Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

TiTUS: Sampling and Summarizing Transmission Trees with Muti-strain Infections

Palash Sashittal [1,*] and Mohammed El-Kebir [2,*]

[1]Department of Aerospace Engineering, University of Illinois at Urbana-Champaign, Urbama, IL 61801, USA and
[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbama, IL 61801, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The combination of genomic and epidemiological data hold the potential to enable accurate pathogen transmission history inference. However, the inference of outbreak transmission histories remains challenging due to various factors such as within-host pathogen diversity and multi-strain infections. Current computational methods ignore within-host diversity and/or multi-strain infections, often failing to accurately infer the transmission history. Thus, there is a need for efficient computational methods for transmission tree inference that accommodate the complexities of real data.
Results: We formulate the Direct Transmission Inference (DTI) problem for inferring transmission trees that support multi-strain infections given a timed phylogeny and additional epidemiological data. We establish hardness for the decision and counting version of the DTI problem. We introduce TiTUS, a method that uses SATISFIABILITY to almost uniformly sample from the space of transmission trees. We introduce criteria that prioritizes parsimonious transmission trees that we subsequently summarize using a novel consensus tree approach. We demonstrate TiTUS's ability to accurately reconstruct transmission trees on simulated data as well as a documented HIV transmission chain.
Availability: `https://github.com/elkebir-group/TiTUS`
Contact: melkebir@illinois.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the advent of cheaper and more powerful sequencing methods, molecular epidemiology has become an indispensable tool for inference of transmission histories of infectious disease outbreaks. Genomic data of pathogen isolates collected from infected hosts is used to assist with the identification of unknown infection sources and transmission chains. Intensive field work generates crucial epidemiological data that provides addition information such as contact history between patients and exposure times of the patients to sources of infection. Methods that can efficiently use genomic and epidemiological data together for accurate inference of transmission history of outbreaks are the key to real-time outbreak management and devising public health policies and disease control strategies for future outbreaks (Dellicour *et al.*, 2018).

There are several challenges that hinder the accurate inference of the transmission history of an outbreak. Phylogeny estimation of the pathogen isolates reveals the evolutionary history of the pathogen during the outbreak. However, due to within-host diversity of the pathogen, branching events in the phylogeny do not correspond to the transmission events during the outbreak (Romero-Severson *et al.*, 2014). Phylogeny-based methods that assume that the transmission events coincide with the branching events in the phylogeny are therefore not applicable in the context of pathogens with low mutation rates, short incubation times and acute infections (Ypma *et al.*, 2011; Harris *et al.*, 2010; Leitner *et al.*, 1996; Cottam *et al.*, 2008).

Another factor that makes outbreak transmission history inference challenging is a *weak transmission bottleneck*, where multiple strains of the pathogen are transmitted from a donor to a recipient through a non-negligibly small inoculum. Due to this, the most recent common ancestor of lineages from the same host need not have arisen in that host. Although large inocula have been observed in a number of diseases (Leonard *et al.*, 2017), most of the existing methods for transmission tree inference

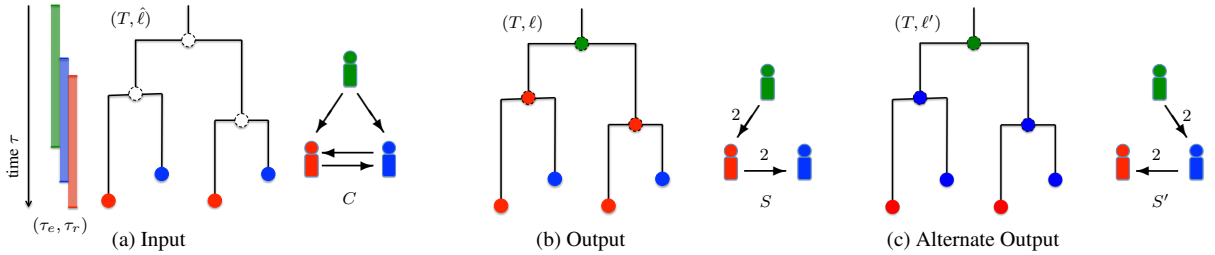


Fig. 1: **Overview of the Direct Transmission Inference (DTI) problem.** (a) The input of the problem consists of a timed phylogeny T that captures the evolutionary history of the pathogen during the course of the outbreak. Each leaf of T corresponds to a sample collected for an individual host and is thus labeled using $\hat{\ell}$ (indicated by colors). The entry and removal times $[\tau_e(s), \tau_r(s)]$ for each host s is also included in the input. (b) Our aim is to label the internal vertices of T with ℓ such that the resulting transmission edges form a transmission tree S (as shown in Fig. 1b). Each edge (s, t) of S is weighted by the number of transmission edges from host s to host t given by the vertex labeling ℓ. (c) An alternative solution to the given DTI instance. It is easy to see that no solution exists under the strong bottleneck constraint whereas under the weak transmission bottleneck there are multiple solutions. All the feasible vertex labelings are shown in Fig. S7.

that account for the within-host diversity do not account for the co-transmission of pathogen strains (Ypma *et al.*, 2013; Didelot *et al.*, 2014; Hall *et al.*, 2015; Didelot *et al.*, 2017). That is, these methods assume a *strong transmission bottleneck* where a single strain of the pathogen is transmitted in an infection. A weak transmission bottleneck is considered in SCOTTI (De Maio *et al.*, 2016) and BadTrIP (De Maio *et al.*, 2018), however they make the simplifying assumption that all the transmissions are independent of each other. SharpTNI (Sashittal and El-Kebir, 2019) considers the weak transmission bottleneck without this assumption under a parsimony based framework for a known phylogeny. However, SharpTNI may yield transmission histories that cannot be represented by a tree due to multiple infections of a single host from distinct donors. Such super-infection are unlikely for pathogens where infected hosts acquire immunity towards further infections of the pathogen (Whittle *et al.*, 1999; Wearing and Rohani, 2009), thus restricting the transmissions history to a tree.

The contributions of this paper are three-fold. First, we consider the problem of counting and sampling uniformly from the set of possible transmission trees for a known phylogeny and epidemiological data. In previous works, this problem is considered by Kenah *et al.* (2016) when the order of infections during the outbreak is completely known and by Hall and Colijn (2019) under the strong transmission bottleneck constraint. In this work, we relax both these constraints and propose a method TiTUS that approximately counts and almost uniformly samples the transmission trees under a weak transmission bottleneck for a given timed phylogeny (Fig. 1). We prove the hardness of the decision and counting versions of this problem and demonstrate the efficiency and accuracy of TiTUS on simulated data. Second, we present a robust criteria for ranking or prioritizing the uniformly sampled candidate transmission trees. In addition to the simulated data, we demonstrate the performance of the selection criteria on an HIV outbreak with a known transmission chain (Vrancken *et al.*, 2014). Third, in practice, the underlying phylogeny has some uncertainty and there can be multiple candidates for the transmission tree for a given phylogeny. It is therefore desirable to have an efficient method to summarize the solution space of transmission trees that are consistent with the genetic and epidemiological data. To this end, we propose a consensus-based method that provides the mean transmission tree for a set of candidate solutions while accounting for the number of distinct strains transmitted in each infection event.

2 Preliminaries

To state the problems we consider in this manuscript, we start by introducing the required concepts and notation. Let T be a rooted tree with vertex set $V(T)$ and edge set $E(T)$. The set of leaves of the tree is given by $L(T)$. The root of the tree is denoted by $r(T)$. We denote the children of a vertex u by $\delta_T(u)$. We write $u \preceq_T v$ if vertex u is ancestral to vertex v, *i.e.* vertex u is present on the unique path from $r(T)$ to vertex v. Note that $\preceq_T$ is reflexive, *i.e.* it holds that $u \preceq_T u$ for all vertices u. We denote the set of m distinct hosts in the outbreak by Σ. In a phylogeographical setting, the set Σ corresponds to m distinct geographical locations.

The evolutionary of all strains of a pathogen in an outbreak is modeled by a timed phylogeny, which we define as follows.

Definition 1. A *timed phylogeny* T is a rooted tree whose vertices are labeled by time-stamps $\tau : V(T) \to \mathbb{R}^{\geq 0}$ such that $\tau(u) \leq \tau(v)$ for all pairs u, v of vertices where $u \preceq_T v$.

Thus, as we can see in the above definition, time moves forward when traversing down a timed phylogeny T starting from the root $r(T)$. It is important to note that the leaves of a timed phylogeny T may occur at distinct time-stamps, *i.e.* T is not necessarily ultrametric.

Each leaf of a timed phylogeny T corresponds to a strain of pathogen that was collected during the outbreak. As such, we know the host from which each individual strain was isolated. This is captured by a leaf labeling, *i.e.* a labeling of the leaves of T by hosts Σ.

Definition 2. A *leaf labeling* of a timed phylogeny T is a function $\hat{\ell} : L(T) \to \Sigma$, assigning a host $\hat{\ell}(u) \in \Sigma$ to each leaf vertex $u \in L(T)$.

While we know the host $\hat{\ell}(u)$ from which each individual leaf u of T was sampled, we do not know the hosts of the internal vertices, which correspond to unsampled, ancestral strains. Here, our goal is to determine the hosts in which these ancestral strains reside.Mathematically, we wish to construct a *vertex labeling* $\ell : V(T) \to \Sigma$, such $\ell(u) = \hat{\ell}(u)$ for all leaves $u \in L(T)$. Given a vertex labeling ℓ, each internal vertex u of T thus corresponds to a strain residing within host $\ell(u)$ at time $\tau(u)$.

In addition to the evolutionary history of all strains in the outbreak, a timed phylogeny T combined with a vertex labeling ℓ gives us information about the transmission history of the outbreak. Transmissions of strains from one host to another correspond to edges (u, v) of T labeled by distinct hosts $\ell(u) \neq \ell(v)$. Formally, we define a *transmission edge* as follows.

Definition 3. Given a timed phylogeny T and vertex labeling ℓ, an edge (u, v) of T is a *transmission edge* if $\ell(u) \neq \ell(v)$.

The vertex labeling that we construct for a given timed phylogeny T and leaf labeling $\hat{\ell}$, must follow certain constraints for a realistic reconstruction of the transmission history of the pathogen. We will now define these epidemiological constraints.

The first constraint that we introduce is called the *direct transmission* constraint, which imposes the following two restrictions. First, the outbreak begins with a single infected host. We call this initial host the *root host* and it labels the root node $r(T)$ of the timed phylogeny. The *root host* is not infected by any other host and therefore if s is the root host, there cannot exist a transmission edge (u, v) such that $\ell(u) \neq s$ and $\ell(v) = s$. Second, the remaining hosts have a unique infector and are thus infected only once in the course of the outbreak. A possible explanation for this phenomenon is diseases where infected hosts acquire immunity towards further infections of the pathogen (Whittle *et al.*, 1999; Wearing and Rohani, 2009). Consequently, there cannot exist two distinct transmission edges (u, v) and (u', v') such that $\ell(v) = \ell(v')$ and $\ell(u) \neq \ell(u')$. However, an infection between any two hosts $s, t \in \Sigma$ may involve the transmission of multiple strains at the same time. This is known as a *weak transmission bottleneck*. Since the transmission of strains must occur concurrently, the time intervals corresponding to any two transmission edges between the same pair (s, t) of hosts must have an non-empty intersection. More formally, we state the *direct transmission* constraint as follows,

Definition 4. For a timed phylogeny T, a vertex labeling ℓ satisfies the *direct transmission constraint* if (i) there does not exist a transmission edge (u, v) such that $\ell(v) = \ell(r(T))$ and (ii) we have $[\tau(u), \tau(v)] \cap [\tau(u'), \tau(v')] \neq \emptyset$ for any two transmission edges (u, v) and (u', v') where $\ell(u) = \ell(u')$ and $\ell(v) = \ell(v')$.

Under the *direct transmission* constraint, the set of transmission edges induced by the vertex labeling ℓ uniquely determines the *transmission tree* S. More formally, the vertex set $V(S)$ of a transmission tree S is the host set Σ, and there is a directed edge from $s \in \Sigma$ to $t \in \Sigma$ if and only if there exists at least one edge $(u, v) \in E(T)$ such that (i) $s \neq t$, (ii) $\ell(u) = s$ and (iii) $\ell(v) = t$. Since every host except the *root host* has a unique infector, the directed edges necessarily form a tree. Each directed edge $(s, t) \in E(S)$ is given a weight $w : E(S) \rightarrow \mathbb{N}$ such that $w(s, t)$ equals the number of transmission edges in T from host s to t. If $w(s, t) = 1$ for all edges $(s, t) \in E(S)$ then each host is infected due to the transmission of a single pathogen strain. This phenomenon is known as a *strong transmission bottleneck*.

Epidemiological data provide two additional types of information. First, for each host s we are given an interval $[\tau_e(s), \tau_r(s)]$ during which the host was present in the outbreak and susceptible for infection. Specifically, $\tau_e(s) \in \mathbb{R}^{\geq 0}$ is the entry time at which host s became susceptible for infection, whereas $\tau_r(s) \in \mathbb{R}^{\geq 0}$ is the *removal time* at which the host was removed from the susceptible and infected populations and placed in treatment or recovering.

Second, there can also be documented geographical constraints that prevent transmissions between any given pair of hosts. We account for all such constraints using a *contact map*. A *contact map* C is a directed graph whose vertex set equals the set Σ of hosts. A directed edge (s, t) represents a possible infection event from host s to host t. If any two hosts are not connected in C then there can be no infection event between that pair of hosts. It can clearly be seen that (i) the contact map C is a subgraph of the interval graph induced by the intervals $[\tau_e(s), \tau_r(s)], \forall s \in \Sigma$ and (ii) the transmission tree S is a spanning arborescence of the contact map C. Thus, even in the absence of documented contacts between hosts, a contact map is induced by the entry and removal times of the hosts.

3 Problem Statement

We focus on inferring the transmission history of an outbreak for a known pathogen phylogeny T. In addition, we are given epidemiological data, which include the contact map C, entry and removal times $[\tau_e(s), \tau_r(s)]$ for each host $s \in \Sigma$ and assume a direct transmission constraint under a weak transmission bottleneck. This leads to the following decision problem.

Problem 1 (Direct Transmission Inference (DTI))**.** Given a timed phylogeny T with time-stamps $\tau : V(T) \rightarrow \mathbb{R}^{\geq 0}$, a leaf labeling $\hat{\ell} : L(T) \rightarrow \Sigma$, a contact map C and entry $\tau_e : \Sigma \rightarrow \mathbb{R}^{\geq 0}$ and removal times $\tau_r : \Sigma \rightarrow \mathbb{R}^{\geq 0}$, find a vertex labeling ℓ that induces a transmission tree S that is a spanning arborescence of C and $\tau(u) \in [\tau_e(s), \tau_r(s)]$ for all hosts s and vertices u where $\ell(u) = s$.

An instance of the DTI problem is shown in Fig. 1a shows an instance of the DTI problem along a with a solution vertex labeling ℓ and induced transmission tree S, where the three hosts are inducated using three colors. Importantly, a DTI problem instance may admit multiple solutions, as shown in Fig. 1b and Fig. 1c. These solutions provide alternative reconstructions of the transmission history, and thus must be taken into consideration in any downstream analysis of the outbreak to devise policy to better manage/prevent future outbreaks. To quantify the number of alternative reconstructions, we formulate the following counting problem.

Problem 2 (# Direct Transmission Inference (#DTI))**.** Given a timed phylogeny T with time-stamps $\tau : V(T) \rightarrow \mathbb{R}^{\geq 0}$, a leaf labeling $\hat{\ell} : L(T) \rightarrow \Sigma$, a contact map C and entry $\tau_e : \Sigma \rightarrow \mathbb{R}^{\geq 0}$ and removal times $\tau_r : \Sigma \rightarrow \mathbb{R}^{\geq 0}$, count the number of vertex labelings ℓ that induce a transmission tree S that is a spanning arborescence of C and $\tau(u) \in [\tau_e(s), \tau_r(s)]$ for all hosts s and vertices u where $\ell(u) = s$.

Let $\mathcal{L}$ be the set of all solutions to a given DTI problem instance. Ideally, we would exhaustively enumerate all solutions to the problem instance. However, worst case, the number of solutions scales exponentially with our input. Thus, to obtain a good overview of the solution space $\mathcal{L}$, we need to consider the sampling version of #DTI problem where we wish to uniformly sample the solution space.

In summary, we defined three versions of the DTI problem: a decision, counting and sampling version. In the following, we will consider a previously defined constrained version of the DTI problem as well as a generalization.

3.1 Related Transmission Tree Inference Problems

We start by considering a version of the DTI problem with one additional constraint. This additional constraint requires that only one pathogen strain is transmitted to a new host in a transmission event, and is known as a *strong transmission bottleneck*. We refer to this problem as Directed Transmission Inference under Strong Bottleneck (DTI-SB), and denote the space of solutions by $\mathcal{L}SB$. This problem was posed by Hall *et al.* (2015).In subsequent work, Hall and Colijn (2019) introduced a polynomial time algorithm to enumerate and uniformly sample from the set $\mathcal{L}_{\mathrm{SB}}$. Since the DTI-SB only has one additional constraint over the original DTI problem, the solution space of DTI-SB is a proper subset of the solution space of DTI for the same timed phylogeny T, leaf labeling $\hat{\ell}$ and epidemiological data. More formally, we have $\mathcal{L}_{SB} \subseteq \mathcal{L}$.

The second problem we consider is a relaxed version of DTI. Specifically, we relax the *direct transmission* constraint for a given instance of DTI. We refer to this problem as rel-DTI and the space of feasible solutions for a given instance by $\mathcal{L}_{\mathrm{REL}}$. Section 5.2.1 introduces a polynomial time dynamic programming algorithm that enumerates, counts and uniformly samples from the set $\mathcal{L}_{\mathrm{REL}}$. Since the rel-DTI problem is a relaxation of the DTI problem, we can use the algorithm introduced in Section 5.2.1 to uniformly sample from the solution space of the DTI problem ($\mathcal{L}$). Fig. 2 shows the relation between the solution spaces of the three transmission tree inference problems.

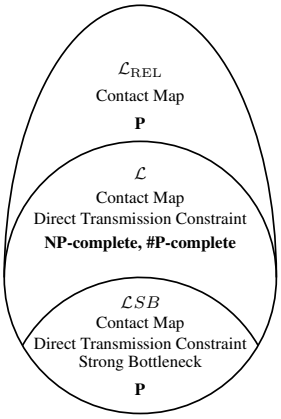


Fig. 2: **Schematic to compare the solution spaces of transmission trees under different constraints for a known timed phylogeny.** We have $\mathcal{L}_{\text{SB}} \subseteq \mathcal{L} \subseteq \mathcal{L}_{\text{REL}}$. $\mathcal{L}_{\text{SB}}$ is the solution space of transmission trees with a strong bottleneck that is considered in the work of Hall and Colijn (2019) where they show that counting the solutions and sampling from this solution space can be performed in polynomial time. $\mathcal{L}$ is the solution space of DTI which we show to be both NP-complete and #P-complete. Finally, $\mathcal{L}_{\text{REL}}$ is the relaxed solution space that is used to construct a polynomial time rejection based naive sampling and counting algorithm in Section 5.2.1.

3.2 Consensus Tree Problem

For the DTI problem described in the previous section, we start with a given pathogen phylogeny T. However, in practice the phylogeny needs to be inferred from genomic sequences of the strains collected from individual hosts Σ. Several methods of phylogeny inference generate either multiple candidates for the phylogeny or a posterior on the solution phylogeny space (Bouckaert *et al.*, 2019; Stamatakis, 2014). Moreover, for each given timed phylogeny, we can get multiple solutions to the DTI problem as shown for a representative instance in Fig. 1. Therefore, there is a need for an efficient method to summarize the candidate transmission trees that explain the disease outbreak.

A common method to summarize the solution space of transmission trees is to aggregate the information from the candidate transmission trees to generate a single graph where each edge is weighted by the number of candidate trees that support that edge (De Maio *et al.*, 2016; Wymant *et al.*, 2017; Didelot *et al.*, 2014). This graph rarely represents a single coherent transmission tree among the set of all hosts in the dataset. For this reason, the resulting graph is called a *relationship graph* (Wymant *et al.*, 2017) and does not provide crucial information about co-occurrence and mutual exclusivity among edges of the candidate transmission trees.

Another line of method summarizes the set of candidate solutions using one or more consensus trees that best represent the solution space (Jombart *et al.*, 2017; Kendall *et al.*, 2018). For instance, Jombart *et al.* (2017) apply pairwise distance metrics on the space $\mathcal{S}$ of transmission trees, not taking into account the number $w(s,t)$ of transmitted strains between pairs of host (s,t). The resulting distance matrix is subsequently embedded into lower dimensional space that the authors then cluster. Finally, each cluster is then assigned a single transmission tree in $\mathcal{S}$ as its representative (Hall and Colijn, 2019). Kendall *et al.* (2018) follow a similar embedding approach, again not taking the number $w(s,t)$ of transmission into account. Thus neither method supports a weak transmission bottleneck. To address this limitation, we define the weighted parent-child distance (WPCD) $d(S_1, S_2)$ between any two transmission trees S_1 and S_2 as follows.

Definition 5. Let $S_1 = (\Sigma, E_1)$ with edge labeling w_1 and $S_2 = (\Sigma, E_2)$ with edge labelings w_2 be two transmission tree on the same vertex set Σ. The *weighted parent-child distance* between the two graphs denoted by $d(S_1, S_2)$ is

$$d(S_1, S_2) = \sum_{(s,t)\in E_1} w_1(s,t) + \sum_{(s,t)\in E_2} w_2(s,t) - 2 \sum_{(s,t)\in E_1\cap E_2} \min\{w_1(s,t), w_2(s,t)\}. \quad (1)$$

In Appendix A.1.1 we show that this distance function induces a metric in the space $\mathcal{S}$ of transmission trees. Note that transmission trees S and S' that have the same topology but different edge weights w and w' will have $d(S, S') > 0$. As a result, WPCD can be used to produce a consensus transmission tree while taking an incomplete transmission bottleneck into account. Under the *strong transmission bottleneck* the *weighted parent-child distance* simplifies to the size of the symmetric difference between the edge sets of the two transmission trees, *i.e.* $d(S, S') = |E'\backslash E| + |E\backslash E'|$. This distance is known as the parent-child distance, and has been used to compare tumor phylogenies (Aguse *et al.*, 2019; Govek *et al.*, 2018). Using WPCD, we define the following consensus tree problem.

Problem 3 (Single Consensus Transmission Tree (SCTT))**.** Given k distinct transmission trees $\mathcal{S} = \{S_1, \cdots, S_k\}$ with edge labelings $\{w_1, \cdots, w_k\}$ find a consensus transmission tree R that minimizes $d(\mathcal{S}, R) = \sum_{i=1}^{k} d(S_i, R)$.

4 Complexity

This section establishes hardness results for the decision and counting versions of the DTI problem.

Theorem 1. DTI is NP-complete.

We show the hardness of DTI by reduction from the 1-in-3SAT problem, which is a known NP-complete problem (Karp, 1972). Details are in Appendix A.2.

It is known that the #1-in-3SAT is a #P-complete problem (Creignou and Hermann, 1993). In order to show that the #DTI is also #P-complete, it suffices to show that there exists a polynomial-time reduction from #1-in-3SAT such that the number of solutions is preserved, which we do in Appendix A.2.

Theorem 2. #DTI is #P-complete.

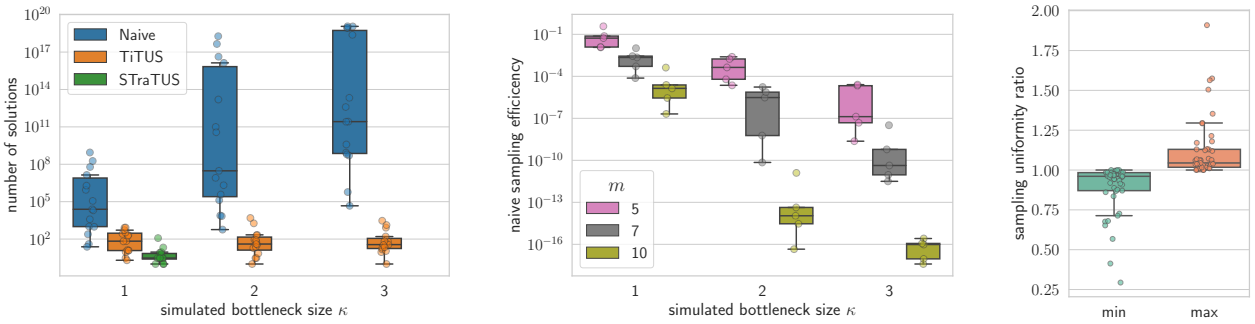
Since the decision problem DTI is NP-complete, there does not exist a fully polynomial randomized approximate scheme (FPRAS) for the counting version of DTI unless NP=RP (Jerrum, 2003; Miklós, 2019).

5 Methods

This sections describes the methods developed to solve the decision, counting and sampling versions of the DTI problem.

5.1 Decision Problem

Since the DTI is NP-complete, we propose to use SATISFIABILITY to solve the decision problem. As such, we construct a Boolean formula ϕ for a given DTI instance $(T, \hat{\ell}, \tau_e, \tau_r, C)$, such that there is a bijection between the solutions of the DTI instance and the corresponding SAT instance ϕ. Solving the SAT instance will then be equivalent to solving the corresponding DTI problem.



(a) (b) (c)

Fig. 3: **TiTUS accurately samples solutions to the DTI problem.** (a) The number of solution to the rel-DTI ($|\mathcal{L}_{\mathrm{REL}}|$), the DTI ($|\mathcal{L}|$), and the DTI-SB ($|\mathcal{L}_{\mathrm{SB}}|$) problems computed using the Naive rejection sampling, TiTUS, and STraTUS respectively. The number of solutions to the rel-DTI problem grows rapidly for increasing values of the simulated bottleneck size κ, while STraTUS fails to provide any solution when κ is greater than 1. (b) The sampling efficiency, defined as the ratio $|\mathcal{L}|$ and $|\mathcal{L}_{\mathrm{REL}}|$ for increasing values of simulated number of hosts m and bottleneck size κ. (c) The ratio between the minimum and maximum observed sampling frequency using TiTUS with the true uniform sampling frequency .

Vertex labeling: Decision variables $\mathbf{x} \in \{0,1\}^{n\times m}$ encode a vertex labeling, *i.e.* $x_{i,s} = 1$ if and only if the node $\ell(v_i) = s$ and $x_{i,s} = 0$ otherwise. We encode uniqueness of the label of each vertex with the following formula.

$$\mathrm{onehot}(\{x_{i,1}, \cdots, x_{i,m}\}), \quad \forall v_i \in V(T). \tag{2}$$

The function $\mathrm{onehot}(X)$ encodes that exactly one binary variable $x \in X$ is true, which can be accomplished by the following constraint.

$$\left[\bigvee_{x\in X} x\right] \wedge \left[\bigwedge_{x,y\in X} (\neg x \vee \neg y)\right]. \tag{3}$$

Transmission edges: We encode the transmission edges using variables $c_{s,t}$ with $s, t \in \Sigma$ and $s \neq t$. We enforce that $c_{s,t} = 1$ if and only if the host t is infected by host s and $c_{s,t} = 0$ otherwise as follows.

$$(x_{i,s} \wedge x_{j,t}) \implies c_{s,t}, \quad \forall (v_i, v_j) \in E(T) \text{ and } s, t \in \Sigma. \tag{4}$$

Root host: To enforce that the host which labels $r(T)$ is not infected by any other host, we have

$$x_{i,t} \implies \neg c_{s,t}, \quad \forall s, t \in \Sigma, s \neq t, \tag{5}$$

where $v_i = r(T)$.

Direct transmission constraint: We enforce that any host cannot be infected by more than one other host. For each host $s \in \Sigma$ we have

$$\neg c_{s,t} \vee \neg c_{s,t'}, \quad t, t' \in \Sigma \text{ and } t \neq t'. \tag{6}$$

We require that all transmission edges from host s to host t must have time intervals that overlap. For all edge pair $(v_i, v_j), (v_k, v_l)$ that do not have overlapping time intervals, i.e. $[\tau(v_i), \tau(v_j)] \cap [\tau(v_k), \tau(v_l)] = \emptyset$, we impose

$$\neg x_{i,s} \vee \neg x_{j,t} \vee \neg x_{k,s} \vee \neg x_{l,t}, \quad \forall s, t \in \Sigma, s \neq t. \tag{7}$$

5.2 Counting and Sampling Problem

5.2.1 Naive Rejection based Method

For a naive rejection sampling algorithm, we relax the *direct transmission constraint* and uniformly sample vertex labelings for the timed phylogeny T such that for all transmission edges (u, v) we have $(\ell(u), \ell(v)) \in E(C)$. As described in Section 3.1, we refer to this as the rel-DTI problem. Let the set of such vertex labelings be $\mathcal{L}_{\mathrm{REL}}$. Drawing a vertex labeling labeling $\ell \in \mathcal{L}_{\mathrm{REL}}$ uniformly at random from the set $\mathcal{L}_{\mathrm{REL}}$ can be done in polynomial time, as we describe in Appendix A.3. The sampled vertex labeling labeling ℓ is rejected unless it satisfies the *direct transmission constraint*, which can be verified in polynomial time. The probability of success for this rejection based sampling algorithm is $1-(|\mathcal{L}|/|\mathcal{L}_{\mathrm{REL}}|)^K$ after K repetitions.

5.2.2 Approximate Counting and Sampling using SAT

Using the SAT formulation shown in Section 5.1, we may use ApproxMC (Chakraborty *et al.*, 2013; Soos and Meel, 2019) to approximate $|\mathcal{L}|$ and UniGen (Chakraborty *et al.*, 2014, 2015) to sample almost uniformly from $\mathcal{L}$. We call the resulting method Transmission Tree Uniform Sampler (TiTUS).

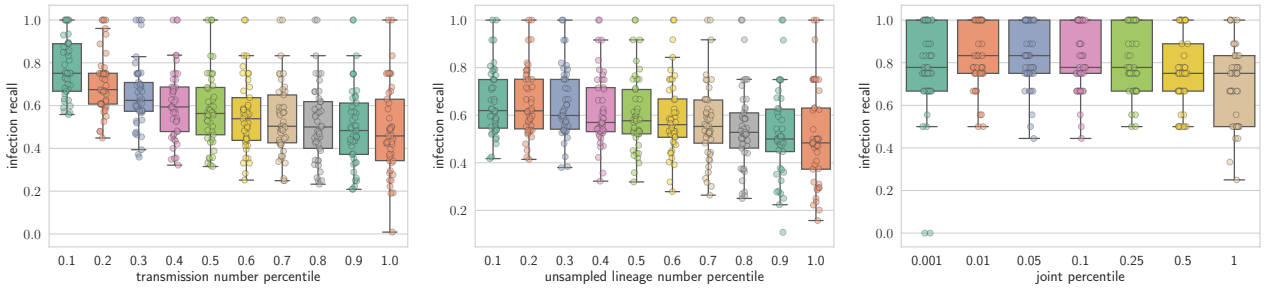
5.3 Consensus Problem

This section introduces a polynomial time algorithm to solve the SCTT problem. The algorithm and the proof for correctness follow the work of (Govek *et al.*, 2018). Let $\mathcal{S} = \{S_1, \cdots, S_k\}$ be a set of k transmission trees with edge weights $\{w_1, \cdots, w_k\}$. Our goal is to find a consensus tree R that minimizes $d(\mathcal{S}, R)$ where $d(\cdot,\cdot)$ is the *weighted parent-child distance*. For any given tree S_i, we define the function $q_i : \Sigma \times \Sigma \to \mathbb{N}$ where

$$q_i(s,t) = \begin{cases} w_i(s,t), & (s,t) \in E(S_i), \\ 0, & \text{otherwise.} \end{cases}$$

Observe that the parent-child distance between two transmission trees S_i and S_j can be re-written as

$$d(S_i, S_j) = \sum_{(s,t)\Sigma\times\Sigma} |q_i(s,t) - q_j(s,t)|.$$



(a) (b) (c)

Fig. 4: The transmission number and number of unsampled lineages of the solutions to the DTI problem are negatively correlated to the infection recall. (a) The infection recall for the uniformly sampled solution within different percentile based on the transmission number. (b) The infection recall for the uniformly sampled solution within different percentile based on the number of unsampled lineages. (c) The infection recall of the consensus transmission trees within different percentiles of both the transmission number and the number of unsampled lineages simultaneously.

To get the optimal weights for the consensus tree, for any pair of hosts $(s,t) \in \Sigma \times \Sigma$, we define

$$w^*(s,t) = \arg\min_{z>0} \sum_{S_i \in \mathcal{S}} |q_i(s,t) - z|.$$

Clearly, $w^*(s,t)$ for every pair of hosts (s,t) is given by $\max\{\text{med}, 1\}$ where med is the median of the set $\{q_1(s,t), \cdots, q_k(s,t)\}$. Thus, we have the following observation.

Observation 1. Given a set $\mathcal{S} = \{S_1, \cdots, S_k\}$ of k transmission trees with edge weights $w_1, \cdots, w_k$, optimal consensus trees R that include the edge (s,t) must assign this edge weight $w^*(s,t)$.

We define the *weighted parent-child graph* P as a complete graph with nodes given by the set Σ and a weight function

$$w_p(s,t) = \sum_{S_i \in \mathcal{S}} (|q_i(s,t) - w^*(s,t)| - |q_i(s,t)|)$$

Observe that the weights of the edges of P can be negative.

Theorem 3. Given a set $\mathcal{S} = \{S_1, \cdots, S_k\}$ of k transmission trees with edge weights $w_1, \cdots, w_k$, a minimum weight spanning arborescence of the corresponding weighted parent-child graph P defines a tree R that is a solution to the SCTT problem with the distance measure used is weighted parent-child distance.

Proof. Provided in Appendix A.4. □

Although edge weights w^* of P can be negative, the requirement of R to be a spanning arborescence of G means that we can solve this problem in polynomial time with standard minimum weight spanning arborescence algorithms.

6 Results

This section presents the results obtained by applying TiTUS to simulated as well as a real dataset.

6.1 Simulations

We employ a two-stage approach to simulate an outbreak, generalizing Didelot *et al.* (2014)'s simulation framework that uses a strong transmission bottleneck to support a weak transmission bottleneck. First,

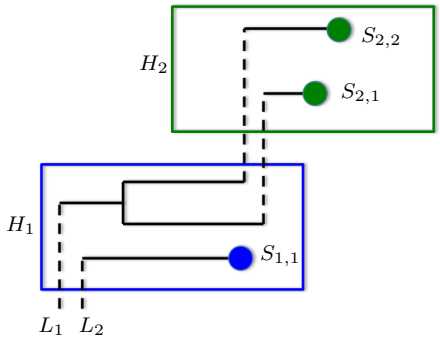


Fig. 5: **Schematic representation of unsampled lineages in outbreaks**. Different hosts H_1 and H_2 are represented by rectangular boxes and the samples taken from the hosts are indicated by blue or green circles inside the boxes respectively. Black lines represent the evolution of pathogen lineages. Solid lines correspond to within-host evolution of the pathogen whereas dashed lines represent the transmission of strains during infection. Two lineages L_1 and L_2 entering host H_1 are shown. Lineage L_1 is an *unsampled lineage* because even though two strains of L_1 are transmitted to host H_2, none of the samples of H_1 belong to the lineage L_1.

we simulate the transmission process between the m hosts using the SIR epidemic model (Allen, 2008). The epidemiological model takes the transmission bottleneck size κ and minimum number n_s of strains/leaves for each host s as input. Given this input, the model generates a transmission tree S with entry $\tau_e(s)$ and removal times $\tau_r(s)$ for each host s as well as the number of transmissions $w(s,t) = \kappa$ between each pair $(s,t) \in E(S)$ of hosts. Given S and w, we then simulate the evolution of the pathogens within each infected host using a simple coalescence model with constant population size (Kingman, 1982). This process yields a forest of timed phylogenies for each individual host s. We construct a single timed phylogeny of all hosts by stitching together individual timed phylogenies using the transmission tree S. For each combination of number $m \in \{5, 7, 10\}$ of hosts and bottleneck size $\kappa \in \{1, 2, 3\}$ we generate five instances, amounting to a total of 45 simulated instances. The cases with $\kappa = 1$ correspond to outbreaks with a strong transmission bottleneck. In order to mimic the uncertainty in epidemiological data seen in practice, we increase the length of the entry and removal time interval $[\tau_e(s) - \Delta, \tau_r(s) + \Delta]$ for each host s, where Δ equals 10% of the total outbreak duration.

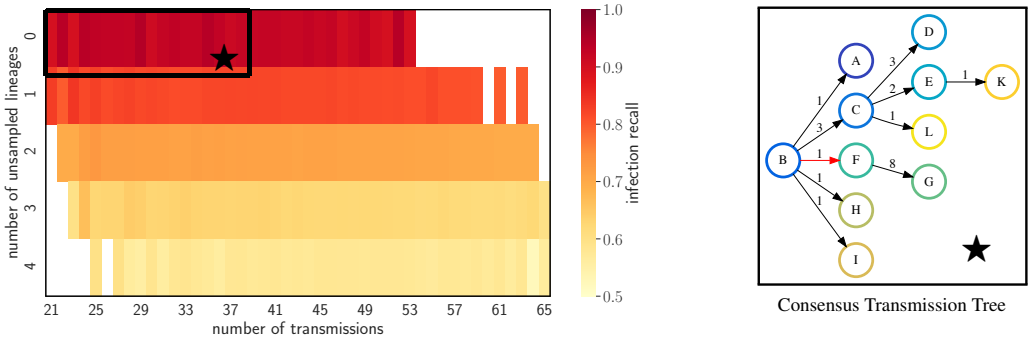


Fig. 6: **Consensus transmission tree computed for the solutions selected using the proposed criteria infers almost the entire transmission chain for the HIV outbreak.** The figure on the left shows the infection recall of the solutions with different transmission numbers and number of unsampled lineages, uniformly sampled using TiTUS. The black box encompasses the solutions selected for the percentile threshold of $\alpha = 0.01$. The figure on the right shows the consensus transmission tree for the selected solutions. Each edge is labeled by the number of strains transmitted from the donor to the recipient host. The incorrectly inferred transmission B→F is highlighted in red.

We find that increasing the number of hosts and bottleneck size in the simulations leads to an increase in the number of vertices n in the phylogenetic trees (Fig. S12a). This leads to a sharp increase in the number of feasible solutions to the rel-DTI (Fig. 3a). The number of solutions to DTI, on the other hand, stays relatively constant for increasing bottleneck size. As a consequence of this, the sampling efficiency of the naive rejection sampling method, defined by the ratio $\mathcal{L}/|\mathcal{L}_{\mathrm{REL}}|$, precipitates with increasing number m of hosts and bottleneck size κ proving it unsuitable for any real applications.

For cases with simulated bottleneck size $\kappa > 1$, STraTUS fails to provide any solutions (Fig. 3a). This shows that when multi-strain infections occur, transmission history inference with a strong bottleneck assumption will fail to provide the true transmission tree topology. Finally, we assess the sampling accuracy of TiTUS by comparing the sampling frequency with $1/|\mathcal{L}|$ where $|\mathcal{L}|$ is computed with sharpSAT (Thurley, 2006). For each unique solution that is sampled, the expected sampling frequency $1/|\mathcal{L}|$ is the same. Fig. 3c shows that the ratio between both the minimum and maximum values of the observed sampling frequencies with their expected values is close to 1.

In summary, our simulations show that methods that assume a strong transmission bottleneck cannot be applied to outbreaks with a weak bottleneck. Moreover, the exponentially increasing gap between the size of the solution space of rel-DTI compared to DTI renders the rejection-based sampling impractical. In contrast, TiTUS almost uniformly samples from the complex solution space of DTI.

6.1.1 Criteria to Prioritize Candidate Transmission Trees

We propose several criteria for ranking the vertex labelings for a given timed phylogeny uniformly sampled by TiTUS. The first criterion is the *number of transmission edges* in the vertex labeling. Based on the parsimony principle, which has been used in previous works for both phylogeny inference (Sankoff, 1975) as well as transmission tree inference (Wymant *et al.*, 2017; Snitkin *et al.*, 2012; Sashittal and El-Kebir, 2019), we expect vertex labelings that have few transmission edges to be closer to the ground truth. The second criterion is the *number of unsampled lineages*, which is the number of transmission edges (u, v) for which there does not exist a descendant leaf v' (*i.e.* $v \preceq_T v'$) labeled by $\ell(v)$. Unsampled lineages are a consequence of multi-strain infections and we expect to see fewer unsampled lineages when the within-host diversity of the infected hosts is adequately sampled. Fig. 5 illustrates this concept.

To assess these criteria, we compare the sampled transmission trees with the ground truth by computing the *infection recall*, defined as the fraction of transmission events between pairs of hosts that are correctly inferred. Fig. 4a shows the value of the *infection recall* for candidate solutions in different percentiles based on the number of transmission edges. Clearly, as we look at solutions with larger transmission numbers, the infection recalls decreases. Fig. 4b show a similar negative correlation between the infection recall and the number of unsampled lineages. We use both the transmission number and the number of lineages to prioritize the uniformly sampled candidate solutions. Specifically, for any given percentile threshold α we include all the vertex labelings whose percentile is at most α for both the transmission number and the number of unsampled lineages. (Thus, setting $\alpha = 1$ will include all sampled vertex labelings.) The selected vertex labelings are then used to compute the consensus transmissions tree. Fig. 4c shows the infection recall of the consensus transmission trees for increasing value of the percentile threshold α. We see that a value of α that is either too small or too large results in a decrease in the *infection recall* Based on the simulated data, we see that $\alpha^* = 0.01$ yields accurate consensus transmission tree solutions. Hence, the two criteria enable accurate prioritization of sampled vertex labelings.

6.2 HIV Outbreak with a Known Transmission Chain

We apply our method TiTUS to infer the transmission history of an HIV-1 outbreak involving 11 patients with a known transmission chain (Vrancken *et al.*, 2014; Lemey *et al.*, 2005). The data consists of 212 samples collected over the span of 18 years from the 11 patients. The direction of transmissions and a relatively narrow time interval for each transmission event were inferred from epidemiological information obtained by patient interviews, clinical data and treatment histories of the patients.

The DTI problem for this HIV dataset is set up as follows. For the timed phylogeny, we use the Maximum Clade Credibility (MCC) tree obtained from the partially sequenced *env* regions presented by Vrancken *et al.* (2014) in their publication. Table 1 in Appendix A.6 shows the sampling times and transmission windows provided in the epidemiological data for each of the hosts. The transmission window of a host is the time interval inside of which the host is expected to have been infected. Transmission windows for host A and host D are incongruent with the given timed phylogeny. By this we mean there is no vertex labeling on the given MCC phylogeny that allows for the known transmissions to host A and host D. We exclude these time windows, while the transmission windows for

the remaining hosts are used to constraint the possible vertex labelings of the MCC tree. We restrict the infection for each host to take place in within the transmission window provided in the epidemiological data. Appendix A.6 shows the details of the implementation of this constraint in the SAT formulation. Note that while using the time window constraints, we only restrict the time of infection and do not utilize information about the known infectors for each infected host. Finally, for each host the entry time is taken as the beginning of its time window of transmission and the removal time is the latest date of sampling (Table 1). We find that STraTUS fails to provide a solution on this dataset. Indeed, a weak transmission bottleneck needs to be considered in order to infer the transmission history.

For this DTI instance, using sharpSAT (Thurley, 2006) we find that there are exactly 30,901,500 feasible vertex labelings. We generate 100,000 samples from this solution space and compute the *infection recall* when compared to the known transmission chain. Fig. 6 shows the values the *infection recall* for solutions with different number of transmission edges and number of unsampled lineages. The infection recall is close to 1 for the solutions that have no unsampled lineages. The number of transmission edges also has a negative, albeit weaker correlation with the infection recall.

For any given percentile threshold α we include all vertex labelings whose percentile is at most α for both the transmission number and the number of unsampled lineages. Based on the simulations, we focus on percentile threshold $\alpha^* = 0.01$. For this threshold value, Fig. 6 shows the consensus transmission tree inferred by TiTUS. The infection recall for this tree is 0.9, *i.e.* we correctly infer $9/10$ transmission from the known transmission chain. We incorrectly infer the transmission B$\rightarrow$F while the known transmission to F based on epidemiological data is A$\rightarrow$F. Fig. S14 shows similar behavior of the infection recall as a function of α as observed in our simulations. Moreover, this figure shows that our method is robust around $\alpha^* = 0.01$.

7 Discussion

In this paper, we formulated the Direct Transmission Inference (DTI) problem of inferring transmission trees for a given timed phylogeny and epidemiological data while supporting a weak transmission bottleneck. Weak transmission bottlenecks are common in the spread of diseases due to pathogens with large inoculum sizes, high mutation rates, long incubation times and chronic infections Leonard *et al.* (2017). Previous studies of counting and sampling transmission trees for a given timed phylogeny assume a strong transmission bottleneck (Kenah *et al.*, 2016; Hall and Colijn, 2019), and are not applicable to outbreaks of pathogens with a weak transmission bottleneck, often failing to return any solution.

We proved that the decision version of the DTI problem is NP-complete and the counting version #DTI is #P-complete. Leveraging recent advances made in approximate counting and sampling of solutions to SATISFIABILITY (Chakraborty *et al.*, 2014, 2013, 2015; Soos *et al.*, 2009), TiTUS, which uses a SATISFIABILITY oracle to almost uniformly sample from the solution space of DTI. In most cases, uniformly sampled candidate solutions from the transmission tree space will deviate considerably from the ground truth. To address this issue, we proposed two criteria that can be used to prioritize the uniformly sampled transmission trees. We demonstrated the performance and robustness of our selection criteria on both simulated data and a real dataset of an HIV outbreak (Vrancken *et al.*, 2014).

Further, we also considered the problem of summarizing a given set of candidate transmission tree solutions of a disease outbreak. We defined a new distance metric *weighted parent-child distance* (WPCD) on the space of transmission multi-trees that capture the transmission of multiple strains between hosts during an outbreak. This distance is an extension of the parent-child distance which is used in previous works to summarize cancer phylogenies (Govek *et al.*, 2018; Aguse *et al.*, 2019). We presented a polynomial time algorithm for finding the consensus transmission tree with minimum total WPCD from the candidate solutions. The performance of the consensus transmission tree of recalling the transmissions that occurred during the outbreak is demonstrated both on simulated and real datasets.

There are several avenues for future research. First, the decision version of the DTI problem can be used to prioritize a posterior distribution of phylogenies, by checking if each phylogeny admits a vertex labeling that induces a transmission tree that is compatible with the given epidemiological data. A similar approach is employed by Sledzieski *et al.* (2019) where they prioritize statistically likely timed phylogenies that admit vertex labelings with fewer transmission edges. By including biological relevant constraints such as a contact map and direct transmission constraints, we expect to obtain high-fidelity phylogenetic and transmission history reconstructions. Second, one limitation of the proposed method is that it assumes that all the infected hosts in the outbreak are sampled. This assumption is only applicable for small outbreaks in regions with perfect surveillance and reporting system in place. An extension of this method to include unsampled hosts would be a useful. Third, akin to (Jombart *et al.*, 2017), we plan to extend the SCTT to simultaneously cluster the set $\mathcal{S}$ of transmission trees and infer a representative consensus transmission tree for each cluster. Finally, we plan to directly include the identified prioritization criteria as constraints in the DTI problem.

Funding

M.E-K. was supported by the National Science Foundation (grant: CCF 18-50502).

References

Aguse, N., Qi, Y., and El-Kebir, M. (2019). Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. *Bioinformatics*, **35**(14), i408–i416.

Allen, L. J. (2008). An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., *et al.* (2019). BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, **15**(4), e1006650.

Chakraborty, S., Meel, K. S., and Vardi, M. Y. (2013). A Scalable Approximate Model Counter. In *Principles and Practice of Constraint Programming*, pages 200–216. Springer, Berlin, Heidelberg, Berlin, Heidelberg.

Chakraborty, S., Meel, K. S., and Vardi, M. Y. (2014). Balancing scalability and uniformity in sat witness generator. In *Proceedings of the 51st Annual Design Automation Conference*, pages 1–6. ACM.

Chakraborty, S., Fremont, D. J., Meel, K. S., Seshia, S. A., and Vardi, M. Y. (2015). On parallel scalable uniform sat witness generation. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 304–319. Springer.

Cottam, E. M., Thébaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P., and Haydon, D. T. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, **275**(1637), 887–895.

Creignou, N. and Hermann, M. (1993). On P completeness of some counting problems. Research Report RR-2144, INRIA.

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

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

Dellicour, S., Baele, G., Dudas, G., Faria, N. R., Pybus, O. G., Suchard, M. A., Rambaut, A., and Lemey, P. (2018). Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature communications*, **9**(1), 2222.

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

Didelot, X., Fraser, C., Gardy, J., and Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, **34**(4), 997–1007.

Govek, K., Sikes, C., and Oesper, L. (2018). A consensus approach to infer tumor evolutionary histories. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 63–72. ACM.

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

Hall, M. D. and Colijn, C. (2019). Transmission trees on a known pathogen phylogeny: enumeration and sampling. *Molecular biology and evolution*, **36**(6), 1333–1343.

Harris, S. R., Feil, E. J., Holden, M. T., Quail, M. A., Nickerson, E. K., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J. A., *et al.* (2010). Evolution of mrsa during hospital transmission and intercontinental spread. *Science*, **327**(5964), 469–474.

Jerrum, M. (2003). *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. *Molecular ecology resources*, **17**(6), 1385–1392.

Karp, R. M. (1972). *Reducibility among Combinatorial Problems*, pages 85–103. Springer.

Kenah, E., Britton, T., Halloran, M. E., and Longini Jr, I. M. (2016). Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. *PLoS computational biology*, **12**(4), e1004869.

Kendall, M., Ayabina, D., Xu, Y., Stimson, J., Colijn, C., *et al.* (2018). Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees. *Statistical Science*, **33**(1), 70–85.

Kingman, J. (1982). b the coalescent. stoch. In *Proc. Appl*, volume 13, pages 235–248.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., and Albert, J. (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences*, **93**(20), 10864–10869.

Lemey, P., Derdelinckx, I., Rambaut, A., Van Laethem, K., Dumont, S., Vermeulen, S., Van Wijngaerden, E., and Vandamme, A.-M. (2005). Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *Journal of virology*, **79**(18), 11981–11989.

Leonard, A. S., Weissman, D. B., Greenbaum, B., Ghedin, E., and Koelle, K. (2017). Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *Journal of virology*, **91**(14), e00171–17.

Miklós, I. (2019). *Computational Complexity of Counting and Sampling*. CRC Press.

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., and Leitner, T. (2014). Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Molecular biology and evolution*, **31**(9), 2472–2482.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, **28**(1), 35–42.

Sashittal, P. and El-Kebir, M. (2019). Sharptni: Counting and sampling parsimonious transmission networks under a weak bottleneck. *bioRxiv*, page 842237.

Sledzieski, S., Zhang, C., Mandoiu, I., and Bansal, M. S. (2019). Treefix-tp: Phylogenetic error-correction for infectious disease transmission network inference. *bioRxiv*, page 813931.

Snitkin, E. S., Zelazny, A. M., Thomas, P. J., Stock, F., Henderson, D. K., Palmore, T. N., Segre, J. A., Program, N. C. S., *et al.* (2012). Tracking a hospital outbreak of carbapenem-resistant klebsiella pneumoniae with whole-genome sequencing. *Science translational medicine*, **4**(148), 148ra116–148ra116.

Soos, M. and Meel, K. S. (2019). BIRD: Engineering an efficient CNF-XOR SAT solver and its applications to approximate model counting. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)(1 2019)*.

Soos, M., Nohl, K., and Castelluccia, C. (2009). Extending sat solvers to cryptographic problems. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 244–257. Springer.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.

Thurley, M. (2006). sharpsat–counting models with advanced component caching and implicit bcp. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 424–429. Springer.

Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., Van Wijngaerden, E., Vandamme, A.-M., Van Laethem, K., and Lemey, P. (2014). The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. *PLoS computational biology*, **10**(4).

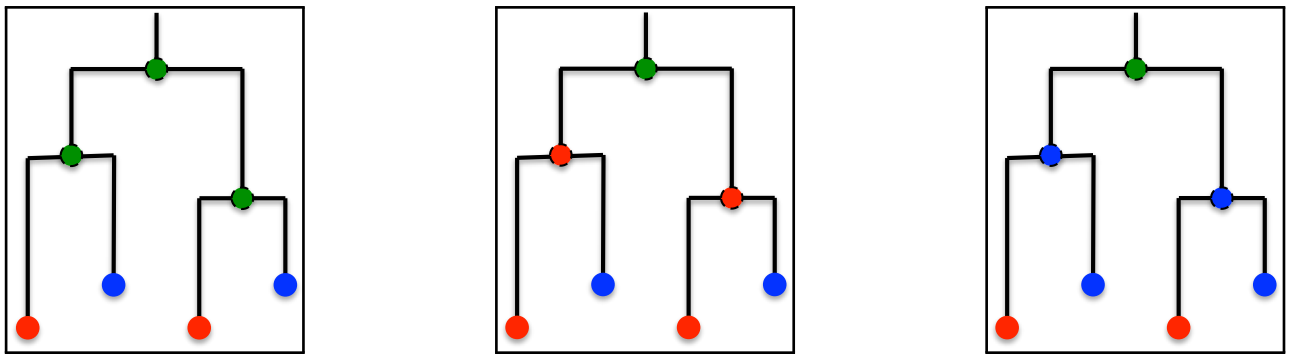
Wearing, H. J. and Rohani, P. (2009). Estimating the duration of pertussis immunity using epidemiological signatures. *PLoS pathogens*, **5**(10).

Whittle, H. C., Aaby, P., Samb, B., Jensen, H., Bennett, J., and Simondon, F. (1999). Effect of subclinical infection on maintaining immunity against measles in vaccinated children in west africa. *The Lancet*, **353**(9147), 98–102.

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, M., Fraser, C., STOP-HCV Consortium, T. M. P. C., and Collaboration, T. B. (2017). Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular biology and evolution*, **35**(3), 719–733.

Ypma, R. J., Bataille, A., Stegeman, A., Koch, G., Wallinga, J., and Van Ballegooijen, W. M. (2011). 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., and Wallinga, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, **195**(3), 1055–1062.



(a) (b) (c)

Fig. S7: **The timed phylogeny shown in Fig. 1a has 3 possible vertex labeling solutions.**

A.1 Background and Theory

In this section we provide the information we could not include in the main text. Fig. S7 shows all the feasible solutions to the representative DTI problem desribed in the Fig. 1.

A.1.1 Transmission Tree Metric

In this section we show that WPCD is a distance metric. To show that WPCD is a distance metric, for any transmission tree S_i, we define the function $q_i : \Sigma \times \Sigma \to \mathbb{N}$ as

$$q_i(s,t) = \begin{cases} w_i(s,t), & (s,t) \in E(S_i), \\ 0, & \text{otherwise.} \end{cases}$$

Observe that, by construction, q_i uniquely determines the transmission tree S_i since for any edge $(s,t) \in E(S_i)$ we have $w_i(s,t) > 0$. Further, the WPCD between any two transmission trees S_1 and S_2 can be alternatively written in terms of q_1 and q_2 as follows,

$$d(S_1, S_2) = \sum_{(s,t)\Sigma\times\Sigma} |q_1(s,t) - q_2(s,t)|.$$

Proposition 1. WPCD is a distance metric on the space of transmission trees $\mathcal{T}$.

Proof. First, we show that for any two transmission trees S_1 and S_2, $d(S_1, S_2) = 0$ if and only if $S_1 = S_2$. Clearly when $S_1 = S_2$, we have $d(S_1, S_2) = 0$. Now, let us consider the case $d(S_1, S_2) = 0$. For any $(s,t) \in \Sigma \times \Sigma$, $|q_1(s,t) - q_2(s,t)| \geq 0$. Therefore, if $d(S_1, S_2)$ then for all $(s,t) \in \Sigma \times \Sigma$ we have $q_1(s,t) = q_2(s,t)$ implying that $S_1 = S_2$.

By definition, WPCD is always nonnegative and symmetric.

We only need to show the triangle inequality, *i.e.* given trees S_1, S_2 and S_3, we must show

$$d(S_1, S_3) \leq d(S_1, S_2) + d(S_2, S_3).$$

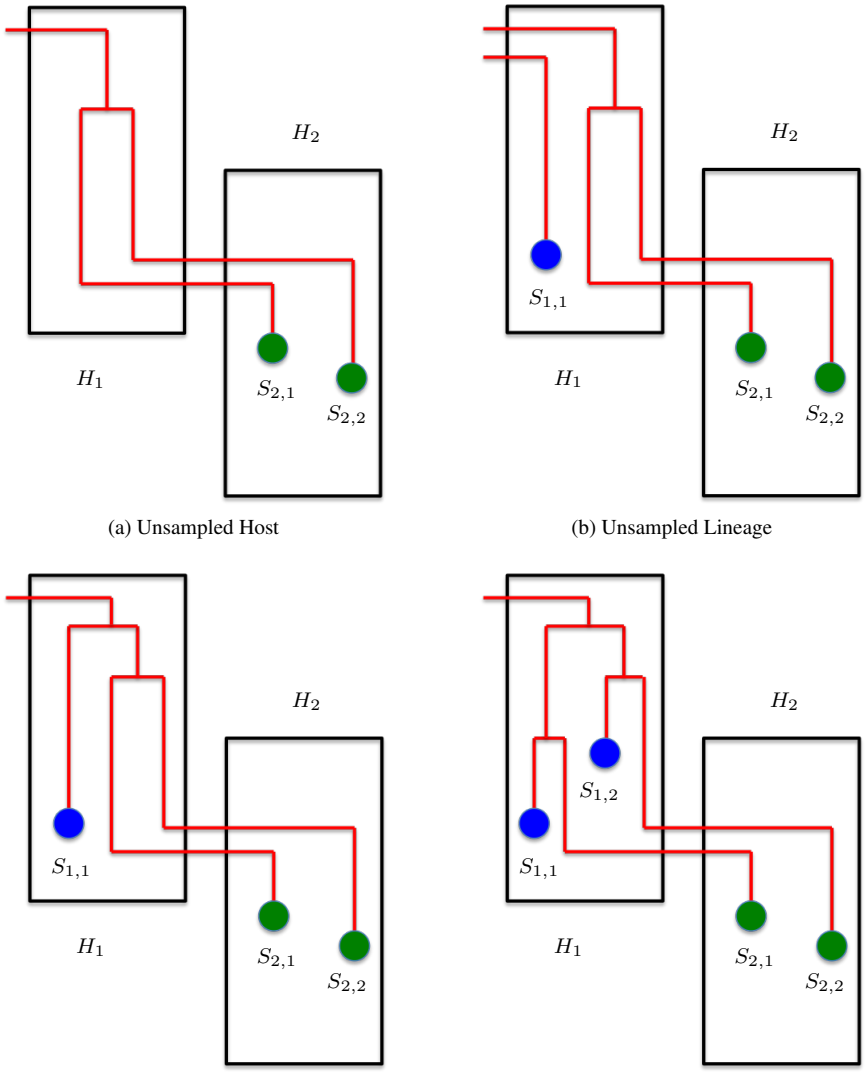
We show this as follows,

$$\begin{aligned} d(S_1, S_3) &= \sum_{(s,t)\in\Sigma\times\Sigma} |q_1(s,t) - q_3(s,t)| \\ &= \sum_{(s,t)\in\Sigma\times\Sigma} |q_1(s,t) - q_2(s,t) + q_2(s,t) - q_3(s,t)| \\ &\leq \sum_{(s,t)\in\Sigma\times\Sigma} (|q_1(s,t) - q_2(s,t)| + |q_2(s,t) - q_3(s,t)|) \\ &= d(S_1, S_2) + d(S_2, S_3). \end{aligned}$$

□

A.1.2 Sampling Scenarios

The weak transmission bottleneck has some interesting implications for the sampling of the within-host diversity of the infected hosts. Fig. S8 gives an overview, with schematic representations, of 4 different scenarios that can occur for real outbreaks.



(a) Unsampled Host (b) Unsampled Lineage

(c) Unsampled Strain (d) Complete Sampling

Fig. S8: **Schematic representation of different sampling scenarios during an outbreak.** Different hosts H_1 and H_2 are represented by rectangular boxes and the samples taken from the hosts are indicated by blue or green circles inside the boxes respectively. Red lines represent the evolution of pathogen lineages. Different scenarios described are (a) *Unsampled Host* scenario where host H_1 is not sampled even though it is part of the outbreak and infects H_2 with multiple strains (b) *Unsampled Lineage* where even though host H_1 is sampled with sample $S_{1,1}$, the lineage that passes two strains into host H_2 remains unsampled (c) *Unsampled Strain* scenario where the host H_1 is sampled and the right lineage is also sampled however the two strains that are transmitted to host H_2 are not sampled (d) *Complete Sampling* scenario where there is no incomplete lineage sorting (ILS) and all the strains transmitted from H_1 to H_2 are sampled.

A.2 Complexity

In this section we show the hardness of the decision and the counting versions of the DTI problem using reduction the one-in-three SAT (1-in-3 SAT).

Problem 4 (1-in-3SAT)**.** Given a Boolean formula $\phi = \bigwedge_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ in 3-conjunctive normal form (3-CNF) with n variables and k clauses, decide whether there exists a truth assignment $\theta : [n] \to \{0, 1\}$ so that each clause has *exactly* one true literal (and thus exactly two false literals).

A.2.1 Decision Problem

To relate literals to variables, we use the function $\nu : [k] \times \{1, 2, 3\} \to [n]$ such that $\nu(i, j)$ is the variable corresponding to literal $y_{i,j}$. We define $\sigma(i, j)$ to be 1 if $y_{i,j}$ is a positive literal (*i.e.* $y_{i,j} = x_{\nu(i,j)}$), otherwise $\sigma(i, j) = 0$ if $y_{i,j}$ is a negative literal (*i.e.* $y_{i,j} = \neg x_{\nu(i,j)}$). A truth assignment θ satisfies ϕ if for each clause $i \in [k]$ there exists a $j \in \{1, 2, 3\}$ such that $\sigma(i, j) = \theta(\nu(i, j))$.

Given ϕ, we construct a timed phylogeny $T(\phi)$ with leaf labeling $\hat{\ell}$, a contact map $C(\phi)$ and time-stamps τ, τ_e, τ_r, as depicted in Fig. S9 and detailed below. We set $\Sigma = \{\perp, x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n\, c_1, \ldots, c_k\}$. Let $\varepsilon > 0$ be a small positive constant. As for entry and removal time-stamps, we set $\tau_e(\perp) = 0, \tau_r(\perp) = \varepsilon$, and $\tau_e(x_i) = \tau_e(\neg x_i) = \varepsilon$ and $\tau_r(x_i) = \tau_r(\neg x_i) = 3\varepsilon$ for each variable $i \in [n]$. For each clause $c_i,\ i \in [k]$ we

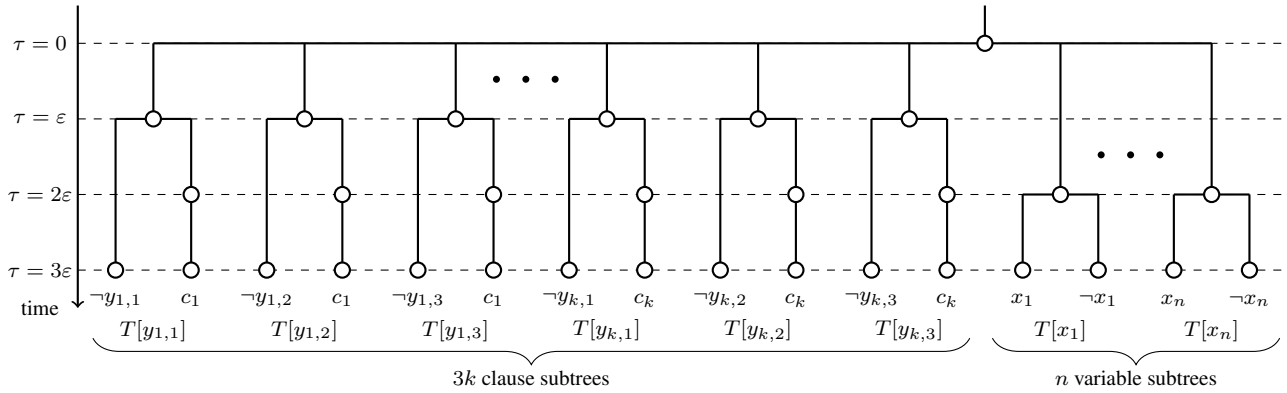


Fig. S9: **Construction of $T(\phi)$ for reduction from 1-in-3SAT to DTI.** Let ϕ be an 1-in-3SAT formula with k clauses and n variables. $T(\phi)$ is built with a root node $r(T(\phi))$ can is connected to $3k$ clause subtrees $\{T[y_{1,1}], T[y_{1,2}], T[y_{1,3}], \cdots, T[y_{k,1}], T[y_{k,2}], T[y_{k,3}]\}$ and n variable subtrees $\{T[x_1], \cdots, T[x_n]\}$. We set $\tau_e(\perp) = 0, \tau_r(\perp) = \varepsilon$, and $\tau_e(x_i) = \tau_e(\neg x_i) = \varepsilon$ and $\tau_r(x_i) = \tau_r(\neg x_i) = 3\varepsilon$ for each variable $i \in [n]$. For each clause $c_i,\ i \in [k]$ we set $\tau_e(c_i) = \tau_r(c_i) = 3\varepsilon$. We prove that there exits a truth assignment so that each clause of ϕ has exactly one true literal if and only if there exists a vertex labeling for $T(\phi)$ that results in a transmission tree that is a spanning arborescence of the contact map $C(\phi)$ (Fig. S10).

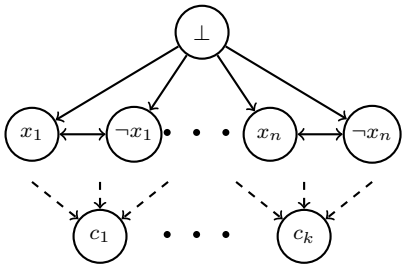


Fig. S10: **Construction of $C(\phi)$ for reduction from 1-in-3SAT to DTI.** Let ϕ be an 1-in-3SAT formula with k clauses and n variables. The host set is $\Sigma = \{\perp, x_1, \cdots, x_n, \neg x_1, \cdots, \neg x_n, c_1, \cdots, c_k\}$. We have a directed edge from $\perp$ to each of the variables $\{x_1, \cdots, x_n, \neg x_1, \cdots, \neg x_n\}$. Each each $i \in [n]$, variable x_i has an outgoing edge to $\neg x_i$ and similarly variable $\neg x_i$ has an outgoing edge to x_i. Finally, each clause c_i has three incoming edges, one from each of the literals that form the clause, i.e. $y_{i,1}, y_{i,2}$ and $y_{i,3}$.

set $\tau_e(c_i) = \tau_r(c_i) = 3\varepsilon$. Timed phylogeny $T(\phi)$ is composed of $3k$ clause gadgets and n variable gadgets, each corresponding to a subtree that is directly attached to the root $r(T(\phi))$. The root vertex has time-stamp $\tau(r(T(\phi)) = 0$. The leaves of T have identical time-stamps 3ε. For each variable $i \in [n]$, we have a subtree $T[x_i]$ whose root has time-stamp $\tau(r(T[x_i])) = 2\varepsilon$. The two children of $r(T[x_i])$ have identical time-stamps 3ε, with one child leading to two leaves labeled by positive literal x_i and the other child leading to two leaves labeled by negative literals $\neg x_i$. Similarly, for each clause $c_i,\ i \in [k]$, we have 3 subtrees $T[y_{i,1}], T[y_{i,2}]$ and $T[y_{i,3}]$. The root of the subtree $T[y_{i,j}]$ has time-stamp ε and two children, one of which is the leaf labeled by $x_{\nu(i,j)}$ if $y_{i,j} = \neg x_{\nu(i,j)}$ and $\neg x_{\nu(i,j)}$ if $y_{i,j} = x_{\nu(i,j)}$. The other child node, denoted as $v_{i,j}$, has time-stamp $\tau(v_{i,j}) = 2\varepsilon$ and has only one child which is a leaf labeled by c_i. The contact map $C(\phi)$ is constructed as follows. The vertex set for the contact map is given by Σ. We have a directed edge from $\perp$ to each of the variables $\{x_1, \cdots, x_n, \neg x_1, \cdots, \neg x_n\}$. For $i \in [n]$, each variable x_i has an outgoing edge to $\neg x_i$ and similarly variable $\neg x_i$ has an outgoing edge to x_i. Finally, each clause c_i has three incoming edges, one from each of the literals that form the clause, i.e. $y_{i,1}, y_{i,2}$ and $y_{i,3}$. For instance, if $c_1 := (x_1 \vee x_2 \vee \neg x_3)$, then we have the directed edges $(x_1, c_1), (x_2, c_1)$ and $\neg x_3, C_1$. Clearly, $T(\phi)$ and $C(\phi)$ can be obtained in polynomial time from ϕ. An example of this reduction is shown in Fig. S11.

Lemma 1. For any vertex labeling ℓ of $T(\phi)$, $\perp$ is the *root host*.

Proof. Under the direct transmission constraint, *root host* is given by the host that labels the root node of the timed phylogeny. The time stamp of the root node of $T(\phi)$ is $\tau(r(T(\phi))) = 0$. The only host that has entry time before $\tau_e \leq 0$ is $\perp$. Therefore, for any vertex labeling we have $\ell(r(T(\phi))) = \perp$, which makes $\perp$ the *root host*. □

Lemma 2. For any variable x, either $\{(\perp, x), (x, \neg x)\} \subseteq E(S)$ or $\{(\perp, \neg x), (\neg x, x)\} \subseteq E(S)$.

Proof. For any variable x, consider the subtree $T[x]$. By construction we have, $\tau(r(T[x])) = 2\varepsilon$ and the node only has two children labeled by x and $\neg x$. From the contact map we know that the only possible infectors for x has $\perp$ and $\neg x$ and similarly for $\neg x$ are $\perp$ and x. Given that $\tau_r(\perp) < \tau(r(T[x]))$, the only remaining choices for $\ell(r(T[x]))$ are x and $\neg x$.

If $\ell(r(T[x])) = x$ then we have $\{(\perp, x), (x, \neg x)\} \subseteq E(S)$ and if $\ell(r(T[x])) = \neg x$ we have $\{(\perp, \neg x), (\neg x, x)\} \subseteq E(S)$. □

Lemma 3. For any clause $c_i = (y_{i,1} \vee y_{i,2} \vee y_{i,3})$, if $(y_{i,j}, c_i) \in E(S)$ then $\ell(r(T[y_{i,j}])) = y_{i,j}$ and $\ell(r(T[y_{i,j'})) = \perp$ for $j' \neq j$.

Proof. Consider the subtree $T[y_{i,j}]$. Let us denote the node that is child of $r(T[y_{i,j}])$ and parent of the leaf of $T[y_{i,j}]$ labeled with c_i as v_j.

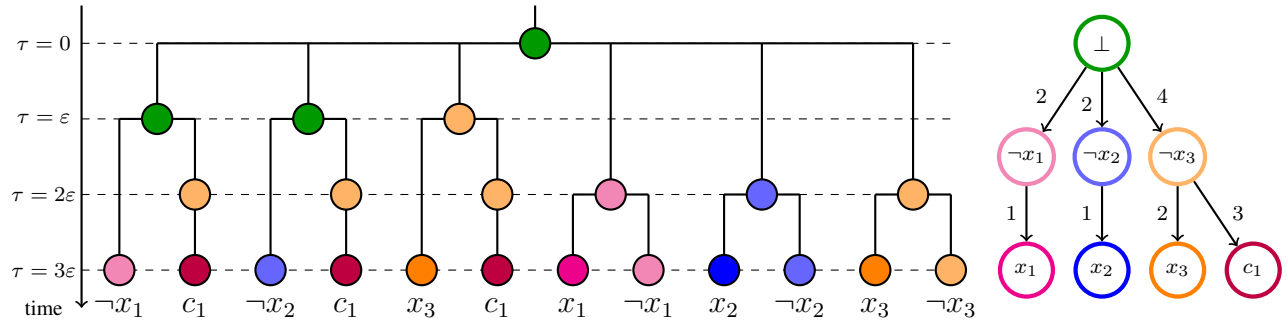


Fig. S11: **Example of reduction.** Consider the 1-in-3SAT Boolean formula $\phi = (x_1 \vee x_2 \vee \neg x_3)$. ϕ is satisfiable with truth assignment $\theta(1) = 0, \theta(2) = 0$ and $\theta(3) = 0$. Figures (on the left) shows a vertex labeling ℓ corresponding to θ. Since the vertex labeling admits a transmission tree (one the right), ϕ is Exactly-1 satisfied with truth assignment θ.

Since S is a spanning arborescence of $C(\phi)$ we have either $(y_{i,1}, c_i)$, $(y_{i,2}, c_i)$ or $(y_{i,3}, c_i)$ in $E(S)$. Without loss of generality, let us assume that $(y_{i,1}, c_i) \in E(S)$.

The edges $(v_1, \delta_T(v_1))$, $(v_2, \delta_T(v_2))$ and $(v_3, \delta_T(v_3))$ need to be transmission edges since $\tau(v_1) = \tau(v_2) = \tau(v_3) < \tau_e(c_i)$. Since $(y_{i,1}, c_i) \in E(S)$, we require $\ell(v_1) = \ell(v_2) = \ell(v_3) = y_{i,1}$. Looking at $r(T[y_{i,2}])$ and $r(T[y_{i,3}])$, since each clause consists of distinct variables, we can only have $\ell(r(T[y_{i,2}])) = \ell(r(T[y_{i,3}])) = \bot$. Consequently, the transmission edges $(r(T[y_{i,2}]), v_2)$ and $(r(T[y_{i,3}]), v_3)$ results in a edge $(\bot, y_{i,1})$ in $E(S)$. By Lemma 2, this also means $(y_{i,1}, \neg y_{i,1}) \in E(S)$ and therefore $\ell(r(T[y_{i,1}])) = y_{i,1}$. □

Lemma 4. For any literal $y_{i,j}$ in clause c_i, $(\bot, y_{i,j}) \in E(S)$ if and only if $(y_{i,j}, c_i) \in E(S)$.

Proof. Consider the subtree $T[y_{i,j}]$. Let us denote the node that is child of $r(T[y_{i,j}])$ and parent of the leaf of $T[y_{i,j}]$ labeled with c_i as v.

($\Rightarrow$) If $(\bot, y_{i,j}) \in E(S)$, then by Lemma 2 we know that $(y_{i,j}, \neg y_{i,j}) \in E(S)$. Therefore, $\ell(r(T[y_{i,j}])) = y_{i,j}$. Given that $\ell(r(T[y_{i,j}])) = y_{i,j}$, $\ell(\delta_T(v)) = c_i$ and $\tau(v) = \varepsilon$, the only feasible label for v is $y_{i,j}$. Therefore $\ell(v) = y_{i,j}$ and $(y_{i,j}, c_i) \in E(s)$.

($\Leftarrow$) If $(y_{i,j}, c_i) \in E(S)$, then since $\tau(v) < \tau_e(c_i)$, we have $\ell(v) = y_{i,j}$. From Lemma 3 we know that $\ell(r(T[y_{i,j}]))$ is either $\bot$ or $y_{i,j}$. If $\ell(r(T[y_{i,j}])) = \bot$, then we will have $\{(\bot, y_{i,j}), (\bot, \neg y_{i,j})\}$ which is not possible due to Lemma 2. Therefore $\ell(r(T[y_{i,j}])) = y_{i,j}$ and consequently $(\bot, y_{i,j}) \in E(S)$. □

Proposition 2. There exists a vertex labeling ℓ of $T(\phi)$ under the direct transmission constraint such that the corresponding transmission tree $S(\ell)$ is a spanning arborescence of $C(\phi)$ if and only if ϕ is satisfiable with a truth assignment θ so that each clause has exactly one true literal.

Proof. ($\Rightarrow$) Let ℓ be a vertex labeling of $T(\phi)$ under the direct transmission constraint such that the corresponding transmission tree S is a spanning arborescence of $C(\phi)$. We construct the corresponding truth assignment θ for ϕ as follows. From Lemma 2 we know that for any variable x, either $(\bot, x) \in E(S)$ or $(\bot, \neg x) \in E(S)$. We set $\theta(i) = 1$ if $(\bot, x_i) \in E(S)$ and $\theta(i) = 0$ if $(\bot, \neg x_i) \in E(S)$. We claim that the this truth assignment satisfies ϕ with exactly one literal for each clause.

We need to show that, for any clause $c_i = (y_{i,1} \vee y_{i,2} \vee y_{i,3})$, exactly one of $(\bot, y_{i,1})$, $(\bot, y_{i,2})$ and $(\bot, y_{i,3})$ is in $E(S)$. From Lemma 4 we know that $(\bot, y_{i,j}) \in E(S)$ if and only if $(y_{i,j}, c_i) \in E(S)$. Since S is a spanning arborescence, exactly one of $(y(i,1), c_i)$, $(y_{i,2}, c_i)$ and $(y_{i,3}, c_i)$ is in $E(S)$. Therefore, exactly one of $(\bot, y_{i,1})$, $(\bot, y_{i,2})$ and $(\bot, y_{i,3})$ is in $E(S)$ which renders the clause c_i satisfied with exactly one literal.

($\Leftarrow$) Consider the truth assignment θ that satisfies ϕ with exactly one literal for each clause in ϕ. We build the vertex labeling ℓ for $T(\phi)$ as follows. From Lemma 1 it is clear that $\bot$ is the root host and therefore $r(S) = \bot$. We set $\ell(T[x_i]) = x_i$ if $\theta(i) = 1$ and $\ell(T[x_i]) = \neg x_i$ if $\theta(i) = 0$. For any clause c_i in ϕ, if $y_{i,j}$ is true we set $\ell(r(T[y_{i,j}])) = y_{i,j}$ and if $\neg y_{i,j}$ is true we set $\ell(r(T[y_{i,j}])) = \bot$. Finally, we set $\ell(v_{i,j}) = y_{i,j}$ for all $j \in \{1, 2, 3\}$. We need to show that constructed vertex labeling satisfies the direct transmission constraint and that the resulting transmission tree is a spanning arborescence of the contact map $C(\phi)$. We do this by first showing that (i) each variable has a unique infector and (ii) all transmission edges between the same pair of hosts have time intervals that overlap.

Consider all the variables that are assigned true by the truth assignment. The infector for all these variables is $\bot$ since $\ell(r(T(\phi))) = \bot$ and $\ell(T[x_i]) = x_i$ if $\theta = 1$ and $\ell(r(T[y_{i,j}])) = \bot$ if $\neg y_{i,j}$ is true. This agrees with $C(\phi)$. The time intervals of the outgoing edges from $r(T(\phi))$ and $r(T[y_{i,j}]), \forall i \in [k], j \in \{1, 2, 3\}$ contain $\tau = \varepsilon$. Therefore, all possible transmission edges from $\bot$ overlap at $\tau = \varepsilon$.

Consider the variables that are assigned false by the truth assignment. From Lemma 2 we know that for any such variable x, they are infected by $\neg x$. This agrees with $C(\phi)$. Moreover, these variables do not label any of the interval vertices of the tree T and all the leaves of T are at the same time-stamp $\tau = 3\varepsilon$. Therefore, all possible transmission edges to any such variable x overlap at $\tau = 3\varepsilon$.

Finally, consider any clause c_i. All the internal vertices $v_{i,j}, j \in \{1, 2, 3\}$ are labeled by the same variable $y_{i,j}$ that renders the clause c_i satisfied. As a result, $y_{i,j}$ is a unique infector of c_i and $(y_{i,j}, c)$ exists in $E(C(\phi))$ by construction. Also, time-stamp of all vertices $v_{i,j}$ are the same $\tau = 2\varepsilon$ and therefore, the transmission edges overlap at $\tau = 2\varepsilon$. □

A.2.2 Counting Problem

This section proves the #P-completeness of the #DTI problem.

Proposition 3. There exists a parsimonious reduction from #1-in-3SAT to #DTI.

Proof. Consider the reduction shown in Section 4. Here we show that this reduction is parsimonious, i.e. it preserves the number of solutions in the solution spaces of the two problems. We show a bijection between the solution space of a 1-in-3SAT and the solution space of the corresponding DTI instance.

Consider the Boolean formula ϕ. For a given truth assignment θ that satisfies each clause of ϕ with exactly one true literal, we construct the vertex labeling of $T(\phi)$ as following. We let $\ell(T[x_i]) = x_i$ if $\theta(i) = 1$ and $\ell(T[x_i]) = \neg x_i$ if $\theta(i) = 0$. We will show that this unique determines the labeling for the rest of the internal vertices of $T(\phi)$. Consider the clause c_i and the corresponding subtrees $T[y_{i,1}], T[y_{i,2}]$ and $T[y_{i,3}]$. Since the truth assignment satisfies each clause with exactly one literal, without loss generality, assume that $y_{i,1}$ is true. Then using Lemma 4, since $(\perp, y_{i,j}) \in E(S)$, we have $(y_{i,j}, c_i) \in E(S)$. For the nodes $v_{i,j}$ we have $\tau(v_{i,j}) < \tau_e(c_i)$ and therefore $\ell(v_{i,j}) = y_{i,j}, \forall j \in \{1, 2, 3\}$. Finally, the vertex labels for the roots of the clause subtrees $\ell(r(T[y_{i,1}])) = \ell(r(T[y_{i,2}])) = \ell(r(T[y_{i,3}])) = y_{i,1}$ due to Lemma 3. Proof of Proposition 2 shows that this vertex labeling is a solution of the DTI problem.

From a given vertex labeling ℓ, we construct the truth assignment as follows. We set $\theta(i) = 1$ if $\ell(r(T[x_i])) = x_i$ and $\theta(i) = 0$ if $\ell(r(T[x_i])) = \neg x_i$. Proof of Proposition 2 shows that this is a truth assignment that satisfies each clause with exactly one true literal.

The construction of θ from ℓ and ℓ from θ are inverses of each other. If we view these constructions as functions then they show a bijection in the solutions spaces of #1-in-3SAT and #DTI. This shows that the number of solutions is preserved. Obviously, the reduction can be performed in polynomial time. Therefore, the reduction is parsimonious. □

A.3 Naive Rejection Sampling Algorithm

Here we descirbe the naive rejection sampling algorithm introduced in Section 5.2.1. Let $h[v, s]$ denote the number of vertex labelings $\ell \in \mathcal{L}_{\text{REL}}$ in the subtree T_v of T rooted at vertex v when $\ell(v) = s$. We define $h[v, s]$ recursively as

$$\begin{cases} 1, & \text{if } v \in L(T),\ \hat{\ell}(v) = s, \\ 0, & \text{if } v \in L(T),\ \hat{\ell}(v) \neq s, \\ 0, & \text{if } v \notin L(T), \tau(v) \notin I(s), \\ \displaystyle\prod_{w \in \delta_T(v)} \sum_{t \in \Gamma_C(s)} h[w, t], & \text{if } v \notin L(T), \tau(v) \in I(s), \end{cases}$$

where $I(s) = [\tau_e(s), \tau_r(s)]$ and $\Gamma_C(s) = \{s, \delta_C(s)\}$. Let $\Sigma^* = \{s_1, \ldots, s_k\}$ be the set of possible labels for the root vertex $r(T)$, *i.e.* $\Sigma^* = \{s \in \Sigma \mid \tau(r(T)) \in I(s)\}$. The number of vertex labelings $|\mathcal{L}_{\text{REL}}|$ is given by $\sum_{s' \in \Sigma^*} h[r(T), s']$.

Using the count matrix $h[u, s]$, we introduce a subroutine that takes a vertex v and host s as input, and uniformly samples a vertex labeling ℓ_u of subtree T_u rooted at u subject to the restriction that $\ell_u(u) = s$ (Algorithm 3). The fraction p_s of the vertex labelings ℓ where $\ell(r(T)) = s$ equals $h[r(T), s] / \sum_{s' \in \Sigma^*} h[r(T), s']$. Thus, to sample *all* vertex labelings uniformly at random, we draw a $s \in \Sigma^*$ according to the categorical probability distribution defined by $(p_1, \ldots, p_k)$. Algorithm 4 is then used on T with $\ell(r(T)) = s$ to sample minimum transmission host labeling ℓ of T uniformly at random. This takes $O(nm)$ time per sample.

For a given phylogeny and vertex labeling (T, ℓ), it is possible to find the minimum number of transmission events in polynomial time (Sashittal and El-Kebir, 2019). The *direct transmission constraint* is satisfied by the vertex labeling when the number of transmission events is $m - 1$, where each transmission event corresponds to an edge of the transmission tree. We can therefore draw vertex labelings from $\mathcal{L}_{\text{REL}}$ and only retain the solutions that belong to $\mathcal{L}$ in polynomial time. Since we are uniformly sampling from $\mathcal{L}_{\text{REL}}$, the retained solutions will also be uniformly sampled from $\mathcal{L}$. For the counting problem we estimate the number of vertex labelings in $\mathcal{L}$ by the success rate of the sampling algorithm. Say after K draws of samples from $\mathcal{L}_{\text{REL}}$, we retain K' vertex labelings that belongs to $\mathcal{L}$. In that case the estimate of the size of $\mathcal{L}$, denote by $\langle|\mathcal{L}|\rangle$, is given by

$$\langle|\mathcal{L}|\rangle = \left(1 - \frac{K'}{K}\right)^{1/K}$$

From the law of large numbers, as $K \to \infty$ we have $\langle|\mathcal{L}|\rangle \to |\mathcal{L}|$. We now present the algorithms for naive rejection based sampling.

Algorithm 1 EnumRelDTI($T, \hat{\ell}, u, s$)

Output: Set $\mathcal{L}_u$ of vertex labelings ℓ of T_u where $\ell(u) = s$

if $u \in L(T)$ **then**
  Let s be the unique host where $\hat{\ell}(u) = s$
  return $\{\{(u, s)\}\}$
else
  Let $v_1, \ldots, v_k$ be the children of v
  $\mathcal{L}_1, \ldots, \mathcal{L}_k \leftarrow \emptyset, \ldots, \emptyset$
  for $v \in \{v_1, \ldots, v_k\}$ **do**
    for $t \in \Gamma((u, v), s)$ **do**
      $\mathcal{L}_v \leftarrow \mathcal{L}_v \cup$ EnumRelDTI(T, g, v, t)
    end for
  end for
  $\mathcal{L}_u \leftarrow \emptyset$
  for $\ell_1, \ldots, \ell_k \in \mathcal{L}_1 \times \ldots \times \mathcal{L}_k$ **do**
    $\mathcal{L}_u \leftarrow \mathcal{L}_u \cup \{\ell_1 \cup \ldots \cup \ell_k \cup \{(u, s)\}\}$
  end for
  return $\mathcal{L}_u$
end if

Algorithm 2 EnumRelDTI(T, g)

Output: Set $\mathcal{L}$ of optimal host labelings ℓ of T

Let Σ^* be the set of hosts s where $\tau(r(T)) \in I(s)$
$\mathcal{L} \leftarrow \emptyset$
for $s \in \Sigma^*$ **do**
  $\mathcal{L} \leftarrow \mathcal{L} \cup$ EnumRelDTI$(T, \hat{\ell}, r(T), s)$
end for
return $\mathcal{L}$

Algorithm 3 SampleRelDTI(T, h, u, s)

Output: Random, optimal host labeling ℓ of T_u where $\ell(u) = s$

Let $\delta_T(u) = \{v_1, \ldots, v_k\}$ be the children of u
for $v \in \{v_1, \ldots, v_k\}$ **do**
  $K \leftarrow \sum_{t \in \Gamma_C(s)} h[v, t]$
  for $t \in \Sigma = \{1, \ldots, m\}$ **do**
    if $t \in \Gamma_C(s)$ **then**
      $p(t) \leftarrow h[v, t]/K$
    else
      $p(t) \leftarrow 0$
    end if
  end for
  Draw host $t^* \in \Sigma$ randomly according to $(p_1, \ldots, p_m)$
  $\ell_v \leftarrow$ SampleRelDTI(T, g, h, v, t^*)
  for $w \in V(T_v)$ **do**
    $\ell(w) \leftarrow \ell_v(w)$
  end for
end for
$\ell(u) \leftarrow s$
return ℓ

Algorithm 4 SampleRelDTI(T, h)

Output: Random, optimal host labeling ℓ of T

Let Σ^* be the set of hosts s where $\tau(r(T)) \in I(s)$
$K \leftarrow \sum_{s \in \Sigma^*} h[r(T), s]$
for $s \in \Sigma$ **do**
  if $s \in \Sigma^*$ **then**
    $p_s \leftarrow h[r(T), s]/K$
  else
    $p_s \leftarrow 0$
  end if
end for
Draw $s^* \in \Sigma$ according to probabilities $p_1, \ldots, p_m$
return SampleRelDTI($T, h, r(T), s^*$)

A.4 Consensus Transmission Tree Algorithm Proof

Theorem 4. Given a set $\mathcal{S} = \{S_1, \cdots, S_k\}$ of k transmission trees with edge weights $w_{S_1}, \cdots, w_{S_k}$, the minimum weight spanning arborescence of the corresponding weighted parent-child graph P defines a tree R that is a solution to the SCTT problem with the distance measure used is weighted parent-child distance.

Proof. Consider the *weighted parent-child graph* P for the set of transmission trees $\mathcal{S}$. Since P is a complete graph, the optimal consensus tree R is necessarily a spanning arborescence of P. The weights of the edges in R are given by w^* due to Proposition 1. The total WPCD of R from the set of transmission trees $\mathcal{S}$ is given by $d(R, \mathcal{S}) = \sum_{S_i \in \mathcal{S}} d(R, S_i)$ where

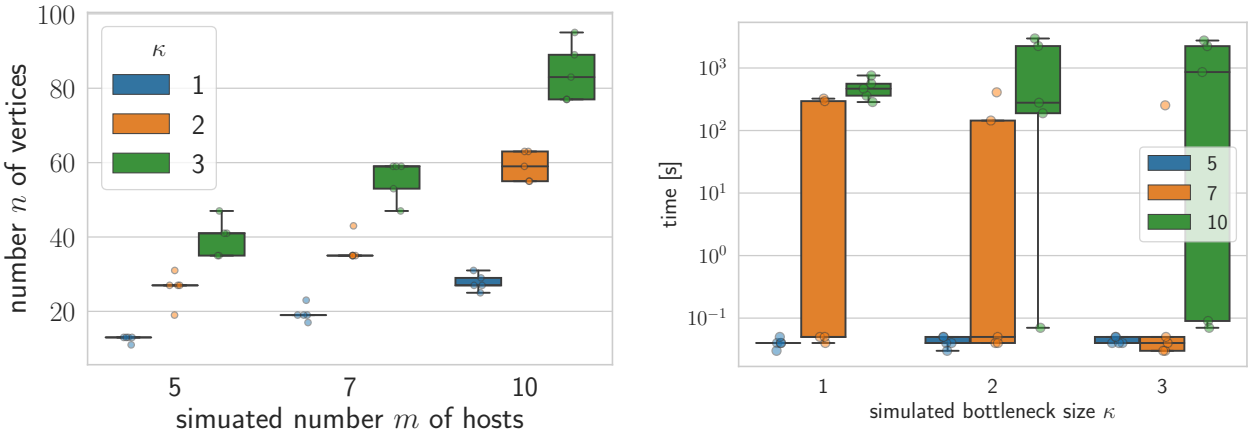
$$\begin{aligned} d(R, S_i) &= \sum_{(s,t) \in E(R)} |q_i(s,t) - w^*(s,t)| + \sum_{s,t \notin E(R)} |q_i(s,t)| \\ &= \sum_{s,t \in E(R)} (|q_i(s,t) - w^*(s,t)| - |q_i(s,t)|) + \\ &\quad \sum_{s,t \in \Sigma \times \Sigma} |q_i(s,t)|. \end{aligned}$$

Consequently,

$$d(R, \mathcal{S}) = \sum_{S_i \in \mathcal{S}} \sum_{s,t \in \Sigma \times \Sigma} |q_i(s,t)| + \sum_{s,t \in E(R)} w_P(s,t),$$

where the first term is a constant with respect to R and minimizing the second term is equivalent to finding the minimum weight spanning arborescence of P. □

A.5 Additional Simulation Results



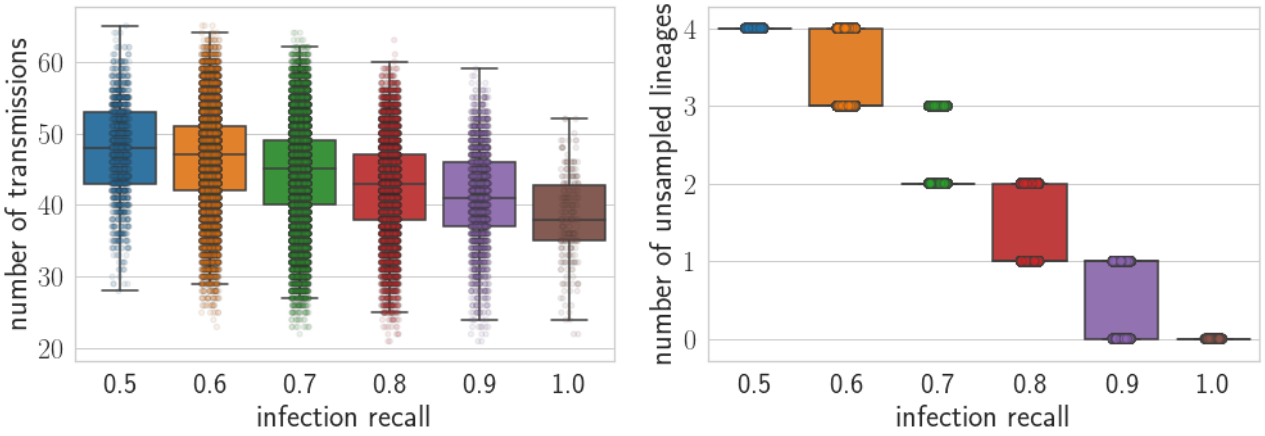
(a) (b)

Fig. S12: (a) The number of vertices n in the timed phylogeny T for increasing number m of simulated hosts and bottleneck size κ. (b) Time taken to generate 100,000 uniformly sampled solutions to the DTI problem using TiTUS for increasing values of simulated bottleneck size κ.

A.6 Additional HIV Data Analysis and Implementation Details

host	transmission window	known infector	latest sample time	entry time	removal time
A	? - 14/05/90	B	7/11/05	$\tau(r(T))$	7/11/05
F	01/02/95 - 02/08/95	A	19/09/05	01/02/95	19/09/05
G	16/01/02 - 16/04/02	F	16/04/02	16/01/02	16/04/02
H	29/06/95 - 24/07/95	B	25/05/98	29/06/95	25/05/98
I	01/02/93 - 28/04/93	B	06/10/99	01/02/93	06/10/99
C	23/09/93 - 10/01/94	B	15/12/03	23/09/93	15/12/03
D	16/03/95 - 01/07/95	C	24/03/03	16/03/95	24/03/03
L	23/09/93 - 12/03/06	C	24/03/06	23/09/93	24/03/06
E	15/06/00 - 01/02/01	C	22/02/06	15/06/00	22/02/06
K	01/06/04 - 15/09/04	E	30/09/04	01/06/04	30/09/04

Table 1. This table shows the epidemiological information provided in the HIV dataset (Vrancken et al., 2014). The transmission window of a host is the expected time-interval during which the host was infected.



(a) (b)

Fig. S13: (a) Transmission number and (b) number of unsampled lineages of all the solutions generated using TiTUS on the HIV dataset vs different infection recall values.

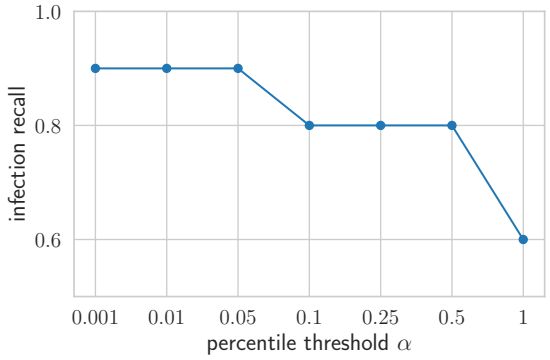


Fig. S14: The infection recall of the consensus transmission tree for solutions sampled using TiTUS on the HIV dataset for increasing values of the percentile threshold α.