Bayesian Inference on Partial Orders from Random Rank-Order Data

Chuxuan (Jessie) Jiang
Christ Church
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2024

This thesis is dedicated to my family:
Jin, Tianshuo and Jianhui.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Geoff Nicholls, whose expertise, guidance, and unwavering support have been instrumental in the completion of this thesis. His insightful feedback and encouragement at every stage of my research have been invaluable, and I am deeply appreciative of his dedication to my academic and professional growth.

I am also immensely grateful to Dr. Kate Lee and Dr. Jessie Wu for their generous support and invaluable input throughout my research journey. Their thoughtful advice and willingness to share their expertise significantly contributed to the development of this work.

Special thanks go to Prof. François Caron and Prof. Arnaud Doucet for their constructive comments and suggestions. Their critical insights helped refine and enhance the quality of this thesis, and I am thankful for their time and effort.

This thesis would not have been possible without the camaraderie of my research group: Laura, Hanwen, Lorenzo, and Schyan. I deeply appreciate their availability for discussions, brainstorming sessions, and their continuous encouragement. Their collaboration made the research process not only more productive but also more enjoyable. I would also like to extend my gratitude to my cohorts from the Statistics department: Silvia, Anya, Desi, Alex, Carlo, Faaiz, Stefano and Dan. Their openness to share ideas, coupled with the laughter and fun memories we created together, made my research journey an enriching and memorable experience.

To my friends, thank you for your constant encouragement, patience, and unwavering belief in me. My heartfelt thanks to my warm Christ Church friends—Sarah, Jenna, Richard, Celine, Milena, Linnea, Shu, Casey, and Peter—for enriching my everyday life with their friendship. I am equally grateful to my OxAI friends—Odhran, Ivan, Yusuf, Puyu, Udai, and Sewook—whose insights have been invaluable not only for my research

but also for my career progression. I feel forever blessed to have met all my friends in Oxford; your support provided the strength I needed to persevere through the challenges of this process.

I want to express my gratitude to my best friend, Annabella, who has consistently supported me through all the highs and lows of both my personal and academic journey.

Finally, I would like to express my deepest appreciation to my family: Jin, Tianshuo, and Jianhui. Without their strong support, this DPhil would not have been possible. I am forever grateful for their unwavering belief in me.

Abstract

Ranking problems are prevalent across various fields, yet most existing models focus on reconstructing total orders rather than partial orders. Models parameterised by partial orders make weaker assumptions about the unknown underlying order relations. This thesis proposes a new class of generic ranking models parameterised by partial orders. We study their properties and develop efficient computational methods for Bayesian inference.

We develop partial order ranking models on both homogeneous and heterogeneous rank-order data. The proposed models are in general marginally consistent[1] and provide control over partial order depth[2], properties that are not inherently provided by traditional models. We extend our basic model in several ways. Ties between elements of the partial order are modeled using non-parametric clustering and the "dimension" of the partial order is estimated using a separate non-parametric approach. Although partial orders offer great generality, inference can be formidable on large datasets. We work with vertex-series-parallel partial orders (VSPs, [88]), a scalable class that can be parameterized using binary decomposition trees. We develop statistical models for random VSPs and propose scalable inference schemes. For heterogeneous setting where rank-orders are provided by various assessors, we introduce a hierarchical partial order model (HPO) with a tree-like parameter-dependence structure. This model specifies a global partial order that governs "leaf" partial orders representing individual assessor preferences, complemented by a VSP approximation approach.

We also address observation errors inherent in rank-order data by investigating various noise models which sit on top of partial order models, including the Mallows model and the Plackett-Luce model. We extend the

[1] See 1.3.

[2] The depth of a partial order h is the length of the longest ordered sequence in h.

"queue-jumping" model [71] to incorporate bi-directional queue-jumping behavior, aligning more closely with real-world scenarios.

While our partial order ranking models are primarily applied to social hierarchy studies, their utility extends to a wide range of ranking problems. We demonstrate their application to analysis of multi-player competition outcomes, user preference data, and more. Comparative analysis shows that our models are advantageous due to their flexibility in capturing diverse preference behaviors.

Contents

Chapter 1

Introduction

Given a "ground set" of items, rank-order data provides a collection of ordered lists from most preferred to least preferred, or from first to last. An ordered list can be a complete ranking of the ground set or an incomplete ranking of a subset of the ground set. Statistical analysis of such rank-order data seeks to identify a "central" ranking parameter of the ordered lists. The class of Mallows models [55] seeks the "central" ranking according to some distance metrics, such as Kendall's tau or Spearman's rho distances. Other models determine the "central" ranking parameter based on latent scores. Notable examples include the related Plackett-Luce [53, 76], Thurstone [86], and Bradley-Terry [7] models. Hierarchical mixture models have been developed for "grouped" data structures, where the rank-orders come from a set of "assessors". Examples include Mallows [62] and Plackett-Luce [50, 65, 87] mixtures. As discussed in [22] and [70], there are two interpretations of the "central" ranking parameter: first, as an "approximation" of the ground truth, and second, as a heuristic summary of the ordered lists. We assume there is some underlying true order relation among the ground set of items (but this need only be a partial order). This work, therefore, fits the first interpretation.

The "central" ranking parameter can be total or partial in the statistical literature. In contrast to *total orders*, which assume strict order relations between any pair of items, *partial orders* (*partially ordered sets* or simply *posets*) allow items to be *incomparable*. A pair of items is incomparable if no order relation exists between them. A total order is a special case of a partial order where all items are comparable. Many ranking models pursue total orders [53, 55, 76]. Some recent statistical literature focuses on partial orders derived from rank-order data, which can be roughly classified into two categories. The first category reconstructs a partial order through graphical learning [13, 21], focusing on inferring edge probabilities (pairwise preference probabilities). The second class of methods treats partial orders as a model

parameter in probabilistic modeling. In this setting, there has been frequentist work [6, 27, 57]. Notably, [57] is among the first statistical literature to place partial orders in a generative model for rank-order data. They consider two classes of partial orders: *vertex-series-parallel partial orders* [88] and *bucket orders*[1], which we will introduce in Section 1.2. Bayesian work in this area is limited to date. [79] provides Bayesian inference for partial orders defining a conjunctive Bayesian network model. [71] and [69, 70] propose latent partial order models for static and time series rank-order data, respectively.

This work proposes probabilistic models for the partial order "parameter" inferred from rank-order data. We infer partial orders in a Bayesian framework. We extend the existing literature [69, 71] by relaxing some key properties of the partial order distribution in chapter 2 (section 2.4), considering various types of noise which might arise in rank-order data (both of these for better model fit) in chapter 2 (section 3.2) and chapter 3 (section 3), and restricting the class of partial orders to VSPs (for scalablity) in chapter 3 (section 2). We further develop a hierarchical partial order model for grouped data in chapter 4 (section 3). We examine the properties of these partial order models and construct Markov Chain Monte Carlo (MCMC)/Bayesian inference schemes to reconstruct the true order relations behind the observed rank-orders.

1.1 Literature Review

1.1.1 Ranking Models

Ranking problems arise in many fields of application and have been extensively studied in the statistical literature. A wide range of ranking models has been proposed; see [1, 51, 59] for comprehensive overviews. Early work includes the Thurstone model [86], which considers pairwise ranking based on order statistics of a Gaussian random vector. The Bradley-Terry model [7] focuses on an exponential family of pairwise comparisons. Further developments in the literature have introduced a variety of ranking models which come in two main classes: distance based and score-based.

The most popular distance-based ranking models are the Mallows-type models. Mallows-type models are specified by a "central" ranking ℓ and a dispersion parameter $\theta \in (0, \infty)$. Given an observed rank-order y, the model can be expressed $p(y|\ell, \theta) \sim exp(-\theta d(\ell, y))$, where the discrepancy between the observed ranking and

[1]Bucket orders are a class of partial orders where incomparable items come in groups or "buckets" which are themselves arranged in a total order.

the "central" ranking is given by a distance function $d(\cdot, \cdot)$. Different choices of distance function lead to different models. For example, Kendall's tau distance is frequently considered in the literature due to its analytical normalizing constant [52, 62] and many good properties [17]. Some studies consider other distance metrics. [89] examines right-invariant distances, including the footrule and Spearman's rho distances, and points out that footrule is often a good modeling choice. Other works explore the Hamming distance [33, 36], the Cayley distance [35] and the reverse major index[2] [23]. The dispersion parameter θ governs the distance distribution between the "central" ranking and the realised rank-orders. It is allowed to vary across rank positions in Generalised Mallows models [24]. This extension allows for emphasis on consensus at certain positions while permitting uncertainty at others. Mallows-type models have been given non-parametric extensions [47], infinite permutations for top-k partial rankings [61], tied-Mallows models [75], and mixture models [49, 52, 62, 89]. Mixture models arise in the context of "grouped" data structures where observed rank-orders come from a heterogeneous population. [63] performs Bayesian non-parametric inference on incomplete rankings for the generalized Mallows model. [4, 46, 49, 52, 89] approximate Bayesian inference for Mallows mixture models. Tied Mallows models consider ties between items in the "central" ranking. This is similar to the concept of "bucket" orders. Existing R-packages on Mallows inference includes `BayesMallows` [94] which performs Bayesian inference on Mallows and Mallows mixtures, and `PerMallows` [34] which estimates the MLE for both the Mallows and Generalised Mallows model. Both packages support various distance metrics.

Score-based ranking models differ fundamentally from distance based models. Extending the Bradley-Terry model to multiple items gives us the Plackett-Luce model [53, 76], which is among the most popular rank-score-based ranking models. Its parameter estimation can be done using either frequentist [32, 60] or Bayesian [12, 29] methodology. The Plackett-Luce model is a multistage (repeated selection) ranking model [25], where the ranking process is performed sequentially from most to least favored. The Extended Plackett-Luce model proposed in [65] relaxes the assumption of the forward ranking process to allow for a flexible rank assignment scheme. It is further studied in [67] and [40]. The Plackett-Luce model possess marginal consistency [32]. This property makes the Plackett-Luce model straightforward to

[2]The reverse major index is similar to Kendall's tau distance but puts more weight on top-position deviations.

apply to incomplete rankings and top-k rankings [28, 31, 66]. There is a rich variety of Plackett-Luce mixture models [93]. This includes the non-parametric mixture model for rank-order clustering [12] with efficient Monte Carlo sampling in [11]. [66] performs Bayesian finite mixture modeling of Plackett-Luce models. [28] allows individual assessors to have partial membership in more than one group. It is worth noting that [50] learns Plackett-Luce mixture from partial order data, and [54] learns the Plackett-Luce model from partitioned preference (bucket orders). This is the reverse of our setting, as we seek to infer partial orders from observed total rank-orders. Existing statistical software packages for the Plackett-Luce model include the R-package `PLMix` [68], which fits finite mixtures of Plackett-Luce models for incomplete top rankings/orderings within the Bayesian framework.

The Plackett-Luce model is the unique rank-score based model (or the so called random utility model [58]) that satisfies the *independence of irrelevant alternatives (IIA)* [53]. Informally, IIA means that adding an additional item to a choice set does not affect the relative choice-probabilities for other items. A Context Dependent Random Utility Model (CDM, due to [80]) extends Plackett-Luce to allow IIA violations, by parameterising each pairwise interaction separately. [81] develops CDMs in a repeated selection framework, leading to their Contextual Repeated Selection (CRS) model. The CRS model captures multimodality and intransitivity which is important in some settings. However, such generalisation can incur a cost of slow convergence with increased data-size, as dropping transitivity looses a lot of information in settings where two items seldom appear in the same list. The CRS model is like a partial order model in some ways, as both are context-dependent and don't assume a single underlying order. However, it is straightforward to show the models parameterise fundamentally different observation models, with transitivity the key distinguishing feature.

Our partial order based parameterisation is closest to the Plackett-Luce model as our latent variable setup is similar to the Plackett-Luce rank score parameterisation. It therefore inherits some nice features from the Plackett-Luce model. Our model flexibly adapts to subset-data where the assessors are presented with different subsets for ranking (due to marginal consistency). It is also a sequential-choice model, and this makes it straightforward to handle top-k applications, where the assessors only rank their top k preferences. Although our VSP models do not use a latent-score parameterisation but exploit the binary tree representation of a VSP they nevertheless share the same convenient properties on subset-data. On the other hand, the Plackett-

Luce model is context independent. The Mallows, CRS and our partial order models are in general context dependent.

1.1.2 Statistical Model on Partial Orders

Work on inferring partial orders from rank-orders were pioneered by [57] and [27]. Both use partial orders for event sequencing, and analyse vertex-series-parallel orders and bucket orders respectively. In both cases, "noise-free" likelihood models are employed, where the observed rank-orders are assumed to be random linear extensions[3] (total orders) respecting an unknown underlying true partial order. [57] adopts a greedy search for partial order discovery, while [27] develops a Bucket Pivot (sorting) algorithm to search for bucket orders. Bucket orders are frequently adopted in statistical literature due to their fast likelihood evaluation (counting linear extensions is at worst linear in the number of items in a rank-order list). [77] and [56] perform Bayesian analysis on bucket orders for fossil sites seriation. [22] infers bucket orders from a collection of possibly noisy complete rankings and develops the Bucket Gap algorithm for bucket order discovery. [43] extends this work to consider both "local" and arbitrary noise in the observed rank-orders. In terms of the general class of partial orders, [6] develops a maximum likelihood estimator for their conjunctive Bayesian network model, a probabilistic graphical model represented by finite posets. [79] further develops a Bayesian inference scheme for such a model. Both studies aim to map order-constraints on the accumulation of cancer-related genetic mutations. A similar study, [26], learns partial orders as the gene signaling network through stochastic sampling. None of the above works consider a latent prior model on random partial orders, which can be desirable for controlling key properties of the partial order distribution. [79] use a uniform prior on partial orders, which is known to be problematic when the number of ranked items is large (the prior concentrates on depth three) but acceptably diffuse for partial orders on any small number of items. Early work proposing a latent random-matrix approach to determine a probability distribution over partial orders is by [91]. This model represents each ranked item in the poset with a pair of real numbers. Item A is ordered above item B if each components of A's latent variables lie above the corresponding component for B. See section 1.4.2 for more details. This approach was generalised in [71] and [69] for better control of partial order properties. [70] extends their work to the time series setting.

[3] A linear extension of the partial order $h_{\mathcal{M}}$ on a ground set of item $\mathcal{M}$ is a permutation of $\mathcal{M}$ that does not violate h.

Partial orders are also implemented in other statistical settings. For example, they are used for causal/structure discovery [30, 72]. [73] uses Metropolis-coupled MCMC and annealed importance sampling to sample posets that support estimation of the posterior distribution of Bayesian Networks. However, the poset is not a parameter in the inference in that setting. Partial orders are also used to summarize uncertainty in ranking. [78] uses the partial order as a summary statistic for order relations between parameter estimates given their confidence intervals. [82] proposes a partial order based probabilistic model to encapsulate possible rankings from score uncertainty. Some literature adopts partial orders as a support structure in computation. For example, [15] applies partial order to sorting and searching problems. Other works use partially ordered sets as input data. [2] searches for optimally approximating bucket orders by minimizing the dissimilarity between the fitted and input posets, utilizing their "total order with ties" structure. Other uses of partial orders include [92] and [74], who consider partial orders in generalised linear models for categorical variables. In our language an assessor is presented with many different sets of choices and chooses one item from each set (top-k with $k = 1$). The authors impose a partially ordered structure (in fact a VSP, but the connection isn't made) on assessor preferences. [38] and [37] propose a non-parametric, choice-based demand model where individual preference is characterized by a partial order. They develop a Plackett-Luce model conditioned on partial orders for top-k ranking predictions, providing exact likelihood evaluation when the partial order can be represented as a forest of directed trees.

1.2 Partial Orders

Denote the ground set of M items as $\mathcal{M}$ so $|\mathcal{M}| = M$. We denote $\mathcal{B}_{\mathcal{M}} = \cup_{S \subseteq \mathcal{M}, S \neq \emptyset}\{S\}$ as the set of all subsets of $\mathcal{M}$ that are not empty. A partial order (partially ordered set) defines an order relation among a set of items.

Definition 1 (Partial Orders) *A partial order (partially ordered set) $h = (S, \prec)$ is a binary relation over a set of items $S \in \mathcal{B}_{\mathcal{M}}$, that is both irreflexive and transitive. Let $i, j, k \in S$,*

- *Irreflexive: the relation $i \prec_h i$ does not exist;*
- *Transitive: if $i \prec_h j$ and $j \prec_h k$, then $i \prec_h k$.*

We denote $\mathcal{H}_S$ the set of partial orders on S.

The partial orders in this thesis are "strong" or "strict" partial orders ("weak" or "non-strict" partial orders are reflexive, anti-symmetric and transitive). It follows from the definition that strong partial orders are anti-symmetric ($i \prec_h j$ implies there is no relation $j \prec_h i$). The relation $j \succ_h i$ is equivalent to $i \prec_h j$.

A partial order can be represented as a Directed Acyclic Graph (DAG). One can also represent a partial order with a binary adjacency matrix as, with an abuse of notation, $h \in \{0,1\}^{|S|\times|S|}$ where $h_{i,i} = 0$ and $h_{i,j} = 1$ if and only if $i \succ_h j$. Two items $i, j \in S$ are *incomparable* if neither $i \succ_h j$ nor $i \prec_h j$, i.e. $h_{i,j} = h_{j,i} = 0$. A partial order is called a *total order* if no incomparable pair of items exists. An *empty order* is a partial order where no order relation holds between any item. A *maximal set* of a partial order $\max(h) = \{i : \nexists j \in \mathcal{M} \text{ s.t. } j \succ_h i\}$ is a set of the *maximal elements* that are not succeeded by any other items in $\mathcal{M}$.

The *transitive closure* of a DAG with vertex set S is the DAG obtained by adding edges implied by transitivity. The set $\mathcal{H}_S$ of all partial orders on the ground set S is in one to one correspondence with the set of all transitively closed directed acyclic graphs with vertex set S. The *transitive reduction* of a DAG is the DAG obtained by removing all edges implied by transitivity. It is the unique DAG with the fewest possible edges that has the same transitive closure as the original DAG. Much of the literature uses the *Hasse diagram* to visualize a partially ordered set. The Hasse diagram is essentially the same as a transitive reduction, except that all edges are arranged upwards for visualization. Figure 1.1 provides an example partial order $h_0 = (S, \prec_{h_0})$ with $S = \{1, 2, \ldots, 5\}$, the adjacency matrix of its transitive reduction, and its transitive reduction/transitive closure DAGs. For the remainder of this thesis, we present all partial orders using their transitive reductions for easier visualisation.

A *linear extension* $l = (l_1, l_2, \ldots, l_{|S|})$ of partial order $h \in \mathcal{H}_S$ is a permutation over S given that its order does not violate h, so for $1 \leq a < b \leq |S|$ we must have either $l_a \succ_h l_b$ or incomparability $l_a \| _h l_b$. Let $\mathcal{P}_S$ be the set of all possible permutations of S. The set of linear extensions of h is then $\mathcal{L}[h] = \{l \in \mathcal{P}_S | h_{l_b,l_a} \neq 1, \forall 1 \leq a < b \leq |S|\}$. The example partial order h_0 in Figure 1.1 has three linear extensions $\mathcal{L}[h_0] = \{(1,2,3,4,5), (1,3,2,4,5), (1,3,4,2,5)\}$.

Given a subset $o \subset \mathcal{M}$ of the ground set and a partial order $h \in \mathcal{H}_\mathcal{M}$ a *sub-order* $h[o]$ of h is the partial order h determined on the elements of o, that is $h[o] = (o, \{\prec: i \prec_h j \; \forall i, j \in o\})$. A *chain* is a sub-order of h that is also a total order (so any path following the directed edges of the DAG representation of h). The *length* of a chain is the number of nodes in this suborder. The length of the longest chain(s) is called the *depth*, $D(h)$ say, of partial order $h \in \mathcal{H}_\mathcal{M}$, with $1 \leq D(h) \leq M$. For example,

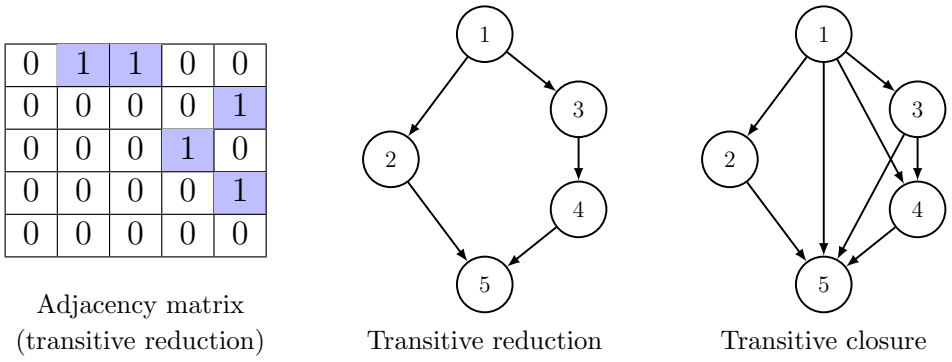


Figure 1.1: An example partial order h_0 (transitive reduction) with $|S| = 5$ items (middle), its adjacency matrix (transitive reduction) (left) and its transitive closure (right).

the longest chain on h_0 in Figure 1.1 is $(1, 3, 4, 5)$, so $D(h_0) = 4$. Partial order depth is an important property in many real-world applications. For example, in the context of social hierarchy, a greater depth indicates a more finely stratified social hierarchy. When proposing probability distributions over random partial orders, it is desirable for the model to offer some control over the partial order depth. This thesis demonstrates several ways to control the prior depth distribution over random partial orders.

1.2.1 Two Special Classes of Partial Orders

There are several subclass of partial orders worth mentioning.

Vertex-Series-Parallel Partial Orders. The vertex-series-parallel partial orders (VSPs) are a class of partial orders $\mathcal{V}_\mathcal{M} \subset \mathcal{H}_\mathcal{M}$ that was discussed in-depth in [88].

Definition 2 (Vertex-Series-Parallel Partial Orders) *The vertex-series-parallel partial orders are formed by repeated series $\otimes$ and parallel $\oplus$ operations. For partial orders h_1 and h_2, let $V(h_1)$ and $V(h_2)$ represent the ground sets of actors for h_1 and h_2 respectively (which we assume are disjoint).*

- *A series partial order, $h = h_1 \otimes h_2$, is the union of all relations in h_1 and h_2, with additional relations $i \succ_h j$ if $i \in V(h_1)$ and $j \in V(h_2)$.*
- *A parallel partial order, $h = h_1 \oplus h_2$, is the union of all relations in h_1 and h_2 with incomparability $i \|_h j$ if $i \in V(h_1)$ and $j \in V(h_2)$.*

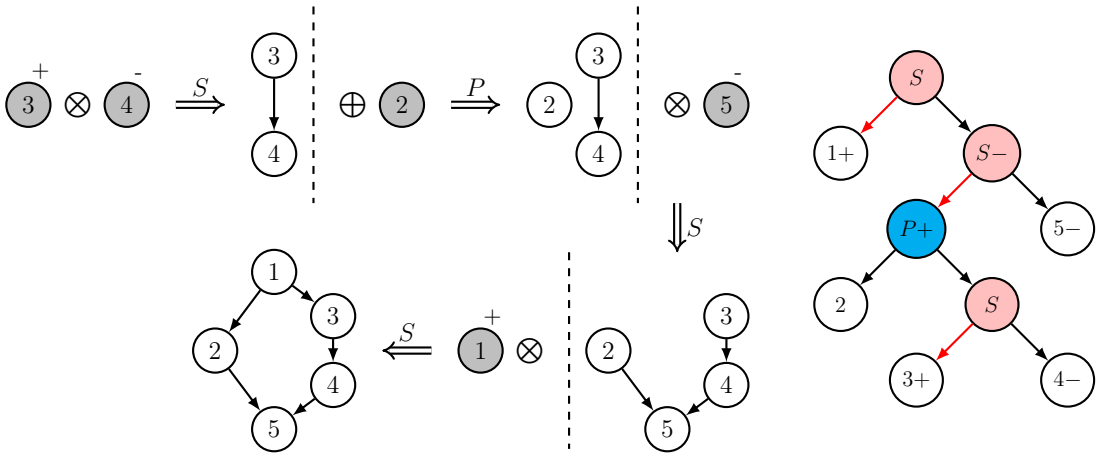


Figure 1.2: (Left) One possible construction procedure for the VSP h_0 in Figure 1.1. The symbols S and P corresponds to the *series* and *parallel* operations respectively. (Right) A BDT t_0 representing h_0 in Figure 1.1, so that $v(t_0) = h_0$. Red edges and '+' signs indicate the upper child.

The set of VSPs $\mathcal{V}_\mathcal{M}$ is defined recursively. If $|V(h)| = 1$ then h is a VSP. If h_1 and h_2 are VSPs then $h_1 \otimes h_2$ and $h_1 \oplus h_2$ are VSPs. The partial order h_0 in Figure 1.1 is a VSP. It can be constructed using the series and parallel operations in Figure 1.2.

VSPs are one to one with binary decomposition trees (BDTs) [88]. Let $\mathcal{T}_\mathcal{M}$ be the set of all BDTs on $\mathcal{M}$. A BDT $t \in \mathcal{T}_\mathcal{M}$ is a directed binary tree graph with M leaves $\mathcal{F}$ and $M-1$ internal nodes $\mathcal{A}$. The full set of nodes is $N(t) = \mathcal{F} \cup \mathcal{A}$. Each leaf corresponds to an element of $\mathcal{M}$. Each internal node has an additional attribute taking values S or P (S for series operation and P for parallel operation). These attributes record the sequence of serial and parallel operations that build up the VSP represented by the BDT. The edges of the BDT $E(t)$ are directed from the root to the leaves. Internal nodes with an S label have an additional attribute indicating which of its child nodes is the "upper child" in the series operation (indicated by a '+' and a red edge in Figure 1.2 (right)), such that the upper child is stacked above the subtree rooted by the other child node (indicated by a '-').

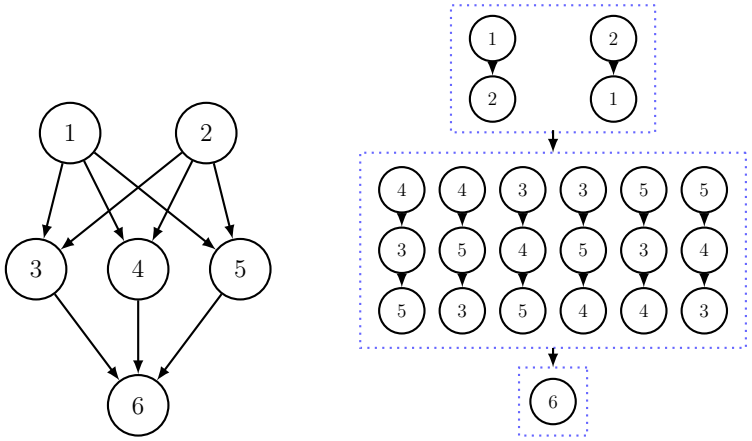
Bucket Orders. The bucket orders (BOs) are a class of partial orders $\mathcal{BO}_\mathcal{M} \subset \mathcal{H}_\mathcal{M}$, in which the actors are arranged in layers.

Definition 3 *A bucket order $h^b = (S, \prec_{h^b})$ on a set of items $S \in \mathcal{B}_\mathcal{M}$ is a partial order on S defined by a partition of S into K buckets $\{S_1, \ldots, S_K\}$ such that (taking $[K] = \{1, 2, \ldots, K\}$)*

- *For $i, j \in S_k,\ k \in [K],\ i \|_{h^b} j$;*

- *For $k_1, k_2 \in [K],\ k_1 > k_2$ and all $i \in S_{k_1},\ j \in S_{k_2}$ the relation $i \succ_{h_b} j$ holds.*

The bucket orders are a subclass of the VSPs, $\mathcal{BO}_\mathcal{M} \subset \mathcal{V}_\mathcal{M}$. One example of bucket order h_0^b with 6 items and its linear extensions $\mathcal{L}[h_0^b]$ is shown in Figure 1.3. Table 1.1 shows the number of posets, VSPs and bucket orders given different numbers of items.



Example bucket order h_0^b Linear extensions $\mathcal{L}[h_0^b]$

Figure 1.3: An example bucket order h_0^b with $M = 6$ actors and 3 buckets $\{\{1,2\},\{3,4,5\},\{6\}\}$ (transitive reduction; left) and its linear extensions $\mathcal{L}[h_0^b]$ (right).

Number of items M	Number of posets $\lvert\mathcal{H}_\mathcal{M}\rvert$	Number of VSPs $\lvert\mathcal{V}_\mathcal{M}\rvert$	Number of BOs $\lvert\mathcal{BO}_\mathcal{M}\rvert$[4]
1	1	1	1
2	3	3	3
3	19	19	13
4	219	195	75
5	4231	2791	541
10	6,611,065,248,783	32,447,734,143	102,247,563
15	~775671710204e14	~71012e14	~2.302832e14

Table 1.1: Number of posets, VSPs and bucket orders given the number of items.

Both the VSPs and BOs admit fast linear extension counting, as we will illustrate in Section 1.2.2.

[4]The number of bucket orders with M labeled nodes is

$$|\mathcal{BO}_\mathcal{M}| = \sum_{k=0}^{M} k! \left\{ {M \atop k} \right\} = \sum_{k=0}^{M} k! \left[\sum_{i=0}^{k} \frac{(-1)^{k-i} i^M}{(k-i)!i!} \right],$$

where $\left\{ {M \atop k} \right\}$ is the Stirling number of the second kind, which counts the number of ways to partition a set of M labelled objects into k nonempty unlabelled subsets.

1.2.2 Counting Linear Extensions

Counting the linear extensions of a poset is a challenging problem studied extensively in the literature. It is an #P-complete task, as shown in [9].

There are both exact and approximate algorithms for linear extension counting. The *approximation* methods include [19], which proposes a fully polynomial randomized approximation via a random walk. This method has asymptotic bounds for the expected running time of $O(\epsilon^{-2}M^3 \log^2 |\mathcal{L}[h]| \log M)$ [5] and $O(\epsilon^{-2}M^5 \log^2 M)$ [83], where ϵ is the allowed error. Subsequent works are based on rapid Markov chain mixing [10, 42]. For example, [83] introduces the adaptive relaxation Monte Carlo method that leverages exact exponential-time counting algorithms. An empirical study [84] shows that these approximation methods often perform better than the worst-case bounds for exact counting. However, these approximation methods can be prohibitive when the number of items M increases, and a low error rate ϵ is required.

Works on *exact* counting algorithms include [45] and [41]. In [45], the number of linear extensions is evaluated by recursively locating and deleting a maximal element of a poset. This method is also implemented in [69]. Some polynomial-time algorithms are proposed for special cases, such as [64] for vertex-series-parallel partial orders, [3] for when the cover graph is a tree, and [16] for when the width of the poset is bounded. The current state-of-the-art in exact linear extension counting is the `lecount` package [41], which uses the tree decomposition of the cover graph of a partial order. If t is the treewidth of the cover graph[5] of a n-element poset then the time complexity of `lecount`() is $\tilde{O}(n^{t+3})$.[6]

The class of VSPs admit fast linear extension counting. For random partial orders h_1 and h_2, let $V(h_1)$ and $V(h_2)$ represent the ground sets of actors for h_1 and h_2 respectively (which we assume are disjoint). [90] gives

$$|\mathcal{L}(h_1 \otimes h_2)| = |\mathcal{L}(h_1)||\mathcal{L}(h_2)| \tag{1.1}$$

$$|\mathcal{L}(h_1 \oplus h_2)| = |\mathcal{L}(h_1)||\mathcal{L}(h_2)| \binom{|V(h_1)| + |V(h_2)|}{|V(h_1)|} \tag{1.2}$$

where $|V(h_1)|$ and $|V(h_2)|$ give the number of items in h_1 and h_2. This may be evaluated recursively in $O(n)$ steps. The class of bucket orders offers easier evaluation on linear extensions, given that the items in a bucket follows an empty order. For a

[5]The cover graph of a poset $h = (S, \prec)$ is the directed graph $(V = S, E)$ where the edge set E is the transitive reduction of h.

[6]The notation $\tilde{O}$ hides polylogarithmic factors.

bucket order h^b with buckets $(S_1, \ldots, S_K)$ has number of linear extensions as

$$|\mathcal{L}[h^b]| = \prod_{i=1}^{K} |S_i|!. \tag{1.3}$$

The bucket order h_0^b in Figure 1.3 has 12 linear extensions, for example.

1.3 Ranking Model Properties

When working with ranking models, there are some desirable properties. Here we list a few.

Marginal consistency. A family of probability distributions is marginally consistent (projective) if all of its marginals belong to the same family. In the context of partially ordered sets, marginal consistency means that we have a family of distributions $p_S,\ S \in \mathcal{B}_\mathcal{M}$ with the property that if $h \sim p_\mathcal{M}$ then $h[S] \sim p_S$, that is, we can sample a random partial order from p_S either directly or by sampling a partial order from $p_\mathcal{M}$ and taking the suborder on S. This tends to arise in latent variable models where the relation between two elements is unaffected by the presence or absence of a third element.

Definition 4 (Marginal Consistency) *The family of probability distributions $p_S(h)$, $S \in \mathcal{B}_\mathcal{M},\ h \in \mathcal{H}_S$ is marginally consistent if, for all $S \in \mathcal{B}_\mathcal{M},\ h \sim p_\mathcal{M}$ implies $h[S] \sim p_S$. Equivalently, for all $g \in \mathcal{H}_S$,*

$$p_S(g) = \sum_{\substack{h \in \mathcal{H}_\mathcal{M} \\ h[S]=g}} p_\mathcal{M}(h).$$

We can see from Table 1.1 that the uniform distribution on partial orders would not be marginally consistent: there are 19 partial orders on three elements $\mathcal{M} = \{1, 2, 3\}$ and 3 on two elements $S = \{1, 2\}$ and since 19 isn't divisible by 3. When we apply the suborder operation which removes element 3 to the partial orders in $\mathcal{H}_{[3]}$ we can't distribute probability uniformly over $\mathcal{H}_{[2]}$.

Marginal consistency is an important property for ranking models. In problems such as some social hierarchy settings, removing an actor/ranked element doesn't have an impact on the order relations among other actors. However, as we will demonstrate in section 1.4.2, marginal consistency is not a property we obtain for free. We show that our latent variable approach supports marginal consistency.

Model consistency. Model consistency is an asymptotic property of a sequence of posterior probability distributions in which the posterior concentrates on the true parameter value when the number of observations goes to infinity. We first define the generative model for rank-order data. For some $h^\dagger \in \mathcal{H}_\mathcal{M}$ let $h^\dagger \sim \pi(\cdot)$ be a random partial order sampled from the prior, so we have the true generative model with no misspecification. Let $\mathbf{y}_{1:N} \sim p(\cdot|h^\dagger)$ be an observation model for N rank-order lists $\mathbf{y}_{1:N}$, so the list data are informed by the underlying true partial order $h^\dagger$. Let $\pi(h|\mathbf{y}_{1:N}) \propto \pi(h)p(\mathbf{y}_{1:N}|h)$ be the posterior probability for $h \in \mathcal{H}_\mathcal{M}$.

Definition 5 (Model Consistency) *A generative model for partial orders and lists is consistent if it determines a posterior satisfying*

$$\lim_{N\to\infty} \frac{\pi(h|\mathbf{y}_{1:N})}{\pi(h^\dagger|\mathbf{y}_{1:N})} = \begin{cases} 0 & \text{when } h \neq h^\dagger; \\ 1 & \text{when } h = h^\dagger. \end{cases} \tag{1.4}$$

It is easy to see that some generative models won't be consistent even if, as we wrote, we have the true generative model for $h^\dagger$. For example, if for every N, the orders fall in two groups which rank non-overlapping subsets of $\mathcal{M}$, then we can never learn about relations between elements belonging to different groups. Consistency guarantees an asymptotic convergence of the statistical ranking model. We demonstrate this property in Chapter 2.

1.4 Ranking Models

Section 1.1 reviewed the literature on a wide range of ranking models - both total order and partial order models. Here we define a few popular ones for later reference.

1.4.1 Total Order Models

As a special class of partial orders, total orders are the most predominant class in the statistical ranking literature. There exists copious probabilistic models on rank-order data that differ in rank generation mechanism or in the parametric space. Two popular total order ranking models are the Mallows model and the Plackett-Luce model. The *Mallows ranking model* discovers a "central" total order based on some distance metrics. Given the "unknown true" total order $l \in \mathcal{P}_\mathcal{M}$, a distance metric $d : \mathcal{P}_\mathcal{M} \times \mathcal{P}_\mathcal{M} \to \mathbb{R}$ and a dispersion parameter $\theta \in (0, \infty)$, the Mallows model defines

$$p(y|l, \theta) = \frac{\exp(-\theta d(l, y))}{\Psi(\theta)}, \quad y \in \mathcal{P}_\mathcal{M}, \tag{1.5}$$

where $\Psi(\theta) = \sum_{z \in \mathcal{P}_\mathcal{M}} e^{-\theta d(l,z)}$ is a normalising constant. We can obtain a closed-form for the normalising constant if $d(\cdot, \cdot)$ is the Kendall's tau distance. Inference for the Mallows model with the maximum likelihood approaches can be difficult given the versatility brought by the distance metric of choice. Some examples are [44, 52] and [85]. [89] considers the Bayesian approach for any right-invariant distance[7] [18] that also works on partial rankings. They further develop Mallows mixture models on heterogeneous population.

In contrast, the Plackett-Luce model seeks the "central" ranking by parameterising the item importance with a set of weights $\alpha = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$. The Plackett-Luce model defines the probability for ranking $y \in \mathcal{P}_\mathcal{M}$ as

$$p(y|\alpha) = \prod_{i=1}^{M} \frac{e^{\alpha_{y_i}}}{\sum_{m=i}^{n} e^{\alpha_{y_m}}}. \tag{1.6}$$

The Plackett-Luce model is a *repeated selection* model, in which an observed order $y \sim p(\cdot|\alpha)$ can be realised in a sequential (forward) process, building from the top of the list to bottom. Inference on the Plackett-Luce model is commonly done with maximum likelihood approaches, for example [32] uses a minorization–maximization algorithm. Some Bayesian inference schemes are proposed in [12, 29] and [11], where [11] proposes to use the Dirichlet process to cluster assessors in grouped data.

1.4.2 Partial Order Models

There have been several probability distributions proposed on random partial orders. [8] gives a good overview. With the exception of [70, 71] and [39] all this work is from the perspective of probability and combinatorics rather than statistics.

One obvious probability distribution on random partial orders is the uniform distribution.

Definition 6 (The Uniform Distribution over Partial Orders) *For partial order $h \in \mathcal{H}_\mathcal{M}$, the uniform distribution over partial orders $\pi_{u,\mathcal{M}}(h)$ say, assigning equal prior probability to each partial order in $\mathcal{H}_\mathcal{M}$, is*

$$\pi_{u,\mathcal{M}}(h) = \frac{1}{|\mathcal{H}_\mathcal{M}|}. \tag{1.7}$$

[7] For a right-invariant distance $d_{r-inv}(\cdot, \cdot)$, the value of d_{r-inv} does not depend on how the objects are indexed. For any permutations $\ell_1, \ell_2, \tau \in \mathcal{P}_\mathcal{M}$, we have $d_{r-inv}(\ell_1, \ell_2) = d_{r-inv}(\ell_1\tau, \ell_2\tau)$ where $\ell\tau$ is defined by $\ell\tau(i) = \ell(\tau(i))$ giving the rank assigned to item i.

However, as is illustrated in Chapter 2 and mentioned above, this class of distribution is not marginally consistent and offers no depth control.

Subsequent improvement produces a class of latent variable models. Originally proposed by [91] for $K = 2$, we call this model the (K, M)-partial order model.

Definition 7 (The (K, M)-Partial Order Model [91]) *Let $Z \in \mathbb{R}^{M\times K}$ be a random latent matrix with fixed column dimension $K \in \mathbb{Z}^+$. The (K, M)-partial order model defines the entries of Z as*

$$Z_{i,k} \overset{i.i.d.}{\sim} \mathcal{U}[0, 1], \forall i, k \in \mathcal{M} \times [K].$$

Write $Z_{i,1:K} \succ Z_{j,1:K}$ if $Z_{i,k} > Z_{j,k}\, \forall k \in \{1, \ldots, K\}, i, j \in \mathcal{M}$. We define the mapping $h : \mathbb{R}^{M\times K} \to \mathcal{H}_{\mathcal{M}}$ such that

$$h(Z) = \{h \in \mathcal{H}_{\mathcal{M}} : i \succ_h j \text{ if and only if } Z_{i,1:K} \succ Z_{j,1:K}\}. \tag{1.8}$$

The latent matrix Z can be interpreted as a matrix of M latent "paths". A path is a piecewise linear continuous curve defined by connecting the chain of points $(k, Z_{i,k})_{k=1}^K$ by line segments. There is one path for each item $i \in \mathcal{M}$ in the partial order. Two items $i, j \in \mathcal{M}$ are incomparable if their paths cross. An order relation $i \succ_h j$ (or $i \prec_h j$) exists if item i's path lies entirely above (or below) the path for j.

The (K, M)-partial order model is marginally consistent. This is straightforward to verify. If we start with a random partial order on $\mathcal{M}$ and then remove item i by taking a suborder $h_{-i} = h[\mathcal{M} \setminus \{i\}]$ then this is the same as removing the row/path $Z_{i,1:K}$ for item i from the Z-matrix to get Z_{-i} say, and then forming $h(Z_{-i})$. Since the rows of Z are independent, the joint distribution of the elements of Z_{-i} is just the same as if i was never present so the partial order $h(Z_{-i})$ was simulated on $\mathcal{M} \setminus \{i\}$. Removing the row $Z_{i,1:K}$ doesn't affect relations between other items (relations between other paths) determined by $h(Z_{-i})$, so we not only have $h_{-i} \sim h(Z_{-i})$ (possibly different but having the same distribution) but $h_{-i} = h(Z_{-i})$ (identically equal) in this construction. The random partial order we get by simulating a partial order on the full set $\mathcal{M}$ and then removing i has the same distribution as a random partial order simulated on the reduced set.

The (K, M)-partial order model also offers depth control by adjusting the number of latent attributes K. The greater K is, the higher chance for two latent paths to cross, therefore, the lower depth a partial order $h(Z)$ would be. However, this depth control can be clumsy. [71] and [69] extend the (K, M)-partial order model to introduce a depth control parameter $\rho \in [0, 1]$.

Definition 8 (The (K, M, ρ)-Partial Order Model [71]) *Let Σ_ρ be a $K \times K$ correlation matrix with $\Sigma_{\rho,i,i} = 1$ and $\Sigma_{\rho,i,j} = \rho,\ i \neq j,\ i, j \in [K]$. In the (K, M, ρ)-model*

$$Z_{i,1:K}|\rho \overset{i.i.d.}{\sim} \mathcal{MVN}(\mathbf{0}, \Sigma_\rho),\ i \in \{1, \ldots, M\}. \tag{1.9}$$

Let

$$\pi_Z(Z|\rho) = \prod_{i=1}^{M} \mathcal{MVN}(Z_{i,1:K}; \mathbf{0}, \Sigma_\rho)$$

denote the joint prior for Z given ρ. The partial order $h = h(Z)$ is constructed according to map (1.8).

The depth control parameter ρ directly governs the variability in the latent paths. A high value of ρ would result in straighter paths, and therefore, less chance for the paths to cross.

It is worth noting that there have been other probabilistic models over random DAGs. Apart from the latent variable models mentioned above, there is also a large class of graph models that can be used to describe DAGs. The simplest one is the $G_{M,p}$-graph order model, which defines a distribution over partial orders by putting an edge from $i \in \mathcal{M}$ to $j \in \mathcal{M}$ independently with some fixed probability, for each each pair of items satisfying $i < j$ (and no edges from larger elements to smaller). Since there are no loops, we can take the transitive closure and get a random partial order.

1.5 Thesis Structure

This dissertation extends the work of [71] and [69] on probabilistic modeling with partial orders. It comprises three self-contained articles on separate but related topics. The overarching theme uniting them is Bayesian inference on partial orders derived from rank-ordered data.

- In Chapter 2: 'Non-Parametric Bayesian Inference for Partial Orders with Ties from Rank Data observed with Mallows Noise', we extend [71]'s latent partial order model (the (K, M, ρ)-partial order model) and introduce a tied structure in the partial orders. We consider noise in the rank-order data by developing the Mallows-ϕ model (with Kendall's tau distance) as an observation model conditioned on partial orders. This is arxived [14] and submitted for publication.

- In Chapter 3: 'Bayesian Inference for Vertex-Series-Parallel Orders', we propose a marginally consistent probability distribution over the vertex-series-parallel partial orders. The novel model exploits the binary decomposition tree representation on VSPs to offer explicit depth control. We extend [71] to propose a bi-directional queue-jumping observation model on observed rank-order data, which will be closer to physical reality in some settings. We further develop an efficient MCMC inference scheme. This paper is paper [39]. It was published in the Conference on Uncertainty in Artificial Intelligence (UAI) 2023.

- In Chapter 4: 'Partial Order Hierarchies', we give a hierarchical partial order model for 'grouped' rank-order data coming from heterogeneous population. This class of model is marginally consistent and nests the Plackett-Luce mixture model as a special case. For model scalability, we propose a VSP approximation for the partial order hierarchy and demonstrate its performance empirically. This paper will be extended with further work prior to publication.

Chapter 2

Non-Parametric Bayesian Inference for Partial Orders with Ties from Rank Data observed with Mallows Noise

Jiang, C. and Nicholls, G. K. (2024). Non-parametric Bayesian inference for partial orders with ties from rank data observed with Mallows noise.
arxiv.org/abs/2408.14661

Non-Parametric Bayesian Inference for Partial Orders with Ties from Rank Data observed with Mallows Noise

Chuxuan (Jessie) Jiang and Geoff K. Nicholls

[a]*Department of Statistics, University of Oxford, Wellington Square, Oxford, OX1 2JD, United Kingdom*

Abstract

Partial orders may be used for modeling and summarising ranking data when the underlying order relations are less strict than a total order. They are a natural choice when the data are lists recording individuals' positions in queues in which queue order is constrained by a social hierarchy, as it may be appropriate to model the social hierarchy as a partial order and the lists as random linear extensions respecting the partial order. In this paper, we set up a new prior model for partial orders incorporating ties by clustering tied actors using a Poisson Dirichlet process. The family of models is projective. We perform Bayesian inference with different choices of noisy observation model. In particular, we propose a Mallow's observation model for our partial orders and give a recursive likelihood evaluation algorithm. We demonstrate our model on the 'Royal Acta' (Bishop) list data where we find the model is favored over well-known alternatives which fit only total orders.

Keywords: Partial Order, DAGs, Ranking, Bayesian Non-Parametric Inference, Mallow's Noise Model, Social Hierarchy

1. Introduction

Ranking methods are widely used in research and industry to provide priority ranks and otherwise summarise data. They play a role in decision support, medical research, chemistry, and many other areas. Order ranking methods can be roughly classified into two categories - total order ranking and partial order ranking. In our application we are interested in recovering hierarchical relations between individuals or *actors*.

A *total order* coincides with one's common understanding of 'ranking' where strict order relations exist between each pair of actors. Correspond-

ingly, most probabilistic ranking models reconstruct an underlying total order. In the first Thurstonian model (Thurstone, 1927), actors are ranked based on random Gaussian attributes. This was followed by paired comparison models, including the Babington-Smith model (Babington-Smith, 1950) and the Bradley-Terry model (Bradley and Terry, 1952). These models assign a probability distribution to total orders given the probabilities of pairwise preference $p_{ij}, i, j \in [n]$ where $[n] = \{1, 2, ..., n\}$. Bradley and Terry (1952) introduced a special parameterisation of this preference such that $p_{ij} = \nu_i/(\nu_i + \nu_j),\ \nu_i, \nu_j \in \mathbb{R}^+$. The Bradley-Terry model is widely studied and there are several extensions and generalisations. Mallows (1957) proposed a tractable variant - the class of Mallows models. Mallows models are built on common ranking distance measures, for example, Spearman's rho (Mallows θ model) and Kendall's tau (Mallows ϕ model). Plackett (1975) and Luce (1959) extended the Bradley-Terry model to compare multiple actors and introduced the Plackett-Luce model. The Mallows and Plackett-Luce models are the most commonly used statistical models for total order ranking in homogeneous populations. Pearce and Erosheva (2024) developed a Rank-Clustered Bradley-Terry-Luce (BTL) model that allows for ties among actors in a BTL model. They used a partition-based spike-and-slab fusion prior. This is relevant to our ties idea as we will explain in section 1.1. However, their model considers ties within total orders while we consider ties in rankings which are only required to be partially ordered.

Mixtures of ranking models have been proposed and may be appropriate for heterogeneous populations. They allow us to fit data with more than one underlying total order in the observation model. Partial order models can be thought of as very general mixture models with one mixture component for every total order that respects the partial order (which can be a huge number, and varies as we change the partial order). Two popular mixture models are the Mallows mixture model and the Plackett-Luce mixture model. Murphy and Martin (2003) applies the Mallows mixture model to ranked list data and includes many useful references to earlier work in the area. Meila and Chen (2012) proposes a Dirichlet process mixture of generalised Mallow Models over discrete incomplete rankings. Lu and Boutilier (2014) learns the Mallow model and its mixture variant from pairwise-preference data. For the Plackett-Luce mixture model, see Mollica and Tardella (2017), where they introduce a Bayesian finite mixture of Plackett-Luce models given partially ranked data. Liu et al. (2019a) learns the Plackett-Luce model and its mixtures from partial order data. Caron et al. (2014) developed a

Dirichlet process mixture in their non-parametric Plackett-Luce model given incomplete (top-k) ranking data.

However, total order based models are not always naturally descriptive for many real-world problems, as one may have pairs of incomparable actors between whom no meaningful underlying order relation exists. There may be a temptation to think of the order on such pairs as simply being uncertain, but the physical reality may be that no order exists. If we want the elements of the model to correspond to elements of reality, relations between actors need to be represented by a partial order.

A *partial order* $h = \{[n], \prec_h\}$ is an order relation that assigns a binary order[1] $\prec_h$ over a set of actor-labels $[n]$. Partial orders on $[n]$ are in one-to-one correspondence with transitively-closed directed acyclic graphs (DAG's) with nodes $V = [n]$ and edges E, such that $(i \prec_h j) \in E$ and $(j \prec_h k) \in E$ gives $(i \prec_h k) \in E$, $i, j, k \in [n]$, $i \neq j \neq k$. Denote by $\mathcal{H}_{[n]}$ the set of all partial orders on actor labels $[n] = \{1, 2, ..., n\}$. An example partial order[2] is shown in Figure 1. A *linear extension* l_h of a partial order h is a permutation of $[n]$ which respects h, so that lesser actors in h never come before greater. See Figure 2 for examples. We denote the set of all linear extensions of partial order h by $\mathcal{L}[h]$.

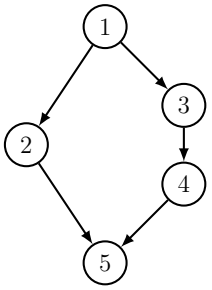


Figure 1: A random partial order h_0 with 5 actors.

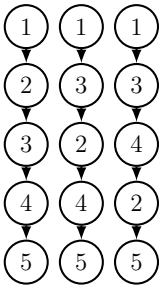


Figure 2: All three linear extensions of partial order h_0 in Figure 1.

Partial orders express fundamental incomparability between actors, rather than a lack of knowledge of the underlying order relation or randomness in

[1]The binary relation $\prec_h$ is both irreflexive (the relation $i \prec_h i$ does not exist) and transitive (if $i \prec_h j$ and $j \prec_h k$, then $i \prec_h k$), where $i, j, k \in [n]$ and $i \neq j \neq k$.

[2]In this article, we visualise partial orders via their transitive reduction - this omits all edges implied by transitivity and is unique.

the realised order. Imagine the partial order h_0 in a hypothetical university setting, where actor 1 is the Head of Department, 2 is the Academic Administrator, 3 and 4 are the Deputy Head of Department and a lecturer respectively, and 5 is a student. Partial order h_0 represents a power hierarchy in which the relation $i \succ_{h_0} j$ expresses the reality that i can direct j and not *vis versa*. The Academic Administrator is not comparable with the Deputy Head of Department as neither manages the other. The lecturer answers to the Deputy Head, and therefore also to the Head, but not the Administrator, and so on.

A total order is just a partial order with no incomparable pairs of actors, so assuming the underlying order is a partial order is weaker than assuming a total order. However, despite being a natural generalisation of rank, the use of partial orders in rank analysis is largely restricted to its use as a tool for summarising order relations obtained by fitting total order models, or as data in which partial orders replace lists, rather than as a component or "parameter" of the model itself. Here we list some exceptions.

In settings in which the partial order is estimated as an explicit "parameter" of the model, both Frequentist and Bayesian methods have been applied. Work on Frequentist estimation includes Beerenwinkel et al. (2007) and Gionis et al. (2006). Beerenwinkel et al. (2007) uses partial orders to map the sequencing relations of different genome mutations. The paper develops a maximum likelihood partial order estimator for conjunctive Bayesian network (CBN) models given sets of binary mutation data occurring in orders constrained by a partial order. Gionis et al. (2006) uses 'bucket' orders[3], a subclass of partial orders, to represent the temporal order of the fossil discovery sites in seriological data analysis. Gionis et al. (2006) and Feng et al. (2008) introduce a pivot optimisation algorithm and a Bucket Gap algorithm respectively to approximate the 'bucket' orders given pairwise/full rankings. Bayesian literature is limited. Nicholls and Muir Watt (2011) estimates a partial order h from list data assuming the data-lists are linear extensions drawn uniformly at random from the set of all linear extensions of h. Nicholls et al. (2023) extends such model to the temporal setting. Unlike the above examples, Nicholls and Muir Watt (2011) and Nicholls et al. (2023) treats

[3]In a 'bucket' order, the actors are arranged in layers - every actor is ordered with respect to actors in other layers, and any pair of actors in the same layer are either unordered or tied.

the partial order h as a random variable.

This paper extends Nicholls and Muir Watt (2011) by proposing new prior and observation models for partial orders. In our setting the unknown partial order h is observed indirectly through possibly incomplete lists of actor labels. The lists respect the order relations imposed by h, so observing the lists tells us something about h. For $i, j \in [n]$, if actor j appears before actor i in the list then we must have either $i \prec_h j$ or $i \| _h j$, i.e. either j beats i or i and j are incomparable in the partial order. Such a list is a linear extension of the partial order h if it includes all n actors and otherwise it is a linear extension of the suborder of actors in that list.

In our Bayesian analysis, we will have a prior over random partial orders and a posterior informed by a set of rank-order lists, each one a permutation of actor labels. We refer to this as *list data*. We will see that one natural observation model relating the lists to the partial order is to treat the lists as uniform random linear extensions of the partial order. This is motivated by considering list data recording the positions of actors in a queue in which higher status individuals come before those of lower status. A status hierarchy is a partial order, so if all queue orders respecting the hierarchy are equally likely, the queue will be a uniform random draw from the linear extensions of the partial order. This model is also justified by writing down a simple stochastic process model for the queue formation process. We return to this point in Section 3.1.

The distinguishing features of our work are of threefold. First, it is Bayesian: we work with probability distributions over partial orders. This makes it straightforward to quantify uncertainty, at least in principle. Second, our models are defined over the set of all partial orders $\mathcal{H}_{[n]}$ of n actors rather than some subset of $\mathcal{H}_{[n]}$ such as bucket orders. Third, our data are linear extensions of suborders, subject to possible recording errors as we discuss later. We further extend the basic latent variable model for partial orders given by Nicholls and Muir Watt (2011) in three ways - we incorporate ties between actors, we allow the dimension of the latent space to be unknown a priori, and we fit observation models in which much more general noise processes are acting. These are needed for model realism.

The paper is structured as follows: in the rest of this section, we motivate our tie and observation error models, and introduce some basic concepts. In Section 2, we write down our partial order model with ties, the (PDP, ρ)-partial order model. In Section 3, we discuss possible choices for the observation model. We consider a Mallows ranking model that allows

error in the observed lists, so the observed lists need not be linear extensions but must be "close" to being linear extensions. In Section 4, we set out Bayesian inference and identify some computational challenges, notably the problem of counting linear extensions in likelihood evaluations. Finally in Section 5, we apply our method to 'Royal Acta', a dataset containing the order of witnesses in legal documents from the twelfth century. In particular, we study the social hierarchy between bishops in two specific time periods. We compare the prior model with ties with other possible ranking models and show our model effectiveness.

1.1. Motivation

We begin by considering ties. Ties between actors are a common phenomenon in reality. For example, Figure 3 shows a hierarchy involving a field marshal (FM), two generals (G1 & G2), an earl (E) and a bishop (B). We suppose their ranking is determined by their titles, and there is no order between military and religious or noble titles, so the generals are equal. This can be interpreted in two ways. First, they must have the same order relations with other actors in the partial order. Secondly, we might additionally require that tied actors appear as an unsplit group in any list, that is, in a random order but sequentially with no other actors between them. More details in Appendix H. In this paper we interpret ties in the first way. The second use of tie models is of interest to us but is future work.

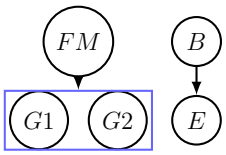


Figure 3: The partial order between a field marshal (FM), two generals ($G1$ & $G2$), a bishop (B) and an earl (E) if their order relation is determined by their titles.

In the example the actor titles can be thought of as attributes or "covariates" which inform the partial order in a simple way. However, discrete attributes of this sort are often unknown to us (as in the examples below) and so we have to learn the tie structure. In our partial order model with ties - the (PDP, ρ)-partial order model - we use a Poisson Dirichlet process to model the clustering process. The partial order describes the order relation between the clusters. When there are unobserved attributes in the setting, this model respects the generative model. From a modelling perspective, it reduces the

dimension of the latent variables needed to represent the partial order, and in turn increases the marginal likelihood. More importantly, the (PDP, ρ)-partial order model assigns higher prior probability on the 'bucket' orders, or partial orders which are "close" to being 'bucket' orders (see Appendix G). This is desirable as bucket orders are common in social hierarchies while allowing adequate prior probability on general partial orders. One nice feature of the families of prior models we write down for partial orders is that they are all *projective* (every marginal of every distribution in the family is also in the family).

We next motivate our models for "noise" in lists. Measurement or recording errors are common in data. For example, the lists we observe may be generated from linear extensions which respected the unknown underlying partial order but were then subject to random exchange of a small number of elements due to recording errors. There is little work fitting partial order models which include a noise component. Beerenwinkel et al. (2007) includes an error component in their mixture model. In Nicholls and Muir Watt (2011) the error model treats errors as a random process of queue-jumping. While this model is physically motivated, it only allows uni-directional queue-jumping in any given list realisation: errors occur when randomly chosen actors are promoted up (or down) the queue. Jiang et al. (2023) extends Nicholls and Muir Watt (2011) to propose a bi-directional queue-jumping model. However, this model is computationally costly and may only be suitable for partial orders with fast linear extension counting, e.g. the vertex-series-parallel partial orders[4] in Jiang et al. (2023).

By contrast there is a large amount of literature reconstructing an underlying total order from noisy list data. For example, in the Mallows model, the observed list will be "close to" but need not be equal to the central ranking under some distance measure. This suggests taking the Mallows model as a noise model, allowing errors in both directions away from a linear extension in a single list realisation. We constrain the central ranking of the Mallows model to be a linear extension of a partial order. The dispersion parameter ϕ governs how "close" the rank is to being a linear extension of the partial order. Depending on the value of ϕ, the generative model can form ranks that are pure noise, or ranks that are exactly linear extensions of partial

[4]The vertex-series-partial orders are a class of partial orders that can be obtained by *series* and *parallel* operations. See Jiang et al. (2023) for details.

order h, or in between. The dispersion parameter controls the level of noise in our model.

1.2. Partial orders

We now provide formal notations on partial orders. So far, we have given two representations of a partial order, as a partially ordered set, and as a transitively closed DAG. For computation it is convenient to code a partial order $h \in \mathcal{H}_{[n]}$[5], $n \geq 1$ as a binary adjacency matrix $h \in \{0,1\}^{n \times n}$. The rows and columns of h correspond to the actors $[n]$. We take $h_{i,i} = 0,\ i \in [n]$ and set $h_{i,j} = 1$ if and only if $i \succ_h j$. Two actors $i, j \in [n]$ are *incomparable* $i \|_h j$, if neither $i \prec_h j$ nor $i \succ_h j$, and in this case $h_{i,j} = h_{j,i} = 0$. A partial order is called a *total order* if no pair of actors is incomparable. An *empty order* is a partial order where no order relation is observed between any actors. An actor $m \in [n]$ is a maximal element of partial order h if there exists no $i \in [n]$ that has $i \succ_h m$ in h. We denote the set of maximal elements of h as $\max(h) = \{m \in [n] : h_{i,m} = 0,\ \forall i \in [n]\}$.

A *linear extension* $l = (l_1, l_2, \ldots, l_n)$ of partial order h is a permutation over $[n]$ given that its order does not violate h, so for $1 \leq a < b \leq n$ we must have either $l_a \succ_h l_b$ or $l_a \| _h l_b$ ("higher status come first"). Let $\mathcal{P}_n$ be the set of all possible permutations of $1, ..., n$. The set of linear extensions of h is then

$$\mathcal{L}[h] = \{l \in \mathcal{P}_n | h_{l_b, l_a} \neq 1, \forall 1 \leq a < b \leq n\}.$$

Figure 2 shows all three possible linear extensions of the partial order h_0 in Figure 1.

A *sub-order* $h[o]$ of a partial order $h \in \mathcal{H}_{[n]}$ that restricts h to a subset $o \subseteq [n],\ o = \{o_1, ..., o_{n'}\}$ for $n' \leq n$ say. All the order relations in h are inherited in $h[o]$ so the matrix representation is $h_{o,o}$ - we simply retain rows and columns $o_1, ..., o_{n'}$ of h. As an example, the suborder of the partial order in Figure 1 restricted to $o = \{2, 3, 5\}$ is shown in Figure 4 with its linear extensions in Figure 5. A *chain* of h is a sub-order that is also a total order. It extracts a strictly ordered sequence in a partial order. The *length* of a chain is the number of nodes in this suborder. The length of the longest chain(s) is called the *depth*, $D(h)$ say, of partial order $h \in \mathcal{H}_{[n]}$, with $1 \leq D(h) \leq n$. For example the depth of h_0 in Figure 1 is four.

[5]We use the same symbols h and $\mathcal{H}_{[n]}$ for a partial order and its space in all three representations; we use H to represent a random variable taking values in $\mathcal{H}_{[n]}$.

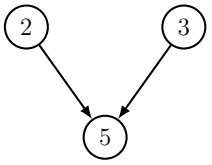


Figure 4: A suborder $h_0[o]$ with actors $o = \{2, 3, 5\}$ of the partial order h_0 in Figure 1.

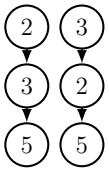


Figure 5: All two linear extensions for the suborder $h_0[o]$ in Figure 4.

The *intersection order* h^{int} of $K \in \mathbb{Z}^+$ permutations $l^{(k)} \in \mathcal{P}_n,\ k = 1, ..., K$ is a partial order $h^{int} \in \mathcal{H}_{[n]}$ such that for $i, j \in [n],\ i \neq j$ we have $i \succ_{h^{int}} j$ if and only if i appears before j in all K permutations. If we think of the permutations as total orders then the intersection order shows all the order relations which are attested in every permutation. If a pair of actors appear in different orders in different permutations then there will be no relation between the actors in the intersection order. If $d_i^{(k)} = \{j \in [n] : l_j^{(k)} = i\}$ is the index of i in list k, then the matrix representing the partial order defined by the intersection of lists satisfies $h_{i,j}^{int} = 1$ if and only if $d_i^{(k)} < d_j^{(k)},\ k = 1, ..., K$.

We define the *dimension* of partial order h to be the minimum number of linear extensions of h whose intersection is h. If the observation model for the data is that the data are uniform random linear extensions of h, and the data is noise-free then the intersection-order is the maximum likelihood estimator of the true partial order (Beerenwinkel et al., 2007). This works because the intersection-order has every order relation allowed by the data, and none which are not, so it is the partial order with the most order relations that admits every list in the data as a linear extension. It has the smallest number of linear extensions, and hence, maximises the likelihood for this observation model.[6] In this setting the intersection order converges in probability to the true partial order as the number of sampled linear extensions goes to infinity.

2. Prior Models for Random Partial Orders

Several probability distributions have been proposed for random partial orders in the combinatorics literature. See Brightwell (1993) for an overview. Of these the most promising for development as a prior for Bayesian statistical work is Winkler (1985)'s latent variable model. Following Nicholls

[6]See Appendix A for proof.

and Muir Watt (2011) we refer to this model as the (K, n)-partial order model. Before defining this model we introduce two prior properties which are desirable in our setting.

First of all in our application, and we expect this to be common, the depth $D(H)$ of the unknown true partial order H is of particular interest. In some cases prior knowledge is given in terms of depth and similarly hypotheses are often statements about partial order depth. We seek a prior $H \sim \pi_{[n]}(\cdot),\ H \in \mathcal{H}_{[n]}$ with the property that H has approximately uniform depth distribution, $D(H) \sim \mathcal{U}\{1, ..., n\}$. This is done so that the prior is non-informative with respect to hypotheses concerning depth. However, the analyst may use this degree of freedom (the prior for ρ, latent dimension and ties) to impose some non-uniform prior weight on the depth.

Secondly, a family of priors $\pi_{[n]}(h)$ is *marginally consistent* or *projective* if every marginal of every distribution in the family is also in the family. For every $n \geq 2$ and every $o, o' \subseteq [n]$ with $o \subset o'$, marginal consistency holds if

$$\pi_o(h) = \sum_{\substack{h' \in \mathcal{H}_{o'} \\ h'[o]=h}} \pi_{o'}(h') \quad \text{for all } h \in \mathcal{H}_o. \tag{1}$$

Physical considerations suggest that our prior beliefs about relations between a subset of actors should be marginally consistent. We are assuming that the relation between any pair of actors in a social hierarchy is not affected by the presence or absence of any other actor.

2.1. Properties of the uniform distribution on Partial Orders

We illustrate a "natural" family of priors which is *not* projective and is additionally strongly informative on depth. It may seem appealing to take as prior the uniform distribution over partial orders.

Definition 1 (The Uniform Distribution over Partial Orders)**.** *For partial order $h \in \mathcal{H}_{[n]}$, the uniform distribution over partial orders $\pi_{u,[n]}(h)$ say, assigning equal prior probability to each partial order in $\mathcal{H}_{[n]}$, is*

$$\pi_{u,[n]}(h) = \frac{1}{|\mathcal{H}_{[n]}|}. \tag{2}$$

First of all the uniform priors $\pi_{u,[n]},\ n \geq 1$ have low sampling depth. As is shown in Kleitman and Rothschild (1975), if $h \sim \pi_{u,[n]}$ then $Pr(D(h) =$

3) $\to 1$ as $n \to \infty$. This surprising result (why 3?) rules out the uniform distribution as a prior for most applied work since, even at small n, it favours partial orders of low depth, and has no parameter we can use to control the distribution over depth.

Secondly, the family of uniform priors $\pi_{u,[n]},\ n \geq 1$ is not projective (Winkler, 1985). For a counter-example, let $h_2 = \{V_2, E_2\} \in \mathcal{H}_{[2]}$ be a partial order on 2 actors with $V_2 = \{1, 2\}$ and edge set E_2. There are 3 choices for E_2: $E_2 = \emptyset$ (the empty partial order), $E_2 = \{(1, 2)\}$ and $E_2 = \{(2, 1)\}$. If there are three actors $h = (V_3, E_3) \in \mathcal{H}_{[3]}$ with $V_3 = \{1, 2, 3\}$ then there are 19 possible edge sets E_3, depicted in the Appendix B. In equation 1, $o = [2]$, $o' = [n] = [3]$ and

$$\text{LHS} = \pi_{u,[2]}(h) = \frac{1}{3}$$

while

$$\begin{aligned}\text{RHS} &= \sum_{\substack{h' \in \mathcal{H}_{[3]} \\ h'[o]=h}} \pi_{u,[3]}(h') \\ &= \frac{|\{h' \in \mathcal{H}_{[3]} : h'[o] = h\}|}{19} \neq \text{LHS},\end{aligned}$$

where $|A|$ denotes the cardinality of set A, and the last step holds because we cannot divide 19 into 3 groups of equal size. This counter-example holds between $n = 2, 3$ and extends straightforwardly to many larger n. The uniform distribution over partial orders is therefore not projective.

2.2. The (K, n)-Partial Order Model

The (K, n)-partial order model is defined in Winkler (1985) and extended in Nicholls and Muir Watt (2011). This generative model adapts a latent matrix approach to represent a random partial order. The (K, n)-partial order model proposes a parameter $K \in \mathbb{Z}^+$ - the fixed column dimension parameter in a random latent matrix $Z \in [0, 1]^{n \times K}$ - to control the depth distribution.

Definition 2 (The (K, n)-Partial Order Model (Winkler, 1985))**.** *Let $Z \in \mathbb{R}^{n \times K}$ be a random latent matrix with fixed column dimension $K \in \mathbb{Z}^+$. The (K, n)-partial order model defines the entries of Z as*

$$Z_{i,k} \overset{i.i.d.}{\sim} \mathcal{U}[0, 1], \forall (i, k) \in [n] \times [K].$$

Write $Z_{i,1:K} \succ Z_{j,1:K}$ *if* $Z_{i,k} > Z_{j,k} \forall k \in \{1,\ldots,K\}, i,j \in [n]$. *We define mapping* $h : [0,1]^{n\times K} \to \mathcal{H}_{[n]}$ *such that*

$$h(Z) = \{h \in \mathcal{H}_{[n]} : i \succ_h j \text{ if and only if } Z_{i,1:K} \succ Z_{j,1:K}\}. \tag{3}$$

(The original paper by Winkler (1985) took $K = 2$*).*

The prior for partial orders under the (K,n)-partial order model is

$$\pi_{[n]}(h) = \int_{[0,1]^{n\times K}} \mathbb{1}_{\{h(Z)=h\}} \pi(Z) dZ. \tag{4}$$

The integral in $\pi(h)$ would be awkward. However, Nicholls and Muir Watt (2011) carries out inference using the latent variable Z. This is in general non-identifiable, but this is not a concern as h is the parameter of interest. Nicholls and Muir Watt (2011) call $Z_{i,1:K} \in [0,1]^K$ the "latent path" (over the column index) for actor i. The condition $Z_{i,1:K} \succ Z_{j,1:K}$ means the paths are non-crossing and $Z_{i,1:K}$ lies entirely above $Z_{j,1:K}$. Figure 6 shows some possible latent paths that generate the partial order h_0 in Figure 1. The mapping $h : [0,1]^{n\times K} \to \mathcal{H}_{[n]}$ is surjective: one partial order can be represented by many latent matrices; however, each latent matrix uniquely determines a partial order.

Any partial order can be realised under this prior if we take K at least $\lfloor n/2 \rfloor$ (Muir Watt, 2015; Nicholls et al., 2023). This can be seen in an equivalent construction. For $j \in [K]$, let $l^{(j)}(Z) = ([n], \succ_{l^{(j)}})$ with

$$\succ_{l^{(j)}} = \{i_1 \succ_{l^{(j)}} i_2 : \ Z_{i_1,j} > Z_{i_2,j}, \ (i_1,i_2) \in [n]\times[n]\}$$

give the total order of the actors according to the Z-values in column j. As discussed in Muir Watt (2015), the rule defining $h = h(Z)$ is identical to taking h to be the intersection order of $l^{(j)}$ over $j = 1,...,K$. Also, all the entries in Z are i.i.d., so $l^{(j)}$ is a uniform random total order of $[n]$ and any set of K total orders $l^{(j)}$ over $j = 1,...,K$ can be generated with non-zero probability. It follows that any partial order can be realised by the process so long as K is at least $\lfloor n/2 \rfloor$: a partial order h on n nodes can always be expressed as the intersection of $\lfloor n/2 \rfloor$ total orders (Hiraguchi, 1951), simply repeating total orders if fewer are needed.

Winkler (1985) shows that the (K,n)-partial order model is projective. This is intuitively obvious: removing an actor is equivalent to removing one

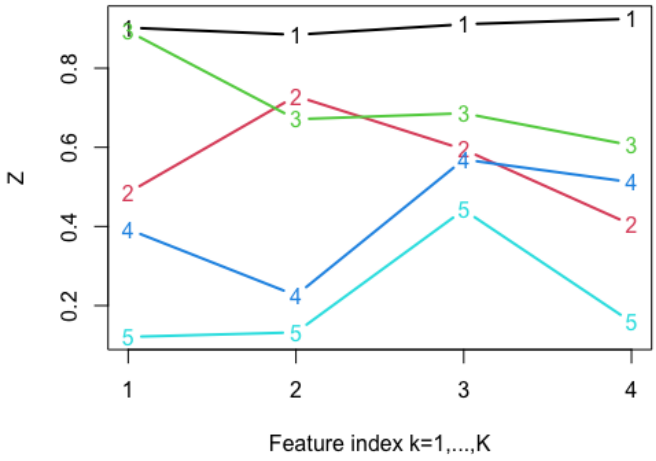


Figure 6: Latent paths of a random Z with $K = 4, n = 5$ defined using the (K, n)-partial order model and yielding the partial order h_0 in Figure 1.

of the paths in Figure 6. Since the paths are independent, and removing one doesn't alter the relations between the paths which remain, the probability to get any particular partial order on $n-1$ labeled actors is the marginal of the probability to get that partial order on any set of n labeled actors containing the original $n-1$ actor labels. This kind of marginal consistency is typical for latent variable models for networks and graphs.

We now state and prove this result, partly to show how the notation works in this simple case. The proof for the more interesting case of the prior with ties below follows a similar path.

Proposition 1. *The (K, n)-partial order model distributions are projective.*

Proof: See Appendix C.

Although projective, the (K, n)-partial order model offers limited control over depth. In this model, K controls the typical depth $D(h(Z))$ of the partial order. When $K = 1$, $\Pr(D(h(Z)) = n) = 1$ since there are no paths to cross. At small $K \geq 1$ the probability that $Z_{i,1:K} \succ Z_{j,1:K}$ is relatively high, as paths are short, so an order between i and j is more likely and partial orders have greater depth. As K increases, random (K, n)-partial orders tend

to have lower depth, and $\lim_{K\to\infty} \Pr(D(h(Z)) > 1) = 0$. However, K needs to be at least $\lfloor n/2 \rfloor$ if we want $h(Z)$ to be able to realise any partial order in $\mathcal{H}_{[n]}$. At these K-values, $D(h(Z)) \ll n$, typically.

2.3. The (K, n, ρ)-Partial Order Model

The issue of a depth distribution concentrated on small values is addressed in Nicholls and Muir Watt (2011) and Muir Watt (2015) where K is fixed and a depth parameter $\rho \in [0, 1)$ is introduced to control depth. This (K, n, ρ)-model for random partial orders is an extension to the (K, n)-model. It replaces the uniform random entries in Z with a distribution which correlates entries within rows (but retains independence between rows). The i-th row of Z, $Z_{i,1:K} = (Z_{i,1}, ..., Z_{i,K})$, gives K attribute-values for actor $i \in [n]$. If the entries in $Z_{i,1:K}$ are more strongly correlated, then the paths $Z_{i,1:K}$ are relatively "flatter", and the probability two paths do not cross is higher, so finally $h(Z)$ tends to have relatively higher depth.

Definition 3 (The (K, n, ρ)-Partial Order Model (Nicholls and Muir Watt, 2011))**.** *Let Σ_ρ be a $K \times K$ correlation matrix with $\Sigma_{\rho,i,i} = 1$ and $\Sigma_{\rho,i,j} = \rho$, $i \neq j$, $i, j \in [K]$. In the (K, n, ρ)-model*

$$Z_{i,1:K}|\rho \overset{i.i.d.}{\sim} \mathcal{MVN}(\mathbf{0}, \Sigma_\rho),\ i \in \{1, \ldots, n\}. \tag{5}$$

Let

$$\pi_Z(Z|\rho) = \prod_{i=1}^{n} \mathcal{MVN}(Z_{i,1:K}; \mathbf{0}, \Sigma_\rho)$$

denote the joint prior for Z given ρ. The partial order $h = h(Z)$ is constructed according to map (3).

There is considerable freedom in choosing the Z-distribution, but correlating entries within rows while maintaining independence across rows seems key to controlling depth whilst ensuring distributions are projective. In Figure 7, we show another set of latent paths for the partial order h_0 in Figure 1 simulated under the (K, n, ρ)-partial order model with depth parameter $\rho = 0.9$.

We would like the marginal prior for H determined by $\rho \sim \pi_\rho(\cdot)$, $Z \sim \pi_Z(\cdot|\rho)$ and $H = h(Z)$

$$\pi_{[n]}(h) = \int_0^1 \int_{Z:h(Z)=h} \pi_Z(Z|\rho)\pi_\rho(\rho) dZ d\rho, \tag{6}$$

to have approximately uniform marginal depth distribution. Simulations reported in Nicholls and Muir Watt (2011) suggest that $\rho \sim \text{Beta}(1, 1/6)$ gives a reasonably flat distribution for $D(H)$ for n in the range of interest to us.

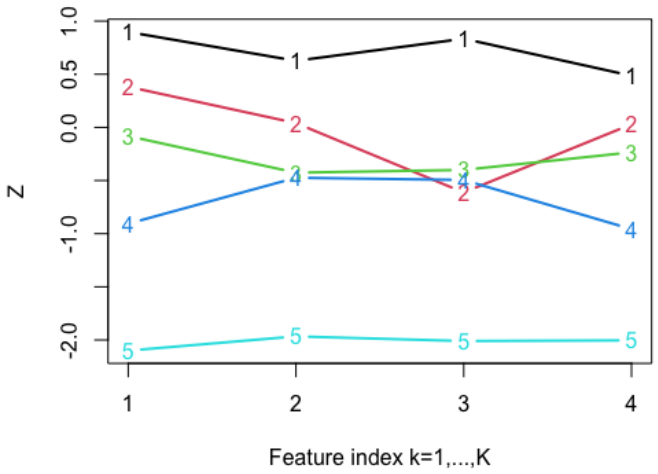


Figure 7: Latent paths of a random Z defined using the (K, n, ρ)-partial order model for the partial order h_0 where $K = 4, n = 5$ and $\rho = 0.9$.

Like the (K, n)-model, the (K, n, ρ)-model can be expressed as an intersection of K linear extensions, which are the rank-vectors of the Z-columns. Any set of rank vectors can be realised, but they are now correlated. The choice of K remains an important model hyperparameter. We discuss below how K can be treated as a parameter and estimated alongside the other parameters. Because the rows of Z are conditionally independent given ρ in the (K, n, ρ)-model, we can show that the family of priors $\pi_{[n]}(h|\rho)$ is projective for every fixed ρ, by the same reasoning given in Proposition 1 for the (K, n)-model.

Proposition 2. *The (K, n, ρ)-partial order distributions are projective.*

Proof. In the notation of Equation 1, if $o \subset o' \subseteq [n]$, $h' \sim \pi_{o'}(\cdot|\rho)$ and $h \sim \pi_o(\cdot|\rho)$ then the conditionals $H'[o]|\rho \sim H|\rho$ are same for every ρ (the proof is essentially unchanged from Proposition 1), so the marginals $H'[o] \sim H$ are also equal in distribution. □

2.4. The (PDP, ρ)-Partial Order Model

Depth control in the (K, n, ρ)-model can be inadequate in certain scenarios. In addition, its continuous latent matrix representation makes it impossible to model equality between actors. This paper extends the (K, n, ρ)-model to consider ties and to provide further control over depth distribution. We call this extended model *the (PDP, ρ)-partial order model*.

2.4.1. Partial order with Ties

Latent variable models are powerful and expressive tools for modelling partial orders. However, in the model we have defined the rows of Z to be almost surely distinct. In some settings we may have prior information that some actors are equivalent in the sense that some groups of actors have the same order relations with other groups of actors. We don't know the groups, or how they are related, but we expect there are groups. As illustrated in the field marshal and generals example in Section 1, this scenario leads to a prior distribution over partial orders where we allow ties between some actors. We define a tied relation as follows.

Definition 4 (Tie Relations in partial orders)**.** *In a partial order with ties h^*, if two actors $i, j \in [n]$ are tied ($i \sim_{h^*} j$), then for any $v \in [n], v \neq i \neq j$,*

1. *$i \succ_{h^*} v$ if and only if $j \succ_{h^*} v$;*
2. *$i \prec_{h^*} v$ if and only if $j \prec_{h^*} v$;*
3. *$i \| _{h^*} v$ if and only if $j \| _{h^*} v$.*

In Section 1.2, we represent a partial order h using a binary adjacency matrix $h \in \{0,1\}^{n \times n}$, such that $h_{i,j} = 1$ if and only if actor $i \succ_h j$. We extend this notation so that if there is a tie between actors i and j in h^*, then $h^*_{i,j} = h^*_{j,i} = 1$. We use $\mathcal{H}^*_{[n]}$ to represent the class of *partial orders with ties*. The original partial order class $\mathcal{H}_{[n]} \subseteq \mathcal{H}^*_{[n]}$.

Definition 5 (Binary Adjacency Representation for Partial Orders with Ties)**.** *A partial order with ties $h^* \in \mathcal{H}^*_{[n]}$ can be represented by a binary adjacency matrix $h^* \in \{0,1\}^{n \times n}$, where for $i, j \in [n], i \neq j$,*

1. *$i \succ_{h^*} j$ if and only if $H_{i,j} = 1$ and $H_{j,i} = 0$;*
2. *$i \sim_{h^*} j$ if and only if $H_{i,j} = H_{j,i} = 1$; and*
3. *$i \|_{h^*} j$ if and only if $H_{i,j} = H_{j,i} = 0$.*

As an example, if we take the partial order h_0 in Figure 1 and tie actors $3 \sim_{h_0^*} 4$ the resulting tied partial order is shown in Figure 8. The new partial order with ties h_0^* is shallower (the depth is 3) and has more linear extensions as is shown in Figure 9.

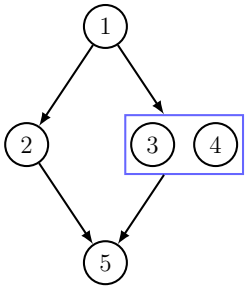


Figure 8: A tied partial order h_0^* of partial order h_0 with cluster $S = (1, 2, (3, 4), 5)$.

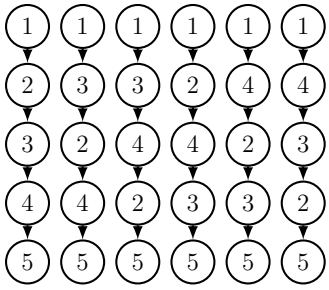


Figure 9: All six linear extensions for the tied partial order h_0^* in Figure 8.

Notice the difference between "equal" $\sim_h$ and "unordered" $\|_h$. If we remove the tie in Figure 8 to get a new partial order in which 3 and 4 are unordered (but all other relations are unchanged) then the linear extensions would be unchanged from those in Figure 9. Since the linear extensions are the data, the new and old partial orders would not be identifiable. So here there is no difference. However, we cannot simply take all unordered pairs in a partial order and replace them with ties, as tied actors must have identical relations to other actors while unordered actors need not. On the other hand, as we saw, if we replace tied relations with unordered relations (maintaining relations to other actors) then the resulting partial order won't be identifiable from the tied order. From the perspective of the observation model, ties don't really change much. However, introducing ties changes the prior: we should think of the ties model as a way to put more weight in our prior on partial orders in which relations are between groups rather than between individuals.

We now give a generative model for random partial orders with ties, which we call the (PDP, K, ρ)-partial order model with fixed latent column dimension K. This model involves partitioning the actors into tied clusters and assigning probability to a random partial order with the clusters as nodes. In other respects the same latent variable setup is used. This ensures marginal consistency and effective control over depth. Actor-partitioning is achieved using a Poisson-Dirichlet process mixture model, a generalisation of

the Dirichlet process to two parameters, over realisations of the rows of Z. The Poisson-Dirichlet process groups the tied actors into clusters with equal Z-paths. Define partition $\mathbf{S} = (S_1, S_2, \ldots, S_C) \in \Xi_{[n]}$ where $C = |\mathbf{S}|$ is the number of clusters and $\Xi_{[n]}$ is the set of all partitions of $[n]$. Let $\eta = (\eta_a, \eta_b)$ be the discount and strength parameters respectively. The process assigns a prior distribution over S according to

$$P_{\eta,[n]}(\mathbf{S}) = \frac{\Gamma(\eta_b)}{\Gamma(\eta_b + n)} \frac{\eta_a^C \Gamma(\eta_b/\eta_a + C)}{\Gamma(\eta_b/\eta_a)} \prod_{c=1}^{C} \frac{\Gamma(n_c - \eta_a)}{\Gamma(1 - \eta_a)}, \tag{7}$$

where $n_c = |S_c|$ denotes the number of actors in cluster S_c. Both $\eta_a \in [0, 1)$ and $\eta_b > -\eta_a$ are constants.

Definition 6 (The (PDP, K, ρ)-Partial Order Model)**.** *Let* $\mathbf{S} \sim PDP(\eta_a, \eta_b) \in \Xi_{[n]}$ *and* $|\mathbf{S}| = C$*. Define latent matrix* $Z^* \in \mathbb{R}^{C \times K}$*, with distribution*

$$Z^*_{c,1:K} | \rho \overset{i.i.d.}{\sim} \mathcal{MVN}(\mathbf{0}, \Sigma_\rho), c \in [C], \tag{8}$$

and Σ_ρ *as defined in Definition 3. For* $i \in [n]$*, let* $c(i; S) = \{c \in [C] : i \in S_c\}$ *be the cluster containing* i*. Define map* $Z : \mathbb{R}^{C \times K} \to \mathbb{R}^{n \times K}$*,*

$$Z(S, Z^*) = (Z^*_{c(i;S),1:K})_{i=1}^n. \tag{9}$$

We extend the map h *to handle ties as follows. Let*

$$G(S, Z^*) = \{i \succ_{h^*} j \text{ iff } Z(S, Z^*)_{i,k} > Z(S, Z^*)_{j,k}, \ \forall\ k \in [K]\}_{i,j \in [n]}$$

and

$$E(S, Z^*) = \{i \sim_{h^*} j \text{ iff } Z(S, Z^*)_{i,k} = Z(S, Z^*)_{j,k}, \ \forall\ k \in [K]\}_{i,j \in [n]}$$

Then

$$h^*(S, Z^*) = ([n], G(S, Z^*) \cup E(S, Z^*)). \tag{10}$$

This reduces to the previous definition in the case of a partial order without ties.

Each row of Z^* represents a cluster. Actors from the same cluster have the same attributes in the latent matrix Z. The (PDP, K, ρ)-partial order model therefore gives a probability distribution over the class of partial order with ties $\mathcal{H}^*_{[n]}$.

2.4.2. Estimating the Latent Matrix Dimension K

We now introduce a second generalisation of the model. This was motivated by our desire to get good control of the depth distribution in the context of a model with ties. We have two parameters influencing depth: K and ρ. The (K, n, ρ) model took K fixed and took a prior distribution for ρ which gave an approximately uniform distribution for $D(H)$. We found, in our (PDP, K, ρ)-partial order model, that we could not get a good control of depth by varying ρ, η_a and η_b alone. We now allow K to vary. We take a prior $K \sim \pi_K(\cdot)$. Small K promotes deeper partial orders.

Definition 7 (The (PDP, ρ)-Partial Order Model)**.** *The full generative model for a random partial order with ties is as follows:* $K \sim \pi_K(\cdot)$, $\rho \sim \pi_\rho(\cdot)$, $S \sim P_{\eta,[n]}(\cdot)$ *(with hyperparameter* $\eta = (\eta_a, \eta_b)$*),*

$$Z^*|\rho, K, S \sim \prod_{c=1}^{C} \mathcal{MVN}(Z^*_{c,1:K}; \mathbf{0}, \mathbf{\Sigma}_\rho), \tag{11}$$

with $\mathbf{0} \in \mathbb{R}^K$ *and* $\mathbf{\Sigma}_\rho$ *as defined above (ie, variance one, covariance* ρ*) and* $H^* = h^*(S, Z^*)$*.*

We now write down the marginal distribution for H^* generated by the process in Definition 7. For $h^* \in \mathcal{H}^*_{[n]}$ and $S \in \Xi_{[n]}$ let

$$Z^*[h^*; S] = \{Z^* \in \mathbb{R}^{C\times K} : h^*(S, Z^*) = h^*\} \tag{12}$$

be the set of Z^* matrices which give h^* for a given actor partition S. The tie-model for random partial orders determines a prior distribution $\pi_{[n]}(h^*)$ over $h^* \in \mathcal{H}^*_{[n]}$. The conditional is

$$\pi_{[n]}(h^*|\rho, K) = \sum_{S\in\Xi_n} \left[\int_{Z^*[h^*;S]} \pi(Z^*|\rho, K, S)dZ^*\right] P_{\eta,[n]}(S), \tag{13}$$

so that marginally,

$$\pi_{[n]}(h^*) = \sum_{k=1}^{\infty} \left[\int_0^1 \pi_H(h^*|\rho, K = k)\pi_\rho(\rho)d\rho.\right] \pi_K(k). \tag{14}$$

The prior $\pi_{[n]}(h^*)$ is not tractable but can be explored by simulating the generative model. When we come to fit the model we work in the latent representation (Z^*, S).

We choose the priors $\pi_K(k) = \text{Geo}(k; 1, \eta_K)$ (with $k \geq 1$) and $\pi(\rho) = \text{Beta}(\rho; 1, \eta_\rho)$. The hyperparameters η_K and η_ρ are chosen by experiment so that the prior predictive distribution for the depth, $D(h^*(S, Z^*))$ determined by taking $\rho \sim \pi_\rho(\cdot)$, $K \sim \pi_K(\cdot)$, $S \sim P_{\eta,[n]}(\cdot)$ and $Z^* \sim \pi(\cdot|\rho, K, S)$ is reasonably flat. The prior distribution for partial order depths for the case $n = 15$ ($\eta_a = 0.7, \eta_b = 3, \eta_\rho = 1/6, \eta_K = 0.0625$) is shown in Figure 10. It is not perfectly uniform but this is not a concern. We simply need to avoid exponentially large weightings which can arise in this setting so that the prior at least "allows" all depths. Any further re-weighting can be made in the analysis using for example Bayes factors. In other settings the priors may need to be adjusted.

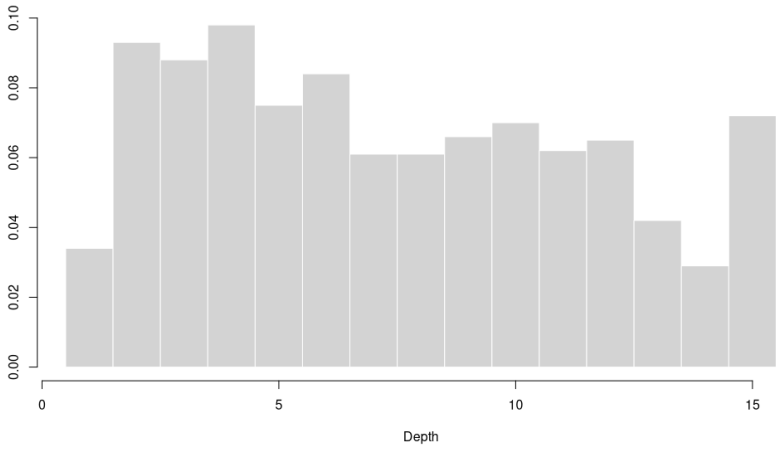


Figure 10: The prior distribution for partial order depths with $n = 15$ actors (with hyperparameters $\eta_a = 0.7, \eta_b = 3, \eta_\rho = 1/6, \eta_K = 0.0625$).

The (PDP, ρ)-partial order model is projective. This follows because the Poisson-Dirichlet process generating Z is projective. The proof is very similar to that for the (K, n)-model in Proposition 1.

Proposition 3. *For $h^* \in \mathcal{H}^*_{[n]}$ let $\pi_{[n]}(h^*) = \Pr(H^* = h^*)$ be the distribution of the random partial order with ties generated by the process in Definition 7 and given in Equation 14. The family of distributions $\pi_{[n]}, n \geq 1$ is projective.*

Proof: see Appendix D.

3. Observation Models for list data

Suppose our data contains N lists, $\mathbf{y} = (y_1, \ldots, y_N)$. Let

$$o_i = (o_{i,1}, o_{i,2}, ..., o_{i,n_i}),\ i = 1, ..., N,$$

with $o_{i,j} < o_{i,j+1}, j = 1, ..., n_i - 1$ be a subset $o_i \subseteq [n]$ of $n_i \geq 2$ actor labels recording the actors present in list i. Let $\mathcal{P}_{o_i}$ be the set of all permutations of o_i. The lists $y_i \in \mathcal{P}_{o_i}$ record the order in which the actors appear in the list. Note the setup: the list y_i is not formed by sampling a distribution on permutations of all n actors and then reducing it to a shorter list containing only those in o_i; rather the list is formed from the start using a distribution over permutations of the actors in o_i only. We condition on the values of $o_i,\ i = 1, ..., N$ throughout. In the context of preference orders, assessors are presented with a choice set o_i and return an ordered preference list y_i which ranks all the elements in the choice set. We will be interested in the distribution of the random lists y_i given the label content o_i. We think of the lists y_i as total orders given by the sequence in which the labels appear from highest to lowest in the order. In the i-th data-list, $y_{i,1}$ came first and y_{i,n_i} last and this will be evidence that actor $y_{i,1}$ is in some sense more important than actor y_{i,n_i}.

3.1. Noise-Free Model

In our observation model for noise-free lists we take y_i to be a linear extension sampled uniformly at random from the set of all linear extensions of the suborder $H[o_i]$ of unknown true partial order H, independently for $i = 1, ..., N$. In this section we write down the likelihood and motivate this choice.

Suppose $H = h$ for some $h \in \mathcal{H}_{[n]}$. Let $o_j \subseteq [n]$ be some generic subset of actor-labels, $j = 1, \ldots, N$. The set of all linear extensions of the sub-order $h[o_j]$ is $\mathcal{L}[h[o_j]]$. Let $y_j \in \mathcal{L}[h[o_j]]$ be some generic noise free 'data', an ordered list of the actor-labels in o_j. The generative model in the noise-free case is

$$p^{(P)}(y_j | h[o_j]) = \frac{\mathbb{1}_{y_j \in \mathcal{L}[h[o_j]]}}{|\mathcal{L}[h[o_j]]|}. \tag{15}$$

This model is motivated by thinking of the list as recording the place of actors $(o_{j,1}, ..., o_{j,n_j})$ in a queue. The queue forms, and then neighboring pairs of actors in the list swap places at random so long as the exchange does not violate any order relation in $h[o_j]$, until the moment the queue is recorded.

This process defines an irreducible Markov chain of random lists converging in distribution to the uniform distribution over linear extensions of $h[o_j]$.

The following result is well known, for example, it appears in Karzanov and Khachiyan (1991) who shows a closely related chain is rapidly mixing.

Proposition 4. *Consider a Markov chain $\{X_t\}_{t\geq 1}$, with state space $X_t \in \mathcal{L}[h[o_j]]$ for some fixed $j \in \{1,\ldots,N\}$. Suppose that at step t we have $X_t = x$. An entry $i \sim U\{1,...,n_j\}$ is chosen uniformly at random. If $i = n_j$ then we reject and set $X_{t+1} = x$ and otherwise $x' = (x_1,...,x_{i-1},x_{i+1},x_i,x_{i+2},...,x_{n_j})$. If $x' \in \mathcal{L}[h[o_j]]$ then we set $X_{t+1} = x'$ and otherwise $X_{t+1} = x$. The process converges in distribution to the uniform distribution over linear extensions of $h[o_j]$, that is, for $l \in \mathcal{L}[h[o_j]]$,*

$$\Pr(X_n = l) \overset{t\to\infty}{\to} \frac{1}{|\mathcal{L}[h[o_j]]|}\mathbb{1}_{l\in\mathcal{L}[h[o_j]]}.$$

Proof: see Appendix E.

3.2. Error Models

Obtaining error-free data is obviously ideal. However measurement or recording errors are often present. In order to incorporate noise in the observation model, Nicholls and Muir Watt (2011) propose a simple queue-jumping error model that allows an actor to jump up the queue with probability $p \in [0,1]$, ignoring any order constraints.

3.2.1. Queue-Jumping Error Model

Let $L(h) = |\mathcal{L}[h]|$ be the number of linear extensions of h and let $L_i(h) = |\{l \in \mathcal{L}[h] : l_1 = i\}|$ give the number of linear extensions headed by actor i. The queue-jumping (upwards) model (Nicholls and Muir Watt (2011)) defines

$$p^{(Q)}(y_j|h,p) = \prod_{i=1}^{n_j-1}\left(\frac{p}{n_j-i+1} + (1-p)\frac{L_{y_i}(h[y_{j,i:n_j}])}{L(h[y_{j,i:n_j}])}\right). \tag{16}$$

We can interpret this as modelling the process by which a list $Y = (Y_1,\ldots,Y_n) \in \mathcal{P}_{[n]}$ is formed. The process is as follows:

1. Set $i = 1$, $s = [n]$ and $h' = h$.
2. Let $q = (L_j(h')/L(h'))_{j\in s}$ be a probability vector weighted by the numbers of linear extensions headed by each actor.

2 With probability p sample $Y_i \sim \mathcal{U}(s)$ (choose one of the remaining actors at random) and otherwise sample $Y_i \sim \text{multinom}(q)$.
3. Set $s \leftarrow s \setminus Y_i$.
4. If s is empty return $Y_1, ..., Y_n$ and otherwise set $i \leftarrow i + 1$, $h' \leftarrow h[s]$ and go to step 2.

The output $(Y_1, ..., Y_n) \sim p^{(Q)}(\cdot|h, p)$ is a random list of n elements distributed according to the queue-jumping model. This follows because the probabilities to choose entries in Y at each step are just the factors in $p^{(Q)}$ in Equation 16. If we set $p = 0$, then we get a telescoping product and $p^{(Q)}(l|h, p = 0) = 1/|\mathcal{L}(h)|$ for $l \in \mathcal{L}[h]$. so the model reduces to the error-free observation model in Equation 15. We can turn the model around and build the list from the bottom, allowing "queue jumping" downwards. Jiang et al. (2023) proposes a 'bi-directional' queue-jumping model that makes displacements in both directions plausible in the similar queue-like setting. If we denote the computational complexity of counting the number of linear extension of a partial order with n nodes as O_n (already at least exponential growth with n), the computational complexity of evaluating the 'bi-directional' queue-jumping likelihood is at least $O(O_n \times 2^n)$. This makes it formidable to evaluate the 'bi-directional' queue-jumping likelihood on general partial orders, as opposed to its subclass - the vertex-series-parallel orders that admit fast linear extension counting - as considered in Jiang et al. (2023).

In order to build a more tractable 'bi-directional' error model, we introduce Mallows and Plackett-Luce noise-model variants, focusing on the Mallows case which we favor. See Appendix I for some remarks on Plackett-Luce and some pros and cons.

3.2.2. The Mallows Model

The Mallows model (Mallows, 1957) assigns probability to a ranking y based on its distance from a 'central ranking' or 'reference ranking' l. Consider the "unknown true" data $l \in \mathcal{P}_{[n]}$ as a linear extension of a generic partial order h where $h \in \mathcal{H}_{[n]}$. We observe $y \in \mathcal{P}_{[n]}$ with "Mallows noise" but "centred" on l (if $y \in \mathcal{P}_o$ then $l \in \mathcal{L}_{h[o]}$ and we are observing a suborder $h[o]$ for some $o \subseteq [n]$, but the setup is the same with $h \to h[o]$ *etc*). In the Mallows-ϕ model, the probability $p(y|l, \theta)$ is given in terms of a symmetric divergence $d(l, y) \geq 0$. Let $\sigma(l, j) = \{k \in [n] : l_k = j\}$ and let the Kendall-tau

distance be $d(y, l)$ so that

$$d(y,l) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{I}_{\sigma(l[y_{i:n}],y_i) > \sigma(l[y_{i:n}],y_j)}.$$

Denote the *dispersion parameter* as $\theta \in (0, \infty)$. The Mallows ϕ-model gives

$$p^{(M)}(y|l,\theta) = \frac{\exp(-\theta d(l,y))}{\Psi_n(\theta)}, \quad y \in \mathcal{P}_{[n]}, \tag{17}$$

where $\Psi_n(\theta)$ is a normalising constant available in closed for as $\Psi_n(\theta) = \sum_{z\in\mathcal{P}_{[n]}} e^{-\theta d(l,z)} = \prod_{i=1}^{n} \psi_i$ with $\psi_i = \sum_{j=1}^{i} e^{-(i-1)\theta}$ under the Kendall-tau distance. The Mallows-ϕ is a sequential choice model as

$$p^{(M)}(y|l,\theta) = \prod_{i=1}^{n-1} \frac{\exp(-\theta \sum_{j=i+1}^{n} \mathbb{I}_{\sigma(l[y_{i:n}],y_i) > \sigma(l[y_{i:n}],y_j)})}{\psi_{n-i+1}(\theta)} \tag{18}$$

$$= \prod_{i=1}^{n-1} q^{(M)}(y_i|l[y_{i:n}],\theta), \tag{19}$$

where $q^{(M)}(y_i|l[y_{i:n}],\theta)$ is the probability y_i is selected next from the remaining choices.

Conditioning on the partial order h, the likelihood is the marginal probability to observe y, if l is a noise-free draw from the linear extensions of h, so we define

$$p^{(M)}(y|h,\theta) \equiv \sum_{l\in\mathcal{L}[h]} p^{(M)}(y|l,\theta) p^{(P)}(l|h) \tag{20}$$

using (15) and the symmetry of $d(l, y)$ in its arguments

$$= \frac{1}{|\mathcal{L}[h]|} \sum_{l\in\mathcal{L}[h]} p^{(M)}(l|y,\theta).$$

We now give a recursion for evaluating the sum. Since $\mathcal{L}[h] = \cup_{k\in\max(h)} \mathcal{L}_k[h]$,

$$\sum_{l\in\mathcal{L}[h]} p^{(M)}(l|y,\theta) = \sum_{k\in\max(h)} \sum_{l\in\mathcal{L}_k[h_{-k}]} p^{(M)}(l|y,\theta),$$

and now l starts with k, so use Equation 19 to split off the first factor

$$= \sum_{k \in \max(h)} q^{(M)}(k|y_{-k}, \theta) \sum_{l' \in \mathcal{L}[h_{-k}]} p^{(M)}(l'|y_{-k}, \theta), \qquad (21)$$

where $h_{-k} = ([n] \setminus \{k\}, \succ_h)$ and similarly for y_{-k}. Let

$$f(y, h, \theta) = \sum_{l \in \mathcal{L}[h]} p^{(M)}(l|y, \theta).$$

The recursion in Equation 21 is

$$f(y, h, \theta) = \sum_{k \in \max(h)} q^{(M)}(k|y_{-k}, \theta) f(y_{-k}, h_{-k}, \theta).$$

Algorithm 1 adapts the algorithm of Knuth and Szwarcfiter (1974) for recursive counting of linear extensions to evaluating this weighted sum. We evaluate the count of linear extensions $|\mathcal{L}[h]|$ in the same pass. Once we are done the likelihood is $p^{(M)}(y|l, \theta) = f(y, h, \theta)/|\mathcal{L}[h]|$.

Algorithm 1: Evaluating $f(y, h, \theta) = \sum_{l \in \mathcal{L}[h} p(l|y, \theta)$ under the Mallows observation model

input: h, θ, y
Function `f(`y,h,θ`)`**:**
  $n \leftarrow$ number of elements in h
  if $D(h) = n$ *(h is a total order)* **then**
    $l \leftarrow$ list ordered as elements of h
    $g \leftarrow p^{(M)}(l|y, \theta)$ (see Equation 18)
    return $(f, \text{count}) = (g, 1)$
  if $D(h) = 1$ *(h is an empty order)* **then**
    return $(f, \text{count}) = (1, n!)$
  Set count $\leftarrow 0$, $f \leftarrow 0$
  foreach $k \in \max(h)$ **do**
    $g_k \leftarrow q^{(M)}(k|y, \theta)$
    $(f_k, c_k) \leftarrow$ `f`(y_{-k}, h_{-k}, θ)
    $f \leftarrow f + g_k \times f_k$ (Equation 21)
    count $\leftarrow$ count $+ c_k$
  return (f, count)

Algorithm 1 gives a recursion in which a function $f(y, h, \theta)$ returns the sum $\sum_{l \in \mathcal{L}[h]} p^{(M)}(l|y, \theta)$ over orders of m elements, by evaluating the sum over

k in the last line of Equation 21 and calling itself to evaluate $f(y_{-k}, h_{-k}, \theta) = \sum_{l' \in \mathcal{L}[h_{-k}]} p^{(M)}(l'|y_{-k}, \theta)$ on orders of length $n-1$. The recursion stops if f is called with h a total order (one extension, so return (17)) or if h is the empty order (then $\mathcal{L}[h] = \mathcal{P}_{[n]}$ so the sum is one as $p^{(M)}(l|y, \theta)$ is normalised over $l \in \mathcal{P}_{[n]}$) so the recursion stops at $n \geq 2$ at lowest.

Note that this paper uses the Kendall-tau distance as the distance metric for the Mallows model. It is chosen not only for its tractable normalising constant, in addition, Diaconis (1988) suggests Kendall-tau is the metric of choice considering its interpretability, tractability, invariance, sensitivity and available theory. However, from a modelling perspective, other distance metric choices may be more natural for a given data set. For example, Vitelli et al. (2018) prefers the footrule distance. Although it is physically well motivated, we didn't explore the footrule distance due to the challenge of its intractable normalising constant in our context.

In our implementation, we impose a Gamma prior over the dispersion parameter θ. We discuss in more detail in Section 5.

4. Bayesian Inference

Given a prior model for partial orders and an observation model for ranking lists, Bayesian inference is straightforward in principle. We demonstrate Bayesian inference for the (PDP, ρ)-partial order model with queue-jumping error model $((PDP, \rho)\backslash Q)$ and the Mallows error model $((PDP, \rho)\backslash M)$. Here we write down the posterior distributions for these two models respectively.

The posterior distribution of the $(PDP, \rho)\backslash Q$ model is

$$\pi(h^*|\mathbf{y}) \propto \sum_{S \in \Xi_n} \sum_{k=1}^{\infty} \int_{Z^*[h^*;S]} \int_{p=0}^{1} \int_{\rho=0}^{1} \pi(Z^*, \rho, k, S, p|\mathbf{y}) d\rho dp dZ^*, \tag{22}$$

where

$$\pi(Z^*, \rho, k, S, p|\mathbf{y}) \propto p^{(Q)}(\mathbf{y}|h(Z), p)\pi(Z^*|\rho, k, S)\pi_\rho(\rho)\pi_K(k)\pi_P(p)P_{\eta,[n]}(S).$$

The posterior distribution of the $(PDP, \rho)\backslash M$ model is

$$\pi(h^*|\mathbf{y}) \propto \sum_{S \in \Xi_n} \sum_{k=1}^{\infty} \int_{Z^*[h^*;S]} \int_{\theta=0}^{\infty} \int_{\rho=0}^{1} \pi(Z^*, \rho, k, S, \theta|\mathbf{y}) d\rho d\theta dZ, \tag{23}$$

where

$$\pi(Z, \rho, k, S, \theta|\mathbf{y}) \propto p^{(M)}(\mathbf{y}|\theta, h(Z))\pi_\theta(\theta)\pi(Z^*|\rho, k, S)\pi_\rho(\rho)\pi_K(k)P_{\eta,[n]}(S).$$

We employ Gibbs Metropolis-Hastings for the inference. Varying K changes the dimension of Z^* and so we use reversible jumping to target the conditional for K. We modify the classic Gibbs sampling algorithm of Neal (2000) for the partition S with a Poisson Dirichlet process as prior and implement simple random-walk Metropolis-Hasting for most of the remaining the parameters. We assume priors as $k \sim Geo(\eta_k), \rho \sim Beta(1, \eta_\rho)$ and $\theta \sim Gamma(\eta_\theta, 1)$. The detailed algorithm is included in Appendix J.

4.1. Asymptotic posterior distributions in the noise free case

It is of interest to consider the behavior of these posteriors as $N \to \infty$, the large data limit. Suppose the true partial order is $h^\dagger$. Since the space of partial orders is discrete, the natural question is whether $\pi(h^\dagger|y) \to 1$ as $N \to \infty$ so the posterior model is a consistent estimator for the unknown true partial order when the observation model is correct (so ignoring model misspecification).

In the first proposition $O = (o_k)_{k=1}^{2^n-1}$ is the set of all non-empty subsets of $[n]$. We are interested in the setting where we can only make observations on suborders $h^\dagger[o_i]$ of $h^\dagger$ for $i \in I$ where $I \subseteq [2^n - 1]$ is the set of subset-indices for subsets of actors appearing together in observable groups. We observe N linear extensions $\mathbf{y}_{1:N}$ in total. We set up the asymptotics so that for $i \in I$ there are N_i lists $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,N_i})$ associated with each suborder $h^\dagger[o_i]$. Here each $y_{i,j} = (y_{i,j,1}, \ldots, y_{i,j,k_i}) \in \mathcal{P}_{o_i}$ is an entire list for each observation $j = 1, \ldots, N_i$ of the group o_i, $i \in I$. We will set $N_i = 0$ for $i \in [2^n - 1] \setminus I$ and take the limit as $\min_{i\in I} N_i \to \infty$. The total number of observed lists is $N = \sum_{i=1}^{2^n-1} N_i = \sum_{i\in I} N_i$.

Proposition 5. *(Consistency for the $(K, n, \rho)\backslash P$ model) Let $h^\dagger \in \mathcal{H}_{[n]}$ be given and suppose $y_{i,j} \sim p^{(P)}(\cdot|h^\dagger[o_i])$ jointly independent for all $i \in I$ and $j = 1, \ldots N_i$. Let $\pi(h|\mathbf{y}_{1:N})$ be the posterior of the $(K, n, \rho)\backslash P$ model with $K \geq \lfloor n/2 \rfloor$, so that*

$$\pi(h|\mathbf{y}_{1:N}) = \int_{Z:h(Z)=h} \int_{\rho=0}^{1} \pi(Z, \rho|\mathbf{y}) d\rho dZ$$

where

$$\pi(Z, \rho|\mathbf{y}_{1:N}) \propto p^{(P)}(\mathbf{y}_{1:N}|h(Z))\pi(Z|\rho)\pi_\rho(\rho).$$

If for each pair of actors $(i, j) \in [n] \times [n]$ *with* $i \neq j$*, there exists* $k \in I$ *such that* $\{i, j\} \subseteq o_k$ *then* $\pi(h^\dagger | \mathbf{y}_{1:N}) \to 1$ *as* $\min_{i \in I} N_i \to \infty$ *for every* $h^\dagger \in \mathcal{H}_{[n]}$.

Proof: see Appendix F.

In order to recover $h^\dagger$ with certainty we must be able to observe at least one list o_k informing the relation between each pair (i, j) of actors in $h^\dagger$ infinitely often. We have an infinite number of observations of a random variable informing every possible edge of the DAG representation of h.

These results are not given for the posteriors arising in a prior model for a partial order with ties. In fact ties are not identifiable as they are treated as unordered in the observation model. When we fit the (PDP, ρ)-partial order model in the noise free observation model, the posterior will concentrate on the *tie class* of the true partial order - the set of partial orders which are equivalent to the true partial order up to ties. The posterior distribution within this tie class of partial orders will just be the prior distribution. There are a number of reasons why the tied model remains useful for us. Firstly, adding ties changes the prior distribution in a fundamental and desirable way, shifting probability mass onto orders with common relations between groups (such as "bucket" orders and Vertex Series Parallel partial orders). Social hierarchies of this sort are common in human society. We do not restrict the class of partial orders we consider to bucket orders, but we can use ties to give them more weight in the prior. Secondly, taken with the randomly variable column dimension K, the latent variables Z^* in a model for partial orders with ties can potentially fit the data with lower parameter dimension ($C \times K$ with $C \leq n$ and potentially $K < n/2$) than the latent variables Z ($n \times n/2$ fixed) in the (K, n, ρ)-model for partial orders.

5. Application and Model Comparison

We demonstrate our models by applying them to a social hierarchy study. We fit both the $(PDP, \rho)\backslash Q$ and the $(PDP, \rho)\backslash M$ on two different datasets from the 'Royal Acta' data (see section 5.1). We analyse the inference results, and conduct reconstruction-accuracy test with synthetic data (section 5.2). We then perform model comparison with 1) models with no ties (K, n, ρ); and 2) other ranking methods not based on partial orders - the Plackett-Luce mixture and the Mallows mixture models.

5.1. The 'Royal Acta' Data

The 'Royal Acta' is a database created 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 royal acts (mainly charters but also writs and other letters) issued in the names of two English kings, William II (reigned 1087 to 1100), and his brother Henry I (reigned 1100 to 1135). Each royal act is identified to a certain time period. Some may be specific to a single year, while some with more uncertainty are assigned to a range of time. Each royal act comes with a witness list - an ordered list of names of individual witnesses. For some examples, see Appendix K.1. The order reflects the relative social importance of individual witnesses. For example, an order Archbishops $\succ$ Bishops $\succ$ Earls can be clearly observed. The historians are interested in studying the social hierarchy among the bishops specifically. We infer these relations by partial orders, which are natural choices for social hierarchy representation given their transitivity and generality compared to total orders. Power hierarchies were rigidly observed at this time. However it is natural to expect some incomparabilities or ties between bishops. We extract the sub-lists that only contain bishops (as these are listed in a block together). For more background and details on data processing, please refer to Nicholls and Muir Watt (2011) and Nicholls et al. (2023).

The data is temporal. We take 'snapshots' for the time periods 1131-1133 and 1100-1103 to study the social hierarchy among bishops in these two time periods. Table 1 summarises some key statistics of the datasets. The full witness lists for these two periods are shown in Appendix K.1. We assign each bishop in a single time period with a unique label. Assume the ground set of bishops in a certain time period is $[n]$. We assume each witness list to be an observed ranking list $y_i, i = 1, \ldots, N$ drawn from the 'true' social hierarchy partial order h with possible recording errors. The lists $Y = \{y_1, \ldots, y_N\}$ are incomplete, in the sense that the membership in list i is $o_i \subset [n]$ in a certain time period. The witness lists are of varied lengths.

Some contradictions can be observed in the witness lists. For example, in 1131-1133, bishops $7, 11, 12$ and 13 appear in both lists 2 and 16 (along with others). However in list 2 the order is $(13, 12, 7, 11)$ while list 16 has 12 and 13 swapped but otherwise the same. Is this just noise, are 12 and 13 unordered or are they tied? Firstly, from a prior modelling perspective but not pretending to any historical expertise, bishop 13 (Henry, de Blois, Bishop of Winchester, 1129-1171) and bishop 12 (Gilbert, the Universal, bishop of London) may be incomparable. They come from two rather independent

Time Period	1131-1133	1100-1103
Number of Lists	21	13
Number of Actors	15	9
Length of the Longest List	8	8
Length of the Shortest List	3	2

Table 1: Information on data structure.

power systems as Winchester at the time was an important and independent administrative centre and seat of power, on a par with London, so may be incomparable (like administrators and academics). It defined a separate hierarchy of tied cities. Secondly, they might be tied, i.e. they are each identified by all other bishops as top powers and possess the same power relation with all other bishops and are ranked with the same importance. Thirdly, the switch in order in the lists might simply be a recording error. Perhaps Henry (13) simply arrived late for witnessing document 16 while his true status was higher than that of Gilbert (12). It should be said that historians do not expect many errors in these data as the deeds involved concerned properties of significant value, and the King or Queen were often themselves signatories.

5.2. Reconstruction-Accuracy Tests

Our list data are incomplete and the number of lists N are not much larger than the number of actors n. In order to measure the reliability of the partial order (social hierarchy) reconstructions and to study the $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models performance on different datasets, we perform reconstruction accuracy tests on both models. In particular, we take the last sampled state of partial order $h^{(T)}$ in section 5.3 and generate synthetic data with the same lengths and list-membership as the real data

$$y_i \sim p(\cdot|h^{(T)}[o_i], p, \theta),\ i = 1, ..., N$$

for both periods 1131-1133 and 1100-1103. If our models are correct then the data contain enough information to reconstruct the truth accurately. We consider the following four synthetic datasets for each time period:

- Simulation 1: error-free, $y_i \sim \mathcal{U}(\mathcal{L}[h^{(T)}[y_i^{obs}]]), \forall i \in [N]$.

- Simulation 2: data-list with random error. For $i \in [N]$, we simulate $y_i \sim \mathcal{U}(\mathcal{L}[h^{(T)}[y_i^{obs}]])$ from the noise-free model; we select a pair of actors $a, b \in y_i$ uniform at random, $a \neq b$, and put them in the same order as they appear in the data. If $a \prec b$ in y_i but $a \succ b$ in y_i^{obs} then exchange the positions of a and b in y_i leaving all else unchanged.[7]

- Simulation 3: data-list with Mallows error given θ^*,

$$y_i \sim p^{(M)}(\cdot|h^{(T)}[o_i], \theta^*), \forall i \in [N].$$

- Simulation 4: data-list with queue-jumping error given p^*,

$$y_i \sim p^{(Q)}(\cdot|h^{(T)}[o_i], p^*), \forall i \in [N].$$

Simulation 2 blends a little of the real data into the synthetic data. It uses a noise process that is neither Mallows nor queue-jumping, so the fitted models are all (mildly) misspecified. Note that the θ^* and p^* values in Simulation 3 and 4 are chosen so that they produce roughly the same level of noise in the data-lists. The experiments presented in this section choose $\theta^* = 2.7$ and $p^* = 0.05$.[8] We fit both the $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models on above simulations. Each MCMC is run for $1e5$ iterations. We omit our convergence analysis on this synthetic data, but ESS values and trace plots showed convergence as good as or better than the convergence and mixing we report for the real data (given below).

We summarise the sampled partial orders using consensus partial orders: $h^{con}(\epsilon)$ includes order relation/edge $(i \succ_{h^{con}(\epsilon)} j)$ if the relation appears more than ϵT times in the T MCMC samples after thinning. The true and consensus partial orders for all experiments in this section are given in Appendix K.4. We calculate the proportion of the true-positive and false-positive relations for $h^{con}(\epsilon), \epsilon \in [0, 1]$ and construct the receiver operating characteristic (ROC) curves in Figure 11 (for simulation 1 and 2) and Figure 12 (for simulation 3 and 4) respectively. The ROC curves show the proportion of false-positive and true-positive relations given concensus order threshold ϵ.

[7] Simulation 2 introduces around 5% error to the synthetic data.

[8] With these values ($\theta^* = 2.7, p^* = 0.05$) the proportion of lists with errors is $\sim 43\%$ error for the 1131-1133 data structure ($N = 21$) and $\sim 24\%$ for the 1100-1103 data structure ($N = 13$; lists tend to be shorter so lower error probability, see Appendix K.1).

The true positive rate (TPR) and false positive rate (FPR) increase with decreasing ϵ from $(0, 0)$ at $\epsilon = 1$ (the consensus order is empty) to $(1, 1)$ at $\epsilon = 0$ (complete graph).

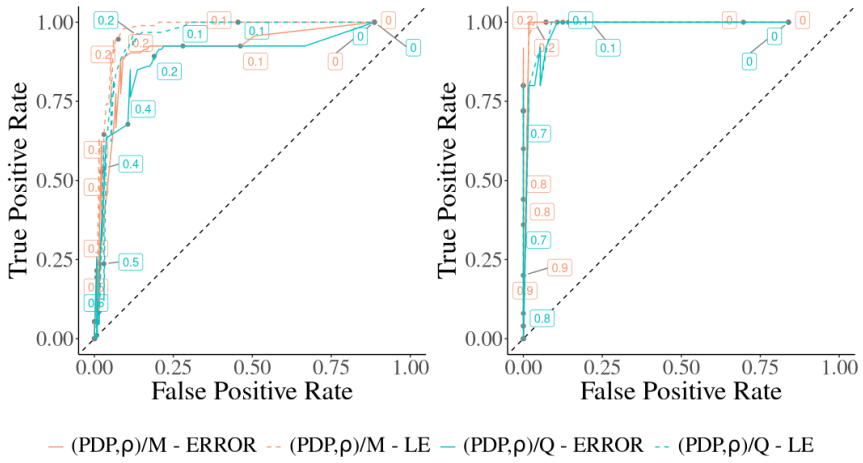


Figure 11: The receiver operating characteristic (ROC) curves for synthetic data - simulation 1 (LE; dashed line) & 2 (ERROR; solid line) - using synthetic data with 1131-1133 (left) and 1100-1103 (right) list membership structures, analysed with the $(PDP, \rho)\backslash M$ (orange) and $(PDP, \rho)\backslash Q$ (blue) models. The true/false positive rates are plotted against ϵ the threshold to construct concensus order $h^{con}(\epsilon), \epsilon \in [0, 1]$.

For each simulated data set there is ϵ giving high true-positive and low false-positive reconstructed relation fractions: if our model is accurate then we reconstruct relations well. Both Figure 11 and Figure 12 have an elbow closest $(0, 1)$ around $\epsilon = 0.2$. This value gives a good balance of true- and false-positives for synthetic reconstruction analyses on the $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models for both time periods. The consensus orders with $\epsilon = 0.2$ and the 'true' partial orders for simulation are shown in Appendix K.4. We use the same threshold when we present consensus orders on real data in section 5.3. From Figure 11 and Figure 12, both models reconstruct the true order relation well. We observe that the $(PDP, \rho)\backslash M$ model slightly outperforms the $(PDP, \rho)\backslash Q$ model at partial order reconstruction in most scenarios. Under the 1131-1133 structured data, the two models appear to be robust to the Mallows or queue-jumping error type and both reconstruct the truth well.

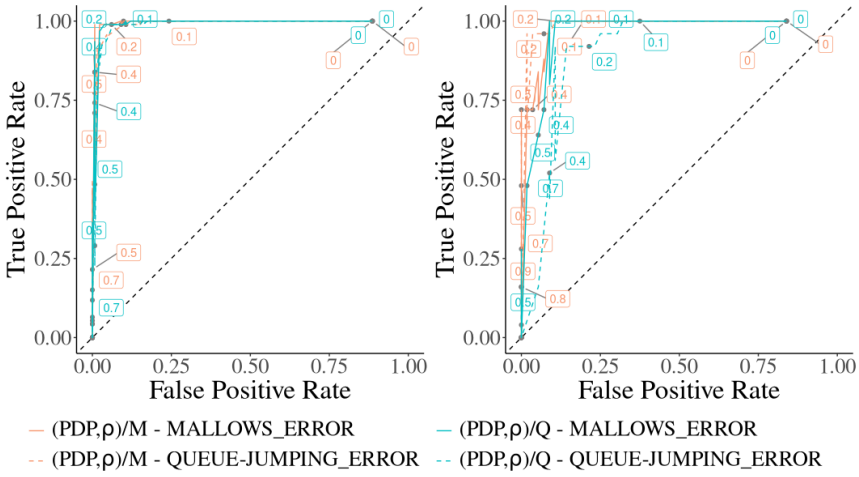


Figure 12: The receiver operating characteristic (ROC) curves for synthetic data - simulation 3 (MALLOWS_ERROR; solid line) & 4 (QUEUE-JUMPING_ERROR; dashed line) - using synthetic data with 1131-1133 (left) and 1100-1103 (right) list membership structures, analysed with the $(PDP, \rho)\backslash M$ (orange) and $(PDP, \rho)\backslash Q$ (blue) models. The true/false positive rates are plotted against ϵ the threshold to construct concensus order $h^{con}(\epsilon), \epsilon \in [0, 1]$.

5.3. Application on the 'Royal Acta' Data

We perform Bayesian inference for the (PDP, ρ)-partial order model with the queue-jumping[9] or Mallows observation model. We employ Metropolis-Hasting MCMC for such inference. Each chain is run for $1e5$ iterations where we record every $2n$ steps. We choose a burn-in period of $500 \times 2n$ for all chains. See Appendix K.2 for Effective Sample Sizes and Appendix K.2.2 for MCMC output traces. These appear to show good MCMC convergence and mixing.

The posterior distributions on partial order depth are shown in Figure K.19. Both the $(PDP, \rho)\backslash M$ and the $(PDP, \rho)\backslash Q$ models conclude similar depths in general. It is clear that the hierarchy is not a total order (depths close to n have low posterior probability in both analyses). The mean posterior depths are around 8.8 for both models in 1131-1133 (with

[9]We choose queue-jumping "up" (rather than "down") for this specific dataset. For example, we know of cases where a lower status bishop is the nephew of a higher status bishop and is promoted to appear immediately after his uncle in the list. Experiments with the two queue-jumping directions in Nicholls et al. (2023) gave very similar results.

$n = 15$ bishops). In 1100-1103 (where $n = 9$), we obtain posterior mean depths of 3.9 for $(PDP, \rho)\backslash M$ and 4.8 for $(PDP, \rho)\backslash Q$. The partial orders in 1100-1103 are flatter in general as reflected on the consensus orders in Figure 13 and Figure 14 for the Mallows and the queue-jumping observation models respectively. The consensus orders from different observation models (Mallows or queue-jumping) are similar despite some slight relation difference, showing the result's robustness to different observation models. Bishop 3 (Robert, de Limesey, bishop of Chester), bishop 5 (Robert, Bloet, bishop of Lincoln) and bishop 8 (William, Giffard, bishop of Winchester, 1100-1129) are strongly tied during 1100-1103. During 1131-1133, both models suggest relatively high posterior probability of ties between Bishops 1 (Roger, Bishop of Salisbury), bishop 12 (Gilbert, the Universal, bishop of London) and bishop 13 (Henry, de Blois, Bishop of Winchester, 1129-1171). These are indicated in the 'heatmaps' in Figure 15 and Figure K.20 which display the probability of different pairs of actors fall in the same cluster.

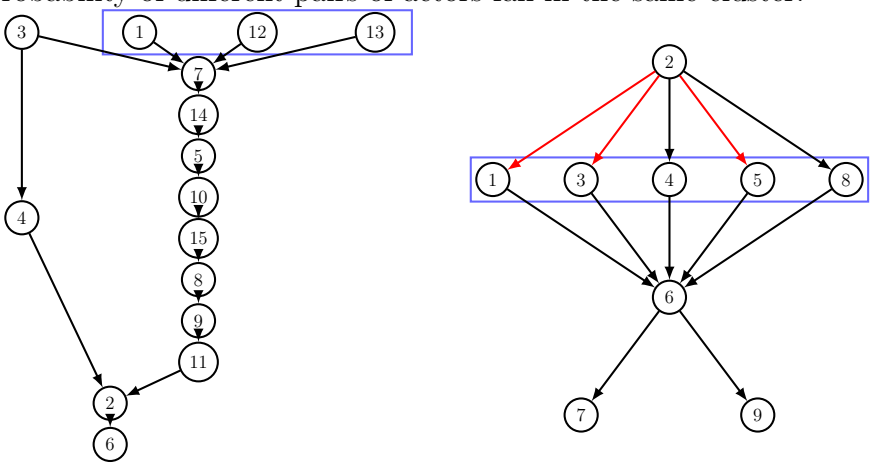


Figure 13: The consensus orders for the $(PDP, \rho)\backslash M$ model on 1131-1133 (left) and 1100-1103 (right) Royal Acta (bishop) data. We conclude an edge if such order relation has more than 0.2 posterior probability (inferred from section 5.2). An edge is colored red if it has more than 0.9 posterior probability. The *blue boxes* indicate the tie relations with more than 0.5 posterior probabilities.

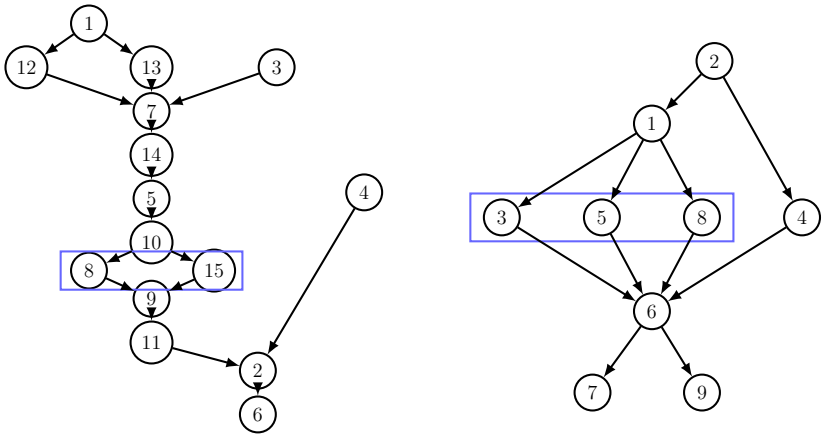


Figure 14: The concensus orders for the $(PDP, \rho)\backslash Q$ model on the 1131-1133 (left) and 1100-1103 (right) 'royal acta' (bishop) witness list data. We conclude an edge if such order relation has more than 0.2 posterior probability (inferred from section 5.2). An edge is colored red if it has more than 0.9 posterior probability. The *blue boxes* indicate the tie relations with more than 0.5 posterior probabilities.

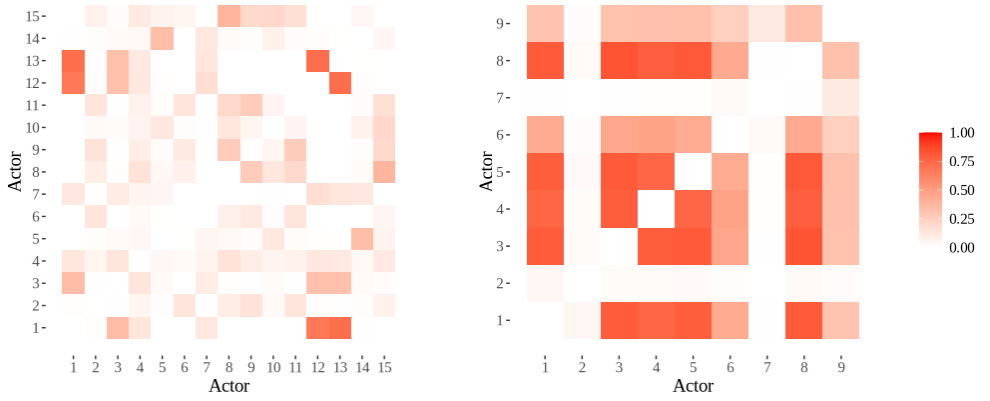


Figure 15: The 'heatmaps' for the estimated clusters for the $(PDP, \rho)\backslash M$ model for time periods 1131-1133 (left) and 1100-1103 (right). Each block indicates the posterior probability for a pair of actors to be in the same cluster. We label higher posterior probability red and lower posterior probability white.

We present the prior and posterior distributions on some key parameters, as is shown in Figure 16. Figure K.19 shows the prior and posterior distributions of the number of clusters and partial order depths in each model. Both the posterior distributions for K and ρ follow similar behaviour across the two models in different time periods. Prior sensitivity analysis in Appendix

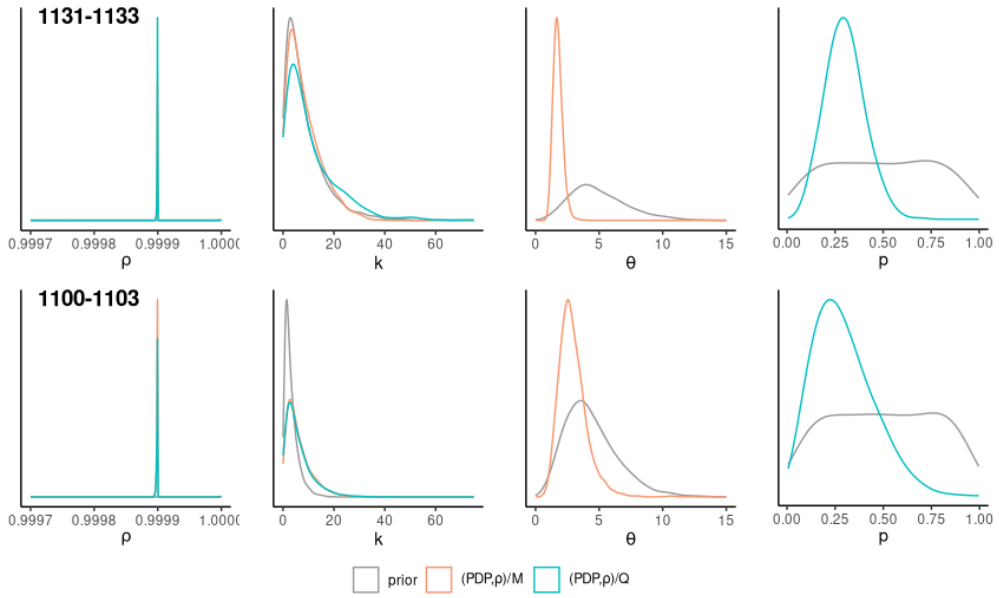


Figure 16: The prior/posterior distributions for key parameters for the $(PDP, \rho)\backslash M$ (pink) and the $(PDP, \rho)\backslash Q$ (blue) models on periods 1131-1133 (upper) and 1100-1103 (lower). The prior distributions are colored in gray.

K.3 shows that the posterior of K is less sensitive to the prior choice during 1131-1133 (where there is more data). A prior mean around $n/2$ appears to be the optimal choice among $\bar{k} \in \{n, n/2, n/4\}$. We therefore present the inference results for prior $\bar{k} = n/2$ in this section. We notice that the posterior for ρ tends to concentrate on values close to one. This is because K and ρ are only weakly identifiable. For any reasonable geometric prior on K, the model can fit well by adjusting the ρ posterior (as K gets larger the hazard for paths to cross increases, so make ρ closer to one for flatter paths and a similar distribution over partial orders). According to our sensitivity analysis to K-prior, when the K-prior concentrates on a small mean-K, the posterior for ρ is supported on lower values. However, the freedom of choosing both K and ρ provides us more flexible control over partial order depth in the prior. The depth distribution in Figure 10 cannot be realised with a single parameter of choice.

In the Mallows observation model, the dispersion parameter θ captures the level of noise in the observed list. The θ priors are adjusted to allow roughly 10% of noise in lists as we expect errors were rare in these data. The

posterior means for θ in the $(PDP, \rho)\backslash M$ models are 1.75 and 2.94 in 1131-1133 and 1100-1103 respectively, somewhat lower than the prior mean, so somewhat higher noise than expected a priori. In particular, the witness lists in period 1131-1133 appear to have higher noise compared to 1100-1103. This observation is also supported in the inference from the $(PDP, \rho)\backslash Q$ model. With the queue-jumping observation model, the parameter p indicates the error probability - the probability for a bishop to jump up the queue. The p-priors are relatively flat, whereas the posterior distributions for p both indicate an error rate of around 30% and are relatively concentrated.

Overall, we obtain consistent inference results from the $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models in both time periods. We conclude that *Roger, Bishop of Salisbury*, *Henry, de Blois, Bishop of Winchester, 1129-1171* (tied) and *Gilbert, the Universal, bishop of London* are of the highest power hierarchy in 1131-1133 while *Everard, bishop of Norwich* is at the bottom of such order relation. Similarly, in 1100-1103, *Maurice, bishop of London* is of the highest hierarchy while *Ranulf, Flambard, bishop of Durham* and *Roger, Bishop of Salisbury* has the lowest. Numerous factors might contribute to this social power relation, e.g. the area one bishop represent, the age of the bishop, etc. We further conclude that there are slight recording errors in the order of witness signature, which can be effectively captured by our model.

5.4. Model Comparison

The (PDP, ρ)-partial order model is preferred over the (K, n, ρ)-partial order model according to the Bayes Factor. If M_0 is the (K, n, ρ)-partial order model (with K fixed to $\lceil n/2 \rceil$) and M_1 is the (PDP, ρ)-partial order model, then M_0 is nested in M_1 (when each actor is in a separate cluster in M_1 so $C(S) = n$ (the number of clusters in S), $S = (\{1\}, \ldots, \{n\})$ and the random K-value in M_1 takes the value $\lceil n/2 \rceil$). The Savage-Dickey Density Ratio $B_{10} = \frac{\pi_{M_1}(C=n, K=\lceil n/2 \rceil)}{\pi_{M_1}(C=n, K=\lceil n/2 \rceil | \mathbf{y})}$, which gives the Bayes factor in this case, is easily estimated using samples from the M_1 posterior and prior. Values bigger than one are evidence for model M_1. As is shown in table 2, the Bayes factors under different observation models all favor the (PDP, ρ)-partial order model for both 1131-1133 and 1100-1103. Clustering and varying K effectively reduce the dimension of Z-matrix without compromising the fit. Ties enforce equal relations to other actors, and it seems the data "likes" a prior that encourages this sort of structure. The Bayes factors values parsimony so this may also contribute to the weighting.

	1131-1133	
	Mallows	Queue-jumping
B_{10}	4.2	3.2
Evidence for M_1	"substantial"	"weak"
	1100-1103	
	Mallows	Queue-jumping
B_{10}	7.1	33
Evidence for M_1	"substantial"	"strong"

Table 2: Bayes factors B_{10} between the (PDP, ρ)-partial order model (M_1 *with ties and variable latent matrix dimension*) and the $(K, \lceil n/2 \rceil, \rho)$-partial order model ($M_0$, with $K = \lceil n/2 \rceil$ fixed) with either the Mallows or queue-jumping observation models in periods 1131-1133 and 1100-1103. A large Bayes factor indicates preference towards the $(PDP.\rho)$-partial order model. Interpretation follows Jefferys.

We further compare our models with the popular ranking models - the Plackett-Luce and Mallows Mixture. It would be hard to estimate Bayes factors in this case so we use the expected log pointwise predictive density (*elpd*, Vehtari et al. (2017)) as our model comparison criterion. This is a predictive loss which can be estimated using the WAIC (Watanabe, 2013). We present $elpd_{waic}$-estimates in table 3. The partial order models have significantly higher $elpd_{waic}$, and are therefore strongly preferred compared to the total order methods on the 1131-1133 and 1100-1103 Bishop witness list data. Under the (PDP, ρ)-partial order model, the Mallows and queue-jumping models perform almost equally well - with the Mallows model being slightly preferred. This is as expected as the Mallows model allows bi-directional errors while the queue-jumping model only allows queue-jumping-up. For the Mallows or Plackett-Luce mixture models, we choose the number of mixture components M that provides the highest $elpd_{waic}(M \in \{1, \ldots, 10\})$. The Mallows mixture appears to outperform the Plackett-Luce mixture.

6. Conclusion and Future Work

We have proposed a new non-parametric statistical model for partial order estimation under the Bayesian framework. The new (PDP, ρ)-partial order model improves on the model set out in Nicholls and Muir Watt (2011) by

	$elpd_{waic}$	
Models	1131-1133	1100-1103
$(PDP, \rho)\backslash M$	-48.2 (9.5)	-23.3 (10.4)
$(PDP, \rho)\backslash Q$	-56.6 (10.4)	-26.0 (11.4)
Mallows Mixture (Kendall-tau)	-72.9 (13.0) (M=1)	-33.9 (12.0) (M=3)
Plackett-Luce Mixture	-173.2 (17.8) (M=4)	-77.5 (10.4) (M=3)

Table 3: Model comparison among $(PDP, \rho)\backslash M$, $(PDP, \rho)\backslash Q$, Plackett-Luce Mixture and Mallows Mixture (with Kendall-tau distance) on time periods 1131-1133 and 1100-1103. We use $elpd_{waic}$ as the model comparison criterion. The results are presented as 'estimation (standard error)'. For the Mallows and Plackett-Luce mixture models, we only consider models with 1-10 mixture components. Here we report the best $elpd_{waic}$ with its corresponding number of mixture component M.

allowing actors to be tied. The more complex model remains projective. The Mallows error model improves on the queue-jumping error model of Nicholls and Muir Watt (2011). We also estimate, rather than fix, the column dimension of the latent matrix Z introduced by Nicholls and Muir Watt (2011) to model the prior for partial orders. The tied-structure of the new model further reduces the dimension of the latent Z-matrix by grouping its rows into equivalence classes. It changes the prior distribution over partial orders to promote more 'bucket' order-like structures. This is appealing from a modeling perspective. Experiments on the 'Royal Acta' data showed that the tied model out-performed the nested model without ties and fixed column dimension.

We also added some theory for the model given in Nicholls and Muir Watt (2011) in the noise free case. We gave necessary and sufficient conditions for the posterior to concentrate on the true partial order. We can ask, if the true model was a partial order model, and we fit say, a Mallows or Plackett-Luce model, what behaviour do we get? Is it possible that the posterior converges to a mixture of rankings rather than a single ranking, and thereby captures some of the "uncertainty" which a partial order allows. For the simplest possible example, consider the partial order h_0 in figure 1 with three linear extensions in figure 2. Let N denote the number of ranked lists sampled uniformly at random from the linear extensions of partial order h_0. We take $N \to \infty$. It is not too hard to see that the Mallows model posterior with Hamming or Cayley distance must concentrate on $(1, 3, 2, 4, 5)$

in the large data limit (rather than say, a posterior which is uniform on the linear extensions of h_0 as we might hope). If we could get a Mallows-mixture to converge to the uniform distribution on the linear extensions of the true partial order it would then be possible to reconstruct the true partial order by intersecting total orders sampled from the posterior. Such an analysis is probably impossible in principle: fitting partial order models is a computational task in $\#P$; fitting a Mallows model with any likelihood which can be evaluated in polynomial time is not in $\#P$. It is unlikely we can solve the harder problem by solving the easier problem. Fundamentally though, the reason we work with partial orders is that we think social hierarchies are well-represented by partial orders, that any total order assumption is too strong, and so if we want to reconstruct partial orders they should be an object in the model.

Future research may consider including covariates in the partial order model, following Nicholls et al. (2023) but adding our non-parametric approach, or exploring a different definition of ties where the tied actors have to appear together in a linear extension.

References

Babington-Smith, B., 1950. Discussion of professor ross's paper. Journal of the Royal Statistical Society B 12, 41–59.

Baker, R., Scarf, P., 2020. Modifying Bradley–Terry and other ranking models to allow ties. IMA Journal of Management Mathematics 32, 451–463. URL: https://doi.org/10.1093/imaman/dpaa027, doi:10.1093/imaman/dpaa027.

Beerenwinkel, N., Eriksson, N., Sturmfels, B., 2007. Conjunctive bayesian networks. Bernoulli , 893–909.

Bradley, R.A., Terry, M.E., 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39, 324–345.

Brightwell, G., 1993. Models of random partial orders. Surveys in combinatorics 5383.

Bubley, R., Dyer, M., 1999. Faster random generation of linear extensions. Discrete mathematics 201, 81–88.

Caron, F., Doucet, A., 2012. Efficient bayesian inference for generalized bradley–terry models. Journal of Computational and Graphical Statistics 21, 174–196.

Caron, F., Teh, Y.W., Murphy, T.B., 2014. Bayesian nonparametric plackett–luce models for the analysis of preferences for college degree programmes. The Annals of Applied Statistics , 1145–1181.

Diaconis, P., 1988. Group representations in probability and statistics. Lecture Notes-Monograph Series 11, i–192. URL: http://www.jstor.org/stable/4355560.

Feng, J., Fang, Q., Ng, W., 2008. Discovering bucket orders from full rankings, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 55–66.

Gionis, A., Mannila, H., Puolamäki, K., Ukkonen, A., 2006. Algorithms for discovering bucket orders from data, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 561–566.

Guiver, J., Snelson, E., 2009. Bayesian inference for plackett-luce ranking models, in: proceedings of the 26th annual international conference on machine learning, pp. 377–384.

Henderson, D.A., 2022. Modelling and analysis of rank ordered data with ties via a generalized plackett-luce model. arXiv preprint arXiv:2212.08543 .

Hiraguchi, T., 1951. On the dimension of partially ordered sets. The science reports of the Kanazawa University 1, 77–94.

Jiang, C., Nicholls, G.K., Lee, J.E., 2023. Bayesian inference for vertex-series-parallel partial orders, in: Uncertainty in Artificial Intelligence, PMLR. pp. 995–1004.

Karzanov, A., Khachiyan, L., 1991. On the conductance of order Markov chains. Order 8, 7–15.

Kleitman, D.J., Rothschild, B.L., 1975. Asymptotic enumeration of partial orders on a finite set. Transactions of the American Mathematical Society 205, 205–220.

Knuth, D.E., Szwarcfiter, J.L., 1974. A structured program to generate all topological sorting arrangements. Information Processing Letters 2, 153–157.

Liu, A., Zhao, Z., Liao, C., Lu, P., Xia, L., 2019a. Learning plackett-luce mixtures from partial preferences, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4328–4335.

Liu, Q., Crispino, M., Scheel, I., Vitelli, V., Frigessi, A., 2019b. Model-based learning from preference data. Annual review of statistics and its application 6, 329–354.

Lu, T., Boutilier, C., 2014. Effective sampling and learning for mallows models with pairwise-preference data. J. Mach. Learn. Res. 15, 3783–3829.

Luce, R.D., 1959. On the possible psychophysical laws. Psychological review 66, 81.

Mallows, C.L., 1957. Non-null ranking models. i. Biometrika 44, 114–130.

Meila, M., Chen, H., 2012. Dirichlet process mixtures of generalized mallows models. arXiv preprint arXiv:1203.3496 .

Mollica, C., Tardella, L., 2017. Bayesian plackett–luce mixture models for partially ranked data. Psychometrika 82, 442–458.

Muir Watt, A., 2015. Inference for partial orders from random linear extensions. Ph.D. thesis. University of Oxford.

Murphy, T.B., Martin, D., 2003. Mixtures of distance-based models for ranking data. Computational statistics & data analysis 41, 645–655.

Neal, R.M., 2000. Markov chain sampling methods for dirichlet process mixture models. Journal of computational and graphical statistics 9, 249–265.

Nicholls, G.K., Lee, J.E., Karn, N., Johnson, D., Huang, R., Muir-Watt, A., 2023. Bayesian inference for partial orders from random linear extensions: power relations from 12th century royal acta. `arXiv:2212.05524`.

Nicholls, G.K., Muir Watt, A., 2011. Partial order models for episcopal social status in 12th century england. IWSM 2011 , 437.

Pearce, M., Erosheva, E.A., 2024. Bayesian rank-clustering. URL: https://arxiv.org/abs/2406.19563, arXiv:2406.19563.

Pitman, J., 2006. Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002. Springer.

Plackett, R.L., 1975. The analysis of permutations. Journal of the Royal Statistical Society: Series C (Applied Statistics) 24, 193–202.

Sharpe, R., Carpenter, D., Doherty, H., Hagger, M., Karn, N., 2014. Charters of william ii and henry i.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological review 34, 273.

Turner, H.L., van Etten, J., Firth, D., Kosmidis, I., 2020. Modelling rankings in R: The PlackettLuce package. Computational Statistics 35, 1027–1057. URL: https://doi.org/10.1007/s00180-020-00959-3, doi:10.1007/s00180-020-00959-3.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and waic. Statistics and computing 27, 1413–1432.

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., Arjas, E., 2018. Probabilistic preference learning with the mallows rank model. Journal of Machine Learning Research 18, 1–49.

Watanabe, S., 2013. A widely applicable bayesian information criterion. Journal of Machine Learning Research 14, 867–897. URL: http://arxiv.org/abs/1208.6338.

Winkler, P., 1985. Random orders. Order 1, 317–331.

Appendix A. The intersection order is the maximum likelihood estimator

The material in the appendix follows the ideas of Beerenwinkel et al. (2007) fairly closely.

Proposition 6. *Given noise-free data lists* $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ *with full lengths, i.e.* $y_j \in \mathcal{P}_n \, \forall j \in [N]$*, and the uniform observation model on linear extensions*

$$f(y_j|h) = \frac{\mathbb{1}_{y_j \in \mathcal{L}[h]}}{|\mathcal{L}[h]|}.$$

The intersection order h_{int} *is the maximum likelihood estimator of the true partial order.*

Proof. The likelihood function given the uniform observation model is

$$L(h; \mathbf{y}) = \prod_{j=1}^{N} \frac{\mathbb{1}_{y_j \in \mathcal{L}[h]}}{|\mathcal{L}[h]|} = \frac{1}{|\mathcal{L}[h]|^N} \prod_{j=1}^{N} \mathbb{1}_{y_j \in \mathcal{L}[h]}$$

The intersection order h_{int} admits $u \succ_{h_{int}} v$ with $u, v \in [n]$ and $u \neq v$ if and only if u appears before v in all $\mathbf{y}$. For $j \in [N]$ and $1 \leq a < b \leq n$, we have either $y_{j,a} \| y_{j,b}$ or $y_{j,a} \succ y_{j,b}$ in h_{int}. Therefore, $\prod_{j=1}^{N} \mathbb{1}_{y_j \in \mathcal{L}[h_{int}]} = 1$.

Let $h_r \in \mathcal{H}_{[n]}$ with $h_r \neq h_{int}$ be some partial order such that $y_j \in \mathcal{L}[h_r] \, \forall j \in [N]$. If h_r admits one more order relation than h_{int}, at least one y_j will be violated. So the condition $y_j \in \mathcal{L}[h_r] \, \forall j \in [N]$ wouldn't hold. That is, for all the order relations in h_{int}, there exists at least one pair of actors $u, v \in [n], u \neq v$ with $(u \succ_{h_{int}} v)$ such that $u \|_{h_r} v$. This gives $|\mathcal{L}[h_{int}]| < |\mathcal{L}[h_r]|$. Therefore,

$$L(h_{int}; Y) > L(h; Y) \; \forall h \in \mathcal{H}_{[n]}, h \neq h_{int}$$

Hence, the intersection order h_{int} is the maximum likelihood estimator of the true partial order given full-length noise-free list data and the uniform observation model on linear extensions. □

Appendix B. Partial orders on three actors

The 19 possible edge sets E_3 are shown in Figure B.17.

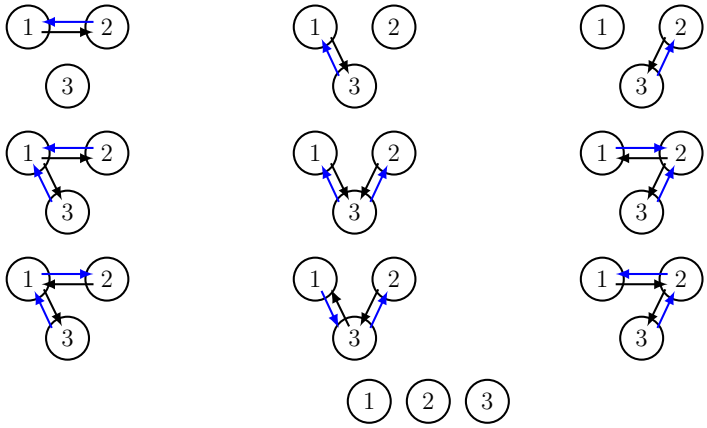


Figure B.17: All possible partial orders given 3 actors. The blue and black edges represent different partial orders respectively.

Appendix C. Proof of Proposition 1

Proposition 1. *The (K, n)-partial order model distributions are projective.*

Proof. For $h \in \mathcal{H}_{[n]}$, let $Z[h] = \{Z' \in R^{n\times K} : h(Z') = h\}$ be the set of Z-matrices generating h so that if $\pi_{[n]}(h) = \Pr(H = h)$ under the (K, n)-model then $\pi_{[n]}(h) = \Pr(Z \in Z[h])$. Let $o = [n-1]$. Let $H_{n-1} \sim \pi_{[n-1]}$ and $H_n \sim \pi_{[n]}$. It is sufficient for marginal consistency to show that $H_n[o] \sim H_{n-1}$, that is $\Pr(H_{n-1} = h) = \Pr(H_n[o] = h)$ for all $h \in \mathcal{H}_{n-1}$. We begin on the RHS.

Let $Z \in [0,1]^{n\times K}$ with $Z_{i,1:K} \sim \mathcal{U}(0,1)^K,\ i = 1, ..., n$ so that $H_n \sim h(Z)$. We have

$$\Pr(H_n[o] = h) = \Pr(h(Z)[o] = h).$$

Next we observe that $h(Z)[o] = h(Z[o, 1{:}K])$ and so

$$\Pr(h(Z)[o] = h) = \Pr(h(Z[o, 1{:}K]) = h).$$

This is because relations between any two rows of Z are not effected by the values in any other rows, so the suborder $h(Z)[o]$ we get by dropping actors outside o is the same as the suborder we get if we compute it on the rows of Z belonging to the actors in o. Now let $Z' \in \mathbb{R}^{(n-1)\times K}$ with

$Z'_{i,1:K} \sim \mathcal{U}(0,1)^K,\ i = 1, ..., n-1$. We have $Z' \sim Z[o, 1{:}K]$ since the rows are all independent. So we may write

$$\Pr(h(Z[o, 1{:}K]) = h) = \Pr(h(Z') = h)$$

and we are done. □

Appendix D. Proof of Proposition 3

Proposition 3. *For $h^* \in \mathcal{H}^*_n$ let $\pi_{[n]}(h^*) = \Pr(H^* = h^*)$ be the distribution of the random partial order with ties generated by the process in Definition 7 and given in Equation 14. The family of distributions $\pi_{[n]}, n \geq 1$ is projective.*

Proof. Let $o = [n-1]$, let $H^*_n \sim \pi_{[n]}(\cdot|\rho, k)$ defined in Equation 13 and let $H^*_{n-1} \sim \pi_{[n-1]}(\cdot|\rho, k)$. We wish to show that $H^*_n[o] \sim H^*_{n-1}$. It is again sufficient to show $H^*_n[o]|\rho, k \sim H^*_{n-1}|\rho, k$ for every fixed ρ and K; if $\Pr(H^*_n[o] = h^*|\rho, k) = \Pr(H^*_{n-1} = h^*|\rho, k)$ then the marginals $\Pr(H^*_n[o] = h^*) = \Pr(H^*_{n-1} = h^*)$ are equal (just average over ρ and K on both sides). In order to verify $H^*_n[o] \sim H^*_{n-1}$ we start with $\Pr(H^*_n[o] = h^*|\rho, k)$. First of all, by the generative model, $H^*_n \sim h(Z(S, Z^*))$ if $S \sim P_{[n],\eta_b,\eta_a}$ and

$$Z^*_{c,1:K} \overset{i.i.d.}{\sim} \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_\rho) \text{ for } c = 1, ..., C.$$

Substituting for H^*_n,

$$\Pr(H^*_n[o] = h^*|\rho, k) = \Pr(h(Z(S, Z^*))[o] = h^*|\rho, k).$$

Now

$$h(Z(S, Z^*))[o] = h(Z(S, Z^*)[o, 1{:}k])$$

because the relations between actors $1, ..., n-1$ determined by $h(\cdot)$ acting on $Z[o, 1{:}k]$ are not effected by the presence or values of $Z[n, 1{:}k]$ so we just get the suborder.

Now let $S^{-n} = (S_1, ..., S_{C^{-n}})$ be the partition S with n removed. Let $Z^{*,-n}$ be the corresponding parameter matrix. If n is in a partition by itself, which must be $S_C = \{n\}$, then that set is deleted and $C^{-n} = C-1$; otherwise $C^{-n} = C$. If $C^{-n} = C$ then $Z^{*,-n} = Z^*$ and otherwise $Z^{*,-n} = Z^*_{1:C-1,1:k}$, so the last row of Z^* is dropped.

In terms of these quantities, $Z(S, Z^*)[o, 1{:}k] = Z(S^{-n}, Z^{*,-n})$, since again, if we remove n from the partition S then use S^{-n} with $Z^{*,-n}$ to determine the rows of Z for $1, ..., n-1$ in Z, we get the same reduced Z as if we determined the entries in Z from (Z^*, S) and then dropped row n from Z. We have then

$$\Pr(H_n^*[o] = h^*|\rho, k) = \Pr(h(S^{-n}, Z(Z^{*,-n})) = h^*|\rho, k).$$

The two parameter Poisson-Dirichlet process generates exchangeable random partitions Pitman (2006). This indicates each item is treated independently, as if they are the 'last arrival'. Given the formulation of the two-parameter Poisson-Dirichlet process , we have $S^{-n} \sim P_{[n-1],\eta_b,\eta_a}$. In addition, the latent matrix setup gives $Z^{*,-n}_{c,1:k} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_\rho)$ i.i.d. for $c = 1, ..., C^{-n}$. It follows that $h(Z(S^{-n}, Z^{*,-n}))$ is a realisation from the generative model at fixed ρ, k on $[n-1]$ and hence $h(Z(S^{-n}, Z^{*,-n})) \sim H^*_{n-1}$. We conclude that $\Pr(H_n^*[o] = h^*|\rho, k) = \Pr(H^*_{n-1} = h^*|\rho, k)$ for every fixed ρ, k, and hence, from the preamble $H_n^*[o] \sim H^*_{n-1}$. We have shown marginal consistency for the special case of $o = [n-1]$ but the actor labels are exchangeable, so we may remove any subset $o \subset [n]$ by removing one entry in $[n] \setminus o$ at a time, sorting it to be the 'last arrival' in the Poisson-Dirichlet process clustering labels. □

Appendix E. Proof of Proposition 4

The key here is to show the chain is irreducible. This is well known, for example, Karzanov and Khachiyan (1991) makes use of this fact for a closely related chain.

Proposition 4. *Consider a Markov chain $\{X_t\}_{t\geq 1}$, with state space $X_t \in \mathcal{L}[h[o_j]]$ for some fixed $j \in \{1, \ldots, N\}$. Suppose that at step t we have $X_t = x$. An entry $i \sim U\{1, ..., n_j\}$ is chosen uniformly at random. If $i = n_j$ then we reject and set $X_{t+1} = x$ and otherwise $x' = (x_1, ..., x_{i+1}, x_i, ..., x_{n_j})$. If $x' \in \mathcal{L}[h[o_j]]$ then we set $X_{t+1} = x'$ and otherwise $X_{t+1} = x$. The process converges in distribution to the uniform distribution over linear extensions of $h[o_j]$, that is, for $l \in \mathcal{L}[h[o_j]]$,*

$$P(X_n = l) \overset{t\to\infty}{\to} \frac{1}{|\mathcal{L}[h[o_j]]|} \mathbb{1}_{l\in\mathcal{L}[h[o_j]]}.$$

Proof. This finite-state Markov chain $\{X_t\}_{t\geq 1}$ is irreducible. Let $l, l' \in \mathcal{L}[h[o_j]]$ with $l \neq l'$ be a pair of linear exentsions in the target space. Let

$a = l_1$ and $b = l'_1$ be the first (or "top") entries in the lists and suppose $a \neq b$, so the top entries in the lists do not match. Suppose $l'_i = a$ with $i > 1$. We will show that we can swap a up to the top of l'. Suppose $c = l'_{i-1}$ with $c = b$ if $i = 2$. Now we must be able to swap a and c in l' without violating an order in $h[o_j]$ as a appears below c in l' but above c in l (as a is top in l so must be above c) so we have both orders $l = (a, ..., c, ...)$ and $l' = (b, ...c, a, ...)$. It follows that we must have $a \|_h c$; they are unordered in h. Iterating this procedure we can move from l' to l by a finite sequence of swaps each of which have probability $1/n_j$.

It is aperiodic (since it is irreducible and rejects when $i = n_j$, at least), and hence ergodic, so it has a unique stationary distribution π over $\mathcal{L}[h[o_j]]$. For $l, l' \in \mathcal{L}[h[o_j]]$ with $l \neq l'$ let $p_{l,l'} = \Pr(X_{t+1} = l' | X_t = l)$ denote the off diagonal elements of the transition matrix. We have

$$p_{l,l'} = \begin{cases} 1/n_j & \text{if } l, l' \text{ differ by a neighbor-swap} \\ 0 & \text{otherwise.} \end{cases}$$

Since p is symmetric, its columns sum to one, so it admits the uniform distribution as a left eigenvector, $\mathbf{1}^T p = \mathbf{1}^T$ with eigenvalue one, where $\mathbf{1}$ is a vector of $|\mathcal{L}[h[o_j]]|$-ones, so

$$\pi_l = \frac{1}{|\mathcal{L}[h[o_j]]|}, \quad l \in \mathcal{L}[h[o_j]].$$

□

Appendix F. Model Consistency for the Partial Order Model with the Error-Free Observation Model

Recall from the notation of Section 4.1 that $O = (o_i)_{i=1}^{2^n-1}$ is the set of all subsets of the actor indices and $I \subseteq [2^n - 1]$ is a set of the indices of sets in O for which we have observations, so we only accumulate lists for the subsets $o_i,\ i \in I$ not for every subset in O. For $i \in I$ we have N_i realisations $y_{i,1}, ..., y_{i,N_i}$ of the noise free observation model $p^{(P)}(\cdot | h^\dagger[o_i]) \propto 1/|\mathcal{L}[h^\dagger[o_i]]|$.

Proposition 5. *Let $h^\dagger \in \mathcal{H}_{[n]}$ be given and suppose $y_{i,j} \sim p^{(P)}(\cdot | h^\dagger[o_i])$ jointly independent for all $i \in I$ and $j = 1, \dots N_i$. Let $\pi(h|\mathbf{y}_{1:N})$ be the posterior of the $(K, n, \rho)\backslash P$ model with $K \geq \lfloor n/2 \rfloor$, so that*

$$\pi(h|\mathbf{y}_{1:N}) = \int_{Z:h(Z)=h} \int_{\rho=0}^{1} \pi(Z, \rho|\mathbf{y}) d\rho dZ$$

where

$$\pi(Z, \rho|\mathbf{y}_{1:N}) \propto p^{(P)}(\mathbf{y}_{1:N}|h(Z))\pi(Z|\rho)\pi_\rho(\rho).$$

If for each pair of actors $(i, j) \in [n] \times [n]$ with $i \neq j$, there exists $k \in I$ such that $\{i, j\} \subseteq o_k$ then $\pi(h^\dagger|\mathbf{y}_{1:N}) \to 1$ as $\min_{i\in I} N_i \to \infty$ for every $h^\dagger \in \mathcal{H}_{[n]}$.

Proof. Assume $h^\dagger \in \mathcal{H}_{[n]}$ to be the true partial order on $[n]$ behind the observed linear extensions $\mathbf{y}_{1:N}$. We denote $o_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,k_i}\} \subseteq [n]$ a subset of actors with size $|o_i| = k_i \leq n$ where i can take values from $\{1, \ldots, 2^n - 1\}$. The suborders of $h^\dagger$ are $h^\dagger[o_i]$.

Assume we observe N linear extensions $\mathbf{y}_{1:N}$. Assume there are N_i lists $(y_{i,1}, \ldots, y_{i,N_i})$ from each suborder $h^\dagger[o_i]$ where $y_{i,j} = (y_{i,j,1}, \ldots, y_{i,j,k_i}) \in \mathcal{P}_{o_i}$ is a list, $j = 1, \ldots, N_i$. Then the total number of observed lists is $N = \sum_{i=1}^{2^n-1} N_i$. For a general partial order $h \in \mathcal{H}_{[n]}$, and under the error-free observation model, we would have

$$y_{i,j} \sim P(y_{i,j}|h[o_i]) = \frac{1}{|\mathcal{L}[h[o_i]]|}\mathbb{1}_{y_{i,j}\in\mathcal{L}[h[o_i]]}. \tag{F.1}$$

Denote by I the indices of suborders appearing in $\mathbf{y}_{1:N}$,

$$Pr(\mathbf{y}_{1:N}|h) = \prod_{i\in I}\prod_{j=1}^{N_i} P(y_{i,j}|h[o_i]). \tag{F.2}$$

Let $M_{i,r} = \sum_{j=1}^{N_i} \mathbb{1}_{y_{i,j}=r}$ be the number of times a list $r \in \mathcal{P}_{o_i}$ appears in $\{y_{i,1}, \ldots, y_{i,N_i}\}$. If we are considering the posterior for h then, since $\mathbf{y}_{1:N}$ would be linear extensions of h, we would have $y_{i,j} \in \mathcal{L}[h[o_i]], \forall i \in I, j \in [N_i]$, which gives $M_{i,r} = 0$ if $r \notin \mathcal{L}[h[o_i]]$. The $(M_{i,r})_{r\in\mathcal{L}[h[o_i]]}$ values follow the multinomial distribution, such that,

$$(M_{i,r})_{r\in\mathcal{L}[h[o_i]]} \sim \text{Multinomial}(N_i, \frac{1}{|\mathcal{L}[h[o_i]]|}, \ldots, \frac{1}{|\mathcal{L}[h[o_i]]|}). \tag{F.3}$$

This gives that for any h,

$$Pr(\mathbf{y}_{1:N}|h) = \prod_{i\in I}\prod_{r\in\mathcal{L}[h[o_i]]} P(r|h[o_i])^{M_{i,r}}. \tag{F.4}$$

To show model consistency, we would like to show

$$\lim_{N\to\infty}\frac{\pi(h|\mathbf{y}_{1:N})}{\pi(h^\dagger|\mathbf{y}_{1:N})} = \lim_{N\to\infty}\frac{Pr(\mathbf{y}_{1:N}|h)\pi(h)}{Pr(\mathbf{y}_{1:N}|h^\dagger)\pi(h^\dagger)} = \begin{cases} 0 & \text{when } h \neq h^\dagger; \\ 1 & \text{when } h = h^\dagger. \end{cases} \tag{F.5}$$

Our prior $\pi(h) > 0, \forall h \in \mathcal{H}_{[n]}$ in equation (6) as $K \geq \lfloor n/2 \rfloor$ so the set $\{Z \in \mathbb{R}^{n \times K} : h(Z) = h\}$ has positive probability for every $h \in \mathcal{H}_{[n]}$. What remains to be shown is that, if $h \neq h^\dagger$, $\frac{Pr(\mathbf{y}_{1:N}|h)}{Pr(\mathbf{y}_{1:N}|h^\dagger)} \to 0$ when $N_i \to \infty, \forall i \in I$. Let $h[o_i]$ denote a suborder of h corresponding to data lists $\mathbf{y}_i$. Consider

$$\frac{Pr(\mathbf{y}_{1:N}|h)}{Pr(\mathbf{y}_{1:N}|h^\dagger)} = \prod_{i \in I} \frac{P(\mathbf{y}_i|h[o_i])}{P(\mathbf{y}_i|h^\dagger[o_i])} = \prod_{i \in I} R_i. \tag{F.6}$$

Since $h^\dagger$ is the true partial order, we have $y_{k,j} \in \mathcal{L}[h^\dagger[o_k]]$. This gives

$$R_i = \prod_{r \in \mathcal{L}[h^\dagger[o_i]]} \left(\frac{P(r|h[o_i])}{P(r|h^\dagger[o_i])} \right)^{M_{i,r}} = \prod_{r \in \mathcal{L}[h^\dagger[o_i]]} \left(\frac{|\mathcal{L}[h^\dagger[o_i]]|}{|\mathcal{L}[h[o_i]]|} \mathbb{1}_{r \in \mathcal{L}[h[o_i]]} \right)^{M_{i,r}}. \tag{F.7}$$

Given equation (F.3), and taking $h = h^\dagger$, we have $(\frac{M_{i,r}}{N_i})_{r \in \mathcal{L}[h^\dagger[o_i]]} \overset{a.s.}{\to} \frac{1}{|\mathcal{L}[h^\dagger[o_i]]|}$ as $N_i \to \infty$. This gives

$$R_i^{1/N_i} \overset{a.s.}{\to} \prod_{r \in \mathcal{L}[h^\dagger[o_i]]} \left(\frac{|\mathcal{L}[h^\dagger[o_i]]|}{|\mathcal{L}[h[o_i]]|} \mathbb{1}_{r \in \mathcal{L}[h[o_i]]} \right)^{1/|\mathcal{L}[h^\dagger[o_i]]|} \tag{F.8}$$

as $N_i \to \infty$. For convenience, set $A = \mathcal{L}[h^\dagger[o_i]]$ and $B = \mathcal{L}[h[o_i]], i \in I$.

There are three different scenarios for the relation between A and B.

1. If $A \setminus B \neq \emptyset$, then $\mathbb{1}_{r \in B} = 0$ for some $r \in A$. We have $R_i^{1/N_i} \overset{a.s.}{\to} 0$. This gives
$$R_i \overset{a.s.}{\to} 0 \text{ as } N_i \to \infty \text{ when } A \setminus B \neq \emptyset. \tag{F.9}$$
2. If $A = B$, then $\mathcal{L}[h^\dagger[o_i]] = \mathcal{L}[h[o_i]]$ then R_i is equal to 1 (and two partial orders with the same linear extensions are equal so $h^\dagger[o_i] = h[o_i]$).
3. If $A \subset B$ then $|A| < |B|$. In that case
$$R_i^{1/N_i} \overset{a.s.}{\to} \prod_{r \in A} \left(\frac{|A|}{|B|} \right)^{1/|A|} = \left(\frac{|A|}{|B|} \right)^{|A|/|A|} = \frac{|A|}{|B|} < 1, \tag{F.10}$$
so $R_i \overset{a.s.}{\to} 0$ in this case.

From cases 1 and 3 we see that when $h^\dagger[o_i] \neq h[o_i]$, $R_i \overset{a.s.}{\to} 0$ as $N_i \to \infty, i \in I$. Assembling these results,

$$\frac{\pi(h|\mathbf{y}_{1:N})}{\pi(h^\dagger|\mathbf{y}_{1:N})} = \frac{\pi(h)}{\pi(h^\dagger)} \prod_{i \in I} R_i \overset{N \to \infty}{\to} \begin{cases} 0 & \text{if } h[o_i] \neq h^\dagger[o_i] \text{ for any } i \in I; \\ \frac{\pi(h)}{\pi(h^\dagger)} & \text{if } h[o_i] = h^\dagger[o_i] \text{ for all } i \in I. \end{cases} \tag{F.11}$$

What's left to show is that if $h[o_i] = h^\dagger[o_i]$ for every $i \in I$ then we must have $h = h^\dagger$. This is where the last condition on I comes in. We want to show

$$\frac{\pi(h|\mathbf{y}_{1:N})}{\pi(h^\dagger|\mathbf{y}_{1:N})} \overset{N\to\infty}{\to} \begin{cases} 0 & \text{when } h \neq h^\dagger; \\ 1 & \text{when } h = h^\dagger, \end{cases} \tag{F.12}$$

for any $h \in \mathcal{H}_{[n]}$ and the true partial order $h^\dagger \in \mathcal{H}_{[n]}$. This requires the following claim.

Claim 1. ***Condition (*)****: $\forall (i,j) \in [n] \times [n]$ with $i \neq j$, $\exists k \in I$ such that $(i,j) \in o_k$. Every pair in $[n] \times [n]$ appears together at least once in the sets for which we have rank information.*

Equation (F.12) holds for every $h, h^\dagger \in \mathcal{H}_{[n]}$ where $h^\dagger$ is the true partial order, if and only if condition (*) holds.

Proof. If condition (*) holds, and $h[o_k] = h^\dagger[o_k]$ for every $k \in I$ then for every $(i,j) \in [n] \times [n]$, we have $\{i,j\} \subseteq o_k$ for some k. It follows that $h[o_k][\{i,j\}] = h^\dagger[o_k][\{i,j\}]$ so $h[\{i,j\}] = h^\dagger[\{i,j\}] \in \{i \succ j, i \prec j, i \| j\}$ (the suborder $h[o_k][\{i,j\}]$ of a suborder $h[o_k]$ is just the suborder $h[\{i,j\}]$ of the original partial order h). So $h = h^\dagger$.

If condition (*) does not hold for some pair (i,j), we give a counter-example constructed so that transitivity doesn't inform the missing relation. Suppose there is no $k \in I$ such that $\{i,j\} \subseteq o_k$ for some $\{i,j\} \subseteq [n]$. Take h and $h^\dagger$ to be empty partial orders and let $h^{i\succ j}$, $h^{i\prec j}$ and $h^{i\|j}$ be three different partial orders obtained by adding order relations $i \succ j$, $i \prec j$ and $i \| j$ to h. Now $h^{i\succ j}$, $h^{i\prec j}$ and $h^{i\|j}$ agree with $h^\dagger$ on all $o_k, k \in I$ so the posterior ratio converges to the prior ratio. Therefore, if condition (*) is not satisfied, equation (F.12) does not hold for every $h^\dagger$. □

The fact that a condition as strong as Condition (*) is needed is surprising as we might hope transitivity would save us. Whilst this would be the case for any individual partial order, we need it to hold simultaneously for all partial orders, as $h^\dagger$ is unknown. □

Appendix G. Prior Properties: Proportion of Vertex-Series-Parallel Orders and Bucket Orders

The proportion of vertex-series-parallel orders (VSPs) and bucket orders (BOs) under priors are shown in Table G.4.

Number of Actors	Tied Prior %VSP	Tied Prior %BO	Non-Tied Prior %VSP	Non-Tied Prior %BO
5	98.436%	79.187%	90.548%	64.787%
10	83.910%	50.833%	49.931%	31.803%
15	64.682%	37.063%	34.776%	21.909%

Table G.4: The probability a random partial order under the tied or non-tied prior is a VSP or BO.

Appendix H. An Alternative Tied Scenario

Ties may be taken to impose an extra constraint in the observation model: tied actors 'must appear together'. Essentially, tied actors enter a list as if they were a single composite actor. In this setup tied actors are treated as a single node in a linear extension with an "internal" permutation over tied nodes taken uniformly at random. In Figure 8, instead of regarding $2\|_{h_0^*}3$ and $2\|_{h_0^*}4$, one may treat actors $(3,4)$ as one node. There will then only be 4 linear extensions as are shown in Figure H.18. This construction does not seem relevant for our application, but may be of interest for further work.

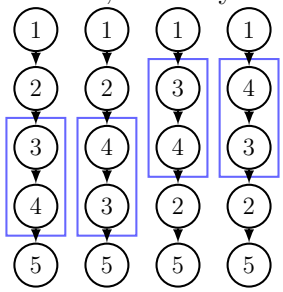


Figure H.18: All four linear extensions for the tied partial order in Figure 8.

Appendix I. The Plackett-Luce Model

This appendix complements a similar presentation introducing the Mallows noise model in Section 3.2.2.

The Plackett-Luce model, presented independently in Plackett (1975) and Luce (1959), is a widely used model for rank data, straightforwardly fitting rank-subset data. The Plackett-Luce model enjoys extensive popularity because of its interpretability and tractability. It is projective so applies straightforwardly to either a complete ranking of all actors, a partial ranking

of a subset of actors or the partial ranking of top actors (Bradley and Terry, 1952). The literature extends the Plackett-Luce model to different inference schemes. For example, Guiver and Snelson (2009) proposed an efficient Bayesian inference scheme for the Plackett-Luce model by applying expectation propagation and Caron and Doucet (2012) gives an efficient MCMC scheme targeting a class of Placket-Luce posterior distributions.

Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a set of actor weights. The Plackett-Luce model defines the probability for ranking $y \in \mathcal{P}_n$ as

$$p(y|\alpha) = \prod_{k=1}^{n} \frac{e^{\alpha_k}}{\sum_{m=k}^{n} e^{\alpha_m}}.$$

We can modify this usual setup for our purposes, as we can have *separate* actor weights for *each* list and integrate over these. This will seem counter-intuitive but is needed to preserve the structure of the underlying noise-free model. For $j = 1, ..., N$, let $\alpha_j = (\alpha_{j,1}, ..., \alpha_{j,n}) \in \mathbb{R}^n$ denote the (logit-scale) weights, so that $\alpha_{j,k}$ is the weight for actor $k = 1, ..., n$ in list j (if they are present). The actor-ordered weights for actors $y_{j,1}, \ldots, y_{j,n_j}$ in data-list $j = 1, ..., N$ are then $\alpha_{j,y_j} = \{\alpha_{j,y_{j,1}}, ..., \alpha_{j,y_{j,n_j}}\}$. The Plackett-Luce model assigns probability mass function

$$p^{(PL)}(y_j|\alpha_j) = \prod_{k=1}^{n_j} \frac{e^{\alpha_{j,y_{j,k}}}}{\sum_{m=k}^{n_j} e^{\alpha_{j,y_{j,m}}}}. \tag{I.1}$$

The Plackett-Luce model is marginally consistent. We can set the prior for α to be $\pi(\alpha|h, \sigma) = \prod_{j=1}^{N} \pi(\alpha_j|h, \sigma)$ where

$$\alpha_j|h, \sigma \propto \mathcal{MVN}(\alpha_j; 0_K, \sigma^2 \mathbb{I}_n) \mathbb{1}_{\{R(\alpha_j) \in \mathcal{L}[h]\}}. \tag{I.2}$$

If we denote the rank vector for α_j as $R(\alpha_j)$ where $R(\alpha_j)_i = \sum_{k=1}^{n} \mathbb{1}_{\{\alpha_{j,k} \leq \alpha_{j,i}\}}$, $R(\alpha_j) \sim \mathcal{U}(\mathcal{L}[h])$ under this prior distribution. In addition, we show that the normalising constant in equation I.2 is $|\mathcal{L}[h]|/n!$. There has been literature that considers the Plackett-Luce model on tied ranking data (Turner et al., 2020) and tied actors (Henderson, 2022; Baker and Scarf, 2020).

Though theoretically appealing, the setup requires $\alpha_j \in \mathbb{R}^n$ for $j = 1, \ldots, N$, which is a large space to explore. We implemented sudo-marginal method to effectively integrate over α, where the approximated list of linear extensions are generated via the Bubley-Dyer algorithm (Bubley and Dyer,

1999). However, the $(PDP, \rho)\backslash M$ model (as discussed in section 3.2.2) still shows a large computational advantage in comparison and seems to show no disadvantages as a model for the data. In addition, Liu et al. (2019b) shows that the Mallow's model is preferred over Plackett-Luce model. The Plackett-Luce model produces a less certain concensus estimation (higher MSE) in their potato ranking example. A similar result is obtained in our model comparison between the Plackett-Luce mixture and Mallows mixture models in table 3. We therefore choose the Mallow's noise observation model instead of Plackett-Luce in this paper, but we include the Plackett-Luce Mixture for model comparison in section 5.

Appendix J. MCMC Algorithm

This section summarises the MCMC algorithm used to sample posterior distributions to the parameters of interest. We implemented the partial order with ties model with two noisy observation models - the Mallow's observation model (M) and the queue-jumping model (Q).

Algorithm 2: MCMC for the Partial Order with Ties Model

input: Z, ρ, S, θ, p

Function `potie(`Z, S`)`**:**

- h $= \mathbf{0}_{n\times n}$;
- K $=$ `ncol` (Z);
- **foreach** $i, j \in [n] \times [n], i \neq j$ **do**
  - **if** $Z_{i,m} > Z_{j,m}\, \forall m \in [K]$ **then**
    - $h_{i,j} = 1$;
- **foreach** $c \in [|S|]$ **do**
  - $h_{i,j} = 0\, \forall i, j \in S_c$;
- **return** h

Update for K & Z

Sample $K^* \in \{K-1, K+1\}$ with $\rho_{K,K+1} = \rho_{K,K-1} = 0.5$;

if $K^* > 0$ **then**

- **if** $K^* = K + 1$ **then**
  - **foreach** $c \in \{1, \ldots, |S|\}$ **do**
    - Set $Z^*_{\cdot,1:K} = Z$;
    - Sample Z^*_{c,K^*} such that $Z^*_{c,\cdot} \sim MVN(\mathbf{0}_{K^*}, \Sigma_\rho)$
- **else**
  - Set $Z^* = Z_{\cdot,1:K^*}$
- $h^* \leftarrow$ `potie(`Z^*, S`)`;
- **if** *model=M* **then**
  - $\phi_K = \frac{\pi(K^*)p^{(M)}(\mathbf{y}|h(Z^*),\theta)}{\pi(K)p^{(M)}(\mathbf{y}|h(Z),\theta)}$
- **else**
  - $\phi_K = \frac{\pi(K^*)p^{(Q)}(\mathbf{y}|h(Z^*),p)}{\pi(K)p^{(Q)}(\mathbf{y}|h(Z^*),p)}$
- Sample $u_K \sim \mathcal{U}(0, 1)$;
- **if** $\phi_K > u_K$ **then**
  - $Z \leftarrow Z^*$; $K \leftarrow K^*$; $h \leftarrow h^*$;

Algorithm 3: MCMC for the Partial Order with Ties Model (part 2)

Update for S & Z

Sample $j \in [n]$;

Set $Z^* \leftarrow Z$;

Simulate $Z^*_{|S^{-j}|+1,\cdot} \sim MVN(\mathbf{0}_K, \Sigma_\rho)$;

if *model=M* **then**

  Sample a new cluster c for $j \in \{1, 2, \ldots, |S^{-j}| + 2\}$ with probability proportional to

$$((n_1^{-j} - \phi_\alpha)p^{(M)}(\mathbf{y}|h(Z_{S^{-j}\cup\{j\in S_1^{-j}\}}), \theta), \ldots, (n^{-j}_{|S^{-j}|} - \phi_\alpha)p^{(M)}(\mathbf{y}|h(Z_{S^{-j}\cup\{j\in S^{-j}_{|S^{-j}|}\}}), \theta),$$

$$(\eta_b + \eta_a|S^{-j}|)p^{(M)}(\mathbf{y}|h(Z^*_{S^{-j}\cup\{j\in S^{-j}_{|S^{-j}|+1}\}}), \theta), (\eta_b + \eta_a|S^{-j}|p^{(M)}(\mathbf{y}|h(Z), \theta));$$

else

  Sample a new cluster c for $j \in \{1, 2, \ldots, |S^{-j}| + 2\}$ with probability proportional to

$$((n_1^{-j} - \phi_\alpha)p^{(Q)}(\mathbf{y}|h(Z_{S^{-j}\cup\{j\in S_1^{-j}\}}, p), \ldots, (n^{-j}_{|S^{-j}|} - \phi_\alpha)p^{(Q)}(\mathbf{y}|h(Z_{S^{-j}\cup\{j\in S^{-j}_{|S^{-j}|}\}}), p),$$

$$(\eta_b + \eta_a|S^{-j}|)p^{(Q)}(\mathbf{y}|h(Z^*_{S^{-j}\cup\{j\in S^{-j}_{|S^{-j}|+1}\}}), p), (\eta_b + \eta_a|S^{-j}|)p^{(Q)}(\mathbf{y}|h(Z), p))$$

if $c = |S^{-j}| + 1$ **then**

  $S \leftarrow S^{-j} \cup \{j \in S^{-j}_{|S^{-j}|+1}\}; Z \leftarrow Z^*$

if $c \in \{1, \ldots, |S^{-j}|\}$ **then**

  $S \leftarrow S^{-j} \cup \{j \in S_c^{-j}\}; Z \leftarrow Z_{S^{-j}\cup\{j\in S_c^{-j}\}}$

Update for Z

Set $Z^* \leftarrow Z$;

Sample $c \in \{1, \ldots, |S|\}$, $k \in \{1, \ldots, K\}$ and simulate $Z^*_{c,k}$ such that $Z^*_{c,\cdot} \sim MVN(Z_{c,\cdot}, \Sigma_\rho)$;

$h^* \leftarrow$ `potie(`Z^*, S`)`;

if *model* $= M$ **then**

  $\phi_Z = \frac{\pi(Z^*|\rho,K,S)p^{(M)}(\mathbf{y}|h(Z^*),\theta)}{\pi(Z|\rho,K,S)p^{(M)}(\mathbf{y}|h(Z),\theta)}$

else

  $\phi_Z = \frac{p^{(Q)}(\mathbf{y}|h(Z^*),p)\pi(Z^*|\rho,K,S)}{p^{(Q)}(\mathbf{y}|h(Z),p)\pi(Z|\rho,K,S)}$

Sample $u_Z \sim \mathcal{U}(0, 1)$;

if $\phi_Z > u_Z$ **then**

  $Z \leftarrow Z^*; h \leftarrow h^*$.

Algorithm 4: MCMC for the Partial Order with Ties Model (part 3)

Update for ρ

Sample $\delta_\rho \sim \mathcal{U}(w_\rho, \frac{1}{w_\rho})$; ▷ $w_\rho \in (0,1)$ is a constant

Set $\rho^* = 1 - \delta_\rho(1-\rho)$ then $\phi_\rho = \frac{\pi(\rho^*)\pi(Z|\rho^*,K,S)}{\pi(\rho)\pi(Z|\rho,K,S)\delta_\rho}$;

Sample $u_\rho \sim \mathcal{U}(0,1)$;

if $u_\rho < \phi_\rho$ & $\rho^* < 1$ **then**

  $\rho \leftarrow \rho^*$.

Update for p

if *model=Q* **then**

  Let $r = log(\frac{p}{1-p})$ and sample $r^* \sim \mathcal{N}(r,1)$;

  Set $p^* \leftarrow \frac{1}{1+e^{-r^*}}$

  $\phi_p = \frac{\pi(p^*)p^{(Q)}(\mathbf{y}|h(Z),p^*)}{\pi(p)p^{(Q)}(\mathbf{y}|h(Z),p)}$;

  Sample $u_p \sim \mathcal{U}(0,1)$;

  if $u_p < \phi_p$ **then**

    $p \leftarrow p^*$.

Update for θ

if *model = M* **then**

  Sample $\theta^* \sim \mathcal{N}(\theta, 0.5)$

  if $\theta^* > 0$ **then**

    $\phi_\theta = \frac{p^{(M)}(\mathbf{y}|h(Z),\theta^*)\pi_\theta(\theta^*)}{p^{(M)}(\mathbf{y}|h(Z),\theta)\pi_\theta(\theta)}$;

    Sample $u_\theta \sim \mathcal{U}(0,1)$;

    if $u_\theta < \phi_\theta$ **then**

      $\theta \leftarrow \theta^*$.

Appendix K. Application - the 'Royal Acta' (Bishops) Data

Appendix K.1. The Data Lists

Witness lists between 1131 and 1133:

[1] 3 4 8
[2] 1 13 12 14 7 11 6
[3] 1 13 7 14 10
[4] 14 11
[5] 1 13 7
[6] 1 13 7 8 15 9
[7] 3 8 4 2
[8] 13 1 5
[9] 13 1 5
[10] 13 7 5
[11] 13 1 5
[12] 1 13 5
[13] 1 7 13 14 5 10 15 8
[14] 15 8
[15] 12 7 9
[16] 12 13 7 10 9 11 6
[17] 3 4 2
[18] 4 3
[19] 7 13 12 1
[20] 1 13 7 5 14 9
[21] 6 1 7

1: *Roger, Bishop of Salisbury*
2: *Richard, Bishop of Bayeux*
3: *John, Bishop of Lisieux*
4: *Ouen, Bishop of Evreux*
5: *Bernard, Bishop of St David's*
6: *Everard, bishop of Norwich*
7: *Alexander, Bishop of Lincoln*
8: *John, Bishop of Sees*
9: *Seffrid, Bishop of Chichester*
10: *John, Bishop of Rochester*
11: *Simon, Bishop of Worcester*
12: *Gilbert, the Universal, bishop of London*
13: *Henry, de Blois, Bishop of Winchester, 1129-1171*
14: *Robert, de Bethune, Bishop of Hereford*
15: *Algar, Bishop of Coutances*

Witness lists between 1100 and 1103:

[1] 2 5 7
[2] 2 8 5 3 1 6 4 7
[3] 2 5 6
[4] 2 9
[5] 5 8
[6] 5 9
[7] 1 2 3 5 4 6 7
[8] 5 8
[9] 1 5
[10] 8 7
[11] 8 7
[12] 2 4 1 3 8
[13] 2 1

1: *Gundulf, bishop of Rochester*
2: *Maurice, bishop of London*
3: *Robert, de Limesey, bishop of Chester*
4: *Ralph, bishop of Chichester*
5: *Robert, Bloet, bishop of Lincoln*
6: *Samson, bishop of Worcester*
7: *Ranulf, Flambard, bishop of Durham*
8: *William, Giffard, bishop of Winchester, 1100-1129*
9: *Roger, Bishop of Salisbury*

Appendix K.2. Additional Results

Four experiments are conducted on the 'Royal Acta' (Bishop) 1131-1133 and 1100-1103 data lists for both the $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models. Each MCMC chain was run for $1e5$ iterations. We record every $2n$ steps and chose a burn-in period of $500(\times 2n)$ for all chains. The list of experiments and effective sample sizes to each key parameter is shown in table K.5.

		Effective Sample Sizes (ESSs)			
Time Period	Model	K	ρ	θ	p
1131-	$(PDP, \rho)\backslash M$	240.60	136.32	1318.60	-
1133	$(PDP, \rho)\backslash Q$	104.32	174.88	-	873.53
1100-	$(PDP, \rho)\backslash M$	605.91	340.02	899.20	-
1103	$(PDP, \rho)\backslash Q$	444.18	358.95	-	1917.23

Table K.5: The effective sample sizes (ESSs) to the key parameters for each experiment.

Appendix K.2.1 and Appendix K.2.2 present the posterior distributions and trace plots for these key parameters respectively.

Appendix K.2.1. Posterior Distributions

This section displays some additional result on prior and posterior distributions. Figure K.19 gives both the prior and posterior distributions on the number of clusters (on actors) and partial order depths.

We further present cluster samples from the MCMC for the $(PDP, \rho)\backslash Q$ model in the heatmap in Figure K.20. The heatmap for $(PDP, \rho)\backslash M$ is in figure 15.

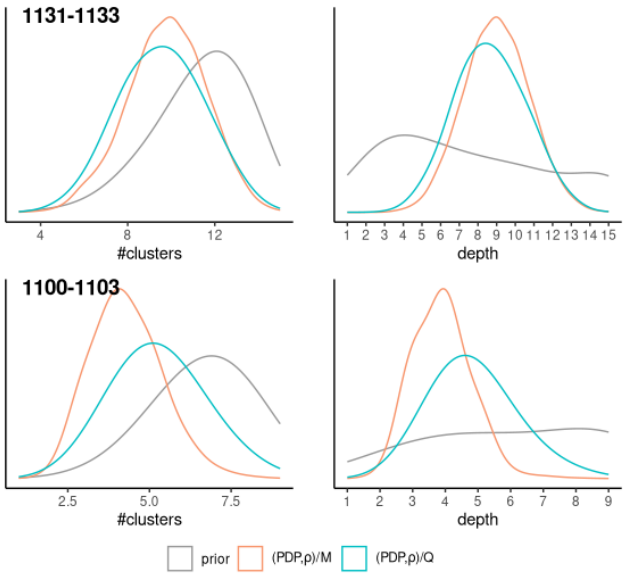


Figure K.19: The prior/posterior distributions for the number of clusters (left) and partial order depths (right) for the $(PDP, \rho)\backslash M$ (pink) and $(PDP, \rho)\backslash Q$ (blue) model on periods 1131-1133 (upper) and 1100-1103 (lower). The prior distributions are colored in gray.

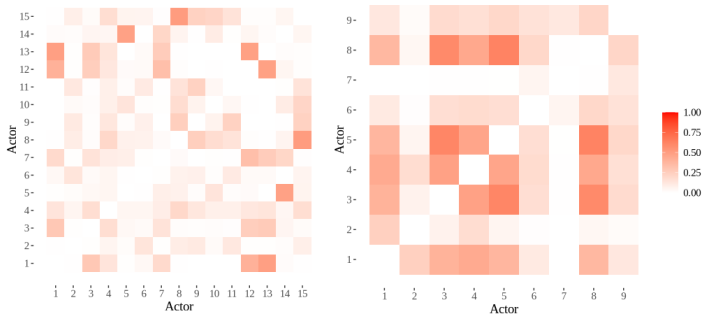


Figure K.20: The 'heatmaps' for the estimated clusters for the $(PDP, \rho)\backslash Q$ model for time periods 1131-1133 (left) and 1100-1103 (right). Each block indicates the posterior probability for a pair of actors to be in the same cluster. We label higher posterior probability red and lower posterior probability white.

Appendix K.2.2. Key Parameter Trace Plots

The trace plots to the key parameters are shown in Figure K.21 for the $(PDP, \rho)\backslash M$ model and Figure K.22 for the $(PDP, \rho)\backslash Q$ model. All trace

plots display great mixing and convergence.

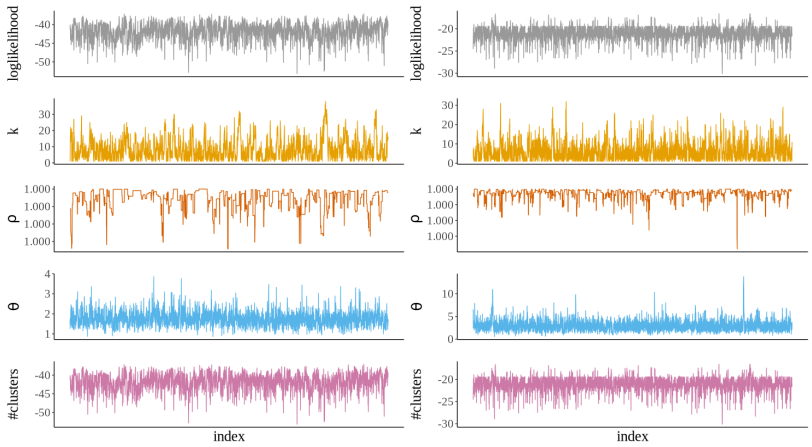


Figure K.21: The trace plots for the $(PDP, \rho)\backslash M$ model on the 1131-1133 (left) and 1100-1103 (right) 'royal acta' (bishop) witness list data. We present the trace plot for the log-likelihood (grey), the number of clusters (pink) and key parameters (latent matrix column dimension parameter K - orange, the depth control parameter ρ - red and the dispersion parameter for the Mallow's observation model θ - blue). The burn-in periods are removed from the trace plots.

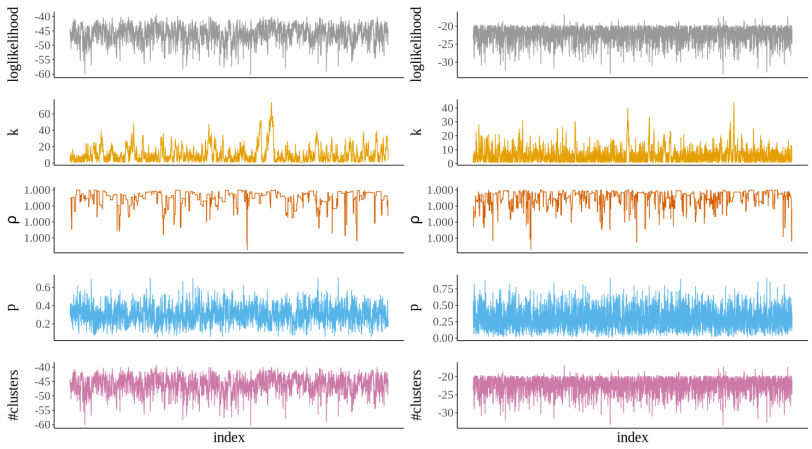


Figure K.22: The trace plots for the $(PDP, \rho)\backslash Q$ model on the 1131-1133 (left) and 1100-1103 (right) 'royal acta' (bishop) witness list data. We present the trace plot for the log-likelihood (grey), the number of clusters (pink) and key parameters (latent matrix column dimension parameter K - orange, the depth control parameter ρ - red and the error probability in the queue-jumping observation observation model p - blue). The burn-in periods are removed from the trace plots.

Appendix K.3. Sensitivity Analysis on K Prior

This section investigates the models' ($(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$) sensitivity to the priors on the latent matrix column dimension parameter K. According to Muir Watt (2015), $k = \lfloor n/2 \rfloor$ is sufficient to recover any unknown partial order. We explicit a geometric prior distribution on K as $k \sim Geo(p_k)$, where p_k is a pre-determined hyperparameter - such that the means of K are in $\{n, n/2, n/4\}$. The priors on other parameters are adjusted accordingly to obtain a relatively flat depth distribution. Table K.6 records the prior distributions on K for different sensitivity analysis on the real data.

We display the prior and posterior distributions for K in Figure K.23 and Figure K.24 for $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models respectively. When $\bar{k} \geq n/2$, the posterior distributions are close to the prior distributions in both periods - indicating their high sensitivity to the prior choice. However, the posterior distributions on K are updated to higher K values for both the $(PDP, \rho)\backslash M$ and $(PDP, \rho)\backslash Q$ models in period 1131-1133. No similar behaviour is observed in period 1100-1103, which is partially a result of data limitation.

Given the result by Muir Watt (2015), a $k \geq n/2$ is sufficient for the

Time Period	Model	K-Prior	PDP-Prior	ρ-Prior
1131-1133 ($n = 15$)	$(PDP, \rho)\backslash M$	$Geo(0.06)(\bar{k} = n)$	$PDP(\eta_b = 3, \eta_a = 0.7)$	$Beta(1, 1/6)$
	$(PDP, \rho)\backslash Q$	$Geo(0.06)(\bar{k} = n)$	$PDP(\eta_b = 3, \eta_a = 0.7)$	$Beta(1, 1/6)$
	$(PDP, \rho)\backslash M$	$Geo(0.12)(\bar{k} = n/2)$	$PDP(\eta_b = 6, \eta_a = 0.6)$	$Beta(1, 1/4)$
	$(PDP, \rho)\backslash Q$	$Geo(0.12)(\bar{k} = n/2)$	$PDP(\eta_b = 6, \eta_a = 0.6)$	$Beta(1, 1/4)$
	$(PDP, \rho)\backslash M$	$Geo(0.21)(\bar{k} = n/4)$	$PDP(\eta_b = 12, \eta_a = 0.55)$	$Beta(1, 1/4)$
	$(PDP, \rho)\backslash Q$	$Geo(0.21)(\bar{k} = n/4)$	$PDP(\eta_b = 12, \eta_a = 0.55)$	$Beta(1, 1/4)$
1100-1103 ($n = 9$)	$(PDP, \rho)\backslash M$	$Geo(0.1)(\bar{k} = n)$	$PDP(\eta_b = 5, \eta_a = 0.4)$	$Beta(1, 1/4)$
	$(PDP, \rho)\backslash Q$	$Geo(0.1)(\bar{k} = n)$	$PDP(\eta_b = 5, \eta_a = 0.4)$	$Beta(1, 1/4)$
	$(PDP, \rho)\backslash M$	$Geo(0.18)(\bar{k} = n/2)$	$PDP(\eta_b = 5, \eta_a = 0.4)$	$Beta(1, 1/3)$
	$(PDP, \rho)\backslash Q$	$Geo(0.18)(\bar{k} = n/2)$	$PDP(\eta_b = 5, \eta_a = 0.4)$	$Beta(1, 1/3)$
	$(PDP, \rho)\backslash M$	$Geo(0.31)(\bar{k} = n/4)$	$PDP(\eta_b = 5, \eta_a = 0.4)$	$Beta(1, 1/3)$
	$(PDP, \rho)\backslash Q$	$Geo(0.31)(\bar{k} = n/4)$	$PDP(\eta_b = 5, \eta_a = 0.4)$	$Beta(1, 1/3)$

Table K.6: Sensitivity analysis on K-priors.

latent matrix model to recover any unknown partial order. Having a large mass on the $k \geq n/2$ values in posterior will give the model enough freedom to explore the space of partial order without much constraints. We therefore conclude a choice of K-prior with $\bar{K} = n/2$ is the most optimal and present the corresponding result in section 5.3.

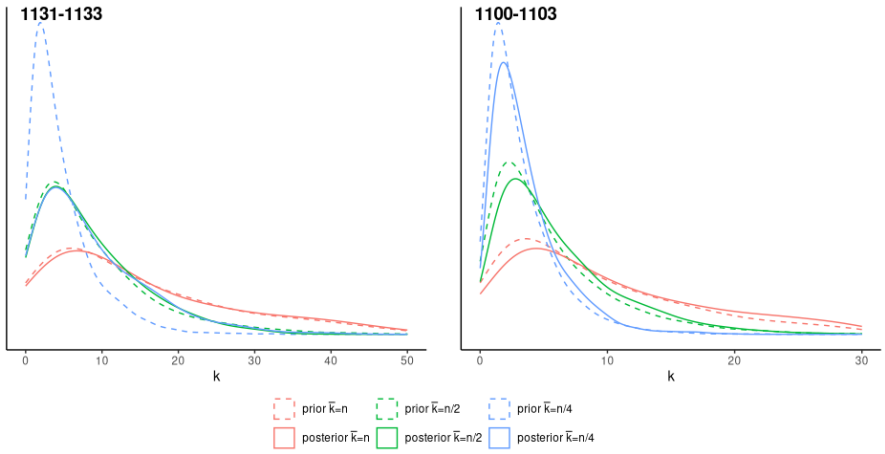


Figure K.23: The prior (dashed) and posterior (solid) distributions for K from the $(PDP, \rho)\backslash M$ model for sensitivity analysis on K-prior. We explicit the K priors to have means of n (red), $n/2$ (green) or $n/4$ (blue) (details in table K.6). The result for time period 1131-1133 is displayed on left, and 1100-1103 on right.

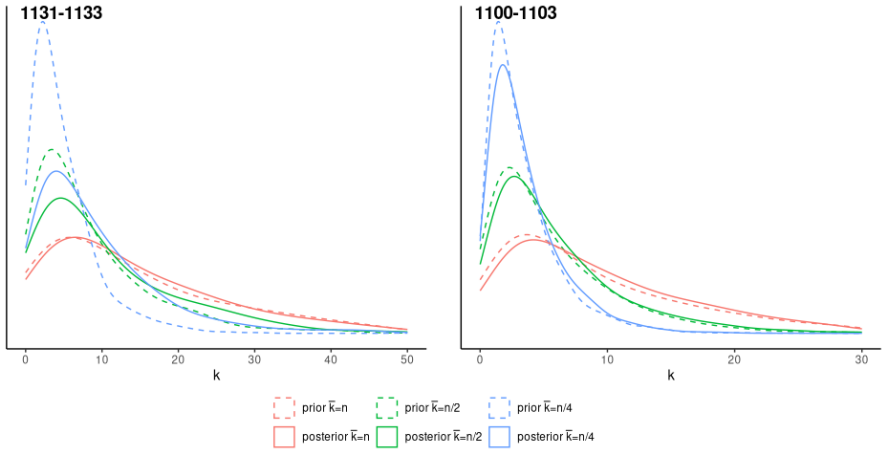


Figure K.24: The prior (dashed) and posterior (solid) distributions for K from the $(PDP, \rho)\backslash Q$ model for sensitivity analysis on K-prior. We explicit the K priors to have means of n (red), $n/2$ (green) or $n/4$ (blue) (details in table K.6). The result for time period 1131-1133 is displayed on left, and 1100-1103 on right.

Appendix K.4. Synthetic Reconstruction Test

The synthetic reconstruction test detailed in section 5.2 is designed to test the model's ability to reconstruct the 'true' partial order given data-lists under different observation models. We consider four scenarios for each time period.

- Simulation 1: error-free, $y_i \sim \mathcal{U}(\mathcal{L}[h^{(T)}[y_i^{obs}]]), \forall i \in [N]$.
- Simulation 2: data-list with random error. For $i \in [N]$, we simulate $y_i \sim \mathcal{U}(\mathcal{L}[h^{(T)}[y_i^{obs}]])$ from the noise-free model; we select a pair of actors $a, b \in y_i$ uniform at random, $a \neq b$, and put them in the same order as they appear in the data. If $a \prec b$ in y_i but $a \succ b$ in y_i^{obs} then exchange the positions of a and b in y_i leaving all else unchanged.
- Simulation 3: data-list with Mallow's error given θ^*,

$$y_i \sim p^{(M)}(\cdot|h^{(T)}[o_i], \theta^*), \forall i \in [N].$$

- Simulation 4: data-list with queue-jumping error given p^*,

$$y_i \sim p^{(Q)}(\cdot|h^{(T)}[o_i], p^*), \forall i \in [N].$$

The simulations are generated based on the true partial order in Figure K.25 for period 1131-1133 and the true partial order in Figure K.28 for period 1100-1103. For the 1131-1133 structured data, we show the concensus orders $h^{con}(\epsilon = 0.2)$ from the $(PDP, \rho)\backslash M$ model in Figure K.26 and the $(PDP, \rho)\backslash Q$ model in Figure K.27. For the 1100-1103 structured data, we show the concensus orders $h^{con}(\epsilon = 0.2)$ from the $(PDP, \rho)\backslash M$ model in Figure K.29 and the $(PDP, \rho)\backslash Q$ model in Figure K.30. The threshold $\epsilon = 0.2$ is chosen based on the elbows in the ROC curves (Figures 11 and 12).

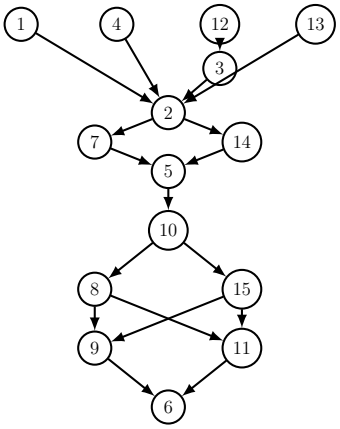


Figure K.25: The true partial order for list simulation (1131-1133 structured data).

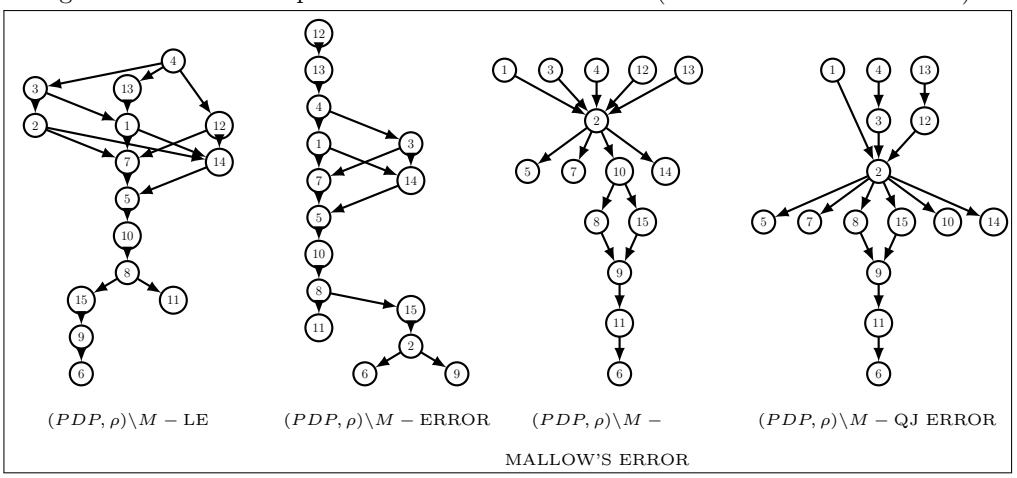


Figure K.26: The concensus orders for the $(PDP, \rho)\backslash M$ model on the synthetic data with 1131-1133 structure (simulation 1-4 from left to right). We conclude an edge if such order relation has more than 0.2 posterior probability (inferred from section 5.2). An edge is colored red if it has more than 0.9 posterior probability.

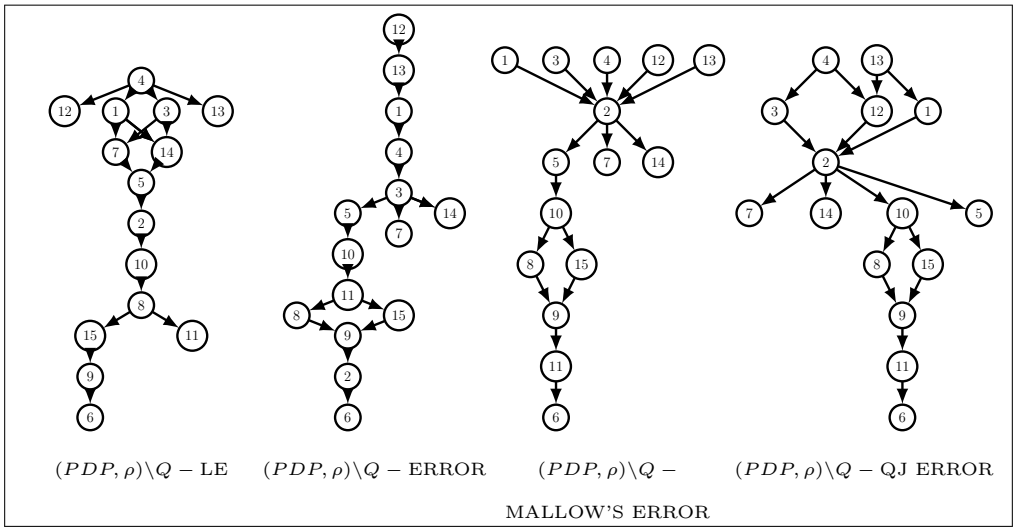


Figure K.27: The concensus orders for the $(PDP, \rho)\backslash Q$ model on the synthetic data with 1131-1133 structure (simulation 1-4 from left to right). We conclude an edge if such order relation has more than 0.2 posterior probability (inferred from section 5.2). An edge is colored red if it has more than 0.9 posterior probability.

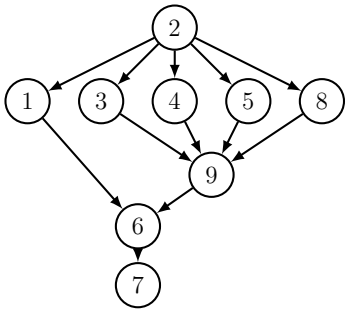


Figure K.28: The true partial order for list simulation (1131-1133 structured data).

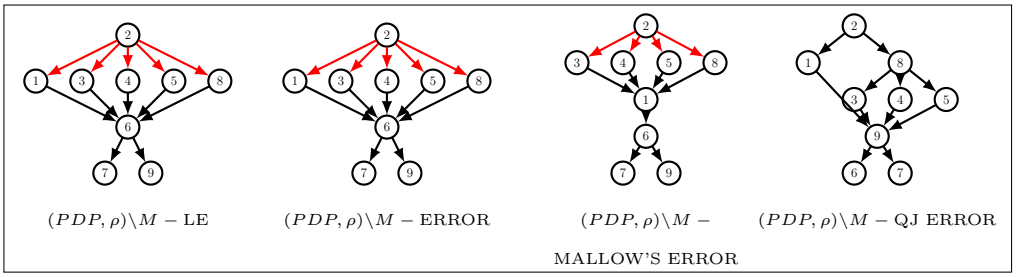


Figure K.29: The concensus orders for the $(PDP, \rho)\backslash M$ model on the synthetic data with 1100-1103 structure (simulation 1-4 from left to right). We conclude an edge if such order relation has more than 0.2 posterior probability (inferred from section 5.2). An edge is colored red if it has more than 0.9 posterior probability.

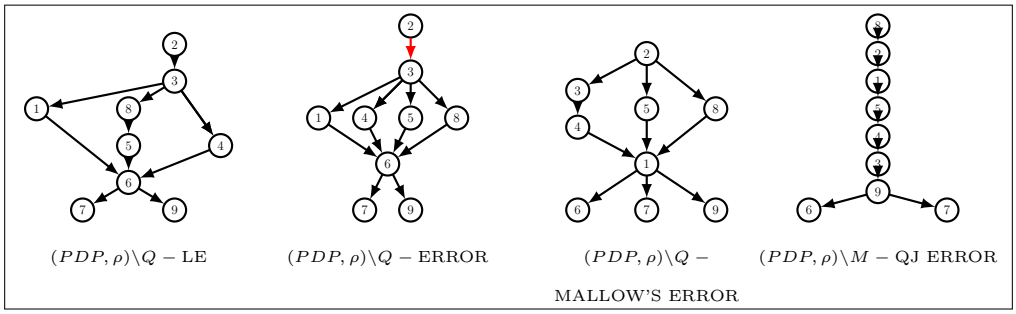


Figure K.30: The concensus orders for the $(PDP, \rho)\backslash Q$ model on the synthetic data with 1100-1103 structure (simulation 1-4 from left to right). We conclude an edge if such order relation has more than 0.2 posterior probability (inferred from section 5.2). An edge is colored red if it has more than 0.9 posterior probability.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Non-Parametric Bayesian Inference for Partial Orders with Ties from Rank Data observed with Mallows Noise
Publication Status	□Published □Accepted for Publication ■Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Jiang, C. and Nicholls, G. K. (2024). Non-Parametric Bayesian Inference for Partial Orders with Ties from Rank Data observed with Mallows Noise. https://arxiv.org/abs/2408.14661.

Student Confirmation

Student Name:	Chuxuan (Jessie) Jiang		
Contribution to the Paper	Joint development of models and theory (with second author), performed all simulation and model experiments, wrote majority of manuscript.		
Signature 姜楚璇		Date	06/08/2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof Geoff Nicholls		
Supervisor comments		
Signature G. K. Nicholls.	Date	06/08/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Bayesian Inference for Vertex-Series-Parallel Orders

Jiang, C., Nicholls, G.K. and Lee, J.E., 2023, July. Bayesian inference for vertex-series-parallel partial orders. *In Uncertainty in Artificial Intelligence* (pp. 995-1004). PMLR.

Bayesian Inference for Vertex-Series-Parallel Partial Orders

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

Abstract

Partial orders are a natural model for the social hierarchies that may constrain "queue-like" rank-order data. However, the computational cost of counting the linear extensions of a general partial order on a ground set with more than a few tens of elements is prohibitive. Vertex-series-parallel partial orders (VSPs) are a subclass of partial orders which admit rapid counting and represent the sorts of relations we expect to see in a social hierarchy. However, no Bayesian analysis of VSPs has been given to date. We construct a marginally consistent family of priors over VSPs with a parameter controlling the prior distribution over VSP depth. The prior for VSPs is given in closed form. We extend an existing observation model for queue-like rank-order data to represent noise in our data and carry out Bayesian inference on "Royal Acta" data and Formula 1 race data. Model comparison shows our model is a better fit to the data than Plackett-Luce mixtures, Mallows mixtures, and "bucket order" models and competitive with more complex models fitting general partial orders.

1 INTRODUCTION

Rank-order data are lists in which a set of elements are ranked. They are analysed in a wide range of areas, including decision support [Beichl et al., 2017], medical research [Beerenwinkel et al., 2007] and chemistry [Pavan and Todeschini, 2008]. We classify ranking methods into two categories - total order ranking and partial order ranking.

Total order models seek a ranking of the elements of the ground set (in our setting, the labels of a group of actors we want to rank) that is "central" to the rank-lists in the data. These models are suitable when we believe that an order relation exists between every pair of actors. The Mallows model [Mallows, 1957], the Plackett-Luce model [Plackett, 1975, Luce, 1959] and related mixture models are models for total orders. However, the real-world relations we are looking to recover may be weaker than a total order: perhaps relations between pairs of actors are not simply weak or uncertain, they don't actually exist. We expect this for some precedence relations that define some social hierarchies.

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

If we want to learn social-order relations between actors by observing their behavior, then the elements of the model we fit should correspond to elements of reality: if relations are incomplete then we should fit a *partial order*. A partial order $h = \{[n], \prec_h\}$ is a (possibly incomplete) set of binary order relations $\prec_h$ over a "ground set" of actors with labels $[n] = \{1, \ldots, n\}$. Our data are records of queues of actors constrained by a social hierarchy h, which is unknown. If we see enough queue realisations we can identify the hierarchy. In this setting the queue is just a *linear extension* (LE) of h, that is, a permutation of actors in $[n]$ that doesn't put an actor ahead of someone of higher precedence.

Partial orders are widely used as a ranking summary tool, or to support efficient computation. For example, partial orders and LEs support efficient computation of marginals in Bayesian networks [Cano et al., 2011, Smail, 2018]. By contrast, in our work the partial order h is the object of inference, so it is a parameter in the likelihood: the data are LEs and the likelihood depends on the number of LEs of h. Counting LEs is an #P-complete task [Brightwell and Winkler, 1991], so work to date in this setting either restricts the class of partial orders [Mannila and Meek, 2000, Gionis et al., 2006, Mannila, 2008] to orders which admit fast counting or works with orders of manageable size [Beerenwinkel et al., 2007, Sakoparnig and Beerenwinkel, 2012, Nicholls and Muir Watt, 2011, Nicholls et al., 2022]. This approach does not scale well with n. We follow Mannila and Meek [2000] and work with *vertex-series-parallel* partial orders (VSPs). These orders are a sub-class of partial orders which can be formed by repeated series and parallel operations on smaller VSPs. They include *bucket orders* [1] as a special case. Valdes et al. [1979] represent VSPs using binary decomposition trees (BDTs). These support counting in a time linear in n [Wells, 1971] and scale to VSPs with hundreds of actors.

VSPs are a well characterised combinatorial class [Wells, 1971, Valdes et al., 1979]. However, work on fitting VSPs to data is limited. Mannila and Meek [2000] learn VSPs from LEs by adapting a greedy search over VSPs. However, there is to date no Bayesian inference and hence no one has given a prior probability distributions over VSPs with good properties for inference. Mannila [2008] gave Bayesian inference for bucket orders, a subclass of VSP, and Sakoparnig and Beerenwinkel [2012] for partial orders, a super-set that doesn't scale.

Contributions. This is the first Bayesian inference for VSPs from LEs and presents some useful new priors and likelihoods. VSPs are equivalent to "transitively closed" Directed Acyclic Graphs (DAGs); when we specify priors over objects of this sort we have to be careful to ensure the prior doesn't impose unwanted weighting and inconsistency.

We specify a prior and give its probability mass function in a simple closed form. Our prior (Sec. 2) is marginally consistent. This property (defined in Definition 1 below) is needed for the model to make sense in our setting. Our prior also represents the information available well: it is non-informative with respect to VSP depth, one of the most interesting summary statistics for a social hierarchy.

Our new observation model (Sec. 3) generalises earlier models for observation noise in records for queue-like data and has a natural physical interpretation in terms of "queue jumping" and "arriving late".

We give MCMC algorithms in Appendix C which target the VSP posterior. We carry out model comparison with the Plackett-Luce and Mallows mixture models in Appendix E.1. We further compare our model with a simple restriction to Bucket-Order models in Appendix E.2 and we compare it with a

[1] Actors are grouped in buckets - every actor is ordered with respect to actors in other groups, and any pair of actors in the same group are incomparible.

more general partial order model [Nicholls et al., 2022] in Appendix E.3.

Finally, our reconstruction of relations between witnesses appearing in Royal Acta (Sec. 5.2) is new. Historians are interested in these relations, but it wasn't possible to reconstruct them all till now as the partial orders were too big to count their LEs (Nicholls et al. [2022] analyse a subset, working in a time-series setting; we give timing comparisons in Appendix F). Our models are relevant for any ranking problem where relations may be partial: in Appendix D.2 we fit Formula 1 race results for the 2021 season. These data show the same preference for our model over other models.

1.1 BACKGROUND

A partial order $h = \{V, \prec_h\}$ is a binary relation[2] $\prec_h$ over a "ground set" of actors V. In our setting the actor labels are $V = [n]$ where $[n] = \{1, 2, ..., n\}$ or some subset. Two actors $i, j \in [n]$ are *incomparable* $i \|_h j$, if neither $i \prec_h j$ nor $i \succ_h j$. Partial orders on $[n]$ are in one-to-one correspondence with transitively-closed DAGs $([n], E)$ with edges $E = \{\langle i, j\rangle \in [n] \times [n] : i \succ_h j\}$. Denote by $\mathcal{H}_{[n]}$ the set of all partial orders on actor labels $[n]$. Let $\mathcal{P}_{[n]}$ be the set of all permutations of $[n]$. A linear extension $l_h \in \mathcal{P}_{[n]}$ is a permutation of actors in $[n]$ that does not violate partial order h. See Fig. 1 for an example partial order[3] and its LEs. We denote the set of all LEs for partial order h as $\mathcal{L}[h]$. A *sub-order* $h[o] = (o, \prec_h)$ of a partial order $h \in \mathcal{H}_{[n]}$ restricts h to a subset $o \subseteq [n],\ o = \{o_1, ..., o_m\}$: all order relations in h are inherited by $h[o]$ so its DAG representation $(o, E[o])$ has edges $E[o] = \{e \in E : e \in o \times o\}$; directed edges incident vertices in $[n] \setminus o$ are removed and all others remain. A *chain* of $h \in \mathcal{H}_{[n]}$ is a sub-order $h[o]$ that is also a total order. The *length* of a chain is the number of nodes $|o|$ in the sub-order. The *depth* $D(h)$ of a partial order is the length of its longest chain, with $1 \leq D(h) \leq n$.

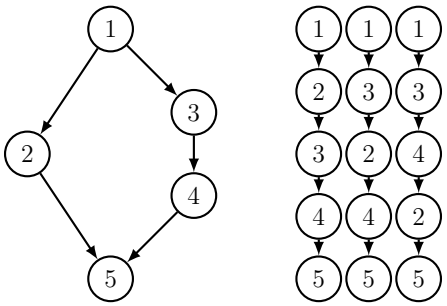


Figure 1: (left) A partial order with 5 actors and depth 4 which is also a VSP, v_0 say, and (right) its three LEs.

The *vertex-series-parallel partial orders (VSP)* on $[n]$ are a class of partial orders $\mathcal{V}_{[n]} \subset \mathcal{H}_{[n]}$ formed by repeated *series* $\otimes$ and *parallel* $\oplus$ operations. For partial orders h_1 and h_2, let $V(h_1)$ and $V(h_2)$ represent the ground sets of actors for h_1 and h_2 respectively (which we assume are disjoint).

- A *series partial order*, $h = h_1 \otimes h_2$, is the union of all relations in h_1 and h_2, with additional relations $i \succ_h j$ if $i \in V(h_1)$ and $j \in V(h_2)$.
- A *parallel partial order*, $h = h_1 \oplus h_2$, is the union of all relations in h_1 and h_2 with incomparability

[2]The binary relation $\prec_h$ is both irreflexive (the relation $i \prec_h i$ does not exist) and transitive (if $i \prec_h j$ and $j \prec_h k$, then $i \prec_h k$), where $i, j, k \in [n]$ and $i \neq j \neq k$.

[3]In this article, we visualise a partial order via its transitive reduction - this omits all edges implied by transitivity and is unique.

$i \|_h j$ if $i \in V(h_1)$ and $j \in V(h_2)$.

The set of VSPs $\mathcal{V}_{[n]}$ is defined recursively: if $|V(h)| = 1$ then h is a VSP; if h_1 and h_2 are VSPs then $h_1 \otimes h_2$ and $h_1 \oplus h_2$ are VSPs. Valdes et al. [1979] show that a partial order is a VSP if it does not contain the "forbidden sub-graph" (Appendix G, Fig. G.1) as a subgraph isomorphism.

The partial order v_0 in Fig. 1 is a VSP. It can be constructed using the series and parallel operations in Fig. 2.

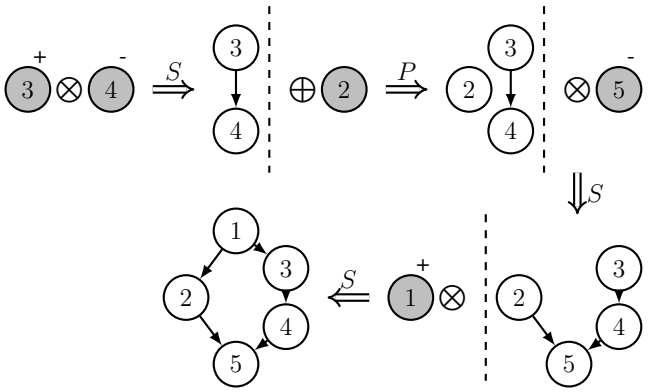


Figure 2: One possible construction procedure for the VSP v_0 shown in Fig. 1.

A VSP on n actors can be parameterised as a Binary Decomposition Tree (BDT) Valdes et al. [1979] - a binary tree $t \in \mathcal{T}_{[n]}$ with n leaves in which nodes have additional attributes (listed below) and edges are directed from the root to the leaves. Let $\mathcal{F}$ and $\mathcal{A}$ be the index sets for the n leaves and $n-1$ internal nodes respectively, with $\mathcal{F} \cup \mathcal{A} = [2n-1]$. Each leaf node index corresponds to a unique actor in the VSP. It is convenient to distinguish between leaf nodes indices and the actor labels to which they correspond. For each leaf node $i \in \mathcal{F}$, let $F_i(t) \in [n]$ give the actor label for the actor corresponding to that leaf node. Internal nodes $i \in \mathcal{A}$ are S nodes if the subtrees rooted by their child nodes are merged in series, otherwise they are P nodes and the subtrees are merged in parallel. Internal nodes with an S label have an additional attribute indicating which of its child nodes is the "upper child": the subtree of this child node (indicated by a '+' and a red edge in Fig. 3) is stacked above the subtree rooted by the other child node (indicated by a '-'). As an example, the VSP v_0 in Fig. 1 can be represented by the BDT t_0 in Fig. 3. Let $S(t) \in [n-1]$ be the number of S-nodes in tree t.

A tree t with edge set $E(t)$ is written $t = (F(t), E(t), L(t))$. Here $L(t) = \{L_i\}_{i \in \mathcal{A}}$ with $L_i(t) = (j, j')$ indicating that internal node i is an S-node with child nodes j, j' and the subtree rooted by j is stacked above that rooted by j', and $L_i(t) = \emptyset$ if i is a P-node. The map from a BDT to the VSP $v : \mathcal{T}_{[n]} \to \mathcal{V}_{[n]}$ is not bijective: for a VSP $v \in \mathcal{V}_{[n]}$, there may exist many BDTs $t \in \mathcal{T}_{[n]}$ which represent it. Let $t(v) = \{t \in \mathcal{T}_{[n]} : v(t) = v\}$ give the set of BDTs representing VSP $v \in \mathcal{V}_{[n]}$.

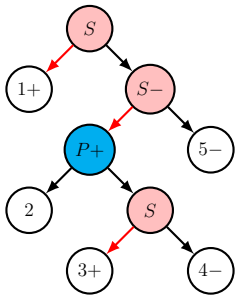


Figure 3: A BDT t_0 representing v_0 in Fig. 1, so that $v(t_0) = v_0$. Red edges and '+' signs indicate the upper child.

Brightwell and Winkler [1991] show that counting the number of LEs of a partial order is a #P-complete problem. However, the subclass of VSP partial orders admits fast counting. Wells [1971] gives

$$|\mathcal{L}(h_1 \otimes h_2)| = |\mathcal{L}(h_1)||\mathcal{L}(h_2)| \tag{1}$$

$$|\mathcal{L}(h_1 \oplus h_2)| = |\mathcal{L}(h_1)||\mathcal{L}(h_2)| \binom{|V(h_1)| + |V(h_2)|}{|V(h_1)|} \tag{2}$$

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.

In the following we make use of one more representation of a VSP: the *Multi-Decomposition Tree* (MDT). These trees are obtained by collapsing edges which connect internal nodes of the same S/P-type in the BDT, as in Fig. 4. Let $\mathcal{M}_{[n]}$ be the set of all MDTs with n distinguisable leaves. A formal definition is given in Appendix A.3.

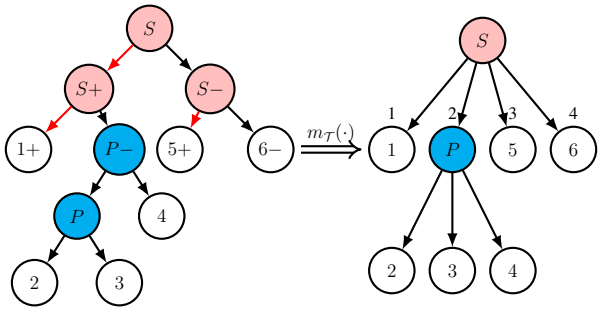


Figure 4: An example BDT t_1 (left) and its corresponding MDT m_1 (right). The child-nodes of any S-node in the MDT are numbered to give the order in which their subtrees are stacked by the BDT.

Valdes [1978] has shown that MDTs are one-to-one with VSPs, so all the BDTs in $t(v)$ representing the VSP v must "collapse down" to give the same MDT. For $m \in \mathcal{M}_{[n]}$ we write $v = v(m)$ for the map to VSPs (relations between any pair of actors in the VSP are simply given by the type of their Most Recent Common Ancestor (MRCA) in m). Let $m_{\mathcal{V}}(v) = \{m \in \mathcal{M}_{[n]} : v(m) = v\}$ be the set of MDTs representing $v \in \mathcal{V}_{[n]}$.

Lemma 1 *The map $m_{\mathcal{V}} : \mathcal{V}_{[n]} \to \mathcal{M}_{[n]}$ is bijective (so that $|m_{\mathcal{V}}(v)| = 1$). See Valdes [1978] for proof and Valdes et al. [1979] for further discussion.*

2 VSP PRIOR

In this section we give a marginally consistent prior $\pi_{\mathcal{V}_{[n]}}(v|q)$ over VSPs on actors in $[n]$, controling the distribution over VSP-depth. We begin by defining a prior probability distribution $\pi_{\mathcal{T}_{[n]}}(t|q)$ over BDTs $t \in \mathcal{T}_{[n]}$.

Our prior on $\mathcal{T}_{[n]}$ has a uniform distribution over trees $([2n-1], E(t))$ with distinguishable leaves. Internal nodes are labelled S with probability q and otherwise P. We choose an "upper child" for each S node at random from its two child nodes, so we have

$$\pi_{\mathcal{T}_{[n]}}(t|q) = \frac{1}{|\mathcal{T}_{[n]}|} \left(\frac{q}{2}\right)^{S(t)} (1-q)^{n-S(t)-1}, \tag{3}$$

where $S(t)$ is the number of S-nodes, $|\mathcal{T}_{[n]}| = (2n-3)!! \equiv (2n-3) \cdot (2n-5)...3 \cdot 1$ is the number of binary tree topologies with n distinguishable leaves, and the types of the $n-1$ internal nodes are independent with a factor $2^{-S(t)}$ for the stacking order of the children of S-nodes.

We get the prior on VSPs $v \in \mathcal{V}_{[n]}$ by summing over all BDTs that represent v,

$$\pi_{\mathcal{V}_{[n]}}(v|q) = \sum_{t \in t(v)} \pi_{\mathcal{T}_{[n]}}(t|q) \tag{4}$$

This simple choice, based on a uniform distribution over tree topologies, determines a prior for VSPs that represents the prior knowledge we want to impose in our setting. If a social hierarchy is built up by making comparisons between groups of people, based for example on their profession, then it will be a VSP. Secondly, the unknown true depth of the social hierarchy we are trying to reconstruct (which is the length of the longest chain in the VSP) is a feature of particular interest, so we don't want the prior to strongly inform depth. We choose a prior distribution over q so that the marginal distribution $\pi_{\mathcal{V}_{[n]}}(v)$ gives a reasonably flat prior distribution for depth $D(v)$ (see Appendix H and Fig.H.1).

We assume that relations between two actors are determined by (unknown) properties intrinsic to those actors (for example, their professions, or ancestry). If that is true then the presence or absence of a third actor should not affect the relations between the first two. It is not straightforward to get this property *and* transitivity. If two actors $1 \| 2$ are unordered and we add actor 3 with relations $1 \succ 3$ and $3 \succ 2$ then $1 \succ 2$ by transitivity: the presence of actor 3 changes the relation between actors 1 and 2. Random VSPs can be built up in many different ways (that is, they are represented by many different BDTs), so we want the prior probability that $1 \succ_w 2$ in a random VSP $w \sim \pi_{\mathcal{V}_{[2]}}$ to be the same as the prior probability that $1 \succ_v 2$ in a random VSP $v \sim \pi_{\mathcal{V}_{[3]}}$. This adds a consistency restriction on any family of prior distributions $\pi_{\mathcal{V}[n]},\ n \geq 1$ we write down.

A family of priors like $\pi_{\mathcal{T}_{[n]}}(t|q)$ or $\pi_{\mathcal{V}_{[n]}}(v|q),\ n \geq 1$ is *marginally consistent* (also known as *projective*) if every marginal of every distribution in the family is also in the family. Marginal consistency is not a property we get for free from the axioms of probability: the uniform distribution on partial orders $h \sim \mathcal{U}(\mathcal{H}_{[n]})$ is not consistent: there are 3 partial orders on the labels $\{1,2\}$ and 19 on $\{1,2,3\}$; since 19 is not divisible by 3, the probability for $1 \succ_h 2$ in $h \sim \mathcal{U}(\mathcal{H}_{[2]})$ cannot equal the marginal probability for $1 \succ_g 2$ in $g \sim \mathcal{U}(\mathcal{H}_{[3]})$.

Definition 1 (Marginal consistency) *Let $\mathcal{O}_{[n]} = \{o \subseteq [n] : |o| > 0\}$ be the set of all subsets of $[n]$ with at least one element. The family of VSP priors $\pi_{\mathcal{V}_o}(v|q),\ o \in \mathcal{O}_{[n]},\ n \geq 1$ 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

$$\pi_{\mathcal{V}_o}(w|q) = \sum_{\substack{v \in \mathcal{V}_{\tilde{o}} \\ v[o]=w}} \pi_{\mathcal{V}_{\tilde{o}}}(v|q) \quad \textit{for all } w \in \mathcal{V}_o. \tag{5}$$

If marginal consistency holds for all q then it holds for marginals $\pi_o(w)$ by taking expectations over q in (5).

The following Theorem is our first main result: we give a closed form expression for the prior for a VSP (we calculate the sum in (4)) and show that the family of priors is marginally consistent. For $v \in \mathcal{V}_{[n]}$, let $t \in t(v)$ be some tree representing v. Partition the internal nodes $\mathcal{A}$ of t into S-*clusters* $C_k^{(S)},\ k = 1, ..., K_S$ and P-*clusters* $C_k^{(P)}, k = 1, ..., K_P$. An S-cluster is a maximal set of internal nodes of type S which are connected by edges in $E(t)$ and corresponds to a node in the MDT-representation. The P-clusters are defined similarly. We will see (in Appendix A.2, proof of Proposition 5) that two BDTs representing the same VSP have the same numbers of S and P clusters, with the same sizes.

Theorem 1 *The family, $\pi_{\mathcal{V}_o}(v|q),\ o \in \mathcal{O}_{[n]}\ n \geq 1$, of VSP priors is marginally consistent. The probability distribution over VSPs $v \in \mathcal{V}_{[n]}$ in (4) is*

$$\pi_{\mathcal{V}_{[n]}}(v|q) = \pi_{\mathcal{T}_{[n]}}(t|q) \prod_{k=1}^{K_P} (2|C_k^{(P)}|-1)!! \prod_{k'=1}^{K^S} \mathcal{C}_{|C_{k'}^{(S)}|} \tag{6}$$

where t may be taken to be any tree $t \in t(v)$ with P- and S-clusters defined above, $\pi_{\mathcal{T}_{[n]}}(t|q)$ is given in (3) and

$$\mathcal{C}_s = \frac{1}{s+1}\binom{2s}{s}, \quad s \geq 0 \tag{7}$$

is the s'th Catalan number [Stanley and Weisstein, 2002].

Proof 1 (Theorem 1) *The proof of Theorem 1 is given in two parts in Appendix A. In Proposition 3 in Appendix A.1 we show that the family of tree-priors $\pi_{\mathcal{T}_{[n]}}(t|q),\ o \in \mathcal{O}_{[n]},\ n \geq 1$ is marginally consistent. This result is used in Proposition 4 in A.1 to show that VSPs are marginally consistent - the first part of Theorem 1.*

The proof of the second part is given in Appendix A.2. We show in Proposition 5 that all trees $t \in t(v)$ have equal values of $\pi_{\mathcal{T}_{[n]}}(t|q)$, so that $\pi_{\mathcal{V}_{[n]}}(v|q) = |t(v)|\pi_{\mathcal{T}_{[n]}}(t|q)$ for any $t \in t(v)$. This is straightforward, as they must all collapse down to the same MDT. Finally, in Proposition 6, we give a formula for $|t(v)|$. We count the number of BDTs that collapse down to a given MDT. Any P-cluster C_k^P of a BDT corresponds to a P-node in its MDT and covers a small sub-tree of the BDT representing an empty partial order on its $|C_k^P| + 1$ labeled leaves. It can be replaced in the BDT by any sub-tree representing the empty partial order without changing the MDT, and there are $(2|C_k^{(P)}|-1)!!$ such trees. Similarly, any S-cluster C_k^S corresponds to a S-node in the MDT and covers a sub-tree of the BDT representing a total order on its leaves. It can be replaced in the BDT by any sub-tree representing the same total order. The Catalan numbers enter because $\mathcal{C}_{s-1}$ gives the number of BDTs representing a total order on s elements (see proof Proposition 6). This last result is new, gives (6) and completes the proof of Theorem 1.

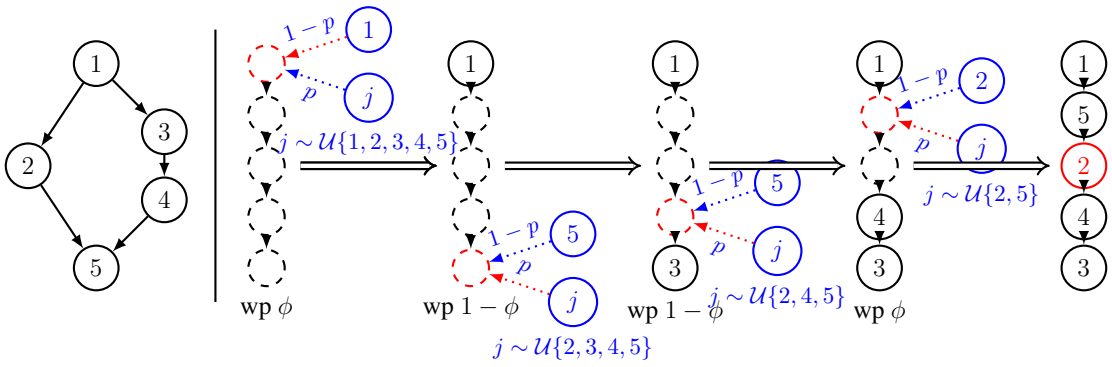


Figure 5: One example list simulation process from the VSP v_0 (left) via the QJ-B observation model. The simulated list is displayed on the right.

Theorem 1 gives the prior for a VSP in terms of the prior for one of the BDTs that represent that VSP. We can also parameterise VSPs using MDTs and this leads to the second MCMC scheme given in Appendix C.2.

Corollary 1 *For $m \in \mathcal{M}_{[n]}$ with internal nodes $\mathcal{A}$, let c_i give the number of children of node $i \in A$ and let $P(m) = \{i \in \mathcal{A} : L_i(m) = \emptyset\}$ and $S(m) = \mathcal{A} \setminus P(m)$ give the sets of $P-$ and $S-$node labels. The prior for VSPs given in (6) is equivalently a prior for MDTs,*

$$\begin{aligned} \pi_{\mathcal{V}_{[n]}}(v(m)|q) &= \pi_{\mathcal{M}_{[n]}}(m|q) \\ &= \frac{1}{(2n-3)!!} \prod_{i \in P(m)} (1-q)^{c_i - 1}(2c_i - 3)!! \\ &\quad \times \prod_{j \in S(m)} \left(\frac{q}{2}\right)^{c_j - 1} \mathcal{C}_{c_j - 1}. \end{aligned} \tag{8}$$

Proof 2 (Corollary 1) *Substitute (3) into (6) and note a tree with c_i leaves has $c_i - 1$ internal nodes. This result gives a convenient representation for prior evaluation.*

3 BI-DIRECTIONAL QUEUE-JUMPING OBSERVATION MODEL

Our data is a collection of N lists. For $j \in [N]$ let $o_j \subseteq [n]$, $o_j = \{o_1, ..., o_{n_j}\}$ be the actors present when the j'th ranking list was observed and let $y_j \in \mathcal{P}_{o_j}$, $y_j = (y_{j,1}, ..., y_{j,n_j})$ be the observed list, just an ordered version of o_j. Let $y = (y_1, ..., y_N)$ be the list of lists. The 'queue-based' observation model given in Nicholls and Muir Watt [2011] models list data as a realisation of a random queue constrained to put higher status individuals before those of lower status. In this model the queue is dynamic. It forms and then unconstrained pairs of actors swap places at random. If this process reaches equilibrium before the queue is read off then the resulting list is a uniform draw from the linear extensions of the constraining social hierarchy [Karzanov and Khachiyan, 1991]. In this noise-free model $y_j \sim \mathcal{U}(\mathcal{L}[v[o_j]])$ independently for $j \in [N]$.

It is unlikely the observations are "error free". In a "queue-jumping" model (QJ-U, see Appendix B.1and Nicholls and Muir Watt [2011] for details) the queue is read from the top: with probability

$p \in [0,1]$ the "next" person in the queue is drawn at random from those remaining, ignoring the social hierarchy; otherwise they are the first person in the remaining LE. The queue can also be read from the bottom up. In this model (QJ-D) actors fall down the queue. We think of these events as actors arriving while the queue is being read.

We would like to have a queue-based model in which displacement in both directions is possible. The resulting "bi-directional queue-jumping" model (QJ-B) is not simply a mixture of QJ-U and QJ-D, as it allows displacement in both directions within a single realisation. The cost of evaluating a QJ-B likelihood is exponential in n. However, for the application in Section 5.1 there is a subset of actors (bishops) known a priori to appear as a group. Separate modelling of this manageable subset ($n \simeq 20$) is well-motivated. Although QJ-B cannot be evaluated for a general partial order (counting LEs is prohibitive) it is fine for a VSP.

Like QJ-U, QJ-B ranks by repeated selection. Fig. 5 provides an example QJ-B list-realisation from VSP v_0. A generic list $x \in \mathcal{P}_{[n]}$ is built up from both ends (see Appendix B.2). Let $z \in \{0,1\}^{n-1}$ with $z_k \sim Bern(\phi)$. Here $z_k = 0$ indicates the k'th actor to be added to the list was placed bottom-up in the QJ-D model and $z_k = 1$ indicates they were placed top-down in the QJ-U model. In Fig. 5, $z = (1,0,0,1)$. If we let $U_0 = 0$ then $U_k = U_{k-1} + z_k$ gives the number of places filled from the top after the k'th actor has arrived, so if $z_k = 1$ then the k'th actor was placed into position $i_k = U_k$ in x. Similarly, if $D_0 = n+1$ then $D_k = D_{k-1} - (1 - z_k)$ tracks places filled from the bottom and gives the placement index $i_k = D_k$ in x when $z_k = 0$, so $i_k = z_k U_k + (1 - z_k) D_k$ gives the position in x into which the k'th actor was added. If $z = (1,0,0,1)$, then $(i_1, ..., i_4) = (1,5,4,2)$ (and $i_5 = 3$, the only remaining place).

Definition 2 (Bi-Directional Queue-Jumping Model) *Let $L_T(v) = |\mathcal{L}[v]|$ be the number of LEs 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 LEs with actor i at the top. Let $B_i(v) = |\{l \in \mathcal{L}[v] : l_n = i\}|$ give the number of LEs with actor i at the bottom. If $z \in \{0,1\}^{n-1}$ is given then $i_k = i_k(z), k = 1, ..., n$ is given above. The observation model for QJ-B for a list $x \in \mathcal{P}_{[n]}$ given z is*

$$\begin{aligned} Q_{bi}(x|z,v,p) = & \prod_{k=1}^{n-1} [\mathbb{1}_{\{z_k=0\}} Q_{bi}(x_{i_k}|x_{i_{1:k-1}}, z_k, v, p) \\ & + \quad \mathbb{1}_{\{z_k=1\}} Q_{bi}(x_{i_k}|x_{i_{1:k-1}}, z_k, v, p)], \end{aligned}$$

where

$$\begin{aligned} & Q_{bi}(x_{i_k}|x_{i_{1:k-1}}, z_k = 0, v, p) = \\ & \frac{p}{n-k+1} + (1-p)\frac{T_{x_{i_k}}(v[x_{[n]\setminus\{i_1,...,i_{k-1}\}}])}{L_T(v[x_{[n]\setminus\{i_1,...,i_{k-1}\}}])}, \end{aligned}$$

$$\begin{aligned} & Q_{bi}(x_{i_k}|x_{i_1:k-1}, z_k = 1, v, p) = \\ & \frac{p}{n-k+1} + (1-p)\frac{B_{x_{i_k}}(v[x_{[n]\setminus\{i_1,...,i_{k-1}\}}])}{L_T(v[x_{[n]\setminus\{i_1,...,i_{k-1}\}}])}, \end{aligned}$$

and marginally,

$$Q_{bi}(x|v,p,\phi) = \sum_{z \in \{0,1\}^n} Q_{bi}(x|z,v,p) p(z|\phi) \tag{9}$$

where $p(z|\phi) = \phi^{\sum_i z_i}(1-\phi)^{n-\sum_i z_i}$.

We give a generative model realising $x \sim Q_{bi}$ in Appendix B.2. This distribution reduces to Q_{up}/QJ-U in Appendix B.1 when $\phi = 1$ (and Q_{down}/QJ-D when $\phi = 0$). We use this nesting to investigate whether QJ-U or QJ-D or QJ-B fits the data better. This is of interest in our application as different error types correspond to obvious physical mechanisms (downwards displacement may be "arrived late" and upwards displacement may be "my friend the King is present"). When $p = 0$ this is the noise free model for every $\phi \in [0, 1]$, so ϕ is not identifiable in the noise free setting.

Counting LEs of a VSP (evaluating $L_T(v)$ etc) is $O(n)$ so the computational complexity for naive evaluation of Q_{bi} using (9) is $O(n^2 2^n)$. We used a recursion (Algorithm B.3) of computational complexity $O(n2^n)$. This avoids repeated evaluation of LE-counts for the same suborders and (by Proposition 7 in Appendix B.3) evaluates Q_{bi}.

4 SUMMARISING THE VSP POSTERIOR

Bayesian inference is straightforward in principle given an explicit prior distribution over VSPs and an observation model $Q = Q_{up}$ or $Q = Q_{bi}$ for our N ranking-lists. We can represent a VSP as a MDT (since the mapping is one-to-one) or carry out Bayesian inference on the latent space of BDTs $t \in \mathcal{T}_{[n]}$ and use the fact that they marginalise to MDTs. We present the posteriors for BDT and VSP. Let $\psi = (p, \phi)$ for QJ-B and $\psi = p$ for QJ-U.

The posterior for the BDT $t \in \mathcal{T}_{[n]}$ is

$$\pi_{\mathcal{T}_{[n]}}(t, q, \psi|y) \propto \pi_{\mathcal{T}_{[n]}}(t|q)\pi(q, \psi)Q(y|v(t), \psi) \tag{10}$$

The posterior distribution for the VSP $v \in \mathcal{V}_{[n]}$ is

$$\pi_{\mathcal{V}_{[n]}}(v, q, \psi|y) \propto \pi_{\mathcal{V}_{[n]}}(v|q)\pi(q, \psi)Q(y|v, \psi), \tag{11}$$

where we use the equivalent MDT posterior with prior given in Corollary 1 for VSPs in (11).

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 (see Appendix A.4 for proof).*

We implemented separate MCMC samplers targeting both (10) and (11). Our MCMC algorithms are given in Appendix C. We checked that the VSP-posterior marginals for the two implementations were equal (up to Monte-Carlo error). We implemented MCMC targeting the BDT posterior (10) first, as BDT data structures are slightly more straightforward to handle than the MDT data structures needed to target the VSP posterior in (11). All results in the next section were computed using the BDT-MCMC.

5 APPLICATIONS

5.1 DATA AND ANALYSES

We analyse a dataset accessed through a database made for "The Charters of William II and Henry I" project by Professor Richard Sharpe and Dr Nicholas Karn [Sharpe et al., 2014]. These data collect

witness lists from legal documents from England and Wales in the eleventh and twelfth century. Witness lists respect a rigid social hierarchy: higher status individuals come ahead of lower status individuals in the lists. Fig. D.1 is an example list.

We represent the hierarchy on actors $[n]$ appearing in the lists as a partial order which is a VSP $v \in \mathcal{V}_{[n]}$ and model a list as the outcome of one of the queuing processes described in Section 3. We imagine the actors lining up to witness the document in a virtual queue.

Lists are witnessed by people from all walks of life and we have their titles. These include "others" (actors who lack titles). Historians are interested in social hierarchies and how they change over time. For illustration we reconstruct hierarchies in three snapshots: the years 1080-84, 1126-30 and 1134-38. The last two cover periods shortly before and after Stephen became King, a time of great change. The 5-year intervals are short enough for any changes in the hierarchy to be slight [Nicholls et al., 2022]. For ease of visualisation we present results for individuals appearing in at least 5 lists (5LPA data) here and results on all actors (1LPA data) in Appendix D.1.1. We fit VSP/QJ-U to all data and fit VSP/QJ-B to 2 of the 3 5LPA data sets (not 1134-38, as QJ-B has runtime growing exponentially with the length of the longest list). However, relations between bishops in 1134-38 are of particular interest so we present VSP/QJ-B results for this subgroup. Table D.1 summarises the data in the different experiments on the Royal Acta data.

In a separate analysis illustrating how our methods apply more generally to any rank-order data, we give an analysis of Formula 1 race outcomes for the 2021 season. Data and results are given in Appendix D.2.

The prior for error probability p and for q (probability for an S-node) is given in Fig. 8. All fitting is done using MCMC in the BDT representation, Algorithm C.1. For any given model we draw MCMC samples $t^{(k)}, p^{(k)}, q^{(k)}, \phi^{(k)} \sim \pi_{\mathcal{T}_{[n]}}(\cdot|y)$ for $k = 1, ..., K$ and set $v^{(k)} = v(t^{(k)})$ per Proposition 1. Example MCMC traces are given in the supplement with Effective Sample Size (ESS) values (Appendix D.1). Sampled VSPs are summarised using consensus VSPs: $V^{con}(\epsilon)$ includes order relation/edge $\langle i, j \rangle$ if the relation appears more than ϵK times in the MCMC output. We color edges black if they are in $V^{con}(\epsilon)$ at $\epsilon = 0.5$ but not $\epsilon = 0.9$ and red if they are supported at $\epsilon = 0.9$. We plot transitive reductions. These omit strongly supported edges from the top to the bottom of the DAG for clarity.

In Sec. 5.2, we fit the QJ-U and QJ-B models to the 5LPA data and make a model comparison using Bayes factors. Consensus orders for the 1LPA data are given in Appendix D.1.1. We additionally compare these models with bucket order models, a Plackett-Luce mixture, Mallows mixture and latent partial order model in Appendix E. We carry out these tests on both the Royal Acta data and the F1 race result data. We report computing time measurements for counting LEs for the latent partial order model and the VSP. They are compared empirically in Appendix F.

5.2 RESULTS

We begin by making reconstruction-accuracy tests on synthetic data. Our list data are incomplete, in the sense that the membership in list $i = 1, ..., N$ is o_i not $[n]$ and the N-values in Table D.1 are not much larger than the number of actors n. In order to measure the reliability of the reconstructions which follow we take representative parameters (parameters sampled from the corresponding posterior, the last sampled state $v^{(K)}, p^{(K)}, q^{(K)}, \phi^{(K)}$) and generate synthetic data with the same list-membership

and length structures as the real data. The ROC curves in Fig. D.12 (5LPA data and QJ-U) and D.15 (5LPA data and QJ-B) for consensus orders $V^{con}(\epsilon)$ show the proportion of inferred false-positive and true-positive relations increasing with decreasing ϵ from $(0, 0)$ at $\epsilon = 1$ (the consensus order is empty) to $(1, 1)$ at $\epsilon = 0$ (complete graph). For each simulated data set there is ϵ giving high true-positive and low false-positive reconstructed relation fractions: if our model is accurate then we reconstruct relations well.

We next report consensus partial orders. Consensus orders for actors color-coded by their professions are shown in Fig. 6 and 7. For both QJ-U and QJ-B models, we observe three clear social hierarchies: King $\succ$ Queen $\succ$ Duke appear at the top, in that order (when they are in the 5LPA data, in 1180-84 and 1134-38); then archbishop/prince $\succ$ bishops; the remaining professions (earl, count, chancellor, other) are ranked lower than bishops in a relatively complex hierarchy.

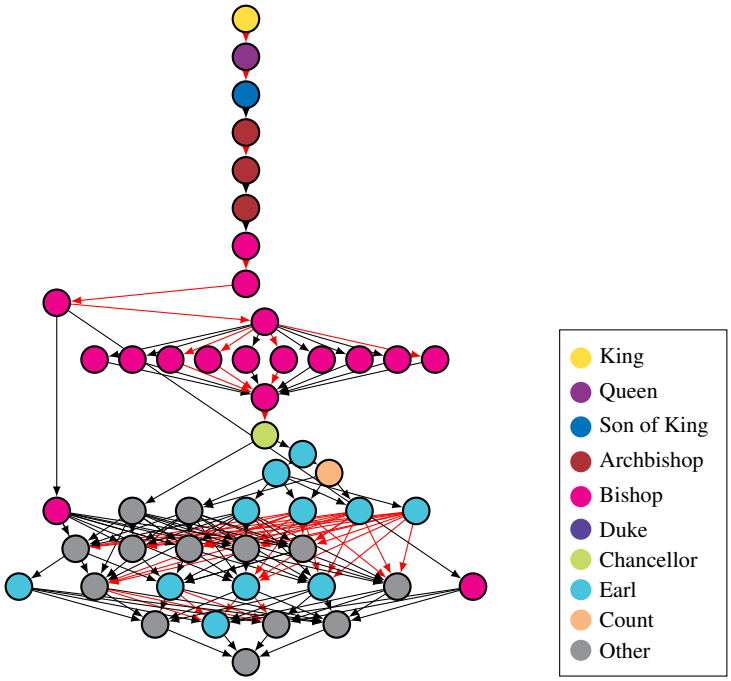


Figure 6: VSP/QJ-U model. Consensus order for 1134-38 5LPA data. Significant/strong order relations are indicated by black/red edges respectively.

Some of this is common sense. However, the web of strongly attested relations between earls and others in 1134-38 is new. There is clear evidence for hierarchies within professions. The bishop-only QJ-U analysis in 1134-38 (top-right graph in Fig. 7) is similar to the bishop subgraph of the full QJ-U analysis for the same period (pink nodes in Fig. 6). The prior is marginally consistent, but information is shared across lists so removing actors changes the data and changes estimated order relations between those that remain. However, the bishops appear as a group in the lists and in Fig. 6 and there are few non-bishops "between" bishops in lists, so this effect is slight. We can attach names to nodes: for example, the top three bishops in 34-38 (in Fig. 6 and in both QJ-U and QJ-B analyses in the rightmost column of Fig. 7) are Henry, de Blois, Bishop of Wincester $\succ$ Roger, Bishop of Salisbury $\succ$ Alexander, Bishop of Lincoln.

The status hierarchies fitted using by QJ-B (bottom row Fig. 7) are simpler and deeper than QJ-U (top row Fig. 7). The data must contain a small number of errors in both directions. A uni-directional model must fit a shallower VSP as it accommodates errors in the "wrong" direction by removing order

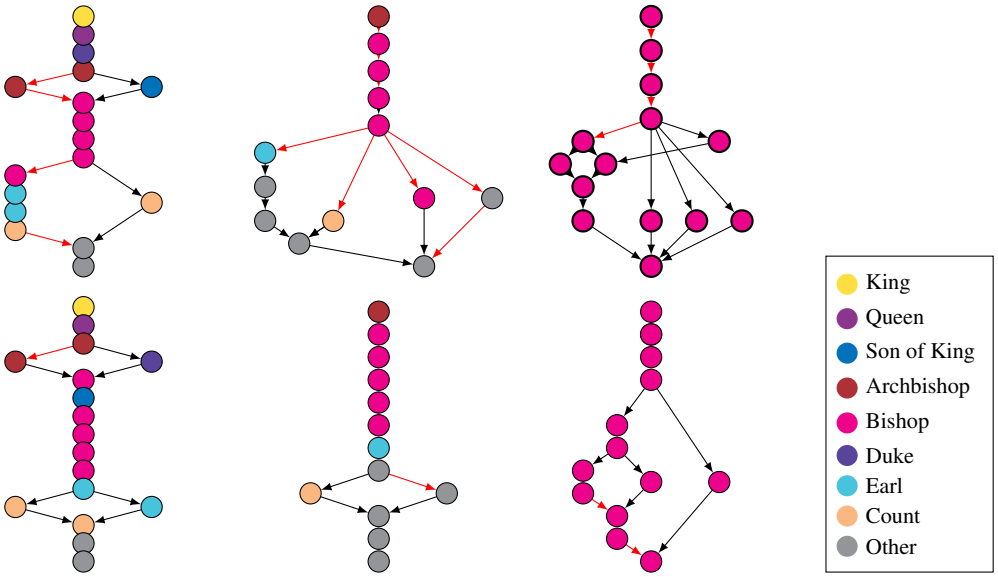


Figure 7: VSP/QJ-U (top row) and VSP/QJ-B (bottom row). Consensus orders for 1080-84, 1126-30 and 1134-38 (bishops) (left to right columns) 5LPA data.

relations in the reconstructed VSP.

We summarise the status of "professions" within VSPs by averaging ranks. Given a partial order $v \in \mathcal{V}_{[n]}$, the rank of actor $i \in [n]$ is the number of actors above them, $\text{rank}_i(v) = 1 + |\{\langle e_1, e_2\rangle \in E(v) : e_2 = i\}|$, and take as our summary the average rank of actors in the profession. The posterior mean ranks given in Table D.5 and D.7 match our remarks on consensus orders.

We next report parameter distributions. Prior and posterior distributions for the probability q for a serial node, error probability p and QJ-B parameter ϕ (equal one for QJ-U and zero for QJ-D) for the three periods are shown in Fig. 8. The p-posteriors are weighted toward smaller values and overlap, though errors are low in 1126-30 and higher in 1180-84 indicating greater respect for the rules of precedence in 1126-30 than in 1180-84. Prior and posterior depth distributions are shown in Fig. D.11 and D.14. The prior depth distributions are fairly flat so any depth-structure in the posterior comes from the data. The probability for a series node in the BDT (q) controls the depth of the fitted order relation. For example, in 1180-84 a relatively high q for QJ-U is associated with relatively high depth VSPs with a mean depth of 14 relative to maximum depth 17 (the number of actors). In contrast, the posterior probabilities for S and P nodes are almost equal in 1134-38 and so we get a relatively shallower hierarchy: the posterior mean depth is about 23 relative to a maximum depth 49 in Fig. D.11.

The QJ-B model for noise in the list data allows actors to jump up or down from a queue-position appropriate for their status. QJ-U is favored if $\phi > 1/2$ and otherwise QJ-D so we see from Fig. 8 that QJ-U is favored in 1134-38(b), while the 1080-84 data supports QJ-D. However, the p-posteriors both favor small p. The displacement direction controlled by ϕ is hard to measure and not identifiable at $p = 0$ so the ϕ-distributions are correspondingly broad.

We next report results of model selection between different queue jumping error models. Preference shifts from downwards to bidirectional to upwards displacement error models over the period 1080-1140. We justify this reading of the results using Bayes factors below. In summary, QJ-D is slightly favored over QJ-B (so we write "$D > B$") in 1080-84 while in 1126-30 models QJ-D and QJ-B are equally

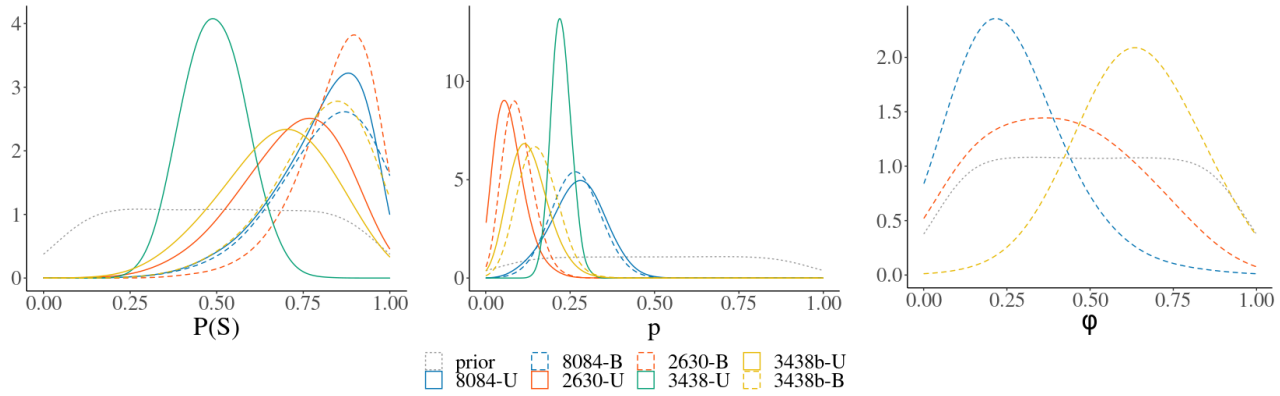


Figure 8: Posterior distributions for $q = P(S)$ (left), error probability p (middle) and QJ-B probability ϕ (right) for the time periods 1180-1184 (blue), 1126-1130 (red), 1134-1138 (green) and 1134-1138(b) (yellow) from both the VSP/QJ-U (solid) and VSP/QJ-B (dashed) models. The prior is represented in grey in all figures.

good ($D \approx B$). Both are clearly favored over QJ-U in these periods ($D, B \gg U$). In 34-38(b) we have $U \approx B$ and $U, B \gg D$.

We can read the Bayes factors we need off Fig. 8 because the models QJ-U and QJ-D are nested in the model QJ-B. The Bayes factor $B_{U,B}$ for QJ-U over QJ-B is

$$\begin{aligned} B_{U,B} &= \lim_{\delta \to 0} \frac{p(y|\phi > 1-\delta)}{p(y|\phi \in (0,1))} \\ &= \lim_{\delta \to 0} \frac{\pi(\phi > 1-\delta|y)}{\pi(\phi \in (0,1)|y)} \frac{\pi(\phi \in (0,1))}{\pi(\phi > 1-\delta)} \\ &= \lim_{\delta \to 0} \frac{\pi(\phi > 1-\delta|y)}{\pi(\phi > 1-\delta)}, \end{aligned}$$

since $\phi \in (0,1)$ with probability one. Similarly,

$$B_{D,B} = \lim_{\delta \to 0} \frac{\pi(\phi < \delta|y)}{\pi(\phi < \delta)},$$

and then $B_{U,D} = B_{U,B}/B_{D,B}$. From Fig. 8, $B_{U,B}$ is close to 0 in periods 1180-84 and 1126-30 as the posterior density is well below the prior density at $\phi \to 1$, providing strong support for QJ-B over QJ-U. In 1134-38(b), we see $B_{U,B} \approx 1$, as the curves meet as $\phi \to 1$ so there is no clear signal from the data. The other comparisons may be justified similarly.

Finally we make model comparisons with other models. Comparisons with a Plackett-Luce mixture model and a Mallows mixture model are given in Appendix E.1, the latent partial order model from Nicholls and Muir Watt [2011] in Appendix E.3 and a simple Bucket Order model in Appendix E.2. When models are nested (Bucket Order) we estimate a Bayes factor. When they are not, we use the Expected Log Pointwise Predictive Density (ELPD, Vehtari et al. [2017]) as our criterion. This is a predictive loss which can be estimated using LOOCV or the WAIC [Watanabe, 2013]. On this basis VSP/QJ (-U and -B) is clearly favoured over Placket-Luce mixture models and Mallows mixture model in Table E.1 ("Royal Acta") and E.3 (Formula 1 race data). With Bayes factors around 2 or 3, Bucket orders are equal or slightly preferred over VSPs in the QJ-B model (Table E.3). Our VSP-based model

QJ-U is clearly preferred over Bucket orders in the QJ-U fit (some very large Bayes factors in favor of VSP).

The support of our VSP model is a subset of the PO support, as POs containing the forbidden sub-graph (Appendix G) are not VSPs. The PO/QJ-U has a slightly larger ELPD (-36.7, see Table E.4) than VSP/QJ-U (-37.8) on the 1126-1130 data with 5LPA. However, the difference is not significant at the precision (± 10) of these estimates so we conclude that VSP/QJ-U models these data as well as PO/QJ-U. It gives similar consensus orders (Fig. E.2) and profession rankings (Table E.5).

A VSP-based analysis is far more computationally efficient than a PO-based model when the number of actors is large. The computing time for counting the LEs of a VSP rises linearly with the number of actors (Fig. F.1) while it increases exponentially for PO (using the best code we could find, *LEcount*, Kangas et al. [2016], but inevitable given Brightwell and Winkler [1991]). We have to count LEs of random POs. In our experience counting LEs on random POs with up to about 25-30 actors is feasible. However, at larger numbers we encounter occasional random POs which are especially "hard" to count and VSP-based analysis is the only way forward at present.

6 DISCUSSION AND CONCLUSION

Our work was motivated by the need to fit relatively large partial orders (up to 200 nodes) to noisy linear-extension data. We saw that, for data on this scale, counting linear extensions in the VSP-tree representation is much faster than current state-of-art counting for general partial orders, enabling our methods to scale. We gave a new consistent and closed form prior distribution over VSPs with a parameter q controlling VSP depth, and a new observation model QJ-B for noisy LEs which generalises QJ-U [Nicholls and Muir Watt, 2011]. We fit the new model to some of the smaller data sets and the old model to all data sets. Neither of these analyses would be possible without the VSP-setup. The data support the new observation model in our application. Our $elpd_{waic}$-based model comparisons also clearly favor VSP/QJ-U and VSP/QJ-B over a Plackett-Luce mixture or a Mallows Mixture. Although we could fit the large data sets, visualising consensus partial orders proved challenging (compare Fig. 7 (top left corner) and Fig. D.4).

We gave MCMC algorithms targeting the posterior for VSPs in both the latent-space (BDT) parameterisation and the integrated MDT parameterisation. We found the BDT-MCMC adequate, though it would be good to make an efficiency comparison with MDT-MCMC, which we expect to be more efficient. These comparisons are underway. BDT updates which don't change the VSP are fast so BDT-MCMC seems to be competitive. For code see `https://github.com/JessieJ315/Bayesian-Inference-for-Vertex-Series-Parallel-Partial-Orders.git`.

In future work we would like to compare our fit with the recently-proposed contextual repeated selection (CRS) model (Seshadri et al. [2020] and Ragain and Ugander [2018]). This is a rich class of models for rank-order data. The elements of the model are not essentially physical, in the sense that a VSP represents a social hierarchy relation by relation. Also, CRS models do not encode transitivity. It is easy to show VSP models cannot be represented as CRS models with "cliques" of size two. CRS models may fit the data well, and a comparison would be worthwhile. However, there is currently no Bayesian CRS analysis so we leave that for future work.

Acknowledgements

We thank Dr. Nicholas Karn for providing the data, and Prof. David Johnson for enlightening discussions.

References

Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Conjunctive Bayesian networks. *Bernoulli*, pages 893–909, 2007.

Isabel Beichl, Alathea Jensen, and Francis Sullivan. A sequential importance sampling algorithm for estimating linear extensions. 2017.

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

Andrés Cano, Manuel Gémez-Olmedo, and Serafén Moral. Approximate inference in Bayesian networks using binary probability trees. *International Journal of Approximate Reasoning*, 52(1):49–62, 2011.

Aristides Gionis, Heikki Mannila, Kai Puolamäki, and Antti Ukkonen. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566, 2006.

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

A. Karzanov and L. Khachiyan. On the conductance of order Markov chains. *Order*, 8:7–15, 1991.

R Duncan Luce. On the possible psychophysical laws. *Psychological review*, 66(2):81, 1959.

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

Heikki Mannila. Finding total and partial orders from data for seriation. In *International Conference on Discovery Science*, pages 16–25. Springer, 2008.

Heikki Mannila and Christopher Meek. Global partial orders from sequential data. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 161–168, 2000.

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

Geoff K. Nicholls, Jeong Eun Lee, Nicholas Karn, David Johnson, Rukuang Huang, and Alexis Muir-Watt. Bayesian inference for partial orders from random linear extensions: power relations from 12th Century Royal Acta, 2022. URL `https://arxiv.org/abs/2212.05524`.

Manuela Pavan and Roberto Todeschini. *Scientific data ranking methods: theory and applications*. Elsevier, 2008.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Stephen Ragain and Johan Ugander. Choosing to rank. *arXiv preprint arXiv:1809.05139*, 2018.

Thomas Sakoparnig and Niko Beerenwinkel. Efficient sampling for Bayesian inference of conjunctive Bayesian networks. *Bioinformatics*, 28(18):2318–2324, 07 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts433. URL https://doi.org/10.1093/bioinformatics/bts433.

Arjun Seshadri, Stephen Ragain, and Johan Ugander. Learning rich rankings. *Advances in Neural Information Processing Systems*, 33:9435–9446, 2020.

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.

L Smail. Junction trees construction: Application to Bayesian networks. In *AIP Conference Proceedings*, volume 2025, page 100007. AIP Publishing LLC, 2018.

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.

Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 8 2013. ISSN 15324435. URL http://arxiv.org/abs/1208.6338.

Mark B Wells. Elements of combinatorial computing. 1971.

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:*

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

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 $\pi_{\mathcal{T}_{[n]}}$.

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:

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

It follows that

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

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.

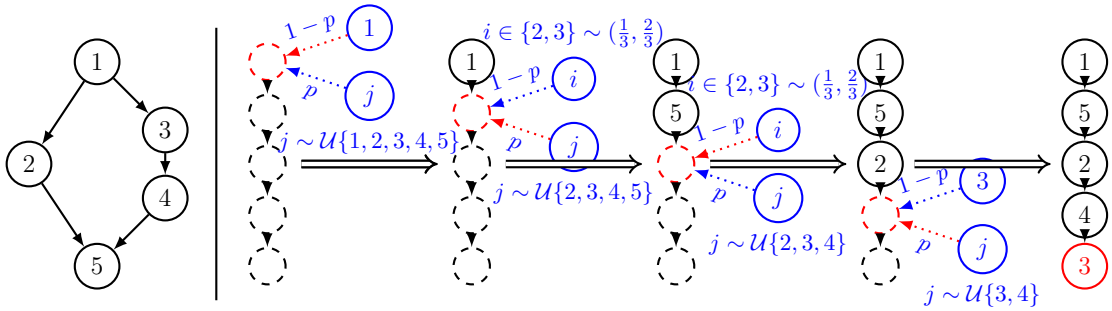


Figure B.1: One example list simulation process from the VSP v_0 (left) via the QJ-U observation model. The simulated list is displayed on the right.

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 ← 1, s ← [n], v' ← v
while |s| > 0 do
    q ← (T_j(v')/L(v'))_{j∈s}
    Sample c ∼ Bern(1 − p)
    if c = 0 then
        Sample x_i ∼ U(s)
    else if c = 1 then
        Sample x_i ∼ multinom(q)
    end if
    s ← s\x_i
    i ← i + 1, v' ← v[s]
end while
return x = (x_1, ..., 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: $x \sim Q_{bi}(x|v,p,\phi)$

$s \leftarrow [n], v' \leftarrow v$
$x \leftarrow (\emptyset,\ldots,\emptyset) \in \{\emptyset\}^n$, $k \leftarrow 1$, $U_0 \leftarrow 0$, $D_0 \leftarrow n+1$
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 $x = (x_1,\ldots,x_n)$

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

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

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.

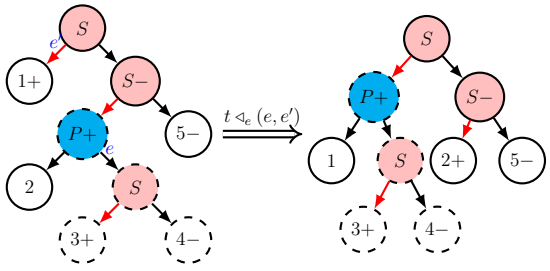


Figure C.1: An example OP-PR edge operation on BDT t_0.

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.

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 $m' = m \lhd_e (e, i)$ creates a new MDT with nodes $\mathcal{F}', \mathcal{A}'$, moving edge $e = \langle e_1, e_2 \rangle$, $e \in E_{-0}(m)$ onto node $i \in (\mathcal{F} \cup \mathcal{A}) \backslash \{e_1, e_2\}$.*

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
   $$E(m') = E(m) \backslash \{e\} \cup \langle i, e_2 \rangle.$$
   Set $L(m') = L(m)$ and make the following changes as needed. If $L_{e_1}(m) \neq \emptyset$ then set $L_{e_1}(m') = L_{e_1}(m) \backslash \{e_1\}$. If $L_i(m) \neq \emptyset$ then suppose $L_i(m) = (j_1, \ldots, j_k)$. Take $L_i(m') \sim \mathcal{U}\{(e_1, j_1, \ldots, j_k), \ldots, (j_1, \ldots$ (insert the subtree below $\langle e_1, e_2 \rangle$ 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
   $$E(m') = E(m) \backslash \{e, \langle \overleftarrow{i}, i \rangle\} \cup \{\langle \overleftarrow{i}, j \rangle, \langle j, i \rangle, \langle j, e_2 \rangle\}.$$
   Set $L(m') = L(m)$ and make the following changes as needed. If $L_{e_1}(m) \neq \emptyset$ then set $L_{e_1}(m') = L_{e_1}(m) \setminus \{e_1\}$. If $L_{\overleftarrow{i}}(m) \neq \emptyset$ (parent is S), suppose $L_{\overleftarrow{i}}(m) = (j_1, \ldots, i, \ldots, j_k)$. Set $L_{\overleftarrow{j}}(m') = (j_1, \ldots, j, \ldots, j_k)$ and $L_j(m') = \emptyset$ (new child is P). Finally, if $L_{\overleftarrow{i}}(m) = \emptyset$ (parent is P), take $L_j(m) \sim \mathcal{U}\{(i, e_2), (e_2, i)\}$ (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
   $$E(m') = E(m) \backslash e \cup \{\langle 0, j \rangle, \langle j, r \rangle, \langle j, e_2 \rangle\}.$$
   Set $L(m') = L(m)$ and make the following changes as needed. If $L_{e_1}(m) \neq \emptyset$ then set $L_{e_1}(m') = L_{e_1}(m) \backslash \{e_1\}$. 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 $L_j(m') \sim \mathcal{U}\{(r, e_2), (e_2, r)\}$ (new node is S).

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

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

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

Ensure:

$$
\begin{aligned}
&t^{(k)} \sim \pi(t|y, q^{(k-1)}, p^{(k-1)}, \phi^{(k-1)}),\\
&q \sim \pi(q|y, t^{(k)}, p^{(k-1)}, \phi^{(k-1)}),\\
&p \sim \pi(q|y, t^{(k)}, q^{(k)}, \phi^{(k-1)}),\\
&\phi \sim \pi(q|y, t^{(k)}, q^{(k)}, p^{(k)})
\end{aligned}
$$

Update for t (local edge operation)

Sample $e \sim \mathcal{U}(E_{-0}(t)), e' \sim \mathcal{U}(E_l(e|t))$

$$
\begin{aligned}
t' &\leftarrow t \triangleleft_e (e, e')\\
\eta_3 &\leftarrow \frac{Q(y|v(t'), p, \phi)\pi_{\mathcal{T}_{[n]}}(t'|q)|E_l(e|t)|}{Q(y|v(t), p, \phi)\pi_{\mathcal{T}_{[n]}}(t|q)|E_l(e|t')|}
\end{aligned}
$$

if $\mathcal{U}(0,1) \leq \eta_3$ **then**
$t \leftarrow t^{(k)} \leftarrow t'$
end if

Updates for q, p and ϕ omitted

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 $\rho(m'|m)$.

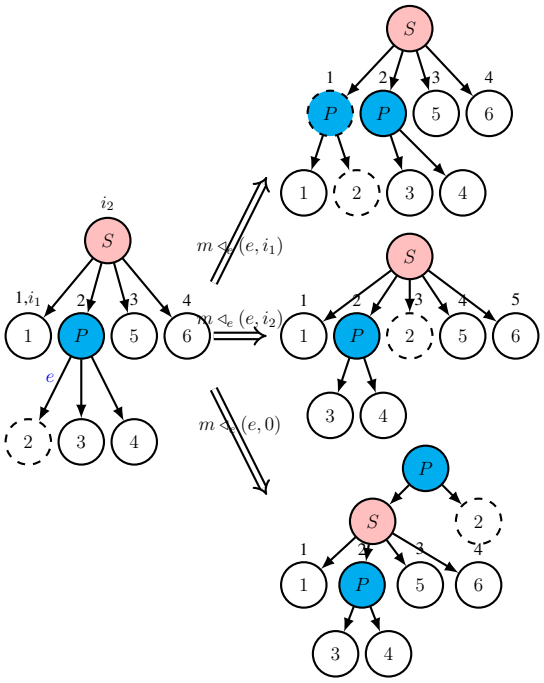


Figure C.2: Some possible operations on the MDT m_1 from Fig. 4. The edge e connected to leaf for actor 2 is reconnected to leaf node i_1 (where it must give a new P node as its neighbor, the parent of i_1, is S), to ancestral node i_2 (and is randomly allocated position 3 among the nodes stacked below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

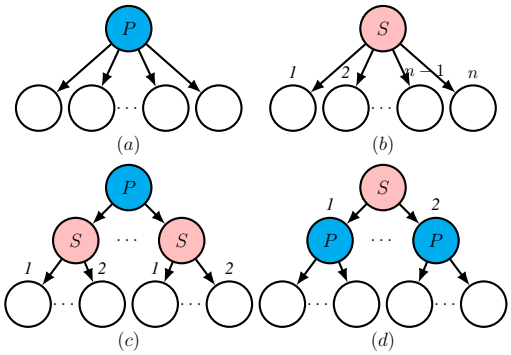


Figure C.3: Four building-block MDT's.

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 $m = m_a \triangleleft_e (e, i)$. 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.3,

omitting the standard q, p and ϕ updates. The algorithm for QJ-U model omits the ϕ-update.

Algorithm C.3 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:

$$
\begin{aligned}
&m^{(k)} \sim \pi(m|y, q, p, \phi),\\
&q^{(k)} \sim \pi(q|y, m^{(k)}, p, \phi),\\
&p^{(k)} \sim \pi(p|y, m^{(k)}, q^{(k)}, \phi),\\
&\phi^{(k)} \sim \pi(\phi|y, m^{(k)}, q^{(k)}, p^{(k)})
\end{aligned}
$$

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* ϕ *omitted*

The queue-jumping probability $p > 0$ (almost surely) so the Hastings ratio $\eta > 0$ in Algorithm C.3 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.

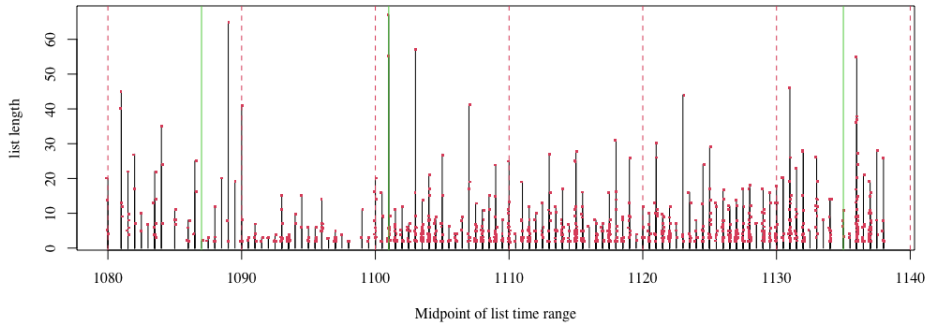


Figure D.2: The midpoint of list time range v.s. list range. Each red dot is a list of length y created in a time range midpointed by x. The bars represents the length of the longest list at time x.

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).

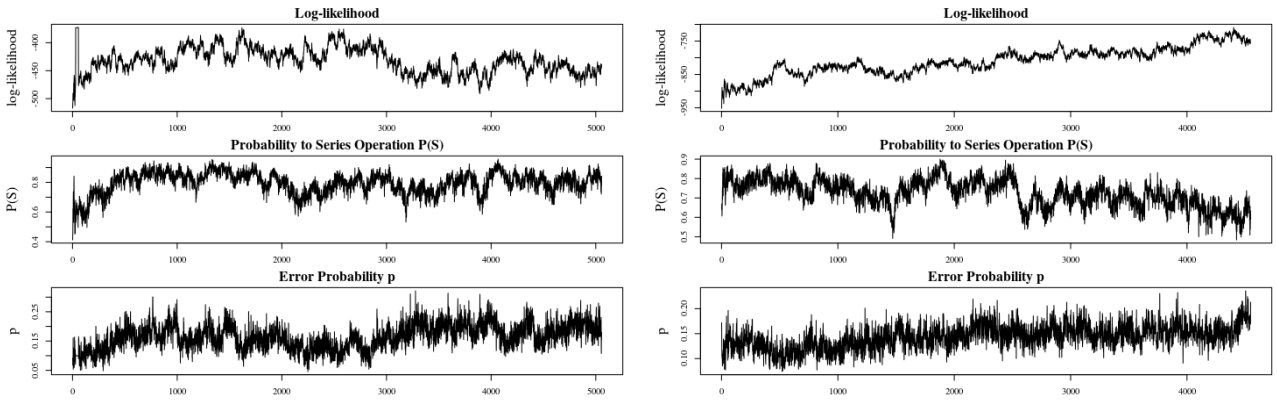
D.1.1 Inference Results on List Data with 1LPA (QJ-U Observation Model)

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.

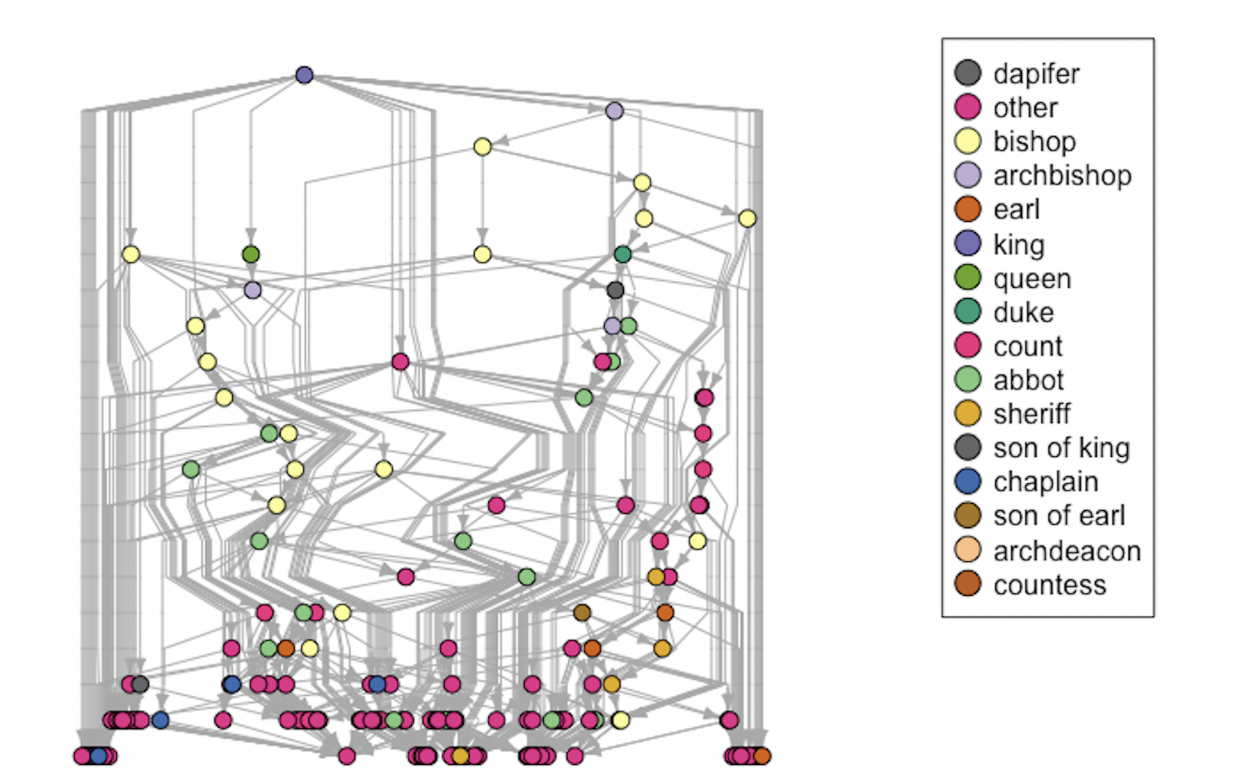


Figure D.4: The consensus order for 1080-1084 (1LPA) data in a VSP/QJ-U analysis.

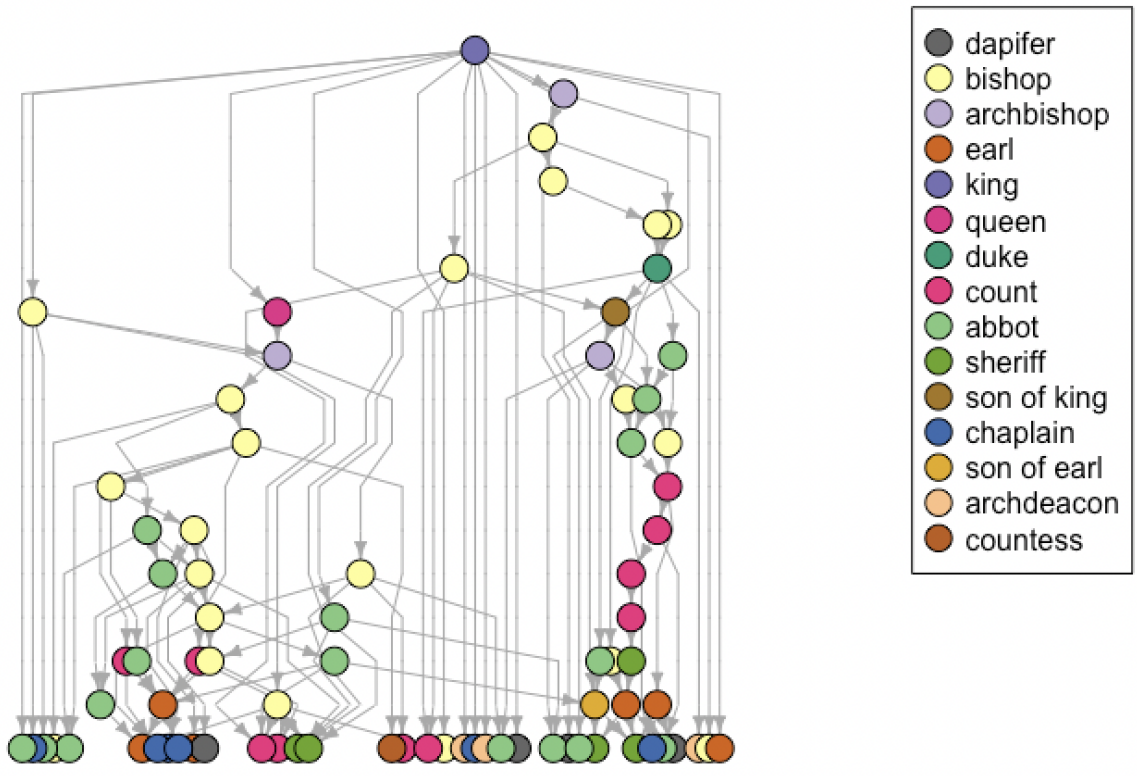


Figure D.6: The consensus order for 1080-1084 (1LPA) data without 'other' actors in a VSP/QJ-U analysis.

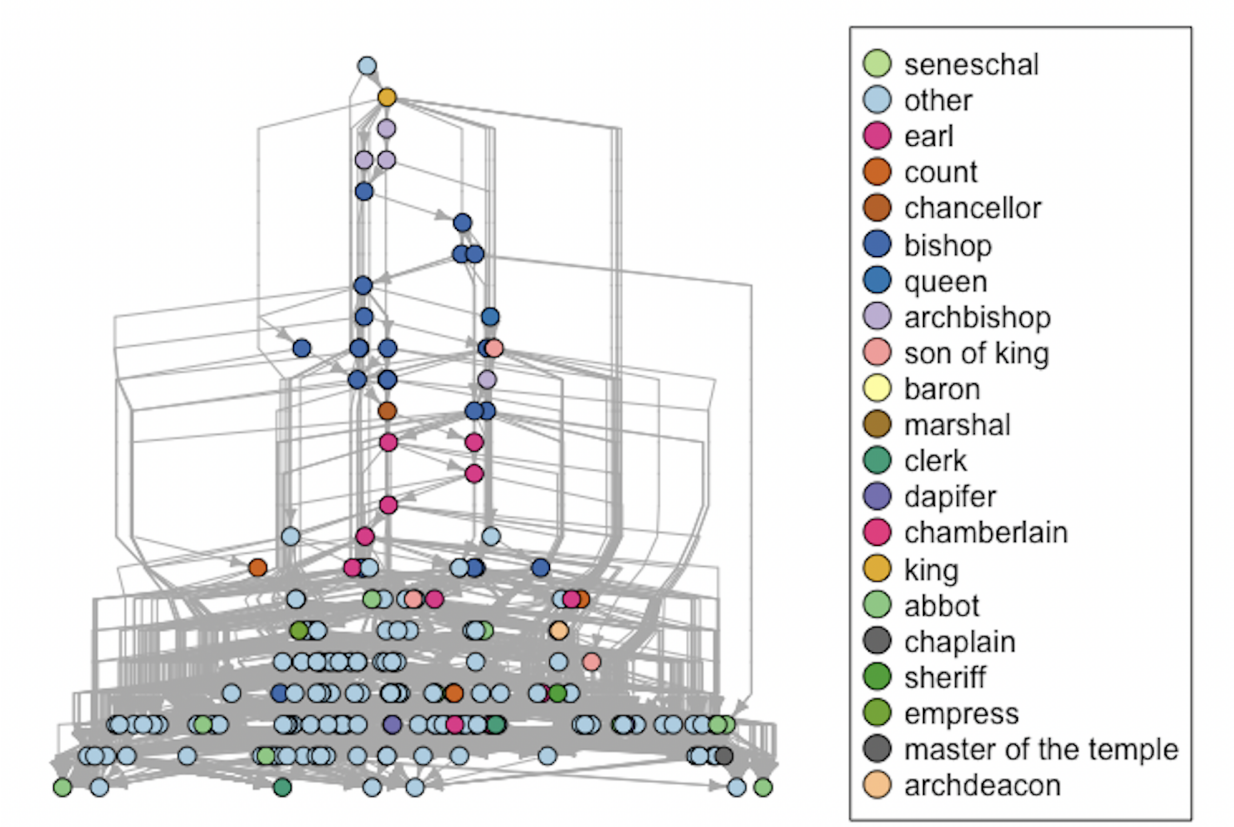


Figure D.5: The consensus order for 1134-1138 (1LPA) data in a VSP/QJ-U analysis.

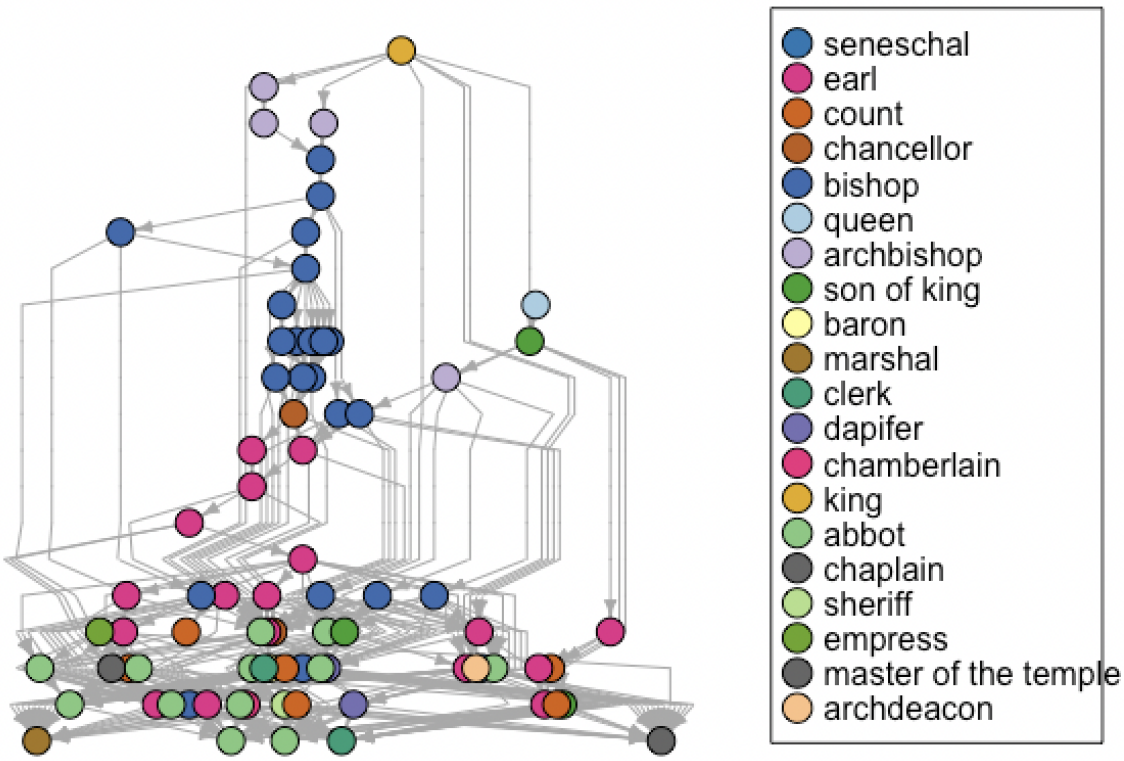


Figure D.7: The consensus order for 1134-1138 (1LPA) data without ‘other’ actors in a VSP/QJ-U analysis.

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.

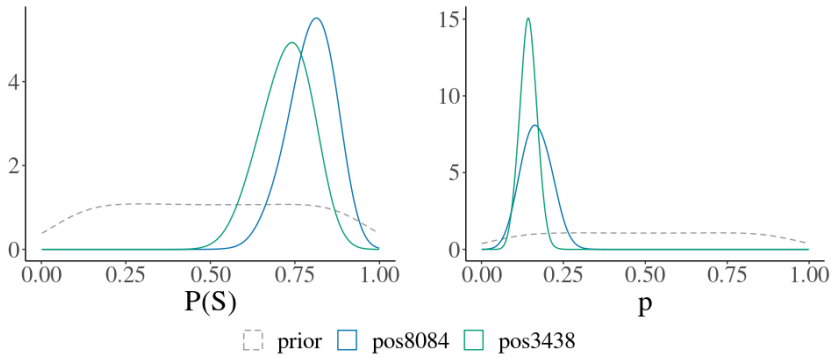


Figure D.8: Prior (grey line) and posterior distributions for $q = P(S)$ (left) and error probability p (right) for the time periods 1080-1084 (1LPA) (blue) and 1134-1138 (1LPA) (green) in a VSP/QJ-U analysis.

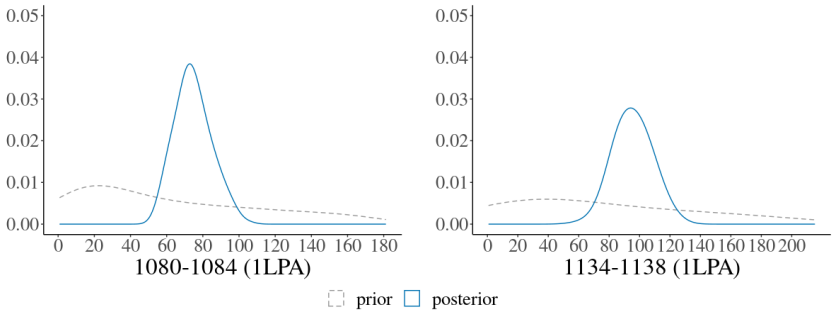


Figure D.9: The prior (grey) and posterior (blue) VSP depth distribution for 1080-1084 (1LPA) (left) and 1134-1138 (1LPA) (right) in a VSP/QJ-U analysis.

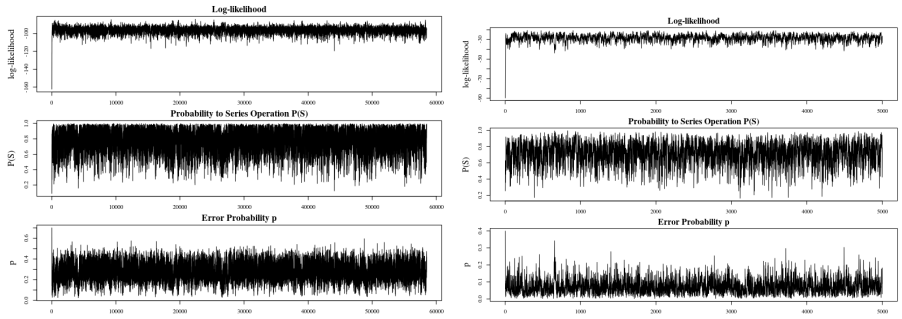
D.1.2 Inference Results on List Data with 5LPA (QJ-U Observation Model)

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

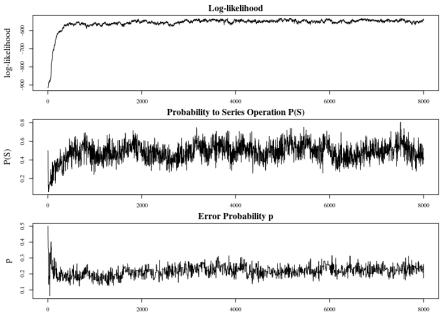
Table D.4: The effective sample sizes for $P(S)$ and error probability p on the four datasets with 5LPA and QJ-U.

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

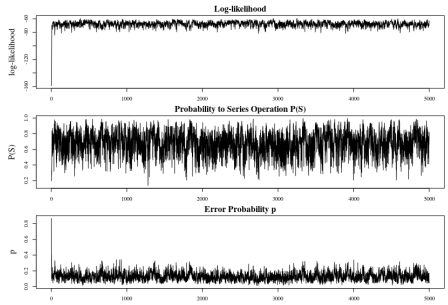


(a) 1080-1084 with 5 LPA

(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.

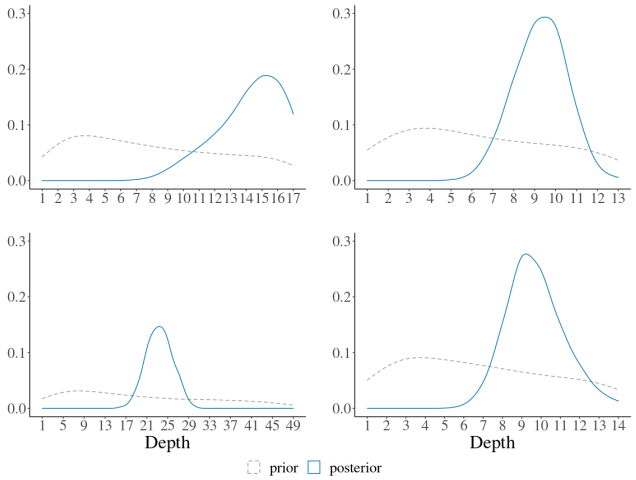


Figure D.11: The prior (grey) and posterior (blue) VSP depth distribution for 1180-1184 (top-left), 1126-1130 (top-right), 1134-1138 (bottom-left) and 1134-1138(b) (bottom-right) with 5LPA and QJ-U.

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.

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.

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

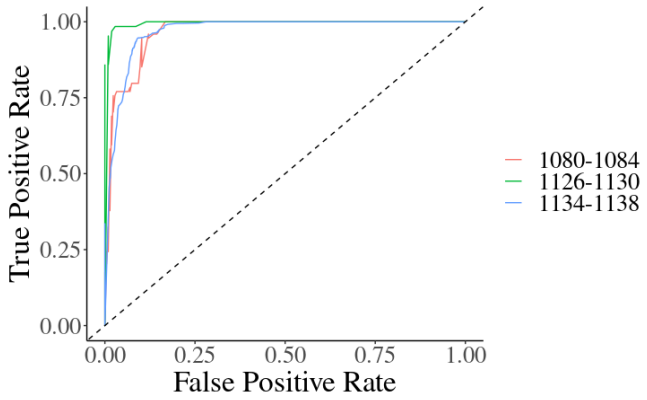
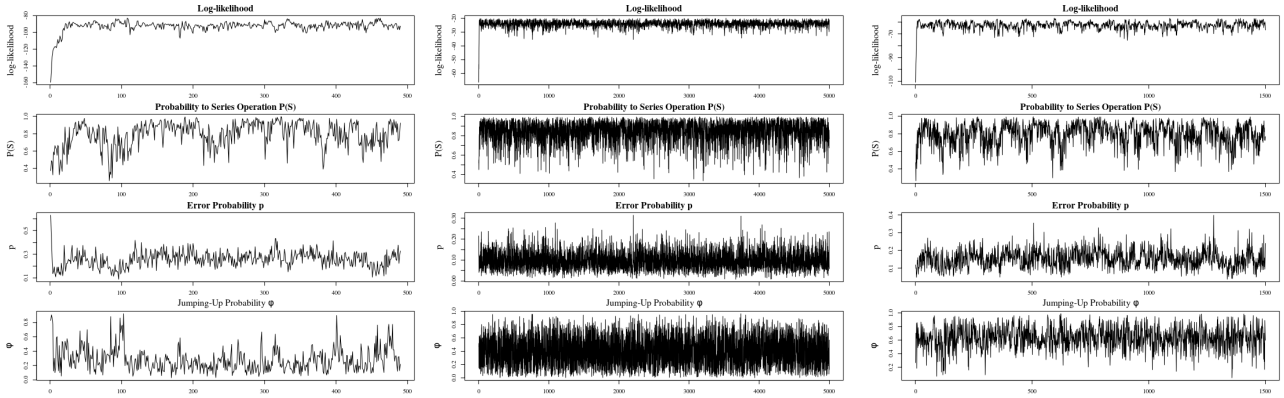


Figure D.12: Receiver operating characteristic (ROC) curves for synthetic data using 1080-1084, 26-30 and 34-38 list membership structures with 5LPA and QJ-U.

	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.



(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 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 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.

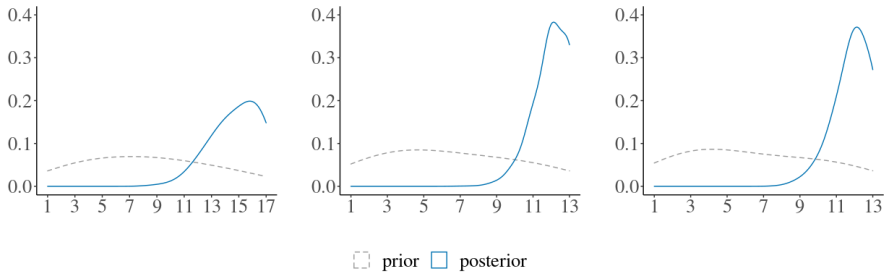


Figure D.14: The prior (grey) and posterior (blue) VSP depth distribution for 1180-1184 (left), 1126-1130 (middle) and 1134-1138(b) (right) with 5LPA data in a VSP/QJ-B analysis.

	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.

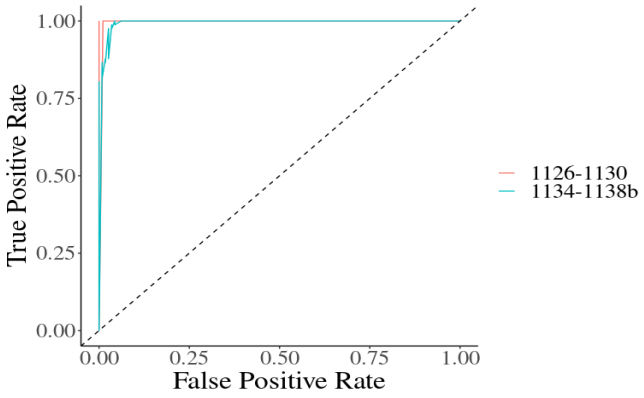


Figure D.15: Receiver operating characteristic (ROC) curves for synthetic data using 1126-1130 and 1134-1138 (bishop) list membership structures with 5LPA and QJ-B.

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

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'.

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".

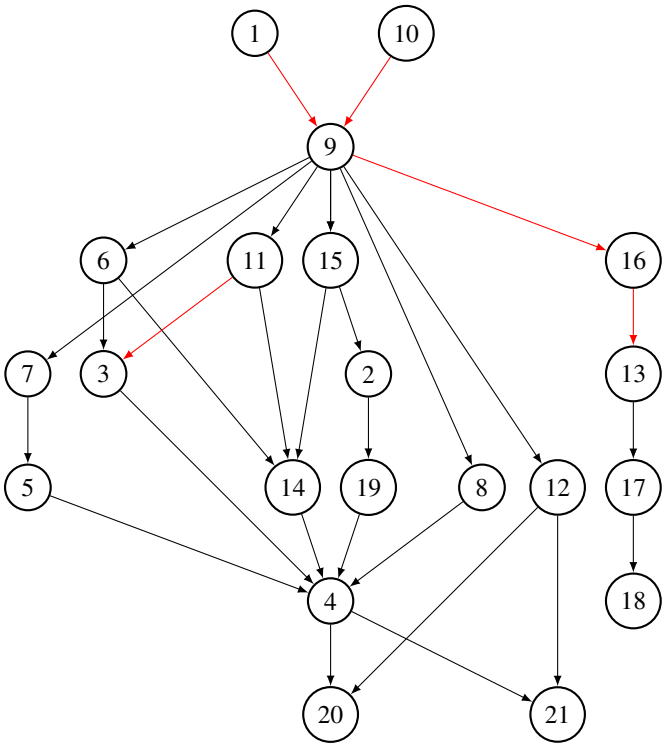


Figure D.16: VSP/QJ-D model. Consensus order for Formula 1 (season 2021) data. Significant/strong order relations are indicated by black/red edges respectively.

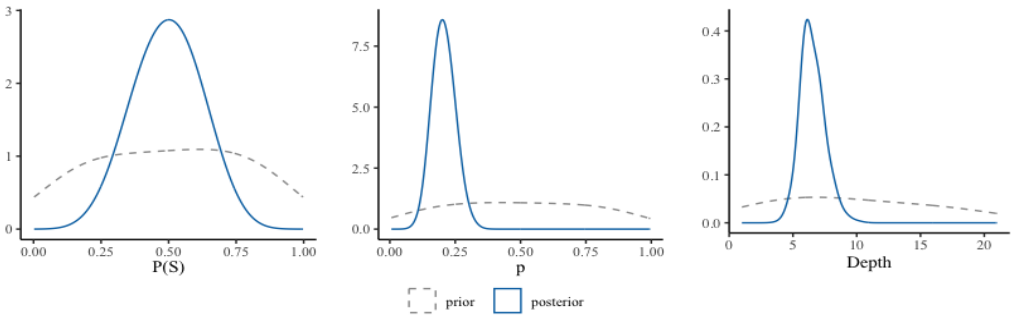


Figure D.17: The prior (grey) and posterior (blue) distributions for $P(S)$ (left), p (middle) and depth (right) for the Formula 1 (season 2021) data.

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 $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$. 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 $\omega = (\omega_1, \ldots, \omega_D)$ 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 $\omega_1, \ldots, \omega_D \sim Dir(1, \ldots, 1)$.

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 taken for ρ, as is suggested in Sørensen et al. [2020] which implements the MCMC proposed in Vitelli et al. [2018].

[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].

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 Mallows, Mallow's mixture, PL and PL-mixture significantly in most cases, except that the Mallow's model outperforms VSP/QJ-U slightly during 1080-1084. 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, similar to our Bayesian inference setting on partial orders.

	$elpd_{waic}$ (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)
Mallows	-174.3 (39.4)	-98.4 (17.7)	-129.6 (33.6)
Mallows-Mix	-225.4 (47.5) (D=2)	-241.3 (37.4) (D=5)	-182.0 (40.6) (D=2)
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)

Table E.1: The estimated $elpd_{waic}$ (se) under six different models - VSP/QJ-U, VSP/QJ-B, Mallow's, Mallow's Mixture (with number of mixture component $D \geq 2$), Plackett-Luce (PL) and 2-mixture Plackett-Luce (PL-Mix2) model.

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.

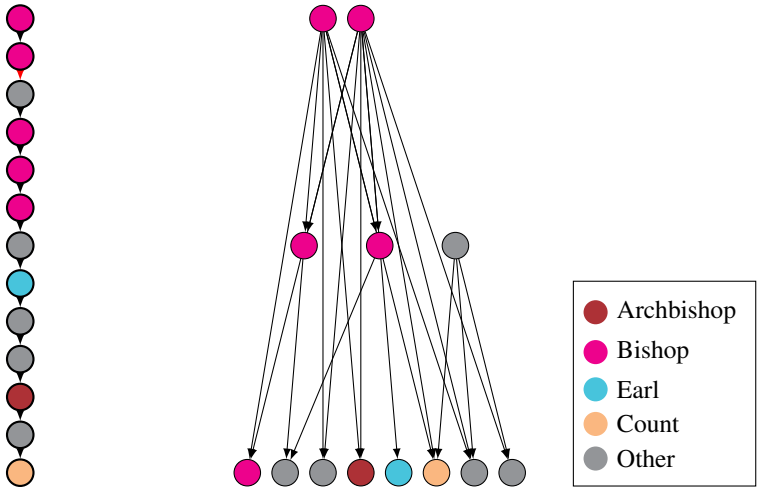


Figure E.1: The estimated consensus orders from the Plackett-Luce (left) and PL-Mixture (D=2) (right) models on the 1126-1130 data. Red edges indicate order relations that posterior probabilities are higher than 0.9.

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.

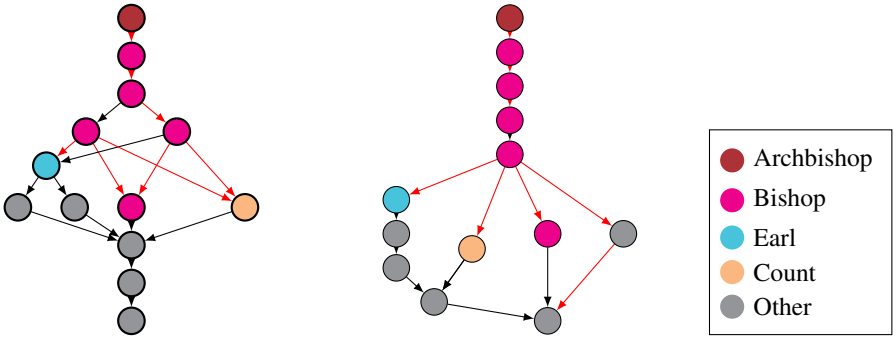


Figure E.2: PO/QJ-U model(left) and VSP/QJ-U model (right; same as Fig. 7). Consensus order for 1126-1130 5LPA data. Significant/strong order relations are indicated by black/red edges respectively.

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.

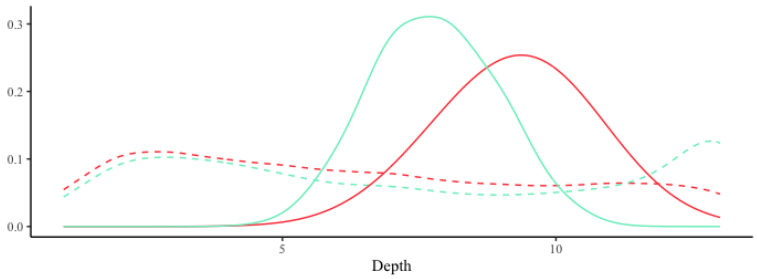


Figure E.3: The prior (dashed) and posterior (solid) distribution over depth for the PO/QJ-U (green) and VSP/QJ-U model (red).

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.

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

	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.

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.

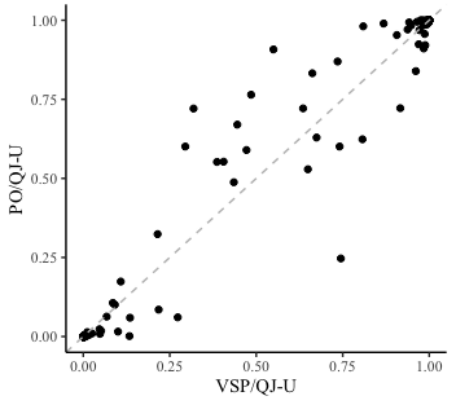


Figure E.4: The comparison plot between the consensus adjacency matrices from the VSP/QJ-U (x-axis) and PO/QJ-U (y-axis) models. The gray dashed line is the $y = x$ reference line.

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).

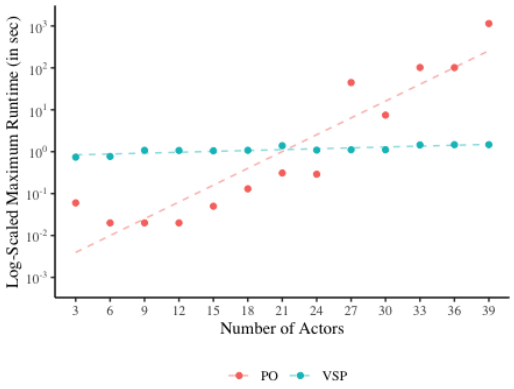


Figure F.1: Run-time analysis between the count approach from tree representation and *LEcount* (Kangas et al. [2016]) on VSPs. The plot compares likelihood (QJ-U) evaluation exploiting the VSP structure (in green) and for a general PO (in red). The log-scaled maximum run-time (in seconds) from the tree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

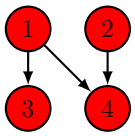


Figure G.1: The 'forbidden sub-graph' to the VSP class of partial orders.

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.

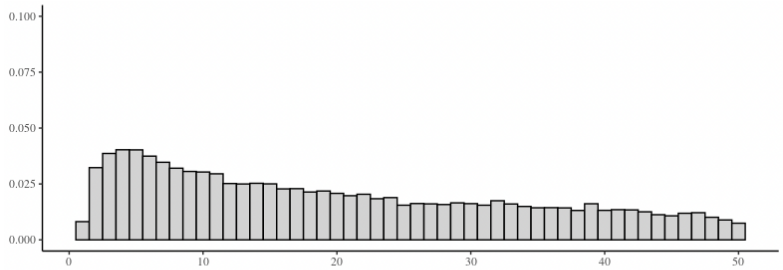


Figure H.1: The prior distribution over depth for partial orders with 50 actors, when $q = \frac{1}{1+e^{-\eta}}, \eta \sim \mathcal{N}(1, 1.5)$.

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.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Bayesian Inference for Vertex-Series-Parallel Orders
Publication Status	■Published □Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Jiang, C., Nicholls, G.K. and Lee, J.E., 2023, July. Bayesian inference for vertex-series-parallel partial orders. *In Uncertainty in Artificial Intelligence* (pp. 995-1004). PMLR.

Student Confirmation

Student Name:	Chuxuan (Jessie) Jiang		
Contribution to the Paper	Joint development of models and theory, performed all simulation and model experiments, drafted manuscript.		
Signature 姜楚璇		Date	06/08/2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof Geoff Nicholls		
Supervisor comments		
Signature G.K. Nicholls.	Date	06/08/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Partial Order Hierarchies

Partial Order Hierarchies

Chuxuan (Jessie) Jiang, Geoff K. Nicholls and Jeong Eun Lee

August 28, 2024

Abstract

Rank-order data often comes from multiple assessors, creating a "grouped" structure in the data. Assessors may treat some items as incomparable and in that setting it is necessary to represent their preferences using a partial order, rather than a total order as is usual. This paper introduces a new hierarchical model for partial orders (HPO) which is designed for grouped preference data with some incomparable items. The partial orders representing each assessor's preferences are correlated with a central partial order in a tree-like hierarchy with a leaf for each assessor. The rank-orders from each assessor are modeled as linear extensions of the the assessor's partial order, with or without noise. The proposed hierarchical partial order model is marginally consistent and nests the Plackett-Luce mixture model. We extend the model to handle unsupervised settings, where the assessor-group labels on rank-orders are not available, leading to a new non-parametric clustering process over rank-order lists. An MCMC-based approach is developed for model inference. We put forward our new model as a natural candidate for generic heterogeneous rank-order data.

However, inference with the HPO-model does not scale well for large datasets. This paper introduces an approximation to the HPO-model using vertex-series-parallel partial orders (VSPs, [1]). The approximation exploits the high scalability of VSPs, demonstrating performance comparable to the full HPO model at much lower asymptotic computational cost. We illustrate our method on 3-D sound spatialisation data where assessor labels are available and sushi preference data where the group structure is unknown.

1 Introduction

Ranking problems arise in a wide variety of fields, including recommender systems [2], social hierarchies [3], voting [4], and sports competitions [5]. For

a given set of lists which rank items, the goal of the analysis is to form a complete or partial ranking of the items which summarises relations between items. When items are ranked multiple times according to slightly different criteria, as is often the case with multiple assessors, a group structure inevitably emerges in the rank-order lists. This type of rank-order data is referred to as "grouped" data. Mixture models for rank-data with assessor labels, such as the Plackett-Luce [6, 7, 8] and Mallows mixture models [9, 10, 11] fit complete orders (or skills scores which define a complete order). However, recent work on rank-order time series [12] and in single-assessor settings [13, 14] has shown that partial order models, which allow incomparability between items, may be superior in some contexts. Therefore, extending partial order models to handle grouped data is well motivated.

Let $\mathcal{M}$ be the "ground set" of items with $|\mathcal{M}| = M$. Let $\mathcal{B}_\mathcal{M} = \cup_{S\subset\mathcal{M}, S\neq\emptyset} S$ be the set of all non-empty subsets of $\mathcal{M}$. A (strong) *partial order* $h = (\mathcal{M}, \succ_h)$ is a binary relation $\succ_h$ over $\mathcal{M}$. The binary relation $\succ_h$ is irreflexive (the relation $i \succ_h i$ does not exist), transitive (if $i \succ_h j$ and $j \succ_h k$, then $i \succ_h k$, where $i, j, k \in \mathcal{M}$ and $i \neq j \neq k$) and antisymmetric ($i \succ_h j$ implies $j \prec_h i$). Figure 1 (left) gives an example of a partial order with five items. Two items $i, j \in \mathcal{M}$ are *incomparable*, and we write $i \|_h j$, if neither $i \prec_h j$ nor $i \succ_h j$. In a compact notation, $\succ_h$ is equivalently a set containing all the pairs $i \succ_h j$ in the ground set which are ordered by h. Denote by $\mathcal{H}_\mathcal{M}$ the set of all partial orders on $\mathcal{M}$.

The preferences of an assessor are represented by some partial order $h \in \mathcal{H}_\mathcal{M}$ particular to that assessor. If $i \succ_h j$ then they prefer i to j and if $i \|_h j$ then they are indifferent. An assessor may be presented with a subset $S \in \mathcal{B}_\mathcal{M}$ of the items in $\mathcal{M}$ for ranking. They will then use the *sub-order* $h[S] = (S, \prec_h)$ of their partial order to rank items. This restricts h to a subset S: all order relations between elements in S present in h are inherited by $h[S]$ and the rest discarded. If for all pairs $i, j \in \mathcal{M}$ either $i \succ_h j$ or $j \succ_h i$ then h is a *complete order*. Let $\mathcal{C}_\mathcal{M}$ be the set of all complete orders of $\mathcal{M}$. Complete orders are one to one with $\mathcal{P}_\mathcal{M}$, the set of all permutations of the elements of $\mathcal{M}$.

A *linear extension* of h is any complete order $l \in \mathcal{C}_\mathcal{M}$ satisfying $j_1 \succ_h j_2 \Rightarrow j_1 \succ_l j_2$, so the linear extension "completes" the partial order $\succ_h$. Denote by $\mathcal{L}[h]$ the set of all linear extensions of h (always non-empty, for this and other classical properties of partial orders see [15] and references therein). In an abuse of notation, we also treat a linear extension as an ordered list $l_{1:M} \in \mathcal{P}_\mathcal{M}$ indexing $l_i,\ i = 1, \dots M$ so that $l_1 \succ_l l_2 \succ_l \cdots \succ_l l_M$. We use the terms list and complete order interchangeably.

A *chain* of $h \in \mathcal{H}_\mathcal{M}$ is any sub-order $h[S],\ S \in \mathcal{B}_\mathcal{M}$ that is also a complete order. The *length* of a chain is the number of nodes $|S|$ in the sub-order.

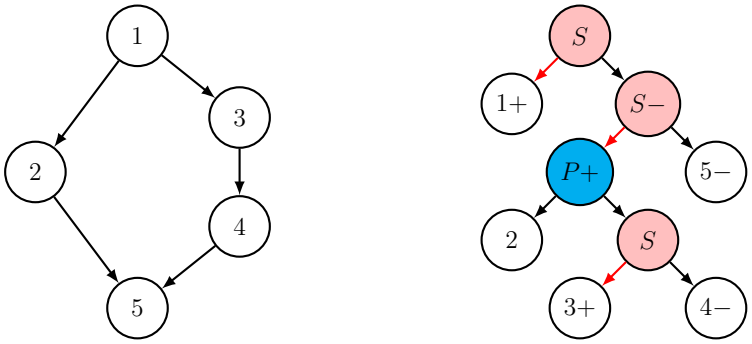


Figure 1: (left) An example partial order v_0 with $M = 5$ items and depth $D(v_0) = 4$ which is also a VSP, with (right) its BDT representation. Red edges and "+" signs indicate the upper child.

The *depth* $D(h)$ of a partial order is the length of its longest chain, $1 \leq D(h) \leq M$ for $h \in \mathcal{H}_{\mathcal{M}}$. The *intersection order* of $K \in \mathbb{Z}^+$ complete orders $l^{(k)} \in \mathcal{C}_{\mathcal{M}},\ k \in [K] = \{1, ..., K\}$ is a partial order $h = (\mathcal{M}, \cap_{k=1}^{K} \prec_{l^{(k)}})$, so $i \succ_h j$ if and only if $i \succ_{l^{(k)}} j$ in every complete order $k \in [K]$. We define the *dimension* of partial order h to be the smallest number of linear extensions of h whose intersection is h (at most $\lfloor M/2 \rfloor$ and at least one).

This work considers the case where we have grouped data from a set of "assessors" $\mathcal{A} = \{1, \ldots, A\}$. Each assessor $a \in \mathcal{A}$ is presented with N_a choice sets $S_{a,i} \in \mathcal{B}_{\mathcal{M}}\ i \in [N_a]$ and produces N_a lists $Y_{a,i} \in \mathcal{C}_{S_{a,i}}$ of possibly varying lengths. We parameterise the preferences of assessor a using a partial order $h^{(a)}$ so we suppose these lists are linear extensions of $h^{(a)}$, or close to being linear extensions if there are errors in the observation process. Let $Y_a = (Y_{a,1}, \ldots, Y_{a,N_a})$ be the preference orders from assessor $a \in \mathcal{A}$ and let $Y = (Y_1, \ldots, Y_A)$ denote all the data. An assessor may not see all the items in $\mathcal{M}$. For $a \in \mathcal{A}$ let $\mathcal{M}_a = \cup_{i=1}^{N_a} S_{a,i}$ be the set of all items they see across all their choice sets. Let $\mathcal{M}_0 = \cup_{a \in \mathcal{A}} \mathcal{M}_a$ be the set of all items ranked by anyone. We only directly learn about assessor a's preferences over items in $\mathcal{M}_a$ and overall we only learn about items in $\mathcal{M}_0$ (in the sound spatialisation example each assessor is presented with pairs of sounds to rank, so $|S_{a,i}| = 2$ for each choice set, however they are all presented with all $M = 12$ sounds eventually, so $\mathcal{M}_a = \mathcal{M}_0 = \mathcal{M}$). We propose a hierarchical partial order model for these grouped data. The model is defined using a tree-like parameter hierarchy, where each assessor's preferences over $\mathcal{M}$ are represented by a "leaf" partial order $h^{(a)} \in \mathcal{H}_{\mathcal{M}_a}$. We assume there is a "root" or global partial order $h^{(0)} \in \mathcal{H}_{\mathcal{M}_0}$ that governs the preference behavior among the individual assessors. This hierarchical setup is presented as a tree in Figure 3.

Partial orders, with properties such as transitivity and incomparability, are well-suited for ranking problems. However, inference with partial orders can be computationally intensive, largely because the likelihood for a partial order in our setup depends on the number of linear extensions it has, and counting the linear extensions of a general partial order (calculating $|\mathcal{L}[h]|$) is #P-complete [16]. This computational challenge limits the applicability of our partial order hierarchical model to no more than perhaps 15-25 items. To enhance model scalability, we consider a special sub-class of partial orders called vertex-series-parallel orders (VSPs, [17, 1]). VSPs are constructed through repeated series and parallel operations. Let $h_1 \in \mathcal{H}_{S_1}$ and $h_2 \in \mathcal{H}_{S_2}$ be two partial orders defined on disjoint choice sets $S_1, S_2 \in \mathcal{B}_{\mathcal{M}}$. We define their *series partial order* by the *series* operation $\otimes$ as the union of all relations in h_1 and h_2, with additional relations $i \succ_h j$ if $i \in S_1$ and $j \in S_2$, so

$$h = h_1 \otimes h_2 = \{S_1 \cup S_2,\ \succ_{h_1} \cup \succ_{h_2} \cup \{i \succ_h j : i \in S_1, j \in S_2\}\},$$

and a *parallel partial order* by the *parallel* operation $\oplus$ as the union of all relations in h_1 and h_2 with incomparability $j \|_h k$ if $i \in S_1$ and $j \in S_2$ so

$$h = h_1 \oplus h_2 = \{S_1 \cup S_2,\ \succ_{h_1} \cup \succ_{h_2}\}.$$

The class of VSPs on $\mathcal{M}$, $\mathcal{V}_{\mathcal{M}} \subset \mathcal{H}_{\mathcal{M}}$, is defined as the set of all partial orders which can be formed by some sequence of series and parallel operations starting from M trivial partial orders $\{(\{i\}, \emptyset)\}_{i \in \mathcal{M}}$ each with a single element. This is strictly smaller than the set of all partial orders when $M \geq 4$.

A VSP $v \in \mathcal{V}_{\mathcal{M}}$ can be conveniently parameterized using a binary decomposition tree (BDT, [17]) $t \in \mathcal{T}_{\mathcal{M}}$. A BDT t representing a VSP v can be thought of as a record of a sequence of parallel and serial operations which form v, starting from the M trivial partial orders. Each leaf in a BDT represents an item $i \in \mathcal{M}$, and each internal node represents a partial order operation, series (S) ($\otimes$) or parallel (P) ($\oplus$), on the VSPs defined by the two sub-trees rooted by its child nodes. Define a map $f_{\mathcal{V}} : \mathcal{T}_{\mathcal{M}} \to \mathcal{V}_{\mathcal{M}}$ from a BDT to its corresponding VSP, such that $f_{\mathcal{V}}(t) = v$. Note that a VSP may be represented by more than one BDT, so $f_{\mathcal{V}}$ is not a bijection. [14] derive a simple expression for the number of trees in the set $f_{\mathcal{V}}^{-1}(v)$. The example partial order in Figure 1 is also a VSP. One of its possible BDT representations is shown alongside it. VSPs allow for fast linear extension counting, with linear computational complexity. This motivates us to approximate the HPO-model by restricting the fitted partial orders to be VSPs. The resulting Hierarchical VSP-model (HVSP) is a potentially useful model in its own right. However, we think of it as an approximation rather than a first-choice model for reasons discussed below. We compare its performance with

the original partial order hierarchy and show that the VSP approximation approach performs well, with a significant gain in computational efficiency.

This paper is structured as follows. In section 2, we introduce a prior probability distribution over a single partial order (so not yet a hierarchy). This will be our model for the "global" partial order at the top of the hierarchy. This prior distribution is similar to that given in [13] but has been modified so that it includes the Plackett-Luce model as a special case. It is straightforward to incorporate covariates. In section 3, we define the conditional distribution over the "leaf" partial orders given the global partial order and provide a prior distribution for the partial order hierarchy when the assessor labels are known. Some candidate observation models and Bayesian inference are discussed in section 4. Section 5 addresses the scenario where assessor labels are unknown, extending the HPO-model to cluster unlabeled rank data with a latent group structure. In section 6, we give our VSP approximation for the HPO-model. In section 7 we compare the performance of the HVSP-model and the HPO-model using labelled 3-D sound spatialisation data and unlabelled sushi preference data as test cases.

2 Distribution over complete and Partial Orders

2.1 The Plackett-Luce Distribution

Most distributions over rank-order data are defined on complete orders, which assume a complete, transitive ranking of all items without ties or incomparabilities. Two popular families of such distributions are the Plackett-Luce distribution and the Mallows distribution.

The Plackett-Luce distribution, often used in the context of sports rankings, elections, and recommender systems, models the probability of a particular ranking by considering a series of conditional probabilities. Each item is assigned a positive weight, and the probability of an item being chosen at each rank depends on its weight relative to the remaining unranked items.

Definition 1 (Plackett-Luce Distribution [18, 19])**.** *Define preference weights* $\alpha_{\mathcal{M}} = (\alpha_1, \ldots, \alpha_M)$, $\alpha_i \in \mathbb{R}$ *for choice* $i \in \mathcal{M}$*. The Placket-Luce distribution defines the probability for a complete order* $y \in \mathcal{C}_S$ *of a choice set* $S \in \mathcal{B}_{\mathcal{M}}$ *with* $|S| = m$ *elements as*

$$p_S(y|\alpha_S) = \prod_{i=1}^{m} \frac{e^{\alpha_{y_i}}}{\sum_{i'=i}^{m} e^{\alpha_{y_{i'}}}}, \tag{1}$$

where $\alpha_S = (\alpha_j)_{j \in S}$ *and the list* $y_{1:m}$ *is indexed so that* $y_1 \succ_y y_2 \succ_y \cdots \succ_y y_m$. *Denote by* $PL(\alpha, S)$ *the distribution in equation 1.*

This multiplicative structure makes the Plackett-Luce model particularly suitable for modeling sequential choice processes. [20] establishes a connection between the Plackett-Luce distribution and the Gumbel distribution. Specifically, the ranking of $|S|$ independent Gumbel-distributed variables follows the Plackett-Luce distribution.

Lemma 1 (Gumbel representation of the Plackett-Luce distribution [20])**.** *Let* $G_j \sim$ *Gumbel*(α_j) *be independent for* $j \in \mathcal{M}$, *where Gumbel*(α_j) *is a distribution with CDF* $F_{\alpha_j}(g) = \exp(-\exp(-(g - \alpha_j)))$, $g \in \mathbb{R}$. *Let* $G = (G_1, \ldots, G_M)$ *and let* $y(G) = (\mathcal{M}, \succ_G)$ *give the complete order for the elements of* G, *that is* $j_1 \succ_G j_2 \Leftrightarrow G_{j_1} > G_{j_2}$. *It holds that* $y(G) \sim PL(\alpha, \mathcal{M})$.

In section 2.2, we extend this parameterisation to the context of partial orders, incorporating the Plackett-Luce distribution as a special case within this broader framework.

The Plackett-Luce distribution is also marginally consistent. A family of distributions is *marginally consistent* (also known as *projective*) if every marginal of every distribution in the family is also a member of the family. In the context of rank-order data, marginal consistency implies that: if sample a random order on a subset S of items by sampling a random order on all items $\mathcal{M}$ and pulling out the sub-order for S, then we get the same distribution as if we just sampled a random order on the items in S in the first place. The property is equivalent to context-independent choice.

Definition 2 (Marginal consistency)**.** *The family of distributions* $p_S(y|\alpha)$, $y \in \mathcal{C}_S$, $S \in \mathcal{B}_\mathcal{M}$ *is marginally consistent if, for all* $S \in \mathcal{B}_\mathcal{M}$, *distributions in the family satisfy*

$$p_S(y|\alpha) = \sum_{\substack{\tilde{y} \in \mathcal{C}_\mathcal{M} \\ \tilde{y}[S] = y}} p_\mathcal{M}(\tilde{y}|\alpha) \quad \forall y \in \mathcal{C}_S. \tag{2}$$

Equivalently, if $\tilde{y} \sim p_\mathcal{M}(\cdot|\alpha)$ *and* $y \sim p_S(\cdot|\alpha)$ *then* $\tilde{y}[S] \sim y$.

It follows from the definition that if $S \subset \tilde{S} \subset \mathcal{M}$ then p_S is the marginal of $p_{\tilde{S}}$, so marginal consistency holds between all pairs of distributions in the family. [21] shows that the Plackett-Luce distribution is marginally consistent by evaluating the sum in equation (2). It also follows from Lemma 1 (removing components from G doesn't change the order or distribution of those which remain).

2.2 Distribution over Partial Orders

The literature on distributions over partial orders is limited in comparison to that for complete orders, and can mainly be found in the literature related to combinatorics and probability distributions over random graphs. See [15] for an overview. Some authors work with the uniform distribution over partial orders $\pi^{(u)}_{\mathcal{M}}(h) = |\mathcal{H}_{\mathcal{M}}|^{-1}$ (employed as a prior in [22]). [23] develops a latent variable model which we call the Uniform-(K, n) partial order model. Further refinement comes from [13], who extend this model by incorporating a depth control parameter ρ, leading to the Normal-(K, n, ρ) partial order model. The latent variable setup makes it easy to bring covariate effects into the prior, as demonstrated in [12]. Here we modify the Normal-(K, n, ρ) model by making a special choice for the distribution of its latent variables. This ensures that our model includes the Plackett-Luce distribution as a special case. We call this the Gumbel-(K, n, ρ) partial order model (PO-model).

Let $U \in \mathbb{R}^{M\times K}$ be a matrix of preference weights with one row $U_{j,\cdot} \in \mathbb{R}^K$ for each object $j \in \mathcal{M}$ and one column $U_{\cdot,k}$ for each "feature" $k = 1, \ldots, K$. Define a mapping $h(U) : \mathbb{R}^{M\times K} \to \mathcal{H}_{\mathcal{M}}$ such that $h(U) = (\mathcal{M}, \succ_U)$ gives a partial order with $i \succ_U j$, if and only if $U_{i,k} > U_{j,k}\ \forall k = 1, \ldots, K$ for a pair of objects $i, j \in \mathcal{M}$. Each column of U defines a complete order $h(U_{\cdot,k}) = (\mathcal{M}, \succ_{U_{\cdot,k}})$ with

$$\succ_{U_{\cdot,k}} = \{i \succ_{U_{\cdot,k}} j : U_{i,k} > U_{j,k}, i, j \in \mathcal{M}\},$$

the ranking of the entries in column k. We equivalently define $h(U)$ as the intersection over the orders of the columns of U, since $i \succ_{h(U)} j \iff i \succ_{h(U_{\cdot,k})} j$ for all $k = 1, \ldots, K$. The map from matrix to partial order is

$$h(U) = \left(\mathcal{M},\ \cap_{k=1}^{K} \succ_{U_{\cdot,k}}\right). \tag{3}$$

Below we modify the normal U-matrix representation of [13] to include the Plackett-Luce distribution, introducing covariates using a similar approach to [12]. In Theorem 1, U is an $M \times K$ matrix of correlated normal random variables which are pushed through a monotone transformation to get a matrix η of correlated Gumbel random variables, one row for each element in the ground set. The rows of η determine the partial order $h(\eta)$ using the rule in equation (3). The construction in Theorem 1 is illustrated in Figure 2.

Theorem 1 (Generative Model for Gumbel-(K, n, ρ) Partial Order Model)**.** *Let $G^{-1}(g) = -\log(-\log(g))$ be the inverse CDF of a standard Gumbel random variable and let Φ be the CDF of a standard normal. Let X be an $M \times p$ design matrix of p covariates with a row x_j for each $j \in \mathcal{M}$, and let $\beta \in \mathbb{R}^p$ be the vector of effect values so that $\alpha = X\beta \in \mathbb{R}^{M\times 1}$ and $\alpha_j = x_j^T\beta$ is the*

linear predictor for $j \in \mathcal{M}$. Let Σ_ρ be a $K \times K$ covariance matrix with unit diagonal $(\Sigma_\rho)_{k,k} = 1$ and constant off diagonal $(\Sigma_\rho)_{k,k'} = \rho$ for $\rho \in [0, 1)$, $k, k' \in \{1, \ldots, K\}$ and $k \neq k'$. We take

$$U_{j,\cdot} \sim N(0_K, \Sigma_\rho), \quad \textit{independent for each } j \in \mathcal{M}, \tag{4}$$

$$\eta_{j,\cdot} = G^{-1}(\Phi(U_{j,\cdot})) + \alpha_j 1_K^T, \quad \textit{and} \tag{5}$$

$$h = h(\eta(U, \beta)), \tag{6}$$

where the functions G^{-1} and Φ are applied to their arguments element by element. Two properties hold for the Gumbel-(K, n, ρ) partial order model.

1. *It holds that $h(\eta_{\cdot,k}) \sim PL(\alpha, \mathcal{M})$ in (1) for each $k = 1, \ldots, K$.*
2. *If data $y \sim \mathcal{U}(\mathcal{L}[h(\eta)])$ are drawn uniformly at random from the linear extensions of $h(\eta)$ then $y \sim PL(\alpha, \mathcal{M})$ when $K = 1$.*

Proof. See Appendix A. □

Figure 2 gives an example of the $G^{-1}(\Phi(U))$ latent matrix and demonstrates the effect of covariates α. In equation (5), $\eta = G^{-1}(\Phi(U)) + \alpha 1_K^T$ is a linear predictor that incorporates the covariate effect by shifting the matrix of preference weights. As each row $\eta_{j,\cdot} = G^{-1}(\Phi(U_{j,\cdot})) + \alpha_j 1_K^T$, a large positive effect $\alpha_j = x_j^T \beta$ for object j tends to move j up in the partial order. We then define the partial order given the shifted preference weight matrix $h = h(\eta)$. We do not include an intercept among the covariates as that duplicates the degree of freedom corresponding to the mean of $U_{j,\cdot}$.

Equations (4) and (5) determine a prior distribution for $h \in \mathcal{H}_S$ for each choice set $S \in \mathcal{B}_\mathcal{M}$, $|S| = m$. Let $U_{j,\cdot} \sim N(0_K, \Sigma_\rho)$ be independent for $j \in S$ so $U \in \mathbb{R}^{m \times K}$ and let $X_S = (X_{j,\cdot})_{j \in S}$ be the submatrix of covariates for choices in S. The latent η for the choices in S is

$$\eta(U, \beta) = G^{-1}(\Phi(U)) + X_S \beta \, 1_K^T,$$

an $m \times K$ matrix with rows $\eta_{j,\cdot}$, $j \in S$. Operators are applied to each matrix element. The random partial order $h(\eta(U, \beta))$ has distribution

$$\pi_S(h|\rho, \beta) = E_U(\mathbb{I}_{h(\eta(U,\beta))=h}), \quad h \in \mathcal{H}_S. \tag{7}$$

This is normalised over the space $\mathcal{H}_S$ of partial orders on S. The prior distribution displays marginal consistency as is shown in corollary 1. The proof is similar to the case for the Normal-(K, n, ρ) family of partial order models given in [12].

Corollary 1. *The family of prior distributions $\pi_S(\cdot|\rho, \beta)$, $S \in \mathcal{B}_\mathcal{M}$ (7) is marginally consistent, that is if $h \sim \pi_\mathcal{M}(\cdot|\rho, \beta)$, then $h[S] \sim \pi_S(\cdot|\rho, \beta)$.*

Proof. For proof see Appendix B. □

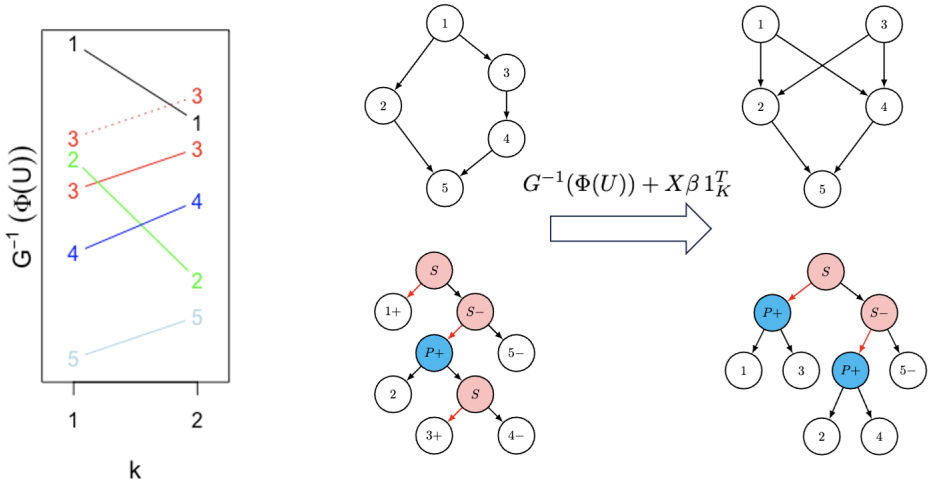


Figure 2: (left) A U-matrix with $M = 5$ rows and $K = 2$ columns representing partial order (also a VSP) v_0 in Figure 1 (first VSP right). Each path plots the sequence of values $G^{-1}(\Phi(U_{j,k})), k = 1, 2$ in a row of feature values for choice $j \in \mathcal{M}$. The path for choice 1 lies entirely above that for choice 2 so $1 \succ_{v_0} 2$. However the path for choice 4 intersects the path for choice 2 so $2\|_{v_0}4$. The effect of covariate is illustrated by the dotted line. If $\alpha_3 > 0$ for choice 3 and $\alpha_j = 0$ elsewhere, choice 3's latent path moves up. As is shown in v_1 (second VSP right), a positive α_3 leads to an intersection between choices 1 and 3 resulting in $1\|_{v_1}3$ after adjusting for covariates.

3 A hierarchical model for grouped data

Grouped data often involve collecting multiple preference rankings from a set of assessors, resulting in a structured dataset where preferences are organized into groups. To model such data, we propose a hierarchical partial order model, which follows a tree-like structure as illustrated in Figure 3. In this hierarchical model, we assume the existence of a "global" partial order that governs the overall preference structure across all assessors. Our goal is to recover this "global" order and understand its influence on individual assessors' rankings. Each assessor's preferences are modeled using a "leaf" partial order. These leaf partial orders are connected to the root partial order through a correlation-like parameter. We give the conditional distribution of the leaf partial orders given the root partial order and define a prior distribution for the hierarchical partial order model (HPO).

3.1 Notation: Grouped Data

Recall the data notation introduced in Section 1. Let $\mathcal{A} = \{1, \ldots, A\}$ be a set of assessors. In grouped data, each assessor $a \in \mathcal{A}$ is presented with N_a choice sets $S_{a,i} \in \mathcal{B}_{\mathcal{M}}, i = 1, \ldots, N_a$ of varying sizes. We condition on the choice sets so they are non-random. Let $m_{a,i} = |S_{a,i}|$ be the number of items in the i'th batch ordered by assessor a, and let $\mathcal{M}_a = \cup_{i=1}^{N_a} S_{a,i}$ be the set of objects assessor a ranked with $m_a = |\mathcal{M}_a|$. Let $S_a = (S_{a,1}, \ldots, S_{a,N_a})$ and $S = (S_1, \ldots, S_A)$. Let $\mathcal{M}_0 = \cup_{a \in \mathcal{A}} \mathcal{M}_a$ be the set of items ranked by at least one assessor. We set $M_a = |\mathcal{M}_a|, a = 0, 1, \ldots, A$. Let $Y_{a,i} \in \mathcal{C}_{S_{a,i}}$ be the complete order assessor a returned for choice set $S_{a,i}$. Let $Y_a = (Y_{a,1}, \ldots, Y_{a,N_a})$ be the data from assessor a. We denote the data as $Y = (Y_1, \ldots, Y_A)$.

3.2 A prior for the partial order hierarchy

Let $U^{(0)} \in \mathbb{R}^{M_0 \times K}$ be the preference weight matrix with a row for each element $j \in \mathcal{M}_0$. We define $h^{(0)} = h(\eta(U^{(0)}, \beta))$ as the global partial order on items in $\mathcal{M}_0$ which were actually ranked by someone. For $K \geq 1$ let

$$\mathcal{H}_{\mathcal{M}_0}^{(K)} = \bigcup_{\eta \in \mathbb{R}^{M_0 \times K}} \{h(\eta)\}$$

be the set of all partial orders on $\mathcal{M}_0$ which can be represented by the model (which is the set of all partial orders on $\mathcal{M}_0$ of dimension at most K, since $h(\eta)$ is the intersection of K complete orders, which can be any K orders depending on η). Correspondingly, we define $U^{(a)} \in \mathbb{R}^{M_a \times K}$ as preference weight matrices for the leaf partial orders $h^{(a)}$ on $\mathcal{M}_a, a \in \mathcal{A}$. We write $U = (U^{(0)}, U^{(1)}, \ldots, U^{(A)})$ and $h(\eta(U, \beta)) = (h^{(0)}, h^{(1)}, \ldots, h^{(A)})$. Let

$$\mathcal{M}_{0:A} = (\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_A) \text{ and } \mathcal{H}_{\mathcal{M}_{0:A}}^{(K)} = \mathcal{H}_{\mathcal{M}_0}^{(K)} \times \mathcal{H}_{\mathcal{M}_1}^{(K)} \cdots \times \mathcal{H}_{\mathcal{M}_A}^{(K)}$$

so that $h(\eta(U, \beta)) \in \mathcal{H}_{\mathcal{M}_{0:A}}^{(K)}$. The setup is illustrated in Figure 3.

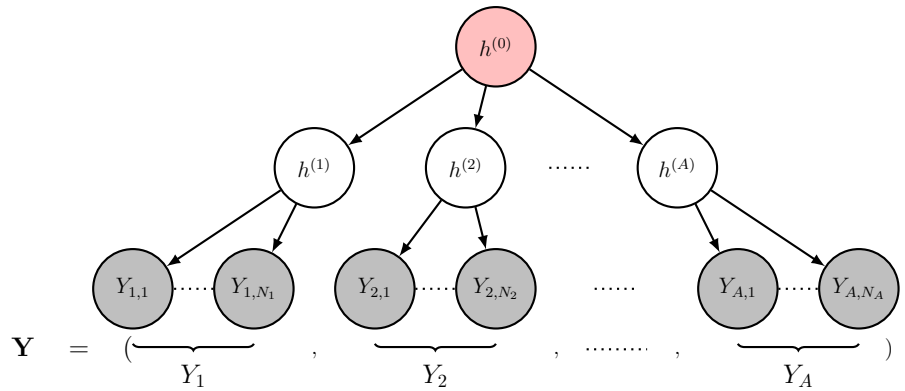


Figure 3: Hierarchical partial order model and grouped data structure.

Definition 3 (Generative Model for Partial Order Hierarchy (HPO))**.** *Let* $0 \leq \rho \leq 1$, $\beta \in \mathbb{R}^p$, $K \geq 1$ *and* $0 < \tau \leq 1$ *be given. For* $\alpha_{\mathcal{M}} = X\beta$ *and* Σ_ρ *as in Section 2, and sets* $\mathcal{M}_a \in \mathcal{B}_{\mathcal{M}}$, $a \in \mathcal{A}$, *let* $\mathcal{M}_0 = \cup_{a=1}^{A} \mathcal{M}_a$. *The generative model* $h \sim \pi_{\mathcal{M}_{0:A}}(h|\rho, \beta, \tau)$, $h \in \mathcal{H}^{(K)}_{\mathcal{M}_{0:A}}$ *is*

$$U^{(0)}_{j,\cdot} \sim N(0, \Sigma_\rho), \text{ independent for each } j \in \mathcal{M}_0, \tag{8}$$

$$U^{(a)}_{j,\cdot} | U^{(0)}_{j,\cdot} \sim N\left(\tau U^{(0)}_{j,\cdot}, (1-\tau^2)\Sigma_\rho\right), \text{ independent for } (j,a) \in \mathcal{M}_a \times \mathcal{A}, \tag{9}$$

$$\eta^{(a)}_{j,\cdot} = G^{-1}(\Phi(U^{(a)}_{j,\cdot})) + \alpha_j 1_K^T \text{ for each } (j,a) \in \mathcal{M}_a \times (\{0\} \cup \mathcal{A}), \tag{10}$$

$$h = h(\eta(U,\beta)) \quad \text{for } \eta(U,\beta) = (\eta^{(0)}, \eta^{(1)}, \ldots, \eta^{(A)}).$$

Notice that the definition of h has changed from section 2.2 as it is now a list of $A+1$ partial orders. The same applies to U, and η and the map $h(\cdot)$ now applies to each $\eta^{(a)}$, $a \in \mathcal{A}$ in the list. We defined the HPO-model so that its partial orders only order items ranked in the data. This works because marginal consistency is preserved (see theorem 2 below). We could enlarge each partial order in the hierarchy so that it orders all the elements of $\mathcal{M}$. However the HPO-prior given above would be the marginal of that more general model, and since the missing elements are never observed, they won't appear in the likelihood. It follows that the posterior for the smaller model will be the marginal of the posterior in the larger model. This saves us ordering elements about which we have no information.

The prior for $U^{(a)}_{j,\cdot}$ given $U^{(0)}_{j,\cdot}$ in (9) is controlled by a correlation-like parameter τ. As $\tau \to 0$ the leaf partial orders $h^{(1:A)}$ become independent of $h^{(0)}$. As $\tau \to 1$ they tend in probability to $h^{(0)}$ itself. The process linking $U^{(a)}_{j,\cdot}$ to $U^{(0)}_{j,\cdot}$ is equivalent to running a K-dimensional Ornstein-Uhlenbeck process $dX_t = -X_t + V^{1/2} dW_t$ with $V = 2\Sigma_\rho$ starting at $X_0 = U^{(0)}_{j,\cdot}$ to get $X_t \sim U^{(a)}_{j,\cdot}$ at $t = -2\log(\tau)$. The stationary distribution of this process is $N(0_K, \Sigma_\rho)$ so if we run the process for no time at all ($\tau = 1$) we get back $U^{(0)}_{j,\cdot}$ and if we run it for an infinite amount of time ($\tau \to 0$) we get an independent draw from $N(0_K, \Sigma_\rho)$. Properties of the HPO-prior are given in Theorem 2.

Theorem 2 (Hierarchical partial order Prior)**.** *The generative model in definition 3 defines a prior distribution over hierarchical partial orders as*

$$\pi_{\mathcal{M}_{0:A}}(h|\rho,\beta,\tau) = E_U(\mathbb{I}_{h(\eta(U,\beta))=h}), \quad h \in \mathcal{H}^{(K)}_{\mathcal{M}_{0:A}}, \tag{11}$$

where $\pi_{\mathcal{M}_{0:A}}(h|\rho,\beta,\tau)$ *is normalised over* $h \in \mathcal{H}^{(K)}_{\mathcal{M}_{0:A}}$. *Let*

$$\pi_{\mathcal{M}_{0:A}}(h^{(a)}|\rho,\beta,\tau) = E_U(\mathbb{1}_{h(\eta^{(a)}(U^{(a)},\beta))=h^{(a)}})$$

be the marginal for $h^{(a)}$. *The prior* $\pi_{\mathcal{M}_{0:A}}$ *has the following properties:*

1. *(single PO marginal) For* $a = 0, \ldots, A$, $\pi_{\mathcal{M}_{0:A}}(h^{(a)}|\rho, \beta, \tau) = \pi_{\mathcal{M}_a}(h^{(a)}|\rho, \beta)$ *where* $\pi_{\mathcal{M}_a}$ *is the single partial-order prior given in* (7)*;*

2. *(PL hierarchy at* $K = 1$*) when* $K = 1$, $h^{(a)} \sim PL(\alpha_{\mathcal{M}_a}, \mathcal{M}_a)$ *for* $a \in \mathcal{A}$*;*

3. *(*$h^{(a)}$ *independent of* $h^{(0)}$ *at* $\tau = 0$*) when* $\tau = 0$ *the prior factorises,*

$$\pi_{\mathcal{M}_{0:A}}(h|\rho, \beta, \tau = 0) = \pi_{\mathcal{M}_0}(h^{(0)}|\rho, \beta, \tau) \prod_{a=1}^{A} \pi_{\mathcal{M}_a}(h^{(a)}|\rho, \beta, \tau);$$

4. *(*$h^{(a)} \to h^{(0)}$ *as* $\tau \to 1$*) If* $\mathcal{M}_a = \mathcal{M}$ *for each* $a = 0, 1, \ldots, A$ *then, for each* $a \in \mathcal{A}$, $\pi_{\mathcal{M}_{0:A}}(h^{(a)}|\rho, \beta, \tau, U^{(0)}) \to \mathbb{1}_{h^{(a)}=h^{(0)}}$ *as* $\tau \to 1$*;*

5. *(marginal consistency) If* $\mathcal{M}_a = \mathcal{M}$ *for each* $a = 0, 1, \ldots, A$ *and* $h \sim \pi_{\mathcal{M}^{A+1}}(\cdot|\rho, \beta, \tau)$ *then* $h[s_{0:A}] \sim \pi_{s_{0:A}}(\cdot|\rho, \beta, \tau)$ *for every* $s_{0:A} \in \mathcal{B}_{\mathcal{M}}^{A+1}$ *with* $s_0 = \cup_{a=1}^{A} s_a$*.*

Proof. See Appendix C. □

In the HPO-model, the "root-to-leaf" correlation parameter $\tau \in (0, 1]$ controls the similarity between the global root partial order and the individual leaf partial orders. A larger value of τ implies a stronger correlation, meaning that the leaf partial orders will more closely resemble the root partial order. This parameter allows us to adjust the degree of alignment between the global and local preferences, providing a way to model varying levels of influence from the global structure on individual assessors. Also, our HPO model includes the Plackett-Luce mixture model as a special case when $K = 1$. The HPO-model generalizes the Plackett-Luce mixture model by incorporating hierarchical structure and depth control, thereby extending its applicability to more complex scenarios involving grouped data.

When we come to fit the HPO-model we will work with the latent U process in the posterior. We extract the partial order hierarchy in post-processing by applying the map $h(\eta(U, \beta))$ to MCMC samples of U and β from the posterior. The U-process prior is

$$\pi(U|\rho, \tau) = \pi(U^{(0)}|\rho) \prod_{a \in \mathcal{A}} \pi(U^{(a)}|U^{(0)}, \rho, \tau) \tag{12}$$

where

$$\pi(U^{(0)}|\rho) = \left[\prod_{j \in \mathcal{M}_0} N\left(U^{(0)}_{j,\cdot}; 0_K, \Sigma_\rho\right) \right]$$

and

$$\pi(U^{(a)}|U^{(0)},\rho,\tau)=\prod_{a\in\mathcal{A}}\prod_{j\in M_a}N\left(U^{(a)}_{j,\cdot};\tau U^{(0)}_{j,\cdot},(1-\tau^2)\Sigma_\rho\right)$$

The HPO posterior is

$$\begin{aligned}\pi_{\mathcal{M}_{0:A}}(\rho,\beta,\tau,U,p|Y)\propto{}&\pi_R(\rho)\pi_B(\beta)\pi_P(\tau)\pi(p)\quad\times\\&\pi(U|\rho,\tau)p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)),p).\end{aligned}\tag{13}$$

The column dimension K of U and the number of assessors/groups are fixed.

4 Observation Model and Bayesian Inference

We now temporarily return to the setting where we have a single partial order $h\in\mathcal{H}_\mathcal{M}$ and consider a single observation $y\in\mathcal{C}_\mathcal{M}$. We seek an observation model for y (with or without error) given h. In the simplest case y is a linear extension of h drawn uniformly at random, so $p(y|h)=|\mathcal{L}[h]|^{-1}$. Alternatively we may observe a noisy copy of the linear extension; [24] gives Plackett-Luce [18, 19] and Mallows observation models for this layer of "noise". In the present paper we use the original "queue-jumping" observation model introduced in [13]. This model incorporates two types of errors: "queue-jumping-up" (QJ-U) and "queue-jumping-down" (QJ-D). In this framework, an observed rank-order y is formed sequentially by taking items off the top of a queue which respects h, so items which are at the top of many linear extensions are more likely to be at the head of the queue. However, as the queue is read from top to bottom, items may jump the queue (with some small probability), ignoring the constraints in h. This seems to give a reasonable balance of realism and computational efficiency. [14] shows this leads to efficient computation when the partial order is a VSP.

Let $L(h)=|\mathcal{L}[h]|$ be the number of linear extensions of $h\in\mathcal{H}_\mathcal{M}$ and for $i\in\mathcal{M}$ let $T_i(v)=|\{l\in\mathcal{L}[h]:l_1=i\}|$ give the number of linear extensions with item i at the top. Let p be the error probability for an item to jump to the head of the queue each time an item is read from the head of the queue. The observation model for QJ-U for a generic list $y\in\mathcal{C}_\mathcal{M}$ (in list notation $y_{1:M}$) is a product over the sequence of steps,

$$Q_{up}(y|h,p)=\prod_{i=1}^{M-1}\left(\frac{p}{M-i+1}+(1-p)\frac{T_{y_i}(h[y_{i:M}])}{L_T(h[y_{i:M}])}\right).\tag{14}$$

This reduces to the noise free case above $Q_{up}(y|h,p=0)=|\mathcal{L}[h]|^{-1}$ when p is zero. See [14] for the QJ-B observation model with distribution $Q_{down}(y|h,p)$ in which the queue is read sequentially from the bottom.

In the context of the HPO-model, let Q represent either Q_{up} or Q_{down}. Since $Y_{a,i}$ is observed on the choice set $S_{a,i}$ it is informed by the suborder $h^{(a)}[S_{a,i}]$ for assessor $a \in \mathcal{A}$ and $Y_{a,i} \in \mathcal{C}_{S_a,i}$. The likelihood is

$$p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)),p) = \prod_{a=1}^{A}\prod_{i=1}^{N_a} Q(Y_{a,i}|h^{(a)}[S_{a,i}],p). \tag{15}$$

The full posterior distribution for the HPO-model is therefore

$$\pi_{\mathcal{M}_{0:A}}(\rho,\beta,\tau,U,p|Y) \propto \pi_R(\rho)\pi_B(\beta)\pi_P(\tau)\pi(p)\pi(U|\rho,\tau)p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)),p),$$

with $\pi(U|\rho,\tau)$ given in equation (12) and $p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)),p)$ given in equation (15). Priors for other parameters are given below. We elicit the prior distributions as $\rho \sim Beta(1,\kappa_\rho), \log(\frac{p}{1-p}) \sim \mathcal{N}(0,\kappa_p)$ and $\tau \sim \mathcal{U}[0,1]$ where κ_ρ and κ_p are hyperparameters.

We use Metropolis-Hasting MCMC to target the posterior distribution over the latent U-matrices. Each parameter is updated sequentially. We update the array of preference weights matrices as follows.

At step t, let $U_t = (U_t^{(0)}, U_t^{(1)}, \ldots, U_t^{(A)})$.

1. Sample $j \in \mathcal{M}_0$ at random.
2. Define U' so that $U'_{i,\cdot} = (U'^{(0)}_{i,\cdot}, U'^{(1)}_{i,\cdot}, \ldots, U'^{(A)}_{i,\cdot}) \in (\mathbb{R}^K)^{A+1}$ for each $i \in \mathcal{M}_0$ and sample $U'_{j,\cdot}$ from the priors in definition 5 with
   - $U'^{(0)}_{j,\cdot} \sim N(0,\Sigma_\rho)$,
   - $U'^{(a)}_{j,\cdot} \sim N(\tau U'^{(0)}_{j,\cdot}, (1-\tau^2)\Sigma_\rho)$,

   and $U'_{i,\cdot} = U_{t,i,\cdot}, \forall i \in \mathcal{M}_0 \setminus j$.
3. Set $U_{t+1} = U'$ with probability $\frac{\pi(U'|\rho,\tau)p_{\mathcal{M}_{1:A}}(Y|h(\eta(U',\beta)),p)}{\pi(U|\rho,\tau)p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)),p)}$ else $U_{t+1} = U_t$.

This has given acceptable rejection rates in problems studied to date.

5 Clustering unlabeled orders

Section 3.2 assumes the assessor labels $a \in \mathcal{A}$ are available, i.e. we have prior knowledge about which assessor produced each observed rank-order. However, in many practical cases, the orders come from a population of orders with an unknown group structure and must be clustered. In this setting the "assessor" labels are just group labels and these are unknown. The number

of groups $|\mathcal{A}| = A$ is also unknown. For N observed rank-orders, we define latent labels $c = (c_1, \ldots, c_N)$ with $c_i \in \mathcal{A}, i \in [N]$ and let the i'th observation come from group c_i. We model the random partition using a Poisson-Dirichlet process [25]. The Poisson-Dirichlet process $PDP(\eta_\theta, \eta_\alpha)$ controls the clustering distribution with a *strength parameter* η_θ and a *discount parameter* η_α. Define partition $\xi = (\xi_1, \ldots, \xi_A) \in \Xi_{[N]}$ where $\xi_a = \{i \in [N] : c_i = a\}, a \in \mathcal{A}$, and $\Xi_{[N]}$ is the set of all partitions of $[N]$. The PDP-prior for ξ is

$$\pi_\Xi(\xi) = \frac{\Gamma(\eta_\theta)}{\Gamma(\eta_\theta + N)} \frac{\eta_\alpha^{|\xi|}\Gamma(\eta_\theta/\eta_\alpha + |\xi|)}{\Gamma(\eta_\theta/\eta_\alpha)} \prod_{\zeta \in \xi} \frac{\Gamma(|\zeta| - \eta_\alpha)}{\Gamma(1 - \eta_\alpha)}. \tag{16}$$

Incorporating clustering modifies the HPO-model by introducing dependence on the partition ξ without changing the basic hierarchical structure. We modify data indexing to handle clustering: the rank-order data is now $Y = (Y_1, \ldots, Y_N)$ with $Y_i \in \mathcal{C}_{S_i}$ so the choice set for Y_i is S_i, $i \in [N]$. Let $\mathcal{M}_a = \cup_{i \in \xi_a} S_i$. The number of clusters is $A(\xi) = |\xi|$ and $\mathcal{A}(\xi) = [A(\xi)]$. Let $Y_{\xi_a} = (Y_i)_{i \in \xi_a}$ be the orders in partition ξ_a. The likelihood is

$$p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)), p, \xi) = \prod_{a=1}^{A(\xi)} \prod_{i \in \xi_a} Q(Y_i|h^{(a)}[S_i], p), \tag{17}$$

where $h^{(a)} = h(\eta(U^{(a)}, \beta))$ as before. We illustrate the Poisson-Dirichlet HPO model (PDP-HPO) in Figure 4.

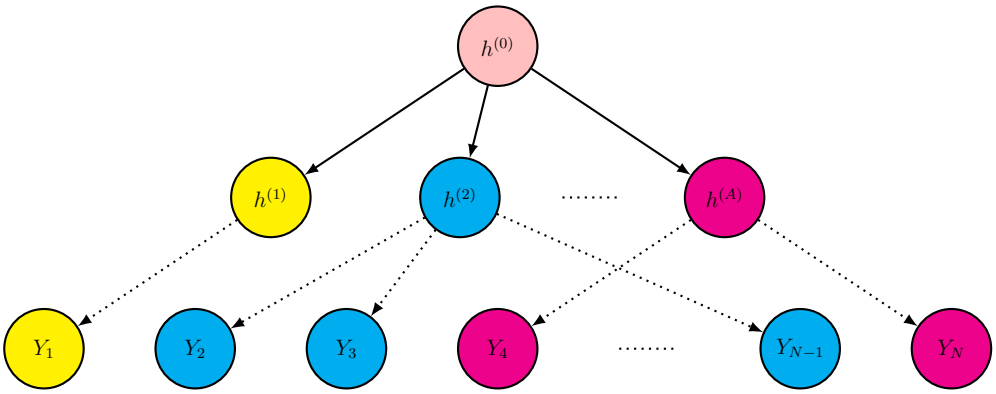


Figure 4: An example partial order hierarchy clustering N orders in a data set $Y = (Y_1, \ldots, Y_N)$ with order partition $\xi = (\{1\}, \{2, 3, N-1\}, \ldots, \{4, N\})$.

The posterior distribution over the arrays of latent U-matrices in the PDP-HPO hierchical model is therefore

$$\begin{aligned}\pi_{\mathcal{M}_{0:A}}(\rho, \beta, \tau, U, p, \xi|Y) \propto \pi_R(\rho)\pi_B(\beta)\pi_P(\tau)\pi(p)\pi_\Xi(\xi) \quad \times \\ \pi(U|\rho, \tau, \xi)p_{\mathcal{M}_{1:A}}(Y|h(\eta(U,\beta)), p, \xi)\end{aligned} \tag{18}$$

where $\pi(U|\rho,\tau,\xi)$ is the same as $\pi(U|\rho,\tau)$ in equation (12) except the product over $a \in \mathcal{A}$ is now a product over $a \in \mathcal{A}(\xi)$ and we should be aware that the sets $\mathcal{M}_a = \cup_{i\in\xi_a} S_i$ depend on the particular collection of orders in ξ_a.

6 Approximation with VSPs

The HPO and PDP-HPO models for general partial orders given in sections 3 and 5 are natural for "grouped" data, particularly when individual assessors' preferences follow a complex order structure. However, statistical inference for such models can be computationally demanding, as the likelihood evaluation has been shown to be #P-complete [23]. [12] notes that MCMC-based inference on partial orders with more than about 25 items is impractical. To address this, we simply condition the partial orders $h^{(a)} = h(\eta(U^{(a)},\beta)),\ a \in \mathcal{A}$ appearing in the likelihood in equation (13) to be VSPs as this allows us to count linear extensions with linear computational complexity [14]. When we condition the U-distribution to realise VSPs we loose marginal consistency so the think of the resulting hierarchical distribution over VSPs as an approximating distribution and not a model in its own right. Although we restrict to VSP, our prior distribution over partial orders was already weighting in favor of VSPs, especially when the matrix dimension $K = 2$. As shown in Table 1, the proportion of VSPs relative to partial orders is significantly higher under the Gumbel-(K, n, ρ) prior.

Number of Actors	Proportion of VSP under prior	Actual Proportion
5	88.940%	61.241%
10	40.624%	0.4733%
15	20.515%	0.000009%
20	12.800%	$\sim 0\%$

Table 1: The actual proportion of VSP and the proportion of VSPs under the Gumbel-$(K, n, \rho)(K = 2)$ prior given by equations (4) and (7). The number of poset is accessed via [26]. We estimate the number of labeled series-parallel posets according to [27].

The revised setup (for HPO with assessor labels) is the same as shown in Figure 3 but each partial order $h^{(a)}$ in the middle layer is replaced by a VSP $v^{(a)} \in \mathcal{V}_{\mathcal{M}_a}$. At each likelihood evaluation, the leaf VSPs $h(\eta(U^{(a)},\beta))$ are converted to their BDT representation. Since only the leaf partial orders are involved in the likelihood evaluation, we do not restrict the root partial order to be a VSP. This allows the model to retain the flexibility and

richness of general partial orders at the root level, while benefiting from the computational advantages of VSPs at the leaf level.

6.1 Prior for VSPs

Under the VSP-approximation of the partial order hierarchical model (HVSP), we restrict the Gumbel-(K, n, ρ) model with covariates (in theorem 1) to the space of VSPs. [1] shows a VSP has a maximum dimension of two so we represent every VSP in $\mathcal{V}_{\mathcal{M}}$ by taking $K = 2$. However, not all partial orders of dimension two are VSPs. For example, [1] shows that a partial order is a VSP if and only if it does not have the "forbidden" suborder $([4], \{1 \succ_h 3,\ 1 \succ_h 4, 2 \succ_h 4\})$ and the dimension of this suborder is two. We therefore restrict the two-dimensional partial orders we can represent with $K = 2$ to ensure they are VSPs. To achieve this we modify the marginal distribution in equation (7) to give a random VSP $v \in \mathcal{V}_S$ with

$$\pi_S(v|\rho, \beta) \propto E_U(\mathbb{I}_{h(\eta(U,\beta))=v}), \quad v \in \mathcal{V}_S, \tag{19}$$

with $U \in \mathbb{R}^{m \times K}$ as in the setup for equation (7), but equation (19) is normalised over the space of VSPs $\mathcal{V}_S$ on S. We call this the VSP-model and summarise it in definition 4.

Definition 4 (Gumbel-(K, n, ρ)-VSP model)**.** *With parameters defined in theorem 1, and $K = 2$, the generative model for Gumbel-(K, n, ρ)-VSPs is*

$$U_{j,\cdot} \sim N(0_K, \Sigma_\rho), \quad \textit{independent for each } j \in \mathcal{M}, \tag{20}$$

$$\eta_{j,\cdot} = G^{-1}(\Phi(U_{j,\cdot})) + \alpha_j 1_K^T, \quad \textit{and} \tag{21}$$

$$g = h(\eta(U, \beta)), \tag{22}$$

where the functions G^{-1} and Φ are applied to their arguments element by element. If $g \in \mathcal{V}_{\mathcal{M}}$ stop and return $v = g$ and otherwise repeat from (20).

If we took $K = 1$ then g is a complete order and the process never rejects, so the Gumbel-(K, n, ρ)-VSP and Gumbel-(K, n, ρ)-PO distributions coincide. However, we must take $K = 2$ to get a prior distribution over VSPs. Every VSP has dimension less that or equal two, and if U is any U-matrix representing a VSP then it can be realised at equation (20) so when $K = 2$ the prior has support over all VSPs in $\mathcal{V}_{\mathcal{M}}$.

It will be helpful when we come to sample the posterior, to have a MCMC proposal for U which determines a Markov Chain which is irreducible over $\mathcal{V}_{\mathcal{M}}$. Given a VSP $v \in \mathcal{V}_{\mathcal{M}}$, there always exists a preference weight matrix U such that $h(\eta(U, \beta)) = v$. We can construct a Markov chain over the

preference weight matrices $MC_U = (U^0, U^1, \dots)$ that evolves by updating a randomly selected row of the matrix.

Markov Chain MC_U update. Let U^t be the state at step t. Update as follows:

1. sample j uniformly at random from $\mathcal{M}$;
2. sample $U'_{j,\cdot} \sim N(0_K, \Sigma_\rho)$ and set $U'_{i,\cdot} = U^t_{i,\cdot}$ for $i \in \mathcal{M}, i \neq j$;
3. if $h(\eta(U', \beta)) \in \mathcal{V}_\mathcal{M}$, then set $U^{t+1} = U'$ else $U^{t+1} = U^t$.

The VSPs represented by MC_U are

$$(v^0 = h(\eta(U^0, \beta)), v^1 = h(\eta(U^1, \beta)), \dots).$$

It is not clear that this process is irreducible over VSPs due to the possibility of rejection at step 3. In lemma 2 we show that this is the case.

Let $U_0 \in \mathbb{R}^{M \times K}$ be any U-matrix representing a VSP $v_0 \in \mathcal{V}_\mathcal{M}$ with $v_0 = h(\eta(U_0, \beta))$ and let $v_{2M} \in \mathcal{V}_\mathcal{M}$ be any other VSP. Let U_{2M} be any U-matrix satisfying $v_{2M} = h(\eta(U_{2M}, \beta))$. There exists an open ball δU_{2M} in $\mathbb{R}^{M \times K}$ centred on U_{2M} such that, for every $U \in \delta U_{2M}$ we have $v_{2M} = h(\eta(U, \beta))$ (imagine perturbing the line segments of U in Figure 2 by ϵ). Our Markov chain is defined on $\mathbb{R}^{M \times K}$ so a sufficient condition for irreducibility is $\Pr(U^{2M} \in \delta U_{2M} | U^0 = U_0) > 0, \ \forall \ U_0, U_{2M} \in \mathbb{R}^{M \times K}$ representing VSPs.

Lemma 2 (Irreducibility of MC_U)**.** *The Markov chain MC_U over U-matrices constrained to represent VSPs is irreducible.*

Proof. See Appendix D. □

6.2 Hierarchical VSP prior

Definition 5 below presents the VSP approximation for $\pi_{\mathcal{M}_{0:A}}(h|\rho, \beta, \tau), \ h \in \mathcal{H}_{\mathcal{M}_{0:A}}$. We simply take the hierarchical prior in Definition 3 and replace the A partial orders for assessors with VSPs. We achieve this by taking $K = 2$ and constraining $h(\eta(U^{(a)}, \beta)) \in \mathcal{V}_M, \ a \in \mathcal{A}$, so the U-matrix for each assessor represents a VSP. However, we do not require the root partial order to be a VSP. From a computational perspective we compute the likelihoods $p(Y_{a,i}|v^{(a)}, p)$ for each $i = 1, \dots, N_a$ and each $a \in \mathcal{A}$, but there is no likelihood factor involving $h^{(0)}$ so we can retain a little more generality and allow it to be a partial order. The space of hierarchical VSPs is then

$$\mathcal{O}_{\mathcal{M}_{0:A}} = \mathcal{H}^{(2)}_{\mathcal{M}_0} \times \mathcal{V}_{\mathcal{M}_1} \cdots \times \mathcal{V}_{\mathcal{M}_A},$$

so that $\mathcal{O}_{\mathcal{M}_{0:A}} \subset \mathcal{H}^{(2)}_{\mathcal{M}_{0:A}}$. In this setting $v \in \mathcal{O}_{\mathcal{M}_{0:A}}$ is our notation for a VSP hierarchy-state so that $v = (h^{(0)}, v^{(1)}, \ldots, v^{(A)})$.

Definition 5 (VSP Approximation for PO Hierarchy (HVSP))**.** *For $\alpha_{\mathcal{M}} = X\beta$ and Σ_ρ as in Section 2, for $0 < \tau \leq 1$ and sets $\mathcal{M}_a \in \mathcal{B}_{\mathcal{M}},\ a \in \mathcal{A}$, let $\mathcal{M}_0 = \cup_{a=1}^{A} \mathcal{M}_a$. The generative model for $v \sim \pi_{\mathcal{M}_{0:A}}(h|\rho, \beta, \tau),\ v \in \mathcal{O}_{\mathcal{M}_{0:A}}$ is*

$$U^{(0)}_{j,\cdot} \sim N(0, \Sigma_\rho),\ \textit{independent for } j \in \mathcal{M}_0, \tag{23}$$

$$U^{(a)}_{j,\cdot} | U^{(0)}_{j,\cdot} \sim N\left(\tau U^{(0)}_{j,\cdot},\ (1-\tau^2)\Sigma_\rho\right),\ \textit{independent for } (j,a) \in \mathcal{M}_a \times \mathcal{A} \tag{24}$$

$$\eta^{(a)}_{j,\cdot} = G^{-1}(\Phi(U^{(a)}_{j,\cdot})) + \alpha_j 1_K^T \textit{ for each } (j,a) \in \mathcal{M}_a \times (\{0\} \cup \mathcal{A}), \tag{25}$$

$$g = h(\eta(U,\beta)) \quad \textit{for } \eta(U,\beta) = (\eta^{(0)}, \eta^{(1)}, \ldots, \eta^{(A)}).$$

If $g \in \mathcal{O}_{\mathcal{M}_{0:A}}$ stop and return $v = g$ and otherwise repeat from (23)*.*

The expression for the prior is similar to that given in equation (19) but now over a VSP-hierarchy,

$$\pi_{\mathcal{M}_{0:A}}(v|\rho,\tau,\beta) \propto E_U(\mathbb{I}_{h(\eta(U,\beta))=v}), \quad v \in \mathcal{O}_{\mathcal{M}_{0:A}},$$

now normalised over the space of VSP-hierarchies. We never actually use the rejection procedure in Definition 5 to simulate VSPs. We use the MC_U update applied to each VSP $v^{(a)},\ a \in \mathcal{A}$.

There is also a VSP-clustering variant of the PDP-HPO model in section 5. The posterior is just the same as equation (18) but constrained to partial order hierarchies which are VSP hierarchies (given the partition $\xi = (\xi_1, \ldots, \xi_A)$ the U matrix is constrained so that $h(\eta(U,\beta)) \in \mathcal{O}_{\mathcal{M}^{A+1}}$). This gives the PDP-HVSP model.

7 Application and Model Comparison

This section applies the Hierarchical Partial Order (HPO) model and its VSP approximation (HVSP) to different scenarios. We begin by investigating the accuracy of PDP-HVSP as an approximation to PDP-HPO on a synthetic dataset with 10 items in section 7.1. The synthetic dataset has 15 unlabeled lists from 3 different assessors. The approximation is good (as evidenced by the Receiver Operating Characteristic (ROC) curve). We then apply HVSP to a real 3-D sound spatialisation dataset [28] in section 7.2 (where assessor labels are known, so the HVSP-model of section 6.2 applies) and the sushi preference data of [29] (where labels are unknown, so we cluster rank lists

using the PDP-HVSP model introduced at the end of section 6.2) in section 7.3. Table 2 summarises the structures of all datasets considered in this paper.

Dataset	# Items	# Lists	# Assessors	model fit
Synthetic Data	10	15	3 (true)	PDP-HPO/VSP
3-D Sound Spatialisation Data	12	1380	46	HVSP
Sushi Preference Data	10	1000	unknown	PDP-HVSP

Table 2: Summary for datasets considered in this paper.

7.1 Application: Synthetic Data

In this section, we cluster 15 synthetic rank-order lists with 10 items each. The unknown true number of assessors is 3 - each producing 5 rankings of all 10 items. The true list partition is $\xi^* = (\{1,\ldots,5\},\{6,\ldots,10\},\{11,\ldots,15\})$. The data are simulated from the HPO generative model with $K = 2, \tau = 0.8$ and $\rho = 0.9$, so the true partial order hierarchy, shown in Figure 5, is in $\mathcal{H}^{(2)}_{\mathcal{M}^{A+1}}$. These orders are all VSPs so the PDP-HVSP model should fit well. The synthetic rank-orders are simulated using the queue-jumping error model with $p = 0$, so if h is the true hierarchy then the data are noise free linear extensions chosen uniformly at random and $Y_i \in \mathcal{L}[h^{(a)}],\ (i,a) \in [N] \times \mathcal{A}$.

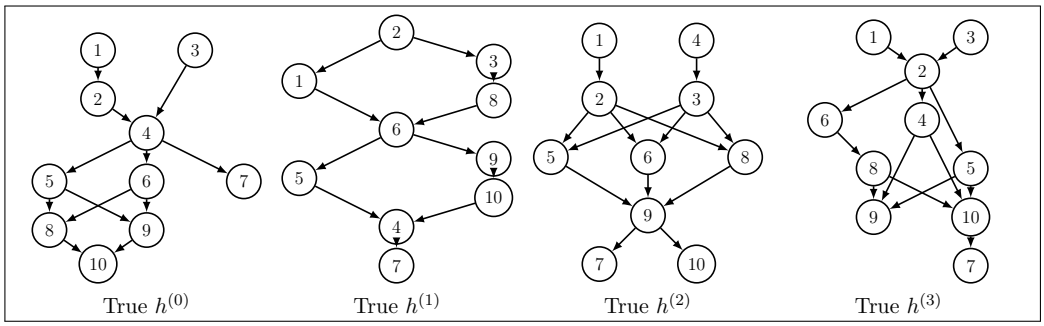


Figure 5: The true "root" partial order $h^{(0)}$ and the true leaf partial orders $h^{(a)}, a \in \{1,2,3\}$ for simulation.

We implement both the PDP-HPO model (with $K = 2$) and PDP-HVSP. We choose the prior distributions $\rho \sim Beta(1, 1/3)$, $\log(\frac{p}{1-p}) \sim \mathcal{N}(0, 1.5)$, $\tau \sim \mathcal{U}[0,1]$ and $\xi \sim PDP(\eta_\theta = 0.5,\ \eta_\alpha = 0.15)$ for a flat prior distribution partial order depth (see Figure 10). All MCMC chains are initialised at random and run for $T = 2e5$ iterations. The trace plots for the key parameters and their effective sample sizes are in appendix F.1. To sum-

marise the sampled partial orders, we define the consensus partial orders $h_{con}(\epsilon)$ that includes order relation/edge $(i \succ_{h_{con}(\epsilon)} j)$ if the relation appears more than ϵT times in the MCMC output. We calculate the proportion of the true-positive (TPR) and false-positive (FPR) relations for $h^{con}(\epsilon), \epsilon \in [0, 1]$ and report ROC curves in Figure 6. These are given for the estimated "root" partial orders $h^{(0)}$ estimated using PDP-HPO and PDP-HVSP. The true positive rate (TPR) and false positive rate (FPR) increase with decreasing ϵ from $(FPR, TRP) = (0, 0)$ at $\epsilon = 1$ (the consensus order is empty) to $(1, 1)$ at $\epsilon = 0$ (complete graph). Both methods shows similar reconstruction accuracy (achieving around 80% TRP at 15% FPR), with the PDP-HVSP ROC-curve lying on the PDP-HPO curve as we would hope given the truth can be represented by the PDP-HVSP model. The ϵ^* that gives (FPR,TPR) closest to $(0, 1)$ is selected (by eye, $\epsilon^* \simeq 0.5$) and used to construct consensus orders below.

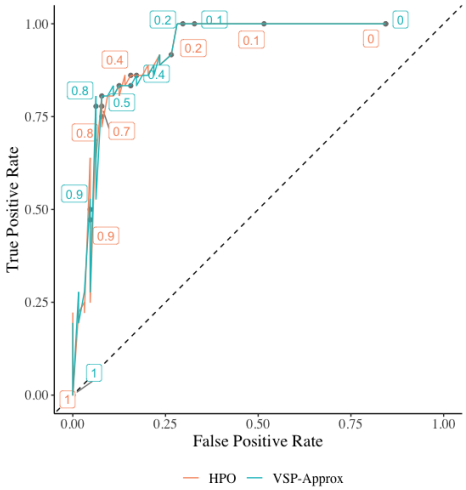


Figure 6: The receiver operating characteristic (ROC) curves for the "root" partial orders in the synthetic data analysed with PDP-HPO (orange) and PDP-HVSP (blue). The true/false positive rates are plotted against the threshold ϵ to construct consensus order $h^{con}(\epsilon), \epsilon \in [0, 1]$.

Figure 7 displays the co-occurrence matrices for cluster allocations from both methods. PDP-HPO and PDP-HVSP perform similarly at clustering. They almost recover the true cluster allocation, except for list 11 and some uncertainty on list 15. This might because that the limited data is not informative enough to recover the true partition. List 11 and list 15 are close to linear extensions of the true $h^{(2)}$ - both are only two pairwise swaps away. Figure 8 gives the consensus "root" and consensus leaf orders from PDP-

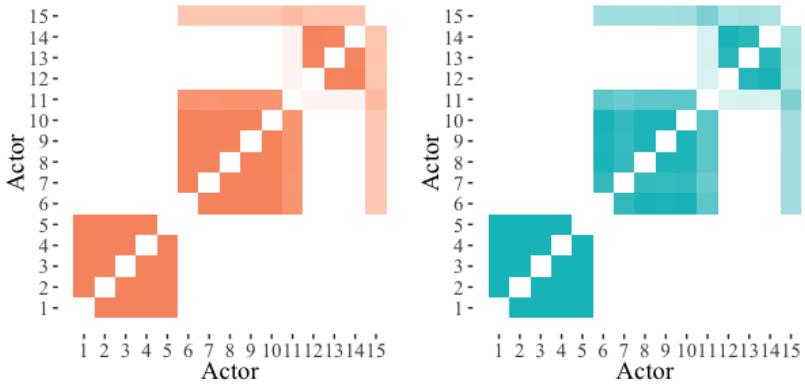


Figure 7: Cluster allocation for PDP-HPO (left) and PDP-HVSP (right). Each cell represents the posterior probability that a pair of items is allocated to the same cluster, taking values from 0 (white) to 1 (orange/blue).

HPO. The ones for PDP-HVSP are shown in Figure 9. The leaf concensus orders are constructed under the most probable list partition in the posterior distribution $\xi = (\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10, 11, 15\}, \{12, 13, 14\})$. Both models give leaf partial orders that are close to the truth, especially for clusters $a = 1$ and $a = 2$.

The prior and posterior distributions for the key parameters (ρ, p, τ and A, the number of clusters, which is an unknown) for both the PDP-HPO and PDP-HVSP models are shown in Figure 10. We also label their true values with the red vertical lines. See their trace plots in Appendix F.1. Both methods find the noise parameter p is close to 0, correctly, as there is no error in the synthetic dataset. The two models perform similar on recovering the truth, while the PDP-HVSP model appears to slightly outperform the PDP-HPO model on recovering the 'root-to-leaf' correlation τ.

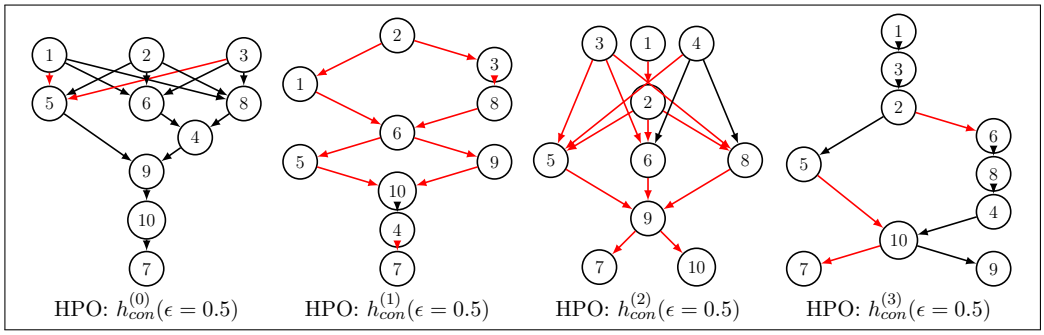


Figure 8: The consensus "root" order $h^{(0)}_{con}(\epsilon = 0.5)$ and the consensus leaf orders $h^{(a)}_{con}(\epsilon = 0.5), a \in \{1, 2, 3\}$ from PDP-HPO. An edge is colored red if its posterior probability is larger than 0.9. The thresholds ϵ's are chosen based on the ROC curves that gives higher TPR and lower FPR. The leaf concensus orders are constructed with regard to the cluster with highest posterior probability $\xi = (\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10, 11, 15\}, \{12, 13, 14\})$.

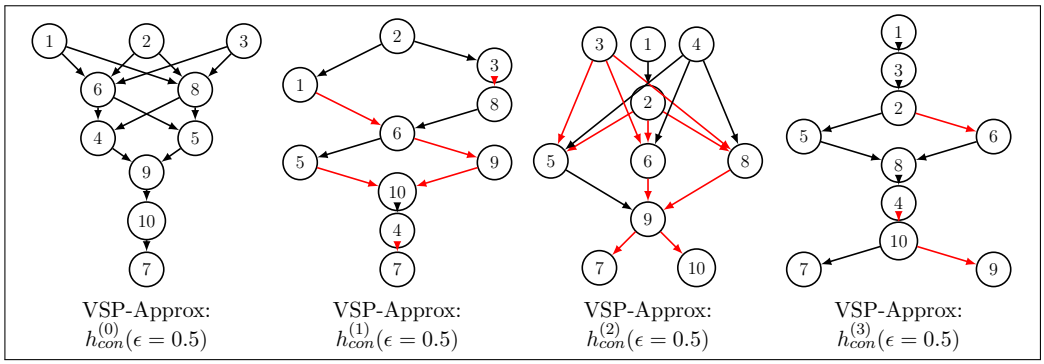


Figure 9: The consensus "root" order $h^{(0)}_{con}(\epsilon = 0.5)$ and the consensus leaf orders $h^{(a)}_{con}(\epsilon = 0.5), a \in \{1, 2, 3\}$ from PDP-HVSP. An edge is colored red if its posterior probability is larger than 0.9. The thresholds ϵ's are chosen based on the ROC curves that gives higher TPR and lower FPR. The leaf concensus orders are constructed with regard to the cluster with highest posterior probability $\xi = (\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10, 11, 15\}, \{12, 13, 14\})$.

The PDP-HVSP approximation performs to a similar standard to the PDP-HPO model. The MCMC for PDP-HPO can be challenging. When we propose a new partition we have to "throw down" an entire new 10×2 dimensional U-matrix for the new cluster. This is not completely hopeless as it is a prior draw conditioned on $U^{(0)}$ so $U'^{(A+1)}$ will resemble the other $U^{(a)}$-matrices in clusters $a = 1, \ldots, A$, at least when the correlation parameter τ is close to one, and this increases the chance it will determine a partial order that makes the data likely. It is nevertheless targeting a relatively high

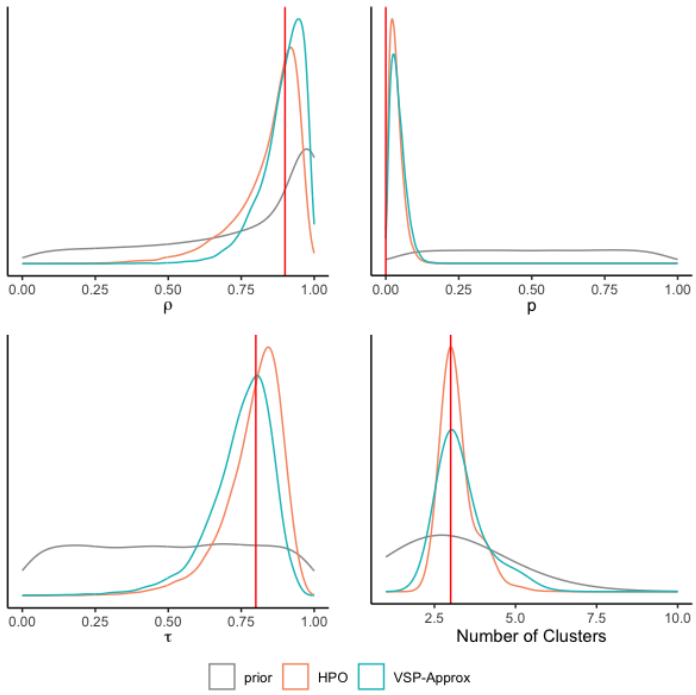


Figure 10: The prior (grey) and posterior distributions for parameters ρ (top left), p (top right), τ (bottom left) and $|\xi|$ (bottom right) from PDP-HPO (orange) and PDP-HVSP (blue). The true values are displayed using the red vertical line.

dimensional conditional distribution. Since PDP-HVSP has to generate a U-matrix for a VSP, which we do via rejection, we might expect it to be even harder. Three further factors work in favor of $PDP-HVSP$: first of all the truth is HVSP so the PDD-HVSP model has a smaller search space which contains the truth; when $K = 2$, the chance we will hit a VSP is in fact quite high so the number of rejections is small (even when τ is small); also, the likelihood evaluation for PDP-HVSP is much faster than that for PDP-HPO (the update is about 3 times faster, even including the rejection step).

7.2 3-D Sound Spatialisation Data

The 3-D sound spatialisation data described and analysed in [28] study the impact of 3-D sound spatialisation on listeners' mental representation of human agency. The study involves $A = 46$ participants (assessors), each presented with $N_a = 30$, $a \in \mathcal{A}$ random and independently chosen pairs of sounds, so $|S_{a,i}| = 2$, $(a, i) \in \mathcal{A} \times [N_a]$. The assessors select the sound,

from each pair, that most evoked a feeling of human causation or human physicality. There are $M = 12$ different test stimuli. Each assessor received each of the M different sounds in at least one pair so $\mathcal{M}_a = \mathcal{M},\ a \in \mathcal{A}$. Stimuli 1 (S1) was sonified so that its mapping ranges were scaled to most clearly enhance features in the data. The remaining 11 stimuli were modified from S1 to either suppressed features of the original movement captured in the data, or reduced the scaling ranges of the sonification or both. Refer to [28] for more details. The 3-D sound spatialisation data therefore contains $N = 1380$ pairwise comparisons over the 12 test stimuli coming from 46 assessors. Given the assessors, we are interested in forming a global order relation among the test sounds sumarising the assessor preferences and to find the test stimuli that simulate the most human agency.

These data were further analysed in [30] using a Mallows model in a method which allows for non-transitivity in the data (as we do also, the queue jumping error model allows this). The analysis clusters assessors. This is different to our PDP-HPO/VSP based clustering, as all the rank-orders associated with an assessor are clustered as a group which cannot be split. In our setup each order in the data enters the clustering separately, so results are not directly comparable.

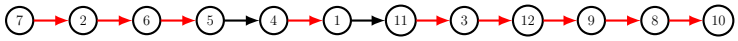


Figure 11: The consensus order $h^{(0)}_{con}(\epsilon = 0.5)$ of the "root" partial order for the 3-D sound spatialisation data is a complete order. An edge is colored red if it appears in more than 90% of posterior samples.

We choose the prior distributions for key parameters as $\rho \sim Beta(1, 1/3)$, $\log(\frac{p}{1-p}) \sim \mathcal{N}(0, 1.5)$ and $\tau \sim \mathcal{U}[0, 1]$ for a relatively flat prior distribution over partial order depth. The MCMC was run for around 1.6×10^5 iterations subsampling at 12 sweeps over all parameters. The results are reported with a burn-in period of 6500×12 iterations. The trace plots and effective sample sizes of the key parameters are in Appendix F.2.1. Convergence seems acceptable but mixing in equilibrium is slow and longer runs may be necessary to get larger ESS values. The consensus order for the "root" partial order is shown in Figure 11. The consensus "root" order is a complete order, where most of its edges are of more than 90% posterior probability. This gives a strong order relation among the 12 test stimuli. The top three test stimuli are stimuli S7, S2 and S6. These three are similar to S1 with only one spatialisation feature removed. In particular, S7, where the pitch variation is removed from S1, significantly outperforms other choices. The bottom three are stimuli S9, S8 and S10. These are where the 3-D spatial

variation is flattened further. For example, S10 flattens the spatial variation in S1 to consist of only three changes in direction between two points in space. It further removes pitch and grain volume variation, resulting in very little variation in grain duration. [30] and [28] analyse the same dataset with a mixture extension of the Mallows model. They perform clustering on the pairwise comparisons, i.e. the assessors with similar perception of the sounds are grouped together. Three clusters are detected, with the top stimuli being S8, S5 and S1 respectively. Their model differs from ours where we assume each assessor to be distinct and produce a concensus order for each assessor in the HVSP model, which are displayed in Appendix F.2.2. More than half of the "leaf" orders are not complete orders, some with more complicated structure. This suggests that different individuals respond differently to a range of 3-D spatialisation features, possibly a result of their different backgrounds, listening abilities and experiences. Our model successfully captures such behaviour and makes both a high-level and individual-level study possible.

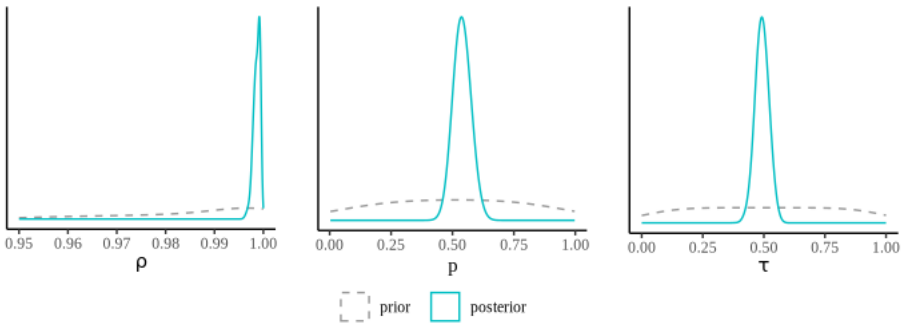


Figure 12: The prior (grey) and posterior distributions (blue) for parameters ρ (left), p (middle) and τ (right) from the HVSP model.

The posterior distribution of key parameters are summarised in Figure 12. The posterior distribution on the depth parameter ρ is concentrated to a value close to 1 giving higher depth, which coincides with the frequent appearance of complete orders in our inference. The posteriors for both the "roof-to-leaf correlation" parameter τ and the queue-jumping error probability p concentrate at around 0.5. This suggests a moderate amount of root-to-leaf shrinkage effect and a relatively high level of error in the participants' responses.

Although the individual reflection on 3-D sound spatialisation features and human agency varies, our study shows a clear order preference among the 12 test stimuli. It seems that it is not necessarily the case that more spatial projection is better. In particular, the sound sonified to optimise features of

the spatial projection (S1) is ranked in the middle. The stimuli with small variation, e.g. with pitch variation removed or where spatial motion is placed in front of the listening location tend to have better performance in general.

7.3 Application: Sushi Preference Data

The sushi preference data [29] is collected by Toshihiro Kamishima and his colleagues on people's preference over selected sushi. A total of 5,000 participants coming from 47 prefectures/11 regions[1] (or overseas) are invited to rank the sushi - producing a set of complete rankings over 10 popular sushi. There are many factors that may affect people's tastebud. For example, [29] collected the participants gender, age, the prefecture/region that they currently live in and the prefecture/region they have lived in the most before 15 years old. We cluster preference orders using the PDP-HVSP model to identify groups with similar sushi preferences. In this example analysis, we do not use the covariates to inform the clustering. We look at the clusters we identify a posteriori and see if any are enriched with particular covariate values. A similar analysis was carried out in [31], [32] and several other papers as this is a classical dataset.

In our study, we apply PDP-HVSP on the sushi rankings. Clustering over 5,000 orders can be relatively costly, we take a random sample of 1,000 rank-orders for demonstration.

We take prior distributions $\rho \sim Beta(1, 1/3), \log(\frac{p}{1-p}) \sim \mathcal{N}(0, 1.5)$, $\tau \sim \mathcal{U}[0, 1]$ and $\xi \sim PDP(\eta_\theta = 0.05, \eta_\alpha = 0.32)$ for a reasonably flat prior distribution over partial order depth and a broadly plausible clustering in prior simulation. The MCMC is run for 1.2×10^5 iterations. We subsample every 10 steps and take a burn-in period of 8000×10 iterations. The trace plots and effective sample sizes of the key parameters are shown in Figure 19 and Table 7 in Appendix F.3. Again, convergence to equilibrium is reasonable but mixing in equilibrium is slow and longer runs may be needed for accurate quantification of uncertainty. The consensus order of the "root" partial order is summarised in Figure 13.

The posterior distributions of the key parameters and the number of clusters is shown in Figure 14. The posterior distribution on the number of clusters concentrated on a far higher value than our prior estimate, with a posterior mean of ~ 77. The posterior cluster allocation is shown in Appendix F.3.3. The heatmap displays the pairwise posterior probability for

[1]The prefectures in Japan rank immediately below the national government and form the country's first level of jurisdiction and administrative division. The prefectures can be categorised into 11 regions: Hokkaido, Tohoku, Hokuriku, Kanto Shizuoka, Nagano Yamanashi, Chukyo, Kinki, Chugoku, Shikoku, Kyushu and Okinawa.

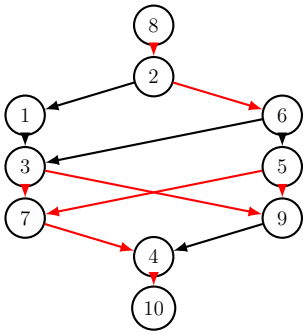


Figure 13: The consensus order $h^{(0)}_{con}(\epsilon = 0.5)$ of the "root" partial order for the sushi preference data. The edge is colored red if it appears more than 90% in the posterior distribution.

any two lists to be in the same cluster. For example, lists 76, 563, 227, 122, 613 and 707 have higher probability to be in the same cluster. We observe a lower value on the depth parameter ρ indicating lower depth in the partial orders in general. Most participants find a few choices of sushi incomparible. Higher posterior values on the "root-to-leaf" correlation parameter τ is observed with a posterior mean of $\bar{\tau} \approx 0.74$. This suggests the leaf partial orders tends to shrink to the root partial order. This indicates a general trend in people's sushi taste, with relatively less local/personal specific influence.

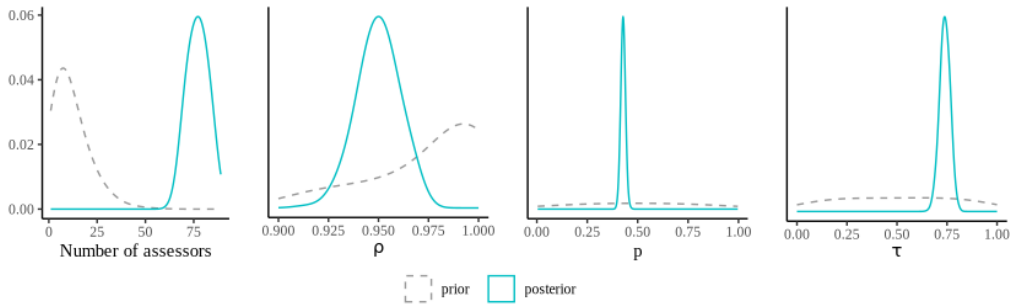


Figure 14: The prior (grey) and posterior distributions (blue) for the number of assessors, the parameters ρ, p and τ from the PDP-HVSP model.

Referring to Figure 13, we see the root order is not a complete order. However, preferences can be discerned. The relatively high shrinkage effect between the leaf and root partial orders suggests the root order is representative of the population as a whole, with a consensus favoring sushi toro (fatty tuna, 8). The anago (sea eel, 2) sushi is in the second place. It appears that the least favourite sushi are kappa maki (cucumber roll, 10) and ika (squid, 4). This is consistent with results in [32] and [31].

We investigate the similarity among assessors that are grouped in the same cluster. We group the rank-order lists based on their clustering posterior distribution into partitions of the same lengths of the number of prefectures (47) or regions (12). However, no geographical separation was found in the above analysis. Geographical separation doesn't seem to have strong influence on people's sushi preference, at least among the set of 10 popular sushi in Japan. Similar analysis is done for other variables. We further performed generalised linear regression on the cluster allocation with different variables. However, no significant relation was found. In future work we would like to introduce covariates in the model and test for an effect directly.

7.4 Model Comparison

In this section we compare our clustering hierarchical VSP model with Mallows and Plackett-Luce mixture models on the synthetic dataset (as a check on our model selection methods) and the sushi preference data. We used the packages `PLMix` [33] for the Plackett-Luce Mixture and `BayesMallows` [34] for the Mallows Mixture model. We didn't make a comparison on the 3-D sound spatialisation data as we could not find code for hierarchical models for Mallows and Plackett-Luce in the supervised setting. We use the expected log pointwise predictive density (*elpd*, [35]) as our model comparison criterion. This is a predictive loss which can be estimated using the WAIC [36]. We give estimated $elpd_{waic}$-values in Table 3.

	$\mathbf{elpd_{waic}}$	
Model	**Synthetic Data**	**Sushi Data**
PDP-HVSP	-187.3 (10.3)	-13218.2 (78.2)
Plackett-Luce Mixture	-362.9 (5.8) ($G = 3$)	-13532.6 (54.1) ($G = 54$)
Mallows Mixture	-801.5 (11.8) ($G = 1$)	-15660.6 (43.6) ($G = 1$)

Table 3: Model comparison between PDP-HVSP, Mallows Mixture (Kendall's tau distance) model (R package `BayesMallows`, [34]) and Plackett-Luce Mixture models (R package `PLMix`, [33]). The model comparison criteria $elpd_{waic}$ is given in the format of *estimate (standard error)*.

The number of mixture components for the Plackett-Luce and Mallows mixture are chosen to give the highest $elpd_{waic}$. The PDP-HVSP outperforms both Plackett-Luce and Mallows mixture models significantly. Better performance on the synthetic dataset is expected because the data is simulated from the generative model for partial order hierarchy. The Plackett-

Luce mixture model outperforms the Mallows Mixture with Kendall's tau distance on the two datasets. It is possible that a Mallows' distance better adapted to the data, such as the footrule distance, would give better results. Overall, we conclude that the PDP-HVSP model is a better fit to the data in consideration.

8 Conclusion

This paper introduces a novel hierarchical partial order model (HPO) designed for "grouped" data, a common scenario in ranking problems. This is the first hierarchical model for partial orders. The model maintains marginal consistency and provides precise depth control over both root and leaf partial orders. The Gumbel-(K, n, ρ) setup enables the partial order hierarchy to incorporate the Plackett-Luce mixture model as a special case. Additionally, we extend the model to handle rank-order clustering through the Poisson-Dirichlet process, accommodating situations where rank-orders are not labeled with their assessors.

To address model scalability, we introduced an approximation approach using VSPs and associated Bayes/MCMC methods. We compared the performance of the clustering PDP-HPO model with its VSP approximation (PDP-HVSP) on a synthetic dataset. The VSP approximation performs no worse than the PDP-HPO model. By approximating leaf partial orders with VSPs, our approach balances computational efficiency and modeling flexibility, allowing for inference on larger and more complex datasets while preserving essential global preference structures through the root partial order. This makes hierarchical ranking models applicable to practical scenarios where computational resources are constrained but preference structures are intricate. We further applied the HVSP and PDP-HVSP approximations to 3-D sound spatialisation data and sushi preference data. Comparison with the Plackett-Luce and Mallows mixture models on the sushi data showed that our PDP-VSP approach gives a better fit to the data there.

The proposed hierarchical partial order model provides a more flexible framework for describing rank preferences, encompassing complete order models as a special case. It serves as a versatile hierarchical model well-suited for various grouped rank-order data structures. Future research could explore incorporating covariates experiments with the models and experiments with HVSP and PDP-HVSP. Additionally, enhancing the efficiency of the MCMC algorithms (perhaps via code optimisation) for both the partial order hierarchy and its VSP approximations could significantly improve performance, particularly for large-scale datasets. Finally, there exist data

sets in which the observations are partial orders [37]. The HVSP model may be used to find a partial order which is central to the partial orders in the data, in the same way that the Mallows model finds a complete order which is central to the complete orders in standard rank-order data.

References

[1] 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.

[2] Harald Steck. Evaluation of recommendations: rating-prediction and ranking. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 213–220, 2013.

[3] Steffen Foerster, Carson M. Franz, Mathias Murray, Ian C. Gilby, Joseph T. Feldblum, Kara K. Walker, and Anne E. Pusey. Chimpanzee females queue but males compete for social status. *Scientific Reports*, page 35404, 2016.

[4] AA Foroughi and M Tamiz. An effective total ranking model for a ranked voting system. *Omega*, 33(6):491–496, 2005.

[5] A Anderson. Maximum likelihood ranking in racing sports. *Applied Economics*, 46(15):1778–1787, 2014.

[6] François Caron, Yee Whye Teh, and Thomas Brendan Murphy. Bayesian nonparametric plackett–luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, pages 1145–1181, 2014.

[7] Ao Liu, Zhibing Zhao, Chao Liao, Pinyan Lu, and Lirong Xia. Learning plackett-luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4328–4335, 2019.

[8] Cristina Mollica and Luca Tardella. Bayesian plackett–luce mixture models for partially ranked data. *Psychometrika*, 82(2):442–458, 2017.

[9] Thomas Brendan Murphy and Donal Martin. Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, 41(3-4):645–655, 2003.

[10] Marina Meila and Harr Chen. Dirichlet process mixtures of generalized mallows models. *arXiv preprint arXiv:1203.3496*, 2012.

[11] Tyler Lu and Craig Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *J. Mach. Learn. Res.*, 15(1):3783–3829, 2014.

[12] Geoff K. Nicholls, Jeong Eun Lee, Nicholas Karn, David Johnson, Rukuang Huang, and Alexis Muir-Watt. Bayesian inference for partial orders from random linear extensions: power relations from 12th century royal acta, 2023.

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

[14] Chuxuan Jiang, Geoff K. Nicholls, and Jeong Eun Lee. Bayesian inference for vertex-series-parallel partial orders. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

[15] G. Brightwell. *Surveys in Combinatorics*, volume 187 of *London Mathematical Society Lecture Note Series*, chapter Models of random partial orders, pages 53–83. Cambridge Univeristy Press, 1993.

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

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

[18] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[19] R Duncan Luce. On the possible psychophysical laws. *Psychological review*, 66(2):81, 1959.

[20] John I. Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.

[21] David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384 – 406, 2004.

[22] Thomas Sakoparnig and Niko Beerenwinkel. Efficient sampling for Bayesian inference of conjunctive Bayesian networks. *Bioinformatics*, 28(18):2318–2324, 07 2012.

[23] Peter Winkler. Random orders. *Order*, 1(4):317–331, 1985.

[24] Chuxuan Jiang and Geoff K. Nicholls. Bayesian inference for partial orders with ties from ranking data with a mallows distribution centred on random linear extensions. 2023.

[25] Kenji Handa. The two-parameter poisson–dirichlet point process. 2009.

[26] 001035: Number of partially ordered sets ("posets") with n labeled elements (or labeled acyclic transitive digraphs). `https://oeis.org/A001035`. Accessed: 2024-05-13.

[27] Steven Finch. Series-parallel networks. `https://oeis.org/A000084/a000084_2.pdf`, July 2003. Accessed: 2024-05-13.

[28] Natasha Barrett and Marta Crispino. The impact of 3-d sound spatialisation on listeners' understanding of human agency in acousmatic music. *Journal of New Music Research*, 47(5):399–415, 2018.

[29] Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 583–588, 2003.

[30] Marta Crispino, Elja Arjas, Valeria Vitelli, Natasha Barrett, and Arnoldo Frigessi. A Bayesian Mallows approach to nontransitive pair comparison data: How human are sounds? *The Annals of Applied Statistics*, 13(1):492 – 519, 2019.

[31] Tyler Lu and Craig Boutilier. Learning mallows models with pairwise preferences. In *Proceedings of the 28th international conference on machine learning (icml-11)*, pages 145–152, 2011.

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

[33] Cristina Mollica and Luca Tardella. Plmix: An r package for modeling and clustering partially ranked data. *arXiv preprint arXiv:1612.08141*, 2016.

[34] Øystein Sørensen, Marta Crispino, Qinghua Liu, and Valeria Vitelli. Bayesmallows: an r package for the bayesian mallows model. *arXiv preprint arXiv:1902.08432*, 2019.

[35] 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.

[36] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 8 2013.

[37] Alberto Arcagni, Alessandro Avellone, and Marco Fattore. Complexity reduction and approximation of multidomain systems of partially ordered data. *Computational Statistics & Data Analysis*, 173:107520, 2022.

[38] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

A Proof of Theorem 1

Theorem 1. *Let $G^{-1}(g) = -\log(-\log(g))$ be the inverse CDF of a standard Gumbel random variable and let Φ be the CDF of a standard normal. Let X be an $M \times p$ design matrix of p covariates with a row x_j for each $j \in \mathcal{M}$, and let $\beta \in \mathbb{R}^p$ be the vector of effect values so that $\alpha = X\beta \in \mathbb{R}^{M\times 1}$ and $\alpha_j = x_j^T\beta$ is the linear predictor for $j \in \mathcal{M}$. Let Σ_ρ be a $K \times K$ covariance matrix with unit diagonal $(\Sigma_\rho)_{k,k} = 1$ and constant off diagonal $(\Sigma_\rho)_{k,k'} = \rho$ for $\rho \in [0, 1)$, $k, k' \in \{1, \ldots, K\}$ and $k \neq k'$. We take*

$$\begin{aligned} U_{j,\cdot} &\sim N(0_K, \Sigma_\rho), \quad \textit{independent for each } j \in \mathcal{M}, \\ \eta_{j,\cdot} &= G^{-1}(\Phi(U_{j,\cdot})) + \alpha_j 1_K^T, \quad \textit{and} \\ h &= h(\eta(U, \beta)), \end{aligned}$$

where the functions G^{-1} and Φ are applied to their arguments element by element. Two properties hold for the Gumbel-(K, n, ρ) partial order model.

1. *It holds that $h(\eta_{\cdot,k}) \sim PL(\alpha, \mathcal{M})$ in (1) for each $k = 1, \ldots, K$.*
2. *If data $y \sim \mathcal{U}(\mathcal{L}[h(\eta)])$ are drawn uniformly at random from the linear extensions of $h(\eta)$ then $y \sim PL(\alpha, \mathcal{M})$ when $K = 1$.*

Proof. Theorem 1 follows from lemma 1. Each of the K components of the correlated normal random variables $U_{j,\cdot} \in \mathbb{R}^K$ is marginally standard normal so $\Phi(U_{j,k}) \sim \mathcal{U}[0, 1], k \in [K], j \in \mathcal{M}$ is a vector of K correlated uniform random variables. They are each mapped to standard Gumbel variables G_j via G^{-1}. Adding the K-component constant vector $\alpha_j 1_K^T$ shifts them so each has mean α_j and so the random variables

$$\eta_{j,k} \sim Gumbel(\alpha_j)$$

have the same distribution for each $(j, k) \in \mathcal{M} \times [K]$. They are independent $\eta_{j_1,k} \perp \eta_{j_2,k}$ for $j_1 \neq j_2$ at any fixed k as they are respectively functions of $U_{j_1,k}$ and $U_{j_2,k}$ alone, where $U_{j_1,k}$ and $U_{j_2,k}$ are independent.

Since it is established that the k'th column $\eta_{\cdot,k} = (\eta_{j,k})_{j\in\mathcal{M}}$ of η is a vector of M independent Gumbel variables, it follows by lemma 1, that the complete order $h(\eta_{\cdot,k})$ is a draw $h(\eta_{\cdot,k}) \sim PL(\alpha, \mathcal{M})$ from the Plackett-Luce distribution in equation (1) so the first claim holds. If $K = 1$ then η only has one column, so $h(\eta)$ itself is a complete order distributed according to $p_{\mathcal{M}}(\cdot|\alpha)$. A complete order only has one linear extension, namely $y = h(\eta)$, and so y is also distributed according to $PL(\alpha, \mathcal{M})$, the second claim. □

B Proof of Corollary 1

Corollary 1. *The family of prior distributions $\pi_S(\cdot|\rho,\beta),\ S \in \mathcal{B}_{\mathcal{M}}$ (7) is marginally consistent, that is if $h \sim \pi_{\mathcal{M}}(\cdot|\rho,\beta)$, then $h[S] \sim \pi_S(\cdot|\rho,\beta)$.*

Proof. Let $S \in \mathcal{B}_{\mathcal{M}}, S \subset \mathcal{M}$ with $|S| = m$. Let U be a random $M \times K$ matrix with independent rows $U_{j,\cdot} \sim N(0_K, \Sigma_\rho),\ j \in \mathcal{M}$ and let $h = h(\eta(U,\beta))$ so $h \sim p_{\mathcal{M}}(\cdot|\rho,\beta)$. Let $\tilde{U}$ be a random $m \times K$ matrix with independent rows $\tilde{U}_{j,\cdot} \sim N(0_K, \Sigma_\rho),\ j \in S$ and let $\tilde{h} = h(\eta(\tilde{U},\beta))$ so $\tilde{h} \sim p_S(\cdot|\rho,\beta)$. Now $h[S] = h(\eta(U,\beta))[S]$ and

$$h(\eta(U,\beta))[S] = h(\eta(U_S,\beta))$$

where $U_S = (U_{j,\cdot})_{j\in S}$. This follows as the relations in $h[S]$ are determined by applying the map $h(\cdot)$ to the rows of η_S and these are the rows of $\eta(U_S,\beta)$. As U_S and $\tilde{U}$ have the same distribution (in both cases the rows are m independent draws from $N(0_K,\Sigma_\rho)$) we have $h(\eta(U_S,\beta)) \sim h(\eta(\tilde{U},\beta))$ and so $h[S] \sim \pi_S(\cdot|\rho,\beta)$. □

C Proof of Theorem 2

Theorem 2. *The generative model in definition 3 defines a prior distribution over hierarchical partial orders as*

$$\pi_{\mathcal{M}_{0:A}}(h|\rho,\beta,\tau) = E_U(\mathbb{I}_{h(\eta(U,\beta))=h}), \quad h \in \mathcal{H}^{(K)}_{\mathcal{M}_{0:A}},$$

where $\pi_{\mathcal{M}_{0:A}}(h|\rho,\beta,\tau)$ is normalised over $h \in \mathcal{H}^{(K)}_{\mathcal{M}_{0:A}}$. Let

$$\pi_{\mathcal{M}_{0:A}}(h^{(a)}|\rho,\beta,\tau) = E_U(\mathbb{1}_{h(\eta^{(a)}(U^{(a)},\beta))=h^{(a)}})$$

be the marginal for $h^{(a)}$. The prior $\pi_{\mathcal{M}_{0:A}}$ has the following properties:

1. *(single PO marginal) For $a = 0,\ldots,A$, $\pi_{\mathcal{M}_{0:A}}(h^{(a)}|\rho,\beta,\tau) = \pi_{\mathcal{M}_a}(h^{(a)}|\rho,\beta)$ where $\pi_{\mathcal{M}_a}$ is the single partial-order prior given in (7);*
2. *(PL hierarchy at $K=1$) when $K=1$, $h^{(a)} \sim PL(\alpha_{\mathcal{M}_a}, \mathcal{M}_a)$ for $a \in \mathcal{A}$;*
3. *($h^{(a)}$ independent of $h^{(0)}$ at $\tau = 0$) when $\tau = 0$ the prior factorises,*

$$\pi_{\mathcal{M}_{0:A}}(h|\rho,\beta,\tau=0) = \pi_{\mathcal{M}_0}(h^{(0)}|\rho,\beta,\tau)\prod_{a=1}^{A}\pi_{\mathcal{M}_a}(h^{(a)}|\rho,\beta,\tau);$$

4. *($h^{(a)} \to h^{(0)}$ as $\tau \to 1$) If $\mathcal{M}_a = \mathcal{M}$ for each $a = 0, 1, \ldots, A$ then, for each $a \in \mathcal{A}$, $\pi_{\mathcal{M}_{0:A}}(h^{(a)}|\rho, \beta, \tau, U^{(0)}) \to \mathbb{1}_{h^{(a)}=h^{(0)}}$ as $\tau \to 1$;*

5. *(marginal consistency) If $\mathcal{M}_a = \mathcal{M}$ for each $a = 0, 1, \ldots, A$ and $h \sim \pi_{\mathcal{M}^{A+1}}(\cdot|\rho, \beta, \tau)$ then $h[s_{0:A}] \sim \pi_{s_{0:A}}(\cdot|\rho, \beta, \tau)$ for every $s_{0:A} \in \mathcal{B}_{\mathcal{M}}^{A+1}$ with $s_0 = \cup_{a=1}^{A} s_a$.*

Proof. We first show that property 1 holds. When $a = 0$ property 1 is obvious by definition. When $a = 1, \ldots, A$, from equation (9), we can write

$$U_{j,\cdot}^{(a)} = \tau U_{j,\cdot}^{(0)} + \epsilon_{j,\cdot}^{(a)}, \text{ where } \epsilon \sim N(0_K, (1-\tau^2)\Sigma_\rho). \tag{26}$$

This gives the marginal covariance of $U_{j,\cdot}^{(a)}$ as

$$\text{cov}(U_{j,\cdot}^{(a)}) = \text{cov}(\tau U_{j,\cdot}^{(0)}) + (1-\tau^2)\Sigma_\rho = \Sigma_\rho,$$

and expectation $E(U_{j,\cdot}^{(a)}) = 0_K$. That is $U_{j,\cdot}^{(a)} \sim N(0_K, \Sigma_\rho)\mathbb{1}_{\{h(U^{(a)}) \in \mathcal{H}_{\mathcal{M}_a}\}}$, the same as the distribution of $U_{j,\cdot}$ for a single Gumbel-(K, n, ρ) PO model in theorem 1. It follows that $\pi_{\mathcal{M}_{0:A}}(v^{(a)}|\rho, \beta, \tau) = \pi_{\mathcal{M}_a}(v^{(a)}|\rho, \beta)$ where $\pi_{\mathcal{M}_a}$ is the single HPO prior given in (7) for $a = 1, \ldots, A$.

Property 2 holds by property 1 and applying theorem 1 to the marginal for a single assessor partial order.

Property 3 follows from equation (26). We have $\text{cov}(U_{j,\cdot}^{(0)}, U_{j,\cdot}^{(a)}) = \tau\Sigma_\rho$ and $\text{cov}(U_{j,\cdot}^{(a)}, U_{j,\cdot}^{(a')}) = \tau^2\Sigma_\rho$ for $a \neq a' \in \mathcal{A}$. When $\tau = 0$, $h^{(a)} = h(\eta(U^{(a)}, \beta))$ and $h^{(a')} = h(\eta(U^{(a')}, \beta))$ are functions of jointly independent normal random variables, giving property 3.

For property 4, notice that

$$\eta_{j,k}^{(a)} - \eta_{j,k}^{(0)} = G^{-1}(\Phi(U_{j,k}^{(a)})) - G^{-1}(\Phi(U_{j,k}^{(0)})),$$

and by equation (9) $U_{j,k}^{(a)} \to U_{j,k}^{(0)}$ in probability as $\tau \to 1$. By the continuous mapping theorem $\eta_{j,k}^{(a)} \to \eta_{j,k}^{(0)}$ also. Since there exists an open ball $B_\epsilon(\eta_{j,k}^{(0)})$, $\epsilon > 0$ in $\mathbb{R}^{M \times K}$ centered on $\eta_{j,k}^{(0)}$ such that $h(x) = h(\eta_{j,k}^{(0)})$ for all $x \in B_\epsilon(\eta_{j,k}^{(0)})$, the distribution of $h^{(a)}$ concentrates on $h^{(0)}$ giving property 4. If we drop the condition that $\mathcal{M}_a = \mathcal{M}$ for each $a \in \mathcal{A}$ then $h^{(a)}$ concentrates on the suborder $h^{(0)}[\mathcal{M}_a]$.

Property 5 follows by similar reasoning to corollary 1. Essentially $h[s_{0:A}] = h(\eta(U_{s_{0:A}}))$ where $h[s_{0:A}] = (h^{(0)}[s_0], \ldots, h^{(A)}[s_A])$ and $U_{s_{0:A}} = (U_{s_0}^{(0)}, \ldots, U_{s_A})$. Now $U_{s_a}^{(a)} \sim \tilde{U}^{(a)}$ where $\tilde{U}^{(a)}$ is a $|s_a| \times K$ matrix with independent rows distributed like $N(0_K, \Sigma_\rho)$. Since $h(\eta(\tilde{U}, \beta)) \sim \pi_{s_{0:A}}(\cdot|\rho, \beta, \tau)$ and $h(\eta(U_{s_{0:A}})) \sim h(\eta(\tilde{U}, \beta))$ we have $h[s_{0:A}] \sim \pi_{s_{0:A}}(\cdot|\rho, \beta, \tau)$.

□

D Proof of Lemma 2

Lemma 2. *The Markov chain MC_U over U-matrices constrained to represent VSPs is irreducible.*

Proof. It may be helpful to refer to Figures 2 and 16. The sequence of U-matrices $U^0, \ldots, U^{2M}$ must realise VSPs at each step so we need $h(\eta(U^t, \beta)) \in \mathcal{V}_\mathcal{M}$ for each $t = 0, \ldots, 2M$. We take a path via some U-matrix U_M of an arbitrary complete order v_M (which is always a VSP) so in the space of VSPs the path is $v_0 \to v_M \to v_{2M}$. We will take v_M to be an arbitrary fixed complete order and show that there is a path with positive probability in $\mathcal{V}_\mathcal{M}$ from v_0 to v_M for any $v_0 \in \mathcal{V}_\mathcal{M}$. Since the chain is reversible there is a path from v_M to v_{2M} and we are done. The setup is illustrated in Figure 15.

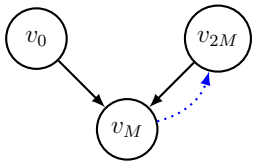


Figure 15: The process communicating from v_0 to v_{2M} through v_M.

Let $(i_1, \ldots, i_M) \in \mathcal{P}_\mathcal{M}$ be an arbitrary permutation of $\mathcal{M}$ and let $v_M = (i_{1:M}, \succ_M)$ where $\succ_M = \{i_a \succ i_b : 1 \leq a < b \leq M\}$ is the complete order ranking elements of $\mathcal{M}$ according to their position in the permutation. Let $u_0 \in \mathbb{R}$ be the least entry in U_0 and U_{2M} so $U_{0,j,k} \geq u_0$ and $U_{2M,j,k} \geq u_0$ for all $(j, k) \in \mathcal{M} \times [K]$ (write $u_0 = \min(U_0, U_{2M})$). Let U_M be any U-matrix satisfying $v_M = h(\eta(U_M, \beta))$ and $\max(U_M) < u_0$ (essentially a ladder of line segments, one for each row of U_M, and segments decreasing from $i_1, \ldots, i_M$ with top rung $U_{M,i_1,k} < u_0, k \in [K]$).

The path is straightforward. In the step from $U^{t-1} = U_{t-1}$ to $U^t = U_t$ select row i_t of U_{t-1} with probability $1/M$ and place it anywhere below u_{t-1} (by setting $U_{t,j,\cdot} = U_{t-1,j,\cdot}$ for $j \neq i_t$ and sampling $U_{t,i_t,\cdot} \sim N(0_K, \Sigma_\rho)$). Set $U^t = U_t$. The probability that $\max(U^t_{i_t,\cdot}) < u_{t-1}$ is always strictly greater than zero. Now set $u_t = \min(U_t, U_{2M})$ and iterate for $t = 1, \ldots, M$. At each step the target for U^t is a set of states which the chain hits with probability greater than zero.

We need to show that $v_t = h(\eta(U_t, \beta))$ is a VSP (so $v_t \in \mathcal{V}_\mathcal{M}$) for each $t = 0, 1, \ldots, M$. We use induction. We have it by definition for $t = 0$ so assume it holds for U_{t-1}. When we update $U_{t-1,i_t,\cdot} \to U_{t,i_t,\cdot}$ we effectively remove i_t from v_{t-1} and place it in a new position below all the elements in $h_{t-1} = v_{t-1}[-i_t]$ (the suborder with i_t removed, and it goes *below* because $G^{-1} \circ \Phi$ is monotone increasing). Since $U_{t,j,\cdot}$, $j \neq i_t$ are unchanged, if $h_t =$

$v_t[-i_t]$ then $h_{t-1} = h_t$. Now define a second partial order $h'_t = (\{i_t\}, \emptyset)$ which is a trivial partial order on the single element i_t. The partial orders h_t and h'_t are both VSPs, h'_t trivially. To show that h_t is a VSP we observe that we have simply removed a leaf from the BDT representing v_{t-1} (refer Figure 16) and this leaves a valid BDT which must define a VSP using the map from BDTs to VSPs (this is the key observation, as it ensures we remain within $\mathcal{V}_\mathcal{M}$). Figure 16 gives an example of removing element 3 from v_0 (it will be placed below the line segment for element 5). The VSP represented by U_t is therefore $v_t = h_t \otimes h'_t$ (the serial operation which stacks h_t above h'_t). This is a VSP by definition.

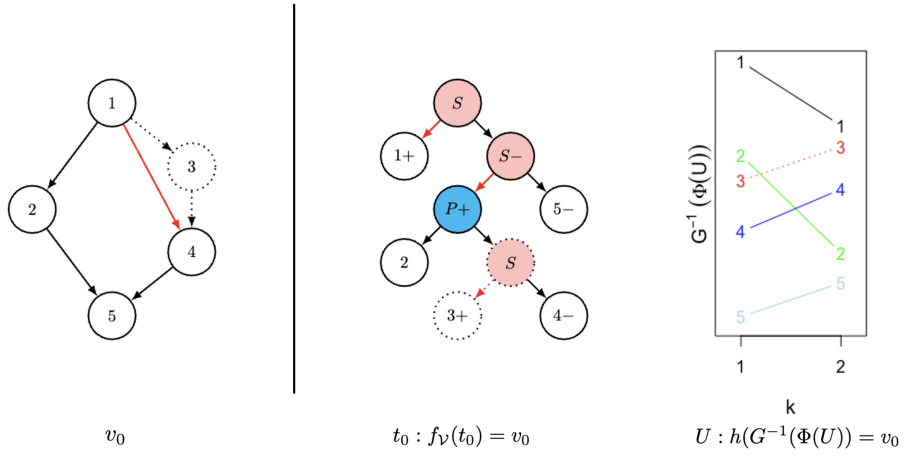


Figure 16: An example item removal operation on v_0. Once we remove item 3 from $\mathcal{M}$, the corresponding leaf node with its connected internal node will be removed in its BDT t_0. In the U-matrix representation, we simply remove the corresponding row. None of these operations affect the order relations among other nodes.

In this way we construct a path of positive probability from any U-matrix U_0 representing the VSP v_0 into the neighborhood δU_M of a U-matrix U_M representing the VSP v_M (which is a complete order). The reverse of this path has positive probability (the measure of the set we have to hit at each step is strictly positive as the elements of U_{2M} are almost surely unequal) so the path $U_0 \to \delta U_M \to \delta U_{2M}$ has $2M$ steps each of which has strictly positive probability of occurring. The point of maintaining a bound $u_t = \min(U_t, U_{2M})$ which respects the ultimate target U_{2M} at each step is that when we come to reverse the chain from U_{2M} into δU_M and run from U_M into δU_{2M} we can be sure that there are no rows of U_{2M} intersecting rows of U_M: that ensures that when we move line segment $U_{t,i_t,\cdot}$ back into place (so

$U^t_{i_t,\cdot} \in \delta U_{2M,i_t,\cdot}$ it cannot intersect a line segment of $U_{t,i_{t'},\cdot}, t' > t$ which has yet to be moved. It follows that $\Pr(U^{2M} \in \delta U_{2M} | U^0 = U_0) > 0$.

The Markov chain over U-matrices constrained to the space of VSPs is therefore irreducible, supporting the MCMC algorithm we use in section 7 to sample the posterior. □

E Sampling Partitions in a PDP-PO model

Denote the number of lists as N. The partition of the lists $\xi = (\xi_1, \ldots, \xi_A) \in \Xi_{[M]}$ follows the Poisson-Dirichlet process $PDP(\eta_\theta, \eta_\alpha)$ with the strength parameter η_θ and discount parameter η_α, such that

$$\pi_{\Xi_{[N]}}(\xi) = \frac{\Gamma(\eta_\theta)}{\Gamma(\eta_\theta + N)} \frac{\eta_\alpha^{|\xi|} \Gamma(\eta_\theta/\eta_\alpha + |\xi|)}{\Gamma(\eta_\theta/\eta_\alpha)} \prod_{a=1}^{A} \frac{\Gamma(|\xi_a| - \eta_\alpha)}{\Gamma(1 - \eta_\alpha)}.$$

We define a Gibbs sampler for the partition ξ and leaf matrix

$$U = (U^{(1)}, \ldots, U^{(A)})$$

below, following [38]. The sampler for PDP-HVSP is the same but in step 2a we must set $K = 2$ and sample $U'^{(A+1)} \sim \pi(\cdot | U_0, \tau, \rho) \mathbb{1}_{h(\eta(U'^{(A+1)}, \beta)) \in \mathcal{V}_\mathcal{M}}$ so the proposed U-matrix must give a VSP (this is done using rejection, which works fairly well for reasons we explain at the end of section 7.1). We use the convention that $a_i = \{a \in \mathcal{A} : i \in \xi_a\}$ is the label of the partition containing list i. For simplicity we suppose $\mathcal{M}_a = \mathcal{M}$ for each $a = 1, \ldots, A$.

Let $X_t = (U, \xi)$.

1. Choose $i \sim \mathcal{U}\{1, \ldots, N\}$ and suppose $i \in \xi_{a_i}$.
2. If $|\xi_{a_i}| > 1$:
   (a) Sample $M \times K$ matrix $U'^{(A+1)} \sim \pi(\cdot|U_0, \tau, \rho)$.
   (b) Let

$$q_a(\xi, a_i) = \begin{cases} (|\xi_{a_i}| - 1 - \eta_\alpha) p(Y_i|U^{(a_i)}, p) & \text{if } a = a_i, \\ (\eta_\theta + A\eta_\alpha) p(Y_i|U'^{(A+1)}, p) & \text{if } a = A+1, \\ (|\xi_a| - \eta_\alpha) p(Y_i|U^{(a)}, p) & \text{if } a \in [A] \backslash \{a_i\}. \end{cases} \tag{27}$$

   (c) Choose new cluster $a' \sim (p_a)_{a=1,\ldots,A+1}$ where $p_a = q_a / \sum_{j=1}^{A+1} q_j$.
      i. If $a' = a_i$, then $(U', \xi') = (U, \xi)$.
      ii. If $a' = A+1$, then $U' = (U, U'^{(A+1)})$ and

$$\xi'_{a_i} = \xi_{a_i} \backslash \{i\}, \xi'_{A+1} = \{i\}, \xi'_a = \xi_a \text{ for } a \in [A] \backslash \{a_i\}.$$

      iii. If $a' \neq a_i$ and $a' \neq A+1$, then $U' = U$ and

$$\xi'_{a_i} = \xi_{a_i} \backslash \{i\}, \xi'_{a'} = \xi_{a'} \cup \{i\}, \xi'_a = \xi_a \text{ for } a \in [A] \backslash \{a_i, a'\}.$$

3. If $|\xi_{a_i}| = 1$:
   (a) Let

$$q_a(\xi, a_i) = \begin{cases} (\eta_\theta + (A-1)\eta_\alpha) p(Y_i|U^{(a_i)}, p) & \text{if } a = a_i, \\ (|\xi_a| - \eta_\alpha) p(Y_i|U^{(a)}, p) & \text{if } a \in [A] \backslash \{a_i\}. \end{cases} \tag{28}$$

   (b) Choose new cluster $a' \sim (p_a)_{a=1,\ldots,A+1}$ where $p_a = q_a / \sum_{j=1}^{A+1} q_j$.
      i. If $a' = a_i$, then $(U', \xi') = (U, \xi)$.
      ii. If $a' \neq a_i$, then $\xi'_{a_i} = \emptyset$ and $\xi'_{a'} = \xi_{a'} \cup \{i\}, \xi'_a = \xi_a$ for $a \in [A] \backslash \{a_i, a'\}$.
      Now remove $\xi'_{a_i} = \emptyset : \xi' \leftarrow \xi' \backslash \xi'_{a_i}$, $U' = U_{-a_i}$ and relabel partition.
4. $X_{t+1} = (U', \xi')$.

F Experiment Results

F.1 Synthetic Data - Trace Plots and Effective Sample Sizes

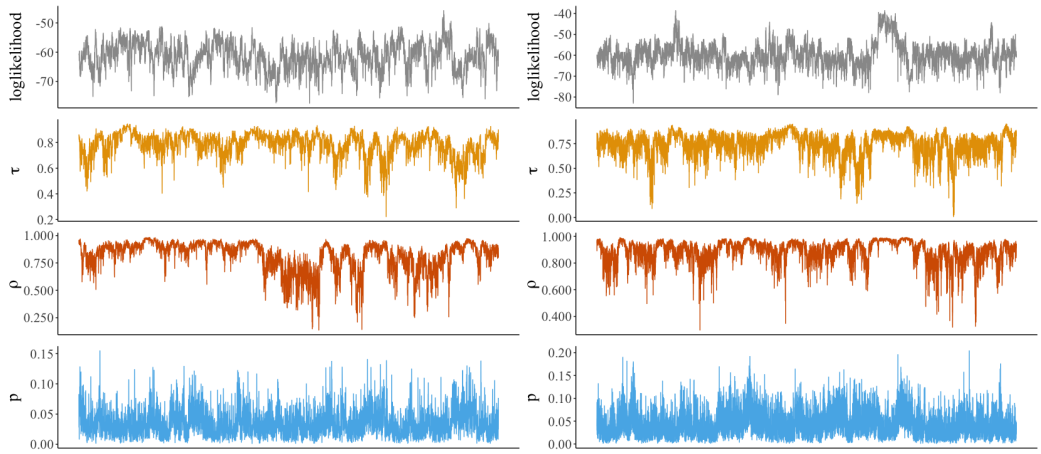


Figure 17: The trace plots for PDP-HPO (left) and PDP-HVSP (right) on the synthetic dataset. We present the trace plot for the log-likelihood (grey), the "roof-to-leaf correlation" parameter τ (yellow), the depth control parameter ρ (blue) and the queue-jumping error probability p (red). The burn-in periods are removed from the trace plots.

	Effective Sample Size	
Parameter	**PDP-HPO**	**PDP-HVSP**
"Roof-to-leaf correlation" parameter τ	87	97
Depth control parameter ρ	88	110
Queue-jumping error probability p	649	246

Table 4: The effective sample sizes (rounded to integer) for key parameters for PDP-HPO and PDP-HVSP on the synthetic rank-order data.

F.2 3-D Sound Spatialisation Data

F.2.1 Trace Plots and Effective Sample Sizes

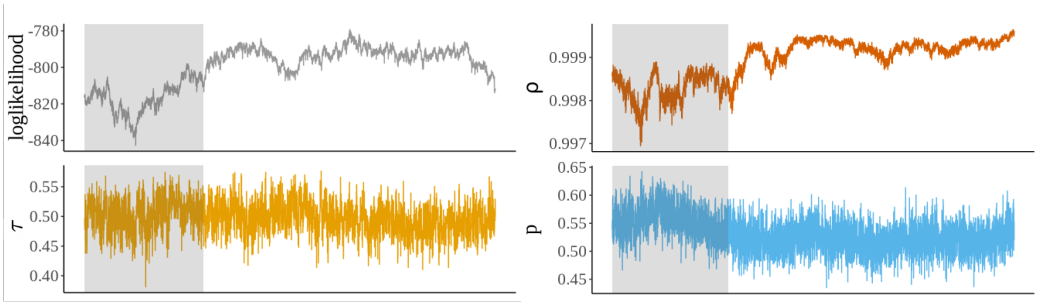


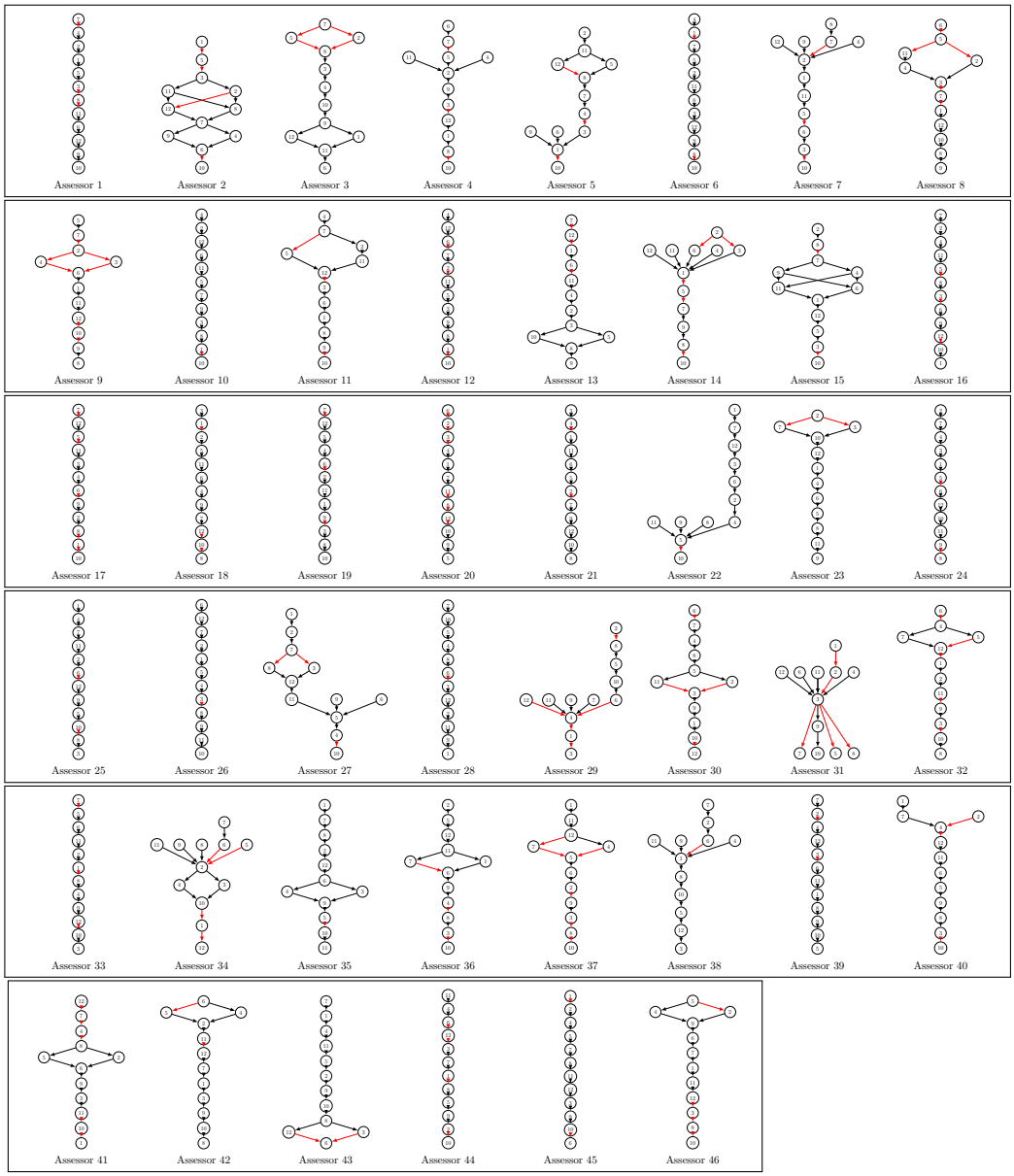
Figure 18: The trace plots for the HVSP model on the 3-D sound spatialisation data. We present the trace plot for the log-likelihood (grey), the "roof-to-leaf correlation" parameter τ (yellow), the depth control parameter ρ (blue) and the queue-jumping error probability p (red). The shaded areas are the burn-in periods. Note the initialisation phase is removed from the trace plots.

Parameter	Effective Sample Size
"Roof-to-leaf correlation" parameter τ	403
Depth control parameter ρ	5
Queue-jumping error probability p	37

Table 5: The effective sample sizes (rounded to integer) for key parameters from the HVSP model on the 3-D sound spatialisation data.

F.2.2 Assessor Concensus Orders

This section displays the concensus orders $h^{con}(\epsilon = 0.5)$ for the 46 individual assessors in the sound spatialisation dataset. We label an edge if it appears more than 50% in the posterior distribution. An edge is labeled red if it appears more than 90% in the VSP posterior samples. All concensus orders are presented as transitive reduction for better visualisation.



F.3 Sushi Preference Data

F.3.1 Sushi item IDs

This section presents the item ID and the corresponding sushi type based on [29].

Item ID	Sushi
1	ebi (shrimp)
2	anago (sea eel)
3	maguro (tuna)
4	ika (squid)
5	uni (sea urchin)
6	ikura (salmon roe)
7	tamago (egg)
8	toro (fatty tuna)
9	tekka maki (tuna roll)
10	kappa maki (cucumber roll)

Table 6: Caption

F.3.2 Trace Plots and Effective Sample Sizes

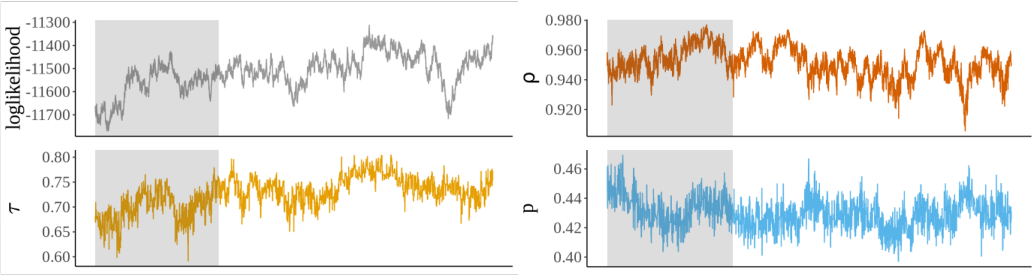


Figure 19: The trace plots for the PDP-HVSP model on the sushi preference data. We present the trace plot for the log-likelihood (grey), the "roof-to-leaf correlation" parameter τ (yellow), the depth control parameter ρ (blue) and the queue-jumping error probability p (red). The shaded areas are the burn-in periods. Note the initialisation phase is removed from the trace plots.

Parameter	Effective Sample Size
"Roof-to-leaf correlation" parameter τ	23
Depth control parameter ρ	23
Queue-jumping error probability p	69

Table 7: The effective sample sizes (rounded to integer) for key parameters from the PDP-HVSP model on the sushi preference data.

F.3.3 Posterior Cluster Allocation

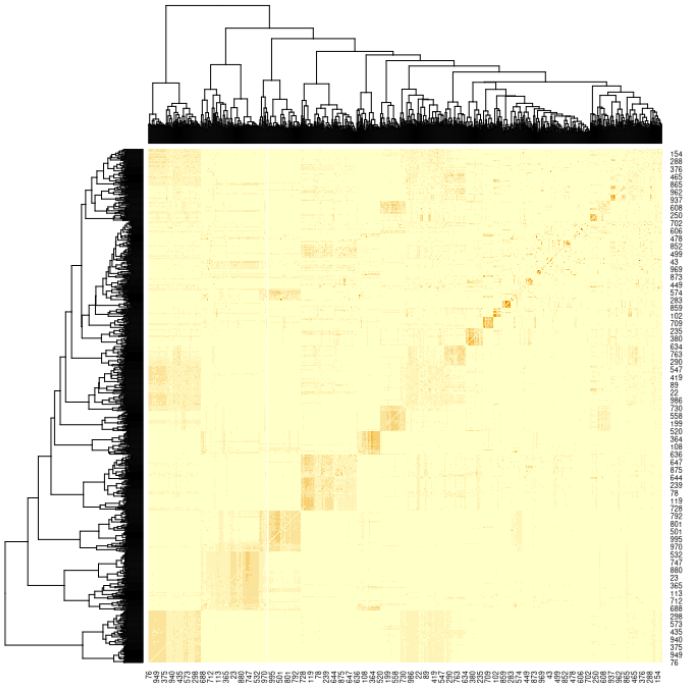


Figure 20: The posterior cluster allocation for the PDP-HVSP model on the sushi preference data. The item pairs with higher posterior probability (orange otherwise yellow) to be in the same cluster are grouped together.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Partial Order Hierarchies
Publication Status	□Published □Accepted for Publication □Submitted for Publication ■Unpublished and unsubmitted work written in a manuscript style
Publication Details	NA

Student Confirmation

Student Name:	Chuxuan (Jessie) Jiang		
Contribution to the Paper	Shared development of models and theory (mainly VSP hierarchy part), performed all simulation and model experiments, co-wrote manuscript.		
Signature 姜楚璇		Date	06/08/2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof Geoff Nicholls		
Supervisor comments		
Signature G.K.Nicholls.	Date	06/08/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Conclusion

This thesis introduced a new class of ranking methods on partial orders and Vertex-Series-Parallel partial orders (VSPs). These models are versatile and applicable across a wide range of ranking scenarios, including social hierarchies, competition analysis, seriation problems, performance reviews, assessor-preference orders and more. As discussed in [22] and [70], ranking methods aim to identify a "central" ranking either 1) as the true order relation that generates the observed rank-orders, or 2) as a heuristic summary of the rank-order data. While the primary focus of this thesis is on the first scenario, where the goal is to uncover the underlying true order relation, our partial order models are also well-suited for applications under the second scenario. These models can be used to provide valuable summaries and insights to understand rank-order data more generally.

We investigated ranking using partial orders based on rank-order data, developing models that accommodate both homogeneous rank-orders and grouped data. Our partial order models are marginally consistent and provide sophisticated control over the depth of random partial orders, a desirable property in many ranking problems. In particular, we focused on Vertex-Series-Parallel (VSP) partial orders, a class that can be reparameterised as binary decomposition trees. These partial orders are advantageous due to their efficient linear extension counting, which enhances the scalability of our models. For the hierarchical partial order model applied to grouped data, we introduced a VSP approximation approach that proves efficient for inference. We also extended models for noise/error in observation of random orders and adapted the Mallows ranking model to the partial order context. We defined a "bi-directional" queue-jumping observation model inspired by [71]. For all proposed models, we developed Markov Chain Monte Carlo (MCMC) algorithms, resulting in an effective inference scheme in practice. Overall, this work addresses a gap in the literature on Bayesian inference on partial orders with rank-order data. Real-world

applications and model comparisons demonstrate that our partial order models are often preferred over total order models due to their greater flexibility in capturing a wide range of order relations.

This thesis comprises three papers that each address different aspects of partial order ranking, with a focus on both theoretical advancements and practical applications. Chapter 2 and Chapter 3 propose new noise models on rank-orders derived from partial orders. Chapter 2 adapts the Mallows model to condition on a partial order, and presents an efficient algorithm for likelihood evaluation. Chapter 3 proposes a novel "bi-directional" queue-jumping noise model, which aligns closely with physical reality. This model is evaluated using Vertex-Series-Parallel (VSP) partial orders, leveraging their computational advantages. Chapter 3 and Chapter 4 explore scalable methods for partial order inference. They work with a scalable class of partial order - VSPs. Both papers highlight the trade-offs between generality and computational efficiency, showing that VSP models can achieve efficiency gains without significantly compromising performance compared to general partial order models. Both Chapter 2 and Chapter 4 utilize non-parametric methods, specifically reversible jump Markov Chain Monte Carlo (MCMC) with the Poisson-Dirichlet process, for clustering. However, the objectives differ: Chapter 2 focuses on clustering items within the partially ordered set (poset) to identify partial orders with ties, while Chapter 4 clusters the observed rank-orders (i.e., the data) to refine the hierarchical partial order model.

This work opens several avenues for future research, which can be categorized into model development, scalable methods, computational improvements, theoretical results, and applications.

Further model development. Our research extends the work of [71] and [69] by incorporating ties and addressing noisy observation lists. However, there is potential for further extension. In Chapter 2, we adapt the Mallows ranking model for partial orders with a uniform dispersion parameter. Future work could explore a Generalized Mallows model, where dispersion parameters vary across different rank positions, allowing for greater flexibility in modeling different degree of uncertainty in ranking positions. Additionally, Chapter 2 only considers the Mallows noise model in the homogeneous setting. Future research could extend this to hierarchical settings (as discussed in Chapter 4), or the time series settings (as seen in [70]). Although [71] explores a random partial order model with queue-jumping noise in a time series context, it is limited to a maximum of around 20 items due to computational constraints. Given the high scalability of the VSP model demonstrated in Chapter 3

and Chapter 4, it would be valuable to investigate VSP's application in time series settings to enhance model scalability.

Regarding the observation model for rank-order data, the current queue-jumping model assumes that all actors have an equal probability of jumping up or down the queue. However, this may not reflect real-world scenarios. For instance, in a rigid social hierarchy, individuals with lower social status are less likely to move up the queue compared to those with higher status. To better capture such hierarchical constraints, we can modify the queue-jumping model by introducing a pairwise queue-jump distance penalty. This penalty would be designed to discourage long-distance jumps in the queue, reflecting the idea that more significant jumps are less likely, especially for individuals with lower status. This adjustment would provide a more realistic representation of queue dynamics in hierarchical contexts.

Incorporating covariates. A stream of research has focused on developing ranking models with covariates. [20, 48] and [70] incorporate covariates on items within a time-series partial order model. Future research could explore integrating covariates into the random VSP model or noise models, such as the queue-jumping model. Specifically, including covariates related to noise in the queue-jumping model could provide a more nuanced understanding of how different factors influence rank-order variability.

Scalable methods. The #P-hard nature of linear extension counting poses significant challenges for model scalability. To address this, we utilize vertex-series-parallel (VSP) partial orders, whose binary decomposition tree (BDT) representation facilitates rapid linear extension counting. To further generalize our approach, we can embed general partial orders as the leaves of the BDT, rather than just individual items. This enhancement could greatly expand the range of partial orders we can represent, while still leveraging the efficient counting capabilities provided by the tree structure.

Another approach is to explore approximation methods for counting the number of linear extensions. One possible avenue is to train graph transformers to estimate the number of linear extensions for large partial orders. Training data can be generated using the class of VSPs for which linear extensions can be efficiently counted, even for hundreds of items. However, most graph neural network and transformer models require node features, which are not directly available in our case. Thus, careful design of node features that capture both local and global structural information is essential. If properly constructed and tuned, this approach could provide accurate estimates of the number of linear extensions. This remains an area for future research.

Computational improvement. Although Chapter 3 and Chapter 4 leverage VSPs for efficient linear extension counting, there is room to enhance likelihood evaluation by exploiting the BDT structure of VSPs further. Additionally, developing a software package for Bayesian inference on partial orders and VSPs with various likelihood models would make our models more accessible to the broader research community.

Theoretical results. Chapter 2 explores the asymptotic behaviour of the (K, n, ρ)-partial order model with the noise-free observation model. Future research could extend these results to include asymptotic consistency for the partial order Mallows and queue-jumping models, offering deeper theoretical insights into these models' performance.

Applications. The partial order and VSP models have been applied to datasets such as social hierarchy studies, sports competition data, and preference rankings. Future research could explore additional applications. For instance, in the bishop data, where witness lists are compiled by "clerks" (assessors) who record the presence of witnesses, there may be variability in how different clerks perceive social importance. Developing a model that accounts for the variable quality of assessors, perhaps incorporating assessor covariates in partial order models could be valuable. These covariates would vary across lists, rather than across the set elements of the partial order as in our work to date. Implementing a hierarchical model in this scenario might be feasible, but it would require careful construction and analysis.

The current models handle complete or incomplete rank-orders. A natural next step is to extend the models to top-k rank-orders. Given that the likelihood models we use—such as the Mallows model, the queue-jumping error model, and the Plackett-Luce model—are all sequential models, it is straightforward to adapt them to consider only the top k items, by stopping the product defining the likelihood for each list at the k'th factor. The partial order would still need to be the full partial order for all elements, not the suborder on the top k. This might limit scalability. However, it would enable more nuanced analysis in situations where only the top rankings are collected.

In summary, the partial order and Vertex-Series-Parallel (VSP) models introduced in this work represent significant advancements in ranking methodologies, with broad applicability to various scenarios such as social hierarchies, competition analysis, and preference rankings. These models not only address the challenge of accurately capturing rank-order relationships but also offer practical solutions for dealing with incomplete or noisy data. Our research contributes to a deeper understanding of how

partial orders can be effectively utilized, providing a robust framework for both theoretical exploration and practical application. The contributions outlined here not only address existing gaps in Bayesian inference for partial orders but also open new avenues for future research. Overall, this research provides a valuable foundation for further exploration and application of partial order models, with the potential to impact a wide range of fields where ranking and order relations are of interest.

Bibliography

[1] Alvo, M. and Philip, L. (2014). *Statistical methods for ranking data*, volume 1341. Springer.

[2] Arcagni, A., Avellone, A., and Fattore, M. (2022). Complexity reduction and approximation of multidomain systems of partially ordered data. *Computational Statistics & Data Analysis*, 173:107520.

[3] Atkinson, M. D. (1990). On computing the number of linear extensions of a tree. *Order*, 7(1):23–25.

[4] Awasthi, P., Blum, A., Sheffet, O., and Vijayaraghavan, A. (2014). Learning mixtures of ranking models. *Advances in Neural Information Processing Systems*, 27.

[5] Banks, J., Garrabrant, S. M., Huber, M. L., and Perizzolo, A. (2018). Using tpa to count linear extensions. *Journal of Discrete Algorithms*, 51:1–11.

[6] Beerenwinkel, N., Eriksson, N., and Sturmfels, B. (2007). Conjunctive bayesian networks. *Bernoulli*, pages 893–909.

[7] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

[8] Brightwell, G. (1993). Models of random partial orders. *Surveys in combinatorics*, 5383.

[9] Brightwell, G. and Winkler, P. (1991). Counting linear extensions. *Order*, 8(3):225–242.

[10] Bubley, R. and Dyer, M. (1999). Faster random generation of linear extensions. *Discrete mathematics*, 201(1-3):81–88.

[11] Caron, F., Teh, W., Yee, M., and Brendan, T. (2014). Bayesian nonparametric plackett-luce models for the analysis of preferences for college. *Annals of Applied Statistics, 8 (2): 1145-1181.*

[12] Caron, F. and Teh, Y. (2012). Bayesian nonparametric models for ranked data. *Advances in Neural Information Processing Systems*, 25.

[13] Cheng, W., Rademaker, M., De Baets, B., and Hüllermeier, E. (2010). Predicting partial orders: ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 215–230. Springer.

[14] Chuxuan, Jiang, and Nicholls, G. K. (2024). Non-parametric bayesian inference for partial orders with ties from rank data observed with mallows noise.

[15] Daskalakis, C., Karp, R. M., Mossel, E., Riesenfeld, S. J., and Verbin, E. (2011). Sorting and selection in posets. *SIAM Journal on Computing*, 40(3):597–622.

[16] De Loof, K., De Meyer, H., and De Baets, B. (2006). Exploiting the lattice of ideals representation of a poset. *Fundamenta Informaticae*, 71(2-3):309–321.

[17] Diaconis, P. (1988a). Group representations in probability and statistics. *Lecture Notes-Monograph Series*, 11:i–192.

[18] Diaconis, P. (1988b). Group representations in probability and statistics. *Lecture notes-monograph series*, 11:i–192.

[19] Dyer, M., Frieze, A., and Kannan, R. (1991). A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17.

[20] Eliseussen, E., Frigessi, A., and Vitelli, V. (2023). Rank-based bayesian clustering via covariate-informed mallows mixtures. *arXiv preprint arXiv:2312.12966.*

[21] Fattore, M., Maggino, F., and Colombo, E. (2012). From composite indicators to partial orders: evaluating socio-economic phenomena through ordinal data. *Quality of life in Italy: Research and reflections*, pages 41–68.

[22] Feng, J., Fang, Q., and Ng, W. (2008). Discovering bucket orders from full rankings. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 55–66.

[23] Feng, Y. and Tang, Y. (2022). On a mallows-type model for (ranked) choices. *Advances in Neural Information Processing Systems*, 35:3052–3065.

[24] Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):359–369.

[25] Fligner, M. A. and Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical association*, 83(403):892–901.

[26] Fröhlich, H., Fellmann, M., Sültmann, H., Poustka, A., and Beißbarth, T. (2007). Large scale statistical inference of signaling pathways from rnai and microarray data. *BMC bioinformatics*, 8:1–15.

[27] Gionis, A., Mannila, H., Puolamäki, K., and Ukkonen, A. (2006). Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566.

[28] Gormley, I. C. and Murphy, T. B. (2009). A grade of membership model for rank data. *Bayesian Analysis*, 4(2).

[29] Guiver, J. and Snelson, E. (2009). Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384.

[30] Heckerman, D., Meek, C., and Cooper, G. (2006). A bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications*, pages 1–28.

[31] Henderson, D. A. and Kirrane, L. J. (2018). A comparison of truncated and time-weighted plackett–luce models for probabilistic forecasting of formula one results. *Bayesian Analysis*, 13(2):335–358.

[32] Hunter, D. R. (2004). Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406.

[33] Irurozki, E., Calvo, B., and Lozano, A. (2014). Sampling and learning the mallows and generalized mallows models under the hamming distance. *Bernoulli (submitted)*, 18:19.

[34] Irurozki, E., Calvo, B., and Lozano, J. A. (2016). Permallows: An r package for mallows and generalized mallows models. *Journal of Statistical Software*, 71(12):1–30.

[35] Irurozki, E., Calvo, B., and Lozano, J. A. (2018). Sampling and learning mallows and generalized mallows models under the cayley distance. *Methodology and Computing in Applied Probability*, 20:1–35.

[36] Irurozki, E., Calvo, B., and Lozano, J. A. (2019). Mallows and generalized mallows model for matchings.

[37] Jagabathula, S., Mitrofanov, D., and Vulcano, G. (2022). Personalized retail promotions through a directed acyclic graph–based representation of customer preferences. *Operations Research*, 70(2):641–665.

[38] Jagabathula, S. and Vulcano, G. (2017). A Partial-Order-based model to estimate individual preferences using panel data. *Management Science*, 64(4):1609–1628.

[39] Jiang, C., Nicholls, G. K., and Lee, J. E. (2023). Bayesian inference for vertex-series-parallel partial orders. In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 995–1004. PMLR.

[40] Johnson, S. R., Henderson, D. A., and Boys, R. J. (2022). On bayesian inference for the extended plackett-luce model. *Bayesian Analysis*, 17(2):465–490.

[41] Kangas, K., Koivisto, M., and Salonen, S. (2020). A faster tree-decomposition based algorithm for counting linear extensions. *Algorithmica*, 82(8):2156–2173.

[42] Karzanov, A. and Khachiyan, L. (1991). On the conductance of order markov chains. *Order*, 8:7–15.

[43] Kenkre, S., Khan, A., and Pandit, V. (2011). On discovering bucket orders from preference data. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 872–883. SIAM.

[44] Khan, M. E., Ko, Y. J., and Seeger, M. (2014). Scalable collaborative bayesian preference learning. In *Artificial Intelligence and Statistics*, pages 475–483. PMLR.

[45] Knuth, D. E. and Szwarcfiter, J. L. (1974). A structured program to generate all topological sorting arrangements. *Information Processing Letters*, 2(6):153–157.

[46] Lebanon, G. and Lafferty, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. In *ICML*, volume 2, pages 363–370.

[47] Lebanon, G. and Mao, Y. (2007). Non-parametric modeling of partially ranked data. *Advances in neural information processing systems*, 20.

[48] Li, X., Yi, D., and Liu, J. S. (2022). Bayesian analysis of rank data with covariates and heterogeneous rankers. *Statistical Science*, 37(1):1–23.

[49] Liu, A. and Moitra, A. (2018). Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 627–638. IEEE.

[50] Liu, A., Zhao, Z., Liao, C., Lu, P., and Xia, L. (2019a). Learning plackett-luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4328–4335.

[51] Liu, Q., Crispino, M., Scheel, I., Vitelli, V., and Frigessi, A. (2019b). Model-based learning from preference data. *Annual Review of Statistics and Its Application*, 6:329–354.

[52] Lu, T. and Boutilier, C. (2014). Effective sampling and learning for mallows models with pairwise-preference data. *J. Mach. Learn. Res.*, 15(1):3783–3829.

[53] Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, 66(2):81.

[54] Ma, J., Yi, X., Tang, W., Zhao, Z., Hong, L., Chi, E., and Mei, Q. (2021). Learning-to-rank with partitioned preference: Fast estimation for the plackett-luce model. In *International Conference on Artificial Intelligence and Statistics*, pages 928–936. PMLR.

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

[56] Mannila, H. (2008). Finding total and partial orders from data for seriation. In *International Conference on Discovery Science*, pages 16–25. Springer.

[57] Mannila, H. and Meek, C. (2000). Global partial orders from sequential data. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 161–168.

[58] Manski, C. F. (1977). The structure of random utility models. *Theory and decision*, 8(3):229.

[59] Marden, J. I. (1996). *Analyzing and modeling rank data*. CRC Press.

[60] Maystre, L. and Grossglauser, M. (2015). Fast and accurate inference of plackett–luce models. *Advances in neural information processing systems*, 28.

[61] Meila, M. and Bao, L. (2010). An exponential model for infinite rankings. *J. Mach. Learn. Res.*, 11:3481–3518.

[62] Meila, M. and Chen, H. (2012). Dirichlet process mixtures of generalized mallows models. *arXiv preprint arXiv:1203.3496*.

[63] Meilă, M. and Chen, H. (2016). Bayesian non-parametric clustering of ranking data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2156–2169.

[64] Möhring, R. H. (1989). Computationally tractable classes of ordered sets. In *Algorithms and order*, pages 105–193. Springer.

[65] Mollica, C. and Tardella, L. (2014). Epitope profiling via mixture modeling of ranked data. *Statistics in Medicine*, 33(21):3738–3758.

[66] Mollica, C. and Tardella, L. (2017). Bayesian plackett–luce mixture models for partially ranked data. *Psychometrika*, 82(2):442–458.

[67] Mollica, C. and Tardella, L. (2018). Algorithms and diagnostics for the analysis of preference rankings with the extended plackett-luce model. *arXiv preprint arXiv:1803.02881*.

[68] Mollica, C. and Tardella, L. (2020). Plmix: An r package for modelling and clustering partially ranked data. *Journal of Statistical Computation and Simulation*, 90(5):925–959.

[69] Muir Watt, A. (2015). *Inference for partial orders from random linear extensions*. PhD thesis, University of Oxford.

[70] Nicholls, G. K., Lee, J. E., Karn, N., Johnson, D., Huang, R., and Muir-Watt, A. (2023). Bayesian inference for partial orders from random linear extensions: power relations from 12th century royal acta.

[71] Nicholls, G. K. and Muir Watt, A. (2011). Partial order models for episcopal social status in 12th century england. *IWSM 2011*, page 437.

[72] Niinimaki, T., Parviainen, P., and Koivisto, M. (2012). Partial order mcmc for structure discovery in bayesian networks. *arXiv preprint arXiv:1202.3753*.

[73] Niinimäki, T., Parviainen, P., and Koivisto, M. (2016). Structure discovery in bayesian networks by sampling partial orders. *Journal of Machine Learning Research*, 17(57):1–47.

[74] Peyhardi, J., Trottier, C., and Guédon, Y. (2016). Partitioned conditional generalized linear models for categorical responses. *Statistical Modelling*, 16(4):297–321.

[75] Piancastelli, L. S. and Friel, N. (2024). Clustered mallows model. *arXiv preprint arXiv:2403.12880*.

[76] Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202.

[77] Puolamäki, K., Fortelius, M., and Mannila, H. (2006). Seriation in paleontological data using markov chain monte carlo methods. *PLoS computational biology*, 2(2):e6.

[78] Rising, J. (2021). Uncertainty in ranking. *arXiv preprint arXiv:2107.03459*.

[79] Sakoparnig, T. and Beerenwinkel, N. (2012). Efficient sampling for bayesian inference of conjunctive bayesian networks. *Bioinformatics*, 28(18):2318–2324.

[80] Seshadri, A., Peysakhovich, A., and Ugander, J. (2019). Discovering context effects from raw choice data. In *International conference on machine learning*, pages 5660–5669. PMLR.

[81] Seshadri, A., Ragain, S., and Ugander, J. (2020). Learning rich rankings. *Advances in Neural Information Processing Systems*, 33:9435–9446.

[82] Soliman, M. A., Ilyas, I. F., and Ben-David, S. (2010). Supporting ranking queries on uncertain and incomplete data. *The VLDB Journal*, 19:477–501.

[83] Talvitie, T., Kangas, K., Niinimäki, T., and Koivisto, M. (2018). Counting linear extensions in practice: Mcmc versus exponential monte carlo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[84] Talvitie, T., Niinimäki, T. M., and Koivisto, M. (2017). The mixing of markov chains on linear extensions in practice. In *IJCAI*, pages 524–530.

[85] Tang, W. (2019). Mallows ranking models: maximum likelihood estimate and regeneration. In *International Conference on Machine Learning*, pages 6125–6134. PMLR.

[86] Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4):273.

[87] Tkachenko, M. and Lauw, H. W. (2016). Plackett-luce regression mixture model for heterogeneous rankings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 237–246.

[88] Valdes, J., Tarjan, R. E., and Lawler, E. L. (1979). The recognition of series parallel digraphs. In *Proceedings of the eleventh annual ACM symposium on Theory of computing*, pages 1–12.

[89] Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., and Arjas, E. (2018). Probabilistic preference learning with the mallows rank model. *Journal of Machine Learning Research*, 18(158):1–49.

[90] Wells, M. B. (1971). Elements of combinatorial computing.

[91] Winkler, P. (1985). Random orders. *Order*, 1(4):317–331.

[92] Zhang, Q. and Ip, E. H. (2012). Generalized linear model for partially ordered data. *Statistics in Medicine*, 31(1):56–68.

[93] Zhao, Z., Liu, A., and Xia, L. (2020). Learning mixtures of random utility models with features from incomplete preferences. *arXiv preprint arXiv:2006.03869*.

[94] Øystein Sørensen, Crispino, M., Liu, Q., and Vitelli, V. (2020). BayesMallows: An R Package for the Bayesian Mallows Model. *The R Journal*, 12(1):324–342.