# A Database Framework for Probabilistic Preferences

Batya Kenig[1], Benny Kimelfeld[1⋆], Haoyue Ping[2], and Julia Stoyanovich[2⋆⋆]

[1] Technion, Israel
[2] Drexel University, USA

## 1 Introduction

Preferences are statements about the relative quality or desirability of items. Ever larger amounts of preference information are being collected and analyzed in a variety of domains, including recommendation systems [2, 16, 18], polling and election analysis [3, 6, 7, 15], and bioinformatics [1, 11, 19].

Preferences are often inferred from indirect input (e.g., a ranked list may be inferred from individual choices), and are therefore uncertain in nature. This motivates a rich body of work on uncertain preference models in the statistics literature [14]. More recently, the machine learning community has been developing methods for effective modeling and efficient inference over preferences, with the Mallows model [13] receiving particular attention [4, 5, 12, 17].

In this paper, we take the position that preference modeling and analysis should be accommodated within a general-purpose probabilistic database framework. Our framework is based on a deterministic concept that we proposed in a past vision paper [8]. In the present work we focus on handing uncertain preferences, and develop a representation of preferences within a *probabilistic preference database*, or *PPD* for short.

This paper is an abbreviated version of our PODS 2017 paper, where an interested reader can find additional details about the formalism and proposed algorithmic solutions.

## 2 Probabilistic Preference Databases

A *preference schema* **S** is a relational schema with some relation symbols marked as *preference symbols* (and others as *ordinary symbols*). Figure 1 gives an example of a preference database instance, with the ordinary symbols **Candidates** and **Voters**, and the preference symbol **Polls**.

An instance over a preference symbol (such as **Polls** in Figure 1) represents a collection of *preferences* among a set of items, where each such preference is

⋆ This work was supported in part by ISF Grant No. 1295/15, BSF Grant No. 2014391 and by the Taub Foundation.
⋆⋆ This work was supported in part by NSF Grants No. 1464327 and 1539856, and BSF Grant No. 2014391.

**Candidates** (o)

| cand | party | sex | edu |
|---|---|---|---|
| Trump | R | M | BS |
| Clinton | D | F | JD |
| Sanders | D | M | BS |
| Rubio | R | M | JD |

**Voters** (o)

| voter | edu | sex | age |
|---|---|---|---|
| Ann | BS | F | 25 |
| Bob | BS | M | 35 |
| Cat | MS | F | 40 |
| Dave | MS | M | 45 |

**Polls** (p)

| voter | date | lcand | rcand |
|---|---|---|---|
| Ann | Oct-5 | Sanders | Clinton |
| Ann | Oct-5 | Sanders | Rubio |
| Ann | Oct-5 | Sanders | Trump |
| Ann | Oct-5 | Clinton | Rubio |
| Ann | Oct-5 | Clinton | Trump |
| Ann | Oct-5 | Rubio | Trump |
| Bob | Oct-5 | Sanders | Rubio |
| Bob | Oct-5 | Sanders | Clinton |
| Bob | Oct-5 | Sanders | Trump |
| Bob | Oct-5 | Rubio | Clinton |
| Bob | Oct-5 | Rubio | Trump |
| Bob | Oct-5 | Clinton | Trump |

A **MAL**-instance over **Polls** (p)

| voter | date | *Preference model* $\mathbf{MAL}(\boldsymbol{\sigma}, \phi)$ |
|---|---|---|
| Ann | Oct-5 | $\langle$Clinton, Sanders, Rubio, Trump$\rangle$, 0.3 |
| Bob | Oct-5 | $\langle$Trump, Rubio, Sanders, Clinton$\rangle$, 0.3 |

**Fig. 1.** An example of a preference database

itself a binary relation called a *session*. A a binary relation $\succ$ over a set $I = \{\sigma_1, \dots, \sigma_n\}$ of *items* is a (strict) *partial order* if it is irreflexive and transitive. A *linear* (or *total*) order is a partial order where every two items are comparable. By a slight abuse of notation, we often identify a linear order $\sigma_1 \succ \cdots \succ \sigma_n$ with the sequence $\langle \sigma_1, \dots, \sigma_n \rangle$, and we call it a *ranking*.

*Example 1.* Our running example is on individual preferences among the set of US presidential candidates $I = \{\texttt{Clinton}, \texttt{Rubio}, \texttt{Sanders}, \texttt{Trump}\}$. The ranking $\boldsymbol{\tau} = \langle \texttt{Clinton}, \texttt{Rubio}, \texttt{Sanders}, \texttt{Trump} \rangle$ is an example ranking over $I$. □

A preference relation instantiates a special relation symbol with a signature of the form $(\beta, A_l, A_r)$, where $\beta$ is the *session signature*, and $A_l$ and $A_r$ are the *left-hand-side* (*lhs*) attribute and *right-hand-side* (*rhs*) attribute, respectively. We use semicolon (;) to distinguish between the different parts and write $(\beta; A_l; A_r)$.

*Example 2.* We use the preference signature (voter, date; lcand; rcand) in our running example. Here the components $\beta$, $A_l$ and $A_r$ are (voter, date), lcand, and rcand, respectively. The table **Polls** of Figure 1 is an instance of this preference signature that contains two sessions. The session (Ann, Oct-5) is associated with the ranking $\langle$Sanders, Clinton, Rubio, Trump$\rangle$. The tuple (Ann, Oct-5; Sanders; Clinton) denotes that in the session of the voter Ann on October 5th, the candidate Sanders is preferred to the candidate Clinton. □

We now make the knowledge about voters' opinions probabilistic, interpreting the preference database of Figure 1 as one possible world of a probabilistic preference database. A *probabilistic preference database* (abbrv. *PPD*) over the preference schema $\mathbf{S}$ is a probability space over preference databases over $\mathbf{S}$. A PPD can be represented by explicitly specifying the entire sample space; however, we wish to allow for standard compact representations of preferences.

A *probabilistic preference model* is a (finite and typically compact) representation $M$ of a probability space over partial orders $\succ$ over a finite set of items; we denote this probability space by $[\![M]\!]$. A *model family* is a collection $\mathcal{M}$ of probabilistic preference models. As prominent examples, we define two model families: **RIM** is the family of RIM [5] models $\mathsf{RIM}(\boldsymbol{\sigma}, \Pi)$, and **MAL** is the family of Mallows [13] models $\mathsf{MAL}(\boldsymbol{\sigma}, \phi)$.

A *Mallows* model [13] $\mathsf{MAL}(\boldsymbol{\sigma}, \phi)$ is parameterized by a reference ranking $\boldsymbol{\sigma} = \langle \sigma_1, \dots, \sigma_m \rangle$ and a dispersion parameter $\phi \in (0, 1]$. The model assigns a non-zero probability to every ranking $\boldsymbol{\tau}$: The higher the Kendal's tau distance [9] of $\boldsymbol{\tau}$ is from $\boldsymbol{\sigma}$, the lower its probability under the model. Lower values of $\phi$ concentrate most of the probability mass around $\boldsymbol{\sigma}$, while $\phi = 1$ corresponds to the uniform probability distribution over the rankings. Doignon [5] showed that $\mathsf{MAL}(\boldsymbol{\sigma}, \phi)$ can be represented as the insertion model $\mathsf{RIM}(\boldsymbol{\sigma}, \Pi)$.

In the PPD representations we explore, termed **RIM**-PPD, each session is associated with the parameters of a RIM model. A **RIM**-PPD represents a probability space over preference databases, where a possible world is obtained by independently sampling a preference from the model of each session. Figure 1 gives an example of a **MAL**-instance over the p-symbol **Polls** that associates each session in **Polls** with a Mallows model. It is straight-forward to extend this representation to a mixture of Mallows, by associating each session with $k$ components $\mathsf{MAL}_1(\boldsymbol{\sigma}_1, \phi_1), \dots, \mathsf{MAL}_k(\boldsymbol{\sigma}_k, \phi_k)$, with the corresponding probabilities $p_1, \dots, p_k$. A possible world would then be obtained by first sampling component $\mathsf{MAL}_i(\boldsymbol{\sigma}_i, \phi_i)$ with probability $p_i$ independently for each session, and then sampling a preference from $\mathsf{MAL}_i(\boldsymbol{\sigma}_i, \phi_i)$.

## 3 Query Evaluation over PPDs

We adopt the semantics of probabilistic databases [20] for query evaluation. Specifically, let $\mathbf{S}$ be a schema, let $Q$ be a query, and let $\mathcal{D} = (\Omega, \pi)$ be a PPD. A *possible answer* for $Q$ is a tuple $\mathbf{a}$ over $sig(Q)$ such that $\mathbf{a} \in Q(D)$ for some sample $D$ of $\mathcal{D}$. We denote by $PosAns(Q, \mathcal{D})$ the set of all possible answers. The *confidence* of a possible answer $\mathbf{a} \in PosAns(Q, \mathcal{D})$, denoted $conf_Q(\mathcal{D}, \mathbf{a})$, is the probability of having $\mathbf{a}$ as an answer when querying a sample of $\mathcal{D}$. If $E$ is an $\mathcal{M}$-PPD for some model class $\mathcal{M}$, then evaluating $Q$ on $E$ is the task of computing the following (finite) set: $Q(E) = \{(\mathbf{a}, conf_Q([\![E]\!], \mathbf{a})) \mid \mathbf{a} \in PosAns([\![E]\!])\}$.

We study the data complexity of evaluating Conjunctive Queries (CQs) over **RIM**-PPDs. We focus on CQs to which we refer as *itemwise*. Intuitively, these are CQs where items are connected only through preferences. We show a natural fragment of CQs where the itemwise CQs are *precisely* the CQs in which query evaluation can be done in polynomial time. In the fragment we consider, we prove that every query that is *not* itemwise is actually #P-hard, and therefore, we establish a dichotomy in complexity.

Let $\mathbf{S}$ be a preference schema, and let $Q$ be a CQ over $\mathbf{S}$. An atomic formula of $Q$ is called a *p-atom* if it is over a p-symbol, and an *o-atom* if it is over an o-symbol. Let $P(s_1, \dots, s_k; t_l; t_r)$ be p-atom of $Q$. Each term $s_i$ for $i = 1, \dots, k$ is said to occur in a *session position*, and each of $t_l$ and $t_r$ is said to occur in an *item position*. A *session variable* of $Q$ is a variable that occurs in a session position, and an *item variable* of $Q$ is a variable that occurs in an item position.

We say that $Q$ is *sessionwise* if all p-atoms of $Q$ refer to the same session; that is, if $P(s_1, \dots, s_k; t_l; t_r)$ and $P'(s'_1, \dots, s'_l; t'_l; t'_r)$ are p-atoms of $Q$, then $P = P'$ and $(s_1, \dots, s_k) = (s'_1, \dots, s'_l)$. We say that $Q$ is *itemwise* if $Q$ is sessionwise, and the joins between item variables occur only inside the p-atoms, or through session variables. Put differently, in an itemwise CQ with a constant session, the o-atoms state properties of individual items but do not draw connections between the items. In [10] we define this property more formally, by means of the *Gaifman graph* of the CQ.

*Example 3.* Consider the following Boolean CQs. The query $Q_1$ asks whether there is a voter with a BS degree who prefers a male Democratic candidate to a female Democratic candidate.

$$Q_1() \leftarrow P(v, \_; l; r), V(v, \mathtt{BS}, \_, \_), C(l, \mathtt{D}, \mathtt{M}, \_), C(r, \mathtt{D}, \mathtt{F}, \_)$$

The query $Q_2$ asks whether there is a voter who prefers a male candidate to a female candidate such that both candidates are of the same political party.

$$Q_2() \leftarrow P(\_, \_; l; r), C(l, p, \mathtt{M}, \_), C(r, p, \mathtt{F}, \_)$$

The query $Q_3$ asks whether there is a voter who prefers a female candidate to both Trump and Sanders.

$$Q_3() \leftarrow P(v, d; l; \mathtt{Trump}), P(v, d; l; \mathtt{Sanders}), C(l, \_, \mathtt{F}, \_)$$

All of these CQs are sessionwise. Indeed, $Q_1$ and $Q_2$ involve a single p-atom (hence, they are sessionwise by definition), and in $Q_3$ both atoms have $(v, d)$ in their session parts. CQs $Q_1$ and $Q_3$ are itemwise, while $Q_2$ is *not* itemwise. □

In [10] we prove the following theorem, which states that every itemwise Boolean CQ can be evaluated in polynomial time, under data complexity.

**Theorem 1.** *Let* $\mathbf{S}$ *be a preference schema, and let* $Q$ *be a Boolean CQ over* $\mathbf{S}$*. If* $Q$ *is itemwise, then* $Q$ *can be evaluated in polynomial time on* **RIM**-*PPDs.*

We also prove that the class itemwise CQs are *precisely* the tractable ones (among the queries in the class), under conventional complexity assumptions. In other words, every Boolean CQ (in the class) that is not itemwise is necessarily hard to evaluate, and therefore, we establish a dichotomy.

**Theorem 2.** *Let* $\mathbf{S}$ *be a preference schema, and let* $Q$ *be a Boolean CQ over* $\mathbf{S}$ *such that* $Q$ *has no self joins and* $Q$ *has a single p-atom. If* $Q$ *is not itemwise, then the evaluation of* $Q$ *on* **RIM**-*PPDs over* $\mathbf{S}$ *is* $\mathrm{FP}^{\#\mathrm{P}}$-*hard.*

In [10] we give is a polynomial-time algorithm for evaluating itemwise CQs. Interestingly, such CQs translate into a natural (and novel) inference problem over RIM. In this problem, every item is associated with one or more labels (e.g., "democratic" party or "comedy" genre), and the goal is to compute the probability that a graph pattern (or equivalently a partial order) over these labels *matches* the random ranking.

## References


1. S. Aerts, D. Lambrechts, S. Maity, P. V. Loo, B. Coessens, F. D. Smet, L.-C. Tranchevent, B. D. Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, 2006.
2. S. Balakrishnan and S. Chopra. Two of a kind or the ratings game? adaptive pairwise preferences and latent factor models. *Frontiers of Computer Science*, 6(2):197–208, 2012.
3. P. Diaconis. A generalization of spectral analysis with applications to ranked data. *Annals of Statistics*, 17(3):949–979, 1989.
4. W. Ding, P. Ishwar, and V. Saligrama. Learning mixed membership mallows models from pairwise comparisons. *CoRR*, abs/1504.00757, 2015.
5. J.-P. Doignon, A. Pekeč, and M. Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.
6. I. C. Gormley and T. B. Murphy. A latent space model for rank data. In *ICML*, 2006.
7. I. C. Gormley and T. B. Murphy. A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, 2(4):1452–1477, 12 2008.
8. M. Jacob, B. Kimelfeld, and J. Stoyanovich. A system for management and analysis of preference data. *PVLDB*, 7(12):1255–1258, 2014.
9. M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
10. B. Kenig, B. Kimelfeld, H. Ping, and J. Stoyanovich. Querying probabilistic preferences in databases. In *PODS*, 2017.
11. R. Kolde, S. Laur, P. Adler, and J. Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.
12. T. Lu and C. Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *J. Mach. Learn. Res.*, 15(1):3783–3829, Jan. 2014.
13. C. L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1-2):114–130, June 1957.
14. J. I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.
15. G. McElroy and M. Marsh. Candidate gender and voter choice: Analysis from a multimember preferential voting system. *Political Research Quarterly*, 63(4):pp. 822–833, 2010.
16. A. D. Sarma, A. D. Sarma, S. Gollapudi, and R. Panigrahy. Ranking mechanisms in twitter-like forums. In *WSDM*, pages 21–30, 2010.
17. J. Stoyanovich, L. Ilijasic, and H. Ping. Workload-driven learning of mallows mixtures with pairwise preference data. In *WebDB*, page 8, 2016.
18. J. Stoyanovich, M. Jacob, and X. Gong. Analyzing crowd rankings. In *WebDB*, pages 41–47, 2015.
19. J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.
20. D. Suciu, D. Olteanu, C. Ré, and C. Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.