# Selective Preference Aggregation

**Shreyas Kadekodi**[*]
Department of Computer Science & Engineering
UC San Diego
skadekodi@ucsd.edu

**Hayden McTavish***
Department of Computer Science
Duke University
hayden.mctavish@cs.duke.edu

**Berk Ustun**
Halıcıoğlu Data Science Institute
UC San Diego
berk@ucsd.edu

## Abstract

Many tasks in machine learning depend on preferences where we aggregate preference data – from recommending products to improving the helpfulness of responses from a large language model. In such tasks, individuals express their preferences over a set of items as votes, ratings, or rankings. Given a dataset of ordinal preferences from a group of individuals, we aggregate them into a single ranking that summarizes the collective preferences as a group. When individuals express conflicting preferences between items, standard methods are designed to arbitrate this dissent to rank one item over another. In this work, we introduce a paradigm for *selective aggregation* in which we *abstain* rather than arbitrate dissent. Given a dataset of ordinal preferences from a group of users, we aggregate their preferences into a *selective ranking* – i.e., a *partial order* over items where every comparison is aligned with at least $1 - \tau\%$ of users. We develop an algorithm to construct selective rankings that achieve all possible trade-offs between comparability and disagreement.

## 1 Introduction

The study of collective preference aggregation has a long history, with formal developments dating back to the 18th century. The Marquis de Condorcet was among the first to formalize the issue of cyclic preferences, now known as Condorcet's Paradox, where group preferences can be inconsistent [9, 11]. Kenneth Arrow extended these ideas in Arrow's Impossibility Theorem [5, 6], which demonstrates that no rank-order voting system can satisfy all fairness criteria simultaneously when aggregating individual preferences.

Rank aggregation was traditionally used in voting to determine a single winner, but modern applications —such as product recommendations, resource allocation, and machine learning—often require consideration of the entire preference order. In these contexts, the goal is to incorporate dissent rather than resolve conflicts, as overruling preferences can reduce the value of collective input. We introduce SPA, a method that creates orderings while preserving dissent by determining non-conflicting aggregate pairwise comparisons. This approach avoids overruling individuals and ensures that the collective preferences reflect differing inputs, preventing the information loss caused by traditional methods, as shown in Fig. 1.

Our main contributions include:

[*] These authors contributed equally to this work.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

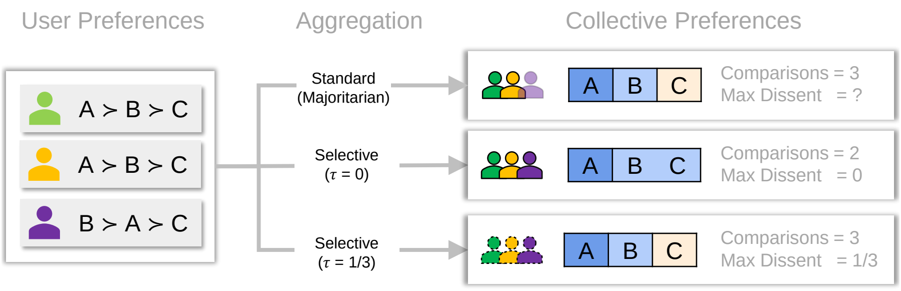

**Figure 1:** Comparison of collective preferences in a task with 3 users and 3 items, contrasting standard aggregation (top) with selective aggregation at $\tau = 0$ (middle) and $\tau = \frac{1}{3}$ (bottom). At $\tau = 0$, dissent is revealed between top items $A$ and $B$, while $\tau = \frac{1}{3}$ allows all comparisons by overruling $\frac{1}{3}$ of judges.

1. We introduce a new paradigm to aggregate ordinal preference data such as votes, labels, rankings, and ratings into a selective ranking that captures collective preferences without arbitrating disagreement.
2. We develop a fast and scalable algorithm to construct selective rankings that balance comparability and disagreement.
3. We evaluate SPA on a real-world dataset, demonstrating its ability to reduce aggregation error while reflecting diverse user preferences in pluralistic AI systems.
4. We provide an open-source Python library for selective rank aggregation at repository.

**Related Work** Our work intersects with social choice theory, rank aggregation, partial orders, and machine learning applications involving conflicting preferences. In social choice, much of the focus has been on resolving individual preferences through voting rules, with less attention to abstention or partial orders—especially in contexts requiring total rankings, such as elections [6, 8]. Similarly, rank aggregation methods typically aim to create a single ranking [2, 12], while our approach addresses cases where preferences contain ties and emphasizes handling dissent without enforcing total order [13].

In contrast, methods that use partial orders or "bucket orderings" [1, 16] offer a framework for expressing collective preferences but often lack the flexibility to handle dissent directly. Our selective ranking approach manages this by grouping items to maximize agreement while limiting dissent to a predefined threshold, balancing comparability and disagreement [15, 18]. This is particularly useful for applications like toxicity detection and personalization, where differing annotations reflect subjective variance rather than factual errors [3, 19]. Unlike existing methods, our algorithm predicts labels that respect diversity in judgments without forcing a consensus.

## 2 Algorithm

We consider a standard preference aggregation task where we wish to order $n$ items in a way that reflects the collective preferences from $p$ judges. We assume that we are given a dataset where each point represents the pairwise preference of a judge between a pair of items.

We define the individual preference function $\pi^k_{i,j}$ for judge $k$ and the aggregate preference function $\pi_{i,j}(T)$ for a set of tiers $T$ as follows, where $i$ and $j$ denote items:

$$\pi^k_{i,j} = \begin{cases} 1 & i \overset{k}{\succ} j, \\ 0 & i \overset{k}{\sim} j, \\ -1 & i \overset{k}{\prec} j \end{cases} \qquad \pi_{i,j}(T) = \begin{cases} 1 \text{ if } & i \in T_l, j \in T_{l'} \text{ where } l < l', \\ -1 \text{ if } & i \in T_l, j \in T_{l'} \text{ where } l > l', \\ \perp \text{ if } & i, j \in T_l \text{ for any } l \end{cases}$$

To summarize collective preferences, we construct a weight matrix $w_{i,j}$, which counts the number of judges who prefer item $i$ over $j$, defined as:

$$w_{i,j} = \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$$

If $\pi^k_{i,j} = 0$, we increment both $w_{i,j}$ and $w_{j,i}$ by 1.

Given a dissent parameter $\tau \in [0, 0.5)$, the *selective ranking* is constructed:

$$\max_{T \in \mathbb{T}} \quad \text{Comparisons}(T) \;\; \text{s.t.} \;\; \text{Disagreements}(T) \leq \tau p \tag{1}$$

Increasing $\tau$ allows for more granular tiers and higher tolerance for conflicting preferences, preserving more disagreements in the ranking.

Where:

- $\text{Comparisons}(T) := \sum_{i,j \in [n]} \mathbb{I}\left[\pi_{i,j}(T) \neq \perp\right]$ counts the number of valid comparisons in a tiered ranking $T$.
- $\text{Disagreements}(T) := \max_{i,j \in [n]} \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] \mathbb{I}\left[\pi_{i,j}(T) = 1\right]$ measures the maximum proportion of preferences overruled by any valid comparison in $T$.

**Algorithm 1** Selective Preference Aggregation

```
Input: pairwise preferences $\{\pi^k_{i,j}\}_{i,j \in [n], k \in [p]}$, dissent parameter $\tau \in [0, 0.5)$
Construct Selective Preference Graph
$V \leftarrow \{1, \ldots, n\}$                                   ▷ vertices are items
$A \leftarrow \{\}$                                              ▷ arcs are collective preferences
for each pair of items $i, j \in [n]$ do
    $w_{i,j} \leftarrow \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$
    if $w_{i,j} \geq \tau p$ then                                ▷ add arcs for comparisons with support ≥ $\tau$
        $A \leftarrow A \cup (i \rightarrow j)$
    end if
end for
Group Vertices by Disagreement
repeat
    Condense strongly connected components   ▷ Group items into tiers based on disagreement
until No supervertices are strongly connected                     ▷ Directed Acyclic Graph
Convert Condensed Graph to selective ranking
Order supervertices from “root/source" to “leaf/sink"              ▷ Topological Sort
Convert ordered supervertices into a selective ranking: $T_l \leftarrow S_l$ for each supervertex $l$
Output: Selective ranking that allows for maximal comparisons without violating Disagreements$(T) \leq \tau p$.
```

## 3 Learning by Agreeing to Disagree

Some of the most salient applications of preference aggregation arise in safety and alignment—e.g., improving the helpfulness or harmlessness of LLM responses [21]. Models are often trained or fine-tuned using labels that encode qualitative characteristics of machine-generated responses. These labels are produced by aggregating judgments from human annotators. In practice, the annotations may exhibit conflict due to noise [22], ambiguity [24], hidden context [21], or subjective disagreement [14, 17]. In some tasks – e.g., toxicity detection – there is no “ground truth," and standard techniques to aggregate labels will return a model that predicts the preferences of the majority [10, 23]. We present an alternative approach in which we aggregate the training labels using selective aggregation. This approach allows us to aggregate in a way that is responsive to dissent, and that can learn models that reflect the preferences of all annotators.

**Setup** We consider a binary classification task to predict the harmfulness of chatbot responses from the Diversity In Conversational AI Evaluation for Safety dataset [4]. We work with `dices350`, which contains harmfulness annotations for $n = 350$ chatbot conversations from $p = 123$ annotators. Each conversation is paired with a set of labels $y_i^k = 1$ if annotator $k$ rates conversation $i$ as toxic. We define a labeled example $(\boldsymbol{x}_i, y_i)$ for each conversation, where each $\boldsymbol{x}_i$ is a feature vector of text embeddings and $y_i$ is one of several training labels:

1. $y_i^{\text{Maj}} = \mathbb{I}\left(\sum_{k=1}^{m_i} y_i^k > \frac{m_i}{2}\right)$: the majority vote among annotators [see e.g., 20].
2. $y_i^{\text{Expert}}$: a harmfulness label from an in-house expert.
3. $y_i^{\text{Borda}}$: aggregate labels produced by applying Borda count [7].
4. $y_i^{\text{SPA}}$: aggregate labels derived from selective rank aggregation. We report results for $\mathsf{SPA}_{49}$, which allows the most dissent and clearest distinctions in output rankings.

We convert these labels into binary by testing each rank as a potential cutoff, calculating the AUC relative to $y^{\text{Expert}}$ with 'Benign' as the threshold, and selecting the rank that achieves the highest AUC as the optimal cutoff.

**Results** In Fig. 2, we summarize how well each approach can aggregate labels and predict toxicity for all individual annotators through the following measures:

- $\text{LabelError}(y^{\text{M}}) := \sum_{k\in[p]} \sum_{i\in[n]} \mathbb{I}\left[y_i^k \neq y_{i,t}^{\text{M}}\right]$, where $y_i^k = 1$ if user $k$ states that conversation $i$ is toxic, and $y_{i,t}^{\text{M}}$ if the label has a toxicity level that exceeds $t_x$.
- $\text{PredictionError}(f^{\text{M}}) = \sum_{k\in[p]} \sum_{i\in[n]} \mathbb{I}\left[y_i^k \neq f^{\text{M}}(x_i)\right]$ where $f^{\text{M}}$ is a classifier trained on $y^{\text{M}}$.

Our results show that $\mathsf{SPA}_{49}$ exhibits substantially lower disagreement compared to expert annotations prior to training. For PredictionError, baseline methods $y^{\text{Maj}}$ and $y^{\text{Borda}}$ show high label errors, each above 40%. In contrast, $\mathsf{SPA}_{49}$ has a label error of 18.4%.

For PredictionError, baseline methods like $f^{\text{Maj}}$, $f^{\text{Expert}}$, and $f^{\text{Borda}}$ retain high levels of disagreement, ranging from 39.5% to 42%. In comparison, $\mathsf{SPA}_{49}$ showed significantly lower PredictionError, at 19.3 %.

These results highlight how SPA consistently outperforms baseline methods in both labeling and prediction accuracy. With $\tau$, SPA offers a customizable balance between minimizing error and enhancing comparability where needed.

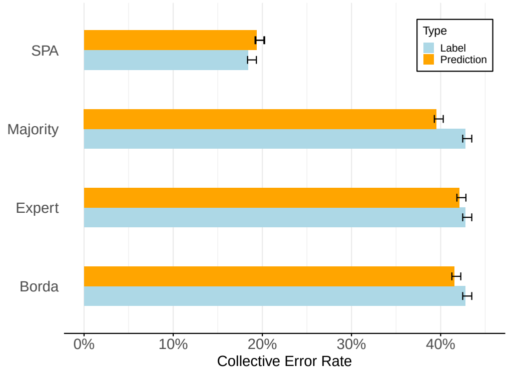

**Figure 2:** Label and Prediction Error for $\mathsf{SPA}_{49}$ are significantly lower than baseline methods

## 4 Concluding Remarks

This work introduces SPA, a new approach to preference aggregation that prioritizes preserving dissent over enforcing consensus. Traditional methods impose strict orders to resolve conflicting preferences, often overruling individual inputs and reducing accuracy. In contrast, SPA effectively aggregates preferences from ties, ratings, and rankings, accommodates non-transitive preferences, and allows for ties, offering practical advantages over other rank aggregation algorithms [8].

One limitation is the conservative handling of incomplete preferences, which may reduce flexibility in certain contexts. Future work could build on existing research by incorporating probabilistic assumptions, as in [1], to enhance robustness when faced with missing or uncertain data.

In many Machine Learning applications, disagreement should be viewed as a "signal not noise" [3]. SPA capitalizes on this by identifying and leveraging these signals. Our approach provides a practical tool for more inclusive decision-making [10, 23].

### Acknowledgements

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

## References

[1] Mastane Achab, Anna Korba, and Stephan Clémençon. Dimensionality reduction and (bucket) ranking: a mass transportation approach. In *Algorithmic Learning Theory*, pages 64–93. PMLR, 2019.

[2] Nir Ailon. Learning and optimizing with preferences. In *International Conference on Algorithmic Learning Theory*, pages 13–21. Springer, 2013.

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

[4] Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[6] Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

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

[8] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[9] Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, page 1785, 1785.

[10] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice for ai alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

[11] Nicolas de Condorcet. Sur la forme des elections. *originale*, pages 0–1, 1789.

[12] Cynthia Dwork, Ravi Kumar, Moni Naor, and D Sivakumar. Rank aggregation revisited, 2001.

[13] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D Sivakumar, and Erik Vee. Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 47–58, 2004.

[14] Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, 2023.

[15] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[16] Aristides Gionis, Heikki Mannila, Kai Puolamäki, and Antti Ukkonen. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 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.

[17] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[18] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

[19] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

[20] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 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*, pages 614–622, 2008.

[21] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

[22] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[23] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

[24] David Stutz, Ali Taylan Cemgil, Abhijit Guha Roy, Tatiana Matejovicova, Melih Barsbey, Patricia Strachan, Mike Schaekermann, Jan Freyberg, Rajeev Rikhye, Beverly Freeman, et al. Evaluating ai systems under uncertain ground truth: a case study in dermatology. *arXiv preprint arXiv:2307.02191*, 2023.

# A Omitted Proofs

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

We will use the following Lemma:

**Lemma 2.** Consider the graph before running condensation or topological sort, but after pruning edges with weight below $\tau$. 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 of Lemma 2.* 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 $i, j$: $w_{i,j} > \tau \iff \sum_{k=1}^{p} 1\left[\pi^k_{i,j} \neq 1\right] > \tau p$ This follows trivially from the definition of $w_{i,j}$ as $\sum_{k=1}^{p} 1\left[\pi^k_{i,j} \neq 1\right]$. 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 $\pi^T_{i,j} = -1$ 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 $\tau$, items can be placed in separate tiers for a tiered ranking based on that same parameter $\tau$, if and only if there is no cycle in the graph involving all of these items.

□

We now use this result to prove the statement of Theorem 1.

*Proof of Theorem 1.* 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 (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 - because 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. Note that the ordering of tiers does not affect the number of comparisons. □

## A.1 On Uniqueness

We restrict $\tau \in [0, 0.5)$ so a selective ranking is aligned with a majority of collective preferences. In this regime, $\mathsf{SPA}_\tau$ returns a unique ranking.

**Theorem 3.** The optimal solution to $\mathsf{SPA}_\tau$ is unique for $\tau \in [0, 0.5)$.

Proof of Theorem 3: The optimal solution to $\mathsf{SPA}_\tau$ is unique for $\tau \in [0, 0.5)$.

*Proof of Theorem 3.* Consider the optimal solution, and note that it is fully specified by the set of items in each tier, and the relative orderings of the tiers.

Now note that swapping the order of any tiers (or any items in different tiers) is guaranteed to violate a constraint for $\tau \in [0, 0.5)$. To see this, consider any pair of items $i, j$ such that $prfijT = 1$ before the swap, but $prfjiT = 1$ 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, we must have at least one of $\sum_{k=1}^{p} 1\left[\pi_{i,j}^{k} \neq 1\right] > \tau p$ or $\sum_{k=1}^{p} 1\left[\pi_{j,i}^{k} \neq 1\right] > \tau p$. In this case, we cannot have $\sum_{k=1}^{p} 1\left[\pi_{i,j}^{k} \neq 1\right] > \tau p$ because the original optimal solution was valid. Thus, we must have that $\sum_{k=1}^{p} 1\left[\pi_{j,i}^{k} \neq 1\right] > \tau p$, which implies that $\text{Disagreements}(T) > \tau p$ for this tiered ranking and violates the constraint. 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 2, which states that items that are part of a cycle in our graph representation of the problem[2], 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. □

### A.2 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.

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

$$\begin{aligned} \text{User 1}: &\quad A \succ B \\ \text{User 2}: &\quad A \succ B \end{aligned}$$

in this case, every $\tau$-selective ranking would $\pi_{A,B}(T) = 1$ for any $\tau \in [0, 0.5)$.

Suppose we elicit pairwise preferences for a third item $C$ and discover that each user asserts that $C$ is equivalent to a different item.

$$\begin{array}{llcccc} \text{User 1}: & A \sim C \succ B & \longleftrightarrow & A \succ B & C \succ B & A \sim C \\ \text{User 2}: & A \succ B \sim C & \longleftrightarrow & A \succ B & B \sim C & A \succ C \end{array}$$

In this case, every $\tau$-selective ranking would place $A$ and $B$ for all $\tau \in [0, \frac{1}{2})$.

**Theorem 5.** Consider a selective rank aggregation task where we construct a tiered ranking using a dataset of complete pairwise preferences from $p$ users over $n$ in the itemset $I^n$. Say we elicit pairwise preferences from all $p$ users with respect to a new item $i_{n+1} \notin I^n$ and constructing a tiered ranking over the new itemset $I^{n+1} := I^n \cup \{i_{n+1}\}$. Let $T^n$ and $T^{n+1}$ denote tiered rankings for $I^n$ and $I^{n+1}$ that we obtain by solving $\mathsf{SPA}_\tau$ for the same dissent parameter $\tau \in [0, \frac{1}{2})$. Given any two items $A, B \in I^n$, we have that $(\pi_{A,B}(T^{n+1}) = \pi_{A,B}(T^n)) \vee (\pi_{A,B}(T^{n+1}) = 0)$.

*Proof.* □

[2] (after pruning edges of weight below $\tau$

### A.3 On the Composition of the Top Tier

**Theorem 6.** Consider a prefrence 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 $\mathsf{SPA}_\tau$ 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. With 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 $\tau$ users believe either of $B \succ A$ or $B \sim A$, because we know $> (1-\tau)$ users believe $A \succ B$. 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 7.** Consider a selective rank 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 tiered ranking we obtain by solving $\mathsf{SPA}_\tau$ will place $i_0$ as the sole item in its top tier.

*Proof.* Let $\alpha$ 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. □

**Correctness** We include a proof of correctness in Theorem 1, and a proof of uniqueness in Appendix A.1. The intuition of the proofs are as follows:

1. The edges with weight $> \tau p$ corresponds to the complete set of constraints to satisfy our disagreement property for any valid tiered ordering. We cannot put an outvertex for such an edge above an invertex.
2. The condensation operation of our algorithm makes the minimum adjustments to the graph to make these constraints satisfiable - it yields the maximal set of tiers and the maximal number of comparisons.
3. For $\tau \in [0, 0.5)$, there are guaranteed to be enough constraints to force a unique output of condensation, and a unique ordering of the condensed vertices.

## B Supplementary Material for Experiments

### B.1 Experimental Results

### B.2 Case Study on the NBA Coach of the Year Award

We explore the application of the SPA algorithm to the 2020-2021 NBA Coach of the Year Award, aiming to illustrate the limitations of the existing system and how slight alterations in the scoring methodology can significantly impact the outcome. Our analysis highlights that these outcomes are heavily influenced by the minute details of each system; SPA can capture consensus while resisting changes due to arbitrary changes in scoring criteria.

**Background** In NBA award voting, rankings are obtained from prominent sports journalists and broadcasters to identify the season's top performers. The traditional point-based system employed by the NBA to determine the Coach of the Year awards points is based upon weighted rankings, where voters express multiple preferences. This system, while prevalent across various awards in sports and other domains, often fails to capture the nuanced opinions of voters, particularly in scenarios where preferences between candidates are narrowly divided.

| Dataset | Metrics | Borda | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Abstentions | 100.0% | 100.0% | 42.9% | 9.5% | 0.0% | 0.0% | 9.5% | 0.0% |
| | Minimum Disagreement per Judge | 0.0% | 0.0% | 4.8% | 4.8% | 4.8% | 4.8% | 4.8% | 4.8% |
| | Med Disagreement per Judge | 0.0% | 0.0% | 0.0% | 9.5% | 9.5% | 9.5% | 9.5% | 9.5% |
| | Maximum Disagreement per Judge | 0.0% | 0.0% | 19.0% | 23.8% | 33.3% | 33.3% | 23.8% | 33.3% |
| | Number of Items with Ties | 1 | 1 | 2 | 2 | 0 | 0 | 2 | 0 |
| | Number of Tiers | 1/7 | 1/7 | 2/7 | 5/7 | 7/7 | 7/7 | 7/7 | 7/7 |
| | Abstentions | 100.0% | 48.8% | 48.8% | 3.5% | 0.0% | 0.0% | 85.4% | 0.1% |
| | Minimum Disagreement per Judge | 0.0% | 0.1% | 0.1% | 0.4% | 3.1% | 1.4% | 0.1% | 3.1% |
| | Med Disagreement per Judge | 0.0% | 1.3% | 1.3% | 7.8% | 9.9% | 9.1% | 0.6% | 9.7% |
| | Maximum Disagreement per Judge | 0.0% | 8.3% | 8.3% | 22.4% | 22.3% | 24.7% | 3.1% | 22.3% |
| | Number of Items with Ties | 1 | 2 | 2 | 4 | 0 | 0 | 1 | 1 |
| | Number of Tiers | 1/40 | 3/40 | 3/40 | 28/40 | 40/40 | 40/40 | 4/40 | 40/40 |
| | Abstentions | 100.0% | 41.2% | 41.2% | 8.0% | 0.0% | 0.0% | 59.1% | 0.6% |
| | Minimum Disagreement per Judge | 0.0% | 3.4% | 3.4% | 2.5% | 3.4% | 3.7% | 0.6% | 3.1% |
| | Med Disagreement per Judge | 0.0% | 1.5% | 1.5% | 8.6% | 11.7% | 11.1% | 4.0% | 11.4% |
| | Maximum Disagreement per Judge | 0.0% | 4.3% | 4.3% | 15.1% | 19.4% | 19.4% | 8.0% | 19.4% |
| | Number of Items with Ties | 1 | 3 | 3 | 3 | 0 | 0 | 3 | 2 |
| | Number of Tiers | 1/26 | 3/26 | 3/26 | 16/26 | 26/26 | 26/26 | 7/26 | 26/26 |
| | Abstentions | 100.0% | 100.0% | 80.0% | 6.7% | 0.0% | 0.0% | 0.0% | 0.0% |
| | Minimum Disagreement per Judge | 0.0% | 0.0% | 2.2% | 17.8% | 20.0% | 22.2% | 22.2% | 20.0% |
| | Med Disagreement per Judge | 0.0% | 0.0% | 2.2% | 31.1% | 33.3% | 33.3% | 33.3% | 33.3% |
| | Maximum Disagreement per Judge | 0.0% | 0.0% | 8.9% | 44.4% | 48.9% | 51.1% | 51.1% | 48.9% |
| | Number of Items with Ties | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| | Number of Tiers | 1/10 | 1/10 | 2/10 | 8/10 | 10/10 | 10/10 | 10/10 | 10/10 |
| | Abstentions | 100.0% | 100.0% | 46.7% | 46.7% | 0.0% | 0.0% | 4.4% | 0.0% |
| | Minimum Disagreement per Judge | 0.0% | 0.0% | 8.9% | 8.9% | 22.2% | 26.7% | 20.0% | 22.2% |
| | Med Disagreement per Judge | 0.0% | 0.0% | 24.4% | 24.4% | 46.7% | 46.7% | 44.4% | 46.7% |
| | Maximum Disagreement per Judge | 0.0% | 0.0% | 42.2% | 42.2% | 68.9% | 68.9% | 68.9% | 68.9% |
| | Number of Items with Ties | 1 | 1 | 1 | 1 | 0 | 0 | 2 | 0 |
| | Number of Tiers | 1/10 | 1/10 | 4/10 | 4/10 | 10/10 | 10/10 | 10/10 | 10/10 |

**Table 1:** Overview of all methods on all datasets.

| | **Votes** | | | **Score** | |
|---|---|---|---|---|---|
| **Candidate** | **1st** | **2nd** | **3rd** | | **NBA** |
| Monty Williams | 45 | 32 | 19 | 340 | 353 |
| Tom Thibodeau | 43 | 42 | 10 | 351 | 352 |
| Quin Snyder | 10 | 23 | 42 | 161 | 148 |
| Doc Rivers | 2 | 2 | 8 | 24 | 24 |
| Nate McMillan | 0 | 0 | 12 | 12 | 12 |
| Steve Nash | 0 | 1 | 4 | 7 | 6 |
| Michael Malone | 0 | 0 | 5 | 5 | 5 |

**Table 2:** Tally of votes and scores for the 2020-21 NBA Coach of the Year. We show scores for the original scoring rule (NBA), which awards 5/3/1 points for each 1st/2nd/3rd place vote.

## Results

- **Impact of Score Function Variability**: The NBA's weighted voting mechanism for the Coach of the Year (COTY) determines outcomes by assigning points: 5 points for a first-place vote, 3 points for a second, and 1 point for a third. This system can lead to a coach with fewer first-place votes but a higher overall ranking across more ballots emerging as the winner. For instance, despite Monty Williams receiving more first-place votes, Tom Thibodeau was declared the winner under the traditional 5-3-1 system. Let us consider an alternative point system of 6,2,1, a subtle yet impactful adjustment from the traditional system. Despite being a relatively minor change, the recalculated total points under the new system led to a dramatic shift in the outcome, with Monty Williams now accumulating 353 points and surpassing Tom Thibodeau, who has 352 points. SPA withstands fluctuations inherent to traditional scoring methods since SPA is based on patterns of consensus across the entire set of rankings, rather than merely tallying points based on position.
- **Limitations in Capturing users' Preferences**: The current point-based system's inability to accommodate ties or express nuanced preferences is a notable limitation. When we consider a hypothetical scenario allowing users to assign tied first-place votes (with each receiving 4 points),

Thibodeau's lead paradoxically increases, despite a scenario suggesting a narrowing preference gap between Williams and Thibodeau. This is because, under the existing system, users who view multiple candidates as equally deserving must still rank them, implicitly suggesting a clear preference hierarchy that may not accurately reflect their views. Introducing the ability to express equivalent preferences for top candidates like Williams and Thibodeau reveals this rigidity, as it leads to an increase in points for Thibodeau.

- **Equivalence** Consider a scenario where users' preferences for Monty Williams and Tom Thibodeau challenge the system's flexibility. Imagine that among the users who originally ranked Williams and Thibodeau as their top 2, 1/3 of them now assign equivalent first-place votes to both coaches, where tied first-place votes are given 4 points each. Including equivalence paradoxically increases the gap between the top 2 increasing Thibodeau's total to 381 points, despite more users showing equal preference for both. This highlights a key flaw: outcomes can significantly shift without any genuine change in opinion or preference among the users.

**Our Solution**:

- SPA outputs a different ranking than the original, highlighting the variability under different scoring systems. By adjusting the dissent $\tau$, we clarify the preference hierarchy, placing Monty Williams as the clear favorite at a dissent value of 0.499, which aligns with his broader support among voters.
- Our ranking explicitly shows the degree of support and opposition for each coach, which are not evident through the traditional voting system. It enables a detailed examination of voter sentiment and produces outcomes that align more closely with the actual consensus.
- This approach is versatile and can be adapted for various decision-making contexts that require an understanding of group preferences. It is designed to handle complex scenarios, such as ties and equal rankings, facilitating more accurate and fair decision outcomes.