Selective Preference Aggregation

Shreyas Kadekodi [*1] **Hayden McTavish** [*2] **Berk Ustun** [1]

Abstract

Many applications in machine learning and decision-making rely on procedures to aggregate human preferences. In such tasks, individuals express ordinal preferences over a set of items through votes, ratings, or pairwise comparisons. We then summarize their collective preferences as a ranking. Standard methods for preference aggregation are designed to return rankings that arbitrate individual disagreements in ways that are faithful and fair. In this work, we introduce a paradigm for *selective aggregation*, where we can avoid the need to arbitrate dissent by abstaining from comparison. We summarize collective preferences as a *selective ranking* – i.e., a partial order where we can only compare items where at least $100 \cdot (1 - \tau)\%$ of individuals agree. We develop algorithms to build selective rankings that achieve all possible trade-offs between comparability and disagreement, and derive formal guarantees on their safety and stability. We conduct an extensive set of experiments on real-world datasets to benchmark our approach and demonstrate its functionality. Our results show selective aggregation can promote transparency and robustness by revealing disagreement and abstaining from arbitration.

1 Introduction

Many of our most important systems rely on procedures where we elicit and aggregate human preferences. In such systems, we ask a group of individuals to express ordinal preferences over a set of items through votes, ratings, or pairwise comparisons. We then use these data to order items in a way that reflects the collective preferences. Over the past century, we have applied this pattern to reap transformative benefits from collective intelligence – in elections [20], search [26], and alignment [21].

*Equal contribution [1]UCSD [2]Duke University. Correspondence to: Berk Ustun <berk@ucsd.edu>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Standard methods for preference aggregation express collective preferences as a *ranking* – i.e., a total order over n items where we can determine the collective preference between items by comparing their positions. Rankings reflect an *approximate* summary of collective preferences. This is because it is impossible to define a coherent order when individuals disagree. This impossibility, which is enshrined in foundational results such as Condorcet's Paradox [20] and Arrow's Impossibility Theorem [9], has framed preference aggregation as an exercise in *arbitration*. "*In tasks where individuals disagree, how can we summarize their collective preferences in a way that is faithful and fair?*"

Over the past few decades, we have developed countless algorithms from this perspective [see 4, 75] to reap benefits from collective intelligence in new use cases:

- Supporting Group Decisions – e.g., to fund grant proposals or hire employees [16, 76],
- Learning Preferences – e.g., to learn consumer preferences over products [17] or content [21].
- Communicating Consensus – e.g., to rank colleges [19] or benchmark language models [61].

In many of these new use cases, however, we do not *need* a total order. When we aggregate preferences to fund grant proposals, a total order can lead to worse decisions as we arbitrarily select the top k items on the list. When we aggregate preferences to rank colleges, a total order can strongly influence where students apply and how institutions invest [see e.g., 43, 28, 27, 68]. When we aggregate preferences to predict helpfulness [25], a total order can lead us to overlook minority views by silently enshrining the views of a slim majority.

In this work, we propose to address these challenges through *selective aggregation*. In this paradigm, we express collective preferences as a *tiered ranking* – i.e., a partial order where we are only allowed to compare items in different tiers. We view tiers as a simple solution to avoid the impossibility of arbitration: given a pair of items where individuals express conflicting preferences, we can place them in the same tier to abstain from comparison. We capitalize on this structure to develop new representation for collective preferences that can reveal disagreement, and new algorithms that can allow us to control it.

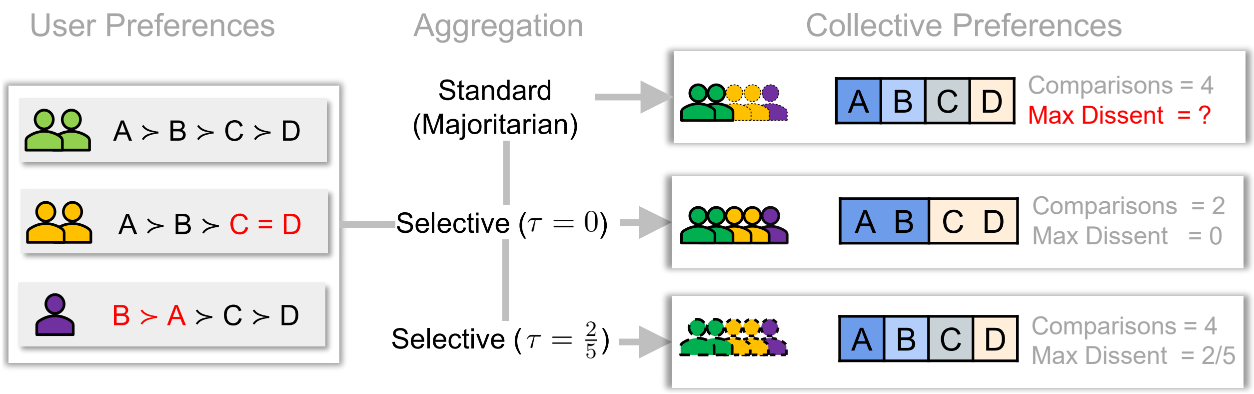


Fig. 1. Comparison of collective preferences for a task where $p = 5$ users express their ordinal preferences over $n = 4$ items as a ranking with ties. Standard methods represent the collective preferences of all users as a ranking – i.e., a total order over n items. The resulting structure orders items in a way that minimizes disagreement but does not reveal its existence or severity. In comparison, selective aggregation represent the collective preferences as a selective ranking – i.e., a partial order with $m \leq n$ tiers where we can only compare items in different tiers, and guaranteed that any such comparison will overrule at most $100\tau\%$ of users. The resulting structure reveals disagreement through its tiers and dissent parameter $\tau \in [0, 0.5)$. Setting $\tau = 0$ reveals that all users unanimously prefer $\{A, B\}$ to $\{C, D\}$. Setting $\tau = \frac{2}{5}$ shows that we can recover a total order when we are willing to overrule 40% of users.

Our main contributions include:

1. We introduce a paradigm for selective preference aggregation in which we summarize collective preferences as a *selective ranking* – i.e., a tiered ranking where each comparison will align with the collective preferences of at least $100(1 - \tau)\%$ of users.
2. We develop algorithms to construct all possible selective rankings for a preference aggregation task. Our algorithms are fast, easy to implement, and guaranteed to behave in ways that are safe and predictable.
3. We conduct a comprehensive empirical study of preferences aggregation in modern use cases with diverse preference data. Our results show how selective rankings can improve transparency and robustness compared to existing approaches.
4. We demonstrate how selective aggregation can learn from subjective annotations through a case study in toxicity detection. Our results show our machinery can improve performance and align predictions with a plurality of users.
5. We provide a Python library for selective preference aggregation on GitHub.

Related Work

Our work is motivated by a growing set of applications where we aggregate conflicting preferences. In machine learning, such issues arise in tasks such as data annotation [7, 53, 29] and alignment [62, 21, 24] as a result of ambiguity, subjectivity, or lack of expertise [57, 62, 77]. In medicine, for example, conflicting annotations reflect uncertainty regarding ground truth [see e.g., 71, 65, 55, 51]. In content moderation, conflicting annotations reflect differences in opinion [39, 34].

Our work is related to an extensive stream of work in social choice [45]. This body of work develops mathematical foundations for preference aggregation by defining salient voting rules and characterizing their properties [see e.g., 14, 49, for a list]. Although much of the effort is driven by the impossibility of reconciling individual preferences [see e.g., 9, 58], few works mention that we could abstain from arbitration by representing collective preferences as a partial order. In effect, abstention is not a viable option in many of the applications that motivate work in this area. In voting, for example, aggregating ballots into a partial order can lead to elections that fail to identify a single winner [50].

On a technical front, our work is related to a stream of research on rank aggregation [13, 26, 36, 3]. Although most work focuses on rankings that represent collective preferences as a total order, some focus on coarser representations such as bucket orderings [see e.g., 2, 6, 33, and references therein]. For example, Achab et al. [2] view bucket orderings as a low-dimensional total order and characterize their potential for recovery. Andrieu et al. [6] view bucket orderings as a vehicle to efficiently combine multiple rankings. In general, these differences in motivation lead to differences in algorithm design and interpretation. For example, items that we would consider "equivalent" in a bucket ordering would be "incomparable" in a tiered ranking.

2 Framework

We consider a standard preference aggregation task where we wish to order n items in a way that reflects the collective preferences of p users. We start with a dataset where each instance $\pi^k_{i,j}$ represents the pairwise preference of a user $k \in [p] := \{1, \dots, p\}$ between a pair of items $i, j \in [n]$:

$$\pi^k_{i,j} = \begin{cases} 1 & \text{if user } k \text{ strictly prefers } i \text{ to } j \Leftrightarrow i \overset{k}{\succ} j \\ 0 & \text{if user } k \text{ is indifferent} \quad \Leftrightarrow i \overset{k}{\sim} j \\ -1 & \text{if user } k \text{ strictly prefers } j \text{ to } i \Leftrightarrow i \overset{k}{\prec} j \end{cases}$$

Pairwise preferences can represent a wide range of ordinal preferences, including labels, ratings, and rankings. In practice, we can convert all of these formats to pairwise preferences as described in Appendix A.2. In doing so, we can avoid restrictive assumptions on elicitation. For example, users can state that items are equivalent by setting $\pi^k_{i,j} = 0$, or express preferences that are intransitive. In what follows, we assume that datasets contain all pairwise preferences from all users for clarity. We describe how to relax this assumption in Section 4, and work with datasets with missing preferences in Section 5.

Collective Preferences as Partial Orders Standard approaches express collective preferences as a *ranking* – i.e., a total order over n items where we can compare any pair of items. We consider an alternative approach in which we express collective preferences as a *tiered ranking*:

Definition 2.1. A *tiered ranking* T is a partial ordering of n items into m *tiers* $T := (T_1, \dots, T_m)$ such that $\cup_{l=1}^m T_l = [n]$ and $T_l \cap T_{l'} = \emptyset$ for all tiers $T_l, T_{l'} \in T$.

Tiers provide a way to *abstain from arbitration*. Given a pair of items where users disagree, we can place them in the same tier and "agree to disagree." Given a tiered ranking, we only make claims about collective preferences by comparing items in different tiers. Formally, we denote the collective preferences as:

$$\pi_{i,j}(T) := \begin{cases} 1 \text{ if } & i \in T_l, j \in T_{l'} \text{ for } l < l', \\ -1 \text{ if } & i \in T_l, j \in T_{l'} \text{ for } l > l', \\ \perp \text{ if } & i, j \in T_l \text{ for any } l \end{cases}$$

Given tiered ranking T, we say that a pairwise comparison between items i, j is *valid* if $\pi_{i,j}(T) \neq \perp$. We refer to a valid pairwise comparison as a *selective comparison*.

Selective Aggregation Given a dataset of pairwise preferences over n items from p users, a *selective ranking* S_τ is a partial order that maximizes the number of comparisons that align with the preferences of at least $1 - \tau\%$ of users.

We can express S_τ as the optimal solution to an optimization problem over the space of all tiered rankings $\mathbb{T}$:

$$\begin{aligned} \max_{T \in \mathbb{T}} \quad & \text{Comparisons}(T) \\ \text{s.t.} \quad & \text{Disagreements}(T) \leq \tau p \end{aligned} \qquad (\text{SPA}_\tau)$$

Here, the objective maximizes the number of valid comparisons in a tiered ranking T

$$\text{Comparisons}(T) := \sum_{i,j \in [n]} \mathbb{I}\left[\pi_{i,j}(T) \neq \perp\right]$$

The constraints restrict the fraction of individual preferences that can be contradicted by any valid comparison in T

$$\text{Disagreements}(T) := \max_{i,j \in [n]} \sum_{k \in [p]} \mathbb{I}\left[\pi_{i,j}(T) = 1, \pi^k_{i,j} \neq 1\right]$$

The *dissent parameter* τ limits the fraction of individual preferences that can be violated by any selective comparison. Given a selective ranking S_τ that places item i in a tier above item j, at most $100 \cdot \tau\%$ of users may have stated $i \not\succ j$.

We restrict $\tau \in [0, 0.5)$ to guarantee that selective ranking S_τ aligns with a majority of users, and is unique (see Appendix B for a proof). In this regime, we can set τ to trade off coverage for alignment as shown in Fig. 2. Setting $\tau = 0$ returns a selective ranking that reflects unanimity by showing all comparisons on which all users agree. Setting τ just shy of 0.5 reflects a selective ranking that maximizes tiers without overruling a majority of users. The trade-off is analogous to the trade-off in selective classification [32, 30, 40]: we output a partial order (selective classifier) that sacrifices "comparisons" (coverage) to reduce "disagreement" (error).

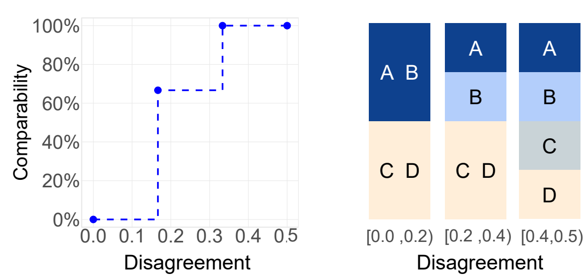


Fig. 2. All possible selective rankings for the task in Fig. 1 where we aggregate the preferences of $p = 5$ users over $n = 4$ items $\{A, B, C, D\}$. We show the comparability and disagreement of each solution to SPA_τ on the left, and their selective rankings on the right. Here, the solution for $\tau \in [0, \frac{1}{5}]$ reveals that all users unanimously prefer $\{A, B\}$ to $\{C, D\}$. The solution for $\tau \in (\frac{1}{5}, \frac{2}{5}]$, reveals that we can recover a single winner if we are willing to make claims that overrule at most 1 user, while the solution for $\tau \in (\frac{2}{5}, \frac{1}{2}]$ reveals we can only recover a total order if we are willing to overrule at most 2 users.

3 Algorithms

We present an algorithm to construct selective rankings in Algorithm 1 and depict its behavior in Fig. 4.

Algorithm 1 Selective Preference Aggregation

Input: $\{\pi^k_{i,j}\}_{i,j\in[n],k\in[p]}$ *preference dataset*
Input: $\tau \in [0, 0.5)$ *dissent parameter*
1: $w_{i,j} \leftarrow \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$ for all $i, j \in [n]$
2: $V_I \leftarrow [n]$
3: $A_I \leftarrow \{(i \to j) \mid w_{i,j} \geq \tau p\}$
4: $V_T \leftarrow \mathsf{ConnectedComponents}(V_I, A_I)$
5: $A_T \leftarrow \{(T \to T') \mid \exists i \in T, j \in T' : (i \to j) \in A_I\}$
6: $l_1, \ldots, l_{|T|} \leftarrow \mathsf{TopologicalSort}(V_T, A_T)$
Output: $S_\tau \leftarrow (T_{l_1}, T_{l_2}, \ldots, T_{l_{|T|}})$ *τ-selective ranking*

Algorithm 1 constructs a selective ranking from a dataset of pairwise preferences and a dissent parameter $\tau \in [0, 0.5)$. The procedure first builds a directed graph over items (V_I, A_I). Here, each vertex corresponds to an item, and each arc corresponds to a collective preference that we must not contradict in a tiered ranking. Given (V_I, A_I), the procedure then builds a directed graph over tiers (V_T, A_T). In Line 4, it calls the ConnectedComponents routine to identify the strongly connected components of (V_I, A_I) which become the set of *supervertices* $V_T = \{T_1, \ldots, T_{|V_T|}\}$, where each supervertex contains items in the same tier. In Line 5, it defines arcs between tiers – drawing an arc from T to T' whose respective elements are connected by an arc in A_I. Given (V_T, A_T), the procedure determines an ordering among tiers by calling the TopologicalSort routine in Line 6. In this case, the graph will admit a topological sort as it is a directed acyclic graph.

Correctness We show that Algorithm 1 recovers the unique optimal solution to SPA_τ in Theorem B.2. The result follows from the fact that the directed graph (V_T, A_T) defines a tiered ranking that is both feasible and optimal with respect to SPA_τ. Specifically, the tiered ranking must obey the disagreement constraint in SPA_τ because we only draw arcs for pairs of items where at least τp users disagree in Line 3. The tiered ranking maximizes the objective of SPA_τ because the ConnectedComponents routine in Line 4 partitions vertices in a way that maximizes the number of tiers, which subsequently maximizes the selective comparisons under the disagreement constraint.

Recovering All Selective Rankings Algorithm 1 is meant to recover a selective ranking in settings where we can set the value of τ a priori (e.g., $\tau = 0\%$ to enforce unanimity). In many applications, we may wish to set τ after seeing the entire path of selective rankings. In a hiring task where we only have the resources to hire 3 candidates, for example, we can choose the smallest value of τ from the solution path such that the top tier contains ≤ 3 candidates. In cases where a top three does not exist, this can lead us to hire fewer candidates or save resources. In a prediction task where labels encode collective preferences, we could aggregate annotations with a selective ranking and treat τ as a hyperparameter to control overfitting.

In these situations, we can produce a *solution path* of selective rankings– i.e., a finite set of selective rankings that covers all possible solutions to SPA_τ for $\tau \in [0, \frac{1}{2}]$ [see e.g., 66]. We observe that a finite solution path must exist as each selective ranking is specified by the arcs in Line 3. In practice, we can compute all selective rankings efficiently by: (1) identifying a smaller subset of dissent parameters to consider as per Proposition 3.1; and (2) re-using the graph of strongly connected components across iterations.

Proposition 3.1. *Given a dataset of pairwise preferences $\mathcal{D}$, let $\mathcal{S}_\mathcal{W}$ denote a finite set of selective rankings for dissent parameters in the set:*

$$\mathcal{W} = \left\{ \tfrac{w}{p} \leq \tfrac{1}{2} \mid w = \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right] \textit{for } i, j \in [n] \right\} \cup \{0\}$$

Let S_τ be a selective ranking for an arbitrary value of $\tau \in [0, \frac{1}{2})$. Then, $\mathcal{S}_\mathcal{W}$ contains a selective ranking $S_{\tau'}$ such that $S_{\tau'} = S_\tau$ for some dissent value $\tau' \leq \tau$.

We describe this procedure in Algorithm 2. Both Algorithms 1 and 2 run in time $\mathcal{O}(n^2 p)$ – i.e., they are linear in the number of individual pairwise preferences elicited (see Appendix B.4). As we show in Fig. 3, however, the resulting approach can lead to an improvement in runtime.

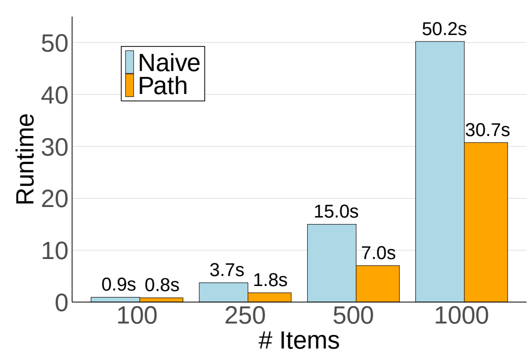


Fig. 3. Runtimes to produce all selective rankings for a synthetic task with $p = 10$ users and n items described in Appendix B. We show results for a naïve approach where we call Algorithm 1 for all possible dissent values, and a solution path algorithm in Appendix B. All results reflect timings on a consumer-grade laptop with 2.3 GhZ and 16 GB of RAM.

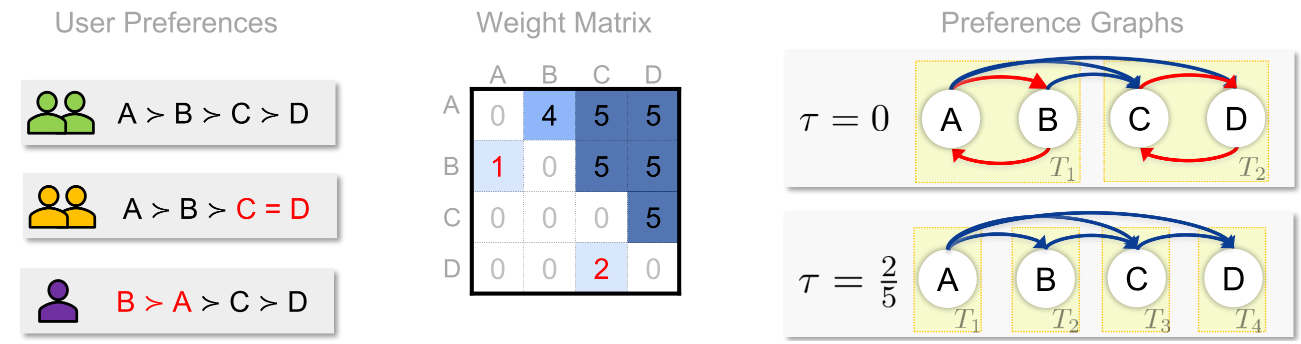


Fig. 4. Graphical representations used to construct selective rankings for the preference aggregation task in Fig. 1. Given a dataset with 5 users and 4 items, we compute a set of weights $w_{i,j} = \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$ in Line 1. Given a dissent parameter τ, we first build a directed graph (V_I, A_I) over items by drawing arcs between aggregate preferences with weight $w_{i,j} \geq \tau p$. We then condense (V_I, A_I) into a directed acyclic graph of supervertices (V_T, A_T) by identifying its strongly connected components (shown in yellow). Here, the selective rankings for $\tau = 0$ and $\tau = \frac{2}{5}$ have 2 and 4 tiers, respectively.

4 Theoretical Guarantees

In this section, we present formal guarantees on the stability and recovery guarantees of selective rankings.

On the Recovery of Condorcet Winners and Smith Sets One of the primary use cases for preference aggregation is to identify items that are collectively preferred to all others. Consider, for example, an application where we aggregate preferences to choose the most valuable player in a sports league or a subset of "top" grant proposals to fund [11]. In Theorem 4.1, we show that we can identify these items from a solution path of selective rankings.

Theorem 4.1. *Consider a preference aggregation task where a majority of users prefer item i_0 to all other items. There exists a threshold value $\tau_0 \in [0, 0.5)$ such that, for every $\tau > \tau_0$, every selective ranking S_τ will place i_0 as the sole item in its top tier.*

Theorem 4.1 provides a formal recovery guarantee that ensures we recover a Condorcet winner or a Smith set [see e.g., 63] when they exist. In practice, the result implies that we can identify such "top items" by constructing and inspecting a solution path of selective rankings.

In tasks where a majority of users prefers an item to all others, the solution path will contain a selective ranking whose top tier consists of a single item. In this case, we can recover the "single winner" and report the threshold value τ_0 as a measure of consensus.

In tasks where such a majority does not exist, every selective ranking S_τ for $\tau \in [0, 0.5)$ will include at least two items in the top tier. In settings where we aggregate preferences to identify a "single winner," we can point to the solution path as evidence that no such winner exists and use it as a signal that further deliberation is required [see e.g., 56].

Stability with Respect to Missing Preferences Standard methods can output rankings that change dramatically once we elicit missing preferences [10, 35, 44]. In Proposition 4.2, we show that we can build a selective ranking that will abstains from unstable comparisons by setting missing preferences to $\pi^k_{i,j} = 0$.

Proposition 4.2. *Consider a preference aggregation task where we are given a dataset with missing preferences $\mathcal{D}^{\text{init}}$. Let $\mathcal{D}^{\text{true}} \supseteq \mathcal{D}^{\text{init}}$ be a complete dataset where we elicit missing preferences, and $\mathcal{D}^{\text{safe}} \supseteq \mathcal{D}^{\text{init}}$ be a complete dataset where we set missing preferences to $\pi^k_{i,j} = 0$. Given $\tau \in [0, \frac{1}{2})$, let S^{safe}_τ and S^{true}_τ denote selective rankings for $\mathcal{D}^{\text{safe}}$ and $\mathcal{D}^{\text{true}}$. Then for any selective comparison $\pi_{i,j}(S^{\text{safe}}_\tau) \in \{-1, 1\}$, we have:*

$$\pi_{i,j}(S^{\text{true}}_\tau) = \pi_{i,j}(S^{\text{safe}}_\tau).$$

This means a selective ranking S^{safe}_τ that we produce using the imputed dataset $\mathcal{D}^{\text{safe}}$ will only include comparisons that will hold on the full dataset.

Proposition 4.2 provides a simple way to ensure stability when working with datasets where we are missing preferences from certain users for certain items. In such cases, we can always build a S is "robust to missingness" in the sense that it will abstain from comparisons that may be invalidated once we elicit missing preferences.

Stability with Respect to New Items In Proposition 4.3, we characterize the stability of selective aggregation as we add a new item to our dataset.

Proposition 4.3. *Consider a task where we start with a dataset of all pairwise preferences from p users over n items, which we then update to include all pairwise preferences for a new $n + 1^{th}$ item. For any $\tau \in [0, \frac{1}{2})$, let S^n_τ and S^{n+1}_τ denote selective rankings over n items and $n + 1$ items,* *respectively. Then for any two items $i, j \in [n]$, we have:*

$$\pi_{i,j}(S_\tau^{n+1}) \in \{-1, 1\}, \pi_{i,j}(S_\tau^n) = \pi_{i,j}(S_\tau^{n+1})$$

The result shows that adding a new item to a selective ranking will either maintain each comparison or abstain. That is, adding a new item can only collapse items that were in different tiers into a single tier. However, it cannot lead items in the same tier to split. Nor can it lead items in different tiers to invert their ordering.

On Setting the Dissent Parameter We can draw on the result in Proposition 4.2 to set the dissent parameter to ensure that selective rankings admit comparisons that are robust to missing or noisy preferences. By treating missing preferences as abstentions, we can build selective rankings that will only admit claims would not be invalidated if we were to elicit missing preferences or correct noisy preferences. In a preference aggregation task where we are missing 5% of preferences, we can set $\tau \geq 0.05$ to ensure that a selective rankings will only support comparisons that would remain valid if we were to elicit missing preferences. In a task where where we elicit noisy preferences, we we can set $\tau \geq 0.05$ to ensure that a selective ranking will only support comparisons that would remain valid if we were to to correct noisy preferences.

5 Experiments

In this section, we present experiments comparing our approach to standard methods in social choice and machine learning. Our goals are twofold: (1) to discuss the properties and behavior of selective aggregation on real-world datasets from modern applications; and (2) to evaluate the stability of selective rankings with respect to missing preferences and adversarial responses. We include details in Appendix D, and code to reproduce our results on GitHub.

Setup We work with 5 datasets from different domains shown in Table 1. Each dataset encodes user preferences over items as votes, ballots, ratings, or rankings. We process each dataset to convert these data into pairwise comparisons – allowing for ties. We then use the same processed dataset to build rankings for our approach and 4 baseline approaches. We construct a solution path of selective rankings for all dissent values using Algorithm 2, and report solutions for 3 values of τ:

- SPA_0, the solution for $\tau = 0$. It captures a selective ranking that reflects unanimous collective preferences.
- $\mathsf{SPA}_{\text{min}}$, the solution for $\tau_{\text{min}} > 0$, i.e., the smallest dissent value that yields a selective ranking with 2+ tiers. It captures the minimum disagreement needed for any collective comparison.
- $\mathsf{SPA}_{\text{maj}}$, the solution for $\tau_{\text{max}} < 0.5$, i.e., the largest dissent value. It captures the most granular collective comparison supported by the data.

We construct rankings using the following baseline methods:

- *Voting Rules*: We consider Borda [12] and Copeland [22], which are voting rules from social choice that rank items based on position or pairwise wins.
- *Median Rankings*: We consider Kemeny [42], which returns a ranking that minimizes disagreement by solving a discrete optimization problem. We use the coranko library [5], and use the 'BioConsert' heuristic for datasets greater than 10 items, due to runtime constraints.
- *Sampling*: We consider MC4, which returns a ranking through a sampling-based approach [26], and can be viewed as an analog of Copeland [31].

Results We summarize the specificity, disagreement, and robustness of rankings from all methods and all datasets in Table 1. In what follows, we discuss these results.

On Selective Rankings Our results highlight different ways a selective ranking can reveal disagreement – e.g., through the dissent parameter, the structure of tiers, or a

		Selective			Standard			
Dataset	**Metrics**	SPA_0	$\mathsf{SPA}_{\text{min}}$	$\mathsf{SPA}_{\text{maj}}$	Borda	Copeland	Kemeny	MC4
nba	Disagreement Rate	0.0%	2.0%	6.4%	8.3%	8.3%	8.1%	7.9%
$n = 7$ items	Abstention Rate	100.0%	42.9%	28.6%	–	–	–	–
$p = 100$ users	# Tiers	1	2	4	7	7	7	6
28.6% missing	# Top Items	7	3	1	1	1	1	1
NBA [52]	Δ Sampling	0.0%	0.0%	0.0%	4.8%	4.8%	4.8%	0.0%
	Δ-Adversarial	0.0%	0.0%	0.0%	19.0%	19.0%	14.3%	19.0%
survivor	Disagreement Rate	0.0%	0.2%	0.2%	6.8%	6.6%	6.7%	6.4%
$n = 39$ items	Abstention Rate	94.9%	42.5%	42.5%	–	–	–	–
$p = 6$ users	# Tiers	2	5	5	39	39	39	39
0.0% missing	# Top Items	1	1	1	1	1	1	1
Purple Rock [54]	Δ Sampling	0.0%	0.0%	0.0%	1.3%	0.8%	0.9%	0.8%
	Δ-Adversarial	0.0%	0.0%	0.0%	2.6%	1.8%	1.6%	3.1%
lawschool	Disagreement Rate	0.0%	0.3%	3.1%	4.7%	4.2%	4.1%	4.2%
$n = 20$ items	Abstention Rate	40.5%	36.8%	4.2%	–	–	–	–
$p = 5$ users	# Tiers	4	6	15	20	20	20	20
0% missing	# Top Items	12	12	2	1	1	1	1
LSData [46]	Δ Sampling	0.0%	0.0%	0.0%	1.6%	1.1%	29.5%	0.5%
	Δ-Adversarial	0.0%	0.0%	0.0%	3.7%	2.6%	45.8%	2.6%
csrankings	Disagreement Rate	0.0%	0.0%	0.1%	12.3%	12.2%	13.7%	12.2%
$n = 175$ items	Abstention Rate	100.0%	98.9%	95.5%	–	–	–	–
$p = 5$ users	# Tiers	1	2	3	175	175	175	175
0% missing	# Top Items	175	1	1	1	1	1	1
csrankings.org [23]	Δ Sampling	0.0%	0.0%	0.0%	0.8%	0.8%	9.0%	0.1%
	Δ-Adversarial	0.0%	0.0%	0.0%	3.1%	1.7%	11.1%	0.1%
sushi	Disagreement Rate	0.0%	13.6%	42.6%	42.6%	42.6%	42.6%	42.6%
$n = 10$ items	Abstention Rate	100.0%	64.4%	0.0%	–	–	–	–
$p = 5,000$ users	# Tiers	1	2	10	10	10	10	10
0.0% missing	# Top Items	10	8	1	1	1	1	1
Kamishima [41]	Δ Sampling	0.0%	0.0%	0.0%	0.0%	0.0%	2.2%	2.2%
	Δ-Adversarial	0.0%	0.0%	0.0%	2.2%	2.2%	11.1%	11.1%

Table 1. Comparability, disagreement, and robustness of rankings for all methods on all datasets. We report the following metrics for each ranking: *Disagreement Rate*, i.e., the fraction of collective preferences that conflict with users; *Abstention Rate*, i.e., the fraction of collective preferences that abstain from comparison; *# Tiers*, the number of tiers or ranks. *# Top Items*, i.e., the number of items in the top tier or rank. Δ-*Sampling*, the average fraction of collective preferences that are inverted when we drop 10% of individual preferences; and Δ-*Adversarial*, the maximum fraction of collective preferences that are inverted when we flip 10% of individual preferences, respectively.

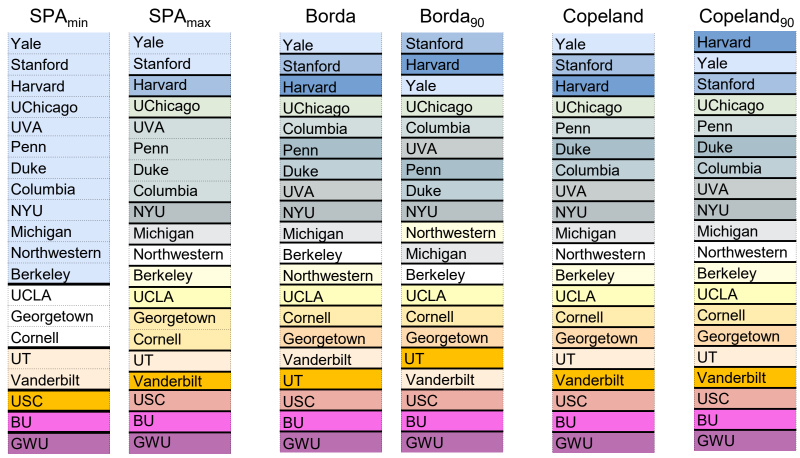


Fig. 5. Consensus rankings of U.S. law schools produced by selective preference aggregation and voting rules on the `lawschool` dataset. Here: the selective rankings for $\mathsf{SPA}_{\text{min}}$ and $\mathsf{SPA}_{\text{maj}}$ correspond to dissent values of $\tau_{\min} = \frac{1}{5}$ and $\tau_{\max} = \frac{2}{5}$, respectively; Borda_{90} corresponds to the ranking from Borda on a dataset where we drop 10% of individual preferences.

combination of both. When seeking a single winner, we can report the threshold dissent for the top tier to contain only one item. This varies across datasets: 0.0 for `survivor` to over 0.48 for `nba`. In general, there is no guarantee that a preference aggregation task will admit a single winner. In `law`, for example, we find that even most granular selective ranking $\mathsf{SPA}_{\text{maj}}$ contains two items in its top tier: Stanford and Yale. In this case, we find that the ranking arises when we set the dissent $\tau_{\max} = \frac{2}{5}$. As we discuss in Theorem 4.1, this implies that these two schools are collectively preferred to all other schools in at least 3 of the 5 rankings.

We can apply a similar line of reasoning to identify dissent values where a selective ranking would achieve a total order. In such cases, the corresponding dissent parameter reflects the number of individual preferences we must be willing to overrule to achieve consensus. For example, $\mathsf{SPA}_{\text{maj}}$ on `sushi` returns a total order for a $\tau = 0.4998$, indicating existing consensus, whereas `law` does not return a total order at any level of dissent, signaling deep underlying differences on certain items, including the top tier.

On Robustness One of the main limitations of standard approaches for preference aggregation is their sensitivity. In effect, it is well-known that such methods can return rankings that change dramatically when we change their inputs [10, 35, 44]. In Table 1, we evaluate the robustness of rankings with respect to two kinds of issues that arise frequently in practice: (1) missingness and (2) misreporting. In particular, we show how the collective preferences for each ranking change when we apply the method on a corrupted dataset where we randomly drop 10% of individual preferences, or randomly invert 10% of preferences. We repeat this process 100 times and report the number of inversions between the collective preferences we obtain using the original dataset and the output produced using the corrupted datasets.

Our results in Table 1 highlight how selective rankings are robust to such effects. In particular, we observe that 0.0% of the collective preferences expressed in a selective ranking will change when we drop or corrupt 10% of preferences. In contrast, we find that existing methods can often exhibit varying degrees of brittleness. On the `nba` dataset, for example, we find that the collective preferences expressed in rankings from Borda and Copeland changed an average of 4.8% when we drop 10% of individual preferences, and up to 19.0% when we flip them.

On the Arbitrariness of Arbitration Our results highlight how principled approaches to preference aggregation can output conflicting summaries for collective preferences. As shown in Fig. 5, voting rules such as Borda and Copeland can identify the same set of top items yet exhibit differences at less salient positions (see e.g., differences in Berkeley, Michigan and Northwestern). In some cases, these effects can arise due to differences in individual preferences. In Fig. 5, for example, we find that Borda and Copeland rank Yale, Stanford and Harvard as the top-3 law schools. However, these rankings will change once we drop 10% of preferences Borda_{90} and $\mathsf{Copeland}_{90}$. In other cases, they may arise due to algorithm design. For instance, in Table 1, Kemeny arbitrarily breaks a tie in preference between coaches Steve Nash and Monty Williams in the `nba` dataset, causing it to have higher disagreement than MC4, which allows for ties. When individuals express conflicting preferences, there are often many equally principled approaches to arbitration. In practice, these effects can compromise the significance and legitimacy of using rankings — as they lead to systems where the top items are determined by differences in algorithm design rather differences in individual preferences.

6 Learning by Agreeing to Disagree

Some of the most salient use cases for preference aggregation in machine learning arise when we wish to align models with the collective preferences of their users. In the simplest case, we would recruit users to label training examples. Given their labels, we would then aggregate them to train a model or fine tune it [47]. We often rely on such approaches in tasks such as medical image segmentation [38] where individuals express conflicting preferences due to ambiguity [65] or subjectivity [34, 29]. In such settings, standard aggregation methods such as majority vote can lead to models whose predictions reflect the collective preferences of the majority [64, 21]. In what follows, we explore how selective aggregation can mitigate these effects by returning training labels that better account for all annotators' views.

Setup We consider a task to build a classifier to detect toxic conversations with a language model. We work with the DICES dataset [8], which contains individual toxicity labels for $n = 350$ conversations from $p = 123$ users. Here, each label is defined as $y_i^k \in \{1, -1, 0\}$ if user k labels conversation i as $\{\texttt{toxic}, \texttt{benign}, \texttt{unsure}\}$, respectively. We randomly split users into two groups: a group of $p^{\text{train}} = 5$ users whose labels we use to train our model; and a group of $p^{\text{test}} = 118$ users whose labels we use to evaluate the predictions of the model at an individual level once it is deployed. We set the relative size of each group to reflect the relative size of annotators and end-users in practice – i.e., where a company would collect labels from a small subset of users to train a model that assigns predictions to a large population.

We use this setup to construct four sets of training labels. We aggregate discordant annotations, where one conversation is labeled as toxic and the other non-toxic. We drop all annotations where a user rates a conversation as "unsure" – i.e., where $y_{i,k} = 0$. This ensures that $y_{i,k} \in \{-1, 1\}$.

- $y_i^{\text{Maj}} := \mathbb{I}\left[\sum_{k\in[p]} \mathbb{I}\left[y_i^k = 1\right] \geq \sum_{k\in[p]} \mathbb{I}\left[y_i^k = -1\right]\right]$, which reflects a common approach to aggregate labels in machine learning [60]. When an item has split votes, y_i^{Maj} is set to toxic.
- $y_i^{\text{Borda}} \in [280]$, aggregate labels from a variant of Borda for pairwise preferences [15].
- $y_i^{\text{SPA}} \in [15]$, which reflects aggregate labels from SPA for the maximum $\tau < 0.5$.
- $y_i^{\text{Exp}} \in \{0, 4\}$, which reflects granular safety labels elicited from an in-house expert. This reflects a baseline where we choose to train a model using annotations from a single human expert.

We process the training labels from each method to ensure that we can use a standard training procedure across similar methods. We use the training labels from each method to fine-tuning a BERT-Mini model [70] that maps tokens to their respective toxicity labels, and denote these models as $f^{\text{SPA}}, f^{\text{Maj}}, f^{\text{Borda}}, f^{\text{Expert}}$.

We evaluate how each method performs with respect to individuals and users in a specific group in terms of the following measures:

$$\text{BER}_k(f^{\text{all}}) := \tfrac{1}{2}\text{TPR}_k(f^{\text{all}}) + \tfrac{1}{2}\text{FPR}_k(f^{\text{all}}) \quad (1)$$

$$\text{LabelError}(y^{\text{all}}) := \frac{1}{p}\sum_{k=1}^{p} \mathsf{BER}_k(y^{\text{all}}) \quad (2)$$

$$\text{PredictError}(f^{\text{all}}) := \frac{1}{p}\sum_{k=1}^{p} \mathsf{BER}_k(f^{\text{all}}) \quad (3)$$

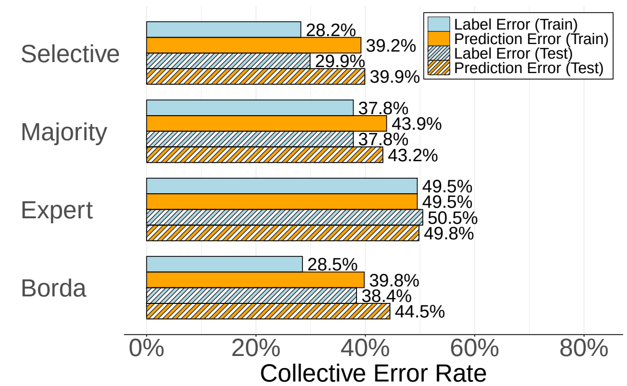
Here: LabelError (2) captures the discrepancy between individual labels and aggregate labels for an average user in a group. PredictError (3) captures the discrepancy between individual labels and predictions of a model trained with aggregate labels. We compute these measures with respect to aggregate labels and predictions after applying thresholds to optimize BER. We report these measures in terms of BER for the sake of clarity, as the data exhibits class imbalance that can vary across users. We include additional details on our setup in Appendix D.5.

Results We summarize our results at a group level in Fig. 6a and an individual level in Fig. 6b.

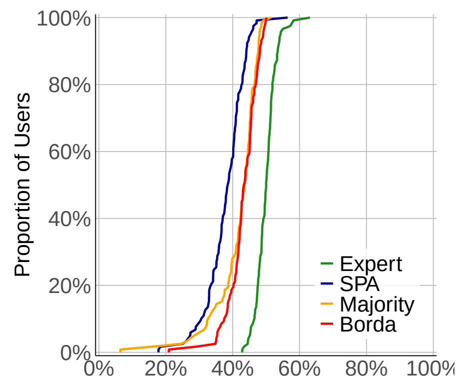
Our results in Fig. 6a highlight how SPA aggregates labels in a way that minimizes disagreement across users – achieving a label error of 28.2% (c.f., 37.8% with y^{Maj}). Moreover, the improved alignment in training labels can lead to propagate into an improved alignment in the predictions of the model. In this case, f^{SPA} has a train prediction error of 29.9% (c.f., 38.4 % on f^{Borda}) and 39.9% test prediction error (c.f., 44.5 % with f^{Borda}).

Our results in Fig. 6b, we show how the prediction error is distributed across the $p^{\text{train}} = 5$ annotators in the train set – i.e., users whose preferences we would collect and observe, as well as the $p^{\text{test}} = 118$ held out annotators, whose preferences we would not be able to know. In this case, we find that roughly 60% of users achieve an individual BER of 40% or less under y^{SPA}, compared to roughly 20% of users for y^{Borda} and y^{Maj}.

Our broadly results highlight a benefit from building models using labels that encode collective preferences. In this case, the large values of label error for y^{Exp} imply that many users disagree with their annotations. The result suggests that there is considerable inherent disagreements among the user population. These findings capture the performance of each approach in a task where we threshold the predictions of each method to optimize the balanced error rate. In practice,



(a) Collective error rates – label error and prediction error – for each method on the DICES dataset. We report values for the train split with annotators and the test set of $p = 118$ held-out users. Selective aggregation achieves the lowest error across all types and splits, and generalizes, with less difference in label and predictive error than both Borda and majority vote.



(b) Cumulative distributions of individual error rates for models built using different methods for label aggregation. For each model f, we plot the fraction of users $p^{\text{test}} = 118$ users in the test set where $\text{BER}(f) \leq \delta$ for $\delta \in [0, 1]$.

we observe similar findings at other salient operating points – e.g., the most accurate model that can achieve a collective TPR of 90%. In such cases, baselines such as majority vote may underperform as their labels can only capture binary information.

7 Concluding Remarks

In many applications where we aggregate human preferences, disagreement should be treated as a "signal, not noise" [7]. We proposed an alternative paradigm to aggregate preferences in such settings—summarizing collective preferences as a partial order. This approach can reveal disagreement to end-users, allow them to reason about it, and control it.

Our work develops foundations for this paradigm. We designed an algorithm that is simple, versatile, and safe. Its main limitation is that it behaves conservatively when datasets are missing many individual preferences. Such datasets are common in settings where elicitation is a bottleneck—either because it is costly or because we must elicit preferences over a large item set.

In these cases, we can still express collective preferences as selective rankings. However, each ranking may collapse into a single tier. This behavior is intentional—it flags where any comparison could be invalidated by missing preferences. But it is also impractical at scale—most datasets are sparse and contain few overlapping ratings. Looking forward, we can extend our paradigm to these settings by adopting probabilistic assumptions [see e.g., 2], or by developing procedures that streamline elicitation [e.g., via RLAIF 48].

Acknowledgements

We thank the following individuals for comments that improved this work: Margaret Haffey; Yang Liu; David Parkes; Ariel Procaccia; Cynthia Rudin; and Devrarat Shah.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences to this work, many of which we have discussed in the manuscript.

References

[1] Above the Law. Top law schools 2023. https://abovethelaw.com/top-law-schools-2023/, 2023.

[2] Achab, M., Korba, A., and Clémençon, S. Dimensionality reduction and (bucket) ranking: a mass transportation approach. In *Algorithmic Learning Theory*, pp. 64–93. PMLR, 2019.

[3] Ailon, N. Learning and optimizing with preferences. In *International Conference on Algorithmic Learning Theory*, pp. 13–21. Springer, 2013.

[4] Alcaraz, J., Landete, M., and Monge, J. F. Rank aggregation: Models and algorithms. In *The palgrave handbook of operations research*, pp. 153–178. Springer, 2022.

[5] Andrieu, P. Corankco: A python package for ranking and consensus. https://github.com/pierreandrieu/corankco, 2023.

[6] Andrieu, P., Brancotte, B., Bulteau, L., Cohen-Boulakia, S., Denise, A., Pierrot, A., and Vialette, S. Efficient, robust and effective rank aggregation for massive biological datasets. *Future Generation Computer Systems*, 124:406–421, 2021.

[7] Aroyo, L. and Welty, C. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.

[8] Aroyo, L., Taylor, A., Diaz, M., Homan, C., Parrish, A., Serapio-García, G., Prabhakaran, V., and Wang, D. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Arrow, K. J. *Social Choice and Individual Values*. John Wiley & Sons, New York, 2nd edition, 1951. Revised edition published in 1963.

[10] Asudeh, A., Jagadish, H., Miklau, G., and Stoyanovich, J. On obtaining stable rankings. *Proceedings of the VLDB Endowment*, 12(3), 2018.

[11] Azoulay, P. and Li, D. Scientific grant funding. Technical Report 26889, National Bureau of Economic Research, March 2020. URL http://www.nber.org/papers/w26889. Revised June 2021.

[12] Borda, J. d. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

[13] Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

[14] Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. *Handbook of computational social choice*. Cambridge University Press, 2016.

[15] Burka, D., Puppe, C., Szepesváry, L., and Tasnádi, A. Voting: A machine learning approach. Technical Report 145, Karlsruhe Institute of Technology (KIT), Institute of Economics, November 2020. URL https://hdl.handle.net/10419/227125.

[16] Cachel, K. *Fair Consensus Decision Making: Preference Aggregation Techniques for Candidate Group Fairness*. PhD thesis, Worcester Polytechnic Institute, 2025.

[17] Chang, A., Rudin, C., Cavaretta, M., Thomas, R., and Chou, G. How to reverse-engineer quality rankings. *Machine learning*, 88:369–398, 2012.

[18] College Kickstart LLC. U.s. news & world report posts 2023 college rankings. https://www.collegekickstart.com/blog/item/u-s-news-world-report-posts-2023-college-rankings, 2022.

[19] Collins, H. W., Jenkins, S. M., Strzelecka, N., Gasman, M., Wang, N., and Nguyen, T. Ranking and rewarding access: An alternative college scorecard. *Penn Center for Minority Serving Institutions*, 2014.

[20] Condorcet, M. d. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, pp. 1785, 1785.

[21] Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. Social choice for ai alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

[22] Copeland, A. H. A reasonable social welfare function. seminar on applications of mathematics to social sciences, university of michigan. *Ann Arbor*, 1951.

[23] csrankings.org. csrankings.org. URL https://csrankings.org/.

[24] Dai, J. and Fleisig, E. Mapping social choice theory to rlhf. *arXiv preprint arXiv:2404.13038*, 2024.

[25] Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00449. URL https://doi.org/10.1162/tacl_a_00449.

[26] Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation revisited, 2001.

[27] Economist, T. Columbia is the latest university caught in a rankings scandal, 2024. URL https://www.economist.com/united-states/columbia-is-the-latest-university-caught-in-a-rankings-scandal/21808445.

[28] Espeland, W. N., Sauder, M., and Espeland, W. *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation, 2016.

[29] Fleisig, E., Abebe, R., and Klein, D. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6715–6726, 2023.

[30] Franc, V., Prusa, D., and Voracek, V. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

[31] Freund, D. and Williamson, D. P. Rank aggregation: New bounds for mcx. *Discrete Applied Mathematics*, 252:28–36, 2019.

[32] Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[33] Gionis, A., Mannila, H., Puolamäki, K., and Ukkonen, A. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 561–566, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933395. doi: 10.1145/1150402.1150468. URL https://doi.org/10.1145/1150402.1150468.

[34] Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., and Bernstein, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

[35] Halpern, D., Kehne, G., Procaccia, A. D., Tucker-Foltz, J., and Wüthrich, M. Representation with incomplete votes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5657–5664, 2023.

[36] Hochbaum, D. S. and Levin, A. Methodologies and algorithms for group-rankings decision. *Management Science*, 52(9):1394–1408, 2006.

[37] Jamieson, K. G. and Nowak, R. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

[38] Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., and Zheng, Y. Learning calibrated medical image segmentation via multi-rater agreement modeling. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12336–12346, 2021. URL https://api.semanticscholar.org/CorpusID:235702932.

[39] Jiang, N.-J. and Marneffe, M.-C. d. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374, 2022.

[40] Joren, H., Marx, C. T., and Ustun, B. Classification with conceptual safeguards. In *The Twelfth International Conference on Learning Representations*, 2023.

[41] Kamishima, T. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 583–588, 2003.

[42] Kemeny, J. G. Mathematics without numbers. *Daedalus*, 88 (4):577–591, 1959.

[43] Larkin, M. How northeastern gamed the college rankings. Boston Magazine, Aug 2014. URL https://www.bostonmagazine.com/news/2014/08/26/how-northeastern-gamed-the-college-rankings/.

[44] Liang, R., Soloff, J. A., Barber, R. F., and Willett, R. Assumption-free stability for ranking problems, 2025. URL https://arxiv.org/abs/2506.02257.

[45] List, C. Social Choice Theory. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

[46] LSData. Law school rankings 2022. https://www.lsd.law/law-school-rankings-2022, 2022.

[47] Lukasik, M., Chen, L., Narasimhan, H., Menon, A. K., Jitkrittum, W., Yu, F. X., Reddi, S. J., Fu, G., Bateni, M., and Kumar, S. Bipartite ranking from multiple labels: On loss versus label aggregation. *arXiv preprint arXiv:2504.11284*, April 2025. doi: 10.48550/arXiv.2504.11284. URL https://arxiv.org/abs/2504.11284. Accepted at ICML 2025.

[48] Mahan, D., Phung, D., Rafailov, R., Blagden, C., Lile, N., Castricato, L., Fränken, J.-P., Finn, C., and Albalak, A. Generative reward models, 2024. URL https://arxiv.org/abs/2410.12832.

[49] Moulin, H. *Axioms of cooperative decision making*. Number 15. Cambridge university press, 1991.

[50] Myerson, R. B. Nash equilibrium and the history of economic theory. *European Economic Review*, 43(4-6):671–697, April 1999.

[51] Nagaraj, S., Liu, Y., Calmon, F., and Ustun, B. Regretful decisions under label noise. In *The Thirteenth International Conference on Learning Representations*, 2025.

[52] NBA. 2020-21 nba coach of the year voter selections. https://ak-static.cms.nba.com/wp-content/uploads/sites/46/2021/06/2020-21-NBA-Coach-of-the-Year-Voter-Selections.pdf, 2021.

[53] Prabhakaran, V., Davani, A. M., and Diaz, M. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

[54] Purple Rock. Survivor season rankings – spoiler-free summaries. https://www.purplerockpodcast.com/survivor-season-rankings-spoiler-free-summaries/, 2023.

[55] Schaekermann, M. *Human-AI Interaction in the Presence of Ambiguity: From Deliberation-based Labeling to Ambiguity-aware AI*. PhD thesis, University of Waterloo, 2020.

[56] Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., and Law, E. Capturing expert arguments from medical adjudication discussions in a machine-readable format. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 1131–1137, 2019.

[57] Schoeffer, J., De-Arteaga, M., and Elmer, J. Perils of label indeterminacy: A case study on prediction of neurological recovery after cardiac arrest. *arXiv preprint arXiv:2504.04243*, 2025.

[58] Sen, A. K. A possibility theorem on majority decisions. *Econometrica: Journal of the Econometric Society*, pp. 491–499, 1966.

[59] Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618*, 2014.

[60] Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614–622, 2008.

[61] Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N., et al. The Leaderboard Illusion. *arXiv preprint arXiv:2504.20879*, 2025.

[62] Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

[63] Smith, J. H. Aggregation of preferences with variable electorate. *Econometrica: Journal of the Econometric Society*, pp. 1027–1041, 1973.

[64] Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

[65] Stutz, D., Cemgil, A. T., Roy, A. G., Matejovicova, T., Barsbey, M., Strachan, P., Schaekermann, M., Freyberg, J., Rikhye, R., Freeman, B., et al. Evaluating ai systems under uncertain ground truth: a case study in dermatology. *arXiv preprint arXiv:2307.02191*, 2023.

[66] Tang, G. and Fan, N. A survey of solution path algorithms for regression and classification models. *Annals of Data Science*, 9(4):749–789, 2022.

[67] TestMax Inc. LSATMax: Comprehensive lsat prep course. https://testmaxprep.com/lsat, 2025.

[68] The Washington Post. Colleges are dropping out of rankings: Here's why yale says it's had enough, Nov 2022. URL https://www.washingtonpost.com/politics/2022/11/18/collegesrankingsyale/.

[69] Top Universities. Qs world university rankings for law and legal studies 2024. https://www.topuniversities.com/university-subject-rankings/law-legal-studies, 2024.

[70] Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

[71] Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

[72] Velocity Test Prep. Top 50 law school rankings & comparisons. https://www.velocitylsat.com/resources/top-law-schools, 2023.

[73] Wang, J. and Shah, N. B. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *arXiv preprint arXiv:1806.05085*, 2018.

[74] Wang, J. and Shah, N. B. Ranking and rating rankings and ratings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13704–13707, 2020.

[75] Wang, S., Deng, Q., Feng, S., Zhang, H., and Liang, C. A survey on rank aggregation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pp. 8281–8289, 2024.

[76] WorldQuant. Voting issues: A brief history of preference aggregation, 2025. URL https://www.worldquant.com/ideas/voting-issuesa-brief-history-of-preference-aggregation/. Accessed: 2025-01-31.

[77] Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

Supplementary Materials
Selective Preference Aggregation

A Supplementary Material for Section 2

A.1 Notation

We provide a list of the notation used throughout the paper in Table 2.

Object	Symbol	Description
Items	$i \in [n] := \{1, \ldots, n\}$	The objects being ordered, for which users have expressed preferences.
Users	$k \in [p] := \{1, \ldots, p\}$	Individuals expressing preferences for given items.
Individual preferences	$\pi^k_{i,j} \in \{-1, 0, 1\}$	Pairwise preference between items i and j for user k.
Tiered ranking	T	The unordered set of supervertices (tiers) created during the creation of a selective ranking.
Collective preference	$\pi_{i,j}(T) \in \{-1, 0, 1\}$	The preference between items i and j in a given ranking.
Selective ranking	S	The ranking outputted by $\mathsf{SPA}_\tau(\mathcal{D})$.
Dissent parameter	$\tau \in [0, \frac{1}{2})$	The admitted dissent between two items i and j.

Table 2. Notation

A.2 Encoding Individual Preferences as Pairwise Comparisons

Representation	Notation	Conversion	Tasks
Labels	$y^k_i \in \{0, 1\}$	$\pi^k_{i,j} = \mathbb{I}\left[y^k_i > y^k_j\right] - \mathbb{I}\left[y^k_i > y^k_j\right]$	Pairwise annotations,, i.e fine-tuning
Ratings	$y^k_i \in [m]$	$\pi^k_{i,j} = \mathbb{I}\left[y^k_i > y^k_j\right] - \mathbb{I}\left[y^k_j > y^k_i\right]$	5-Star Ratings, i.e Product Ratings
Rankings	$r^k : [n] \to [n]$	$\pi^k_{i,j} = \mathbb{I}\left[r^k(i) > r^k(j)\right] - \mathbb{I}\left[r^k(i) < r^k(j)\right]$	Item Orderings, i.e Grant Proposals [11]

Table 3. Data structures that capture ordinal preferences over n items. Each representation can be converted into a set of $\binom{n}{2}$ pairwise preferences in a way that ensures (and assumes) transitivity. Item-level representations require fewer queries but may be subject to calibration issues between annotators.

One of the benefits in developing machinery to aggregate preferences is that it can provide practitioners with flexibility in deciding how to elicit and aggregate the preferences. In practice, such choices involve trade-offs that we discuss briefly below. Specifically, eliciting pairwise preferences from users requires more queries than other approaches [37]. However, it may recover a more reliable representation of ordinal preferences than ratings or rankings [i.e., 59]. In tasks where we work with a few items, we can elicit preferences as ratings, rankings, or pairwise comparisons. In tasks where we elicit rankings, we can convert them into pairwise comparisons without a loss of information. In this case, eliciting pairwise comparisons can test implicit assumptions such as transitivity. In tasks where we elicit labels and ratings, the conversion is lossy – since we are converting cardinal preferences to ordinal preferences. In practice, this conversion can resolve issues related to calibration across users [see e.g, 74, 73]. In theory, it may also resolve disagreement [58].

B Supplementary Material for Section 3

This appendix provides supplementary material for Section 3, including proofs of the claims in this section and an description of the solution path algorithm.

B.1 Proof of Correctness

Lemma B.1. Consider the graph before running condensation or topological sort, but after pruning edges with weights below τp. Items can be placed in separate tiers without violating Disagreements$(T) \leq \tau p$ if and only if there is no cycle in the graph involving those items.

Proof. We start by connecting the edges in a graph to conditions on the items in a tiered ranking and eventually expand that connection to show the one-to-one correspondence between cycles and tiers.

First note that for any items i, j: $w_{i,j} > \tau \iff \sum_{k=1}^{p} 1\left[\pi_{i,j}^{k} \neq 1\right] > \tau p$. This follows trivially from the definition of $w_{i,j} := \sum_{k=1}^{p} 1\left[\pi_{i,j}^{k} \neq 1\right]$. From this, we know that if and only if there exists an arc (i, j) that is not pruned before condensation, we cannot have a tiered ranking with $\pi_{i,j}^{T} = -1$ without violating Disagreements$(T) \geq \tau p$.

If there exists a cycle in this graph, then we know the items in that cycle must be placed in the same tier. To show this, consider some edge i, j in the cycle. We know item j cannot be in a lower tier than i without violating the disagreements property, from the above. So item j must be in the same or a higher tier. But item j has an arrow to another item, k, which must be in the same or a higher tier than both j and i, and so on, until the cycle comes back to item i. This corresponds to the constraint that all items must be in the same tier.

If a set of items is not in a cycle, then these items do not need to be placed in the same tier. If the items are not in a cycle, then there exists a pair of items (i, j) such that there is no path from j to i. Thus i can be placed in a higher tier than j without violating any disagreement constraints. Thus not all items in this set need to be placed in the same tier.

Thus we have shown that for a graph pruned with a given value of τ, items can be placed in separate tiers for a tiered ranking based on that same parameter τ, if and only if there is no cycle in the graph involving all of these items.

□

We draw on this Lemma to prove the main result:

Theorem B.2. Given a preference aggregation task with n items and p users, Algorithm 1 returns the optimal solution to SPA_τ for any dissent parameter $\tau \in [0, \frac{1}{2})$.

Proof of Theorem B.2. Consider that items in our solution are in the same tier if and only if they are part of a cycle in the pruned graph (i.e., if and only if they are in the same strongly connected component). So items are in the same tier if and only if they must be in the same tier for the solution to be feasible. No other feasible tiered ranking could have any of these items in separate tiers. So no other tiered ranking could have any more tiers, or any more comparisons. To do so would require placing some same-tier items in different tiers. Thus, our solution is maximal with respect to the number of tiers, and with respect to the number of comparisons. □

B.2 Proof of Uniqueness

Theorem B.3. The optimal solution to SPA_τ is unique for $\tau \in [0, 0.5)$.

Proof of Theorem B.3. Let T denote an optimal solution to SPA_τ. We will show that the optimality T is fully specified by: (1) the items in each tier and (2) the ordering between tiers. That is, if we were to produce a tiered ranking T' that assigns different items to each tier, or that orders tiers in a different way would be suboptimal or infeasible.

Consider a tiered ranking T that is feasible with respect to SPA_τ for some $\tau \in [0, 0.5)$. Let T' denote a tiered ranking where we swap the order of two tiers in T. We observe that the T' must violate a constraint. To see this, consider any pair of items i, j such that $\pi_{i,j}(T) = 1$ before the swap, but $\pi_{j,i}(T') = 1$ after the swap. One such pair must exist for any swapping of tier orders, because all tiers are non-empty. Because we elicited complete preferences, one of the following conditions must hold:

$$\sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right] > \tau p \tag{4}$$

$$\sum_{k\in[p]} \mathbb{I}\left[\pi_{j,i}^k \neq 1\right] > \tau p \tag{5}$$

Assuming that T was an optimal solution to SPA_τ, we observe that the condition in Eq. (4) must be violated because the original optimal solution was valid. Thus, we must have that $\sum_{k\in[p]} \mathbb{I}\left[\pi_{j,i}^k \neq 1\right] > \tau p$. This implies that $\text{Disagreements}(T') > \tau p$ for this tiered ranking. Thus, swapping the order of tiers violates constraints because $\tau < 0.5$.

Now note that any separation of items from within the same tier is not possible without violating a constraint. This follows from Lemma B.1, which states that items that are part of a cycle in our graph representation of the problem[1], must be in the same tier for a solution to be valid. And, as specified in our algorithm, we know our optimal solution has tiers only where there are cycles in the graph representation of the problem. So any tiers in the optimal solution cannot be separated.

We can still merge two tiers together without violating constraints, but such an operation reduces the number of comparisons and would no longer be optimal. And after merging two tiers, the only valid separation operation would be simply to undo that merge (since any other partition of the items in that merged tier, would correspond to separating items that were within the same tier in the optimal solution). So we cannot use merges as part of an operation to reach a valid alternative optimal solution.

So we know that for the optimal solution, we cannot separate out any items within the same tier, and we cannot reorder any of the tiers. Merging, meanwhile, sacrifices optimality. Thus, the original optimal solution is unique. □

B.3 Constructing All Possible Selective Rankings

We start with a proof for Proposition 3.1.

Proof of Proposition 3.1. Recall that Algorithm 1, an edge (i, j) with weight $w_{i,j}$ is excluded if at least τp users disagree with the preference $j \succ i$. We observe that $w_{i,j} = \sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \geq 0\right]$ corresponds the number of users who disagree with the preference $j \succ i$. Given a dataset, denote the set of dissent values that could lead to different outputs as:

$$\mathcal{W} = \{0\} \cup \left\{\tau' \mid \exists i, j : \tau' = \left(\frac{1}{p}\sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \geq 0\right]\right) < \tfrac{1}{2}\right\}$$

This corresponds to the set of unique $w_{i,j}/p$ for all i, j, with the value 0 included as well. To see this, note $w_{i,j} = \sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \geq 0\right]$. We will now show the following Lemma, which will resolve the original claim.

Lemma B.4. *Given any two adjacent elements $a, b \in \mathcal{W} \cup \{\frac{1}{2}\}$. All dissent values in $\tau \in [a, b)$ lead to the same selective ranking as the selective ranking for $\tau = a$.*

Proof. To show this, note that there exists no edge $i \to j$ such that $ap < w_{i,j} < bp$. If there did exist, then we would have

$$a < \frac{w_{i,j}}{p} < b.$$

This would imply that $\mathcal{W}$ would have to include an additional between a and b. But a and b are adjacent in $\mathcal{W}$. This is a contradiction.

Since there exists no edge $i \to j$ such that $ap < w_{i,j} < bp$, there exists no edge such that the decision to include its arc in the graph changes based on what value of dissent we select in $[a, b)$. Recall that we exclude $i \to j$ iff $w_{ij} \geq \tau p$ □

Now that we know that for any two adjacent values a, b in $\mathcal{W} \cup \{\frac{1}{2}\}$, all dissent values in $[a, b)$ lead to the same tiered ranking as with dissent value a, we know that for any dissent value $\tau \in [0, \frac{1}{2})$, the largest value of $\tau' \in \mathcal{W}$ that is $\leq \tau$ will

[1] after pruning edges of weight below τ

lead to the same tiered ranking. Simply substitute τ in for a, and the smallest value above τ in $\mathcal{W} \cup \{\frac{1}{2}\}$ for b (such a value exists, on both sides, because 0 and $\frac{1}{2}$ are both $\in \mathcal{W} \cup \{\frac{1}{2}\}$, and $\tau \in [0.\frac{1}{2})$).

Thus we have shown the required claim.

□

Algorithm We present an algorithm to construct a solution path of selective rankings in Algorithm 2.

Algorithm 2 Solution Path Algorithm

Input: $\mathcal{D} = \{\pi^k_{i,j}\}_{i,j\in[n],k\in[p]}$ — *preference dataset*
1: $\mathcal{S} = \{\}$ — *initialize solution path*

Construct Initial Preference Graph for $\tau = 0$

2: $w_{i,j} \leftarrow \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$ for all $i, j \in [n]$ — *$w_{i,j}$ = # preferences claiming $i \succeq j$*
3: $V_I \leftarrow [n]$ — *Vertices represent items*
4: $A_I \leftarrow \{(i \to j) \mid w_{i,j} \geq 0\}$ — *Arcs for observed preferences*

Construct Selective Rankings for All Possible Dissent Values

5: $\mathcal{W} \leftarrow \{w_{i,j}$ for all $i, j \in [n] \mid w_{i,j} < \lceil \frac{p}{2} \rceil\} \cup \{0\}$ — *Set of dissent parameters (see Proposition 3.1)*
6: **for** $\tau \in \mathcal{W}$ **do**
7: $\quad A_I \leftarrow A_I / \{(i \to j) \in \mid w_{i,j} \geq \tau p\}$ — *Add arcs with support $\geq \tau p$*
8: $\quad V_T \leftarrow \mathsf{ConnectedComponents}((T, A_T))$ — *Group items into tiers*
9: $\quad A_T \leftarrow \{(T \to T') \mid \exists i \in T, j \in T' : (i \to j) \in A_I\}$ — *Add edges between items to supervertex*
10: $\quad (l_1, \ldots, l_{|V_T|}) \leftarrow \mathsf{TopologicalSort}((V_T, A_T))$ — *Sort components based on directed edges*
11: $\quad S_\tau \leftarrow (T_{l_1}, \ldots, T_{l_{|V_T|}})$
12: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{S_\tau\}$
13: **end for**
Output: $\mathcal{S}$ — *Selective rankings that cover the comparison-disagreement frontier*

Given a preference dataset Algorithm 2 returns a finite collection of selective rankings $\mathcal{S}$ that achieve all possible trade-offs of comparability and dissent. The procedure improves the scalability by restricting the values of the dissent parameter τ as per Proposition 3.1 in Line 2, and by reducing the overhead of computing graph structures. In this case, we construct the preference graph once in Line 4, and progressively add arcs with sufficient support in Line 7.

Algorithm 2 assumes a complete preference dataset – i.e., where we have all pairwise preferences from all users. In practice, we can satisfy this assumption by imputing missing preferences to 0 as described in Proposition 4.2. Alternatively, we can also add an additional step after Line 7 to check that the item graph (V_I, A_I) remains connected.

Details on Synthetic Dataset in Fig. 3 We benchmarked Algorithm 2 to Algorithm 1 in Fig. 3 on synthetic preference aggregation tasks where we could vary the number of users and items. We fixed the number of users to $p = 10$ users. For each user $k \in [p]$, we sampled their pairwise preferences as $\pi^k_{i,j} \sim \mathsf{Uniform}(1, 0, -1)$.

B.4 Proofs of Algorithm Runtime

Algorithm 1 Line 1 computes a sum while visiting each pairwise preference for each judge, taking $\mathcal{O}(n^2 p)$ time. All subsequent steps are linear in the graph size: both ConnectedComponents and TopologicalSort are linear in input size, and the other steps are just operations on each edge. So the total runtime is $\mathcal{O}(n^2 p)$.

Algorithm 2 Note that $|\mathcal{W}| = \lceil \frac{p}{2} \rceil$, because w_{ij} only takes integer values and there are $\lceil \frac{p}{2} \rceil$ integers between 0 and $\lceil \frac{p}{2} \rceil$ inclusive of 0 and exclusive of $\lceil \frac{p}{2} \rceil$. so the for loop runs $\lceil \frac{p}{2} \rceil$ times, and everything in the loop runs in time linear in the graph size, so $\mathcal{O}(n^2)$. Thus the whole runtime of the loop is $\mathcal{O}(n^2 p)$. The preprocessing, as before, is $\mathcal{O}(n^2 p)$ time. Note that computing $\mathcal{W}$ can be done in $\mathcal{O}(n^2 p)$ time: just iterate through all w_{ij} for each of the $\lceil \frac{p}{2} \rceil$ possible distinct values, and add the value to $\mathcal{W}$ if it occurs at least once. Thus the total runtime is the sum of a constant number of $\mathcal{O}(n^2 p)$ steps, meaning the total runtime is $\mathcal{O}(n^2 p)$.

C Supplementary Material for Section 4

This appendix provides proofs and additional results to support the claims in Section 4.

C.1 On the Top Tier

Theorem C.1. *Consider a preference aggregation task where at most $\alpha < \frac{1}{2}$ of users strictly prefer one item over all other items. Given any $\tau \in [0, \frac{1}{2})$, the tiered ranking from SPA_τ will include at least two items in its top tier.*

Proof. We show the contrapositive: having $> (1-\tau)$ users rank an item first guarantees having only one item in the top tier. Without loss of generality, call an item with $> (1-\tau)$ users rating a specific item first A. Consider WLOG any other item B. No more than τ users claim either of $B \succ A$ or $B \sim A$, because we know $> (1-\tau)$ users claim $A \succ B$. So for any tiered ranking that places some other item B in the same tier as A, we could instead place A above all other items in that tier, and have one more item. Since the result of our algorithm must have the maximal number of tiers, we cannot have a case where A is in the same tier as any other item. □

Lemma C.2. *Consider a preference aggregation task where a majority of users strictly prefer an item i_0 over all items $i \neq i_0$. There exists some threshold dissent $\tau_0 \in [0, \frac{1}{2})$ such that for all $\tau > \tau_0$, every selective ranking we obtain by solving SPA_τ will place i_0 as the sole item in its top tier.*

Proof. Let α denote the fraction of users who strictly prefer i_0 over all items. Since $\alpha > \frac{1}{2}$, we observe that at most $1 - \alpha < 1 - \frac{1}{2}$ users can express a conflicting preference. Given any item $i \neq i_0$, let $\tau_0 = 1 - \alpha$ denote the fraction who users who believe either of $i \succ i_0$ or $i \sim i_0$. For any tiered ranking that places i_0 and i in the same tier, we could instead place i above all other items in that tier, and have one more tier. Since our algorithm returns a tiered ranking with the maximal number of tiers, we cannot have a case where i is in the same tier as any other item. □

C.2 On Missing Preferences

Proof of Proposition 4.2. If we are missing preferences, our algorithm's behavior is to assume all missing preferences would be in disagreement with any asserted ordering. This exactly corresponds to the actual disagreement if the true values are all asserted equivalence/indifference, and an upper bound on dissent if the preferences are directional. By doing this, we guarantee that the disagreement property will be satisfied under any possible missingness mechanism, even a worst-case adversarial mechanism. We denote missingness as $\pi_k(i,j) =?$ if the preference is missing. This property is trivial to show. Consider that

$$\begin{aligned}
\text{Disagreements}(T) &:= \max_{\substack{i,j \in T, T' \\ T \succ T'}} \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] \\
&\leq \max_{\substack{i,j \in T, T' \\ T \succ T'}} \sum_{k \in [p]} 1\left[\pi^k_{i,j} \in \{0, -1, ?\}\right] \\
&= \max_{\substack{i,j \in T, T' \\ T \succ T'}} \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \in \{0, -1\}\right] \text{ if we we set all missing values } \pi^k_{i,j} =? \text{ to } \pi^k_{i,j} = 0
\end{aligned}$$

Given that overall disagreement when preferences are imputed cannot increase, we have that $\pi_{i,j}(S^{\text{true}}_\tau) = \pi_{i,j}(S^{\text{safe}}_\tau)$.

More formally: from the disagreements argument above, we know that $\mathcal{D}^{\text{safe}}$ has the same or more disagreements for any preference than does $\mathcal{D}^{\text{true}}$. Every selective comparison in S^{safe}_τ corresponds to a pair of items in distinct strongly connected components under the constraints from $\mathcal{D}^{\text{safe}}$ (see Lemma B.1). When we relax to only the constraints from $\mathcal{D}^{\text{true}}$, we cannot have more disagreement for any preferences, so those items will remain in distinct strongly connected components. Since they remain in distinct strongly connected components, Lemma B.1 tells us the two items will not be in the same tier.

To show that the two items will have the same ordering in both tiered rankings, note that even under $\mathcal{D}^{\text{true}}$ there must be a constraint on one of the two directions of the preference[2]. And that constraint will still hold under $\mathcal{D}^{\text{safe}}$, which is no less constrained than $\mathcal{D}^{\text{true}}$. Thus, S^{true}_τ cannot have a preference in the opposite direction from S^{safe}_τ

[2]Given a dataset of complete pairwise preferences and $\tau \in [0, \frac{1}{2})$, we must have that at least one of the following holds: $\sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] > \tau p$ or $\sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \neq -1\right] > \tau p$. (This is because for the former claim to be true, we'd need at least

□

C.3 On the Distribution of Dissent

A selective ranking only allows comparisons that violate at most τp of preferences in a dataset. In practice, these violations may be disproportionately distributed across users or items. For example, we may have a task with $\tau = \frac{1}{p}$ where the same user disagrees with all comparisons in a dataset. Alternatively, the violations may be equally distributed across users – so that there is no coalition of users who agrees with all preferences. In Remark C.3, we bound the number of users who can disagree with a selective ranking.

Remark C.3. A τ-selective ranking contradicts the preferences of at most $\frac{p^2}{4} \cdot \tau p$ users.

The result in Remark C.3 only applies in tasks where the number of users exceeds the number of selective comparisons. In other tasks – where the number of selective comparisons exceeds the number of users – the statement is vacuous as we cannot rule out a worst-case where every user disagrees with at least one comparison.

Proof. We observe that a selective ranking with a single tier makes no claims. Thus we can restrict our attention to cases where the τ-selective ranking contains at least two tiers. Given a selective ranking with more than 2 tiers, then any user who disagrees with the ranking of items from non-adjacent tiers, also disagrees with the ranking of two items in adjacent tiers. So every user with a conflict must disagree about the ordering of at least one pair of items in adjacent tiers. This bounds the number of users who disagree as τ times the number of distinct pairs of items in adjacent tiers. This is because no more than τ proportion of users can disagree with any one pairing.

The number of distinct, adjacent-tier pairs is of the form $\sum_{l=1}^{|T|-1} n_l n_{l+1}$ where tier ; contains n_l items, and all the tiers together contain all n items ($\sum_{i=l} |T| n_l = n$). This quantity is maximized when we have $|T| = 2$ tiers that contain $\frac{n}{2}$ items each (rounding if n is odd). In this case, the maximum value is $\frac{n}{4}$ (or slightly below if n is odd). The worst case is tight, achieved with two tiers, each with half the items, and an even number of items. □

C.4 On Stability with Respect to New Items

We start with a simple counterexample to show that selective rankings do not satisfy the "independence of irrelevant alternatives" axiom [9].

Example C.4 (Selective Rankings do not Satisfy IIA)**.** Consider a preference aggregation task where we have pairwise preferences from 2 users for 2 items i and j where both users agree that $i \succ j$.

$$\text{User 1}: \quad i \succ j$$
$$\text{User 2}: \quad i \succ j$$

In this case, every τ-selective ranking would be $\pi_{i,j}(T) = 1$ for any $\tau \in [0, 0.5)$.

Suppose we elicit preferences for a third item z, and discover that each user asserts that z is equivalent to a different item:

$$\text{User 1}: \quad i \sim z \succ j \quad \longleftrightarrow \quad i \succ j \quad z \succ j \quad i \sim z$$
$$\text{User 2}: \quad i \succ j \sim z \quad \longleftrightarrow \quad i \succ j \quad j \sim z \quad i \succ z$$

In this case, every τ-selective ranking would be $\pi_{i,j}(T) = 0$ for all $\tau \in [0, \frac{1}{2})$. This violates IIA because the relative comparison $\pi_{i,j}(T)$ changes depending on the preferences involving z.

Proposition C.5. *Consider a preference aggregation task where for a given $\tau \in [0, \frac{1}{2})$ we construct a selective ranking S_n using a dataset $\mathcal{D}$ of complete pairwise preferences from p users over n items in the itemset $[n]$. Say we elicit pairwise preferences from all p users with respect to a new item $n+1 \notin [n]$ and construct a selective ranking S_{n+1} for the same τ over the new itemset $[n+1] := [n] \cup \{n+1\}$.*

Given any two items $i, j \in [n]$, we have that $(\pi_{i,j}(S_{n+1}) = \pi_{i,j}(S_n)) \vee (\pi_{i,j}(S_{n+1}) = 0)$.

Proof. It is sufficient to show the following:

$(1-\tau)p$ preferences to be 1, which then forces the latter claim to be false because we've set $(1-\tau)p > \tau p$ values to be something other than -1).

- When $\pi_{i,j}(S_n) \neq -1$, we never have $\pi_{i,j}(S_{n+1}) = -1$
- When $\pi_{i,j}(S_n) \neq 1$, we never have $\pi_{i,j}(S_{n+1}) = 1$.

Given a dataset of complete pairwise preferences and $\tau \in [0, \frac{1}{2})$, at least one of the following conditions must hold:

$$\text{Condition A:} \qquad \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] > \tau p$$

$$\text{Condition B:} \qquad \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq -1\right] > \tau p$$

This is because for Condition A to be False, we would need at least $(1-\tau)p$ preferences to be 1, which then forces Claim B to be true because we have set $(1-\tau)p > \tau p$ values to be something other than -1.

Consider WLOG that Condition A holds. If $\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] > \tau p$, then we know that $\pi_{i,j}(S_n) \neq 1$. Otherwise we would violate the disagreement constraint in SPA_τ. Note that eliciting preferences for a new item does not change $\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right]$. So we still have $\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] > \tau p$, and we still have $\pi_{i,j}(S_{n+1}) \neq 1$. Thus, we have that both $\pi_{i,j}(S_n) \neq 1$ and $\pi_{i,j}(S_{n+1}) \neq 1$. We can apply a symmetric argument to show Condition B holds. In this case, we would have that $\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq -1\right] > \tau p$ and see that both $\pi_{i,j}(S_n) \neq -1$ and $\pi_{i,j}(S_{n+1}) \neq -1$.

This guarantees that the claim of Proposition 4.3 cannot be violated. When $\pi_{i,j}(S_n) = 0$ so too does $\pi_{i,j}(S_{n+1}) = 0$. When $\pi_{i,j}(S_n) \neq -1$ we never have $\pi_{i,j}(S_{n+1}) = -1$, when $\pi_{i,j}(S_n) \neq 1$ we never have $\pi_{i,j}(S_{n+1}) = 1$. Thus we have proven the claim by cases. □

Proposition C.6. *Consider a preference aggregation task where we have a complete dataset $\mathcal{D}$ with n items and p users. Let:*

- $w_{ij} := \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq -1\right]$ *denote the number of users who disagree with the claim $j \succ i$.*
- $m_{ij} \in \{1, \ldots, w_{ij} - \tau p\}$
- S'_τ *be the selective ranking on a dataset with m_{ij} preferences between items i and j having been inverted.*

Then for any pair of items $i, j \in [n]$ where

$$m_{i,j} < w_{i,j} - \tau p.$$

We have that:

$$\pi_{i,j}(S_\tau) = 1 \implies \pi_{i,j}(S'_\tau) \neq -1$$

That is, a collective preference expressed in a selective ranking between items i and j cannot be inverted unless $m_{i,j} + 1$ preferences are inverted.

Since $w_{i,j} \geq 0.5p$ when $\pi_{i,j}(S_\tau) = 1$, we can also say that if $\pi_{i,j}(S_\tau) = 1$, then $\pi_{i,j}(S'_\tau) \neq -1$ provided $m(\mathcal{D}) + p\tau < 0.5 \cdot p$

Proof. Let $w_{i,j}(\mathcal{D}) := \mathbb{I}\left[\pi^k_{i,j} \neq -1\right]$ denote the number of users who disagree with $j \succ i$ in the dataset $\mathcal{D}$. Let m denote the number of preferences that are flipped in the dataset – we assume a worst case outcome, where all flipped preferences are between i and j (which we denote $m_{i,j}$, and set equal to m). In a dataset where we flip m preferences the number of users who disagree with $j \succ i$ is no lower than $w - m$. A comparison $i \succ j$ can only invert to $j \succ i$ if the proportion of disagreement with $j \succ i$ falls below τ.

$$\frac{w - m}{p} < \tau.$$

We can re-arrange this inequality to obtain:

$$m > w - p\tau.$$

Thus, a comparison $i \succ j$ will invert to $j \succ i$ only if $m > w - p\tau$. □

D Supplementary Material for Sections 5 and 6

In what follows, we include additional details and results for the experiments in Section 5 and our demonstration in Section 6.

D.1 Descriptions of Datasets

Dataset	n	p	Format	Description
`nba`	101 Voters	7 Coaches	Ballots	2021 NBA Coach of the Year Award, where sports journalists vote for the top 3 coaches
`lawschool`	26 Schools	5 Rankings	Rankings	Top U.S. law schools ranked by 5 organizations based on academic performance, reputation, and other metrics in 2023.
`survivor`	40 Seasons	6 Fans	Rankings	Rankings task where 6 fans of the show Survivor rank seasons 1-40 from best to worst.
`sushi`	10 Sushi Types	5,000 Respondents	Pairwise	Benchmark recommendation dataset collected in Japan, where participants provided pairwise preferences over 10 different types of sushi: ebi (shrimp), anago (sea eel), maguro (tuna), ika (squid), uni (sea urchin), ikura (salmon roe), tamago (egg), toro (fatty tuna), tekka-maki (tuna roll), and kappa-maki (cucumber roll).
`csrankings`	175 Departments	5 Subfields	Rankings	Rankings of computer science departments from csrankings.org based on research output in AI, NLP, Computer Vision, Data Mining, and Web Retrieval.

Table 4. Overview of datasets. We consider five datasets from salient use cases of preference aggregation.

D.2 List of Metrics

In what follows, we provide detailed descriptions of the metrics in Table 1.

Metric	Formula	Description
AbstentionRate(T)	$\frac{1}{n(n-1)}\sum_{i,j\in[n]}\mathbb{I}\left[\pi_{i,j}(T)=\bot\right]$	Given a selective ranking over n items T, the abstention rate represents the fraction of pairwise comparisons where we abstain.
DisagreementRate($T, \mathcal{D}$)	$\frac{1}{n(n-1)p}\sum_{k\in[p]}\sum_{i,j\in[n]}\mathbb{I}\left[\pi^k_{i,j}\neq\pi_{i,j}(T),\pi_{i,j}(T)\neq\bot\right]$	Given a ranking over n items T, the *disagreement rate* represents the fraction of individual preferences in $\mathcal{D}$ that disagree with the collective preferences in T.
#Tiers(S_τ)	$\lvert S_\tau\rvert$	Given a selective ranking S_τ, the number of tiers. For standard methods, each rank is converted to a tier.
#TopItems(S_τ)	$\lvert T_1\rvert$	Given $S_\tau=(T_1,\ldots,T_m)$, the number of items in the top tier. For standard methods, each rank is converted to a tier.
DisagreementPerUser($T, \mathcal{D}$)	$\underset{k\in[p]}{\text{median}}\,\frac{1}{n(n-1)/2}\sum_{i,j\in[n]}\mathbb{I}\left[\pi^k_{i,j}\neq\pi_{i,j}(T)\right]$	The median fraction of preference violations across users.
Δ Sampling ($T, \mathcal{D}$)	$\underset{b\in\{1,\ldots,N_b\}}{\text{median}}\left[\frac{\sum_{i,j\in[n]}\mathbb{I}\left[T_{i,j}\neq T^b_{i,j}\wedge T_{i,j}\neq 0\wedge T^b_{i,j}\neq 0\right]}{\sum_{i,j\in[n]}\mathbb{I}\left[T_{i,j}\neq 0\right]}\right]$	Given the ranking produced on the full dataset T, the median proportion of collective preferences that are inverted when we drop 10% of preferences. We construct a bootstrap estimate by applying the method to N_b datasets where we randomly drop 10% of all preferences and obtain N_b rankings $\{T^1,\ldots,T^{N_b}\}$.
Δ Adversarial ($T, \mathcal{D}$)	$\underset{b\in\{1,\ldots,N_b\}}{\max}\left[\frac{\sum_{i,j\in[n]}\mathbb{I}\left[T_{i,j}\neq T^b_{i,j}\wedge T_{i,j}\neq 0\wedge T^b_{i,j}\right]\neq 0}{\sum_{i,j\in[n]}\mathbb{I}\left[T_{i,j}\neq 0\right]}\right]$	Given the original ranking T, the *maximum* proportion of collective preferences inverted when we flip 10% of individual preferences. We construct a bootstrap estimate where we first apply the method to N_b datasets where we randomly flip 10% of all preferences and obtain N_b rankings $\{T^1,T^2,\ldots,T^{N_b}\}$.

Table 5. Metrics used to evaluate comparability, disagreement, and robustness of rankings in Table 1 and Appendix D.4

D.3 Selective Ranking Paths

We present the solution paths of selective rankings for each dataset in Section 5 in Fig. 7 to Fig. 11.

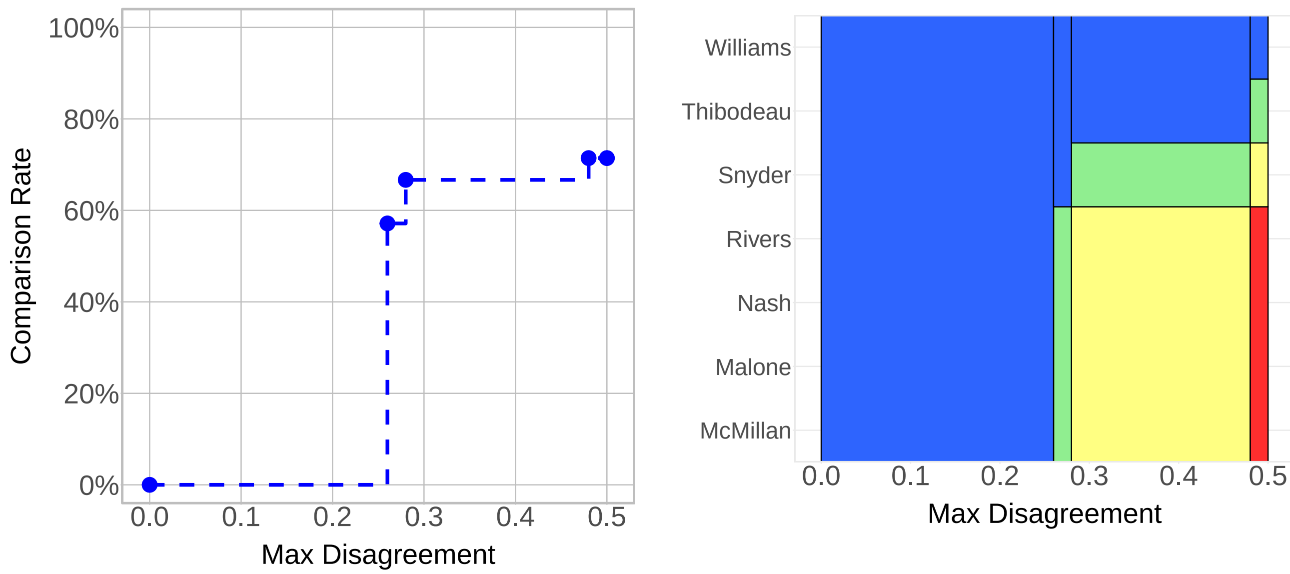


Fig. 7. Selective rankings for the `nba` dataset ($n = 7$ items and $p = 100$ users). We show the tradeoff between comparision and disagreement (left) and the unique rankings over the dissent path (right).

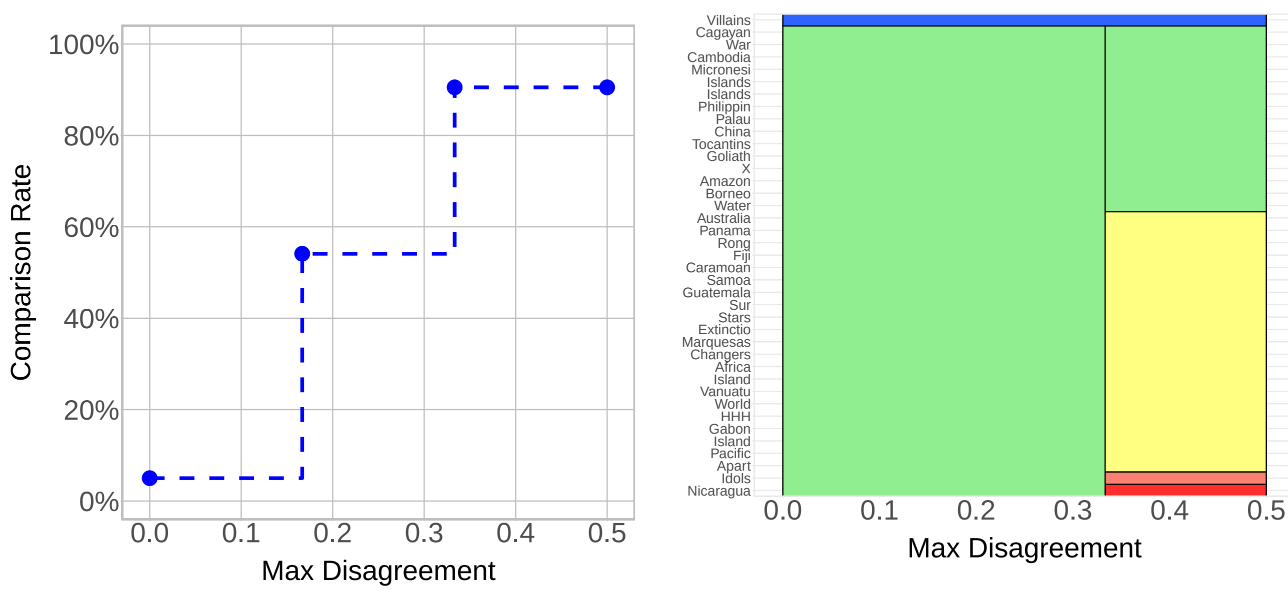


Fig. 8. Selective rankings for the `survivor` dataset ($n = 39$ items and $p = 6$ users). We show the tradeoff between comparision and disagreement (left) and the unique rankings over the dissent path (right).

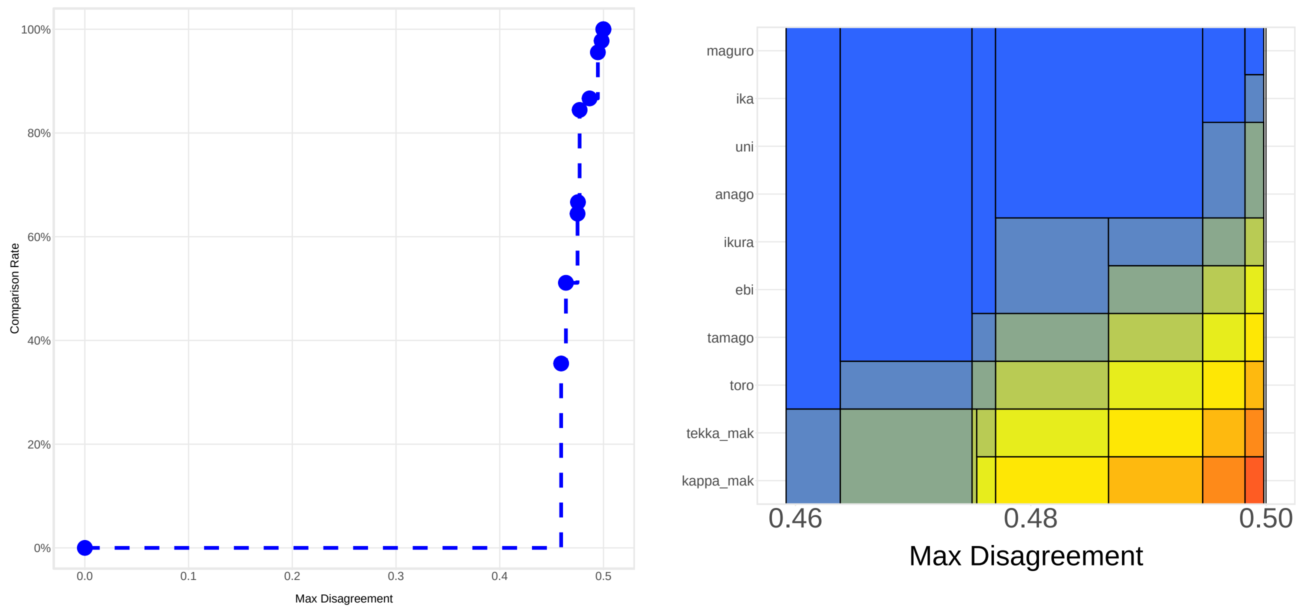


Fig. 9. Selective rankings for the `Sushi` dataset ($n = 10$ items and $p = 5000$ users). We show the tradeoff between comparision and disagreement (left) and the unique rankings over the dissent path (right). Note that only a subset of dissent values are shown for clarity.

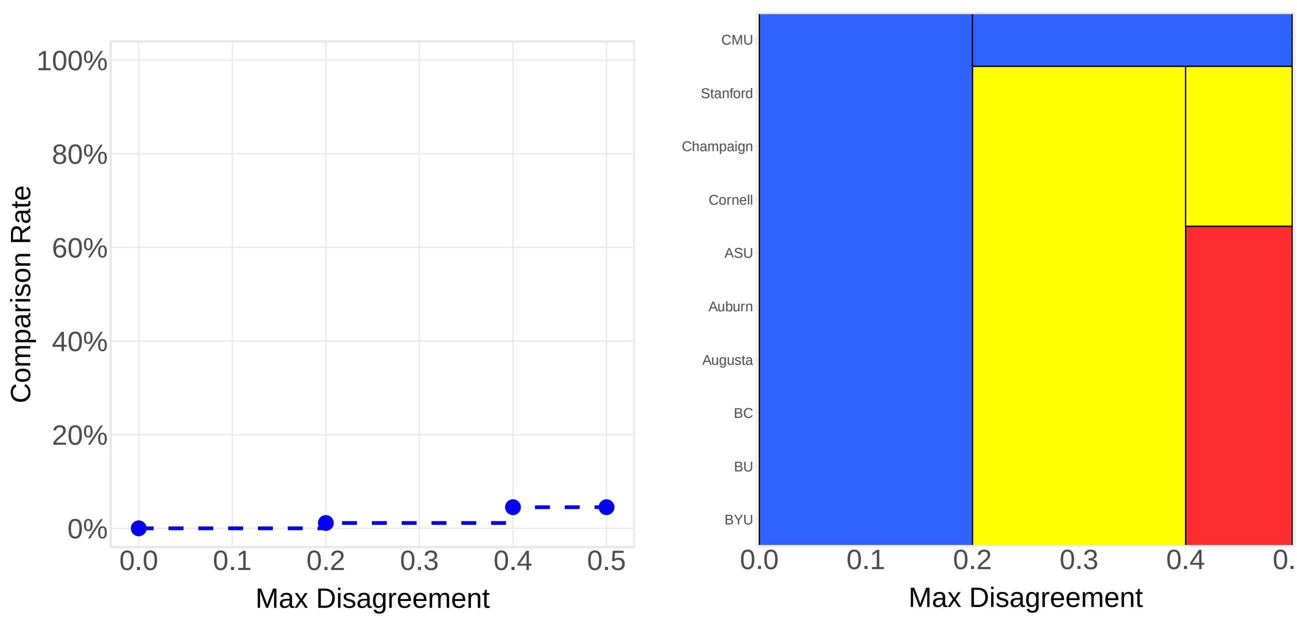


Fig. 10. Selective rankings for the `csrankings` dataset ($n = 175$ items and $p = 5$ users). We show the tradeoff between comparision and disagreement (left) and the unique rankings over the dissent path (right).

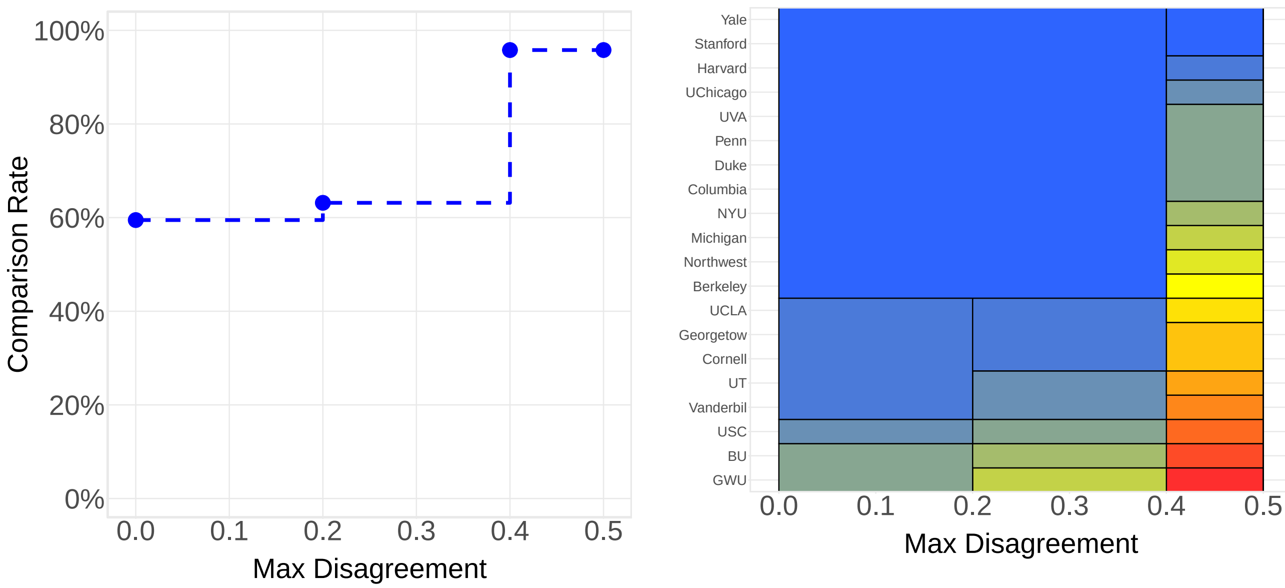


Fig. 11. Selective rankings for the `lawschool` dataset ($n = 20$ items and $p = 5$ users). We show the tradeoff between comparision and disagreement (left) and the unique rankings over the dissent path (right).

D.4 Expanded Table of Results

We include an expanded version of our results for all methods and all datasets in Appendix D.4. This table covers the same results as in Table 1, but includes the following additional metrics:

1. Δ *Abstentions [Intervention]*, which measures the proportion of strict collective preferences (e.g., $A \succ B$ or $A \prec B$) that turn into ties or abstentions in the ranking that we obtain after running the method on a modified dataset.
2. Δ *Specifications [Intervention]*, which measures the proportion of ties or abstentions that turn into ties or abstentions in the ranking that we obtain after running the method on a modified dataset.

We report these values for same interventions we consider in Section 5, namely: *Sampling*, where we run the method on a dataset where we randomly omit 10% of individual preferences; and *Adversarial*, where we run the method on a dataset where we randomly flip 10% of individual preferences. Each value corresponds to a bootstrap estimates where we perform the same estimate 100 times. For clarity, we list the $\Delta -$ Sampling as $\Delta -$ Inversions $- -$Sampling, and $\Delta -$ Adversarial $- -$Inversions.

		Selective			Traditional			
Dataset	**Metrics**	SPA_0	SPA_{min}	SPA_{maj}	Borda	Copeland	Kemeny	MC4
nba $n = 7$ items $p = 100$ users 28.6% missing NBA [52]	Disagreement Rate	0.0%	2.0%	6.4%	8.3%	8.3%	8.1%	7.9%
	Median Disagreement per User	0.0%	0.0%	4.8%	4.8%	4.8%	9.5%	9.5%
	Abstention Rate	100.0%	42.9%	28.6%	0.0%	0.0%	0.0%	4.8%
	# Tiers	1	2	4	7	7	7	6
	# Top Items	7	3	1	1	1	1	1
	Dissent	0.0000	0.2600	0.4900	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	4.8%	4.8%	4.8%	14.3%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	19.0%	19.0%	14.3%	19.0%
	Δ Specifications Sampling	0.0%	9.5%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	9.5%	0.0%	0.0%	0.0%	0.0%	4.8%
	Δ Abstentions Sampling	0.0%	0.0%	28.6%	0.0%	0.0%	0.0%	9.5%
	Δ Abstentions Adversarial	0.0%	19.0%	28.6%	0.0%	4.8%	0.0%	33.3%
survivor $n = 39$ items $p = 6$ users 0.0% missing Purple Rock [54]	Disagreement Rate	0.0%	0.2%	0.2%	6.8%	6.6%	6.7%	6.4%
	Median Disagreement per User	0.0%	0.1%	0.1%	7.2%	7.1%	7.1%	6.8%
	Abstention Rate	94.9%	42.5%	42.5%	0.0%	0.4%	0.0%	0.0%
	# Tiers	2	5	5	39	39	39	39
	# Top Items	1	1	1	1	1	1	1
	Dissent	0.0000	0.1667	0.3333	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	1.3%	0.8%	0.9%	0.8%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	2.6%	1.8%	1.6%	3.1%
	Δ Specifications Sampling	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.1%
	Δ Specifications Adversarial	0.0%	5.1%	0.0%	0.0%	0.4%	0.0%	0.3%
	Δ Abstentions Sampling	0.0%	52.4%	57.5%	0.0%	0.1%	0.0%	0.1%
	Δ Abstentions Adversarial	0.0%	57.5%	57.5%	0.0%	0.4%	0.4%	0.7%
lawschool $n = 20$ items $p = 5$ users 0% missing LSData [46]	Disagreement Rate	0.0%	0.3%	3.1%	4.7%	4.2%	4.1%	4.2%
	Median Disagreement per User	0.0%	0.0%	1.6%	4.2%	2.6%	2.1%	2.6%
	Abstention Rate	40.5%	36.8%	4.2%	0.0%	0.0%	0.0%	0.5%
	# Tiers	4	6	15	20	20	20	20
	# Top Items	12	12	2	1	1	1	1
	Dissent	0.0000	0.2000	0.4000	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	1.6%	1.1%	29.5%	0.5%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	3.7%	2.6%	45.8%	2.6%
	Δ Specifications Sampling	0.0%	11.1%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Sampling	59.5%	28.2%	95.8%	0.0%	0.0%	0.0%	0.5%
	Δ Abstentions Adversarial	59.5%	0.0%	95.8%	0.0%	1.6%	0.0%	1.6%
csrankings $n = 175$ items $p = 5$ users 0% missing csrankings.org [23]	Disagreement Rate	0.0%	0.0%	0.1%	12.3%	12.2%	13.7%	12.2%
	Median Disagreement per User	0.0%	0.0%	0.1%	12.3%	12.6%	13.5%	12.3%
	Abstention Rate	100.0%	98.9%	95.5%	0.0%	0.0%	0.0%	0.0%
	# Tiers	1	2	3	175	175	175	175
	# Top Items	175	1	1	1	1	1	1
	Dissent	0.0000	0.2000	0.4000	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	0.8%	0.8%	9.0%	0.1%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	3.1%	1.7%	11.1%	0.1%
	Δ Specifications Sampling	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
	Δ Abstentions Sampling	0.0%	1.1%	4.5%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Adversarial	0.0%	0.0%	4.5%	0.0%	0.1%	0.0%	0.0%
sushi $n = 10$ items $p = 5,000$ users 0.0% missing Kamishima [41]	Disagreement Rate	0.0%	13.6%	42.6%	42.6%	42.6%	42.6%	42.6%
	Median Disagreement per User	0.0%	13.3%	42.2%	42.2%	42.2%	42.2%	42.2%
	Abstention Rate	100.0%	64.4%	0.0%	0.0%	0.0%	0.0%	0.0%
	# Tiers	1	2	10	10	10	10	10
	# Top Items	10	8	1	1	1	1	1
	Dissent	0.0000	0.0020	0.4998	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	0.0%	0.0%	2.2%	2.2%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	2.2%	2.2%	11.1%	11.1%
	Δ Specifications Sampling	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Sampling	0.0%	35.6%	100.0%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Adversarial	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	15.6%

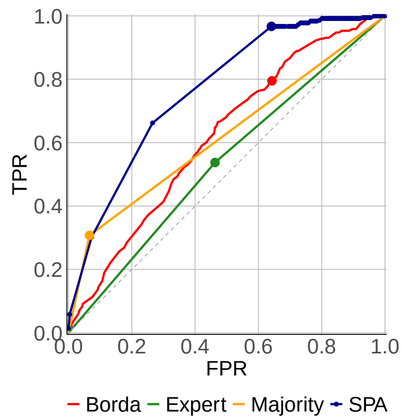


Fig. 12. ROC model curves on the training set for all four methods. We highlight the label for each method closest to tpr$>90\%$ on labels with a large dot. f^{SPA} is the only method whose chosen operating point keeps the true-positive rate above 80 % on the model output while controlling FPR.

D.5 Supplementary Material for Section 6

Selective Aggregation with Binary Annotations A key challenge in applying SPA to the DICES dataset is that it elicits categorical labels for each item individually, rather than comparative ratings. This conversion can create unnecessary equivalence, where a pairwise preference is inferred as a tie ($\pi^k_{i,j} = 0$). This is not a reflection of a user's true judgment but an artifact of two limitations: (1) users annotate items individually rather than comparing them, and (2) the annotations are restricted to $\{0, 1\}$ instead of granular ratings. For example, a user may believe item A is significantly more toxic than item B, but the conversion results in a tie if both were labeled "toxic" a distinction that is lost in this setting.

We address this by running a variant of selective aggregation where we construct aggregate labels from users who express a strict preference between items – $i \succ j$ or $j \succ i$. In addition, we assume that users who have not asserted an opinion (because of dataset scope) are "deferring judgment" to those who have.

For each pair of items $i, j \in [n]$, we define:

- $s_{i,j} := \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} = 1\right]$ denote number of users who strictly prefer item i to item j
- $s_{j,i} := \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} = -1\right]$ denote the number of users who strictly prefer item j to item i.
- The aggregate preference weight $w_{i,j}$ as the proportion of users who strictly prefer i to j among those who expressed a strict preference, scaled to n items. Note that all item pairs had at least 1 preference:

$$w_{i,j} := n \cdot \frac{s_{i,j}}{s_{i,j} + s_{j,i}}$$

In this setup, the dissent parameter τ no longer maintains its standard interpretation because users may not assign a preference to each item, and items may be assigned different weights. As a result, we produce selective rankings for all possible dissent parameters that lead to a connected graph in Algorithm 2. In this case, the maximum dissent value is specified to a threshold value where Line 4 returns a disconnected graph.

D.6 Model Training

All experiments used 5-fold cross-validation on the training split. We fine-tuned a BERT-Mini model; all fine-tuning experiments used 5-fold cross-validation on the training split. We optimized with a learning rate of 2×10^{-5} for up to 25 epochs, employing early stopping. We trained in mini-batches of size 16 and enabled oversampling of minority classes in each batch.