Procedures for enumerating and uniformly sampling transmission trees for a known phylogeny

Matthew Hall[1]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK

July 11, 2017

Abstract

One approach to the reconstruction of infectious disease transmission trees from pathogen genomic data has been to annotate the internal nodes of a phylogeny with information about the host that each ancestral lineage was infecting. If the transmission bottleneck is complete, the set of all possible ways of making this annotation is equivalent to the set of partitions of the nodes of the phylogeny such that the nodes in each partition element induce a connected subgraph of the tree. However, the mathematical properties of this space remain largely unexplored. Here, a procedure by which the cardinality of the set of partitions for a given phylogeny can be calculated is described, and also I show how to uniformly sample from that set. The procedure is outlined, first, for situations where one sample is available from each host and trees do not have branch lengths, and it is then extended to allow incomplete sampling, multiple sampling, and the application to time trees in a situation where limits on the period during which each host could have been infected are known. The sampling algorithm is available as an `R` script.

1 Introduction

The use of genetic data to reconstruction pathogen transmission trees has been the subject of considerable interest in recent years. Many different approaches have been proposed, both phylogenetic and non-phylogenetic [1, 5, 10]. The phylogenetic approaches can broadly be divided into two categories: those that assume that internal nodes in the phylogenetic tree correspond to transmission events [7–9], and those that do not [2–4, 6]. In the former case, the phylogeny and the transmission tree are effectively the same object.

The assumption of coinciding lineage coalescences and transmission events may be unwise, and in particular it does not take into account within-host pathogen diversity [11]. Several approaches have been taken that do not make it, one of which to note that if a phylogeny from a completely sampled outbreak has its nodes annotated with the hosts in which each lineage was present, the transmission tree is known [2–4]. In particular, Hall, Woolhouse, and Rambaut [4] noted that the set of transmission trees for a known phylogeny, with complete sampling and assuming transmission is a complete bottleneck, is equivalent to the set of partitions of its nodes such that each partition element contains at least one tip and the subgraph induced by the nodes in each partition element is connected. However, for the most part the mathematical properties of this space of partitions remain unexplored.

Here, I provide procedures for counting the total number of partitions (and hence the total number of transmission trees) for a known phylogeny. I also show how an algorithm can be written to sample uniformly from the set of partitions. Initially I assume that the phylogeny is binary, sampling is complete, that each host provided one sample, and that nothing is known about the timings of each infection, but I go on to individually relax each of these assumptions. The procedures outlined here may be useful to researchers wishing to explore the structure that the phylogeny imposes on transmission tree space. An R implementation of the algorithms described is available at `http://github.com/twoseventwo/TTsampler`.

2 Complete, single sampling

2.1 Preliminaries

Let the phylogeny $\mathcal{T}$ be an unlabelled rooted binary tree, initially without branch lengths. Let $\mathcal{T}^*$ represent the *unrooted* binary tree obtained from $\mathcal{T}$ by attaching a single extra tip to the root of $\mathcal{T}$ by a single edge. (Note that two distinct $\mathcal{T}$s can have the same $\mathcal{T}^*$.) $\mathcal{T}^*$, importantly, has one more tip than $\mathcal{T}$.

I follow the correspondence described by Hall, Woolhouse, and Rambaut [4] between transmission trees and partitions of the node set of $\mathcal{T}$ such that all tips derived from the same host are members of the same partition element and the subgraph induced by each partition element is connected. This assumes that sampling is complete, which I later relax, and that transmission is a complete bottleneck, which is a more fundamental assumption. Furthermore, I assume for now that only one tip is derived from each host. See figure 1 for an example.

In what follows, "subtree" has its normal phylogenetic meaning; a subgraph of tree $\mathcal{T}$ consisting of a node of $\mathcal{T}$, all its descendants (if any), and the edges between them. If u is a node then we will denote the subtree rooted at u by $\mathcal{T}_u$; this is defined even if u is a tip.

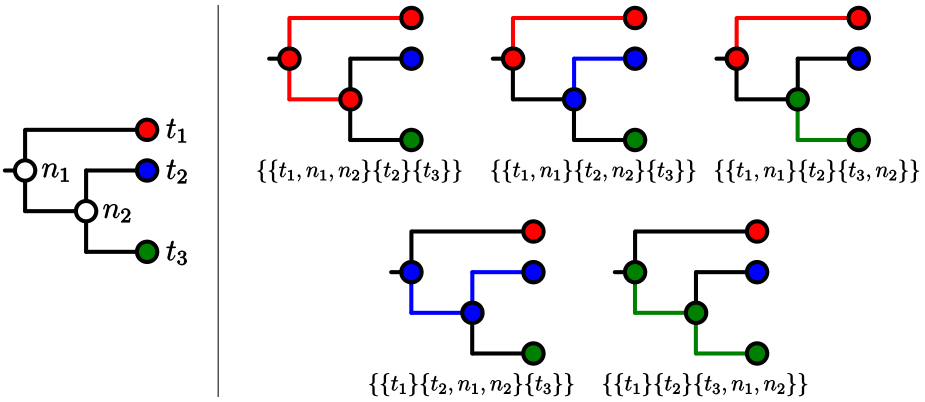


Figure 1: A rooted phylogeny (left) and the five compatible transmission trees expressed as partitions of its node set (right). Coloured branches connect members of the same partition element.

2.2 Counting transmission trees

With $\mathcal{T}$ fixed and having n tips, suppose we wish to count the number of partitions, as defined above, of its node set $N(\mathcal{T})$. If the set of such partitions is $\mathbf{P}(\mathcal{T})$ (this is a set of sets of sets), we wish to calculate $|\mathbf{P}(\mathcal{T})|$. Nothing about the definition of a partition requires a rooted tree, so $\mathbf{P}(\mathcal{T}^*)$ is defined similarly. It is trivial that if $n = 1$ then $|\mathbf{P}(\mathcal{T})| = |\mathbf{P}(\mathcal{T}^*)| = 1$.

If $\mathcal{T}_u$ is a subtree, we can $\mathbf{P}(\mathcal{T}_u)$ in the obvious way by regarding it as a tree in its own right. We need to define another set of partitions of $N(\mathcal{T}_u)$, however, which is the full set of its intersections with all the elements of each member of the full set of partitions of $N(\mathcal{T})$. This is different because it allows an internal node of $\mathcal{T}_u$ to share its partition element with no tip of $\mathcal{T}_u$, as it was constructed by taking the intersection of $N(\mathcal{T}_u)$ with an element of a partition of $N(\mathcal{T})$ that contains a tip of $\mathcal{T}$ which is not a tip of $\mathcal{T}_u$

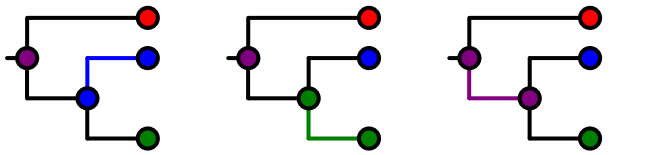
Suppose $\mathfrak{A}$ is a partition of $N(\mathcal{T})$ and there exists $A \in \mathfrak{A}$ such that $A \cap N(\mathcal{T}_u)$ is nonempty and contains no tip of $\mathcal{T}_u$. Then:

1. $u \in A \cap N(\mathcal{T}_u)$ because if it were not then the A would not obey the connectedness requirement for being an element of a partition of $N(\mathcal{T})$. If $v \in A \cap N(\mathcal{T}_u)$ and t is the tip of $\mathcal{T}$ in A then the path from v to t must intersect u.

2. $A \cap N(\mathcal{T}_u)$ is the only member of the set $\{B \cap N(\mathcal{T}_u) : B \in \mathfrak{A}\}$ that contains no tips of $\mathcal{T}_u$, because u can belong to only one member of it.

Let $\mathbf{Q}(\mathcal{T})$ be the set of partitions of $\mathcal{T}$ which allow an extra partition element containing $\mathcal{T}$'s root. Figure 2 shows an example of the extra elements of $\mathbf{Q}(\mathcal{T})$ which are not already elements of $\mathbf{P}(\mathcal{T})$ (and hence already displayed in figure 1).

To look at $\mathbf{Q}(\mathcal{T}_u)$ $(u \neq r)$ in another way, suppose $\sim$ is an equivalence relation on the elements of $\mathbf{P}(\mathcal{T})$ such that $\mathfrak{A} \sim \mathfrak{B}$ if $\mathfrak{A} \cap N(\mathcal{T}) = \mathfrak{B} \cap N(\mathcal{T})$,



$\{\{t_1\}\{t_2,n_2\}\{t_3\}\{n_1\}\}$ $\{\{t_1\}\{t_2\}\{t_3,n_2\}\{n_1\}\}$ $\{\{t_1\}\{t_2\}\{t_3\}\{n_1,n_2\}\}$

Figure 2: For the tree in figure 1, the three members of $\mathbf{Q}(\mathcal{T})$ which are not members of $\mathbf{P}(\mathcal{T})$.

where $\mathfrak{A} \cap N(\mathcal{T})$ is the set of pairwise intersections of $N(\mathcal{T})$ with elements of $\mathfrak{A} \in \mathbf{P}(\mathcal{T})$. $\mathbf{Q}(\mathcal{T}_u)$ can be seen as the set of equivalence classes. An important point to note is that even if the tip of $\mathcal{T}$ which shares a partition element with u is different in $\mathfrak{A}$ and $\mathfrak{B}$, $\mathfrak{A} \sim \mathfrak{B}$ is still possible. The partition of the nodes in $N(\mathcal{T}) \backslash N(\mathcal{T}_u)$ does not matter. $|\mathbf{Q}(\mathcal{T}_u)|$ is the number of ways of partitioning the nodes of $\mathcal{T}_u$ allowing for the possibility that that an unspecified extra partition element could "creep" down onto it from above.

The exact correspondence of $\mathbf{Q}(\mathcal{T})$ with $\mathbf{P}(\mathcal{T}^*)$ is obvious, as $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by attaching a single tip to $\mathcal{T}$'s root. Compare figure 3 with the full set of partitions displayed in figures 1 and 2 as an example.

If n is at least 2, then $\mathcal{T}$ has a left subtree $\mathcal{T}_{rL}$ rooted at the left child rL of its root node r and a right subtree $\mathcal{T}_{rR}$ rooted at the right child rR.

Proposition 2.1. *If $\mathcal{T}$ has at least two tips, then* $|\mathbf{P}(\mathcal{T})| = (|\mathbf{P}(\mathcal{T}_{rL})| \times |\mathbf{P}(\mathcal{T}^*_{rR})|) + (|\mathbf{P}(\mathcal{T}_{rR})| \times |\mathbf{P}(\mathcal{T}^*_{rL})|)$.

Proof. Since $\mathcal{T}$ is not the tree with one node, its root r is internal. First we count the number of partitions where r is in the same partition element as a tip of $\mathcal{T}_{rL}$. If this is true then rL is in that same element by the connectedness requirement for elements: if it were not then the path from that tip to r would go through a node in a different element. The connectedness requirement then also insists that no node of $\mathcal{T}_{rL}$ is in the same partition element as a tip of $\mathcal{T}_{rR}$, and the number of ways of partitioning the nodes of $\mathcal{T}_{rL}$ as a subtree of $\mathcal{T}$ such that this is true is just $|\mathbf{P}(\mathcal{T}_{rL})|$. For each of those ways, the number of ways of partitioning the nodes of $\mathcal{T}_{rR}$ is $|\mathbf{Q}(\mathcal{T}_{rR})|$, since some nodes of $\mathcal{T}_{rR}$ can be in the same element as r (and hence rL) and if any are then rR is. Since $|\mathbf{Q}(\mathcal{T}_{rR})| = |\mathbf{P}(\mathcal{T}^*_{rR})|$ the number we are looking for is hence $|\mathbf{P}(\mathcal{T}_{rL})| \times |\mathbf{P}(\mathcal{T}^*_{rR})|$.

An identical argument shows that the number of partitions where r is in the same element as a tip of $\mathcal{T}_{rR}$ is $|\mathbf{P}(\mathcal{T}_{rR})| \times |\mathbf{P}(\mathcal{T}^*_{rL})|$, so the total number is $(|\mathbf{P}(\mathcal{T}_{rL})| \times |\mathbf{P}(\mathcal{T}^*_{rR})|) + (|\mathbf{P}(\mathcal{T}_{rR})| \times |\mathbf{P}(\mathcal{T}^*_{rL})|)$. □

Proposition 2.2. *If $\mathcal{T}$ has at least two tips, then* $|\mathbf{P}(\mathcal{T}^*)| = |\mathbf{P}(\mathcal{T})| + (|\mathbf{P}(\mathcal{T}^*_{rL})| \times |\mathbf{P}(\mathcal{T}^*_{rR})|)$.

Proof. The root r of $\mathcal{T}$ has become an internal node of $\mathcal{T}^*$ connected to a new tip, t. In some partitions of $N(\mathcal{T}^*)$, t is the only member of its element and

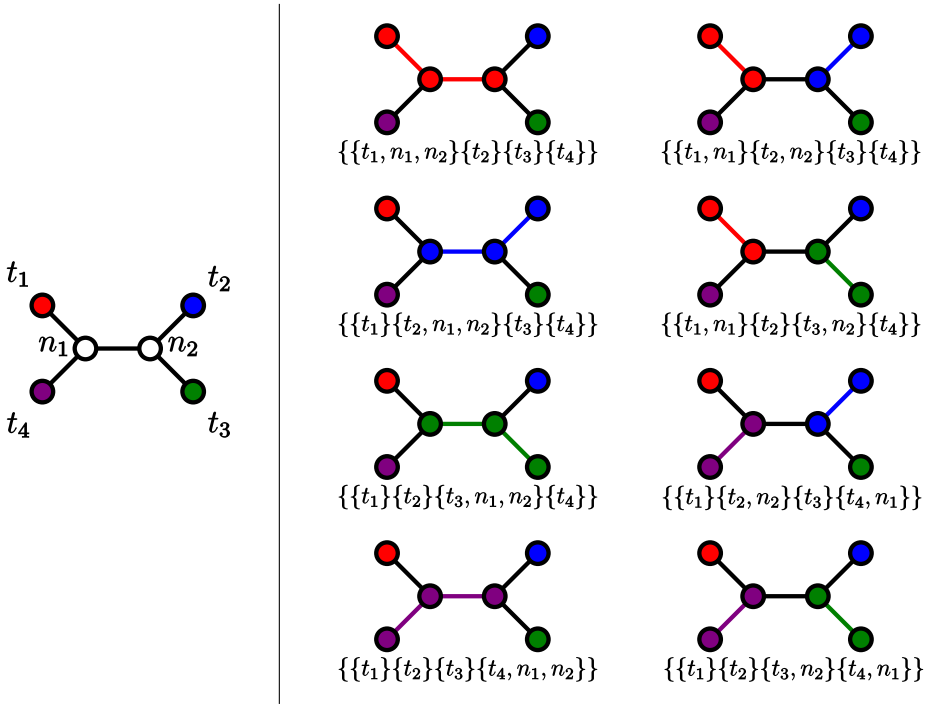


Figure 3: An unrooted phylogeny (left) and the eight partitions of its node set (right). Coloured branches connect members of the same partition element.

there are obviously $|\mathbf{P}(\mathcal{T})|$ of these, because counting them is the same problem as counting partitions of the tree rooted at r with t excised.

If t is not the solo member of its element, r is in the same element by the connectedness requirement. The number of ways of partitioning $\mathcal{T}_{rL}$ and $\mathcal{T}_{rR}$ as subtrees of $\mathcal{T}$ if this is true are $|\mathbf{Q}(\mathcal{T}_{rL})|$ $(= |\mathbf{P}(\mathcal{T}^*_{rL})|)$ and $|\mathbf{Q}(\mathcal{T}_{rR})|$ $(= |\mathbf{P}(\mathcal{T}^*_{rR})|)$ respectively. The total number of such partitions is the product of these. □

Since $\mathbf{P}(\mathcal{T})$ and $\mathbf{P}(\mathcal{T}^*)$ are trivially known when $\mathcal{T}$ has one tip, $|\mathbf{P}(\mathcal{T})|$ can now be calculated for any $\mathcal{T}$ by doing a post-order tree traversal. Specifically, at each node u, if u is a tip then we record $|\mathbf{P}(\mathcal{T}_u)| = |\mathbf{P}(\mathcal{T}^*_u)| = 1$. Otherwise if uL and uR are u's children, we have already recorded $|\mathbf{P}(\mathcal{T}_{uL})|$, $|\mathbf{P}(\mathcal{T}^*_{uL})|$, $|\mathbf{P}(\mathcal{T}_{uR})|$ and $|\mathbf{P}(\mathcal{T}^*_{uR})|$. Calculate $|\mathbf{P}(\mathcal{T}_u)|$ and $|\mathbf{P}(\mathcal{T}^*_u)|$ by propositions 2.1 and 2.2 and record them. The last node visited is r, and at this point we can stop with the calculation of $|\mathbf{P}(\mathcal{T})|$. See figure 4 for an example.

2.2.1 Extension to the multifurcating case

The modification is fairly trivial If $\mathcal{T}$ is not binary. If the root r has p children then and $\mathcal{T}_{rk}$ is the subtree rooted at its kth child, then:

$$|\mathbf{P}(\mathcal{T})| = \sum_{1\leq i\leq p} \left(|\mathbf{P}(\mathcal{T}_{ri})| \prod_{\substack{1\leq j\leq p \\ j\neq i}} |\mathbf{P}(\mathcal{T}^*_{rj})| \right)$$

and

$$|\mathbf{P}(\mathcal{T}^*)| = |\mathbf{P}(\mathcal{T})| + \prod_{1\leq i\leq p} |\mathbf{P}(\mathcal{T}^*_{ri})|$$

and the traversal works as before.

2.3 Counting root elements

We now want to determine, of the $|\mathbf{P}(\mathcal{T})|$ partitions of $\mathcal{T}$'s node set, what number have r in the same partition element as a tip t.

Let $\{t_1, \ldots, t_n\}$ be the tips of $\mathcal{T}$, ordered as they would appear in a post-order traversal, in particular such that the tips from any subtree make up a consecutive run of indexes. We wish to calculate the elements of n-tuple $\mathbf{v}(\mathcal{T}) = (v_1(\mathcal{T}), \ldots, v_n(\mathcal{T}))$ where $v_i(\mathcal{T})$ is the number of partitions of $N(\mathcal{T})$ with r in the same element as t_i; then $\sum_{1\leq i\leq n} v_i(\mathcal{T}) = |\mathbf{P}(\mathcal{T})|$. If $\mathcal{T}$ has one tip t_1, then obviously $v_1(\mathcal{T}) = 1$ and $\mathbf{v}(\mathcal{T})$ is the 1-tuple (1). For any other $\mathcal{T}$, define $\mathbf{v}(\mathcal{T}_{rL})$ and $\mathbf{v}(\mathcal{T}_{rR})$ as the same counts when $\mathcal{T}_{rL}$ and $\mathcal{T}_{rR}$ are considered trees in their own right; these tuples thus have only the same number of elements as $\mathcal{T}_{rL}$ and $\mathcal{T}_{rR}$, respectively, have tips. Suppose the tips of $\mathcal{T}_{rL}$ occur first in the ordering of the tips of $\mathcal{T}$, and there are z of the former (and hence $n - z$ tips of $\mathcal{T}_{rR}$).

Proposition 2.3. *Suppose $\mathcal{T}$ has at least two tips. Then:*

$$v_i(\mathcal{T}) = \begin{cases} v_i(\mathcal{T}_{rL}) \times |\mathbf{P}(\mathcal{T}^*_{rR})| & t_i \text{ is descended from } rL \\ v_{i-z}(\mathcal{T}_{rR}) \times |\mathbf{P}(\mathcal{T}^*_{rL})| & t_i \text{ is descended from } rR \end{cases}$$

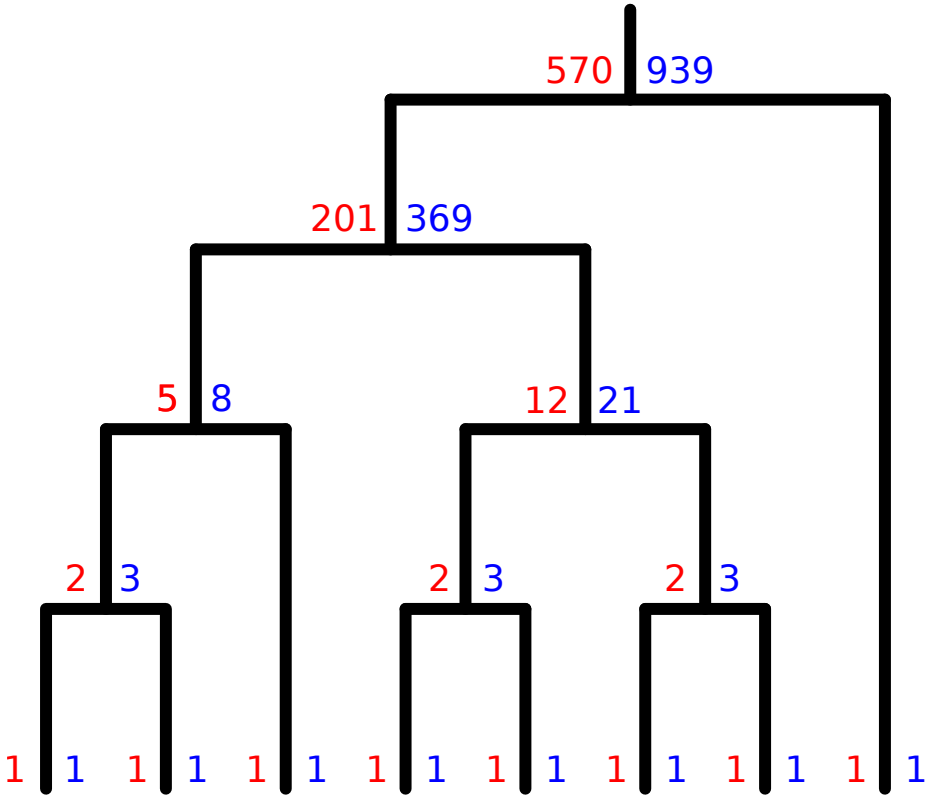


Figure 4: How to count partitions. At each node u, if $\mathcal{T}_u$ is the subtree rooted at u, then the red number is $|\mathbf{P}(\mathcal{T}_u)|$ and the blue $|\mathbf{P}(\mathcal{T}_u^*)|$. If an internal u has children uL and uR, $|\mathbf{P}(\mathcal{T}_u)|$ is calculated as $(|\mathbf{P}(\mathcal{T}_{uL})| \times |\mathbf{P}(\mathcal{T}_{uR}^*)|) + (|\mathbf{P}(\mathcal{T}_{uR})| \times |\mathbf{P}(\mathcal{T}_{uL}^*)|)$ (the sum of the product of the blue number at uL and the red number at uR, and the product of the blue number at uR and the red number at uL). $|\mathbf{P}(\mathcal{T}_u^*)|$ is calculated $|\mathbf{P}(\mathcal{T}_u)| + (|\mathbf{P}(\mathcal{T}_{uL}^*)| \times |\mathbf{P}(\mathcal{T}_{uR}^*)|)$ (the sum of the red number at u and the product of the blue numbers at its children).

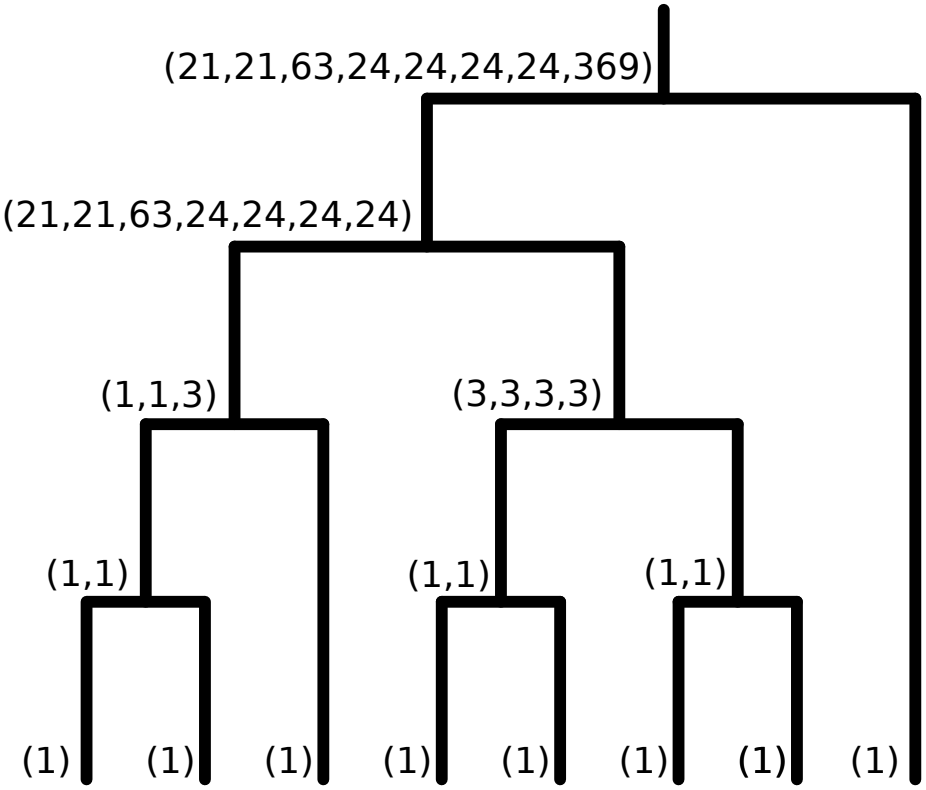


Figure 5: The calculation of $\mathbf{v}(\mathcal{T})$ if $\mathcal{T}$ is the tree in figure 4. Each internal node u is annotated with $\mathbf{v}(\mathcal{T}_u)$, with orders within tuples occurring in the order of the descendant tips of u as displayed from left to right.

Proof. First suppose r is in the same element as a tip t_i of $\mathcal{T}_{rL}$. (Because the tips of rL occur first in the ordering, the index of t_i as a tip of $\mathcal{T}$ is the same as that of it as a tip of $\mathcal{T}_{rL}$.) This forces rL to be in the same partition element as r by the connectedness requirement, and the number of members of $\mathcal{T}_{rL}$ that have this property is $v_i(\mathcal{T}_{rL})$. Such a choice of partition of $\mathcal{T}_{rL}$ leaves $|\mathbf{Q}(\mathcal{T}_{rR})| = |\mathbf{P}(\mathcal{T}^*_{rR})|$ ways of partitioning the nodes of $\mathcal{T}_{rR}$. If r is instead the same element as a tip of $\mathcal{T}_{rR}$, then the argument is the same except that the ith tip of $\mathcal{T}$ is the $(i-z)$th tip of $\mathcal{T}_{rR}$. □

All entries of $\mathbf{v}(\mathcal{T})$ can be calculated by a similar post-order traversal to that described in the previous section. At each internal node u, $\mathbf{v}(\mathcal{T}_u)$ can be formed as described above, $|\mathbf{P}(\mathcal{T}_u)|$ obtained by summing its elements, and $|\mathbf{P}(\mathcal{T}^*_u)|$ calculated from $|\mathbf{P}(\mathcal{T}_u)|$ and the partition counts for u's child subtrees. See figure 5 for an example.

2.4 Sampling uniformly from $\mathbf{P}(\mathcal{T})$

If the post-order traversal above is complete, sampling a random partition requires a single *pre*-order traversal. The vector $\mathbf{v}(\mathcal{T})$ consists of probability weights for a draw of partition element for r, as it determines how many of the $\mathbf{P}(\mathcal{T})$ total partitions have r in each element. Subsequently, when the traversal reaches another node u with parent uP, and we have already placed uP in a partition element containing a tip t, then u must also be in that element if t is one of its descendants, by connectedness, or if t is u itself. Otherwise, there are $|\mathbf{P}(\mathcal{T}_u^*)|$ ways in which $\mathcal{T}_u$ can be partitioned given that uP has already been allocated an element. $|\mathbf{P}(\mathcal{T}_u^*)| - |\mathbf{P}(\mathcal{T}_u)|$ of these have u in the same element as uP, while the remaining $|\mathbf{P}(\mathcal{T}_u)|$ do not. The entries of $\mathbf{v}(\mathcal{T}_u)$ give the numbers of ways in which u can be placed in the same element as each of its descendant tips. The partition element for u can then be sampled with probability determined by a weight vector that has the entries of $\mathbf{v}(\mathcal{T}_u)$ for the elements containing the tips descended from u, $|\mathbf{P}(\mathcal{T}_u^*)| - |\mathbf{P}(\mathcal{T}_u)|$ for the already determined parent partition element, and 0 for any other element.

3 Incomplete sampling

Now suppose that not every host in the transmission tree was sampled, but instead that that there were m unsampled individuals, all of which are ancestral to at least one sampled individual. It is not sufficient merely add m extra partition elements that contain no tips. This is because, to provide a simple example, if an unsampled host b was infected by a sampled host a and directly infects only one other host c which was also sampled, the region of the phylogeny corresponding to the infection of b exists only along a branch (the branch whose parent node is in the partition element containing a's tip and whose child node is in the one containing c's tip) and no internal nodes are associated with b at all. Nonetheless, the procedure for counting, and sampling from, the set of tree partitions with m extra elements turns out to lead to the more general answer as a byproduct of the calculations.

Now suppose $\mathbf{P}_m(\mathcal{T})$ is this set. ($\mathbf{P}(\mathcal{T})$ as described above is $\mathbf{P}_0(\mathcal{T})$). Let $\mathbf{PS}_m(\mathcal{T})$ be the subset of $\mathbf{P}_m(\mathcal{T})$ where the root of $\mathcal{T}$ shares its partition element with a tip, and $\mathbf{PU}_m(\mathcal{T})$ the subset where it does not. $\mathbf{P}_m(\mathcal{T}^*)$ can be defined, although as $\mathcal{T}^*$ has no root $\mathbf{PS}_m(\mathcal{T}^*)$ and $\mathbf{PU}_m(\mathcal{T}^*)$ cannot be. $\mathbf{Q}_m(\mathcal{T})$ can also be defined and again is exactly analogous to $\mathbf{P}_m(\mathcal{T}^*)$.

Call the partition elements containing tips the sampled elements, and those not containing tips the unsampled elements.

If $\mathcal{T}$ has a single tip then $\mathbf{P}_m(\mathcal{T}) = 0$ and $\mathbf{P}_m(\mathcal{T}^*) = 0$ for all $m > 0$. No internal nodes exist to be assigned to unsampled elements.

3.1 Counting transmission trees

Proposition 3.1. *If $\mathcal{T}$ has at least two tips, then*

$$|\mathbf{PS}_m(\mathcal{T})| = \sum_{i=0}^{m}((|\mathbf{P}_i(\mathcal{T}_{rL})| \times |\mathbf{P}_{m-i}(\mathcal{T}^*_{rR})|) + (|\mathbf{P}_{m-i}(\mathcal{T}_{rR})| \times |\mathbf{P}_i(\mathcal{T}^*_{rL})|))$$

Proof. Since r is not part of an unsampled element, the m such elements must be split between the subtree descended from its left child and that descended from its right child. The summation expresses the number of ways to make this split. Apart from this adjustment the argument is the same as in proposition 2.1, as if $m = 0$ then r is always in a sampled element.

□

Proposition 3.2. *If $\mathcal{T}$ has at least two tips, then*

$$|\mathbf{PU}_m(\mathcal{T})| = \sum_{i=0}^{m-1}(|\mathbf{P}_i(\mathcal{T}^*_{rL})| \times |\mathbf{P}_{m-1-i}(\mathcal{T}^*_{rR})|)$$

Proof. Since r is in an unsampled element, one of the m of those is accounted for. The remaining $m-1$ are split amongst the left and right subtrees as above.

If we consider $\mathcal{T}_{rL}$ in isolation, we want to count the number of ways of partitioning its nodes with i unsampled elements for certain and possibly one extra which, if it exists, must contain rL. (This possible element is the intersection of an element of a partition of the nodes of $\mathcal{T}$ which includes r, and $N(\mathcal{T}_{rL})$.) This number is clearly $|\mathbf{Q}_i(\mathcal{T}_{rL})| = |\mathbf{P}_i(\mathcal{T}^*_{rL})|$. Since exactly the same applies to $\mathcal{T}_{rR}$, the product of $|\mathbf{P}_i(\mathcal{T}^*_{rL})|$ and $|\mathbf{P}_{m-i}(\mathcal{T}^*_{rL})|$ is the desired number for a known i.

□

Proposition 3.3. *If $\mathcal{T}$ has at least two tips, then*

$$|\mathbf{P}_m(\mathcal{T}^*)| = |\mathbf{P}_m(\mathcal{T})| + \sum_{i=0}^{m}(|\mathbf{P}_i(\mathcal{T}^*_{rL})| \times |\mathbf{P}_{m-i}(\mathcal{T}^*_{rR})|)$$

Proof. Identical to proposition 2.2 except that we again allow for all the ways that the m unsampled elements can be distributed.

□

Propositions 3.1 to 3.3 then allow us to calculate $|\mathbf{P}_m(\mathcal{T})| = |\mathbf{PU}_m(\mathcal{T})| + |\mathbf{PS}_m(\mathcal{T})|$ by a post-order traversal; note that at every internal node u we must calculate $|\mathbf{P}_i(\mathcal{T}_u)|$ and $|\mathbf{P}_i(\mathcal{T}^*_u)|$ for all i with $0 \leq i \leq m$, not just $|\mathbf{P}_m(\mathcal{T}_u)|$ and $|\mathbf{P}_m(\mathcal{T}^*_u)|$.

3.2 Counting root sampled elements

Once again, let $\{t_1, \ldots, t_n\}$ be the tips of $\mathcal{T}$, ordered as they would appear in a post-order traversal and that the first z tips are descended from $\mathcal{T}_{rL}$. Define $\mathbf{V}(\mathcal{T})$ be the $n \times (m+1)$ matrix whose ijth entry $v_{ij}(\mathcal{T})$ is the number of partitions of $\mathcal{T}$ such that r is in the same element as t_i if there are $j-1$ unsampled elements.

Proposition 3.4.

$$v_{ij}(\mathcal{T}) = \begin{cases} \sum_{k=0}^{j-1} v_{ik}(\mathcal{T}_{rL}) \times |\mathbf{P}_{j-1-k}(\mathcal{T}^*_{rR})| & t_i \text{ is descended from } rL \\ \sum_{k=0}^{j-1} v_{(i-z)k}(\mathcal{T}_{rR}) \times |\mathbf{P}_{j-1-k}(\mathcal{T}^*_{rL})| & t_i \text{ is descended from } rR \end{cases}$$

Proof. Analagous to proposition 2.3 after counting the ways the i unsampled elements can be split between the two child subtrees. □

3.3 Sampling uniformly from $\mathbf{P}_m(\mathcal{T})$

The previous sections allows us, for $m \in \mathbb{N}$ and a binary $\mathcal{T}$, to calculate $\mathbf{PU}_i(\mathcal{T})$, $\mathbf{PS}_i(\mathcal{T})$, $\mathbf{P}^*_i(\mathcal{T})$ and $\mathbf{V}(\mathcal{T})$ for all i with $0 \leq i \leq m$ by a post-order traversal. The pre-order sampling procedure works by, first, at r, choosing an element using a vector of probability weights consisting of the $(m+1)$th row of $\mathbf{V}(\mathcal{T})$ for the sampled elements and $\mathbf{PU}_m(\mathcal{T})$ for an unsampled element. Once this is done, we must randomly choose how the remaining unsampled elements are divided between $\mathcal{T}_{rL}$ and $\mathcal{T}_{rR}$.

If we chose the element containing a tip t_i for r and t_i is descended from rL, then the number of partitions which have j unsampled elements amongst the nodes of $\mathcal{T}_{rL}$ (and hence $m-j$ amongst the nodes of $\mathcal{T}_{rR}$) is $v_{ij}(\mathcal{T}_{rL}) \times |\mathbf{P}_{m-j}(\mathcal{T}^*_{rR})|$.

If we chose the element containing a tip t_i for r and t_i is descended from rR, then the number of partitions which have j unsampled elements amongst the nodes of $\mathcal{T}_{rL}$ is $v_{i(m-j)}(\mathcal{T}_{rR}) \times |\mathbf{P}_j(\mathcal{T}^*_{rL})|$.

If we chose an unsampled element for r, then the number of partitions which have j *other* unsampled elements amongst the nodes of $\mathcal{T}_{rL}$ (and hence $m-1-j$ amongst the nodes of $\mathcal{T}_{rR}$) is $|\mathbf{P}_j(\mathcal{T}^*_{rL})| \times |\mathbf{P}_{m-1-j}(\mathcal{T}^*_{rR})|$.

In all cases we have a set of weights which we can use to randomly select j.

When the traversal arrives at a new node u whose parent uP has been assigned an element and we know that there are k previously unencountered unsampled elements in the partition of the nodes of $\mathcal{T}_u$, then we are forced to put u in the same element as uP if that element is sampled and contains a tip descended from u. Otherwise, the $(k+1)$th row of $\mathbf{V}(\mathcal{T}_u)$ gives the weights for being in a sampled element along with a tip of $\mathcal{T}_u$, $\mathbf{PU}_k(\mathcal{T}_u)$ the weight for a transition to a new unsampled element (whether or not the element containing uP is unsampled) and $\mathbf{P}_k(\mathcal{T}^*_u) - \mathbf{PS}_k(\mathcal{T}_u)$ the weight for continuing in the same element as uP whether than element was sampled or unsampled. We choose an element for u in this way and then, if u is not a tip, split the $k-1$ (if we assigned u to a new unsampled element) or k (if we did not) remaining unencountered

unsampled elements between u's left and right descendant subtrees as above, except that there is an extra case where we chose a *sampled* element for u but the tip in that element is descended from neither of u's children. The weight for j of k remaining unsampled elements going to u's left subtree $\mathcal{T}_{uL}$ is $|\mathbf{P}_j(\mathcal{T}^*_{uL})| \times |\mathbf{P}_{k-j}(\mathcal{T}^*_{uR})|$.

3.4 Sampling uniformly from the set of transmission trees with m unsampled hosts

To complete the picture, we now must consider the case where only l of the m unsampled hosts actually correspond to partition elements, and the remaining $m-l$ appear only along branches. If we have sampled an element of $\mathbf{P}_l(\mathcal{T})$ as above, we need to distribute the additional $m-l$, and there are $l+n$ branches on which these can be put, which are the branches ending in the earliest appearances of in the tree of each of the $l+n$ partition elements. The number of ways of doing this is the number of ways of assigning $m-l$ identical objects to $l+n$ possibly empty groups, i.e. $\binom{m+n-1}{l+n-1}$.

If we calculate $|\mathbf{P}_l(\mathcal{T})|$ for $0 \leq l \leq m$, then we know that there are $|\mathbf{P}_l(\mathcal{T})| \times \binom{m+n-1}{l+n-1}$ transmission trees with m unsampled hosts where l of those hosts have partition elements. We can randomly select an l using those counts as probability weights, randomly generate an element of $\mathbf{P}_l(\mathcal{T})$ as above, and finally randomly assign the extra $m-l$ elements to branches.

4 Multiple sampling

Removing unsampled hosts from consideration, I now relax the assumption that each partition element contains only a single tip. Fix a partition $\mathfrak{P}$ of the tip set $E(\mathcal{T})$ of $\mathcal{T}$ and want to investigate the set $\mathbf{P}(\mathcal{T};\mathfrak{P})$ which is the set of partitions $\mathfrak{A}$ of $N(\mathcal{T})$ such that $\{A \cap E(\mathcal{T}) : A \in \mathfrak{A}\} = \mathfrak{P}$ (i.e. the partitions of $N(\mathcal{T})$ which agree with $\mathfrak{P}$ on the tips). Each element of $\mathfrak{P}$ contains all the tips sampled from a single host. $\mathbf{P}(\mathcal{T})$ as in section 2 is $\mathbf{P}(\mathcal{T} : \mathfrak{I})$ where $\mathfrak{I}$ is the partition of singletons of $E(\mathcal{T})$.

For a set $A \in \mathfrak{P}$ define the bridge $b(A)$ of A to be the minimal subset of $N(\mathcal{T})$ such that $A \subseteq b(A)$ and the subgraph of $\mathcal{T}$ induced by $b(A)$ is connected. This contains all elements of A, the MRCA of A, and all nodes on the paths between them. Obviously if $|A| = 1$ then $b(A) = A$.

If any two elements of $\mathfrak{P}$ have bridges whose intersections are nonempty, then $|\mathbf{P}(\mathcal{T};\mathfrak{P})| = 0$; there are simply no possible transmission trees because the connectedness requirement would insist that some nodes be part of more than one partition element. So assume that this is not true. For any partition, being a bridge node forces a node to be a member of the same element as those tips whose bridge it belongs to.

Note there are two intuitive ways to consider $\mathbf{P}(\mathcal{T}_u;\mathfrak{P})$ for a subtree $\mathcal{T}_u$ of $\mathcal{T}$ rooted at u. The first would be to use the set $\mathfrak{P}_u = \{A \cap E(\mathcal{T}_u) : A \in \mathfrak{P}\}$ as a

partition of the tips of $\mathcal{T}_u$. In this case, bridge nodes determined by $\mathfrak{P}$ are not necessarily determined as such by $\mathfrak{P}_u$.

The second way is to retain the restrictions on the partition elements to which a node of $\mathcal{T}_u$ can belong that are determined by $\mathfrak{P}$ even when we move to counting partitions of $\mathcal{T}_u$. This is the version which is useful for our purposes. For example, if $A \in \mathfrak{P}$ consists of the two tips t_1 and t_2, but only t_1 is a tip of $\mathcal{T}_u$, then in enumerating and sampling partitions we still consider the intersection $b(A) \cap N(\mathcal{T}_u)$ to be bridge nodes, even though $\{t_1\}$ is a singleton element of $\mathfrak{P}_u$. (In fact u must be a member of $b(A)$.)

So for any node u of $\mathcal{T}$, we define $\mathbf{P}(\mathcal{T}_u; \mathfrak{P})$ to be the number of ways to partitioning $\mathcal{T}$'s nodes that respects the set of bridge nodes that $\mathfrak{P}$ requires.

It should be fairly obvious that if an internal node is a bridge node then one or both of its children must be as well. $\mathbf{P}(\mathcal{T}^*; \mathfrak{P})$ has the definition one would expect, with the extra node forming a singleton extra element of the tip partition. For a node u, $\mathbf{P}(\mathcal{T}^*_u; \mathfrak{P})$ again respects the bridge nodes imposed by $\mathfrak{P}$ on $\mathcal{T}$.

Now suppose $\mathcal{T}$ has at least two tips and let u be any internal node of $\mathcal{T}$, including r. Its children are uL and uR.

Proposition 4.1.

$$|\mathbf{P}(\mathcal{T}^*_u; \mathfrak{P})| = \begin{cases} |\mathbf{P}(\mathcal{T}_u; \mathfrak{P})| \\ \quad +(|\mathbf{P}(\mathcal{T}^*_{uL}; \mathfrak{P})| \times |\mathbf{P}(\mathcal{T}^*_{uR}; \mathfrak{P})|) & u \textit{ is a not a bridge node} \\ |\mathbf{P}(\mathcal{T}_u; \mathfrak{P})| & u \textit{ is a bridge node} \end{cases}$$

Proof. If u is a bridge node then it cannot belong to the partition element containing the extra node, so the number of partitions is exactly the number with that node excised. Otherwise, the argument is as proposition 2.2. □

Proposition 4.2.

$$|\mathbf{P}(\mathcal{T}_u; \mathfrak{P})| = \begin{cases} (|\mathbf{P}(\mathcal{T}_{uL}; \mathfrak{P})| \times |\mathbf{P}(\mathcal{T}^*_{uR}; \mathfrak{P})|) \\ \quad + (|\mathbf{P}(\mathcal{T}_{uR}; \mathfrak{P})| \times |\mathbf{P}(\mathcal{T}^*_{uL}; \mathfrak{P})|) & u \textit{ is not a bridge node} \\ |\mathbf{P}(\mathcal{T}^*_{uL}; \mathfrak{P})| \times |\mathbf{P}(\mathcal{T}^*_{uR}; \mathfrak{P})| & u \textit{ is a bridge node} \end{cases}$$

Proof. If u is a bridge node then at least one of its children is too. At least one of those children, in fact, must be part of the same bridge as itself; suppose this is uL. Now u and uL must be in the same partition element so the $|\mathbf{P}(\mathcal{T}_{uL}); \mathfrak{P}|$ partitions of uL determine which element u belongs to, and because uL is a bridge node $|\mathbf{P}(\mathcal{T}_{uL}); \mathfrak{P}| = |\mathbf{P}(\mathcal{T}^*_{uL}); \mathfrak{P}|$ by proposition 4.1. If uR is not a bridge node then there are $|\mathbf{P}(\mathcal{T}^*_{uR}); \mathfrak{P}|$ ways of partitioning the nodes of $\mathcal{T}_{uR}$ by a by now very familiar logic. If it is, then there are $|\mathbf{P}(\mathcal{T}_{uR}); \mathfrak{P}|$ partitions of those nodes since the element that uR belongs to is fixed, but $|\mathbf{P}(\mathcal{T}_{uR}); \mathfrak{P}| = |\mathbf{P}(\mathcal{T}^*_{uR}); \mathfrak{P}|$ by proposition 4.1. Obviously the same applies with uR and uL reversed. If u is not a bridge node then the argument of proposition 2.1 still applies. □

If $\mathfrak{P}$ has l elements, number them in an arbitrary way. Define $\mathbf{v}(\mathcal{T};\mathfrak{P}) = (v_1(\mathcal{T};\mathfrak{P}),\ldots,v_l(\mathcal{T};\mathfrak{P}))$ where $v_i(\mathcal{T};\mathfrak{P})$ is the number of partitions of the nodes of $\mathcal{T}$ where r shares a partition element with the members of the ith element A_i of $\mathfrak{P}$. For a subgraph $\mathcal{T}_u$, let $v_i(\mathcal{T}_u;\mathfrak{P})$ be the number of partitions of the nodes of $\mathcal{T}_u$ where r shares a partition element with the intersection $E(\mathcal{T}_u) \cap A_i$; this may be 0 where that intersection is empty. Once again, $v_i(\mathcal{T}_u;\mathfrak{P})$ counts only partitions that respect the bridge nodes imposed by $\mathfrak{P}$.

Proposition 4.3. *Suppose $|A_i \cap E(\mathcal{T}_u)| > 0$; that is, some tips of $\mathcal{T}_u$ are members of A_i.*

$$v_i(\mathcal{T}_u;\mathfrak{P}) = \begin{cases} v_i(\mathcal{T}_{uL};\mathfrak{P}) \times |\mathbf{P}(\mathcal{T}^*_{uR};\mathfrak{P})| & A_i \cap E(\mathcal{T}_{uR}) = \emptyset \text{ and either } u \text{ is not a bridge node or } u \in b(A_i) \\ v_i(\mathcal{T}_{uR};\mathfrak{P}) \times |\mathbf{P}(\mathcal{T}^*_{uL};\mathfrak{P})| & A_i \cap E(\mathcal{T}_{uL}) = \emptyset \text{ and either } u \text{ is not a bridge node or } u \in b(A_i) \\ v_i(\mathcal{T}_{uL};\mathfrak{P}) \times v_i(\mathcal{T}_{uR};\mathfrak{P}) & A_i \cap E(\mathcal{T}_{uL}) \neq \emptyset \text{ and } A_i \cap E(\mathcal{T}_{uR}) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

Proof. First note that if $A_i \cap E(\mathcal{T}_{uL}) \neq \emptyset$ and $A_i \cap E(\mathcal{T}_{uR}) = \emptyset$ then u is a bridge node and in particular a member of $b(A_i)$. If u is not a bridge node at all, then we must be in one of the first two cases. The argument in that case differs from proposition 2.3 only in terms of notation.

If $u \in b(A_i)$ but only one child of u has any descendant tips that are members of A_i, then suppose uL is the child that does. Then uR is either not a bridge node or a member of $b(A_j)$ for $j \neq i$. The number of ways of partitioning the nodes of $\mathcal{T}_{uL}$ with uL sharing a partition element with the members of A_i is $v_i(\mathcal{T}_{uL})$ and, because u shares a partition element with uL, each of those once again results in $|\mathbf{P}(\mathcal{T}^*_{uR};\mathfrak{P})|$ ways of partitioning the nodes of $\mathcal{T}_u$. The same goes with uL and uR reversed.

If both children have descendant tips that are members of A_i then the number of partitions with u sharing an element with the members of A_i is just the product of the number of ways of partitioning its child subtrees in the same way. All three nodes are forced to be part of the same partition element.

Under any other situation $u \in b(A_j)$ for $j \neq i$ and there cannot be any partitions that have it sharing an element with the members of A_i.

□

If t is a tip then $|\mathbf{P}(\mathcal{T}^*_t;\mathfrak{P})| = 1$, $|\mathbf{P}(\mathcal{T}_t;\mathfrak{P})| = 1$, and $v_i(\mathcal{T}_t;\mathfrak{P}) = 1$ if $t \in A_i$ and 0 otherwise. This is all that is necessary to set up traversals analogous to those described in section 2.

5 Trees with timings

Finally, I return once more to the case where sampling is single and complete. I now give $\mathcal{T}$ branch lengths, which means a height function $h : N(\mathcal{T}) \to \mathbb{R}^+$

can be defined such that for all nodes u with parent uP, $h(u) < h(uP)$. Branch lengths and heights are intended to be in units of calendar time, not genetic distance. We extend h to $N(\mathcal{T}^*)$ by setting $h(t) = \infty$ if t is the extra tip; while $\mathcal{T}^*$ remains formally unrooted, there is only one way to display it that makes sense.

Each tip t_i is now associated with a closed interval $I_i = [\alpha_i, \beta_i]$ such that, for any partition $\mathfrak{A}$ of the nodes of $\mathcal{T}$, if $\{u, t_i\} \subseteq A_i \in \mathfrak{A}$ then $h(u) \in I_i$. (Obviously no partitions exist without $t_i \in I_i$ for all i.)

The I_is determine minimum and maximum heights for all the nodes in each partition element. This is useful if infection is expected to end with sampling, or if a maximum time from infection to sampling is known. If $\mathbf{I}$ is the complete set of intervals, then let $\mathbf{P}(\mathcal{T}; \mathbf{I})$ by the set of partitions subject to these additional restrictions. For a subtree $\mathcal{T}_u$, $\mathbf{P}(\mathcal{T}_u; \mathbf{I})$ is the set of partitions subject to the restrictions where they are appropriate (i.e. the I_i where t_i is actually a tip of $\mathcal{T}_u$).

Let $\mathbf{P}(\mathcal{T}^*; \mathbf{I} \cup [\gamma, \infty))$ be the set of partitions of $\mathcal{T}^*$ such that the element containing the extra tip contains only nodes of heights greater than γ, and the restrictions imposed by $\mathbf{I}$ still apply to the other elements.

Proposition 5.1. *If $\mathcal{T}$ has at least two tips, then*

$$|\mathbf{P}(\mathcal{T}^*; \mathbf{I} \cup [\gamma, \infty))| = \begin{cases} |\mathbf{P}(\mathcal{T}; \mathbf{I})| & h(r) < \gamma \\ (|\mathbf{P}(\mathcal{T}^*_{rL}; \mathbf{I} \cup [\gamma, \infty))| \times |\mathbf{P}(\mathcal{T}^*_{rR}; \mathbf{I} \cup [\gamma, \infty))|) & \\ \quad + |\mathbf{P}(\mathcal{T}; \mathbf{I})| & h(r) \geq \gamma \end{cases}$$

Proof. If $h(r) < \gamma$ then r cannot be in the same partition element as the extra tip, so the number of partitions is the same as in the rooted case. Otherwise, see proposition 2.2. □

I omit a single expression for $|\mathbf{P}(\mathcal{T}; \mathbf{I})|$ and instead suggest calculating it by calculating the v_i first and adding them up. So let $v_i(\mathcal{T}; \mathbf{I})$ be the number of partitions of $\mathcal{T}$ where r is in the same element as t_i and the restrictions imposed by $\mathbf{I}$ apply. Returning to the notation of proposition 2.3, so $\mathcal{T}_{rL}$ has z of the n tips and those descended from it come first in the ordering:

Proposition 5.2. *If $\mathcal{T}$ has at least two tips, then*

$$v_i(\mathcal{T}; \mathbf{I}) = \begin{cases} 0 & h(r) \notin I_i \\ v_i(\mathcal{T}_{rL}; \mathbf{I}) \times |\mathbf{P}(\mathcal{T}^*_{rR}; \mathbf{I} \cup [\alpha_i, \infty))| & h(r) \in I_i \text{ and } t_i \text{ is descended from } rL \\ v_{i-z}(\mathcal{T}_{rR}; \mathbf{I}) \times |\mathbf{P}(\mathcal{T}^*_{rL}; \mathbf{I} \cup [\alpha_i, \infty))| & h(r) \in I_i \text{ and } t_i \text{ is descended from } rR \end{cases}$$

Proof. If $h(r)$ lies outside I_i then the answer is trivially zero. If not, and t_i is descended from rL, then there are $v_i(\mathcal{T}_{rL}; \mathbf{I})$ ways of partitioning the nodes of $\mathcal{T}_{rL}$ such that rL is in the same element as t_i. For each of these, we need the the number of ways of partitioning the nodes of $\mathcal{T}_{rR}$ such that an extra

element can creep down from the root, but that that element cannot contain any nodes whose heights are smaller than the lower limit of I_i, i.e. α_i. This is $|\mathbf{P}(\mathcal{T}^*_{rR}; \mathbf{I} \cup [\alpha_i, \infty))|$. As usual, an identical argument applies with rL and rR reversed. □

Then $|\mathbf{P}(\mathcal{T}; \mathbf{I})| = \sum_{i=1}^{n} v_i(\mathcal{T}; \mathbf{I})$. If $\mathcal{T}$ has one tip, then both $\mathbf{P}(\mathcal{T}; \mathbf{I})$ and $\mathbf{P}(\mathcal{T}^*; \mathbf{I} \cup [\gamma, \infty))$ (regardless of γ) have one element if the tip lies within its own interval and zero if it does not. As in the previous section, the traversals described in section 2 can be used to count the full set of partitions and sample uniformly from it.

References

[1] M. Aldrin et al. "Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates". In: *Journal of The Royal Society Interface* 8.62 (2011), pp. 1346–1356.

[2] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. "Bayesian inference of infectious disease transmission from whole genome sequence data". In: *Molecular Biology and Evolution* 31.7 (2014), pp. 1869–1879.

[3] Xavier Didelot et al. "Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks". In: *Molecular Biology and Evolution* 34.4 (2017), pp. 997–1007.

[4] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. "Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set". In: *PLOS Computational Biology* 11.12 (2015), e1004613.

[5] Thibaut Jombart et al. "Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data". In: *PLOS Computational Biology* 10.1 (2014), e1003457.

[6] Don Klinkenberg et al. "Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks". In: *PLOS Computational Biology* 13.5 (2017), e1005495.

[7] Max S. Y. Lau et al. "A systematic Bayesian integration of epidemiological and genetic data". In: *PLOS Computational Biology* 11.11 (2015), e1004633.

[8] Nardus Mollentze et al. "A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1782 (2014), p. 20133251.

[9] Marco J. Morelli et al. "A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data". In: *PLOS Computational Biology* 8.11 (2012), e1002768. (Visited on 12/13/2012).

[10] Pavel Skums et al. "QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data". In: *Bioinformatics* (2017). DOI: 10.1093/bioinformatics/btx402.

[11] Rolf J. F. Ypma, W. Marijn van Ballegooijen, and Jacco Wallinga. "Relating phylogenetic trees to transmission trees of infectious disease outbreaks". In: *Genetics* 195.3 (Nov. 1, 2013), pp. 1055–1062.

[12] Guangchuang Yu et al. "ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data". In: *Methods in Ecology and Evolution* 8.1 (Jan. 1, 2017), pp. 28–36.