase in distance. As a result, transmission likelihood declines progressively as individuals become more socially distant.

To assess the robustness of the model under uncertainty, we introduce noise by randomly altering 5%, 10%, and 20% of network ties. This involves flipping the adjacency matrix entries for a selected proportion of node pairs, followed by recalculating network distances and refitting the ERGM to evaluate its stability and sensitivity to data perturbations.

Table 3.4: ERGM summary output for the inferred transmission tree using social distance as an edge covariate.

Term	**Estimate**	**Std. Error**	**z-value**	**Pr(>|z|)**
edges	-4.87930	0.14055	-34.715	<1e-04
edgecov.Distance	-0.06836	0.01476	-4.631	<1e-04

Null Deviance	172940 on 124750 degrees of freedom
Residual Deviance	6485 on 124748 degrees of freedom
AIC	6489
BIC	6509

Table 3.5: ERGM results for different levels of network noise (5%, 10%, and 20%). Each noise level is repeated three times.

Metric	**5% Noise**			**10% Noise**			**20% Noise**		
	Run 1	**Run 2**	**Run 3**	**Run 1**	**Run 2**	**Run 3**	**Run 1**	**Run 2**	**Run 3**
Estimate	-0.9	-0.15	-0.10	-0.15	-0.12	-0.07	-0.13	-0.14	-0.001
SE	0.05	0.05	0.05	0.068	0.068	0.070	0.079	0.079	0.079
p-value	0.076	0.004	0.057	0.025	0.085	0.333	0.164	0.065	0.994
BIC	6527	6522	6526	6525	6527	6529	6527	6527	6530

As network noise increases from 5% to 20%, the ERGM results show a growing tendency to classify social distance as an insignificant factor in transmission dynamics in Table 3.5. At 5% noise, the standard error (SE) remains low (0.05), and the covariate is largely significant. However, at 10% noise, SE increases

(0.068), and the p-values show variability, with some runs reaching 0.333, indicating weaker evidence for an effect. At 20% noise, SE continues to rise (0.079), and the p-values become highly inconsistent, with some exceeding 0.99, suggesting that the ERGM frequently fails to detect a significant relationship between social distance and transmission.

These findings validate that as noise increases, ERGM correctly reflects the weakening role of social distance in transmission, as the added perturbations disrupt the original structure of the network. This suggests that under high levels of noise, the observed transmission links become less dependent on social distance, leading to an expected loss of significance in the ERGM results. Thus, the model effectively captures the impact of noise on network structure, confirming its ability to adapt to changes in connectivity patterns.

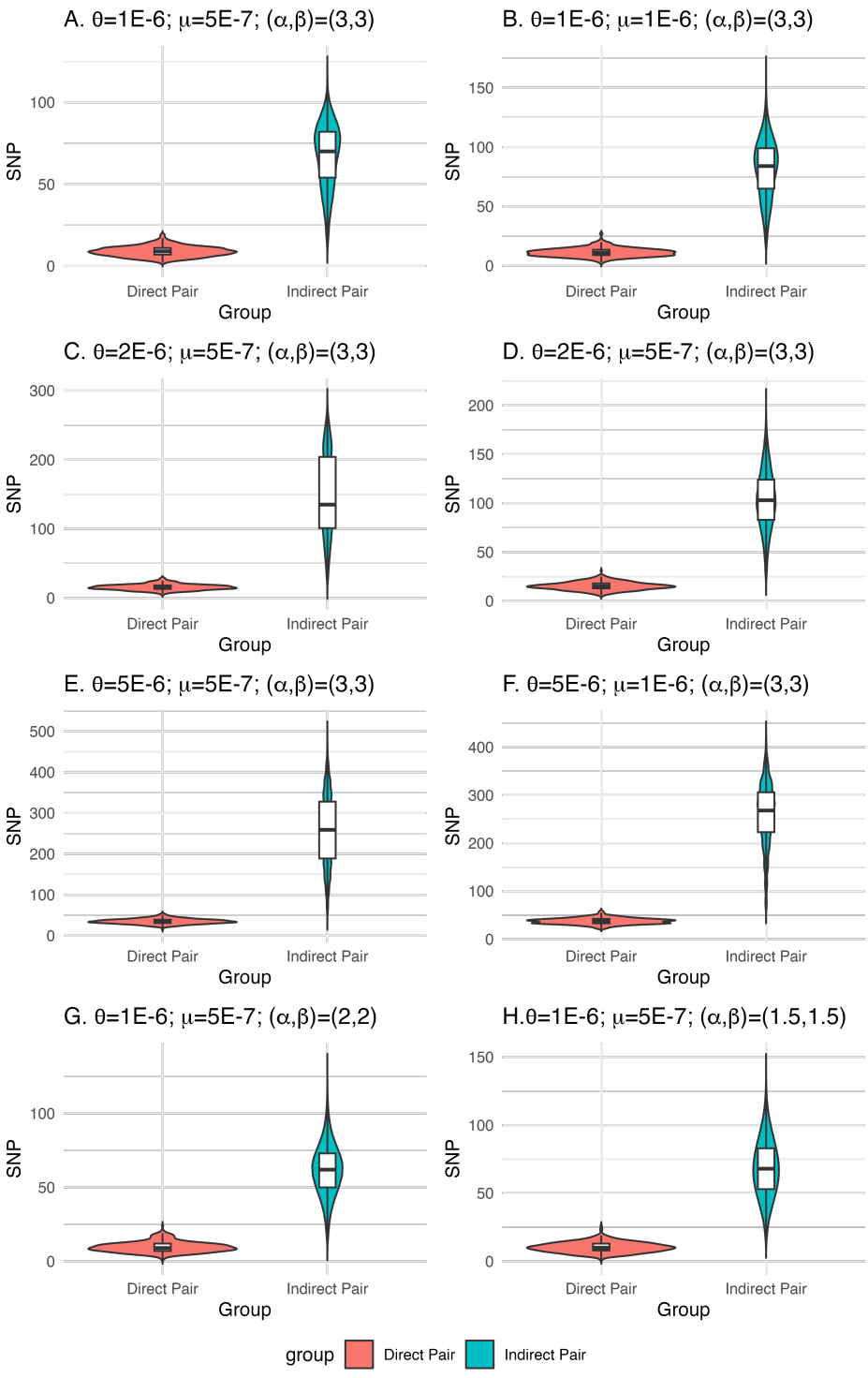


Figure 3.5: Violin plots of SNP distributions for direct and indirect pairs under eight parameter settings. Each subfigure (A–H) corresponds to a unique combination of θ, μ, α, and β.

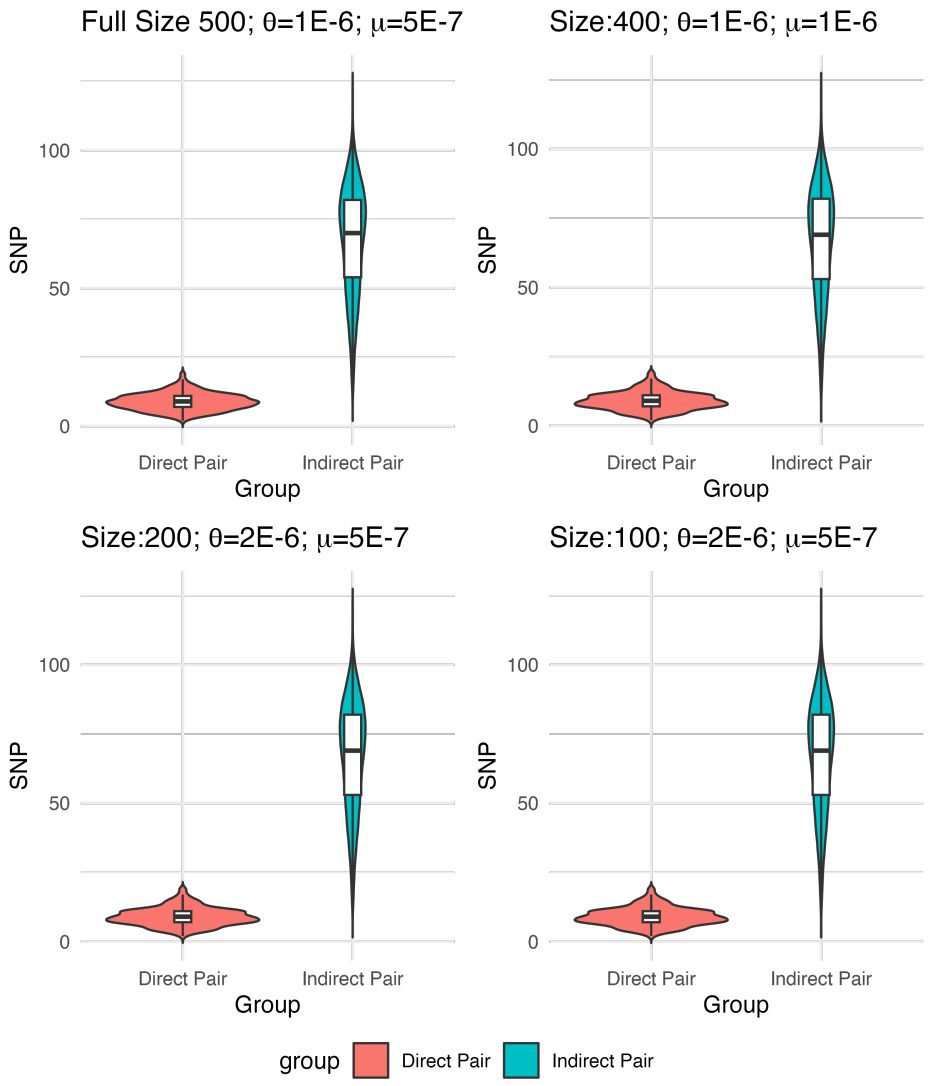


Figure 3.6: Violin plots of SNP distributions for direct and indirect pairs at varying sample sizes. Parameters are set to $\theta = 1 \times 10^{-6}$, $\mu = 5 \times 10^{-7}$, and $(\alpha, \beta) = (3, 3)$.

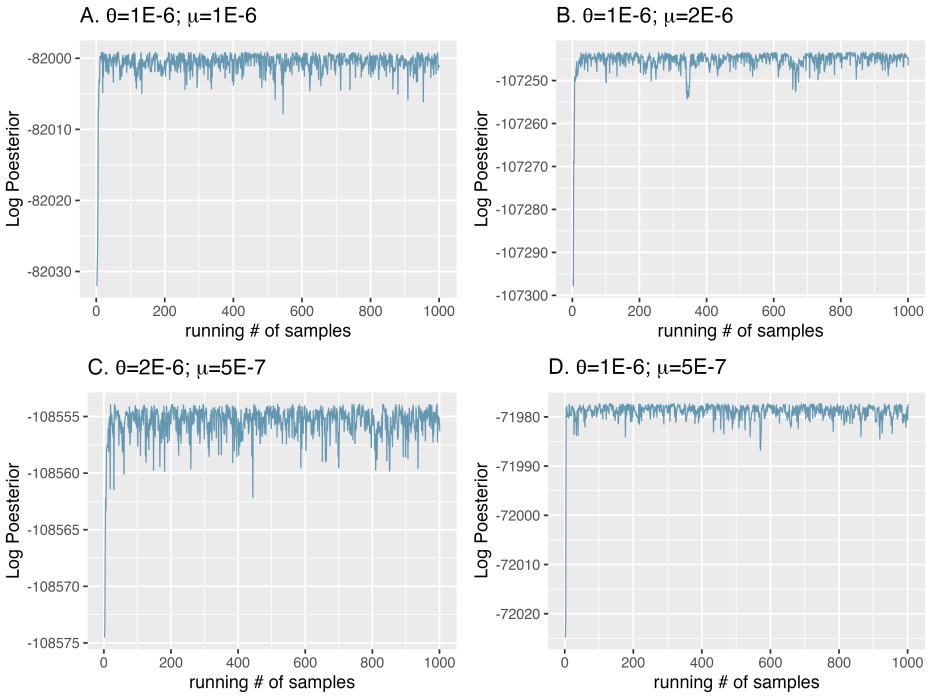


Figure 3.7: Trace plots of the log posterior probability for four different parameter configurations, demonstrating convergence after the burn-in period.

Simulation Type \ Bayesian Framework	Bayesian Model (Without Network Data)			Bayesian Model (With Network Data)		
With Network Data	Insufficient	Full	88%	Sufficient	Full	95%
		400	80%		400	86%
		200	58%		200	62%
		100	42%		100	44%
Without Network Data	Sufficient	Full	93%	Misleading	Full	72%
		400	86%		400	68%
		200	60%		200	46%
		100	38%		100	30%

Figure 3.8: Overall Accuracy under different simulation schemes and Bayesian models with genome length 1×10^6.

Simulation Type \ Bayesian Framework	Bayesian Model (Without Network Data)			Bayesian Model (With Network Data)		
With Network Data	Insufficient	Full	100%	Sufficient	Full	100%
		400	87%		400	88%
		200	57%		200	57%
		100	38%		100	38%
Without Network Data	Sufficient	Full	100%	Misleading	Full	77%
		400	88%		400	66%
		200	54%		200	40%
		100	32%		100	25%

Figure 3.9: Overall Accuracy under different simulation schemes and Bayesian models with genome length 4.4×10^6.

Simulation Schemes \ Bayesian Framework	Bayesian Model (Without Network Data)			Bayesian Model (With Network Data)		
With Network Data	Insufficient	Full	88%	Sufficient	Full	95%
		400	93%		400	97%
		200	94%		200	98%
		100	95%		100	100%
Without Network Data	Sufficient	Full	93%	Misleading	Full	72%
		400	96%		400	75%
		200	98%		200	74%
		100	100%		100	80%

Figure 3.10: Accuracy over direct pairs under different simulation schemes and Bayesian models with genome length 1×10^6.

Simulation Schemes \ Bayesian Framework	Bayesian Model (Without Network Data)			Bayesian Model (With Network Data)		
With Network Data	Insufficient	Full	100%	Sufficient	Full	100%
		400	100%		400	100%
		200	100%		200	100%
		100	99%		100	100%
Without Network Data	Sufficient	Full	100%	Misleading	Full	77%
		400	100%		400	76%
		200	100%		200	74%
		100	100%		100	81%

Figure 3.11: Accuracy over direct pairs under different simulation schemes and Bayesian models with genome length 4.4×10^6.

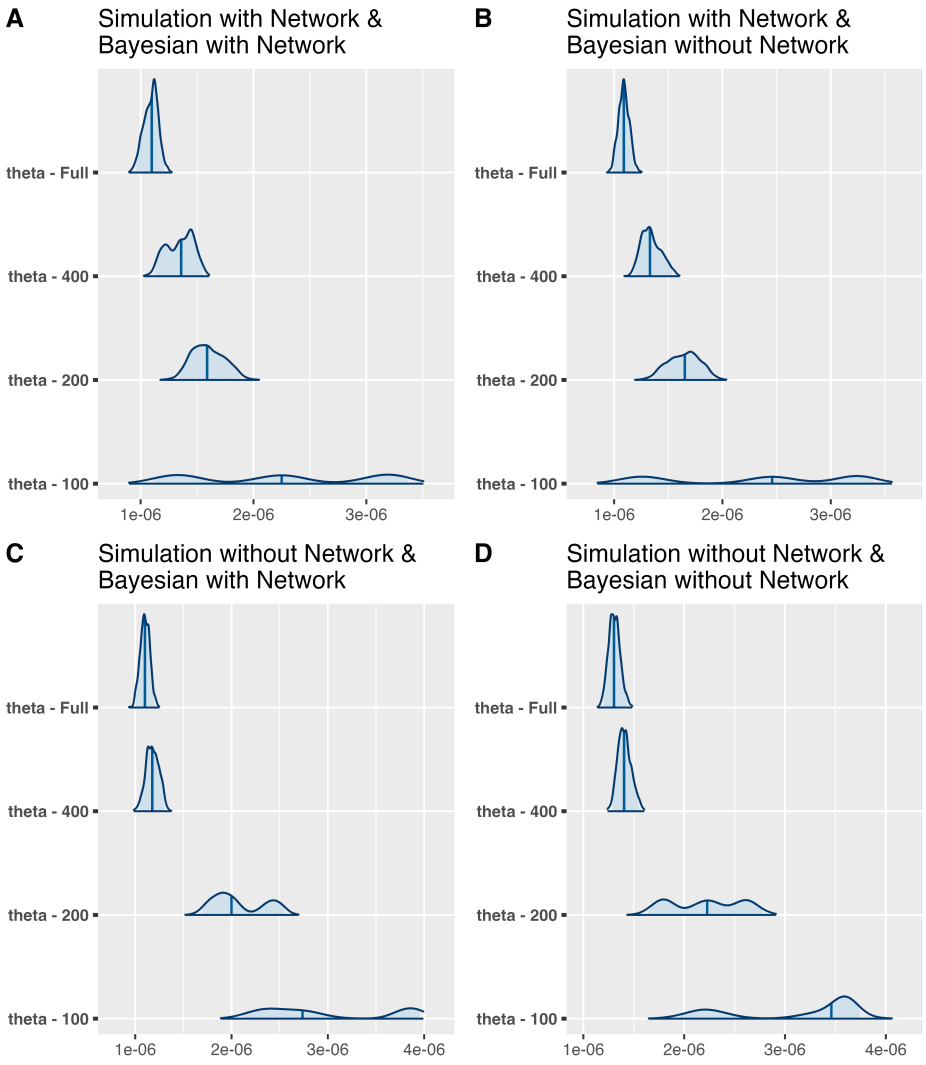


Figure 3.12: Posterior probability distributions of parameter θ across different combinations of simulation and Bayesian inference scenarios, both with and without incorporating network effects.All scenarios employed parameter settings $(\alpha, \beta) = (3, 3)$, $\theta = 1 \times 10^{-6}$, and $\mu = 5 \times 10^{-7}$. Distributions are displayed across decreasing sample sizes (Full dataset, $n = 400$, $n = 200$, and $n = 100$), highlighting the impact of different modeling assumptions and sample limitations on posterior estimates of θ.

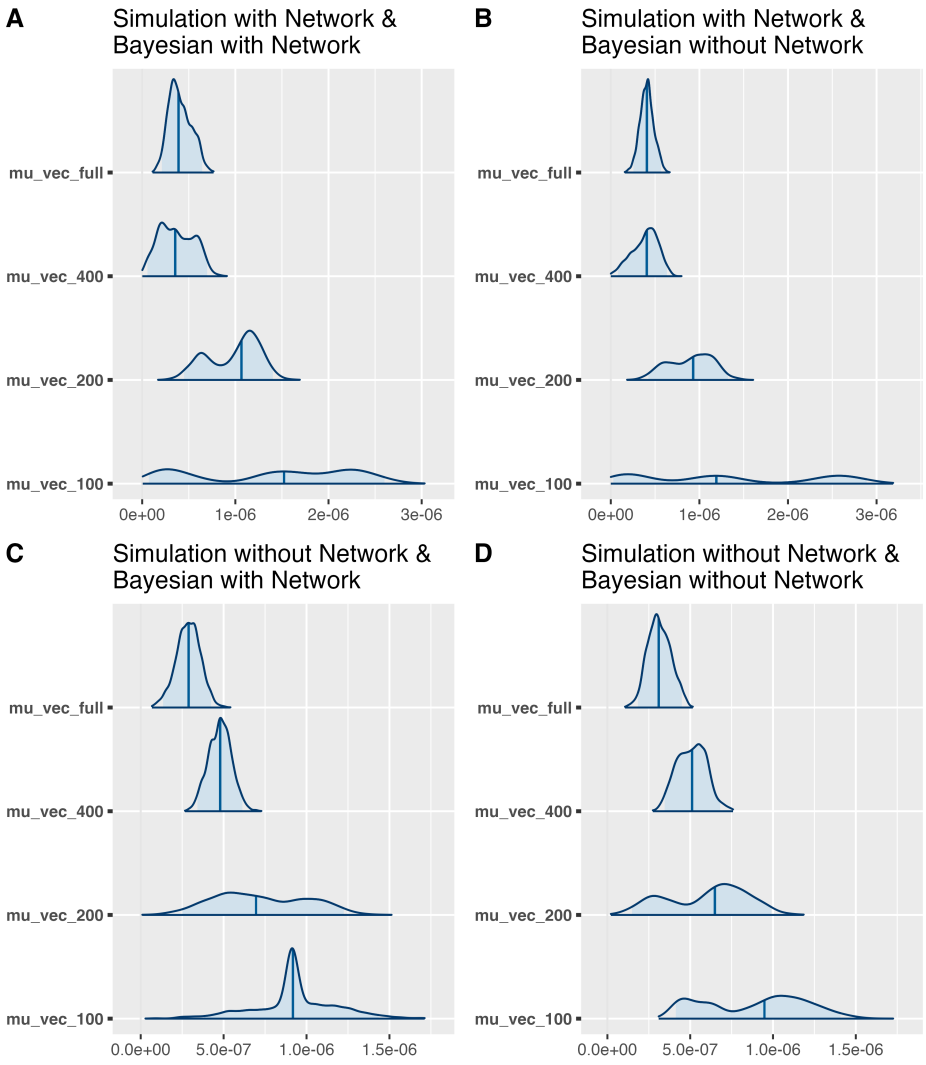


Figure 3.13: Posterior probability distributions of parameter μ across different combinations of simulation and Bayesian inference scenarios, both with and without incorporating network effects.All scenarios employed parameter settings $(\alpha, \beta) = (3, 3)$, $\theta = 1 \times 10^{-6}$, and $\mu = 5 \times 10^{-7}$. Distributions are displayed across decreasing sample sizes (Full dataset, $n = 400$, $n = 200$, and $n = 100$), highlighting the impact of different modeling assumptions and sample limitations on posterior estimates of θ.

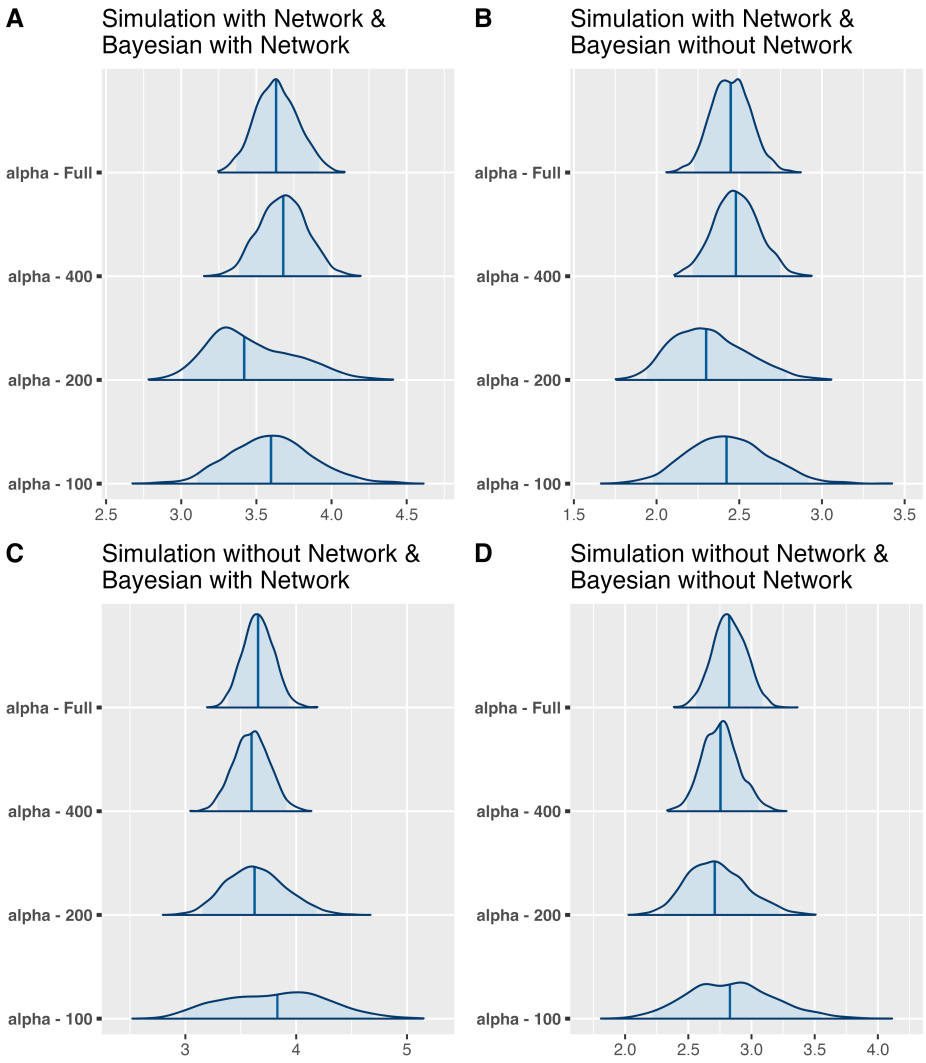


Figure 3.14: Posterior probability distributions of parameter α across different combinations of simulation and Bayesian inference scenarios, both with and without incorporating network effects.All scenarios employed parameter settings $(\alpha, \beta) = (3, 3)$, $\theta = 1 \times 10^{-6}$, and $\mu = 5 \times 10^{-7}$. Distributions are displayed across decreasing sample sizes (Full dataset, $n = 400$, $n = 200$, and $n = 100$), highlighting the impact of different modeling assumptions and sample limitations on posterior estimates of θ.

Simulation Schemes \ Bayesian Framework	Bayesian Model (Without Network Data)			Bayesian Model (With Network Data)		
With Network Data	Insufficient	Full	---	Sufficient	Full	---
		400	68%→83%		400	68%→83%
		200	50%→86%		200	50%→86%
		100	48%→88%		100	48%→87%
Without Network Data	Sufficient	Full	---	Misleading	Full	---
		400	75%→89%		400	78%→93%
		200	55%→92%		200	64%→93%
		100	62%→94%		100	69%→94%

Figure 3.15: Sensitivity Before and After Correction of Hypothesis Testing: genome length 4.4×10^6.

Simulation Schemes \ Bayesian Framework	Bayesian Model (Without Network Data)			Bayesian Model (With Network Data)		
With Network Data	Insufficient	Full	---	Sufficient	Full	---
		400	2%→7%		400	2%→7%
		200	0%→8%		200	0%→8%
		100	1%→10%		100	1%→9%
Without Network Data	Sufficient	Full	---	Misleading	Full	---
		400	2%→11%		400	2%→10%
		200	1%→9%		200	1%→10%
		100	0%→13%		100	0%→15%

Figure 3.16: False Positive Proportion Before and After Correction of Hypothesis Testing: genome length 4.4×10^6.

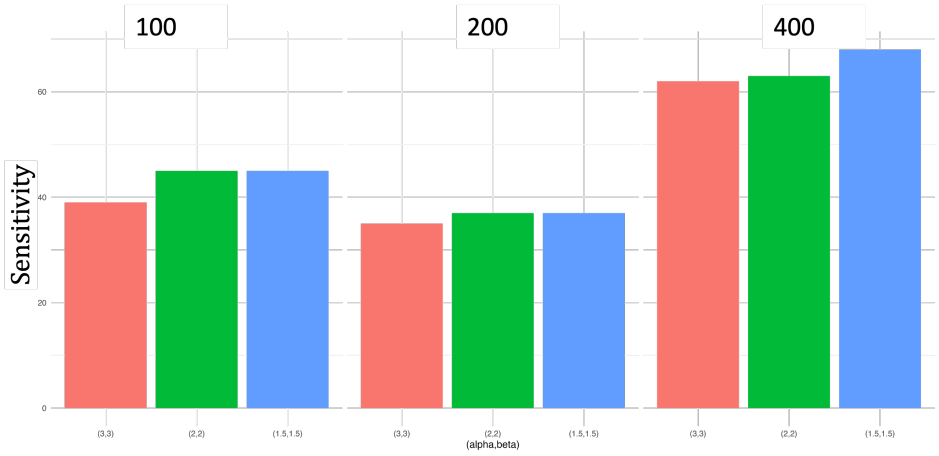


Figure 3.17: Impact of Infection and Removal Rates on Statistical Sensitivity. Average sensitivity is shown for three paired (α, β) settings—$(3, 3)$ (red), $(2, 2)$ (green), and $(1.5, 1.5)$ (blue)—across sample sizes of 100, 200, and 400, with genome length fixed at 4.4×10^6.

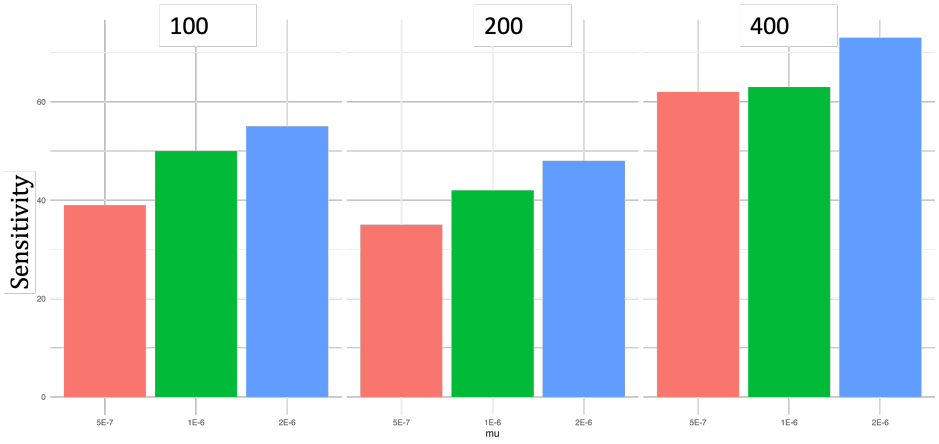


Figure 3.18: Effect of Mutation Rate on Statistical Sensitivity. The plot presents average sensitivity for three mutation rate values—5×10^{-7} (red), 1×10^{-6} (green), and 2×10^{-6} (blue)—grouped by sample sizes of 100, 200, and 400, with $\theta = 1 \times 10^{-6}$ and $\alpha = \beta = 3$.

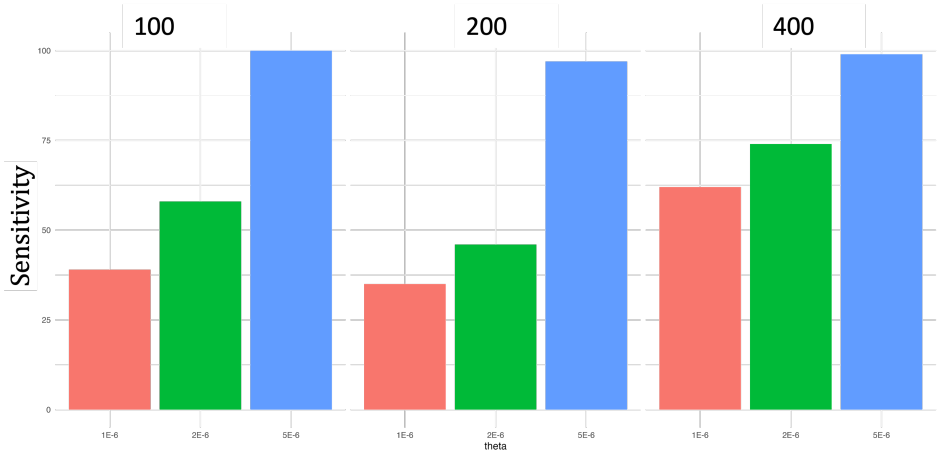


Figure 3.19: Impact of Effective Population Size on Statistical Sensitivity. Average sensitivity is depicted for three effective population sizes—1×10^{-6} (red), 2×10^{-6} (green), and 5×10^{-6} (blue)—across sample sizes of 100, 200, and 400, with $\mu = 5 \times 10^{-7}$ and $\alpha = \beta = 3$.

Chapter 4

Real World Application

4.1 Real-World Data

4.1.1 Study Population and Data Sources

Kakaire et al., 2021 investigated tuberculosis (TB) transmission dynamics in Kampala, Uganda, by analyzing both household and extra-household contacts of TB cases. In their study, 123 TB cases and 124 controls were enrolled; the controls were frequency-matched to the index cases by age group, sex, and parish and were recruited through door-to-door surveys. Whole-genome sequencing (WGS) data — covering a genome of 411,532 base pairs — was obtained along with detailed temporal and geographical information for these samples, and 69 TB patients with complete genomic and temporal data were subsequently selected for analysis using a Bayesian model (Xu et al., 2025). Following the additional acquisition of whole-genome sequencing (WGS) data and refinements in analytical methodologies, the present study now comprises 93 TB patients with comprehensive genomic and temporal data.

In parallel, Miller et al., 2021 constructed a comprehensive social network based on the same set of index participants (123 TB cases and 124 controls) using a two-step egocentric sampling approach. Initially, index participants listed their immediate contacts- household members and individuals with whom they had close, regular interactions. In the second step, these first-level contacts provided the names of their own contacts, thereby generating a second-level egocentric network. By merging overlapping individuals across these two levels, the researchers assembled an extensive socio-centric network comprising 11,840 individuals.

In the present study, we focus on these 93 TB patients—each with complete genomic and temporal data—and leverage the full socio-centric network to derive network-based measures (e.g., network distance) among these 93 cases.

4.1.2 Exploratory Data Analysis

4.1.2.1 Temporal Data

The study, which involved 93 patients, lasted 4.4 years – from the onset of the first patient to the removal of the last.

The infectious periods range from 0.08 to 2.08 years. The median and first quartile are both 0.25 years (approximately 91 days), while the third quartile is 0.332 years (around 120 days), and the mean is 0.37 years (roughly 135 days). The mean exceeding the median indicates a right-skewed distribution driven by a few cases with substantially longer infectious periods. Notably, the first two patients exhibited infectious periods exceeding 2 years (about 760 days), which likely contributed to the extensive transmission observed in the study.

4.1.2.2 Genomic Data

The dataset includes 93 strains with SNP counts ranging from 0 to 1935, a median of 662, and a mean of 731. The interquartile range is 333 to 1167, suggesting a modest right-skew driven by a few high values.

We analyzed genetic distances from two complementary perspectives. First, we generated a density plot of all pairwise comparisons among 93 isolates (a total of $\binom{93}{2}$ comparisons). In this overall analysis, we highlighted a threshold of 20 SNP differences and found that only 50 pairs—less than 1% of all comparisons—fall below this cutoff (as shown by the light blue area in Figure 4.1).

In a second, patient-level analysis, for each patient, we identified the smallest SNP difference with a potential transmitter (defined as an isolate with a smaller ID and an earlier onset time), which yielded 92 values. Applying the same 20 SNP threshold in this context, we observed that 28 of these patient-level comparisons (approximately 30% of the patients) have fewer than 20 SNP differences.

These analyses reveal that only a few strain pairs exhibit the extremely low genetic distances (fewer than 20 SNPs) characteristic of direct transmission. This suggests that many direct transmitters may be missing from the dataset, causing a larger proportion of patients to appear linked through indirect transmission routes when inferring the network.

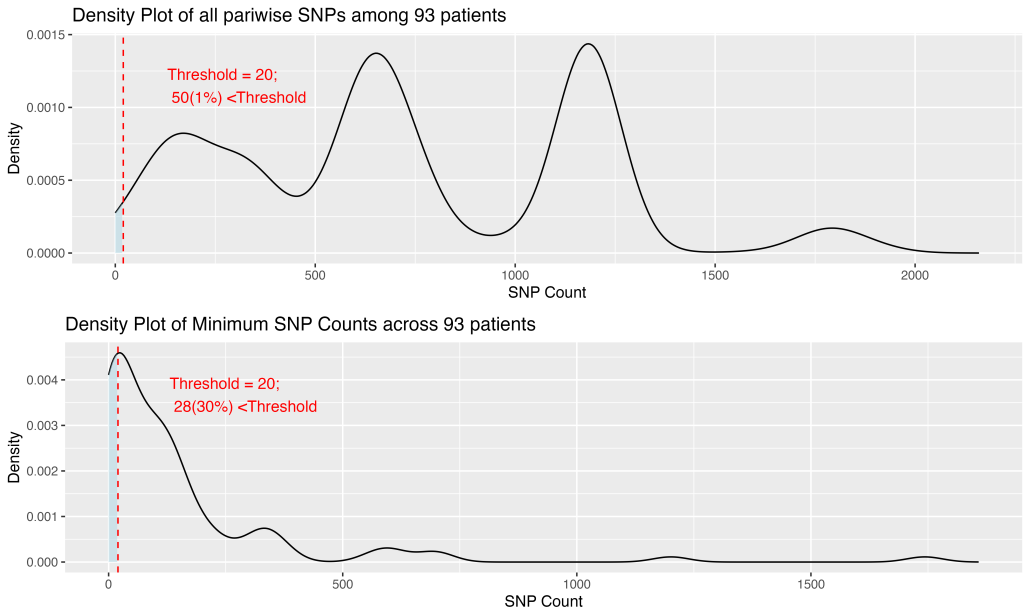


Figure 4.1: Density plots of (top) all pairwise SNP differences among 93 patients and (bottom) the minimum SNP differences per patient. The dashed red line at 20 SNPs highlights the small fraction of observations below this threshold.

4.1.2.3 Network Data

In subsection 3.2.1, we examined the sociometric network of 11,840 individuals and found a giant component comprising 84% of the population, along with 46 smaller components. The probability of a direct connection is only 0.02%, and 29% of all pairs lack any path within the network.

To explore the network distances among the 93 TB patients, we evaluated all $\binom{93}{2} = 4278$ pairs, calculating each pair's shortest path and storing these values in a 93×93 distance matrix. We then summarized the pairwise distance frequencies by assigning the distances into discrete levels and visualized this distribution using a pie chart (Figure 4.2), which shows that 51% of pairs are disconnected and only 3% have distances between 1 and 5—indicating that very few patients are closely connected in the network.

4.2 Results

Based on the robust performance observed for the Bayesian model incorporating both temporal and genomic data across different simulation scenarios, we

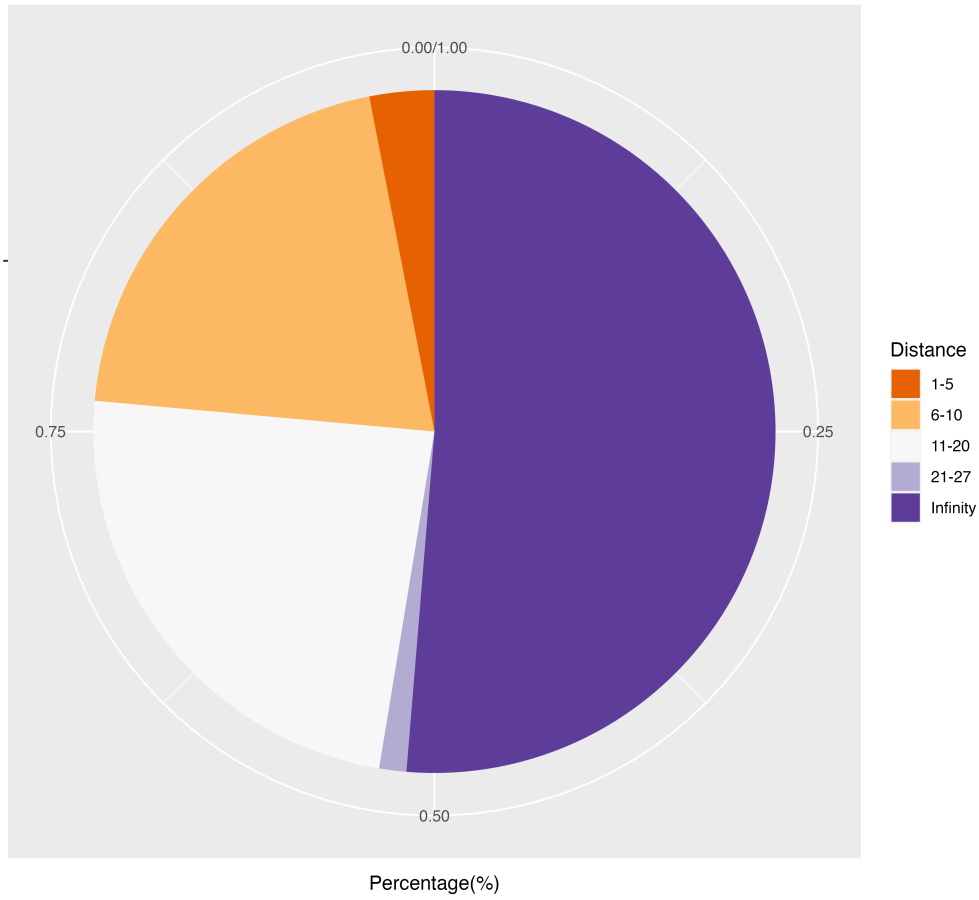


Figure 4.2: Pie chart illustrating the distribution of pairwise network distances among the 93 TB patients.

selected this framework for real-world data analysis. In our implementation, the MCMC algorithm was run for 1,000,000 iterations, including a 20,000-iteration burn-in period. Parameter estimates were recorded every 100 iterations following the burn-in, and convergence was assessed using three independent runs with different initial parameter values. This process was performed for each Bayesian framework, both with and without network data.

4.2.1 Convergence

The trace plot 4.3 of the log-posterior probabilities over 1,000,000 iterations demonstrates that the Markov chain Monte Carlo (MCMC) chains for both Bayesian models reach convergence by the 20,000th iteration, designated as the burn-in period. Discarding these initial samples helps remove transient effects before the chain settles into its stationary distribution. After this burn-in phase, the trace plots for all three independent chains remain stable and show substantial overlap, indicating that the chains have converged to the same posterior region.

We combined the posterior samples after burn-in periods from three independent runs. The posterior mean for the effective population size parameter θ is 8.70×10^{-7} with the 95% credible interval of $[4.45 \times 10^{-7}, 1.28 \times 10^{-6}]$. For the mutation rate, μ, the posterior mean is 2.36×10^{-6} with the 95% credible interval of $[1.62 \times 10^{-6}, 3.24 \times 10^{-6}]$. The posterior mean for the infection rate is 2.74 with the 95% credible interval of $[2.21, 3.33]$.

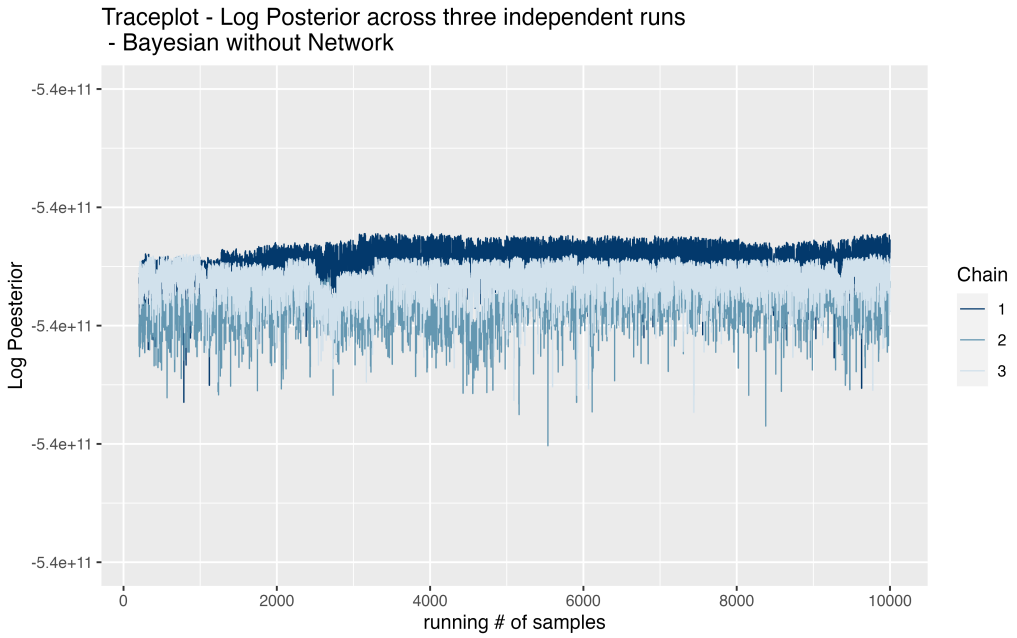


Figure 4.3: The trace plot of log-posterior probabilities for the Bayesian model without network over 1,000,000 iterations where the first 20,000 iterations (burn-in) are omitted.

4.2.2 Hypothesis testing

Hypothesis testing for direct transmissions was performed by comparing the observed number of SNPs to the number expected under the Bayesian model for direct transmission. A 95% confidence interval $[0, u]$ for the number of SNPs was constructed. If the observed SNP count for an edge in the inferred network exceeded or matched the upper bound u of the 95% confidence interval, that edge was classified as an indirect transmission. The hypothesis test identified 16 direct transmissions (Table 4.1).

Under the Bayesian framework without network data, 28 direct transmission pairs were identified across three independent runs. The infection IDs for these 28 patients were identical in every run, indicating consistent inference. Moreover, 26 pairs were classified as direct transmissions in all three runs, underscoring robust and reliable inference for the majority of cases. In contrast, one pair was identified as direct in only two runs and another in just one run, suggesting some variability, particularly in borderline situations. This variability could be attributed to uncertainties in parameter estimation, including parameters like θ and the latent period t^L, which may have a greater impact when the evidence for direct transmission is marginal.

Under the Bayesian framework without network data, a majority vote across three independent runs identified 27 direct transmission pairs—each classified as direct in at least two out of three runs, thereby ensuring a robust consensus—while one additional pair was classified as direct in only a single run.

Network distance among 28 pairs Among the 28 direct transmission pairs identified using the Bayesian transmission tree without network data, 14 exhibited finite network distances. Table 4.2 presents the distribution of pairwise SNP differences for these pairs, categorized into three intervals: 2, 6–10, and 11–15. Notably, 64.3% of the pairs fall within the 6–10 SNP range, suggesting that only a small proportion of infections occur at the neighborhood level (i.e., with a network distance of 2 SNP differences). These genomic findings imply that a significant share of TB transmissions may occur through extra-household contacts rather than solely within immediate neighborhoods.

This inference is supported by broader empirical evidence indicating that tuberculosis transmission predominantly occurs outside the household. For example, Martinez et al., 2017 estimated that less than 20% of transmission is attributable to household exposure, and Verver et al., 2004 found that only 19% of community transmission occurs within households in high-incidence settings. Moreover, studies in African urban settings by Kakaire et al., 2021 and Kiwanuka et al., 2024 underscore the substantial role of extra-household

Table 4.1: Identification of direct transmissions in the transmission network of 69 strains. The Bayesian 95% confidence interval $[0, u]$ for the number of SNPs associated with each of the 68 edges in the estimated network. Sixteen direct transmissions were identified as the observed number of SNPs was less than or equal to the upper bound u.

Case ID	**Case ID (transmitter)**	**# of SNPs**	**Upper bound u**		
			Run 1	**Run 2**	**Run 3**
17	6	8	25	20	25
18	16	1	18	15	15
24	3	1	21	23	19
25	20	27	28	29	30
26	16	2	20	20	15
27	2	1	30	21	42
28	17	10	18	20	19
30	19	4	21	22	20
32	29	3	16	15	15
33	19	0	16	18	15
37	30	0	22	23	19
49	33	0	20	18	15
50	35	4	16	16	14
52	28	1	20	15	16
53	26	2	20	21	19
54	44	0	15	16	16
58	17	13	23	24	19
59	13	1	19	21	25
62	49	3	18	17	17
66	39	15	16	16	18
67	63	17	17	17	16
72	53	10	16	18	17
75	20	31	27	25	33
78	36	8	28	28	27
79	39	17	26	31	25
83	69	0	19	20	20
86	53	13	24	25	27
87	45	20	28	31	27

contacts and broader social networks in driving the TB spread. Collectively, these findings highlight the need for public health interventions that target community-level transmission dynamics.

Table 4.2: Frequency distribution grouped by levels.

Group	Distance	Count	Percentage
2		**1**	**7.1%**
	2	1	7.1%
6–10		**9**	**64.3%**
	6	2	14.3%
	7	1	7.1%
	8	1	7.1%
	9	4	28.6%
	10	1	7.1%
11–15		**4**	**28.6%**
	11	1	7.1%
	13	1	7.1%
	15	2	14.3%

4.2.3 Posterior Probability of Transmission Events

To evaluate the posterior probability of the transmission tree and its corresponding transmission events, we merged the infection ID estimates from three independent runs after the burn-in period, yielding a total of 29,400 estimates for each individual. For each infection, we selected the transmitter with the highest frequency among these estimates and computed its posterior probability by dividing the frequency by 29,400. This probability is taken as the measure of confidence in the inferred transmitter.

Table 4.3 demonstrates that the posterior probabilities are highly concentrated at the lower and upper extremes, with few values observed in the intermediate range. This bimodal distribution suggests that the model differentiates clearly between transmission pairs with minimal support and those with near-certain support, with only a minority of cases exhibiting moderate levels of confidence.

The observed bimodal distribution of posterior probabilities appears to be driven by the underlying genetic distances between individuals and their potential transmitters. In cases where the posterior estimate is low (i.e., <0.1), potential transmitters typically exhibit substantial SNP differences from the individual, indicating a lack of sufficient genetic similarity to support a direct

Table 4.3: Frequency and percentage of posterior estimates for transmission confidence levels.

Posterior Porbability	Count	Percentage
1	36	39%
$[0.8, 1)$	1	1%
$[0.5, 0.8)$	2	2%
$[0.2, 0.5)$	3	3%
$[0.1, 0.2)$	5	5%
$[0, 0.1)$	44	48%

transmission link. To illustrate this phenomenon, Table 4.4 presents the minimum SNP distances among potential transmitters for individuals with posterior estimates below 0.1.

Among individuals with extremely low posterior probabilities (i.e., <0.1), candidate transmitters consistently exhibit substantial SNP differences and unfavorable temporal profiles relative to the focal case. This pronounced lack of genetic and temporal support results in negligible posterior confidence for any direct transmission link. Table 4.4 substantiates this observation by showing that 98% of these individuals have a minimum SNP distance exceeding 50. In addition, 71% of the individuals have a minimum SNP distance exceeding 100, with the maximum observed minimum SNP distance reaching 1734. These findings underscore the significant genetic divergence between individuals and their potential transmitters, thereby reinforcing the extremely low posterior estimates.

In contrast, individuals with a posterior probability of 1 represent cases where the inference has stabilized after the burn-in period, indicating high confidence in the transmission link. For these cases, the genetic data show markedly lower distances between the individual and the potential transmitters. Specifically, among the 36 transmission pairs in this category, 72% exhibit a minimum SNP distance of less than 20, and 89% have a distance below 40.

These findings underscore the bimodal nature of the posterior estimates: one extreme is associated with significant genetic divergence and minimal posterior support, while the other is characterized by strong genetic similarity and robust confidence in the inferred transmission link.

Table 4.4: Minimum SNP distances among potential transmitters for individuals with posterior estimates < 0.1.

Min. SNP Distance	Frequency	Percentage
$[16, 50)$	1	2%
$[50, 100)$	12	27%
$[100, 200)$	16	36%
$[200, 500)$	11	25%
$[500, 1743)$	5	11%

4.2.4 Transmission Tree

Based on the hypothesis testing results and the posterior probabilities from the transmission tree, we constructed a transmission network Figure 4.4. In this visualization, edges are rendered as solid lines when they converge (i.e., posterior probability ≥ 0.8) and as dashed lines when they do not. Furthermore, edge colors differentiate transmission types: red indicates inferred direct transmissions, while black denotes inferred indirect transmissions. Notably, all direct transmission pairs exhibit convergence.

The transmission tree reveals several key findings. For example, Patient 16 emerges as the source of a lineage comprising 10 converged pairs, including 7 direct transmissions, whereas Patient 6 is identified as the source of a lineage with 4 direct transmissions. These individuals are highlighted as high-priority candidates for further investigation.

4.2.5 ERGM

In this analysis, we employ an exponential random graph model (ERGM) to examine how distance within a social network influences the likelihood of a tie forming between two nodes. By incorporating distance as a covariate, each estimated coefficient in the ERGM contributes additively to the log odds of a tie. Once the model is fitted, we transform these log odds via the logistic function to obtain $\Pr(\text{tie} = 1 \mid \text{covariate})$. We then extend this calculation by applying a weighted sum over the distribution of the distance covariate values, thereby deriving $\Pr(\text{tie} = 1, \text{covariate} = \text{specific value})$. This approach allows us to quantify both the conditional probability of tie formation given a particular distance and the overall probability of observing a tie at that distance level, thus elucidating the role of social distance in shaping network structure.

The ERGM results indicate that the baseline propensity for tie formation is captured by the **edges** term, which is both negative and highly significant ($p < 1 \times 10^{-4}$). This negative coefficient reflects the inherent sparsity of the network, in part because our model restricts each individual to a single primary infection source (i.e., coinfection is not permitted), thereby reducing the overall likelihood of tie formation. In contrast, the coefficient for the distance-based covariate (Distance) is positive, suggesting that increased distance is associated with a higher probability of tie formation, although this effect is not statistically significant ($p = 0.313$). Notably, these results are consistent with our preliminary findings: among direct pairs with a distance of ≤ 2, only 1 out of 28 pairs formed a connection, while nearly half remained disconnected. Overall, the model fit—as indicated by a reduction in deviance and acceptable AIC (891.4) and BIC (904.2) values—demonstrates that the ERGM reasonably captures the network structure. These results imply that while network sparsity is robustly captured by the **edges** term, additional covariates may be required to fully explain the dynamics of tie formation.

Building on this robust network characterization, our analysis reveals that most transmission occurs via weak ties—casual contacts between individuals who do not know each other—rather than within closely connected social networks (defined as groups with paths 2). This finding aligns with Granovetter's key insight that weak ties serve as critical bridges connecting disparate social groups, thereby facilitating the flow of information—or, in this context, pathogens—across otherwise isolated clusters Granovetter, 1973.

Table 4.5: ERGM summary output for Model 1.

Term	Estimate	Std. Error	z-value	Pr(>|z|)
edges	-3.933e+00	1.598e-01	-24.615	<1e-04
edgecov.Distance	1.813e-05	1.797e-05	1.009	0.313

Null Deviance	5930.6 on 4278 degrees of freedom
Residual Deviance	887.4 on 4276 degrees of freedom
AIC	891.4
BIC	904.2

Computing Transmission Probabilities by Distance To quantify the likelihood of transmission occurring within a specific distance threshold (e.g., ≤ 2), we follow these steps:

Step 1: Conditional Probability: Using the fitted ERGM, derive the conditional probability of tie formation given a particular distance,

$$Pr(\text{tie} = 1 \mid \text{distance} = g),$$

where g represents discrete distance bins (e.g., $g = 1, 2, \ldots$).

If the edge covariate equals 1, the log odd of a tie is calculated as:

$$\text{log-odds} = \beta_{\text{edges}} + \beta_{\text{edgecov}} \times 1 = -3.933 + 1.813 \times 10^{-5} = -3.93298187.$$

Next, we convert this log-odds value to a probability using the logistic function:

$$\Pr(\text{tie} = 1 \mid \text{covariate} = 1) = \frac{\exp(\text{log-odds})}{1 + \exp(\text{log-odds})} = \frac{\exp(-3.93298187)}{1 + \exp(-3.93298187)}.$$

Numerically, $\exp(-3.93298187) \approx 0.01948$, so

$$\Pr(\text{tie} = 1 \mid \text{covariate} = 1) \approx 0.019 \quad (\text{i.e., } 1.9\%).$$

Step 2: Empirical Frequency: Compute the empirical frequency of each distance value, $\Pr(\text{distance} = g)$, from the observed distance matrix. As shown in Table 4.6, nearly half of all pairs (51.3%) are disconnected, underscoring the inherent sparsity of the network.

Step 3: Joint Probability: Calculate the joint probability of a tie and a specific distance as

$$\Pr(\text{tie} = 1, \text{distance} = g) = \Pr(\text{tie} = 1 \mid \text{distance} = g) \times \Pr(\text{distance} = g),$$

for $g = 1, 2, \ldots, \infty$. In cases where a pair is disconnected, we assign a distance of 11,839, which corresponds to one less than the total number of nodes in the complete social network (11,840).

Step 4: Relative Contribution: Determine the relative contribution of transmission events at a specific distance (e.g., $g = 1$) by computing

$$\frac{\Pr(\text{tie} = 1, \text{distance} = 1)}{\sum_g \Pr(\text{tie} = 1, \text{distance} = g)}.$$

This ratio represents the proportion of transmission events occurring at distance 1 relative to all transmission events across the range of distances.

Table 4.6: Frequency, percentage, $\Pr(\text{tie} \mid \text{distance})$, and $\Pr(\text{tie}, \text{distance})$ for each distance.

Distance	**Frequency**	**Percentage**	$\mathbf{P}(\text{tie} \mid \text{distance})$	$\mathbf{P}(\text{tie}, \text{distance})$
1	8	0.2%	0.01920897	3.59214e-05
2	19	0.4%	0.01920932	8.531486e-05
3	25	0.6%	0.01920966	1.122584e-04
4	28	0.7%	0.01921000	1.257316e-04
5	51	1.2%	0.01921034	2.290153e-04
6	86	2.0%	0.01921068	3.861895e-04
7	126	2.9%	0.01921102	5.658226e-04
8	186	4.3%	0.01921136	8.352767e-04
9	224	5.2%	0.01921171	1.005943e-03
10	255	6.0%	0.01921205	1.145178e-03
11	243	5.7%	0.01921239	1.091307e-03
12	207	4.8%	0.01921273	9.296483e-04
13	173	4.0%	0.01921307	7.769662e-04
14	127	3.0%	0.01921341	5.703842e-04
15	92	2.2%	0.01921376	4.131991e-04
16	62	1.4%	0.01921410	2.784652e-04
17	40	0.9%	0.01921444	1.796582e-04
18	34	0.8%	0.01921478	1.527121e-04
19	27	0.6%	0.01921512	1.212736e-04
20	15	0.4%	0.01921546	6.737540e-05
21	19	0.4%	0.01921581	8.534369e-05
22	19	0.4%	0.01921615	8.534521e-05
23	5	0.1%	0.01921649	2.245967e-05
24	4	0.1%	0.01921683	1.796805e-05
25	6	0.1%	0.01921717	2.695256e-05
26	1	0.0%	0.01921751	4.492173e-06
27	1	0.0%	0.01921786	4.492253e-06
11839	2195	51.3%	0.02369852	1.215948e-02

This structured approach clearly delineates each component of the analysis, ensuring that the methodology is both transparent and accessible.

Here, when we are interested in the accountability of the neighbourhood level (i.e, g<=2) among transmission, we may have

$$
\begin{aligned}
&\frac{\sum_{g=1,2} \Pr(\text{tie} = 1, \text{distance} = g)}{\sum_{g} \Pr(\text{tie} = 1, \text{distance} = g)} \\
=&\frac{3.592141e-05 + 8.531486e-05}{0.02151417} \\
=&0.6\%
\end{aligned}
$$

Or if we would further expand the radius to a distance 5, we may have

$$
\begin{aligned}
&\frac{\sum_{g=1,2,3,4,5} \Pr(\text{tie} = 1, \text{distance} = g)}{\sum_{g} \Pr(\text{tie} = 1, \text{distance} = g)} \\
=&\frac{0.0005882416}{0.02151417} \\
=&2.7\%
\end{aligned}
$$

This structured approach clearly delineates each component of our analysis, ensuring that the methodology is both transparent and accessible. To assess the contribution of transmission events occurring within local neighborhoods (i.e., distances $g \leq 2$), we compute the ratio

$$
\frac{\sum_{g=1}^{2} \Pr(\text{tie} = 1, \text{distance} = g)}{\sum_{g} \Pr(\text{tie} = 1, \text{distance} = g)}.
$$

Substituting the corresponding values yields

$$
\frac{3.592141 \times 10^{-5} + 8.531486 \times 10^{-5}}{0.02151417} \approx 0.6\%.
$$

If we expand the neighborhood radius to include distances up to $g = 5$, the ratio becomes

$$
\frac{\sum_{g=1}^{5} \Pr(\text{tie} = 1, \text{distance} = g)}{\sum_{g} \Pr(\text{tie} = 1, \text{distance} = g)} = \frac{0.0005882416}{0.02151417} \approx 2.7\%.
$$

These results suggest that only a small fraction of transmission events occur within the immediate neighborhood, and this proportion increases slightly when a broader distance threshold is considered.

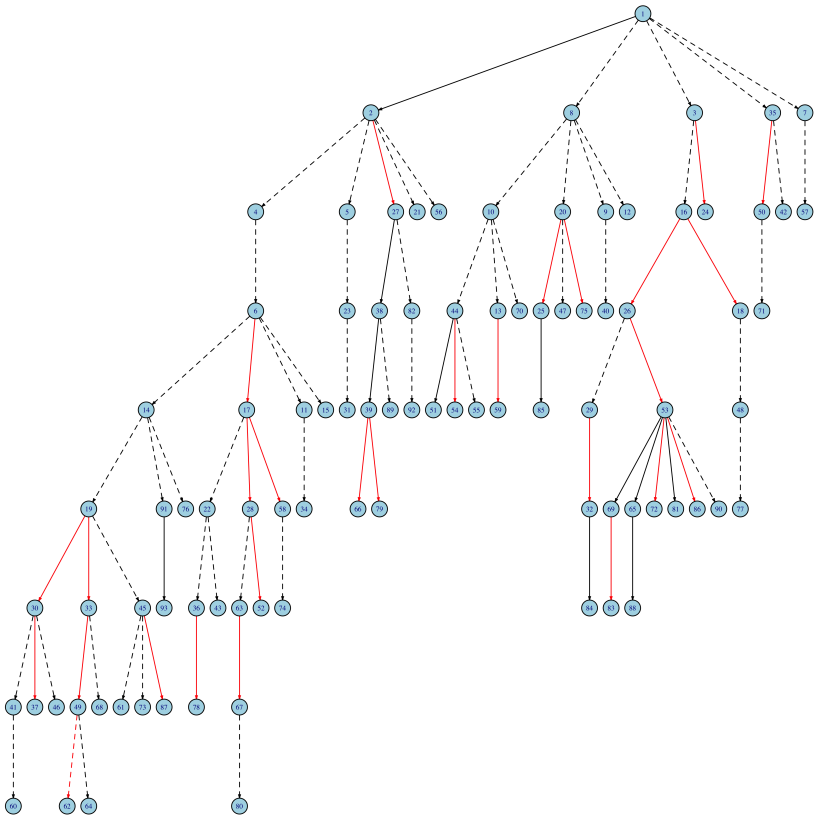


Figure 4.4: The transmission tree of 93 TB patients. Solid black edges denote inferred indirect transmission events with a posterior probability ≥ 0.8, solid red edges denote inferred direct transmission events with a posterior probability ≥ 0.8, and dashed black edges denote inferred indirect transmissions with a posterior probability < 0.8.

Chapter 5

Conclusion

5.1 Summary of Findings

Reconstructing transmission networks is crucial for identifying critical epidemiological factors, such as superspreaders and high-risk locations, which inform effective strategies for pandemic prevention and control. In this dissertation, we propose two Bayesian frameworks designed to reconstruct infectious disease transmission networks by integrating genomic and temporal data, with one framework further incorporating network data. The Bayesian transmission models account for within-host genetic diversity, unobserved infection times, and incomplete sampling of infected individuals. Transmission network inference is accomplished through the estimation of posterior probabilities of transmission events among infected cases. Our Bayesian framework simultaneously integrates sequence data, phylogenetic trees, and the transmission tree. By analytically marginalizing the phylogenetic tree during posterior probability computations, our approach circumvents the computational burdens commonly associated with existing methods.

To evaluate the impact of incorporating network data, we designed two simulation scenarios—one with network information and one without—to assess the performance of the two Bayesian frameworks. Simulation results indicate that the Bayesian transmission model without network data accurately identifies direct transmission pairs, achieving 93% accuracy at a genome length of 10^6 and 100% accuracy at 4.4×10^6. Conversely, the Bayesian model incorporating network data demonstrates only limited improvement when network data are included and substantially decreased accuracy when network data are absent. Thus, the simulations confirm the robustness and reliability of the Bayesian transmission model without network data, especially in scenarios lacking ground truth knowledge.

Hypothesis testing identifies direct transmission events with an average false positive rate of approximately 1% across diverse sample sizes and scenarios. Notably, sensitivity declines from 78% to 50% as the sample size decreases, primarily because smaller samples are more likely to omit direct transmitters, leading to indirect transmitters with larger SNP distances being misclassified as direct. In addition, we observe that applying Nelder–Mead optimization—implemented to correct an overestimated parameter in the hypothesis test—boosts sensitivity by 30% while increasing the false positive rate by 10%. These findings illustrate the inherent trade-off between sensitivity and specificity, particularly in smaller datasets.

Furthermore, we fitted an Exponential Random Graph Model (ERGM) to the inferred transmission tree, incorporating social distance as an edge covariate to capture its impact on tie formation. The model revealed a low baseline transmission probability, with each unit increase in social distance reducing tie probability by approximately 6.6%. When network noise was introduced at levels of 5%, 10%, and 20%, the uncertainty of the social distance effect increased, eventually rendering it statistically insignificant. These results suggest that under higher noise conditions, the impact of social distance on transmission dynamics becomes less detectable.

For the real-world data analysis, we applied a Bayesian model that integrates temporal and genomic data to reconstruct the transmission tree for 93 tuberculosis patients. We assessed edge confidence by designating those with posterior probabilities above 0.8 as converged. In addition, hypothesis testing identified 28 converged direct transmission pairs, with only one (4%) linked to neighborhood-level transmission (social distance 2), confirming the rarity of within-neighborhood infections. We further fitted an exponential random graph model (ERGM) that incorporated social distance as an edge covariate and accounted for network sparsity—partly due to the constraint that each individual has a single primary infection source. Although increased social distance was associated with a higher probability of tie formation, this effect was not statistically significant.

5.2 Limitations and Future Directions

Limited pathogen genome availability significantly impacts the Bayesian transmission model's ability to accurately estimate key parameters, including infection rate, mutation rate (μ), and effective population size (θ). In smaller samples, indirect transmissions are more prevalent, leading the model to overestimate both θ and μ. This overestimation consequently decreases hypothesis-testing

sensitivity. Although our correction procedure improves sensitivity by over 30%, it simultaneously increases the false positive proportion by approximately 10%. Further refinement is therefore required to enhance sensitivity while effectively controlling false positives.

Nevertheless, our simulations demonstrate that, despite these estimation challenges, the model robustly reconstructs transmission networks and reliably identifies direct transmission pairs. This finding confirms that accurate inference of transmission dynamics remains achievable even with limited sample sizes.

Currently, the Bayesian model assumes uniform effective population size across hosts, an assumption that could be relaxed to accommodate host-specific variations. However, allowing such variability would substantially increase parameter complexity and necessitate larger sample sizes for reliable estimation. Under the Jukes-Cantor model, pairwise SNP distances suffice for model fitting. Employing more complex substitution models would allow the Bayesian approach to utilize additional genomic information via sequence-based likelihood functions, thereby enhancing inference accuracy.

Presently, we adopt social network data to inform heterogeneous infection probabilities, yet this method may overlook random encounters between strangers. Our next phase entails integrating social network data with predicted location trajectories obtained from GPS and cell phone data through machine learning. This integration will enable us to accurately infer instances of collocation among individuals and improve estimates of contact likelihood. By uniquely combining personal network data with trajectory predictions, our approach refines the identification of frequent or close interactions and represents the first application of network analysis to tuberculosis transmission in an endemic region. This innovative framework ultimately deepens our understanding of disease spread dynamics. Additionally, adopting non-uniform priors could offer targeted insights into public health interventions, enabling authorities to focus resources effectively on specific social groups or communities.

Bayesian transmission network analysis can be computationally demanding due to the calculation of posterior distributions over extensive parameter spaces, particularly in large datasets. Leveraging parallel computing techniques and optimized software significantly reduces computational time by distributing calculations across multiple processors. Thus, effective computational strategies can substantially enhance the practicality and scalability of Bayesian methods in epidemiological investigations.

Bibliography

Alizon, S., Luciani, F., & Regoes, R. R. (2011). Epidemiological and clinical consequences of within-host evolution. *Trends in microbiology*, *19*(1), 24–32.

Almutiry, W., & Deardon, R. (2021). Contact network uncertainty in individual level models of infectious disease transmission. *Statistical Communications in Infectious Diseases*, *13*(1), 20190012.

Andersen, P., & Doherty, T. M. (2005). The success and failure of bcg—implications for a novel tuberculosis vaccine. *Nature Reviews Microbiology*, *3*(8), 656–662.

Ayabina, D., Ronning, J. O., Alfsnes, K., Debech, N., Brynildsrud, O. B., Arnesen, T., Norheim, G., Mengshoel, A.-T., Rykkvin, R., Dahle, U. R., et al. (2018). Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population. *Microbial genomics*, *4*(10), e000219.

Bagcchi, S. (2023). Who's global tuberculosis report 2022. *The Lancet Microbe*, *4*(1), e20.

Bandoy, D. D. R., & Weimer, B. C. (2021). Analysis of sars-cov-2 genomic epidemiology reveals disease transmission coupled to variant emergence and allelic variation. *Scientific Reports*, *11*(1), 7380.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Bayes, T. (1991). An essay towards solving a problem in the doctrine of chances. 1763. *MD computing: computers in medical practice*, *8*(3), 157–171.

Behr, M. A., Edelstein, P. H., & Ramakrishnan, L. (2018). Revisiting the timetable of tuberculosis. *Bmj*, *362*.

Calmette, A. (1922). *L'infection bacillaire et la tuberculose chez l'homme et chez les animaux* [Available at the Internet Archive and Gallica digital library.]. Masson et Cie. https://archive.org/details/calmettetuberculose

Campbell, F., Cori, A., Ferguson, N., & Jombart, T. (2019). Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology*, *15*(3), e1006930.

Campbell, F., Strang, C., Ferguson, N., Cori, A., & Jombart, T. (2018). When are pathogen genome sequences informative of transmission events? *PLoS pathogens*, *14*(2), e1006885.

Carson, J., Keeling, M., Wyllie, D., Ribeca, P., & Didelot, X. (2024). Inference of infectious disease transmission through a relaxed bottleneck using multiple genomes per host. *Molecular Biology and Evolution*, *41*(1), msad288.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S., Eiglmeier, K., Gas, S., Barry Iii, C., et al. (1998). Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, *396*(6707), 190–190.

Coll, F., Raven, K. E., Knight, G. M., Blane, B., Harrison, E. M., Leek, D., Enoch, D. A., Brown, N. M., Parkhill, J., & Peacock, S. J. (2020). Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant staphylococcus aureus: A genomic epidemiology analysis. *The Lancet Microbe*, *1*(8), e328–e335.

Dawson, D., Rasmussen, D., Peng, X., & Lanzas, C. (2021). Inferring environmental transmission using phylodynamics: A case-study using simulated evolution of an enteric pathogen. *Journal of the Royal Society Interface*, *18*(179), 20210041.

De Maio, N., Worby, C. J., Wilson, D. J., & Stoesser, N. (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS computational biology*, *14*(4), e1006117.

De Maio, N., Wu, C.-H., & Wilson, D. J. (2016). Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology*, *12*(9), e1005130.

Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R., & Wilson, D. J. (2018). Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic acids research*, *46*(22), e134–e134.

Didelot, X., Gardy, J., & Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution*, *31*(7), 1869–1879.

Doege, T. C. (1965). Tuberculosis mortality in the united states, 1900 to 1960. *JAMA*, *192*(12), 1045–1048.

Duault, H., Durand, B., & Canini, L. (2022). Methods combining genomic and epidemiological data in the reconstruction of transmission trees: A systematic review. *Pathogens*, *11*(2), 252.

Erdős, P., & Rényi, A. (1959). On random graphs i. *Publ. math. debrecen*, *6*(290-297), 18.

Ferguson, N. M., Donnelly, C. A., & Anderson, R. M. (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain. *Nature*, *413*(6855), 542–548.

Fonkwo, P. N. (2008). Pricing infectious disease: The economic and health implications of infectious diseases. *EMBO reports*, *9*(S1), S13–S17.

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., Johnston, J. C., Gardy, J., Lipsitch, M., & Fortune, S. M. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics*, *45*(7), 784–790.

Fujikura, Y., Hamamoto, T., Kanayama, A., Kaku, K., Yamagishi, J., & Kawana, A. (2019). Bayesian reconstruction of a vancomycin-resistant enterococcus transmission route using epidemiologic data and genomic variants from whole genome sequencing. *Journal of Hospital Infection*, *103*(4), 395–403.

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine*, *364*(8), 730–739.

Gilbertson, M. L., Fountain-Jones, N. M., & Craft, M. E. (2018). Incorporating genomic methods into contact networks to reveal new insights into animal behaviour and infectious disease dynamics. *Behaviour*, *155*(7-9), 759–791.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360–1380.

Hall, M., Woolhouse, M., & Rambaut, A. (2015). Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLoS computational biology*, *11*(12), e1004613.

Haydon, D. T., Chase–Topping, M., Shaw, D., Matthews, L., Friar, J., Wilesmith, J., & Woolhouse, M. (2003). The construction and analysis of epidemic trees with reference to the 2001 uk foot–and–mouth outbreak. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1511), 121–127.

Hu, H. (2023). *Bayesian inference of transmission networks for infectious disease modeling* [Ph.D. dissertation]. University of Georgia. https://esploro.libs.uga.edu/esploro/outputs/doctoral/Bayesian-Inference-of-Transmission-Networks-for/9949467412202959?institution=01GALI_UGA

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., & Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology*, *10*(1), e1003457.

Kakaire, R., Kiwanuka, N., Zalwango, S., Sekandi, J. N., Quach, T. H. T., Castellanos, M. E., Quinn, F., & Whalen, C. C. (2021). Excess risk of tuberculosis infection among extra-household contacts of tuberculosis cases in an african city. *Clinical Infectious Diseases*, *73*(9), e3438–e3445.

Ke, Z., & Vikalo, H. (2023). Graph-based reconstruction and analysis of disease transmission networks using viral genomic data. *Journal of Computational Biology*, *30*(7), 796–813.

Kendall, M., & Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular biology and evolution*, *33*(10), 2735–2743.

Kiwanuka, N., Zalwango, S., Kakaire, R., Castellanos, M. E., Quach, T. H. T., & Whalen, C. C. (2024). M. tuberculosis infection attributable to exposure in social networks of tuberculosis cases in an urban african community. *Open Forum Infectious Diseases*, *11*(5), ofae200.

Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., & Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology*, *13*(5), e1005495.

KOCH, R. (1882). Die aetionlogie der tuberculose. berlin. klin. *Wchnschur*, *19*, 221–230.

Laplace, P. S. (1774). Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*.

Lau, M. S., Marion, G., Streftaris, G., & Gibson, G. (2015). A systematic bayesian integration of epidemiological and genetic data. *PLoS computational biology*, *11*(11), e1004633.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, *438*(7066), 355–359.

Luke, D. A., & Harris, J. K. (2007). Network analysis in public health: History, methods, and applications. *Annual review of public health*, *28*(1), 69–93.

Martin, M. A., Lee, R. S., Cowley, L. A., Gardy, J. L., & Hanage, W. P. (2018). Within-host mycobacterium tuberculosis diversity and its utility for inferences of transmission. *Microbial Genomics*, *4*(10), e000217.

Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017). Transmission of mycobacterium tuberculosis in households

and the community: A systematic review and meta-analysis. *American journal of epidemiology*, *185*(12), 1327–1339.

Martin-Hughes, R., Vu, L., Cheikh, N., Kelly, S. L., Fraser-Hurt, N., Shubber, Z., Manhiça, I., Mbendera, K., Girma, B., Pambudi, I., et al. (2022). Impacts of covid-19-related service disruptions on tb incidence and deaths in indonesia, kyrgyzstan, malawi, mozambique, and peru: Implications for national tb responses. *PLOS Global Public Health*, *2*(3), e0000219.

Miller, P. B., Zalwango, S., Galiwango, R., Kakaire, R., Sekandi, J., Steinbaum, L., Drake, J. M., Whalen, C. C., & Kiwanuka, N. (2021). Association between tuberculosis in men and social network structure in kampala, uganda. *BMC Infectious Diseases*, *21*, 1–9.

Montazeri, H., Little, S., Mozaffarilegha, M., Beerenwinkel, N., & DeGruttola, V. (2020). Bayesian reconstruction of transmission trees from genetic sequences and uncertain infection times. *Statistical applications in genetics and molecular biology*, *19*(4-6), 20190026.

Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T., & Soubeyrand, S. (2012). A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS computational biology*, *8*(11), e1002768.

Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., Croucher, N. J., Choi, S. Y., Harris, S. R., Lebens, M., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, *477*(7365), 462–465.

Newman, M. E. (2008). The mathematics of networks. *The new palgrave encyclopedia of economics*, *2*(2008), 1–12.

Nübel, U., Dordel, J., Kurt, K., Strommenger, B., Westh, H., Shukla, S. K., Žemličková, H., Leblois, R., Wirth, T., Jombart, T., et al. (2010). A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant staphylococcus aureus. *PLoS pathogens*, *6*(4), e1000855.

Parker, J., Rambaut, A., & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution*, *8*(3), 239–246.

Ratmann, O., Hodcroft, E. B., Pickles, M., Cori, A., Hall, M., Lycett, S., Colijn, C., Dearlove, B., Didelot, X., Frost, S., et al. (2017). Phylogenetic tools for generalized hiv-1 epidemics: Findings from the pangea-hiv methods comparison. *Molecular biology and evolution*, *34*(1), 185–203.

Rodrigues, L. C., & Smith, P. G. (1990). Tuberculosis in developing countries and methods for its control. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *84*(5), 739–744.

Schatz, A., Bugle, E., & Waksman, S. A. (1944). Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. *Proceedings of the Society for Experimental Biology and Medicine*, *55*(1), 66–69.

Seung, K. J., Keshavjee, S., & Rich, M. L. (2015). Multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis. *Cold Spring Harbor perspectives in medicine*, *5*(9), a017863.

Sharma, S., Mohan, A., & Kadhiravan, T. (2005). Hiv-tb co-infection: Epidemiology, diagnosis & management. *Indian Journal of Medical Research*, *121*(4), 550–567.

Skums, P., Mohebbi, F., Tsyvina, V., Baykal, P. I., Nemira, A., Ramachandran, S., & Khudyakov, Y. (2022). Sophie: Viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework. *Cell systems*, *13*(10), 844–856.

Stimson, J., Gardy, J., Mathema, B., Crudu, V., Cohen, T., & Colijn, C. (2019). Beyond the snp threshold: Identifying outbreak clusters using inferred transmissions. *Molecular biology and evolution*, *36*(3), 587–603.

Takayama, K., Wang, L., & David, H. L. (1972). Effect of isoniazid on the in vivo mycolic acid synthesis, cell growth, and viability of mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, *2*(1), 29–35.

Van der Roest, B. R., Bootsma, M. C., Fischer, E. A., Klinkenberg, D., & Kretzschmar, M. E. (2023). A bayesian inference method to estimate transmission trees with multiple introductions; applied to sars-cov-2 in dutch mink farms. *PLoS Computational Biology*, *19*(11), e1010928.

Verver, S., Warren, R. M., Munch, Z., Richardson, M., van der Spuy, G. D., Borgdorff, M. W., Behr, M. A., Beyers, N., & van Helden, P. D. (2004). Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *The Lancet*, *363*(9404), 212–214.

Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., De Angelis, D., Cartwright, E. J., Peacock, S. J., & Cooper, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*, *10*(1), 395.

Xu, J., Hu, H., Ellison, G., Yu, L., Whalen, C. C., & Liu, L. (2025). Bayesian estimation of transmission networks for infectious diseases. *Journal of Mathematical Biology*, *90*(3), 1–19.

Yang, C., Lu, L., Warren, J. L., Wu, J., Jiang, Q., Zuo, T., Gan, M., Liu, M., Liu, Q., DeRiemer, K., et al. (2018). Internal migration and transmission dynamics of tuberculosis in shanghai, china: An epidemiological, spatial, genomic analysis. *The Lancet Infectious Diseases*, *18*(7), 788–795.

Ypma, R. J., Bataille, A., Stegeman, A., Koch, G., Wallinga, J., & Van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1728), 444–450.

Ypma, R. J., van Ballegooijen, W. M., & Wallinga, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, *195*(3), 1055–1062.

Zhang, Y., Leitner, T., Albert, J., & Britton, T. (2020). Inferring transmission heterogeneity using virus genealogies: Estimation and targeted prevention. *PLoS computational biology*, *16*(9), e1008122.