Bayesian Inference for Vertex-Series-Parallel Partial Orders (Supplementary Material)

Chuxuan (Jessie) Jiang[1] **Geoff K. Nicholls**[1] **Jeong Eun Lee**[2]

[1]Department of Statistics, University of Oxford, United Kingdom
[2]Department of Statistics, University of Auckland, New Zealand

A PROOF OF THEOREM 1

This Appendix states and proves the propositions referred to in the proof of Theorem 1 given in Section 2.

A.1 PART I: MARGINAL CONSISTENCY

We first prove marginal consistency for our VSP prior. Intuitively, relations between actors in a VSP $v \in \mathcal{V}_{[n]}$ are determined by the type of their "Most Recent Common Ancestor" (MRCA) in any BDT $t \in t(v)$ representing v. For example the MRCA of actors 2 and 4 in the tree t_0 in Fig. 3 is the blue P-node, so $2\|_{v_0}4$ in the VSP v_0 in Fig. 1. Adding or removing a leaf in the BDT doesn't change relations between other actors because it doesn't change the types of their MRCA's. This property leads to marginal consistency of trees and VSPs.

We begin by giving a stochastic process realising $t \sim \pi_{\mathcal{T}_{[n]}}(t|q)$ in which leaves are added to the tree one at a time. This construction appears in Valdes [1978] but without the random element.

Definition A.1 (Leaf Insertion and deletion) *If $t' \in \mathcal{T}_{[n-1]}$, $t' = (F(t'), E(t'), L(t'))$, is a tree on actors $(i_1, ..., i_{n-1})$ with $\mathcal{F}' \cup \mathcal{A}' = [2n-3]$ then the* leaf-insertion *operation $t = t' \triangleleft (e, i_n)$ at edge $e = \langle e_1, e_2 \rangle,\ e \in E(t')$, gives a tree $t = (F(t), E(t), L(t))$ with two new nodes $j' = 2n-2$ and $j = 2n-1$, leaves $\mathcal{F} = \mathcal{F}' \cup \{j'\}$, internal nodes $\mathcal{A} = \mathcal{A}' \cup \{j\}$, leaf-to-actor map $F_{\mathcal{F}}(t) = F_{\mathcal{F}}(t')$ and $F_{j'}(t') = i_n$, edge set*

$$E(t) = E(t') \setminus \{e\} \cup \{\langle j, j' \rangle, \langle e_1, j \rangle, \langle j, e_2 \rangle\}$$

and $L(t) = L(t') \cup L_j$ where $L_j = (j', e_2), (e_2, j')$ or $\emptyset$ with probabilities $q/2, q/2$ and $1-q$ respectively. The leaf deletion operation $t' = t \triangleright i_n$ reverses this operation, pruning the leaf for actor i_n (and removing its parent node).

Definition A.2 (Generative model for BDTs) *Let $(i_1, \ldots, i_n) \in \mathcal{P}_{[n]}$ be the actor list taken in any order. Simulate $t \sim \pi_{\mathcal{T}_{[n]}}(t|q)$ as follows:*

1. *Set $\mathcal{F} = \{0, 1\}$, $\mathcal{A} = \emptyset$, $F_1 = 1, E = \{\langle 0, 1 \rangle\}, L = \emptyset$ and $t_{(1)} = (F, E, L)$ (a single-edge tree);*
2. *For $k = 2 : n$, (add the actors one at a time)*
   (a) choose an edge $e \sim \mathcal{U}\{E(t_{(k-1)})\}$ at random;
   (b) set $t_{(k)} = t_{(k-1)} \triangleleft (e, i_k)$;
3. *$E(t_{(n)})$ contains an edge $e = \langle 0, e_2 \rangle$. Return the BDT $t = (F(t_{(n)}), E(t_{(n)}) \setminus \{e\}, L(t_{(n)}))$ with leaf labels $\mathcal{F} \leftarrow \mathcal{F} \setminus \{0\}$.*

If we run this generative model we get a random tree distributed according to $\pi_{\mathcal{T}_{[n]}}$.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Proposition 2 (Prior Probability Distribution over $\mathcal{T}_{[n]}$) *The probability distribution over BDTs determined by the process in Definition A.2 is given by (3).*

Proof A.1 (Proposition 2) *Each distinct topology is determined by a unique sequence of edge choices at step 2a in Definition A.2, and at step k an edge is chosen uniformly over the $2k-3$ edges of a tree with k leaves (recall there is a temporary leaf $0 \in \mathcal{F}$ which is removed at step 3). The types of internal nodes are independent so it makes no difference if we set them as we build the tree or at the end.*

We now define sub-trees of BDTs. At the end of step k in the tree-generation process in Definition A.2 the "current tree" is $t_{(k)} \in \mathcal{T}_o$ with $o = (i_1, ..., i_k)$ and at the end of step $k' > k$ it is $t_{(k')} \in \mathcal{T}_{\tilde{o}}$ with $\tilde{o} = (i_1, ..., i_k, i_{k+1}, ..., i_{k'})$. If, for $o, \tilde{o} \in \mathcal{O}_{[n]}$ with $o \subseteq \tilde{o}$, we fix $\tau \in \mathcal{T}_o$ and $t \in \mathcal{T}_{\tilde{o}}$ then the conditional probability

$$\pi_{\mathcal{T}_{\tilde{o}}|\mathcal{T}_o}(t|\tau, q) = \Pr(t_{(k')} = t|t_{(k)} = \tau, q)$$

is the probability to realise $t_{(k')} = t$ when $t_{(k)} = \tau$.

Definition A.3 (Sub-trees and containing trees) *Tree τ is a sub-tree of t (and t contains τ) if $\pi_{\mathcal{T}_{\tilde{o}}|\mathcal{T}_o}(t|\tau, q) > 0$. Let*

$$\mathcal{T}_{\tilde{o}}(\tau) = \{t \in \mathcal{T}_{\tilde{o}} : \pi_{\mathcal{T}_{\tilde{o}}|\mathcal{T}_o}(t|\tau, q) > 0\}$$

give the set of trees in $\mathcal{T}_{\tilde{o}}$ containing a given tree $\tau \in \mathcal{T}_o$.

If t contains τ then t can be realised from τ by a sequence of edge insertions $\triangleleft$ and τ can be recovered from t removing the actors in $\tilde{o} \setminus o$ using the pruning operator $\triangleright$.

The family of prior distributions over trees $\pi_{\mathcal{T}_o}(\tau|q),\ o \in \mathcal{O}_{[n]},\ n \geq 1$ is marginally consistent if, for all $n \geq 1$ and all $o, \tilde{o} \in \mathcal{O}_{[n]}$ with $o \subseteq \tilde{o}$, distributions in the family satisfy

$$\pi_{\mathcal{T}_o}(\tau|q) = \sum_{t \in \mathcal{T}_{\tilde{o}}(\tau)} \pi_{\mathcal{T}_{\tilde{o}}}(t|q) \quad \text{for all } \tau \in \mathcal{T}_o\ . \tag{A.1}$$

Proposition 3 *The probability distribution over BDTs given in (3) is marginally consistent.*

Proof A.2 (Proposition 3) *It is sufficient show marginal consistency holds for $\tilde{o} = [n]$ and $o = [n] \setminus \{i\}$ for any single actor $i \in [n]$ as Eqn. A.1 follows for any pair of subsets of $[n]$ by pruning leaves one at a time using the $\triangleright$ operator.*

Since $\pi_{\mathcal{T}_{[n]}}(t|q)$ in Eqn. 3 does not depend on the order $i_1, \ldots, i_n$ in which we add actors, we can make node $i_n = i$ the last arrival. If t_{-i} is the tree at the end of the penultimate loop then

$$\pi_{\mathcal{T}_o}(t_{-i}|q) = \sum_{e \in E(t_{-i})} \pi_{\mathcal{T}_{\tilde{o}}}(t_{-i} \triangleleft (e, i)|q). \tag{A.2}$$

Now take $\tau = t_{-i}$. Since leaf deletion reverses edge insertion, the set of trees $\mathcal{T}_{[n]}(\tau)$ that contain τ is the set of trees that are obtained from τ by some edge addition,

$$\mathcal{T}_{[n]}(\tau) = \bigcup_{e \in E(\tau)} \{\tau \triangleleft (e, i)\}$$

and so

$$\pi_{\mathcal{T}_o}(\tau|q) = \sum_{t \in \mathcal{T}_{[n]}(\tau)} \pi_{\mathcal{T}_{\tilde{o}}}(t|q).$$

which is marginal consistency for addition of one actor.

Proposition 4 *The probability distribution over VSPs given in (4) is marginally consistent.*

Proof A.3 (Proposition 4) *It is sufficient to show that Eqn. 5 holds for $\tilde{o} = [n]$ and $o = [n] \setminus \{i\}$ and any $i \in [n]$ in Definition 1 since Eqn. 5 follows for any pair of subsets of $[n]$ by removing actors one at a time.*

In this case $v[o]$ is the suborder obtained from $v \in \mathcal{V}_{[n]}$ by removing actor i and we want to verfiy

$$\pi_{\mathcal{V}_o}(w|q) = \sum_{\substack{v \in \mathcal{V}_{[n]} \\ v[o]=w}} \pi_{\mathcal{V}_{[n]}}(v|q) \quad \textit{for all } w \in \mathcal{V}_o. \tag{A.3}$$

Picking up the RHS of Eqn. A.3 we have from Eqn. 4

$$\sum_{\substack{v \in \mathcal{V}_{[n]} \\ v[o]=w}} \pi_{\mathcal{V}_{[n]}}(v|q) = \sum_{\substack{v \in \mathcal{V}_{[n]} \\ v[o]=w}} \sum_{t \in t(v)} \pi_{\mathcal{T}_{[n]}}(t|q).$$

Referring to Definition A.3, the sum on the right is a sum over all trees "containing" a tree in $t(w)$, that is, the set of all trees which can be constructed by taking a tree $\tau \in t(w)$ and adding actor i to the tree by edge insertion at any edge in τ:

$$\bigcup_{\substack{v \in \mathcal{V}_{[n]} \\ v[o]=w}} \bigcup_{t \in t(v)} \{t\} = \bigcup_{\tau \in t(w)} \bigcup_{e \in E(\tau)} \{\tau \lhd (e,i)\}.$$

It follows that

$$\begin{aligned} \sum_{\substack{v \in \mathcal{V}_{[n]} \\ v[o]=w}} \pi_{\mathcal{V}_{[n]}}(v|q) &= \sum_{\tau \in t(w)} \sum_{e \in E(\tau)} \pi_{\mathcal{T}_{[n]}}(\tau \lhd (e,i)|q) \\ &= \sum_{\tau \in t(w)} \pi_{\mathcal{T}_o}(\tau|q), \ \textit{(by Eqn. A.2)} \\ &= \pi_{\mathcal{V}_o}(w|q) \qquad\qquad \textit{(by Eqn. 4)}, \end{aligned}$$

which is the LHS of Eqn. A.3.

This concludes the first part of Theorem 1. We now prove the second part.

A.2 PART II: CLOSED FORM PRIOR

The following proof makes use of the MDT representation of a VSP introduced in Section 1.1 and detailed in A.3 below.

We next observe that all BDTs representing the same VSP have equal prior probabilities (they collapse to the same MDT and that fixes $S(t)$). This makes it easy to do the sum in (4) as the summand is constant.

Proposition 5 (Probability Distribution over VSPs) *The prior probability for a VSP with n nodes is*

$$\pi_{\mathcal{V}_{[n]}}(v|q) = |t(v)| \pi_{\mathcal{T}_{[n]}}(t|q),$$

for any tree $t \in t(v)$.

Proof A.4 (Proposition 5) *For $v \in \mathcal{V}_{[n]}$, any two trees $t, t' \in t(v)$ are both in $\mathcal{T}_{[n]}$. They also satisfy $S(t) = S(t')$. This follows from Lemma 1: if these numbers differ then the S-clusters of t and t' cannot all have equal sizes; the S-cluster sizes of a BDT determine of the numbers of children of the S-nodes in its MDT; it follows that $m = t_{\mathcal{M}}(t)$ and $m' = t_{\mathcal{M}}(t')$ cannot be isomorphic (identifying leaves by actor labels); but m and m' are then distinct MDT's for v which contradicts Lemma 1. Referring to Eqn. 3 we see that $\pi_{\mathcal{T}_{[n]}}(t|q)$ is constant over $t \in t(v)$ so the sum in Eqn. 4 just counts trees in $t(v)$.*

Finally, we count trees in $t(v)$ and this gives us the closed form we seek. This seems to be new.

Proposition 6 *Let $t \in t(v)$ be an arbitrary BDT of a VSP $v \in \mathcal{V}_{[n]}$ with P- and S-clusters defined as in Theorem 1. The number of BDTs of v is*

$$|t(v)| = \prod_{k=1}^{K_P} (|2C_k^{(P)}| - 1)!! \prod_{k'=1}^{K^S} \mathcal{C}_{|C_{k'}^{(S)}|} \tag{A.4}$$

with $\mathcal{C}_s, \ s \geq 0$ given in (7).

Proof A.5 (Proposition 6) *By Lemma 1 the set of BDT trees $t(v)$ for any $v \in \mathcal{V}_{[n]}$ is identical to the set $t_{\mathcal{M}}(m) = \{t \in \mathcal{T}_{[n]} : m_{\mathcal{T}}(t) = m\}$ when $m = m_{\mathcal{V}}(v)$ so we need to count the number of BDT's that collapse down to the same MDT. Let $m = (F, E, L)$ be an MDT with leaves $\mathcal{F}$ and internal nodes $\mathcal{A}$.*

A P-node $i \in \mathcal{A}$ in m having c child nodes is generated by collapsing some P-cluster $C_k^{(P)}$ of a BDT $t \in t_{\mathcal{M}}(m)$ with $|C_k^{(P)}| = c - 1$ nodes "internal" to the P-cluster. This P-cluster corresponds to a sub-tree $t_k = (V(t_k), E(t_k))$ with vertices $V(t_k) = C_k^{(P)}$ and edges

$$E(t_k) = E(t) \cap (C_k^{(P)} \times C_k^{(P)}).$$

The sub-tree t_k has $c = |C_k^{(P)}| + 1$ leaves. If we replace t_k with any tree with $|C_k^{(P)}| + 1$ labelled leaves then it collapses to a MDT node with in- and out-edges isomorphic to those of node i in m. The number of such trees is $(2|C_k^{(P)}| - 1)!!$.

An S-node $i \in \mathcal{A}$ of the MDT with s child nodes and stacking data $L_i(m) = (i_1, ..., i_s)$ is generated by collapsing some S-cluster $S_k^{(S)}$ of a BDT. Again, that cluster covers $|S_k^{(P)}| = s - 1$ internal nodes in the BDT. This S-cluster corresponds to a sub-tree of t with $s = |S_k^{(P)}| + 1$ leaves. Since all the internal nodes of the sub-tree are of type S and its leaf nodes are labelled, this sub-tree is a BDT representing the fixed total order $i_1 \succ i_2 ... \succ i_s$ on its leaf nodes. If we replace this subtree with any tree with s labelled leaves representing the same total order then it collapses to a MDT node with in- and out-edges isomorphic to i and the same stacking data. The number of such trees is given by the Catalan number $\mathcal{C}_{s-1} = \mathcal{C}_{|S_k^{(P)}|}$. This can be shown by the following induction.

The number of BDT's representing a total order on 1 or 2 elements is one and indeed $\mathcal{C}_0 = \mathcal{C}_1 = 1$. Suppose the number of BDT's representing a total order $1 \succ 2 \succ ... \succ s$ is $\mathcal{C}_{s-1}$ and consider a BDT representing $1 \succ 2 \succ ... \succ s+1$. The root of such a BDT must partition the leaves into $1, ..., k$ and $k+1, ..., s+1$ for some $1 \leq k \leq s$ so that the root stacks $1, ..., k$ above $k+1, ..., s+1$. By the induction hypothesis the number of subtrees representing $1 \succ 2 \succ ... \succ k$ is $\mathcal{C}_{k-1}$ and the number representing $k+1 \succ 2 \succ ... \succ s+1$ is $\mathcal{C}_{s-k-1}$, so the number of BDT's splitting the leaves into $1, ..., k$ and $k+1, ..., s+1$ is $\mathcal{C}_{k-1}\mathcal{C}_{s-k-1}$. The total number of BDT's representing $1 \succ 2 \succ ... \succ s+1$ is then

$$\begin{aligned} \sum_{k=1}^{s} \mathcal{C}_{k-1}\mathcal{C}_{s-k-1} &= \sum_{k=0}^{s} \mathcal{C}_k \mathcal{C}_{s-k} \\ &= \mathcal{C}_s, \end{aligned}$$

where the last step is given in Stanley and Weisstein [2002].

The total number of BDT's is given by the product over the internal nodes of the MDT of the numbers of BDT sub-trees which collapse to give those nodes. This gives Eqn. A.4 and completes the proof of Theorem 1.

A.3 MULTI-DECOMPOSITION TREES

A MDT $m \in \mathcal{M}_{[n]}$ is a tree $m = (F(m), E(m), L(m))$ with n leaves and edges $E(m)$ directed from the root to the leaves. Let $\mathcal{F}$ and $\mathcal{A}$ be the index sets for the leaves and internal nodes, such that $|\mathcal{F}| = n$ and $1 \leq |\mathcal{A}| \leq n - 1$. An internal node $i \in \mathcal{A}$ of a MDT may have any number of child nodes between two and $n - 1$. For $i \in \mathcal{F}$ and $m \in \mathcal{M}_{[n]}$, the array $F_i(m) \in [n]$ records the actor represented by leaf node i. The internal nodes $i \in \mathcal{A}$ are either of type S or type P. The key defining feature of an MDT is that the internal nodes of an MDT which are adjacent must have unequal types.

Let $S(m)$ be the number of S-nodes in multi-tree $m \in \mathcal{M}_{[n]}$. For $m \in \mathcal{M}_{[n]}$ let $v(m) :\in \mathcal{V}_{[n]}$ map an MDT to its corresponding VSP and for $i \in \mathcal{F} \cup \mathcal{A}$ let $m_i(m)$ denote the sub-tree rooted by node i. If $i \in \mathcal{A}$ is of type P with k children $j_1, \ldots, j_k$, then

$$v(m_i(m)) = v(m_{j_1}(m)) \oplus \cdots \oplus v(m_{j_k}(m)).$$

If $i \in \mathcal{A}$ is of type S with k child nodes $\{j_1, \ldots, j_k\} = \{j \in \mathcal{F} \cup \mathcal{A} : \langle i, j \rangle \in E(m)\}$, an ordered set $L_i = (j_1, \ldots, j_k)$ gives the stacking order (with j_1 at the top) for the sub-trees rooted by the children of i. It follows that

$$v(m_i(m)) = v(m_{j_1}(m)) \otimes \cdots \otimes v(m_{j_k}(m)).$$

Let $L(m) = \{L_i\}_{i \in \mathcal{A}}$ with $L_i = \emptyset$ if i is a P-node. Adjacent internal nodes have unequal type so if $\langle i, j \rangle \in E(m)$ then exactly one of L_i and L_j is empty. In this notation a MDT tree is a BDT if all its internal nodes have two child nodes and a BDT is an MDT if all adjacent internal nodes have different S/P-types.

An MDT can be formed from a BDT by collapsing edges between internal nodes in the BDT which have the same type while preserving information about stacking order at S-nodes. This collapses P- and S-clusters to a single node. A set of BDT's can be recovered from an MDT by "unpacking" internal nodes of the MDT with more than two child nodes in different ways. For $t \in \mathcal{T}_{[n]}$ let $m_{\mathcal{T}}(t) \in \mathcal{M}_{[n]}$ map the BDT t to its corresponding MDT. See Figure 4 for an example.

Counting linear extensions in the MDT formulation is similar to the BDT case (Eqns. 1 & 2).

$$|\mathcal{L}(h_1 \otimes \cdots \otimes h_n)| = |\mathcal{L}(h_1)| \times \cdots \times |\mathcal{L}(h_n)| \tag{A.5}$$

$$|\mathcal{L}(h_1 \oplus \cdots \oplus h_n)| = |\mathcal{L}(h_1)| \times \cdots \times |\mathcal{L}(h_n)| \binom{|V(h_1)| + \cdots + |V(h_n)|}{|V(h_1)|, \ldots, |V(h_n)|} \tag{A.6}$$

where $|V(h_1)|$ and $|V(h_2)|$ give the number of actors in h_1 and h_2. This may be evaluated recursively in $O(n)$ steps.

A.4 PROOF OF PROPOSITION 1

Proposition 1 (Posterior Marginals) *Sampling the BDT posterior $(t, q, \psi) \sim \pi_{\mathcal{T}_{[n]}}(\cdot|y)$ gives samples $(v(t), q, \psi) \sim \pi_{\mathcal{V}_{[n]}}(\cdot|y)$ from the VSP posterior.*

Proof A.6 (Proposition 1) *Eqn. 11 is the marginal over $t \in t(v)$ of Eqn. 10: if $(t, q, \psi) \sim \pi_{\mathcal{T}_{[n]}}(\cdot|y)$ then the new joint distribution at $v(t) = v$ is*

$$\begin{aligned} p(v, q, \psi) &\propto \sum_{t' \in t(v)} \pi_{\mathcal{T}_{[n]}}(t'|q)\pi(q, \psi)Q(y|v(t'), \psi) \\ &= \pi(q, \psi)Q(y|v, \psi) \sum_{t' \in t(v)} \pi_{\mathcal{T}_{[n]}}(t'|q) \\ &= \pi_{\mathcal{V}_{[n]}}(v, q, \psi|y) \end{aligned}$$

as $Q(y|v(t), \psi) = Q(y|v, \psi)$ is a constant for $t \in t(v)$ and the prior marginalises to $\pi_{\mathcal{V}_{[n]}}(v|q)$ by Eqn. 4.

B QUEUE-JUMPING MODELS

B.1 QUEUE-JUMPING UP/DOWN OBSERVATION MODEL

Let $L_T(v) = |\mathcal{L}[v]|$ be the number of linear extensions of VSP $v \in \mathcal{V}_{[n]}$ and for $i \in [n]$ let $T_i(v) = |\{l \in \mathcal{L}[v] : l_1 = i\}|$ give the number of linear extensions with actor i at the top. The observation model for QJ-U for a generic list $x \in \mathcal{P}_{[n]}$ is

$$Q_{up}(x|v,p) = \prod_{i=1}^{n-1} \left(\frac{p}{n-i+1} + (1-p) \frac{T_{x_i}(v[x_{i:n}])}{L_T(v[y_{i:n}])} \right). \tag{B.1}$$

We can interpret this as the distribution over lists determined by a process in which the list is formed by building it up one element at a time from the top, choosing the next actor at random from those that remain with probability p and otherwise choosing the next actor as the first actor in a list drawn from the noise free model (beginning of Section 3) applied to the remaining actors. Fig. B.1 gives an example list realisation for VSP v_0. We give the generative model alg.B.1.

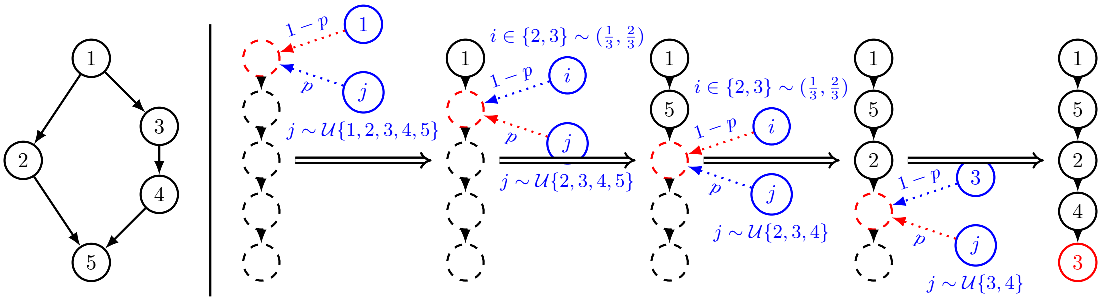

1
2
3
4
5
$1-p$
p
1
j
$j \sim \mathcal{U}\{1,2,3,4,5\}$
$i \in \{2,3\} \sim (\frac{1}{3}, \frac{2}{3})$
1
$1-p$
i
p
j
$j \sim \mathcal{U}\{2,3,4,5\}$
1
5
$i \in \{2,3\} \sim (\frac{1}{3}, \frac{2}{3})$
$1-p$
i
p
j
$j \sim \mathcal{U}\{2,3,4\}$
1
5
2
$1-p$
3
p
j
$j \sim \mathcal{U}\{3,4\}$
1
5
2
4
3

Figure B.1: One example list simulation process from the VSP v_0 (left) via the QJ-U observation model. The simulated list is displayed on the right.

Algorithm B.1 Simulation algorithm for QJ-U.

```
Require: $v \in \mathcal{V}_{[n]}, p \in [0,1]$
Ensure: $x \sim Q_{up}(\cdot|v,p)$
  $i \leftarrow 1, s \leftarrow [n], v' \leftarrow v$
  while $|s| > 0$ do
      $q \leftarrow (T_j(v')/L(v'))_{j \in s}$
      Sample $c \sim Bern(1-p)$
      if $c = 0$ then
          Sample $x_i \sim \mathcal{U}(s)$
      else if $c = 1$ then
          Sample $x_i \sim$ multinom$(q)$
      end if
      $s \leftarrow s \backslash x_i$
      $i \leftarrow i+1, v' \leftarrow v[s]$
  end while
  return $x = (x_1, \dots, x_n)$
```

The output $x \sim Q_{up}(\cdot|v,p)$ is a random list of n elements distributed according to Q_{up}. This follows because the probabilities to choose entries in x at each step are just the factors in Q_{up} in Eqn. B.1. We can turn the model around and build the list from the bottom, allowing "queue jumping-down". If we set $p = 0$, we get a telescoping product and $Q_{up}(l|v, p = 0) = 1/L_T(v)$ for $l \in \mathcal{L}[v]$, so we recover the error-free model. Lists are assumed to be drawn independently, and the actors present $o_j,\ j = 1, ..., N$ are known, so the likelihood is

$$Q(y|v,p) = \prod_{j=1}^{N} Q(y_j|v[o_j], p).$$

Here $Q = Q_{up}$ (and $Q = Q_{bi}$ in the next section).

B.2 BI-DIRECTIONAL QUEUE-JUMPING MODEL

Similar to QJ-U, QJ-B ranks by repeated selection - but from both ends. We either rank from the top with probability ϕ or from the bottom with probability $1-\phi$. From the top (bottom) of the list, the next actor is chosen at random from those that remain with probability p, and otherwise as the first (last) actor in a list drawn from the noise free model. An example simulation process from VSP v_0 is visualised in Fig. 5.

Algorithm B.2 gives the simulation algorithm for the bi-directional queue jumping model. It introduces one extra step in each loop of algorithm B.1 in which we randomly choose the top/bottom fill-direction to place the next actor in the realised list with probability ϕ.

Algorithm B.2 Simulation algorithm for QJ-B.

```
Require: $v \in \mathcal{V}_{[n]}, p \in [0,1], \phi \in [0,1]$
Ensure: $x \sim Q_{bi}(x|v,p,\phi)$
  $s \leftarrow [n], v' \leftarrow v$
  $x \leftarrow (\emptyset,\dots,\emptyset) \in \{\emptyset\}^n$, $k \leftarrow 1$, $U_0 \leftarrow 0$, $D_0 \leftarrow n+1$
  while $|s| > 0$ do
      Sample $z_k \sim Bern(1-\phi)$
      $U_k \leftarrow U_{k-1} + z_k$, $D_k \leftarrow D_{k-1} - (1-z_k)$
      $i_k \leftarrow z_k U_k + (1-z_k) D_k$
      Sample $c_k \sim Bern(1-p)$
      if $c_k = 0$ then
          Sample $a \sim \mathcal{U}(s)$
      else
          if $z_k = 0$ then
              $q \leftarrow (T_a(v')/L_T(v'))_{a \in s}$
              Sample $a \sim$ multinom$(q)$
          else if $z_k = 1$ then
              $q \leftarrow (B_a(v')/L_T(v'))_{a \in s}$
              Sample $a \sim$ multinom$(q)$
          end if
      end if
      Set $x_{i_k} \leftarrow a,\ k \leftarrow k+1,\ s \leftarrow s \backslash a,\ v' \leftarrow v[s]$
  end while
  return $x = (x_1,\dots,x_n)$
```

B.3 RECURSIVE EVALUATION ALGORITHM FOR QJ-B

This sub-section gives Algorithm B.3, an algorithm for recursive evaluation of the QJ-B likelihood.

Algorithm B.3 Recursion evaluating Q_{bi} in Eqn. 9

```
procedure $f(v, x, p, \phi)$
    $n = |v|$
    if $n = 1$ then
        return 1
    end if
    if $\phi > 0$ then
        $l_0 \leftarrow \frac{p}{n} + (1-p)\frac{T_{x_1}(v)}{L_T(v)}$
        $x \leftarrow x_{2:n}, v \leftarrow v[x]$
        $P_0 = \phi \times l_0 \times f(v, x, p, \phi)$
    else $L_0 = 0$
    end if
    if $\phi < 1$ then
        $l_1 \leftarrow \frac{p}{n} + (1-p)\frac{B_{x_n}(v)}{L_T(v)}$
        $x \leftarrow x_{1:n-1}, v \leftarrow v[x]$
        $P_1 = (1-\phi) \times l_1 \times f(v, x, p, \phi)$
    else $P_1 = 0$
    end if
    return $P_0 + P_1$
end procedure
```

We now show this algorithm is correct.

Let $X \sim Q_{bi}$ be a random list with realisation $X = x$. For sub-list $x_{a:b},\ 1 \le a < b \le n$ let

$$P_{a|a:b} = p(X_a = x_a | z_k = 0, v[x_{a:b}], p),$$
$$P_{b|a:b} = p(X_b = x_b | z_k = 1, v[x_{a:b}], p),$$
$$P_{a:b} = p(X_{a:b} = x_{a:b} | v[x_{a:b}], p, \phi),$$

so that $P_{1:n} = Q_{bi}(x|v, p, \phi)$ and $P_a = 1$ when $a = b$.

Proposition 7

$$P_{a:b} = \phi P_{a|a:b} P_{a+1:b} + (1-\phi) P_{b|a:b} P_{a:b-1}, \tag{B.2}$$

and $f(v, x, p, \phi)$ in Algorithm B.3 returns $Q_{bi}(x|v, p, \phi)$.

Proof B.1 (Proposition 7) *First of all, if Eqn. B.2 holds then a call to $f(v[x_{a:b}], x_{a:b}, p, \phi)$ evaluates $l_0 = P_{a|a:b}$, $l_1 = P_{b|a:b}$ and returns the sum of $\phi l_0 f(v[x_{a:b}], x_{a:b}, p, \phi)$ and $(1-\phi) l_1 f(v[x_{a:b-1}], x_{a:b-1}, p, \phi)$. Then since $f(v[x_a], x_a, p, \phi) = P_a = 1$ we have by induction (and Eqn. B.2) that $f(v[x_{a:b}], x_{a:b}, p, \phi) = P_{a:b}$ and*

$$f(v, x, p, \phi) = Q_{bi}(x|v, p, \phi).$$

We now show Eqn. B.2) holds for the distribution of sub-lists $X_{a:b}$ of $X \sim Q_{bi}$. If $a : b$ remain to be realised then $a-1+n-(b-1)$ entries in X have been realised and this would occur as we enter step $k = n+a-b+1$ of Algorithm B.2. Partitioning on the value of z_k,

$$\begin{aligned}
P_{a:b} &= p(X_{a:b} = x_{a:b} | v[x_{a:b}], p, \phi) \\
&= p(z_k = 0|\phi) p(X_{a:b} = x_{a:b} | z_k = 0, v[x_{a:b}], p, \phi) \\
&\quad + \ p(z_k = 1|\phi) p(X_{a:b} = x_{a:b} | z_k = 1, v[x_{a:b}], p, \phi) \\
&= \phi P_{a|a:b} p(x_{a+1:b} | v[x_{a+1:b}], p, \phi) \\
&\quad + \ (1-\phi) P_{b|a:b} p(x_{a:b-1} | v[y_{a:b-1}], p, \phi), \\
&= \phi P_{a|a:b} P_{a+1:b} + (1-\phi) P_{b|a:b} P_{a:b-1}.
\end{aligned}$$

C MCMC SAMPLER

We use Metropolis-Hasting MCMC to sample posterior distributions. We can target either distribution in Proposition 1.

C.1 MCMC SAMPLER IN THE BDT REPRESENTATION

We start with MCMC targeting BDT. This was the method we implemented as the data structures seem slightly simpler. However, we would expect MCMC targeting the VSP posterior directly to be a little more efficient, as MCMC targeting the BDT posterior wastes time exploring latent subspaces $t(v)$ without changing v. Tree sampling requires edge operations on trees (called "subtree prune and regraft" (OP-PR) in the phylogenetics literature). For this purpose we assume the 0-node with an edge to the root of the BDT is restored, so $0 \in \mathcal{F}$ for a regraft above the root. Let $\mathcal{F}_{-0} = \mathcal{F} \setminus \{0\}$ and $E_{-0}(t) = E(t) \setminus \{\langle e_1, e_2\rangle \in E(t) : e_1 = 0\}$.

Definition C.1 (Subtree Prune and Regraft on a BDT) *For $t = (F(t), E(t), L(t))$, $t \in \mathcal{T}_{[n]}$ a BDT with leaf node labels $\mathcal{F}$ and internal node labels $\mathcal{A}$, an edge operation $t' = t \lhd_e (e, e')$ moves edge $e = \langle e_1, e_2\rangle$, $e \in E_{-0}(t)$ to edge $e' = \langle e_1', e_2'\rangle$, $e' \in E(t')$. The leaf-to-actor map $F(t') = F(t)$ is unchanged. Let*

$$f_p(j|t) = \{i \in \mathcal{A} | \langle i, j\rangle \in E(t)\}$$

give the parent of $j \in \mathcal{F}_{-0} \cup \mathcal{A}$ with $f_p(r|t) = 0$ if r is the root. Let

$$f_c(i|t) = \{j_1, j_2 \in \mathcal{F} \cup \mathcal{A} | \{\langle i, j_1\rangle, \langle i, j_2\rangle \subset E(t)\}$$

give the children of $i \in \mathcal{A}$.Let $\overleftarrow{e}_1 = f_p(e_1|t)$ give the parent of e_1 and $\vec{e}_2 = f_c(e_1|t)\backslash\{e_2\}$ give the "sibling" of e_2 in t (the child of e_1 which is not e_2). Then

$$\begin{aligned} E(t') = E(t)\backslash\{e', \langle \overleftarrow{e}_1, e_1\rangle, \langle e_1, \vec{e}_2\rangle\} \\ \cup \{\langle e_1', e_1\rangle, \langle e_1, e_2'\rangle, \langle \overleftarrow{e}_1, \vec{e}_2\rangle\}. \end{aligned}$$

Set $L(t') = L(t)$ and make the following replacements as needed. If $L_{\overleftarrow{e}_1}(t) \neq \emptyset$ then $L_{\overleftarrow{e}_1}(t)$ is an ordered set containing two edges. Set $L_{\overleftarrow{e}_1}(t') = L_{\overleftarrow{e}_1}(t)\backslash\{e_1\} \cup \{\vec{e}_2\}$ where the replacement enters the vacated position in the ordered set. If $L_{e_1'}(t) \neq \emptyset$, $L_{e_1'}(t') = L_{e_1'}(t)\backslash\{e_2'\} \cup \{e_1\}$. If $L_{e_1}(t) \neq \emptyset$ then take $L_{e_1}(t') \sim \mathcal{U}\{(e_2, e_2'), (e_2', e_2)\}$.

The edge operation $t \lhd_e (e, e')$ moves the sub-tree rooted by e_2 into edge e', breaking that edge and inserting node e_1. The S/P-type of e_1 travels with e_1, and if it is S we must assign a stacking order to the subtrees rooted by e_2' and e_2. Figure C.1 illustrates an example edge operation.

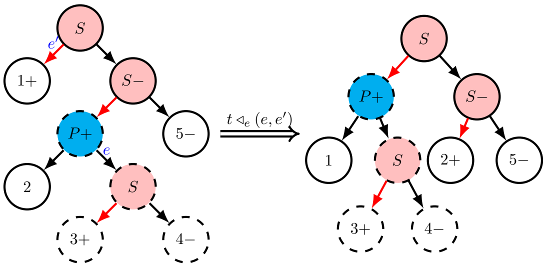


Figure C.1: An example OP-PR edge operation on BDT t_0.

The tree updates in our MCMC admit both local and global edge operations. In the local edge operation, an edge can only be moved to a neighboring edge, i.e. if $e = \langle e_1, e_2\rangle$, e' is selected from e's neighboring edges $E_l(e|t)$ such that

$$E_l(e|t) = \{\langle e_1', e_2'\rangle \in E(t) \mid e_2' = \overleftarrow{e}_1 \text{ or } e_1' = \vec{e}_2 \text{ or } e_2' = \vec{e}_1\}.$$

These "small" changes have a higher acceptance rate. The global edge operation moves an edge e to any $e' \in E(t)\backslash e$. For $t \in \mathcal{T}_{[n]}$, we typically perform 1 global edge operation for every n local edge operations. We present the MCMC algorithm for BDT with the QJ-B observation model in Algorithm C.1, omitting the standard q, p and ϕ updates. A simple internal node type update is included. The algorithm for QJ-U observation model is similar but without the ϕ-update.

Algorithm C.1 The MCMC algorithm for the BDT with QJ-B observation model at step k.

```
Require: $y, t^{(k-1)} = t, q^{(k-1)} = q, p^{(k-1)} = p, \phi^{(k-1)} = \phi$ with $t = (F(t), E(t), L(t)),\ t \in \mathcal{T}_{[n]}$.
Ensure:
    $t^{(k)} \sim \pi(t|y, q, p, \phi),$
    $q^{(k)} \sim \pi(q|y, t^{(k)}, p, \phi),$
    $p^{(k)} \sim \pi(p|y, t^{(k)}, q^{(k)}, \phi),$
    $\phi^{(k)} \sim \pi(\phi|y, t^{(k)}, q^{(k)}, p^{(k)})$

  function TYPE($i|t$)
      if $L_i(t) = \emptyset$ then
          return $P$
      else
          return $S$
      end if
  end function
  ______________ Update for t (internal node type) ______________
  $t' \leftarrow t^{(k)} \leftarrow t$
  Sample $i \sim \mathcal{U}(\mathcal{A})$
  if TYPE($i|t$)=$P$ then
      Sample $z \sim \mathcal{U}\{0, 1\}$
          $L_i(t') \leftarrow z f_c(i|t)[(1, 2)] + (1 - z) f_c(i|t)[(2, 1)]$
          $\eta_1 \leftarrow \frac{2 \times Q(y|v(t'), p, \phi)\pi_{\mathcal{T}_{[n]}}(t'|q)}{Q(y|v(t), p, \phi)\pi_{\mathcal{T}_{[n]}}(t|q)}$
  else if TYPE($i|t$)=$S$ then
          $L_i(t') \leftarrow \emptyset$
          $\eta_1 \leftarrow \frac{Q(y|v(t'), p, \phi)\pi_{\mathcal{T}_{[n]}}(t'|q)}{2Q(y|v(t), p, \phi)\pi_{\mathcal{T}_{[n]}}(t|q)}$
  end if
  if $\mathcal{U}(0, 1) \leq \eta_1$ then
      $t \leftarrow t^{(k)} \leftarrow t'$
  end if
  ______________ Update for t (global edge operation) ______________
  Sample $e \sim \mathcal{U}(E_{-0}(t)), e' \sim \mathcal{U}(E(t) \backslash e)$
          $t' \leftarrow t \triangleleft_e (e, e')$
          $\eta_2 \leftarrow \frac{Q(y|v(t'), p, \phi)\pi_{\mathcal{T}_{[n]}}(t'|q)}{Q(y|v(t^{(k)}), p, \phi)\pi_{\mathcal{T}_{[n]}}(t|q)}$
  if $\mathcal{U}(0, 1) \leq \eta_2$ then
      $t \leftarrow t^{(k)} \leftarrow t'$
  end if
  ______________ Update for t (local edge operation) ______________
  Sample $e \sim \mathcal{U}(E_{-0}(t)), e' \sim \mathcal{U}(E_l(e|t))$
          $t' \leftarrow t \triangleleft_e (e, e')$
          $\eta_3 \leftarrow \frac{Q(y|v(t'), p, \phi)\pi_{\mathcal{T}_{[n]}}(t'|q)|E_l(e|t)|}{Q(y|v(t), p, \phi)\pi_{\mathcal{T}_{[n]}}(t|q)|E_l(e|t')|}$
  if $\mathcal{U}(0, 1) \leq \eta_3$ then
      $t \leftarrow t^{(k)} \leftarrow t'$
  end if
  ______________ Updates for $q, p$ and $\phi$ omitted ______________
```

C.2 MCMC SAMPLER IN THE MDT REPRESENTATION

We can target the VSP-posterior directly. Since MDT's are one-to-one with VSP's, we can parameterise using MDT's and define (in Defn. C.2) a sub-tree prune and regraft operator for MDT's.

Definition C.2 (Subtree Prune and Regraft on a MDT) *For $m = (F(m), E(m), L(m)),\ m \in \mathcal{M}_{[n]}$ a MDT with leaf node labels $\mathcal{F}$ and internal nodes labels $\mathcal{A}$, an edge operation $m' = m \triangleleft_e (e, i)$ creates a new MDT with nodes $\mathcal{F}', \mathcal{A}'$, moving edge $\mathrm{e} = \langle \mathrm{e}_1, \mathrm{e}_2 \rangle,\ \mathrm{e} \in E_{-0}(m)$ onto node $i \in (\mathcal{F} \cup \mathcal{A}) \backslash \{\mathrm{e}_1, \mathrm{e}_2\}$.*

We need at most $2n$ node labels below. Assume $\mathcal{F}_{-0} \cup \mathcal{A} \subset [2n]$ and let $pop(\mathcal{F}, \mathcal{A}) = \min([2n] \setminus (\mathcal{F} \cup \mathcal{A}))$ be a function we call when we need a new node label. There are three types of edge operation.

1. *$i \in \mathcal{A}$: we connect e to node i.*
   Here $F(m') = F(m)$ and
   $$E(m') = E(m) \backslash \{\mathrm{e}\} \cup \langle i, \mathrm{e}_2 \rangle.$$
   Set $L(m') = L(m)$ and make the following changes as needed. If $L_{\mathrm{e}_1}(m) \neq \emptyset$ then set $L_{\mathrm{e}_1}(m') = L_{\mathrm{e}_1}(m) \backslash \{\mathrm{e}_1\}$. If $L_i(m) \neq \emptyset$ then suppose $L_i(m) = (j_1, \dots, j_k)$. Take $L_i(m') \sim \mathcal{U}\{(\mathrm{e}_1, j_1, \dots, j_k), \dots, (j_1, \dots, j_k, \mathrm{e}_1)\}$ (insert the subtree below $\langle \mathrm{e}_1, \mathrm{e}_2 \rangle$ uniformly in the stack under i).
2. *$i \in \mathcal{F}$: we connect e into edge $\langle \overleftarrow{i}, i \rangle$ with $\overleftarrow{i} = f_p(i|m)$ and add an additional internal node $j = pop(\mathcal{F}, \mathcal{A})$.*
   Here $F(m') = F(m)$ and
   $$E(m') = E(m) \backslash \{\mathrm{e}, \langle \overleftarrow{i}, i \rangle\} \cup \{\langle \overleftarrow{i}, j \rangle, \langle j, i \rangle, \langle j, \mathrm{e}_2 \rangle\}.$$
   Set $L(m') = L(m)$ and make the following changes as needed. If $L_{\mathrm{e}_1}(m) \neq \emptyset$ then set $L_{\mathrm{e}_1}(m') = L_{\mathrm{e}_1}(m) \setminus \{\mathrm{e}_1\}$. If $L_{\overleftarrow{i}}(m) \neq \emptyset$ (parent is S), suppose $L_{\overleftarrow{i}}(m) = (j_1, \dots, i, \dots, j_k)$. Set $L_{\overleftarrow{j}}(m') = (j_1, \dots, j, \dots, j_k)$ and $L_j(m') = \emptyset$ (new child is P). Finally, if $L_{\overleftarrow{i}}(m) = \emptyset$ (parent is P), take $L_j(m) \sim \mathcal{U}\{(i, \mathrm{e}_2), (\mathrm{e}_2, i)\}$ (new child is S).
3. *$i = 0$: connect e into the edge above the root, $r = f_c(0|m),\ r \in \mathcal{A}$ and add an additional internal node $j = pop(\mathcal{F}, \mathcal{A})$ which will root m'.*
   Here $F(m') = F(m)$ and
   $$E(m') = E(m) \backslash \mathrm{e} \cup \{\langle 0, j \rangle, \langle j, r \rangle, \langle j, \mathrm{e}_2 \rangle\}.$$
   Set $L(m') = L(m)$ and make the following changes as needed. If $L_{\mathrm{e}_1}(m) \neq \emptyset$ then set $L_{\mathrm{e}_1}(m') = L_{\mathrm{e}_1}(m) \backslash \{\mathrm{e}_1\}$. If $L_r(m) \neq \emptyset$ (child is S), we define $L_j(m') = \emptyset$ (new node is P). Otherwise, if $L_r(m) = \emptyset$ (child is P), we take $L_j(m') \sim \mathcal{U}\{(r, \mathrm{e}_2), (\mathrm{e}_2, r)\}$ (new node is S).

Figure C.2 illustrates an example edge operation on a MDT. Moving an edge $e = \langle e_1, e_2 \rangle$ may increase or decrease the number of edges and internal nodes. For example, if in case (1) $f_c(e_1|m) = \{e_2, \vec{e}_2\}$, moving e replaces $\langle \overleftarrow{e}_1, e_1 \rangle, \langle e_1, \vec{e}_2 \rangle$ with $\langle \overleftarrow{e}_1, \vec{e}_2 \rangle$ and e_1 is removed. If e is attached in an existing internal node $i \in \mathcal{A}$ then the number of nodes and edges each go down by one.

If we take $e \sim \mathcal{U}(E_{-0}(m))$ and $i \sim \mathcal{U}[(\mathcal{F} \cup \mathcal{A}) \setminus \{e_1, e_2\}]$ and set $m' = m \triangleleft_e (e, i)$ as given in Defn. C.2 then the proposal probability $\rho(m'|m)$ depends on e and i. A simple generic expression is

$$\rho(m'|m) = \frac{1}{|E(m)|} \times \frac{1}{|\mathcal{F} \cup \mathcal{A}| - 2} \times \rho_{m,m'} \tag{C.1}$$

where $\rho_{m,m'}$ is given as follows: (Case 1) $\rho_{m,m'} = 1/(c_i + 1)$ if i is internal and has c_i child nodes and type S (e_1 must be placed in the stack below i) and $\rho_{m,m'} = 1$ if i is internal and type P; (Case 2) $\rho_{m,m'} = 1/2$ if i is a leaf and $\overleftarrow{i}$ is type P (as i and e_2 must be stacked) and $\rho_{m,m'} = 1$ if i is leaf and $\overleftarrow{i}$ is type S; (Case 3) $\rho_{m,m'} = 1/2$ if $i = 0$ and $r = f_c(0|m)$ is type P (as r and e_2 must be stacked) and $\rho_{m,m'} = 1$ if $i = 0$ and r is type S.

Not every operation is admissible: if $f_c(e_1|m) = \{e_2, \vec{e}_2\}$ and $\vec{e}_2$ is not a leaf, then $\vec{e}_2$ and $\overleftarrow{e}_1$ must have the same type. An edge $\langle \overleftarrow{e}_1, \vec{e}_2 \rangle$ would then connect two internal nodes of the same type and so $m' \notin \mathcal{M}_{[n]}$. In Eqn. C.1, $\rho(m'|m)$ has a simple form because we do not "keep trying till we get $m' \in \mathcal{M}_{[n]}$". We know $m' \notin \mathcal{M}_{[n]}$ is a possible outcome for m', but we don't try to write down $\rho(m'|m)$ in this case as these proposals will be rejected without the need to evaluate $\rho(m'|m)$.

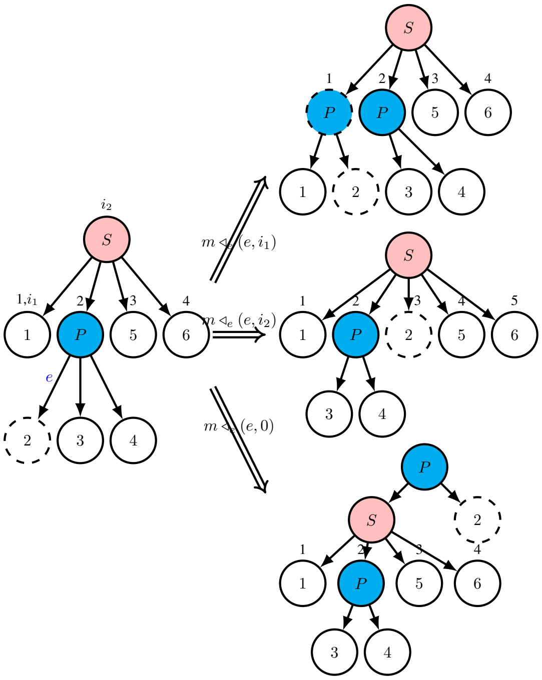
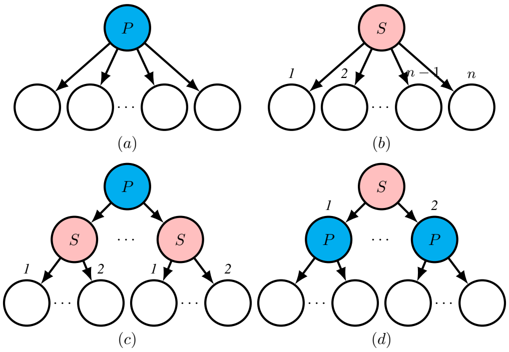


Figure C.2: Some possible operations on the MDT m_1 from Fig. 4. The edge e connected to leaf for actor 2 is reconnected to leaf node i_1 (where it must give a new P node as its neighbor, the parent of i_1, is S), to ancestral node i_2 (and is randomly allocated position 3 among the nodes stacked below the S-node i_2), and to node 0 (where it is added above the root as a P-node, as its neighbor the ex-root node is S).

Some operations are inadmissible, so we need to check our proposal defines an irreducible Markov chain on its own, or add other operations.

Proposition 8 (Posterior Marginals) *Consider the MDT Markov chain $M_k,\ k \geq 0$ with $M_0 \in \mathcal{M}_{[n]}$ formed by repeated random updates defined as follows: let $M_t = m$; let $e \sim \mathcal{U}(E_{-0}(m))$ and $i \sim \mathcal{U}[(\mathcal{F} \cup \mathcal{A}) \setminus \{e_1, e_2\}]$; Let $m' = m \lhd_e (e, i)$ be given by Defn. C.2; if $m' \in \mathcal{M}_{[n]}$ set $M_{k+1} = m'$ and otherwise $M_{k+1} = m$. This proposal-chain is irreducible.*

Proof C.1 (Proposition 8) *Consider the two building-block MDT's m_a, m_b shown in the top row of Fig. C.3. These have a single internal node with n leaves. Any MDT $m \in \mathcal{M}_{[n]}$ has a root node which must be of type P or S. We show that every MDT with a root of type P (or S) intercommunicates with m_a (respectively m_b) and that m_a intercommunicates with m_b and hence $\mathcal{M}_{[n]}$ is a closed communicating class.*



(a) (b) (c) (d)

Figure C.3: Four building-block MDT's.

We first show $m_a \to m_b$. We use the 0 node but there are many paths. Let r_a be the label of the root node in m_a and r_b in m_b. Suppose $L_{r_b}(m_b) = (i_1, ..., i_n)$ gives the stacking data for the children of the S node r_b. Label nodes of m_a so $f_c(r_a|m_a) = \{i_1, ..., i_n\}$ and $F_{i_k}(m_a) = F_{i_k}(m_b),\ k = 1, ..., n$. Now take $e = \langle r_a, i_1 \rangle$ in m_a and $i = 0$ and set $m = m_a \lhd_e (e, i)$. This creates a new node j of type S above the root. Let the stacking data of this new node be

$L_j(m) = (i_1, r)$. Now apply $m \leftarrow m \triangleleft_e (\langle r, i_k\rangle, j)$ for each $k = 2, ..., n-1$, adding i_k into position k in the list $L_j(m)$. When we do the last node $k = n-1$, node r is removed and j connects directly to i_n with i_n in the correct position in $L_j(m)$. This gives $m = m_b$. All these operations are admissible and have non-zero probability. The same scheme can be reversed, so we can take a MDT of type m_b and reorder the entries in $L_{r_b}(m_b)$ by going to m_a and back, placing the leaves in any desired order in $L_j(m)$ as we pass back.

Now take a general $m^* \in \mathcal{M}_{[n]}$. Its root r^* matches m_a or m_b by type. The root of m^* partitions the leaves into K sets $\{s_1, ..., s_K\}$ where K is the number of child nodes of r^* and $s_k = (s_{k,1}, ..., s_{k,c_k}),\ k = 1, ..., K$.

If the root type of m^* is S then these partitions are ordered. In this case we permute the leaves of m_b so that $L_{r_b}(m_b) = (s_{1,1}, ..., s_{K,c_K})$. Let $m = m_b$ with root r. If $i \in s_{k'}$ is a child of r^* which is a leaf then $s_{k'} = \{i\}$ and we are done. All the other partitions s_k correspond to child nodes i_k of r^* which are P nodes. We pull the edges $\langle r, i\rangle,\ i \in s_k$ of m down one at a time to create a P node with child nodes s_k matching the leaf-descendants of i_k in m^*. This gives a new m matching m^* down to all nodes of depth less than or equal to two. The passage from m_b to the new $m = m_d$ is illustrated bottom right in Fig. C.3.

If the root type of m^* is P then the partitions are $\{s_1, ..., s_K\}$ are unordered. The same process is repeated for $m = m_a$, pulling down the edges $\langle r, i\rangle,\ i \in s_k$ one at a time to build an S-node with leaves s_k matching the leaf-descendants of i_k and their order in m^*.

The process can now be repeated, as the problem of changing an MDT m so that it matches m^* to depth three when it already matches m^* to depth two is the problem of changing the MDT's in m rooted by $i_1, ..., i_K$ to match the corresponding subtrees of m^* to depth two. This task is the same as the original task and we have shown we can match to depth two. Since we can always increase the depth of the match and the depth is finite, we can change m_a or m_b to match m^*.

It is straightforward to check that these processes can be reversed and so the MDT proposal Markov chain formed by repeated edge operation defined in Defn. C.2 is irreducible.

Our MCMC algorithm for MDT with the QJ-B observation model is given in Algorithm C.2, omitting the standard q, p and ϕ updates. The algorithm for QJ-U model omits the ϕ-update.

Algorithm C.2 The MCMC algorithm for the MDT with QJ-B observation model at step k.

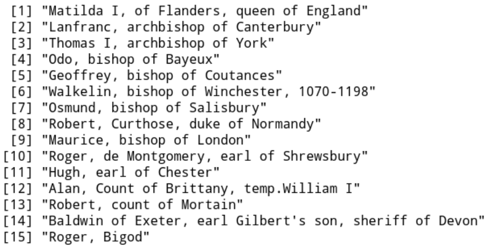
```
Require: $y, m^{(k-1)}{=}m, q^{(k-1)}{=}q, p^{(k-1)}{=}p, \phi^{(k-1)}{=}\phi$ with $m{=}(F(m), E(m), L(m)),\ m{\in}\mathcal{M}_{[n]}$
Ensure: $m^{(k)} \sim \pi(m|y, q, p, \phi),$
        $q^{(k)} \sim \pi(q|y, m^{(k)}, p, \phi),$
        $p^{(k)} \sim \pi(p|y, m^{(k)}, q^{(k)}, \phi),$
        $\phi^{(k)} \sim \pi(\phi|y, m^{(k)}, q^{(k)}, p^{(k)})$
    ———————— Update for $m$ ————————
  $m' \leftarrow m^{(k-1)} \leftarrow m$
  Sample $e \sim \mathcal{U}(E_{-0}(m))$ and $i \sim \mathcal{U}[(\mathcal{F} \cup \mathcal{A})\backslash\{e_1, e_2\}]$
  $m' \leftarrow m \triangleleft_e (e, i)$
  if $m' \in \mathcal{M}_{[n]}$ then
      $\eta_1 \leftarrow \frac{Q(y|v(m'), p, \phi)\pi_{\mathcal{M}_{[n]}}(m'|q)\rho(m|m')}{Q(y|v(m), p, \phi)\pi_{\mathcal{M}_{[n]}}(m|q)\rho(m'|m)}$
      if $\mathcal{U}(0,1) \leq \eta_1$ then
          $m \leftarrow m^{(k)} \leftarrow m'$
      end if
  end if
    ———————— Updates for $q, p$ and $\phi$ omitted ————————
```

The queue-jumping probability $p > 0$ (almost surely) so the Hastings ratio $\eta > 0$ in Algorithm C.2 is not zero for all $m, m' \in \mathcal{M}_{[n]}$ connected by an update. Since the proposal chain $M_k,\ k \geq 0$ in Proposition 8 is irreducible, it follows that our MDT-MCMC is irreducible.

D DATA BACKGROUND AND ADDITIONAL RESULTS

D.1 THE ‘ROYAL ACTA’ DATA

The “Royal Acta” data is a database made for “The Charters of William II and Henry I” project by the late Professor Richard Sharpe and Dr Nicholas Karn [Sharpe et al., 2014]. It collects dated witness lists from legal documents in England and Wales in the eleventh and twelfth century. Each witness list is dated though the dating is sometimes uncertain (a few years is typical). Lower and upper bounds on the date of a list are part of the data. Each individual is associated with a profession (title) such as Queen, Archbishop, etc. We assign witnesses with no title as “other”. Fig. D.1 gives an example of such witness list. The data records different number of lists with various lengths over time - summarised in Figure D.2.



```
 [1] "Matilda I, of Flanders, queen of England"
 [2] "Lanfranc, archbishop of Canterbury"
 [3] "Thomas I, archbishop of York"
 [4] "Odo, bishop of Bayeux"
 [5] "Geoffrey, bishop of Coutances"
 [6] "Walkelin, bishop of Winchester, 1070-1198"
 [7] "Osmund, bishop of Salisbury"
 [8] "Robert, Curthose, duke of Normandy"
 [9] "Maurice, bishop of London"
[10] "Roger, de Montgomery, earl of Shrewsbury"
[11] "Hugh, earl of Chester"
[12] "Alan, Count of Brittany, temp.William I"
[13] "Robert, count of Mortain"
[14] "Baldwin of Exeter, earl Gilbert's son, sheriff of Devon"
[15] "Roger, Bigod"
```

Figure D.1: An example witness list from 1080, extracted from the “Royal Acta” data. The witnesses names are entered by a clerk in order from top to bottom.

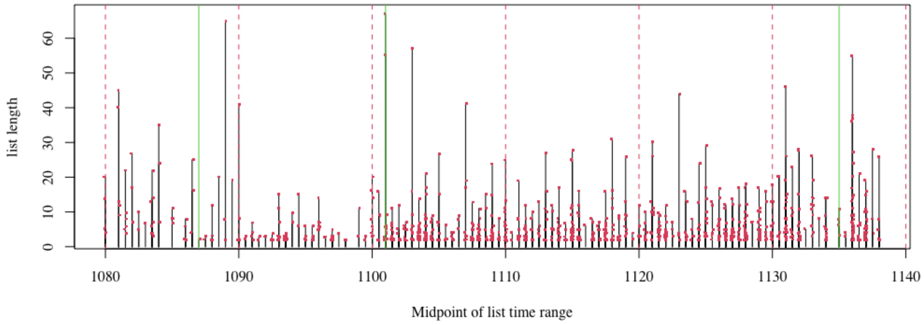


Figure D.2: The midpoint of list time range v.s. list range. Each red dot is a list of length y created in a time range midpointed by x. The bars represents the length of the longest list at time x.

In Section 5, we limit the number of lists per actor (LPA) participate in to be at least 5 for ease of presentation. However, it is possible to fit our model on much larger datasets. We chose time periods with a large number of lists with relatively long lengths - 1080-1084 and 1136-1138, and extract the lists with 1LPA. Table D.1 summarises the data in the different experiments. In Section D.1.1, we carry out Bayesian inference on the 1LPA datasets. In Sections D.1.2 and D.1.3, we present MCMC traceplots and effective sample sizes for MCMC samples of key parameters in the analysis on 5LPA data, from the VSP/QJ-U and VSP/QJ-B models respectively.

	5LPA				**1LPA**	
	80-84	26-30	34-38	34-38(b)	80-84	34-38
n	17	13	49	14	181	216
N	20	30	82	37	27	95
$\max(y)$	17	8	35	14	45	55

Table D.1: Data content for time periods of interest including the number of actors (n), number of lists (N) and the length of their longest list ($\max(y)$). Data analysed with both VSP/QJ-U and VSP/QJ-B are marked in blue. The 1134-1138 bishop-only data is 34-38(b).

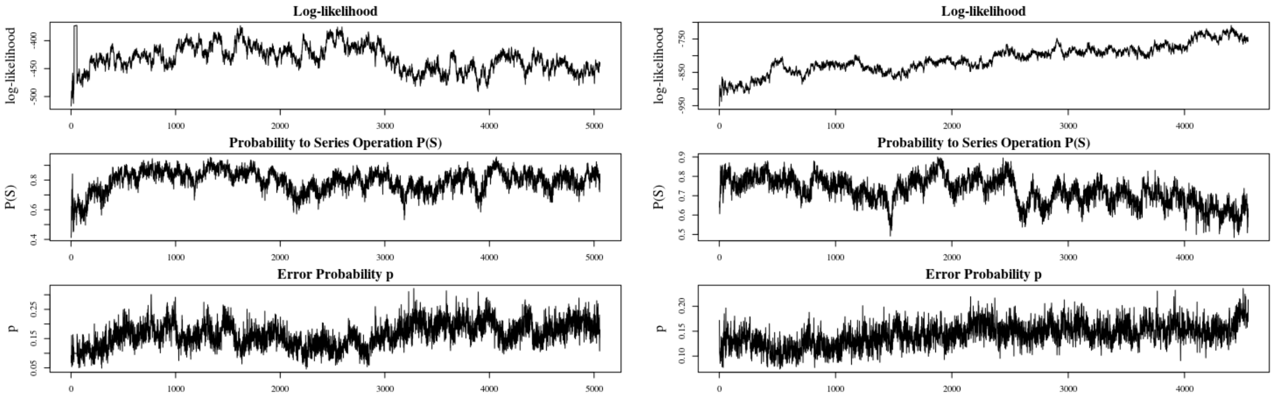
D.1.1 Inference Results on List Data with 1LPA (QJ-U Observation Model)

Using the full-data lists (allowing $LPA = 1$), we arrive at much larger datasets with 181 actors (1080-1084) and 216 actors (1134-1138) respectively, as is summarised in table D.1. Though QJ-B observation model has higher flexibility, it is rather computationally demanding when we move to large datasets. In this section, we fit the VSP/QJ-U model on both data lists instead.

We perform 50,000 MCMC iterations on 1080-1084 (1LPA) data and 48,000 iterations on 1134-1138 (1LPA) data. For details of the MCMC algorithm, see Algorithm C.1. Every 10 steps is recorded from the MCMC. The effective sample sizes and traceplots for the key parameters p and $P(S) = q$ from the MCMC samples are shown in Table D.2 and Figure D.3. The MCMC on the 1080-1084 (1LPA) data displays fair mixing, however, the MCMC for 1134-1138 (1LPA) is yet to be fully mixed. We are aware the effective sample sizes are relatively small, here we only present the current results as a demonstration.

	ESS	
Parameter	1080-1084	1134-1138
$P(S)$	41	25
p	32	47

Table D.2: The effective sample sizes for $P(S)$ and error probability p on four datasets with 1LPA.



(a) 1080-1084 with 1 LPA

(b) 1134-1138 with 1 LPA

Figure D.3: Traceplots for log-likelihood, $P(S)$ and error probability p for the two data sets of interest here - 1080-1084 (a) and 1134-1138 (b) with 1 LPA data.

We present the consensus orders $V^{con}(\epsilon)$ in Figure D.4 for 1080-1084 (1LPA) and Figure D.5 for 1134-1138 (1LPA). We choose a threshold of $\epsilon = 0.6$ in order to represent readable consensus orders graphically. Considering the large number of actors in both time periods, we also extract the non-'other' actors and reconstruct the consensus orders in Figure D.6 for 1080-1084 (1LPA) and Figure D.7 for 1134-1138 (1LPA).

A clear order relation for king $\succ$ queen $\succ$ archbishop $\succ$ bishop is observed in both time periods. The actors roughly appear in the "group" of their professions.

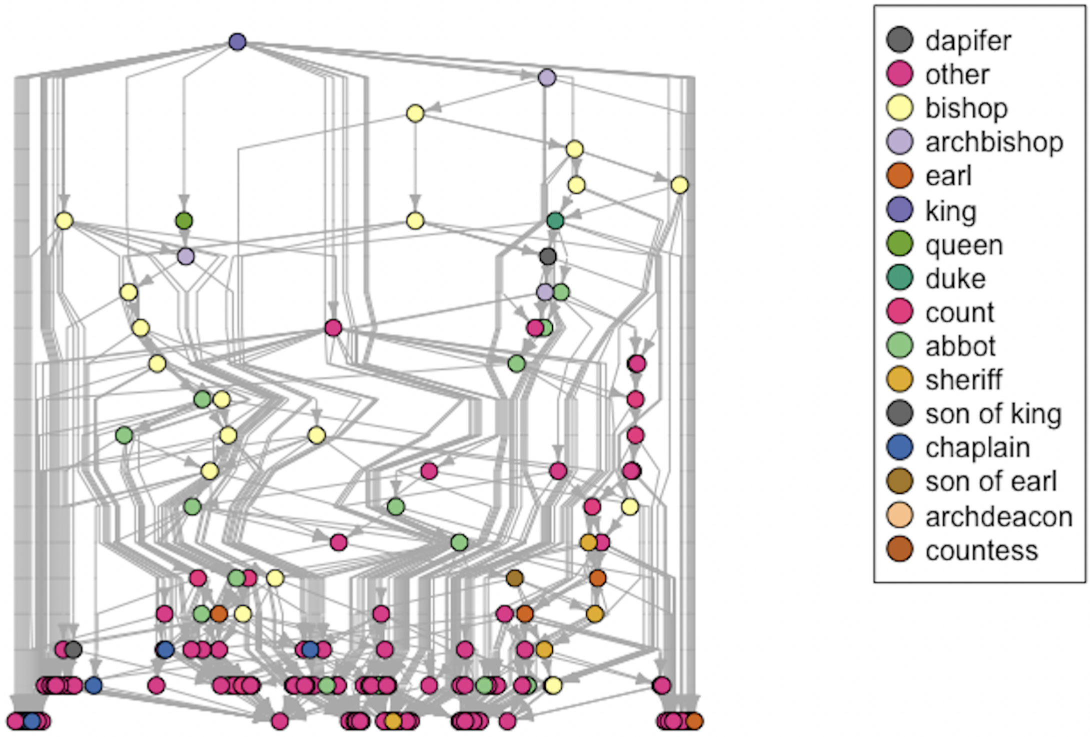


Figure D.4: The consensus order for 1080-1084 (1LPA) data in a VSP/QJ-U analysis.

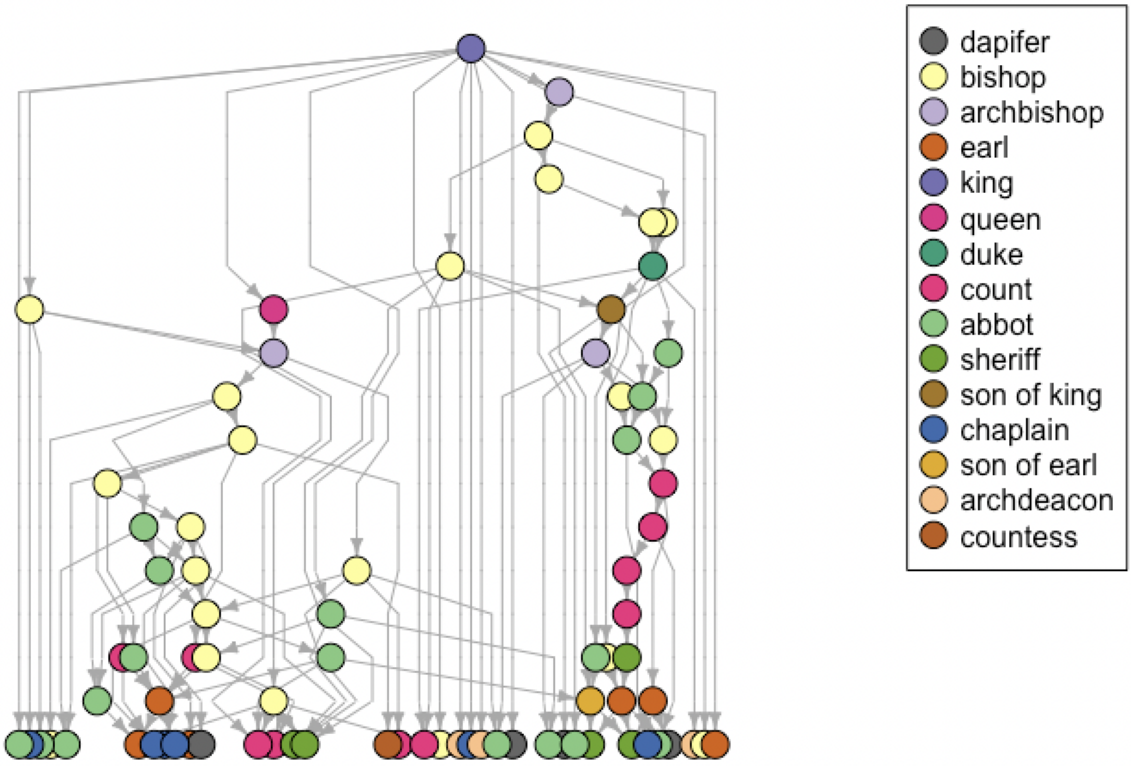


Figure D.6: The consensus order for 1080-1084 (1LPA) data without ‘other’ actors in a VSP/QJ-U analysis.

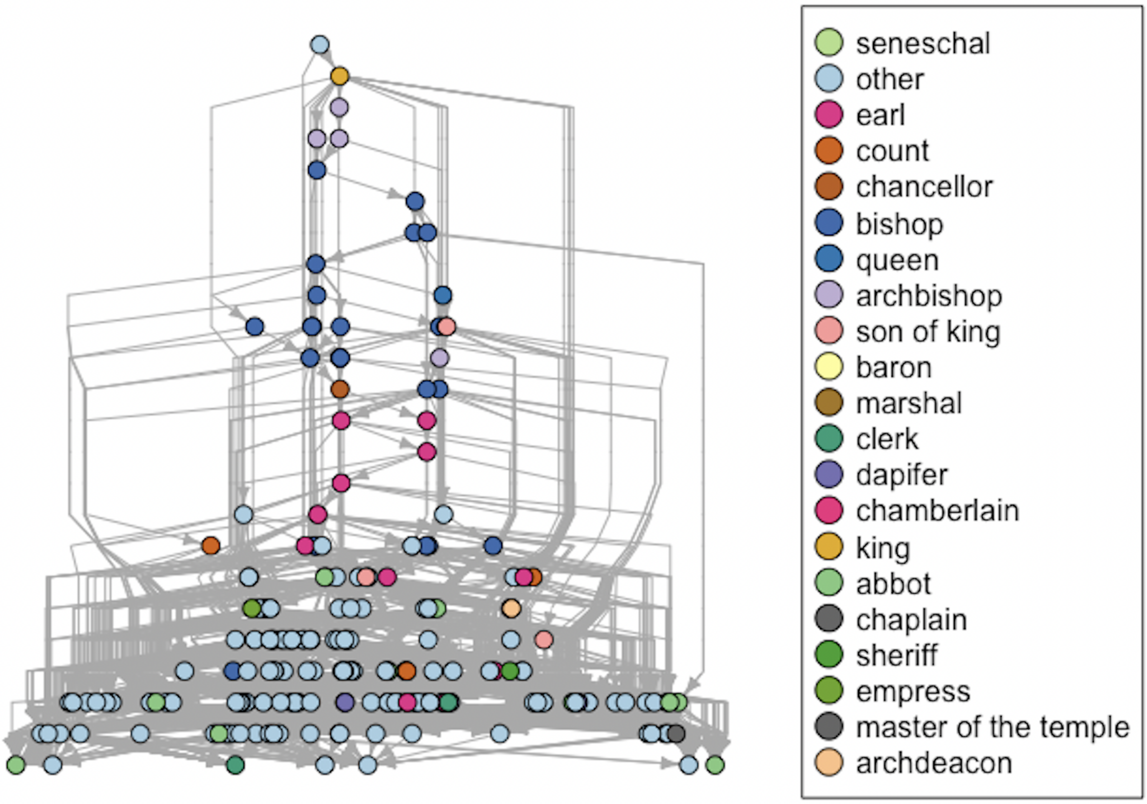


Figure D.5: The consensus order for 1134-1138 (1LPA) data in a VSP/QJ-U analysis.

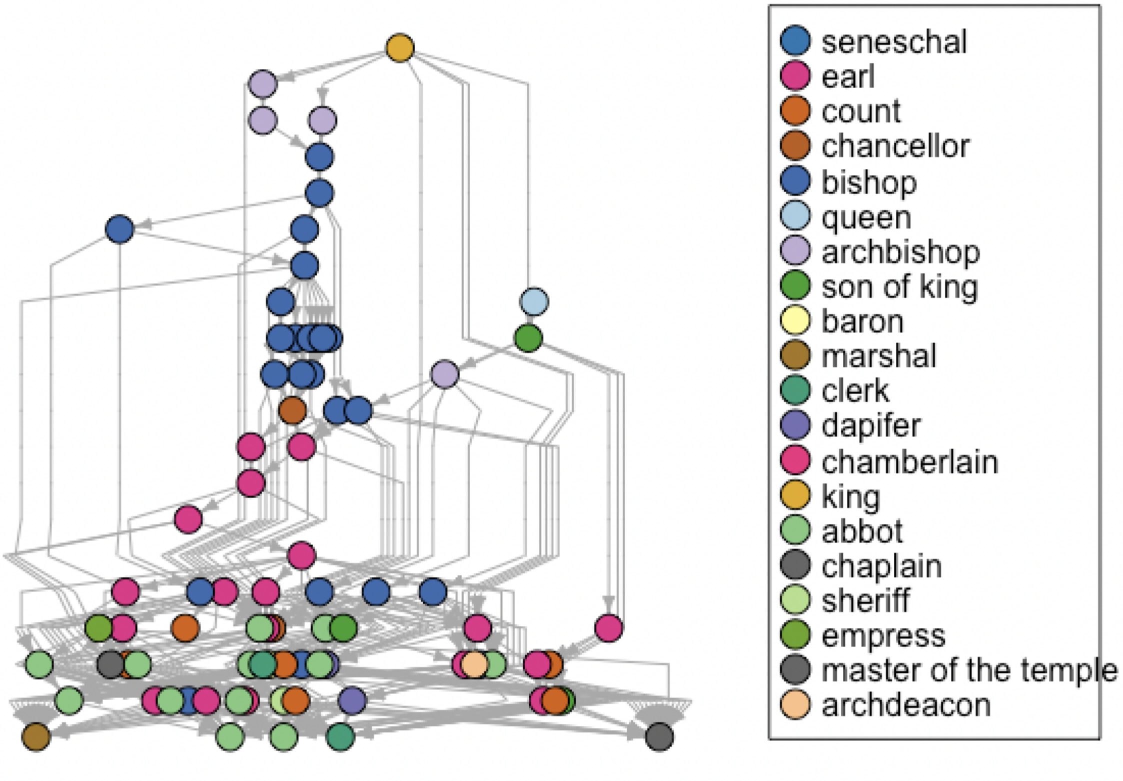


Figure D.7: The consensus order for 1134-1138 (1LPA) data without ‘other’ actors in a VSP/QJ-U analysis.

Table D.3 presents the average rankings of different professions for 1080-1084 (1LPA) and 1134-1138 (1LPA). The average rankings support our observations above. Interestingly, abbots tend to be ranked higher during 1080-1084 than 1134-1138, and the archdeacon is ranked higher in 1134-1138 than 1080-1084.

	Average Rank	
Profession	1080-1084	1134-1138
King	1.21 (0.007)	3.73 (0.02)
Queen	4.81 (0.03)	4.97 (0.02)
Archbishop	9.70 (0.05)	8.89 (0.04)
Empress	NA	16.0 (0.07)
Duke	15.4 (0.08)	NA
Bishop	18.7 (0.10)	20.8 (0.10)
Son of King	18.8 (0.10)	24.0 (0.11)
Seneschal	NA	28.0 (0.13)
Abbot	32.8 (0.18)	88.0 (0.41)
Countess	39.0 (0.22)	NA
Count	43.1 (0.24)	33 (0.15)
Son of Earl	43.5 (0.24)	NA
Earl	44.3 (0.24)	44.3 (0.20)
Dapifer	44.5 (0.25)	81.3 (0.38)
Archdeacon	48.7 (0.27)	35.3 (0.16)
Chancellor	NA	43.6 (0.20)
Other	50.1 (0.28)	79.2 (0.37)
Chaplain	50.3 (0.28)	44.7 (0.21)
Baron	NA	78.4 (0.36)
Sheriff	60.5 (0.33)	95.7 (0.44)
Chamberlain	NA	101 (0.47)
Clerk	NA	114 (0.53)
Master of the temple	NA	137 (0.63)
Marshal	NA	150 (0.70)

Table D.3: The professions and their average rankings for 1080-1084 (1LPA) and 1134-1138 (1LPA). NA means the profession of interest does not appear in this time period.

Posterior distributions for the key parameters in Figure D.8 show that witness lists in 1080-1084 tend to respect a stronger social hierarchy than in 1134-1138 with larger $P(S)$. The error probabilities p are relatively smaller for witness lists in 1134-1138. This agrees with the results for 5LPA presented in Fig. 5, Section 5.2. The prior and posterior VSP depth distributions are shown in Fig. D.9. Despite the roughly uniform prior distribution over the VSP depth, the posterior depths appear to concentrate around 75 for 1080-1084 and 90 for 1134-1138.

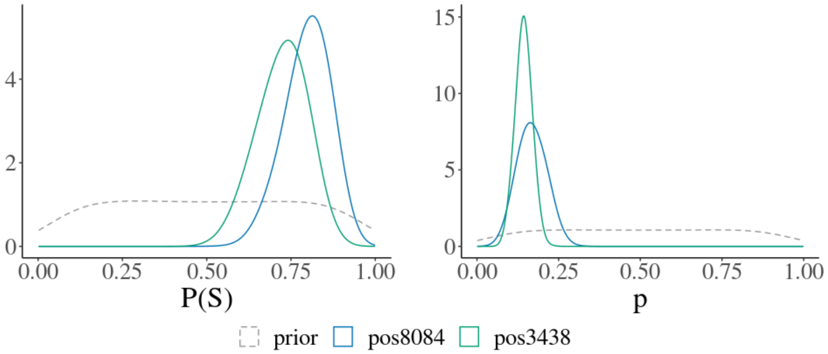


Figure D.8: Prior (grey line) and posterior distributions for $q = P(S)$ (left) and error probability p (right) for the time periods 1080-1084 (1LPA) (blue) and 1134-1138 (1LPA) (green) in a VSP/QJ-U analysis.

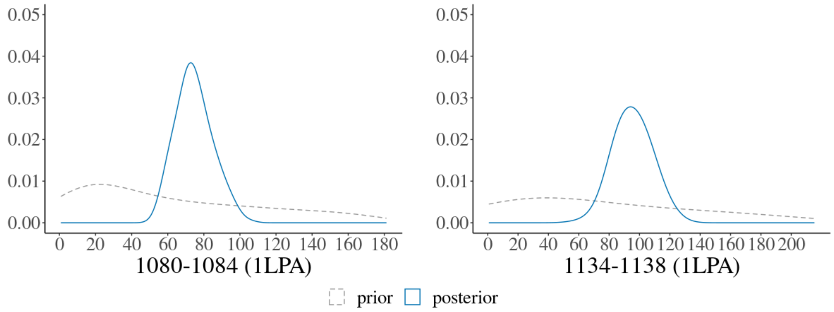


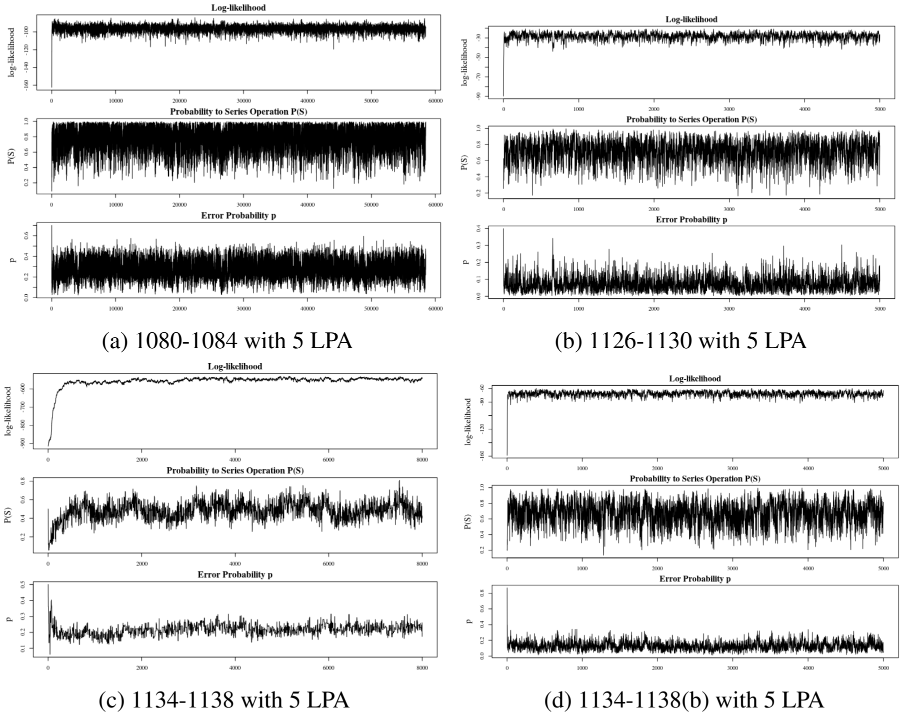
Figure D.9: The prior (grey) and posterior (blue) VSP depth distribution for 1080-1084 (1LPA) (left) and 1134-1138 (1LPA) (right) in a VSP/QJ-U analysis.

D.1.2 Inference Results on List Data with 5LPA (QJ-U Observation Model)

Fig. 6 and Fig. 7 (top-row) show the consensus orders V^{con} for 1134-1138 (5LPA), 1080-1084 (5LPA), 1126-1130 (5LPA) and 1134-1138 (bishop) (5LPA) under the VSP/QJ-U model. The MCMC converge well. Here we estimate and report effective sample sizes (ESS, Table D.4) and inspect MCMC traces (Fig. D.10). Both the high ESSs and the traceplots indicate good convergence to the posterior distribution.

	ESS			
Parameter	1080-1084	1126-1130	1134-1138	1134-1138(b)
$P(S)$	1676	1477	95	648
p	1297	1426	262	586

Table D.4: The effective sample sizes for $P(S)$ and error probability p on the four datasets with 5LPA and QJ-U.



(a) 1080-1084 with 5 LPA

(b) 1126-1130 with 5 LPA

(c) 1134-1138 with 5 LPA

(d) 1134-1138(b) with 5 LPA

Figure D.10: Traceplots for log-likelihood, $P(S)$ and error probability p for the four list data of interest - 1080-1084 (a) and 1126-1130 (b), 1134-1138 (c) and 1134-1138 (bishops) (d) with 5 LPA data and a VSP/QJ-U analysis.

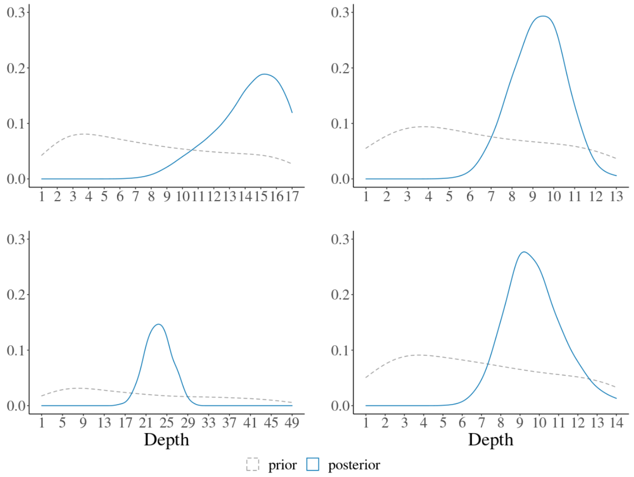


Figure D.11: The prior (grey) and posterior (blue) VSP depth distribution for 1180-1184 (top-left), 1126-1130 (top-right), 1134-1138 (bottom-left) and 1134-1138(b) (bottom-right) with 5LPA and QJ-U.

The posterior distributions for both p and $q = P(S)$ are shown in Fig. 8. We also present the posterior depth-distributions for the datasets in Figure D.11. It appears that 1080-1084 (5LPA) admits the most rigid social hierarchy, while 1134-1138 (5LPA) has less hierarchy with respect to n. The average rankings per profession are reported in Table D.5. Similar to the consensus orders (Fig. 6 and Fig. 7), king $\succ$ queen $\succ$ archbishop $\succ$ bishop. The three time periods show similar hierarchical structure, although the power gap between count and earl is relatively narrower in 1126-1130.

	Average Rank		
Profession	1080-1084	1126-1130	1134-1138
King	1.02 (0.06)	NA	1.01 (0.02)
Queen	2.15 (0.13)	NA	2.01 (0.04)
Duke	2.79 (0.16)	NA	NA
Son of King	4.63 (0.27)	NA	3.11 (0.06)
Archbishop	4.45 (0.26)	1 (0.08)	4.55 (0.09)
Bishop	8.25 (0.49)	4.02 (0.31)	11.10 (0.23)
Chancellor	NA	NA	21.40 (0.44)
Count	10.90 (0.64)	5.92 (0.45)	24.00 (0.49)
Earl	12.20 (0.72)	5.98 (0.46)	28.10 (0.57)
Other	15.30 (0.90)	8.80 (0.68)	33.10 (0.68)

Table D.5: The professions and their average rankings for all three time periods with 5LPA and QJ-U. NA means the profession of interest does not appear in this time period.

As discussed, we perform reconstruction accuracy tests on each dataset to assess the reliability of our estimations. This is done by taking representative parameters (the last sample state of the parameters sampled from the corresponding posterior), and generating synthetic data with the same list-memberships and lengths as the real data. We carry out or standard analysis on these synthetic datasets, fitting the same model used to simulate the data, and construct the corresponding consensus orders $V^{con}(\epsilon)$ with $\epsilon \in [0, 1]$. The results are summarised using receiver operator characteristic (ROC) curves. The ROC curve shows the relation between the proportion of inferred false-positive order relations (x-axis) and true-positive relations (y-axis) for different ϵ. The existence of a ϵ that gives high true-positive and low false-positive reconstructed fraction means reconstruction accuracy is high.

Fig. D.12 shows ROC curves for such a reconstruction test on the 1080-1084 (5LPA), 1126-1130 (5LPA) and 1134-1138 (5LPA) data in a VSP/QJ-U model. The proportion of inferred false-positive (x-axis) and true-positive (y-axis) relations increases with decreasing ϵ from (0, 0) at $\epsilon = 1$ (the consensus order is empty) to (1, 1) at $\epsilon = 0$ (complete graph). For all time periods, we observe ϵ that gives high true-positive and low false-positive reconstructed fraction, indicating our model's high reliability to reconstruct relations.

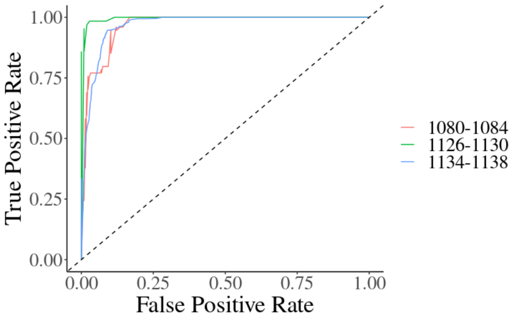


Figure D.12: Receiver operating characteristic (ROC) curves for synthetic data using 1080-1084, 26-30 and 34-38 list membership structures with 5LPA and QJ-U.

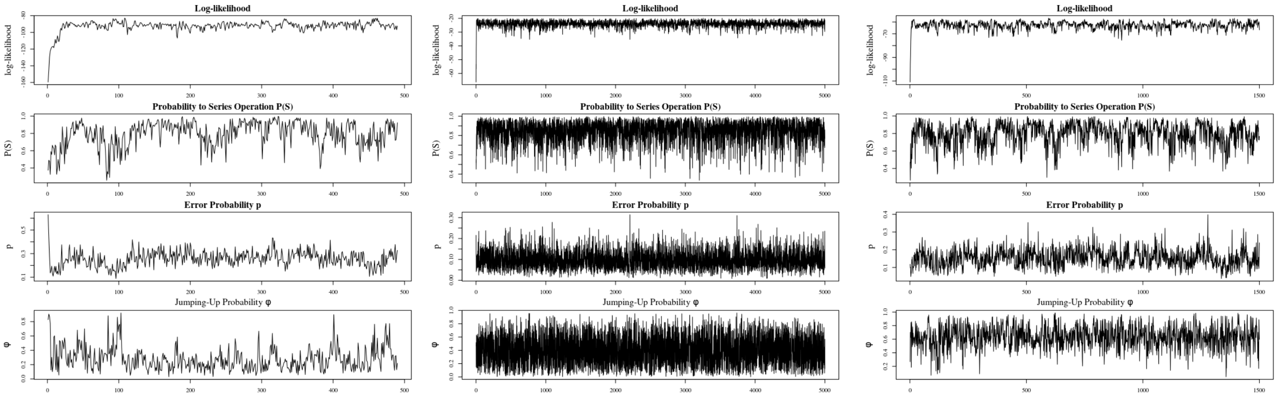
D.1.3 Inference Results on List Data with 5LPA (QJ-B Observation Model)

In this section, we fit the VSP/QJ-B data on the datasets 1080-1084 (5LPA), 1126-1130 (5LPA) and 1134-1138 (bishop) (5LPA). See algorithm C.1 for the MCMC details. Traceplots for the log-likelihood, $P(S)$, error probability p and bi-directional top/bottom insertion probability ϕ are all presented in Figure D.13. They all display reasonable convergence. In table D.6 we estimate effective sample sizes (ESS) for key parameters. Mixing for the key parameters are fair during time period 1080-1084 and 1134-1138 (bishop), and the agreement (to some extent) to the analyses in Section D.1.2 supports our conclusion that the samples are representative.

	ESS		
Parameter	1080-1084	1126-1130	1134-1138(b)
$P(S)$	47	1875	121
p	61	3401	197
ϕ	69	3428	728

Table D.6: The effective sample sizes for $P(S)$ and error probability p on the three datasets with 5LPA fitting VSP/QJ-B.

Consensus orders $V^{con}(\epsilon)$ with $\epsilon = 0.5$ are shown in Fig. 7 (bottom-row). We report the average rankings per profession for 1080-1084 (5LPA) and 1126-1130 (5LPA) in Table D.7. The posterior distributions for the key parameters p, $q = P(S)$ and ϕ are shown in Fig. 8. Here we display the posterior depth distribution for the three time periods in Fig. D.14. All periods favour higher VSP depths. By comparing the consensus orders, the bi-directional queue-jumping model seems to fit a more rigid social hierarchy than the queue-jumping-up model, especially during periods 1126-1130 and 1134-1138. This is also illustrated by higher posterior means on $q = P(S)$ for both the 1126-1130 (5LPA) and 1134-1138 (bishop) (5LPA) data. It is surprising that earl $\succ$ count in 1126-1130 under the QJ-B model, although the opposite is observed under QJ-U. Both QJ-U and QJ-B models conclude similar posterior distribution on p, the error probability in the data-lists. By inspecting the



(a) 1080-1084 with 5 LPA and QJ-B

(b) 1126-1130 with 5 LPA

(c) 1134-1138(b) with 5 LPA

Figure D.13: Traceplots for the log-likelihood, $P(S)$ and error probability p for the three list data sets of interest - 1080-1084 (a) and 1126-1130 (b) and 1134-1138 (bishops) (c) with 5LPA data and a VSP/QJ-B analysis.

posterior distributions on ϕ, it appears that QJ-D is slightly preferred for 1080-1084 (5LPA) while QJ-U/QJ-B is preferred for 1134-1138 (bishop) (5LPA). This is justified by the Bayes Factors in section 5.

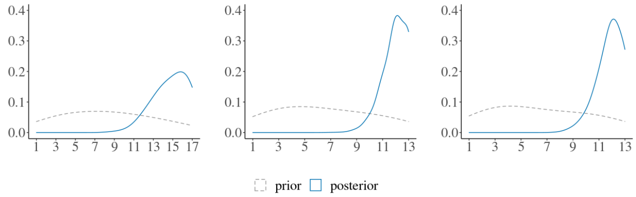


Figure D.14: The prior (grey) and posterior (blue) VSP depth distribution for 1180-1184 (left), 1126-1130 (middle) and 1134-1138(b) (right) with 5LPA data in a VSP/QJ-B analysis.

	Average Rank	
Profession	1080-1084	1126-1130
King	1.03 (0.06)	NA
Queen	1.95 (0.11)	NA
Duke	4.29 (0.25)	NA
Son of King	6.18 (0.36)	NA
Archbishop	3.88 (0.23)	1 (0.08)
Bishop	8.38 (0.49)	3.99 (0.31)
Earl	12.40 (0.73)	6.93 (0.53)
Count	13.00(0.77)	8.94 (0.69)
Other	15.90 (0.94)	10.40 (0.80)

Table D.7: The professions and their average rankings for all three time periods with 5LPA data and QJ-B. NA means the profession of interest does not appear in this time period.

Figure D.15 displays ROC curves from a reconstruction accuracy test using VSP/QJ-B to simulate and fit synthetic data matching the 1126-1130 and 1134-1138 5LPA data, as described in Section 5. Again, we see the proportion of inferred false-positive and true-positive relations increasing while decreasing ϵ from $(0, 0)$ at $\epsilon = 1$ to $(1, 1)$ at $\epsilon = 0$. The ϵ's that give high true-positive and low false-positive reconstruction fraction can be easily identified in Fig. D.15. This indicates our model's high accuracy in reconstruction order relations.

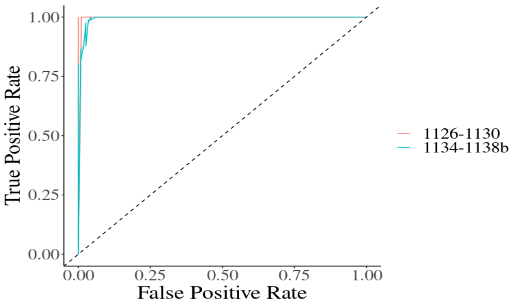


Figure D.15: Receiver operating characteristic (ROC) curves for synthetic data using 1126-1130 and 1134-1138 (bishop) list membership structures with 5LPA and QJ-B.

D.2 THE FORMULA 1 RACE DATA

The Formula 1 race data (2017 - 2022) F1D records information about every formula 1 race in the past five seasons. The data gives the top 20 drivers in each Grand Prix race in each season. One typical list, for the British Grand Prix (Silverstone Circuit) in 2021, is as follows

> 1 – HAM, 2 – LEC, 3 – BOT, 4 – NOR, 5 – RIC, 6 – SAI, 7 – ALO, 8 – STR, 9 – OCO, 10 – TSU, 11 – GAS, 12 – RUS, 13 – GIO, 14 – LAT, 15 – RAI, 16 – PER, 17 – MAZ, 18 – MSC, R – VET, R – VER.

Each abbreviation is a unique code for a driver (see table D.8), e.g. ‘HAM’ stands for Lewis Hamilton, who was the winner of this race. The drivers are ordered based on their finishing position. The label ‘R’ indicates special circumstances, e.g. collision, accident, retirement, etc.

We are interested in the order relations between these drivers and construct a VSP map of their performance in a specific season. This is an intersting test of the method as a heuristic model (in the sense that Plackett-Luce and Mallows are in general heuristic). There is no constraint other than car speed and skill to stop one driver overcoming another so it is not clear that the order relations we recover correspond to any element of reality. One feature that is characteristic of a PO-style analysis (such as ours with VSPs) is that the race resembles a queue in which drivers exchange places subject to skill and car-speed. In a race, a driver can fall down the order with a certain probability due to unexpected circumstances (poor tyre management, problems in the pits, small collisions, time penalties etc). However, there is no obvious mechanism promoting a driver up the race order. We therefore believe the QJ-D observation model is natural.

In this analysis, we take a snapshot of 2021, assuming relative car-quality and skill are roughly constant over a year. The Formula 1 (F1) 2021 data consists of 22 lists corresponding to the 22 Grand Prix races. Each list is has at most 20 elements. We disregard the ‘R’ positions, so the lists are of unequal length. There are a total of 21 drivers participating in season 2021. We assign each of them a unique Driver ID, listed in table D.8.

We analyse the data-lists from season 2021 between the 21 actors using the VSP/QJ-D model. The consensus order for the drivers in this season is shown in Fig. D.16. Both Lewis Hamilton and Max Verstappen are ranked at top of the consensus VSP for the 2021 season, with high posterior probability (more than 0.9).

The posterior distributions for individual parameters and the depth are shown in Fig. D.17. The effective sample sizes are 567 for $q = P(S)$ and 130 for p. The posterior for $P(S)$ concentrates at around 0.5, showing a relatively relaxed ranking relation. The posterior distribution for p concentrates at a lower value at 0.15. This suggests the VSP model relatively accurately represents the strength of each driver-car pairing. The VSP depths are relatively low for this data. We are not observing a ranking as deep as the social hierarchy for witnesses in “Royal Acta”.

Driver ID	Code	Name	DOB	Nationality
1	HAM	Lewis Hamilton	07/01/85	British
2	ALO	Fernando Alonso	29/07/81	Spanish
3	RAI	Kimi Raikonnen	17/10/79	Finnish
4	KUB	Robert Kubica	07/12/84	Polish
5	VET	Sebastian Vettel	03/07/87	German
6	GAS	Pierre Gasly	07/02/96	French
7	PER	Sergio Perez	26/01/90	Mexican
8	RIC	Daniel Ricciardo	01/07/89	Australian
9	BOT	Valtteri Bottas	28/08/89	Finnish
10	VER	Max Verstappen	30/09/97	Dutch
11	SAI	Carlos Sainz	01/09/94	Spanish
12	OCO	Esteban Ocon	17/9/96	French
13	STR	Lance Stroll	29/10/98	Canadian
14	GIO	Antonio Giovinazzi	14/12/93	Italian
15	LEC	Charles Leclerc	16/10/97	Monegasque
16	NOR	Lando Norris	13/11/99	British
17	RUS	George Russell	15/02/98	British
18	LAT	Nicholas Latifi	29/06/95	Canadian
19	TSU	Yuki Tsunoda	11/05/00	Japanese
20	MAZ	Nikita Mazepin	02/03/99	Russian
21	MSC	Mick Schumacher	22/03/99	German

Table D.8: The list of drivers in Formula 1 season 2021. Each driver is assigned a unique 'Code' and 'Driver ID' in our analysis. We also include further information of the drivers, including their date of birth ('DOB') and 'Nationality'.

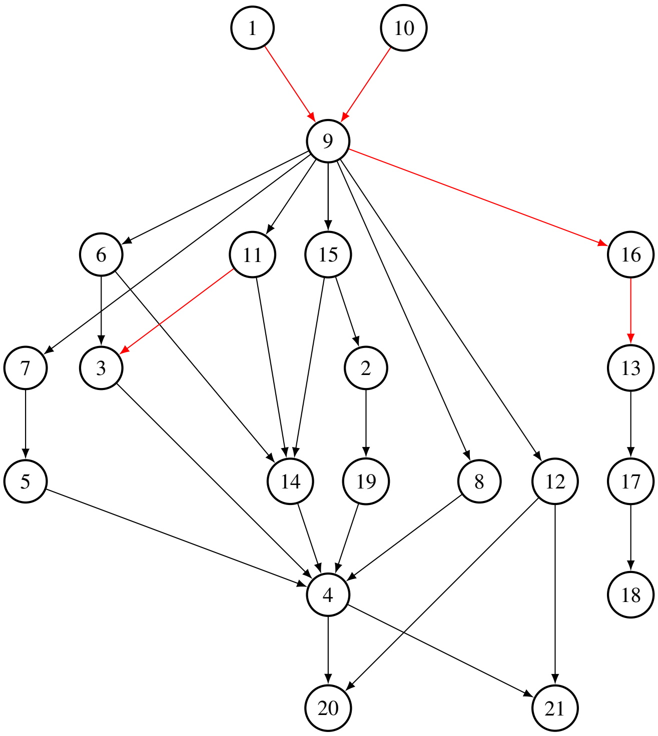


Figure D.16: VSP/QJ-D model. Consensus order for Formula 1 (season 2021) data. Significant/strong order relations are indicated by black/red edges respectively.

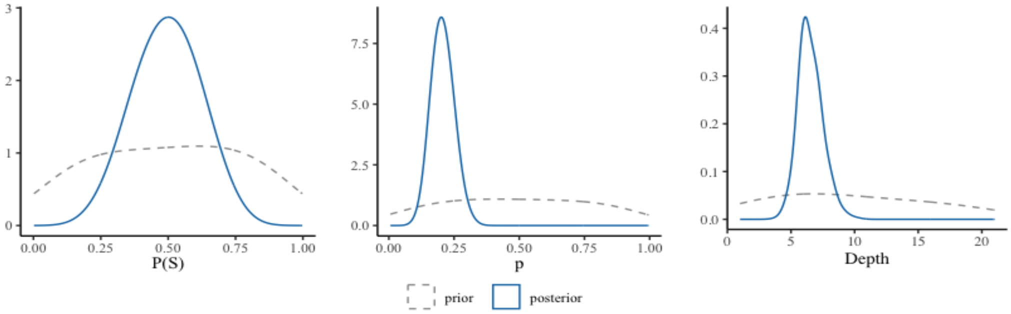


Figure D.17: The prior (grey) and posterior (blue) distributions for $P(S)$ (left), p (middle) and depth (right) for the Formula 1 (season 2021) data.

E MODEL COMPARISON

E.1 MODEL COMPARISON WITH PLACKETT-LUCE AND MALLOWS

The Plackett-Luce model, the Mallows model, and their mixture-models are two categories of model widely used for ranking and partial ranking. In this section, we compare the VSP/QJ-U and VSP/QJ-B models with the two PL-models[1] and the two Mallows models[2] using the WAIC. This estimates the expected log pointwise predictive density (ELPD, Vehtari et al. [2017]). It is a principled criterion for model comparison which is relatively easily estimated.

The Plackett-Luce model defines a distribution over ranked lists $y_i \in \mathcal{P}_{[n]},\ i \in [N]$ with actor attributes $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$. Taking into account the list membership sets $o_i,\ i \in [N]$, the likelihood is

$$PL(y|\lambda) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{e^{\lambda_{y_{i,o_j}}}}{\sum_{k=j}^{n_i} e^{\lambda_{y_{i,o_k}}}}. \tag{E.1}$$

The Plackett-Luce mixture assumes the lists are sampled from a heterogeneous population composed of D sub-populations. Each mixture component has a Plackett-Luce distribution over lists with actor attributes $\lambda^{(d)} \in \mathbb{R}^n,\ d \in [D]$. A finite mixture of Plackett-Luce models was proposed as a robust model for ranked data with incomplete lists in Mollica and Tardella [2017, 2020]. Let $\Lambda = (\lambda^{(d)})_{d \in [D]} \in \mathbb{R}^{n \times D}$ give the matrix of actor attributes and $\omega = (\omega_1, \dots, \omega_D)$ give the weights of mixture components with $\sum_{d=1}^{D} \omega_d = 1$. The D-component mixture Plackett-Luce model likelihood is

$$PL_{mix}(y|\Lambda, \omega) = \prod_{i=1}^{N} \sum_{d=1}^{D} \omega_d PL(y_i|\lambda^{(d)}). \tag{E.2}$$

Non-informative priors suggested by Mollica and Tardella [2020] are assigned with $e^{\lambda_j^{(d)}} \sim \text{Gamma}(1, 0.001)$ for $j \in [n]$ and $d \in [D]$ and $\omega_1, \dots, \omega_D \sim Dir(1, \dots, 1)$.

The Mallows model Mallows [1957] is typically controlled by a *location parameter (consensus ranking)* $\rho \in \mathcal{P}_n$ and a *scaling parameter* $\alpha \in (0, \infty)$. Letting $d(\cdot, \cdot) : \mathcal{P}_n \times \mathcal{P}_n \to \mathbb{R}_+$ be a *discrepancy function* between two permutations, the Mallows model is

$$P_d(y|\rho, \alpha) = \prod_{i=1}^{N} \frac{1}{Z_n(\alpha)} e^{-\frac{\alpha}{n} d(\rho, y_i)}, \tag{E.3}$$

where $Z_n(\alpha) := \sum_{y \in \mathcal{P}_n} e^{-\frac{\alpha}{n} d(\rho, y)}$ is the normalising constant. A typical distance choice is the Kendall's tau distance. Let $\sigma(l, a) = \{k \in [n] : l_k = a\}$. The Kendall's tau distance counts the number of pairwise disagreements between two permutations, $d(y, l) = \sum_{i<j} \mathbb{1}_{\sigma(l, y_i) > \sigma(l, y_j)}$, and this gives a tractable normalising constant $Z_n(\alpha)$. We use the Mallows ϕ model in our model comparison. A truncated exponential prior is specified for α and a uniform prior $\pi(\rho)$ on $\mathcal{P}_n$ is

[1] We use the MCMC sampler available in the R-package `PLmix` Mollica and Tardella [2017]. This uses a data augmentation scheme due to Caron and Doucet [2012].

[2] We use the MCMC sampler available in the R-package `BayesMallows` Sørensen et al. [2020].

taken for ρ, as is suggested in Sørensen et al. [2020] which implements the MCMC proposed in Vitelli et al. [2018]. The `BayesMallows` R-package deals with partial ranking by applying data augmentation techniques before fitting the full Mallows model.

Similar to the Plackett-Luce Mixture, the finite Mallows mixture allows for heterogeneity. Let $\{\rho_d, \alpha_d\}_{d=1,\ldots,D}$ be the set of parameters for cluster d and let $z_1, \ldots, z_N \in \{1, \ldots, D\}$ be the cluster labels that assign each list to one cluster. The D-component mixture Mallows likelihood is

$$P(y|\{\rho_d, \alpha_d\}_{d=1,\ldots,D}, \{z_i\}_{i=1,\ldots,N}) = \prod_{i=1}^{N} \frac{1}{Z_n(\alpha_{z_i})} e^{-\frac{\alpha_{z_i}}{n} d(y_i, \rho_{z_i})}. \tag{E.4}$$

Independent truncated exponential priors and independent uniform priors are specified for α and ρ respectively. Following Sørensen et al. [2020], $z_1, \ldots, z_N$ follow a uniform multinomial distribution and are assumed conditionally independent given the cluster parameters.

The ELPD measures the posterior predictive accuracy of a model. It is a natural choice for goodness-of-fit and model comparison. We use the WAIC to estimate the ELPD for a generic model ("A" say). The estimator resembles the AIC and BIC,

$$\widehat{elpd}_{waic}(A|y) = \sum_{i=1}^{N} \log p_A(y_i|y) - p_{waic}, \tag{E.5}$$

where

$$p_A(y_i|y) = \int p_A(y_i|\theta) p_A(\theta|y) d\theta \tag{E.6}$$

with θ representing all parameter in model A. The predictive probability in Eqn. E.6 is estimated using MCMC samples. For a MCMC sample (after burn-in) of length K targeting $p_A(\theta|y)$,

$$\widehat{p_A}(y_i|y) = \frac{1}{k} \sum_{k \in [K]} p_A(y_i|\theta^{(k)}).$$

The term p_{waic} is the effective number of parameters. If $V_{k=1}^{K} a_k = \frac{1}{K-1} \sum_{k=1}^{K} (a_k - \bar{a})^2$, then p_{waic} is estimated using $\hat{p}_{waic} = \sum_{i=1}^{N} V_{k=1}^{K} (\log(p(y_i|\theta^{(k)})))$. The `waic` function from R package `loo` [Vehtari et al., 2017] is used for $elpd_{waic}$ estimation.

The `PLmix` package in R provides a range of model selection criterion to select the optimal number of mixture components D. We use the Deviation Information Criterion to select the optimal model on a given data. Similar model selection procedures are implemented for the Mallows model.

E.1.1 Model comparison on the 'Royal Acta' Data

Table E.1 summarises the estimated $elpd_{waic}$ for the six models, on three signature dataset - 1080-1084, 1126-1130 and 1134-1138(b) (5PLA). The VSP/QJ models outperforms the PL, PL-mixture, Mallows and Mallows moxture models significantly in all time periods. The VSP/QJ-B model is relatively favourable compared to VSP/QJ-U. We note that we made no careful choice of priors on the PL models and the Mallows models. Non-informative priors are adapted in both cases so it is possible the performance of these models could be improved. However, they have a long way to go to catch up.

We estimate consensus orders for both the PL and PL-Mixture models. This is done by first sampling from the posterior distribution of ranking(s). We turn the rankings into partial order representations. For a PL-mixture, we calculate the intersection order that records the order relation appearing in all rankings. The consensus order is then constructed from this 'posterior distribution of partial orders'. The estimated consensus orders for the PL and PL-Mixture (D=2) models are shown in Figure E.1.

	$elpd_{waic}$ (se)		
Model	**1080-1084**	**1126-1130**	**1134-1138(b)**
VSP/QJ-B	-103.5 (26.0)	-28.6 (9.6)	-72.2 (21.9)
VSP/QJ-U	-197.2 (77.8)	-37.8 (10.8)	-86.3 (27.6)
PL	-316.5 (38.5)	-270.4 (25.8)	-336.2 (35.6)
PL-Mix2	-291.1 (37.2)	-267.6 (24.7)	-318.6 (36.3)
Mallows	-601.9 (6.8)	-624.5 (3.0)	-770.2 (7.6)
Mallows-Mix	-613.9 (4.1) (D=4)	-604.7 (1.9) (D=6)	-820.7 (4.8) (D=4)

Table E.1: The estimated $elpd_{waic}$ (se) under six different models - VSP/QJ-U, VSP/QJ-B, Plackett-Luce (PL) and 2-mixture Plackett-Luce (PL-Mix2) model.

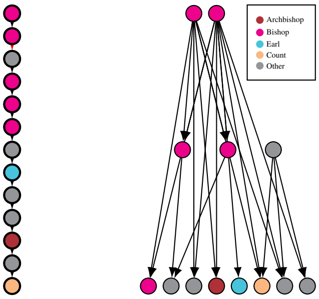


Figure E.1: The estimated consensus orders from the Plackett-Luce (left) and PL-Mixture (D=2) (right) models on the 1126-1130 data. Red edges indicate order relations that posterior probabilities are higher than 0.9.

Both the PL and PL-Mixture (D=2) model are not designed to reconstruct partial orders in the way we use it here. It was of interest to see if they did capture the same or similar relations to those we find with VSP models. This is not the case. Although we don't know the true partial order, we do expect a fairly deep social hierarchy in the 12th century. Neither model reflects such a feature.

E.1.2 Model comparison on the Formula 1 Race Data

We compare the VSP/QJ-D model with the Placket-Luce and Mallows model, and their mixtures on the Formula 1 dataset. The comparison result using $elpd_{waic}$ is shown in table E.2. The VSP/QJ-D model outperforms both the Plackett-Luce, the Mallows and their mixtures significantly.

Model	$elpd_{waic}$ (se)
VSP/QJ-D	-597.1 (25.2)
PL	-847.4 (18.6)
PL-Mix2	-821.6 (17.4)
Mallows	-973.7 (3.4)
Mallows-Mix3	-963.5 (3.9)

Table E.2: The estimated $elpd_{waic}$ (se) under five different models for the Formula 1 Racing Data - VSP/QJ-D, Plackett-Luce (PL) and 2-mixture Plackett-Luce (PL-Mix2), Mallows and 3-Mixture Mallows (Mallows-Mix3) model.

E.2 MODEL COMPARISON VSP V. BUCKET ORDER

Bayes factors B_{01} for bucket orders (see Section 1 over VSPs can be estimated using the Savage-Dickey Ratio. Results are summarized in Table E.3 for both models QJ-U and QJ-B and both 1LPA and 5LPA datasets. Numbers above one support bucket orders. Numbers below one support VSPs. For 1PLA dataset, we observe strong support for VSPs. For 5LPA data there is a very slight preference for bucket orders "barely worth mentioning" over QJ-B. Presumably the extra model complexity of QJ-B is costing something here. For QJ-U and the period 1180-84 there is no strong preference - the consensus order in Fig. 7 is “nearly” a bucket order. However, for QJ-U, 1126-30 and 1134-38 and 1134-38(b) the consensus orders are more complex and VSP’s are strongly preferred over Bucket orders.

	Bayes Factor B_{01}	
Dataset	**VSP/QJ-U**	**VSP/QJ-B**
1080-1084	1.73	2.83
1126-1130	0.18	2.83
1134-1138	0.00	NA
1134-1138(b)	0.33	2.59

	Bayes Factor B_{01}
Dataset	**VSP/QJ-U**
1080-1084	0.00
1134-1138	0.00

Table E.3: The Bayes factors B_{01} for ‘bucket’ order over VSP on all datasets 5LPA (Left) and 1LPA (Right).

E.3 MODEL COMPARISON WITH THE LATENT PARTIAL ORDER MODEL

Nicholls and Muir Watt [2011] proposes a latent partial order model, which can be applied to fit general partial orders to rank-order list-data. Though their method is not scalable to datasets of more than around 20 actors, we are interested in comparing the performance between their partial order (PO) model and the VSP class of models proposed in this paper. We choose the same observation model, QJ-U, to make the test. We choose a relatively small dataset, 1126-1130 with 5LPA, for this comparison, so the full PO model is tractable. We chose priors $\rho \sim Beta(1, \frac{1}{6})$ as suggested in Nicholls and Muir Watt [2011] and a non-informative prior for the error probability $p = \frac{e^r}{1+e^r}$ where $r \sim \mathcal{N}(0, 1.5)$ in order to get a reasonably flat depth distribution for the PO-prior.

The consensus order from the PO/QJ-U model is shown in Fig. E.2 (left). We also copy the result from the VSP/QJ-U model here for comparison. The two models indicates similar social hierarchy. However, the PO/QJ-U model presents a less strict hierarchy among bishops.

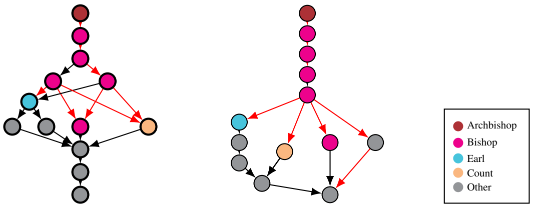


Figure E.2: PO/QJ-U model(left) and VSP/QJ-U model (right; same as Fig. 7). Consensus order for 1126-1130 5LPA data. Significant/strong order relations are indicated by black/red edges respectively.

The consensus order from the PO/QJ-U model is actually a VSP. Fig. E.3 shows the prior and posterior depth distributions for both the PO/QJ-U and VSP/QJ-U models. Although the prior distributions over depth are all relatively flat for the two models, the PO/QJ-U model favour partial orders with relatively lower depth.

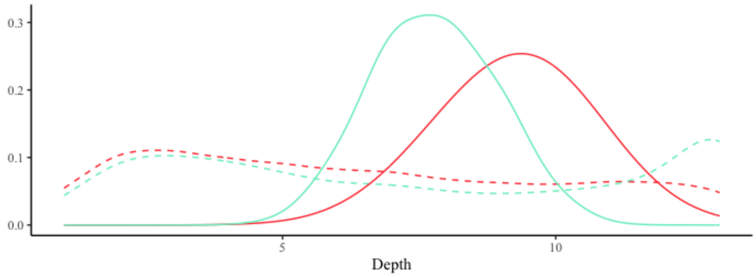


Figure E.3: The prior (dashed) and posterior (solid) distribution over depth for the PO/QJ-U (green) and VSP/QJ-U model (red).

The posterior probability to get a VSP given the PO/QJ-U model is $p_{PO/QJ-U}(h \in \mathcal{V}_{[n]}|\mathbf{y}) = 0.31$ so there is a reasonable chance in the more general model that the unknown true social hierarchy is a VSP. The model comparison performed in

Table E.4 indicates similar $elpd_{waic}$ for both models. Considering the uncertainty in our estimation, we conclude both models fitting the data equally well.

Model	$elpd_{waic}$ **(***se***)**
VSP/QJ-U	-37.8 (10.8)
PO/QJ-U	-36.7 (10.1)

Table E.4: The estimated $elpd_{waic}$ (se) for the VSP/QJ-U and PO/QJ-U models.

We compare the average ranking for different professions in table E.5 and observe the same ranking order in professions although ranking scales are slight different.

	Average Rank	
Profession	PO/QJ-U	VSP/QJ-U
Archbishop	1 (0.08)	1 (0.08)
Bishop	3.76 (0.29)	3.99 (0.31)
Earl	5.75 (0.44)	6.93 (0.53)
Count	6.04 (0.46)	8.94 (0.69)
Other	9.28 (0.71)	10.40 (0.80)

Table E.5: The professions and their average rankings under the PO/QJ-U and VSP/QJ-U models for time period 1126-1130.

We summarise the posterior distributions over POs/VSPs using the consensus adjacency matrix m, such that

$$m_{i,j} = p(i \succ j|\mathbf{y}), i, j \in [n].$$

The consensus orders are inferred from the consensus adjacency matrix by setting a certain threshold. This paper chooses a threshold of 0.5. Fig. E.4 plots the entries of the two consensus adjacency matrices against each other. The points roughly scatter along the reference line $y = x$, and show a positive monotone trend. Based on Fig. E.4, the two consensus adjacency matrices roughly agree with each other, highlighting the fact that although the VSP is a more restricted model, it works as well as a flexible and scalable partial order model in social hierarchy scenarios.

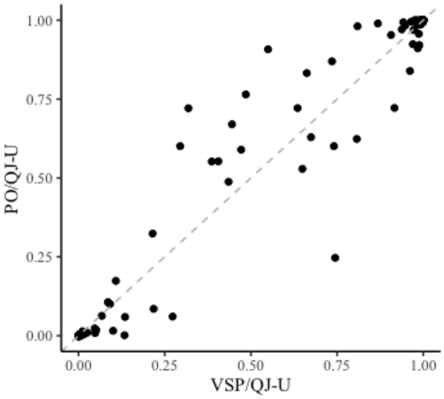


Figure E.4: The comparison plot between the consensus adjacency matrices from the VSP/QJ-U (x-axis) and PO/QJ-U (y-axis) models. The gray dashed line is the $y = x$ reference line.

F SCALING ANALYSIS

Counting the number of linear extensions of a general partial order is known to be #P-complete (Brightwell and Winkler [1991]). *LEcount* by Kangas et al. [2016] seems to be the most computationally efficient counting tool available. *LEcount* chooses between two algorithms, one counts by recursion in $O(2^n n)$ operations and the other by variable elimination in $O(n^{t+4})$ where t is the treewidth of the cover graph. The linear-extension counting algorithm we use exploits the tree representation (1, 2) so it only works for VSPs, but it is more reliable and faster than *LEcount* especially for the complicated and large VSPs at the right end of Fig. F.1.

The likelihood evaluation involves substantial computation of the number of linear extensions, and is an essential part of our MCMC analysis. We compare the computational cost to the likelihood evaluations under either the VSP tree representation or *LEcount*. This is done by simulating $N = 20$ full length lists on VSPs of increasing size $n = 3, 6, ..., 39$ from our VSP prior. For each group of N lists we evaluate the likelihood for the VSP used in simulation. We repeat this 50 times for each VSP size n for each method to derive an estimated distribution over run-times. The log-scaled maximum run-time (in seconds) for each sample size is shown in Fig. F.1. The log-scaled maximum run-time appears to be linear for the tree representation and exponential for *LEcount*. The optimised *LEcount* approach outperforms the tree representation LE evaluation when we have VSPs less than 25 actors. However, VSP-based counting significantly outperforms *LEcount* when we move to much larger datasets (completely as expected, all that matters is that we are comparing a simple implementation of a fast VSP algorithm with a well optimised implementation of a PO algorithm and the simple VSP implementation still beats the optimised PO implementation at large enough VSP sizes because the VSP algorithm only works for a subset of POs, so there is no criticism of LEcount here).

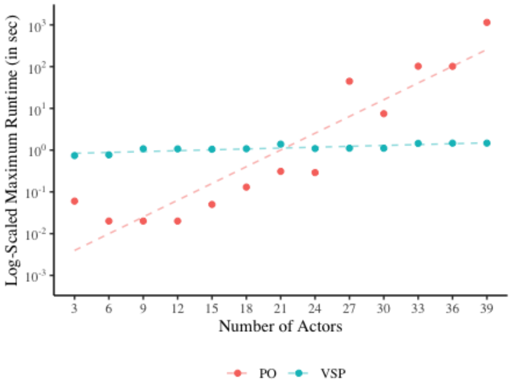


Figure F.1: Run-time analysis between the count approach from tree representation and *LEcount* (Kangas et al. [2016]) on VSPs. The plot compares likelihood (QJ-U) evaluation exploiting the VSP structure (in green) and for a general PO (in red). The log-scaled maximum run-time (in seconds) from the tree representation (green) and the *LEcount* is shown in y-axis, and the number of actors in VSP is shown in the x-axis.

The scaling analysis demonstrates the high scalability of the VSP counting method via the tree representation. This enables our model to work on datasets with more than 200 actors, see Section D.1.1.

G DETECTING VSP'S

Valdes et al. [1979] proposes an efficient way to recognise VSP's by detecting the so-called *forbidden sub-graph* (Fig. G.1).

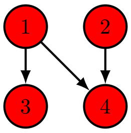


Figure G.1: The 'forbidden sub-graph' to the VSP class of partial orders.

A partial order $h \in \mathcal{H}_{[n]}$ is a VSP if it does not contain a set of vertices $o = \{j_1, \ldots, j_4\} \subset [n]$ with sub-graph $h = h[o]$ that is isomorphic to the 'forbidden sub-graph' $F = ([4], \{\langle 1, 3\rangle, \langle 1, 4\rangle, \langle 2, 4\rangle\})$. If two graphs are isomorphic, F and h' in our case, they must be identical after vertex relabelling. This means edges absent in F must also be absent in h'. This makes it straightforward to test if a partial order is a VSP.

H PRIOR DISTRIBUTION ON DEPTH

Our VSP-prior gives good control over partial order depth. We can choose the prior distribution over q so that the marginal distribution $\pi_{\mathcal{V}_{[n]}}(v)$ has a reasonably flat distribution over the depth $D(v)$ of the VSP-partial order v. This ensures the prior is non-informative with respect to partial-order depth, a property of a social hierarchy on actors which is of particular interest. After some experimentation we found that taking $\eta \sim \mathcal{N}(1, 1.5)$ and setting $q = \frac{1}{1+e^{-\eta}}$ gave a reasonably non-informative depth distribution. Fig. H.1 shows an example prior depth distribution for partial orders with 50 actors under this prior.

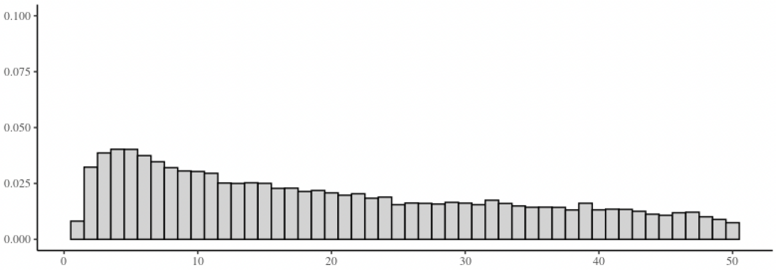


Figure H.1: The prior distribution over depth for partial orders with 50 actors, when $q = \frac{1}{1+e^{-\eta}}, \eta \sim \mathcal{N}(1, 1.5)$.

References

Formula 1 template. https://www.spreadsheet.com/template/formula-1. Accessed: 2023-04-23.

Graham Brightwell and Peter Winkler. Counting linear extensions. *Order*, 8(3):225–242, 1991.

Francois Caron and Arnaud Doucet. Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

Kustaa Kangas, Teemu Hankala, Teppo Mikael Niinimäki, and Mikko Koivisto. Counting linear extensions of sparse posets. In *IJCAI*, pages 603–609, 2016.

Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

Cristina Mollica and Luca Tardella. Bayesian Plackett-Luce mixture models for partially ranked data. *Psychometrika*, 82(2): 442–458, 2017. ISSN 0033-3123. doi: 10.1007/s11336-016-9530-0.

Cristina Mollica and Luca Tardella. PLMIX: An R package for modelling and clustering partially ranked data. *Journal of Statistical Computation and Simulation*, 90(5):925–959, 2020.

Geoff K Nicholls and Alexis Muir Watt. Partial order models for episcopal social status in 12th century England. *IWSM 2011*, page 437, 2011.

R. Sharpe, D. Carpenter, H. Doherty, M. Hagger, and N. Karn. The Charters of William II and Henry I. Online: Last accessed 27 October 2022, 2014.

Øystein Sørensen, Marta Crispino, Qinghua Liu, and Valeria Vitelli. BayesMallows: An R package for the Bayesian Mallows model. *The R Journal*, 12(1):324–342, 2020. doi: 10.32614/RJ-2020-026.

Richard Stanley and Eric W. Weisstein. *Catalan Number*. https://mathworld.wolfram.com/CatalanNumber.html, 2002. MathWorld–A Wolfram Web Resource.

Jacobo Valdes. *Parsing Flowcharts and Series-Parallel Graphs.* PhD thesis, Stanford, CA, USA, 1978. AAI7905944.

Jacobo Valdes, Robert E Tarjan, and Eugene L Lawler. The recognition of series parallel digraphs. In *Proceedings of the eleventh annual ACM symposium on Theory of computing*, pages 1–12, 1979.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017.

Valeria Vitelli, Øystein Sørensen, Marta Crispino, Arnoldo Frigessi Di Rattalma, and Elja Arjas. Probabilistic preference learning with the mallows rank model. *Journal of Machine Learning Research*, 18(158):1–49, 2018.