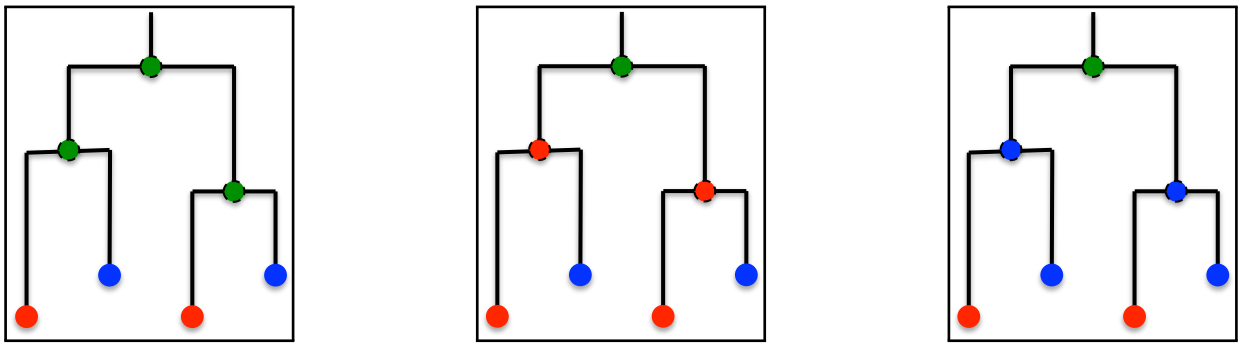
Supplementary Material — Sampling and Summarizing Transmission Trees with Multi-strain Infections

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

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

Contents

*To whom correspondence should be addressed.



(a) (b) (c)

Figure S1: **The timed phylogeny shown in Main Text Fig. 1a has 3 possible vertex labeling solutions.**

A Background and Theory

In this section we provide the information we could not include in the main text. Fig. S1 shows all the feasible solutions to the representative DTI problem described in the Main Text Fig. 1.

A.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,

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

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.

$$d(S_1, S_3) = \sum_{(s,t)\in\Sigma\times\Sigma} |q_1(s,t) - q_3(s,t)|$$

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

□

A.2 Sampling Scenarios

The weak transmission bottleneck has some interesting implications for the sampling of the within-host diversity of the infected hosts. Fig. S2 gives an overview, with schematic representations, of 4 different scenarios that can occur for real outbreaks.

B Complexity

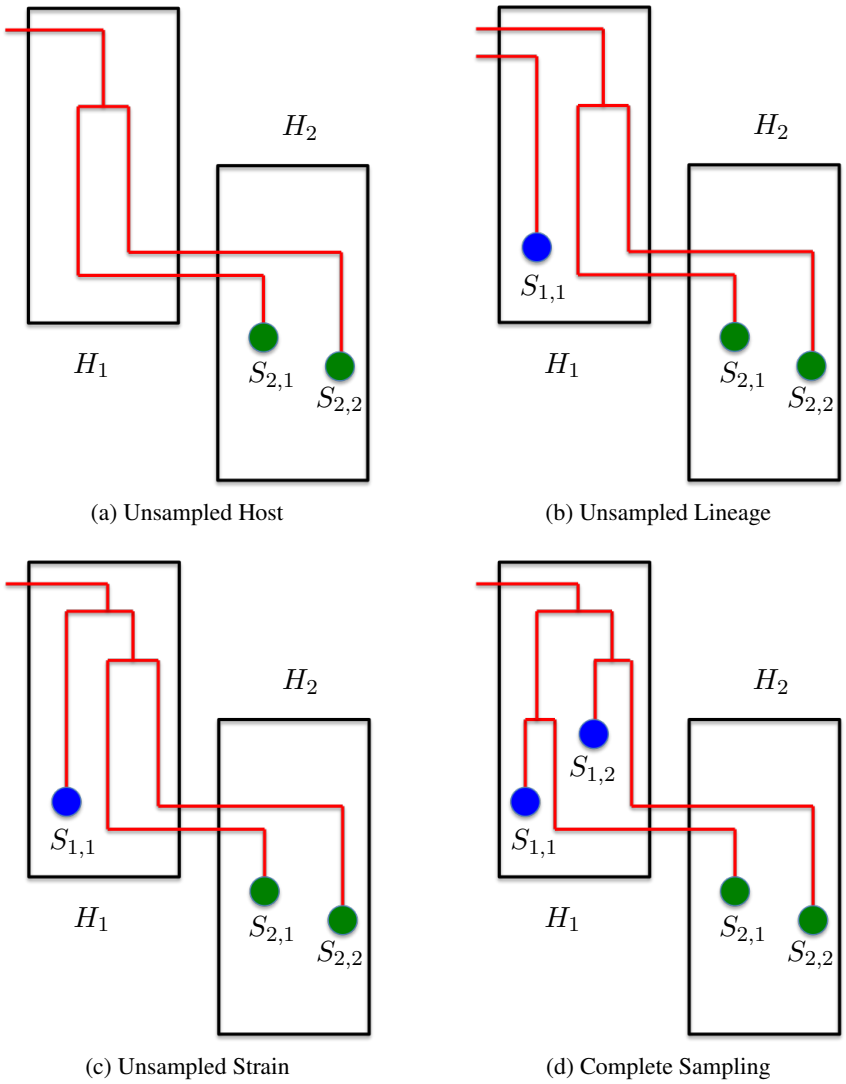
This section shows the hardness of the decision and the counting versions of the DTI problem by reduction from the one-in-three SAT (1-in-3 SAT).

Problem 1 (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).

B.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. S3 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 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 $c_1 \coloneqq (x_1 \vee x_2 \vee \neg x_3)$, 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. S5.



(a) Unsampled Host

(b) Unsampled Lineage

(c) Unsampled Strain

(d) Complete Sampling

Figure S2: **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.

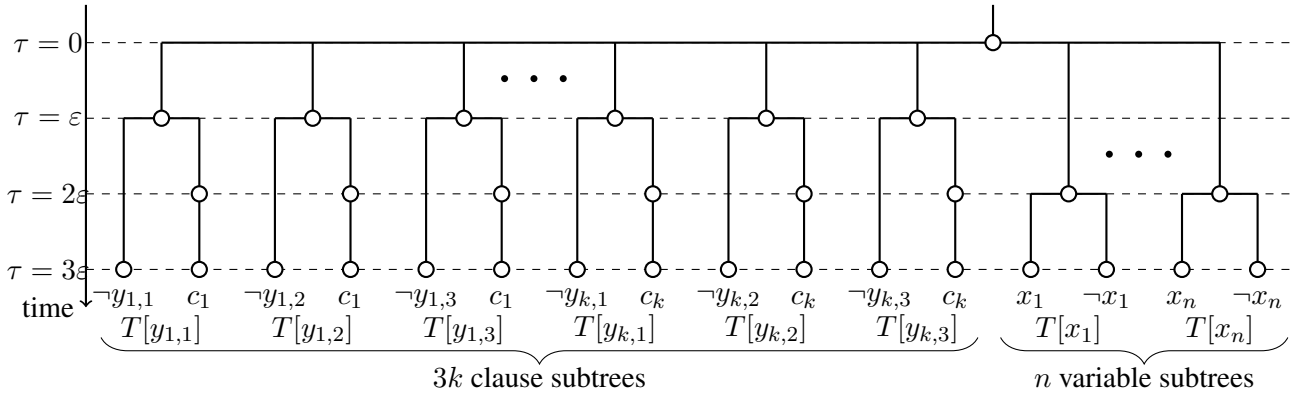


Figure S3: **Construction of $T(\phi)$ for reduction from 1-in-3SAT to DTI.** Let ϕ be an 1-in-3SAT formula with k clauses and n variables. $T(\phi)$ is built with a root node $r(T(\phi))$ can is connected to $3k$ clause subtrees $\{T[y_{1,1}], T[y_{1,2}], T[y_{1,3}], \cdots, T[y_{k,1}], T[y_{k,2}], T[y_{k,3}]\}$ 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. S4).

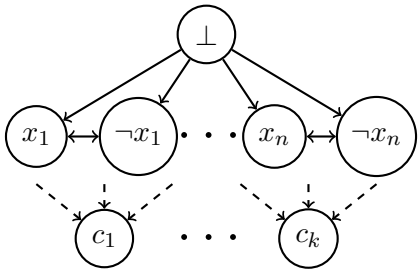


Figure S4: **Construction of $C(\phi)$ for reduction from 1-in-3SAT to DTI.** Let ϕ be an 1-in-3SAT formula with k clauses and n variables. The host set is $\Sigma = \{\perp, x_1, \cdots, x_n, \neg x_1, \cdots, \neg x_n, c_1, \cdots, c_k\}$. We have a directed edge from $\perp$ to each of the variables $\{x_1, \cdots, x_n, \neg x_1, \cdots, \neg x_n\}$. 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 $y_{i,3}$.

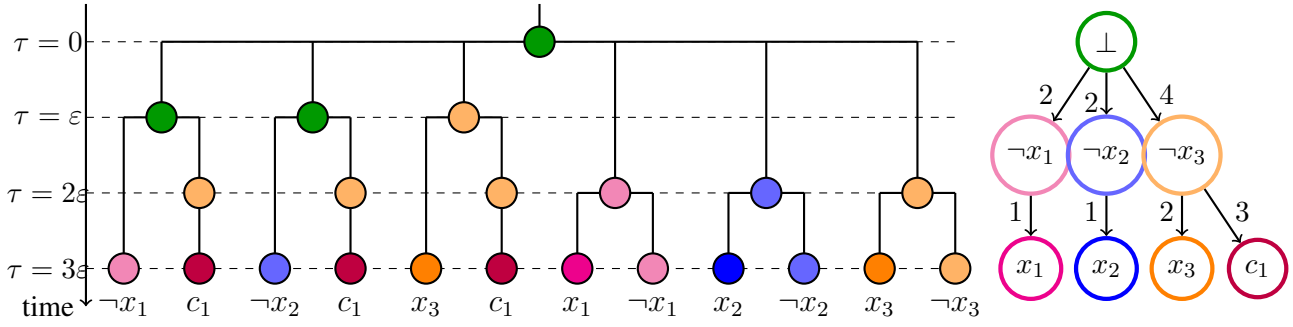


Figure S5: **Example of reduction.** Consider the 1-in-3SAT Boolean formula $\phi = (x_1 \vee x_2 \vee \neg x_3)$. ϕ is satisfiable with truth assignment $\theta(1) = 0, \theta(2) = 0$ and $\theta(3) = 0$. Figures (on the left) shows a vertex labeling ℓ corresponding to θ. Since the vertex labeling admits a transmission tree (one the right), ϕ is Exactly-1 satisfied with truth assignment θ.

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.

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 $\ell(r(T[y_{i,2}])) = \ell(r(T[y_{i,3}])) = \perp$. Consequently, the transmission edges $(r(T[y_{i,2}]), v_2)$ and $(r(T[y_{i,3}]), v_3)$ results in a edge $(\perp, y_{i,1})$ 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, $(\perp, y_{i,j}) \in E(S)$ 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 $(\perp, y_{i,j}) \in E(S)$, 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 $\perp$ or $y_{i,j}$. If $\ell(r(T[y_{i,j}])) = \perp$, then we will have $\{(\perp, y_{i,j}), (\perp, \neg y_{i,j})\}$ which is not possible due to Lemma 2. Therefore $\ell(r(T[y_{i,j}])) = y_{i,j}$ and consequently $(\perp, y_{i,j}) \in E(S)$. □

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 $(\perp, x) \in E(S)$ or $(\perp, \neg x) \in E(S)$. We set $\theta(i) = 1$ if $(\perp, x_i) \in E(S)$ and $\theta(i) = 0$ if $(\perp, \neg x_i) \in E(S)$. 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 $(\perp, y_{i,1}), (\perp, y_{i,2})$ and $(\perp, y_{i,3})$ is in $E(S)$. From Lemma 4 we know that $(\perp, y_{i,j}) \in E(S)$ 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 $(\perp, y_{i,1}), (\perp, y_{i,2})$ and $(\perp, y_{i,3})$ 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 $\perp$ is the root host and therefore $r(S) = \perp$. 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 $\ell(r(T[y_{i,j}])) = \perp$. 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 $\perp$ since $\ell(r(T(\phi))) = \perp$ and $\ell(T[x_i]) = x_i$ if $\theta = 1$ and $\ell(r(T[y_{i,j}])) = \perp$ 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 $\perp$ 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 $\tau = 2\varepsilon$. □

B.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 B. 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. □

C Naive Rejection Sampling Algorithm

Here we describe the naive rejection sampling algorithm introduced in Main Text Section 5.2.1. Let $h[v, s]$ denote the number of vertex labelings $\ell \in \mathcal{L}_{\mathrm{REL}}$ 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 $|\mathcal{L}_{\mathrm{REL}}|$ 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 $\mathcal{L}_{\mathrm{REL}}$ and only retain the solutions that belong to $\mathcal{L}$ in polynomial time. Since we are uniformly sampling from $\mathcal{L}_{\mathrm{REL}}$, 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 $\mathcal{L}_{\mathrm{REL}}$, 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$(T, h, r(T), s^*)$

D Consensus Transmission Tree Algorithm Proof

Theorem 1. 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^* (Main Text Lemma 1).

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

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

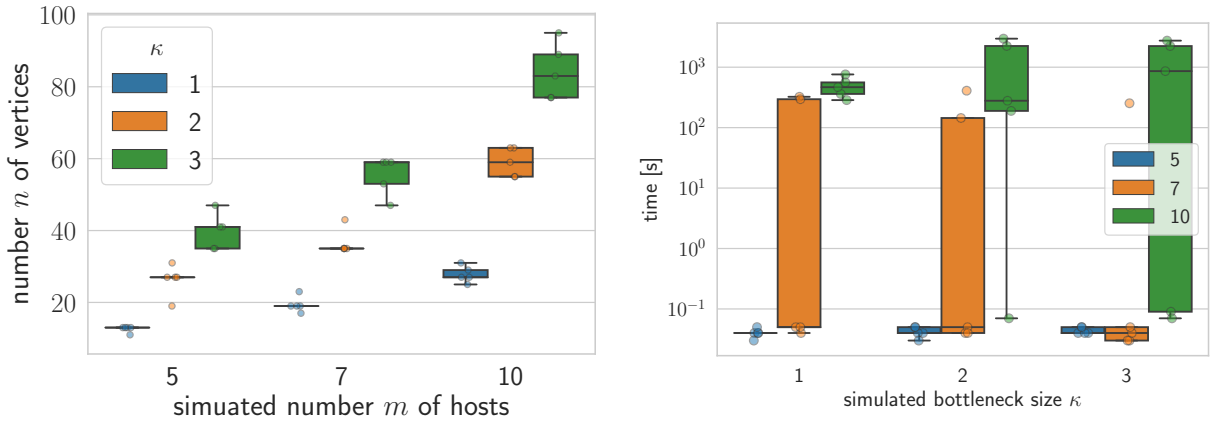
$$\begin{aligned} d(R, S_i) &= \sum_{(s,t) \in E(R)} |q_i(s,t) - w^*(s,t)| + \sum_{(s,t) \notin E(R)} |q_i(s,t)| \\ &= \sum_{(s,t) \in E(R)} (|q_i(s,t) - w^*(s,t)| - |q_i(s,t)|) + \sum_{(s,t) \in \Sigma \times \Sigma} |q_i(s,t)|. \end{aligned}$$

Consequently,

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

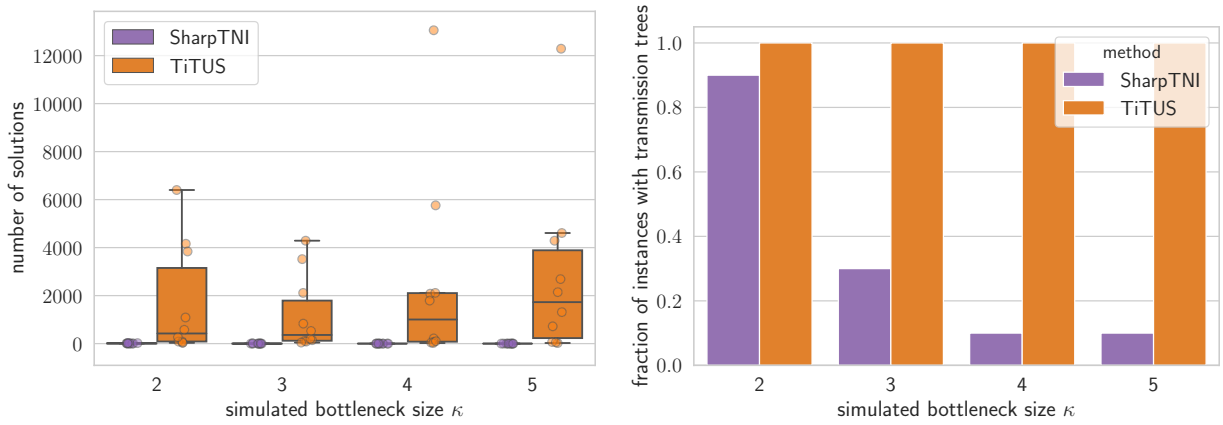
where the first term is a constant with respect to R and minimizing the second term is the sum of the weights of a minimum weight spanning arborescence R of P. □

E Additional Simulation Results



(a) (b)

Figure S6: (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.



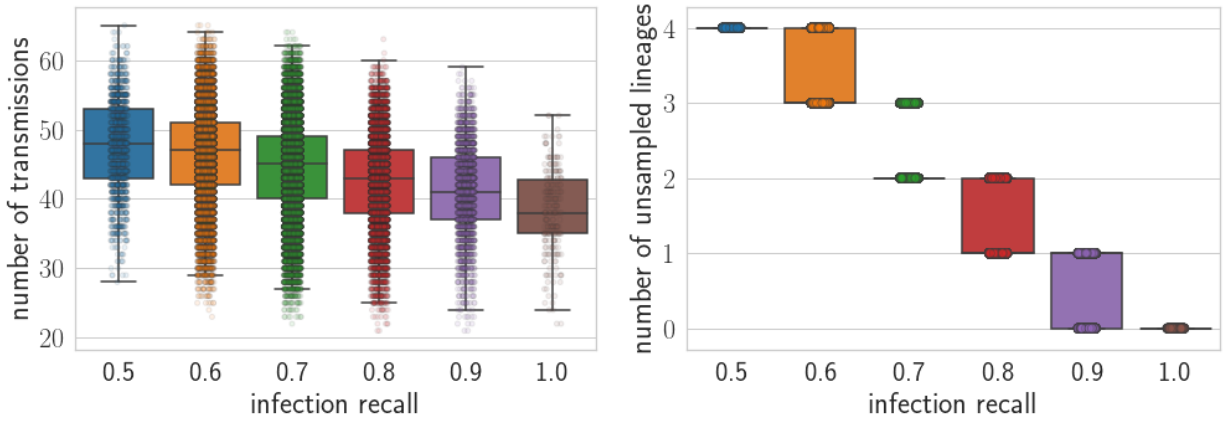
(a) (b)

Figure S7: **Comparison of SharpTNI and TiTUS on simulated, partially sampled outbreaks.** Transmissions in simulated instances followed a tree-like pattern (*i.e.* direct transmission) but not all strains within a host were sampled. (a) The number of solutions where the vertex labeling induces a transmission tree (rather than a general graph) for increasing bottleneck size κ. (b) The fraction of simulation instances for which each method identified a transmission tree, for increasing values of simulated bottleneck size κ.

F 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 S1: 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)

Figure S8: (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.

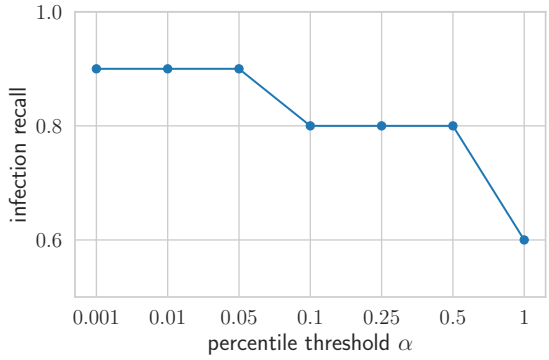


Figure S9: The infection recall of the consensus transmission tree for solutions sampled using TITUS on the HIV dataset for increasing values of the percentile threshold α.

References

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

Vrancken, B. *et al.* (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).