A Probabilistic Graphical Model for Mallows Preferences

Batya Kenig and Benny Kimelfeld
Technion, Israel Institute of Technology

Abstract

Reasoning about preference distributions is an important task in various areas (e.g., recommender systems, social choice), and is critical for learning the parameters of the distribution. We consider the Mallows model with arbitrary pairwise comparisons as evidence. Existing inference methods are able to reason only with evidence that abides to a restrictive form. We establish the conditional independences in the Mallows model, and apply them to develop a Bayesian network that enables querying and sampling from the Mallows posterior (the conditional probability space with the evidence incorporated). While inference over the Mallows posterior is computationally hard in general, our translation allows to utilize the wealth of tools for inference over Bayesian networks. Moreover, we show how our translation gives rise to new results on significant cases with a polynomial-time inference.

1 Introduction

Preferences drive our choices, making them prevalent in applications and systems that support user decisions and act on their behalf. Preferences reflect the relative quality or desirability of the outcomes over some domain of possible choices, and can arise in a variety of forms, including total orders or rankings, top-k lists, and partial orders. The proliferation and volume of preference data gives rise to special analytical tasks over them. Examples include rank aggregation in *genomic data* [Boulesteix and Slawski, 2009; Kolde *et al.*, 2012], management and analysis of elections [Gormley and Murphy, 2008; McElroy and Marsh, 2009], and recommendation systems in *electronic commerce* [Das Sarma *et al.*, 2010]. Incorporating sophisticated analytics into applications that operate over modern volumes of preference data poses a new challenge to knowledge representation and inference formalisms required to model and query such data.

Statistical models over ranking data, developed in the statistics and psychometrics literature [Marden, 1995], are used in order to form recommendations, make predictions, and uncover trends. Current approaches to querying preference distributions accommodate only very restricted forms of evidence about individual user preferences, like complete rankings, top-k/bottom-k alternatives, and *partitioned preferences* [Lebanon and Mao, 2008]. However, preference data will often take the form of arbitrary pairwise comparisons (or partial orders), as expected in scenarios of search, product comparison, and online advertising [Huang *et al.*, 2012; Lu and Boutilier, 2014].

Statistical models for generating ranking data assume a group of users with a *reference ranking* (often called *modal*). The ranking corresponding to each user is a noisy version of the reference ranking that is generated independently for each user. The *Mallows model* [Mallows, 1957] has received much attention from the statistics and machine learning communities [Lu and Boutilier, 2014; Awasthi *et al.*, 2014; Lebanon and Mao, 2008]. This model is parametrized by *(a)* a reference ranking $\sigma = \sigma_1, \ldots, \sigma_m$ over a set A of m alternatives, *and (b)* a *dispersion parameter* $\phi \in (0, 1]$. The probability of a ranking r in this model is proportional to $\phi^{d(\sigma,r)}$, where $d(\sigma, r)$ denotes the Kendall's tau distance between permutations (see Section 2 for a formal definition).

The conditional (posterior) distribution implied by incorporating evidence into a Mallows model is termed the *Mallows posterior* (see Equation (5)). In this paper we make no assumptions regarding the evidence, which can take the form of any partial order. Computing the probability of evidence, and sampling from the Mallows posterior, are intractable under standard complexity assumptions [Lu and Boutilier, 2014]. The artificial intelligence community is no stranger to such computational challenges, and has developed tools for representing distributions over a combinatorially large space, along with a suite of algorithms to efficiently perform probabilistic inference over these distributions.

The *Repeated Insertion Model (RIM)* [Doignon *et al.*, 2004], is a generative process that gives rise to a family of distributions over rankings and offers a constructive, simple, and useful way to sample complete rankings from the Mallows distribution. Probabilistic graphical models have been successfully applied to the task of modeling and reasoning about generative processes, including speech recognition [Zweig and Russell, 1998], document classification [Hofmann, 2001], medical diagnosis [Heckerman, 1989], social network analysis [Farasat *et al.*, 2015] and object tracking [Pavlović *et al.*, 1999]. Using probabilistic graphical models to represent the RIM process facilitates the application of state-of-the-art algorithms for the task of reasoning

about partial orders in the Mallows distribution.

The task we address in this paper is that of computing the probability that a ranking abides to a given partial order. This is essentially the *partition function* (i.e., the normalization factor) associated with the Mallows posterior. Furthermore, a Bayesian network representation of the Mallows posterior enables applying known and well tested sampling techniques, which are critical for learning the Mallows model parameters [Lu and Boutilier, 2014; Stoyanovich *et al.*, 2016; Awasthi *et al.*, 2014; Huang *et al.*, 2012].

Contributions. Our contributions are as follows.

1. We introduce a probabilistic model for RIM and use it to characterize the conditional independence relationships inherent in the RIM process.
2. We use these conditional independence relationships in order to construct a Bayesian network that enables inferring the probability of evidence in an arbitrary form.
3. We establish new fragments of partial orders that do not fall into any of the known tractable cases, but that allow for efficient inference using our proposed approach

Roadmap. In Section 2 we introduce required notation, and provide an overview of the Mallows model and Bayesian networks. We introduce the probabilistic model for RIM in Section 3. In Section 4 we show how the Mallows posterior distribution is represented using Bayesian networks, and introduce a new fragment of partial orders whose Bayesian network representation enables performing inference in polynomial time. We conclude in Section 5.

2 Preliminaries

2.1 Mallows and the Repeated Insertion Model

Basic notation. We follow the notation of Lu and Boutilier [2014]. We assume an agent has a total order (or ranking), $\succ$, over a set $A = \{a_1, a_2, \ldots, a_m\}$ of m alternatives. The relationship $x \succ y$ means that the agent prefers alternative x to y. The ranking is represented as a bijection $\sigma : A \to [m]$, where $\sigma(a)$ is the position of item a in the ranking. We denote by $\sigma = \sigma_1, \ldots, \sigma_m$ a ranking with the i-th ranked alternative $\sigma_i \in A$, and the induced preference relation as $\succ_\sigma$.

In many cases the agent's complete ranking is unknown, and we have only partial information about it in the form of a set of pairwise *preferences*: $\nu = \{x_1 \succ y_1, \ldots, x_k \succ y_k\}$. The semantics of ν is that of its transitive closure, denoted ν^*. If ν^* is a total order on A, then we say that the set ν is *complete*. Given a ranking $\sigma = \sigma_1, \ldots, \sigma_m$ and preferences ν, we define the *distance* between the two to be the number of pairs in ν that are misordered relative to σ:

$$d(\nu, \sigma) = \sum_{1 \le i < j \le m} \mathbb{1}[\sigma_j \succ \sigma_i \in \nu^*] \tag{1}$$

where $\mathbb{1}$ is the indicator function that returns 1 for all pairs in ν^* (the transitive closure of ν) and 0 on all other pairs. When ν is a complete ranking, $d(\nu, \sigma)$ is the classic *Kendall's tau* metric on rankings [Kendall, 1938].

The Mallows model is parameterized by a ranking σ called a *reference ranking*, and a number $\phi \in (0, 1]$ called a *dispersion parameter*. The probability of observing a ranking r is given by

$$P(r) = P(r|\sigma, \phi) = \frac{1}{Z}\phi^{d(r,\sigma)} \tag{2}$$

where $d(r, \sigma)$ is the distance between r and σ as defined in (1), and Z, the normalization constant, is given by

$$Z = (1 + \phi)(1 + \phi + \phi^2) \ldots (1 + \cdots + \phi^{m-1}) . \tag{3}$$

The Mallows model has been phrased via a *rejection sampling procedure*, which results in the probability distribution of Equation (2) [Mallows, 1957]. This procedure, however, is highly inefficient due to the high rejection rate of intransitive rankings. An efficient sampling procedure is the RIM process [Doignon *et al.*, 2004] that we describe next.

Algorithm RIM(σ,ϕ)

1: Let r be an empty ranking
2: **for all** $i \in 1 \ldots m$ **do**
3: Insert σ_i into r at position $j \le i$ with probability $\frac{\phi^{i-j}}{(1+\phi+\phi^2+\cdots+\phi^{i-1})}$

Figure 1: RIM Sampling of Mallows

The Repeated Insertion Model (RIM) is a generative process that gives rise to the distribution of Equation (2) and provides a practical way to sample rankings from a Mallows model [Doignon *et al.*, 2004]. The RIM procedure, presented in Figure 1, assumes a reference ranking $\sigma = \sigma_1, \ldots, \sigma_m$, and dispersion parameter ϕ.

RIM generates a new output ranking, r, by inserting items $\sigma_1, \ldots, \sigma_m$ into r, according to their order in σ (the reference ranking). After $i - 1$ rounds, the ranking r will contain items $\sigma_1, \ldots, \sigma_{i-1}$. Then, in round i, the probability of inserting σ_i into position $j \in [1, i]$ in r is:

$$p_{ij} = \frac{\phi^{i-j}}{1 + \phi + \phi^2 + \cdots + \phi^{i-1}} \tag{4}$$

Critically, the insertion probabilities are independent of the ordering of the previously inserted alternatives.

The Mallows posterior is the conditional distribution that results from incorporating evidence, ν, in the form of a partial order, into the Mallows distribution. Formally, for the Mallows distribution given by σ and ϕ, the Mallows posterior under the evidence ν is given by

$$P_\nu(r) = \frac{\phi^{d(r,\sigma)}}{\sum_{r' \in \Omega(\nu)} \phi^{d(r',\sigma)}} \cdot \mathbb{1}[r \in \Omega(\nu)] \tag{5}$$

where $\Omega(\nu)$ is the set of complete rankings over A that are consistent with ν. The probability of the partial order ν is the denominator of Equation (5) divided by Z, the normalization constant (Equation (3)); formally, it is given by

$$P(\nu) = \frac{1}{Z} \sum_{r' \in \Omega(\nu)} \phi^{d(r',\sigma)} .$$

RIM is an efficient sampler, but it does not provide an efficient way to sample from the Mallows posterior, which is important when the goal is to reason about the probability of partial orders.

2.2 Bayesian Networks

We denote Random Variables (RVs) by capital letters (e.g., X, Y, Z), and their values by lowercase letters (e.g., x, y, z). Sets of RVs are denoted by bold capital letters (e.g., $\mathbf{X}, \mathbf{Y}$) and their values by bold lowercase letters (e.g., $\mathbf{x}, \mathbf{y}$).

Two RVs X and Y are *independent*, denoted $X \perp Y$, if

$$P(X, Y) = P(X)P(Y)$$

Two RVs X and Y are *conditionally independent* given a set $\mathbf{Z}$ of RVs, denoted $X \perp Y \mid \mathbf{Z}$, if

$$P(X, Y \mid \mathbf{Z}) = P(X \mid \mathbf{Z})P(Y \mid \mathbf{Z})$$

A Probabilistic Graphical Model (PGM) is a compact representation of a probability distribution that is too large to be handled using traditional specifications such as tables and equations. Probabilistic graphical models exploit the structure induced by the probabilistic independence properties that exist in many distributions modeling real-world phenomena. This structure, captured graphically, enables factoring the representation of the distribution into modular components that lead to a compact representation of high-dimensional distributions. Graphical models are equipped with a suite of inference algorithms that enable to automatically infer implications of the represented information.

A special case of a PGM is a Bayesian Network (BN) that represents a joint probability distribution over a set $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ of RVs. The variables are represented as nodes in a directed acyclic graph, and an edge between two nodes represents a direct probabilistic dependency between them, such that the joint distribution is given by the *chain rule*

$$P(x_1, \ldots, x_n) = \prod_i P(x_i | \text{parents}(X_i))$$

where $\text{parents}(\text{X}_\text{i})$ are the nodes with outgoing edges to X_i. Each node has a *Conditional Probability Table* (CPT) that enumerates $P(X_i|\text{parents}(X_i))$ for all possible values of X_i and $\text{parents}(X_i)$.

Although probabilistic inference is generally a #P-complete problem, existing algorithms are able to take advantage of the conditional independences encoded in the graphical structure. Specifically, the complexity of exact inference algorithms is exponential in the BN's *treewidth*, a graph-theoretic parameter that measures the resemblance of a graph to a tree [Darwiche, 2010]. As the network becomes denser, its treewidth increases, as does the complexity of inference. This lead to the development of approximation algorithms for performing inference [Darwiche, 2010]. Casting the Mallows posterior as a BN enables to capitalize on the vast progress made in the field of PGMs over the last few decades.

3 A representation for RIM

In what follows we consider the ranking, generated by RIM, at each time $t \in \{1, 2, \ldots, m\}$.

Definition 1 (Position RV)**.** *We denote by X_i^t the position of item $\sigma_i \in A$ at time t, right after the insertion of item σ_t into the ranking. We denote by $\mathbf{X}^t = \{X_1^t, X_2^t, \ldots, X_t^t\}$ the positions of items $\sigma_1, \ldots, \sigma_t$ in the ranking at time t. The* time span *of X_i^t is the interval $[i, t]$.*

A position RV (Definition 1) whose subscript and superscript are identical (i.e., X_i^i) is simply the insertion position of σ_i, and is thus termed *insertion RV*.

Definition 2 (Ranking at time t)**.** *We denote by r^t the ranking over the items $\sigma_1, \ldots, \sigma_t$, as generated by the RIM process, at time t.*

We note that for any reference ranking σ there is a bijection between r^t and $\mathbf{X}^t$.

Definition 3 (Insertion Probabilities [Doignon *et al.*, 2004])**.** *We denote by w_i^k the probability that item $\sigma_i \in A$ is inserted into position $k \leq i$ at time i. Therefore:*

$$w_i^k = P(X_i^i = k) = \phi^{i-k} \left(\frac{1 - \phi}{1 - \phi^i} \right)$$

We use the following property of the RIM process. The insertion of the ith alternative, σ_i, at some position in the ranking r is independent of the positioning of all previously inserted alternatives $\sigma_1, \ldots, \sigma_{i-1}$ [Doignon *et al.*, 2004]. The independence always holds for RIM. When considering the posterior distribution this is no longer the case because the position of σ_i may depend on the positions of items $\{\sigma_j \in A : j < i\}$, previously placed in the ranking.

3.1 Position Probabilities

One of the building blocks we will need for generating a BN for partial orders over Mallows involves calculating the probability that item σ_i at time $j > i$ is at position l (i.e., $X_i^j = l$). The steps outlined in Equation (6) define a dynamic programming algorithm that calculates the probability that $X_i^j = l$. The event $X_i^j = l$ can take place in one of two ways. First, σ_i was at position $l - 1$ at the previous time $(j - 1)$ and σ_j was inserted somewhere before it (i.e., $X_i^{j-1} = l-1$ and $X_j^j \leq l$). Second, it must have already been in position l at time $j - 1$, and σ_j was inserted after it (i.e., $X_i^{j-1} = l$ and $X_j^j > l$). The rest of the transitions in Equation (6) are due to the RIM independence property that states that the insertion probabilities do not depend on the placement of the previous members. The base case is when $j = i$ and then $P(X_i^i = l) = w_i^l$.

The complexity of the dynamic programming algorithm defined by Equation (6) is determined by the number of distinct states that require computation. Each state is defined by a time slice $t \in [i, j]$, and a possible position of item σ_i in the range $[1, l]$, at time t. Overall, this leads to a complexity of $O((j - i) \cdot l)$.

$$
\begin{aligned}
&P(X_i^j = l) = \\
&P(X_i^{j-1} = l-1, X_j^j \le l-1) + P(X_i^{j-1} = l, X_j^j > l) = \\
&P(X_j^j \le l-1)P(X_i^{j-1} = l-1) + P(X_j^j > l)P(X_i^{j-1} = l) = \\
&P(X_i^{j-1} = l-1)\sum_{q=1}^{l-1} P(X_j^j = q) + \\
&\qquad\qquad P(X_i^{j-1} = l)\sum_{s=l+1}^{j} P(X_j^j = s) = \\
&P(X_i^{j-1} = l-1)\sum_{q=1}^{l-1} w_j^q + P(X_i^{j-1} = l)\sum_{s=l+1}^{j} w_j^s
\end{aligned}
\tag{6}
$$

The dynamic programming algorithm of Equation (6) can easily be extended to compute the conditional probability $P(X_i^j = l \mid X_i^k = l')$, where $i \le k < j$ and $l' \le l$. The only difference is that the base case is different and occurs when $j = k$ (instead of $j = i$). The complexity of computing the conditional probability $P(X_i^j = l \mid X_i^k = l')$ is $O((j-k)(l-l'))$.

Proposition 1 states that our model enables computing the probability of a single preference efficiently.

Proposition 1 (Single Ordered Pair)**.** *Let $\sigma = \sigma_1, \ldots, \sigma_m$ denote the reference ranking of a Mallows distribution over a set A of m items. Let $\sigma_i, \sigma_j \in A$ such that w.l.o.g $i < j$. The probability of the partial order $\nu = \{\sigma_j \succ \sigma_i\}$ can be computed in time $O(m^3)$.*

Proof. We show how to apply the dynamic programming algorithm from Equation (6) in order to compute the desired probability. The key observation is that for any pair of items $\sigma_i, \sigma_j \in A$ such that (w.l.o.g) $i < j$, the relative ordering between σ_i and σ_j is determined when σ_j is inserted into the ranking, and this relative ordering remains unchanged through out the RIM process. Therefore:

$$
\begin{aligned}
P(X_i^m > X_j^m) &= P(X_i^{j-1} \ge X_j^j) \\
&= \sum_{l=1}^{j-1} P(X_i^{j-1} = l, X_j^j \le l) \\
&= \sum_{l=1}^{j-1} P(X_j^j \le l \mid X_i^{j-1} = l) \cdot P(X_i^{j-1} = l) \\
&= \sum_{l=1}^{j-1} P(X_i^{j-1} = l) \cdot \sum_{q=1}^{l} P(X_j^j = q \mid X_i^{j-1} = l) \\
&= \sum_{l=1}^{j-1} P(X_i^{j-1} = l) \cdot \sum_{q=1}^{l} P(X_j^j = q) \\
&= \sum_{l=1}^{j-1} P(X_i^{j-1} = l) \cdot (w_j^1 + \cdots + w_j^l)
\end{aligned}
\tag{7}
$$

The probability $P(X_i^{j-1} = l)$ can be computed using Equation (6) in time $O((j-i)\cdot l)$, leading to an overall complexity of $\sum_{l=1}^{j-1} O((j-i)\cdot l) = O(j^3)$. □

We note that the partial order of Proposition 1, despite being modest in size, does not fall into any of the known restrictive forms of evidence for which the probability can be computed efficiently.

3.2 Temporal Independence in RIM

The Bayesian network we will develop for inferring the Mallows posterior exploits the inherent temporal independences in the RIM process. Let $\sigma_i \in \sigma$, and let $t > i$ be some time. Intuitively, we see that the position of σ_i at some time $t' > t$, will not depend on any event that occurred before time t given that we know the value of X_i^t, (i.e., σ_i's position at this time). We formalize this intuition in the following Lemma (proofs are deferred to the full version of this paper).

Lemma 1. *Let $\sigma_i, \sigma_j \in \sigma$ and let t be a time such that $t > \max(i, j)$. Then:*

$$X_i^{t+k} \perp X_j^{t-l} \mid X_i^t$$

where k, l are integers such that $k \ge 0$ and $1 \le l \le t - j$. This proposition also holds for $i = j$.

Lemma 1 is used to prove the following lemma.

Lemma 2. *Let $\sigma_i, \sigma_j \in \sigma$ such that $i < j$. Then:*

$$X_i^{j-l} \perp X_j^{j+k}$$

where k, l are integers such that $k \ge 0$ and $1 \le l \le j - i$.

The following theorem is applied in the algorithm for generating the Bayesian network.

Theorem 1. *Let X_i^s, X_j^t, $i \ne j$, denote two position RVs with disjoint time spans, (i.e., $[i, s] \cap [j, t] = \emptyset$). Then $X_i^s \perp X_j^t$.*

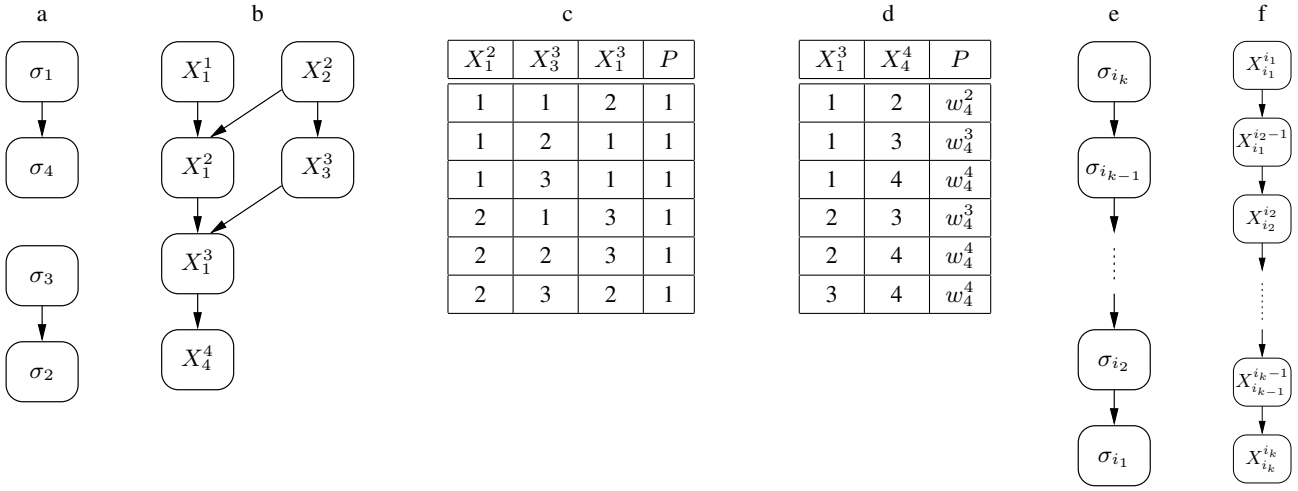
Proof. Assume w.l.o.g that $i < j$. Since $[i, s] \cap [j, t] = \emptyset$ then $s < j$, that is, $s = j - l$ for some $l \ge 1$. Also, $t \ge j$ and therefore $t = j + k$ for some $k \ge 0$. The result follows immediately from Lemma 2. □

4 A Bayesian Network for RIM

In this section we show how the method of computing an item's position (Section 3.1) and the temporal independences in RIM are applied to the generation of a BN for RIM. In what follows, the reference ranking is denoted by σ and the partial order by ν.

The BN generation procedure can be viewed as tracking the RIM process (Figure 1), while enforcing the constraints dictated by ν. Specifically, at each step $i \in [1, m]$, the BN is augmented with the constraints that relate the position of item σ_i with the positions of $\sigma_1, \ldots, \sigma_{i-1}$, which are already part of the ranking.

In the complete version of this paper we shall provide a detailed algorithm for constructing the BN from a given partial order and prove all of the conditional independences captured in the network. We will also show that the size of the BN is polynomial in m, the size of the reference ranking. Specifically, we prove that the BN has a polynomial number of variables and that each CPT contains a polynomial number of entries. This does not mean that we can perform inference in polynomial time because the complexity of inference depends



X_1^2	X_3^3	X_1^3	P
1	1	2	1
1	2	1	1
1	3	1	1
2	1	3	1
2	2	3	1
2	3	2	1

X_1^3	X_4^4	P
1	2	w_4^2
1	3	w_4^3
1	4	w_4^4
2	3	w_4^3
2	4	w_4^4
3	4	w_4^4

Figure 2: (a) The partial order $\nu = \{\sigma_1 \succ \sigma_4, \sigma_3 \succ \sigma_2\}$ (b) BN for ν, the partial order in Figure 2a (c) CPT for the family $\{X_1^2, X_3^3, X_1^3\}$ in the BN of Figure 2b, representing the affect of X_1^2 and X_3^3 on the position of σ_1 at time 3 (i.e., X_1^3) (d) CPT for the family $\{X_1^3, X_4^4\}$ in the BN of Figure 2b, that represents the constraint $X_4^4 > X_1^3$. (e) Graphical representation of partial order $\xi = \{\sigma_{i_k} \succ \sigma_{i_{k-1}} \succ \sigma_{i_{k-2}} \succ \cdots \succ \sigma_{i_1}\}$ (f) BN for ξ, the partial order in Figure 2e

exponentially on the *width* of the BN, which is not necessarily fixed [Darwiche, 2010]. There are partial orders that, under our model, induce BNs that admit polynomial time inference algorithms. Theorem 2 provides a characterization of such a class of partial orders.

In this section, we focus on how the ordering constraints are included in the CPTs of the BN. We begin with an example of a BN that represents two pairwise preferences. A single pairwise preference was considered in Theorem 1.

Example 1. *We consider a Mallows model over five items, with $\sigma = \sigma_1\sigma_2\sigma_3\sigma_4\sigma_5$ and dispersion parameter ϕ. The partial order, $\nu = \{\sigma_1 \succ \sigma_4, \sigma_3 \succ \sigma_2\}$, is presented in Figure 2a. A RIM-generated ranking, r, can abide to this constraint only if at time $t = 4$, σ_4 is inserted into a position that is larger than the position of σ_1, and, at time $t = 3$, σ_3 is inserted into a position that is not larger than the position of σ_2, or formally, only if $X_4^4 > X_1^3$, and $X_3^3 \leq X_2^2$. These constraints are captured in the BN of Figure 2b.*

The CPT in Figure 2d contains only entries that abide to the pairwise constraint $\sigma_1 \succ \sigma_4$, or equivalently, $X_4^4 > X_1^3$. The probability column of these entries incorporates the insertion probability of σ_4 into the ranking r.

At a first glance, it would seem that we could have constructed our BN without the random variables X_1^1 and X_1^2, since these do not directly take part in any of the constraints (see Figure 2a). However, the value of X_1^3, which is required for modeling the ordering constraints, is affected by the insertion positions of both σ_2 and σ_3 (i.e., X_2^2 and X_3^3). This affect is captured using Equation (8) and presented in tabular form in Figure 2c.

Equation (8), and its corresponding CPT (Figure 2c), essentially encode the functional dependency of σ_i's position at time k (i.e., X_i^k) on its position in the previous time slice (i.e., X_i^{k-1}), and the insertion position of item σ_k (i.e., X_k^k).

$$P(X_i^k | X_i^{k-1}, X_k^k) = \begin{cases} 1 & \text{if } X_k^k \leq X_i^{k-1} \text{ and } X_i^k = X_i^{k-1} + 1 \\ 1 & \text{if } X_k^k > X_i^{k-1} \text{ and } X_i^k = X_i^{k-1} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The BN generation procedure is outlined as follows.

Compute the set of position and insertion RVs (Definition 1) that are required to represent the constraints imposed by the partial order. This step is carried out by examining the constraints that are relevant at each time $i \in [1, m]$, and generating the associated RVs. In Example 1, ν^2, the partial order relevant to time $i = 2$, is empty. On the other hand, the partial order $\nu^3 = \{\sigma_3 \succ \sigma_2\}$, induces the creation of RVs X_2^2 and X_3^3 (see Figure 2b), which are required to enforce the constraint that $X_3^3 \leq X_2^2$.

Apply the conditional independence propositions from Section 3.2 in order to generate the Directed Acyclic Graph (DAG) of the BN. Theorem 1 provides a characterization of when two position RVs X_i^s and X_j^t, are independent. In the full version of this paper we will describe how to relate between these position RVs when the condition of the theorem is not met (that is, $[i, s] \cap [j, t] \neq \emptyset$).

Compute the CPTs for each family, $\{X_i \cup \text{parents}(X_i)\}$, in the DAG generated in the previous step. We partition the CPTs of the BN into two types.

The first type of CPT specifies the possible values of an *insertion RV*, X_i^i, based on the assignment to $\text{parents}(X_i^i)$, the nodes with outgoing edges to X_i^i. (We remind the reader that an insertion RV, X_i^i, represents the insertion position of item σ_i into the ranking.)

In CPTs of this type, the parents represent certain positions in the ranking at time $i - 1$, right before the insertion of σ_i into the ranking takes place. The parent RVs determine the possible range of values that the insertion RV can be as-

signed. Specifically, X_i^i, the insertion RV, contains at most two parents, which represent the lower and upper bounds on the positions that X_i^i can be assigned. For example, assume that ν, the partial order, contains the following constraint $\sigma_j \succ \sigma_i \succ \sigma_k$, where $j < i$ and $k < i$. In order to generate a ranking that complies to this constraint, item σ_i needs to be inserted before σ_k and after σ_j. Translating this constraint into our model we arrive at $X_j^{i-1} < X_i^i \leq X_k^{i-1}$. Therefore, in the BN, we will have that $\text{parents}(X_i^i) = \{X_j^{i-1}, X_k^{i-1}\}$. An example of such a CPT, in tabular form, is presented in Figure 2d.

In many cases, the boundaries of the insertion RV are a function of other positions. For example, assume the following set of pairwise constraints, $\nu = \{\sigma_i \succ \sigma_k, \sigma_i \succ \sigma_j, \sigma_i \succ \sigma_l\}$. Applying our model, this constraint is translated to $X_i^i \leq \min(X_j^{i-1}, X_k^{i-1}, X_l^{i-1})$. In the complete version of this paper we will describe how to generate CPTs that represent more complex constraints on the insertion RV.

The second type of CPT enables tracking the position of an item throughout the RIM process. These CPTs represent the relationship between two position RVs, X_i^s and X_i^t, associated with the same item, σ_i. The entries in these CPTs are computed using the dynamic programming algorithm of Equation (6).

The fact that the CPTs of the resulting BN are polynomial in size has important implications regarding the study of fragments of partial orders that allow for efficient reasoning. For example, partial orders that induce polytrees (a directed acyclic graph in which the subgraph reachable from any node forms a tree[1]), can be reasoned with efficiently by applying one of the well known inference algorithms, *Belief Propagation* [Pearl, 1989] or *Bucket Elimination* [Dechter, 1999]. In the full version of this paper we will prove Theorem 2 that considers *linear partial orders*.

Theorem 2. *Consider a Mallows model over a set A of m elements, with a reference ranking $\sigma = \sigma_1, \ldots, \sigma_m$, and dispersion parameter ϕ. Let $\xi = \{\sigma_{i_1} \succ \sigma_{i_2}, \sigma_{i_3} \succ \sigma_{i_4}, \ldots, \sigma_{i_{n-1}} \succ \sigma_{i_n}\}$. The probability of ξ can be computed in polynomial time if the item indexes can be arranged in either:*

1. *Ascending order:*

$$i_1 < i_2 \leq i_3 < i_4 \cdots i_{n-3} < i_{n-2} \leq i_{n-1} < i_n$$

2. *or, descending order:*

$$i_1 > i_2 \geq i_3 > i_4 \cdots i_{n-3} > i_{n-2} \geq i_{n-1} > i_n$$

The example that follows shows how to represent and reason about evidence in the form of a linear series.

Example 2. *We consider a Mallows model over m items, $\sigma = \sigma_1, \ldots, \sigma_m$ and dispersion parameter ϕ. Let $\xi = \{\sigma_{i_k} \succ \sigma_{i_{k-1}} \succ \sigma_{i_{k-2}} \succ \cdots \succ \sigma_{i_1}\}$ where $i_k > i_{k-1} > \cdots > i_1$. A RIM-generated ranking, r, can abide to this constraint only if for every consecutive pair of indexes $\{i_{j-1}, i_j\}$ where $j \in [2, k]$, we have that $X_{i_j}^{i_j} \leq X_{i_{j-1}}^{i_j - 1}$. For example, if $\xi = \{\sigma_8 \succ \sigma_4 \succ \sigma_1\}$ then we have that $X_4^4 \leq X_1^3$ and $X_8^8 \leq X_4^7$.*

[1] https://en.wikipedia.org/wiki/Polytree

The partial order and corresponding BN are presented in Figures 2e and 2f, respectively. The BN in Figure 2f contains two types of CPTs. The first relates between RVs that represent the item's position at different times, (i.e., X_i^t and $X_i^{t'}$). The size of these factors is limited to $O(m^2)$, and each entry can be computed in polynomial time using the dynamic programming algorithm of Equation (6)*. The second type of CPT enforces the constraint that item σ_{i_j} will be inserted into a position that is not larger than the position of $\sigma_{i_{j-1}}$ (i.e., $X_{i_j}^{i_j} \leq X_{i_{j-1}}^{i_j - 1}$). The size of this factor is also in $O(m^2)$.*

It is easily seen that the BN of Figure 2f has $2k$ factors, each of polynomial size. This, combined with the fact that this BN is a linear path, enables performing inference in polynomial time.

To the best of our knowledge, ours is the first approach which enables computing the probability of linear partial orders exactly and efficiently. For example, the approximation algorithm `AMP` [Lu and Boutilier, 2014] was shown to preform arbitrarily bad on such partial orders, with an approximation error of $O(m)$, where m is the cardinality of σ, the reference ranking.

Theorem 2 illustrates the potential of the proposed approach in representing and reasoning about the Mallows posterior distribution efficiently. An accurate representation of the posterior enables answering queries over this distribution, and sampling from it. In cases where the partial order induces a Bayesian network with a width too large to be handled exactly, we can resort to approximate reasoning techniques [Mateescu *et al.*, 2010; Gogate and Dechter, 2011] to obtain an approximate representation of the posterior.

5 Conclusions

In this paper we analyzed the RIM process that provides an effective sampling procedure over Mallows distributions. We identified the conditional independences in the RIM process, and applied them to the generation of a Bayesian network that represents RIM with partial order constraints. We presented a dynamic programming algorithm that enables calculating the position of an item at a certain time, and applied the algorithm to the construction the network's CPTs. Finally, we presented a new fragment of partial orders that admits efficient inference.

In an ongoing research we pursue the utilization of our modeling by deploying the wealth of research, algorithms and software for inference over Bayesian networks. With that, our hope is to substantially advance the state of the art on inference over Mallows distributions.

Acknowledgments

The authors are very grateful to Lovro Ilijasic, Haoyue Ping and Julia Stoyanovich for helpful discussions. This research is supported by the US-Israel Binational Science Foundation, Grant #2014391, the National Science Foundation, Grant #1539856, and the Israeli Science Foundation, Grants #1295/15 and #1308/15. Benny Kimelfeld is a Taub Fellow, supported by the Taub Foundation.

References

[Awasthi *et al.*, 2014] Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning mixtures of ranking models. *CoRR*, abs/1410.8750, 2014.

[Boulesteix and Slawski, 2009] Anne-Laure Boulesteix and Martin Slawski. Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, 10(5):556–568, 2009.

[Darwiche, 2010] Adnan Darwiche. Bayesian networks. *Commun. ACM*, 53(12):80–90, 2010.

[Das Sarma *et al.*, 2010] Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Ranking mechanisms in twitter-like forums. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 21–30, New York, NY, USA, 2010. ACM.

[Dechter, 1999] Rina Dechter. Bucket elimination: A unifying framework for reasoning. *Artif. Intell.*, 113(1-2):41–85, 1999.

[Doignon *et al.*, 2004] Jean-Paul Doignon, Aleksandar Peke, and Michel Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.

[Farasat *et al.*, 2015] Alireza Farasat, Alexander Nikolaev, Sargur N. Srihari, and Rachael Hageman Blair. Probabilistic graphical models in modern social network analysis. *Social Network Analysis and Mining*, 5(1):1–18, 2015.

[Gogate and Dechter, 2011] Vibhav Gogate and Rina Dechter. Samplesearch: Importance sampling in presence of determinism. *Artificial Intelligence*, 175(2):694 – 729, 2011.

[Gormley and Murphy, 2008] Isobel Claire Gormley and Thomas Brendan Murphy. A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.*, 2(4):1452–1477, 12 2008.

[Heckerman, 1989] David Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In Max Henrion, Ross D. Shachter, Laveen N. Kanal, and John F. Lemmer, editors, *UAI*, pages 163–172. North-Holland, 1989.

[Hofmann, 2001] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.

[Huang *et al.*, 2012] Jonathan Huang, Ashish Kapoor, and Carlos Guestrin. Riffled independence for efficient inference with partial rankings. *J. Artif. Intell. Res. (JAIR)*, 44:491–532, 2012.

[Kendall, 1938] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[Kolde *et al.*, 2012] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, February 2012.

[Lebanon and Mao, 2008] G. Lebanon and Y. Mao. Non-Parametric Modeling of Partially Ranked Data. *Journal of Machine Learning Research*, 9:2401–2429, 2008.

[Lu and Boutilier, 2014] Tyler Lu and Craig Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *Journal of Machine Learning Research*, 15(1):3783–3829, 2014.

[Mallows, 1957] C. L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1-2):114–130, June 1957.

[Marden, 1995] John I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

[Mateescu *et al.*, 2010] Robert Mateescu, Kalev Kask, Vibhav Gogate, and Rina Dechter. Join-graph propagation algorithms. *Journal of Artificial Intelligence Research*, 37:279–328, 1 2010.

[McElroy and Marsh, 2009] Gail McElroy and Michael Marsh. Candidate gender and voter choice: Analysis from a multimember preferential voting system. *Political Research Quarterly*, 2009.

[Pavlović *et al.*, 1999] Vladimir Pavlović, James M Rehg, Tat-Jen Cham, and Kevin P Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 94–101. IEEE, 1999.

[Pearl, 1989] Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

[Stoyanovich *et al.*, 2016] Julia Stoyanovich, Lovro Ilijasic, and Haoyue Ping. Workload-driven learning of mallows mixtures with pairwise preference data. In *Proceedings of the 19th International Workshop on Web and Databases*, WebDB'16. ACM, 2016.

[Zweig and Russell, 1998] Geoffrey Zweig and Stuart J. Russell. Speech recognition with dynamic bayesian networks. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA.*, pages 173–180, 1998.