Selective Preference Aggregation

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

Hayden McTavish
Department of Computer Science
Duke University
hmctavish@cs.duke.edu

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

Abstract

Many tasks in machine learning are shaped by procedures where items are ordered based on the preferences of a group – from funding proposals, recommending products, or improving the helpfulness of responses from a large language model. In such settings, individuals express their preferences over items as votes, ratings, and rankings. Given a dataset of individual preferences, preference aggregation methods rank the items in a way that summarizes their collective preferences as a group. Standard methods for preference aggregation are designed to arbitrate dissent. When individuals express conflicting preferences between items, methods will rank one item over another—resolving disagreement based on axioms of social choice. 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 judges, we aggregate their preferences into a *selective ranking*—i.e., a *partial order* over items where every comparison is aligned with $1 - \tau\%$ of judges. We handle missing data effectively, ensuring robust performance across diverse settings, suiting alignment tasks. 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 [8, 10]. Kenneth Arrow extended these ideas in Arrow's Impossibility Theorem [3, 4], 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, we can better align models with human behavior by reflecting collective preferences rather than arbitrating disagreement. 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.

Our main contributions include:

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

1. We introduce a new paradigm to aggregate ordinal preference data such as votes, labels, rankings, ratings. Our paradigm aggregate these into a selective ranking that captures collective preferences without arbitrating disagreement.
2. We develop an algorithm to construct selective rankings that is fast, scalable, and easy to implement. It can output selective rankings that achieve all possible trade-offs between comparability and dissent.
3. We present a comprehensive empirical study of selective rankings on real-world and synthetic datasets. Our results show how selective rankings can promote transparency by revealing disagreement. We demonstrate how selective rankings can be used to learn from human annotations.
4. We provide an open-source Python library for selective rank aggregation at repository.

2 Algorithm

We consider a standard preference aggregation task where we aim to order n items to reflect collective preferences from p users. Each user provides pairwise preferences between items, represented as a dataset $\mathcal{D}$, where each point corresponds to a user's preference between a specific pair of items.

Individual and Aggregate Preference Functions For a given user $k \in [p] := \{1, \ldots, p\}$ and items $i, j \in [n] := \{1, \ldots, n\}$, we define the individual preference function $\pi^k_{i,j}$ as follows:

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

To express collective preferences while respecting disagreements, we introduce a *tiered ranking*:

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

In this setup, items within the same tier are not strictly ranked, which allows us to abstain from arbitrating conflicting preferences. We then define the aggregate preference function $\pi_{i,j}(T)$ based on the tiers as:

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

Here, $\perp$ denotes that the ranking abstains from comparison when i and j belong to the same tier.

Selective Aggregation We aim to produce a *selective ranking*, a tiered ranking that maximizes valid comparisons while limiting the proportion of overruled preferences. For a dissent parameter $\tau \in [0, 0.5)$, selective ranking optimizes the following objective:

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

We refer to this task as *selective preference aggregation*, denoted by SPA_τ, with:

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

The dissent parameter τ thus restricts the proportion of preferences that can be contradicted by any comparison in SPA_τ.

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 between vertices 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**

    $A \leftarrow A \cup (i \rightarrow j)$ ▷ add arcs for comparisons with support $\geq \tau p$

  end if

end for

Group Vertices by Disagreement

repeat

  Condensation procedure in notation ▷ Connected Components

until No two supervertices in condensed graph share edge ▷ Condensed Graph is DAG

Convert Condensed Graph to Tiered 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 maximizes comparisons while ensuring Disagreements$(T) \leq \tau p$

Algorithm for Selective Preference Aggregation In Algorithm 1, each item pair with at least $(1 - \tau)$ fraction of user agreement is included, and tiers are formed by grouping items within strongly connected components of the graph. This approach balances maximizing comparisons and minimizing overruled preferences, producing a unique topological order of tiers.

3 Experiments

In this section, we compare selective rankings to standard methods of preference aggregation. Our goal is to characterize the effects of on handling dissent with respect to consensus and robustness. We include code to reproduce our results in our repository and provide additional details in Appendix B.

We work with seven datasets that differ in terms of their characteristics, assumptions, and downstream applications. This includes five datasets from salient real-world applications of preference aggregation: voting (`nba`), consensus rankings (`law`), and quality assessments (`survivor`). We also consider two synthetic datasets to evaluate each method in tasks with common challenges: truth and noisy annotation.

On Selective Rankings At $\tau = 0$, SPA rankings reflect unanimous preferences, often producing multiple tiers. The number of tiers depends on agreement levels within the dataset; stronger consensus generally leads to more tiers, while varied preferences group more items together. The top tier usually includes the most universally preferred item, but when no clear favorite exists, multiple items may share the top tier.

As τ increases, selective rankings incorporate more preferences, allowing limited overruling. The dissent required to isolate a single winner varies across datasets: a small increase can sometimes suffice, though some datasets retain ties at the top tier even up to $\tau = 0.5$, indicating no clear consensus. A complete order is often infeasible, as forcing it would overrule a majority of users, as in 1.

In SPA, disagreement is capped overall, ensuring minority views are preserved across the dataset. In datasets where most judges agree, dissent may overrule some judges entirely if their preferences conflict with the majority. Where disagreement is more distributed, per-judge disagreement stays close to τ.

At low levels of dissent, SPA maintains high alignment with judges across all datasets, showing low median disagreement rates (0.0%-1.5%) for SPA_0 and SPA_{min} reflecting SPA's preference for consensus at low dissent rates. For example, in `law`, SPA_{min} creates 3 tiers with limited judge disagreement by preserving preferences even among minority judges, contrasting with baseline

methods that prioritize consensus at the cost of higher disagreements. As dissent tolerance increases, SPA captures more granular preferences. In `survivor`, SPA_{maj} partitions items into 28 out of 40 possible tiers, reflecting diverse levels of item support.

Dataset	Metrics	SPA_0	SPA_{min}	$\text{SPA}_{\text{unique}}$	SPA_{maj}	Borda	Kemeny	MC4	Copeland
	Abstentions	100.0%	100.0%	42.9%	9.5%	0.0%	0.0%	9.5%	0.0%
`nba`	Minimum Disagreement per Judge	0.0%	0.0%	4.8%	4.8%	4.8%	4.8%	4.8%	4.8%
$n = 7$ items	Med Disagreement per Judge	0.0%	0.0%	0.0%	9.5%	9.5%	9.5%	9.5%	9.5%
$p = 100$ judges	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%
`survivor`	Minimum Disagreement per Judge	0.0%	0.1%	0.1%	0.4%	3.1%	1.4%	0.1%	3.1%
$n = 101$ items	Med Disagreement per Judge	0.0%	1.3%	1.3%	7.8%	9.9%	9.1%	0.6%	9.7%
$p = 15$ judges	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%
`lawschool`	Minimum Disagreement per Judge	0.0%	3.4%	3.4%	2.5%	3.4%	3.7%	0.6%	3.1%
$n = 27$ items	Med Disagreement per Judge	0.0%	1.5%	1.5%	8.6%	11.7%	11.1%	4.0%	11.4%
$p = 5$ judges	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%
`btl`	Minimum Disagreement per Judge	0.0%	0.0%	2.2%	17.8%	20.0%	22.2%	22.2%	20.0%
$n = 10$ items	Med Disagreement per Judge	0.0%	0.0%	2.2%	31.1%	33.3%	33.3%	33.3%	33.3%
$p = 100$ judges	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%
`btl-modified`	Minimum Disagreement per Judge	0.0%	0.0%	8.9%	8.9%	22.2%	26.7%	20.0%	22.2%
$n = 10$ items	Med Disagreement per Judge	0.0%	0.0%	24.4%	24.4%	46.7%	46.7%	44.4%	46.7%
$p = 100$ judges	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.

Even at high dissent levels, SPA maintains low disagreement, especially compared to abstention rates. In the `law` dataset, disagreement remains minimal while abstention is high, with SPA_{maj} showing only 8.6% disagreement while keeping abstention under 10%. This stands in contrast to methods like Borda and Kemeny, which enforce full ordering and produce higher disagreement in datasets with varied preferences, such as in `law`, where Borda, Kemeny, and Copeland exceed 11% disagreement.

4 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 [19]. 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 [20], ambiguity [22], hidden context [19], or subjective disagreement [13, 15]. In some tasks – e.g., toxicity detection – there is no "ground truth," and standard techniques to aggregate labels (e.g., majority vote) will return a model that predicts the preferences of the majority [9, 21]. 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 benchmark [2]. 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 harmful. 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. $y^{\text{Maj}} = \mathbb{I}\left(\sum_{k=1}^{m_i} y_i^k > \frac{m_i}{2}\right)$: the majority vote among annotators [see e.g., 18].
2. y^{Expert}: a gold-standard harmfulness label from in-house experts.

3. y^{spa}, y^{Borda}, y^{MC}, y^{Maj} and y^{Copeland}: aggregate labels derived from different methods of handling rank preferences, including SPA, Borda, MC4, Majority Vote, and Copeland methods. We report results for selective rankings with τ of 0.48 and 0.49 (SPA_{48} and SPA_{49}), at which an acceptable granularity in ranking is achieved.

We convert these labels into a binary format by setting a threshold to maximize concordance with y^{Expert}, aligning the cutoffs with the percentage of responses deemed 'Benign' (the lowest ranking).

Results In Fig. 1, we summarize how well each approach can aggregate labels and predict toxicity for all individual annotators through the following measures:

- $\text{LabelError}(y^{\text{M}}) := \sum_{k\in[p]} \sum_{i\in[n]} \mathbb{I}\left[y_i^k \neq {y^{\text{M}}}_{i,t}\right]$, where $y_i^k = 1$ if user k states that conversation i is toxic, and ${y^{\text{M}}}_{i,t} = 1$ 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^{M} is a classifier trained on y^{M}.

Our results show that SPA exhibits substantially lower disagreement compared to expert annotations prior to training. For PredictionError, baseline methods like y^{Maj}, y^{Borda}, y^{MC}, and y^{Copeland} showed higher labeling errors—ranging from 20.5% to 25.1%. In contrast, SPA methods demonstrated lower error rates, with 14.9% for both SPA_{48} and 49.

For PredictionError, baseline methods like f^{Maj}, f^{Borda}, f^{MC}, and f^{Copeland} showed higher mismatch rates, ranging from 39.5% to 48.7%. In comparison, the SPA models achieved notably lower PredictionError, with 24.0% for SPA_{48} and 29.9% for SPA_{49}.

These results demonstrate how SPA's approach of refining label aggregation supports more accurate model predictions. By incorporating dissenting views into the labeling process, SPA reduces the reliance on strict majority rule, producing labels that better capture the range of perspectives within the data. By creating more accurate labels before model training, SPA models are better able to predict toxicity.

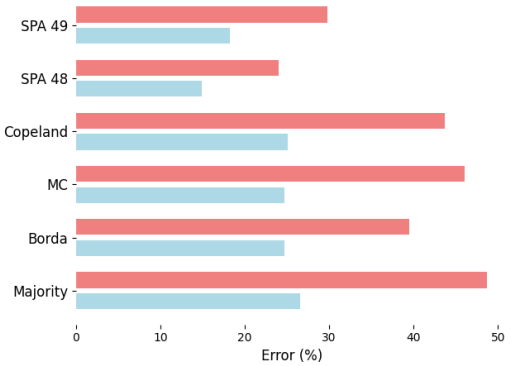

Figure 1: Comparison of label and prediction errors between Baseline and SPA models. Light blue bars represent label errors, while lighter red bars represent prediction errors. SPA shows lower errors than baselines.

5 Concluding Remarks

This work introduces SPA, a novel approach to preference aggregation that prioritizes preserving dissent rather than enforcing consensus. Traditional methods often impose strict orders, overruling individual inputs and reducing alignment with diverse perspectives found in human behaviors. In contrast, SPA effectively handles preferences derived from ties, ratings, and rankings, accommodates non-transitive preferences, and allows for ties, providing practical advantages over existing rank aggregation algorithms [5].

In many modern machine learning applications, disagreement should be treated as a "signal, not noise" [1]. SPA demonstrates clear advantages over full-order methods by using dissent as insight, highlighting diverse perspectives that baseline approaches often obscure. This approach ensures better alignment with minority views in aggregation tasks, thus providing a simple and safe method to aggregate human feedback in contexts where diverse preferences are prevalent [9, 21].

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

[2] 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.

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

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

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

[6] Jonah E. Bromwich. Former temple university business dean convicted of fraud. The New York Times, Nov 2021. URL https://www.nytimes.com/2021/11/29/us/temple-university-moshe-porat-fraud.html.

[7] Allison Chang, Cynthia Rudin, Michael Cavaretta, Robert Thomas, and Gloria Chou. How to reverse-engineer quality rankings. *Machine learning*, 88:369–398, 2012.

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

[9] 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.

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

[11] The Economist. 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.

[12] Wendy Nelson Espeland, Michael Sauder, and Wendy Espeland. *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation, 2016.

[13] 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.

[14] Malcolm Gladwell. The order of things. *The New Yorker*, Feb 2011. URL https://www.newyorker.com/magazine/2011/02/14/the-order-of-things.

[15] 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.

[16] Max Larkin. 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/.

[17] Michael Luca and Jonathan Smith. Salience in quality disclosure: Evidence from the us news college rankings. *Journal of Economics & Management Strategy*, 22(1):58–77, 2013.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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/. Accessed: date-of-access.

A Omitted Proofs

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

We will use the following Lemma:

Lemma 3. Consider the graph before running condensation or topological sort, but after pruning edges with weight below τ. 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 3. 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 $w_{i,j}$ as $\sum_{k=1}^{p} 1\left[\pi_{i,j}^k \neq 1\right]$. 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 now use this result to prove the statement of Theorem 2.

Proof of Theorem 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 (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, SPA_τ returns a unique ranking.

Theorem 4. The optimal solution to SPA_τ is unique for $\tau \in [0, 0.5)$.

Proof of Theorem 4: The optimal solution to SPA_τ is unique for $\tau \in [0, 0.5)$.

Proof of Theorem 4. 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 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 3, 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. □

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 5 (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 τ-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.

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

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

Theorem 6. 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 SPA_τ 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. □

[1](after pruning edges of weight below τ

A.3 On the Composition of the Top Tier

Theorem 7. 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 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. 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 τ 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 8. 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 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. □

Correctness We include a proof of correctness in Theorem 2, 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 Case Study on Law School Rankings

One of the most prominent applications of rankings is in higher education, where companies rank colleges and professional based on a weighted combination of features related to their educational program, resources, and reputation [12, 14]. It is well-known that score-based rankings are highly sensitive to small changes in their input data [7]. In this setting, such changes have major impact on applications, enrollment, and recruiting [see e.g., 17, who show a 1-rank drop leads to a 1% drop in applications] –forcing schools to invest in changes to maintain their ranking incentivizing *gaming* over *improvement* [see e.g., 6, 11, 16].

We demonstrate how selective rankings can provide an alternative way to rank schools in a way that can moderates these perverse incentives. Specifically, by presenting information in a format that abstains from comparisons that are subject to noise or to dissent. This is aligned with empirical results Luca and Smith [17] who show that the 'response to the information represented in [college] rankings depends on the way in which that information is presented." Here, we study these issues in the context of US law school rankings, where the influence of *US News* (USN) rankings has recently led top-ranked schools to opt out of rankings [23].

Setup We consider a strategic classification task where each school can manipulate the data reported to USN to improve their position in the 2022 USN Rankings. Our goal is to estimate the net reduction in "manipulation" by switching from the existing ranking to a selective ranking – i.e., how much each school would have to change the data reported to US News. We identify influential features by reverse-engineering a score function from dataset of historical USN rankings and school features reported from 2017–2021. The dataset contains X labelled examples $\{(x_{i,t}, y_{i,t})\}$ where $y_{i,t} \in [n]$ denotes the rank of school i in year t and $x_{i,t} \in \mathbb{R}^d$ contains $d = XX$ features used to rank them (e.g., `LSAT_score`) We use data from 2017-2021 to train a RankSVM model whose parameters we tune via k-CV. Our model correctly predicts XX.0%/80.0% of pairwise preferences on training and test data – which is reliable enough to identify the weights of each feature.

We construct selective rankings by aggregating rankings from 5 companies – treating each company as a user. In this setup, schools in the top tier would not need to manipulate their features to improve or maintain their position. Thus, we treat the manipulation as a baseline.

Dataset used to Reverse-Engineer USN Rankings We constructed our dataset used to train the score function D_{score} by compiling historical rankings from US News across seven years with data reported to the American Bar Association (ABA 509 disclosures). The ABA 509 disclosures provide information on law schools, encompassing graduation rates, bar passage rates, employment outcomes, student demographics, tuition, and financial aid.

Dataset used to construct Selective Ranking Dataset We constructed tiered rankings using data from 5 companies. Our final dataset covers X colleges. D_{tra}.

Law School	LSD	Velocity	QS	Above The Law	Cleo
Stanford	1	2	3	8	8
Yale	1	1	2	6	7
Harvard	3	3	1	9	1
Columbia	4	4	5	13	2
University of Chicago	4	4	13	1	10
NYU	6	6	4	16	4
UPenn	6	6	9	7	18
UVA	8	8	13	2	5
UC Berkeley	9	9	6	11	14
Duke University	10	10	10	3	9
University of Michigan	10	10	11	5	16

Table 2: Ranking of Law Schools with in Tier 1

Reverse-Engineered Score

Results Given a set of schools I and specific gameable features J_G, we measured the overall cost of manipulation as $\Delta^{\text{total}} = \sum_{i \in I} \sum_{j \in J} (x'_{ij} - x_{i,j})$ where $x'_{i,j}$ and $x_{i,j}$ and values of feature j before and after manipulation.

We consider how much each school can improve its ranking by at least one position. For features where no amount of modification lifted the school's rank sufficiently, $x'_{ij} - x_{ij}$ was set to 0. The change for each feature to increase each school's rank 5 places or ascend to the top rank was \$130.80.

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.

	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 3: 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 τ, 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.