Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ranking Abuse via Strategic Pairwise Data Perturbations

Junyi Yao

j.yao@wustl.edu

Washington University in St. Louis

Zihao Zheng

Washington University in St. Louis

Jiayu Long

Washington University in St. Louis

Research Article

Keywords: Pairwise ranking, Bradley–Terry model, adversarial manipula- tion, social choice, ranking robustness

Posted Date: April 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9459163/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Ranking Abuse via Strategic Pairwise Data Perturbations

Junyi Yao[1] Zihao Zheng[1] Jiayu Long[1] Ying Xu[1]

[1]Department of Computer Science and Engineering,
Washington University in St. Louis, USA
j.yao@wustl.edu, z.zihaogary@wustl.edu,
jiayujacqueline@wustl.edu, x.ying1@wustl.edu

Abstract

Pairwise ranking systems based on Maximum Likelihood Estimation (MLE), such as the Bradley-Terry model, are widely used to aggregate preferences from pairwise comparisons. However, their robustness under strategic data manipulation remains insufficiently understood.

In this paper, we study the vulnerability of MLE-based ranking systems to adversarial perturbations. We formulate the manipulation task as a constrained combinatorial optimization problem and propose an Adaptive Subset Selection Attack (ASSA) to efficiently identify high-impact perturbations.

Experimental results on both synthetic data and real-world election datasets show that MLE-based rankings exhibit a sharp phase-transition behavior: beyond a small perturbation budget, a limited number of strategic voters can significantly alter the global ranking. In particular, our method consistently outperforms random and greedy baselines under constrained budgets.

These findings reveal a fundamental sensitivity of MLE-based ranking mechanisms to structured perturbations and highlight the need for more robust aggregation methods in collective decision-making systems.

Keywords: Pairwise ranking, Bradley–Terry model, adversarial manipulation, social choice, ranking robustness

1 Introduction

Social choice theory centers on the structural aggregation of individual preferences to facilitate collective consensus. While traditionally character-

ized by theoretical abstractions, modern implementations—spanning from ranked-choice voting in public governance to collective decision-making systems, including voting processes and online preference aggregation platforms—confront the intensifying dual challenges of ensuring algorithmic fairness and maintaining resilience against strategic manipulation.

A prominent statistical paradigm for inferring collective preferences is the utilization of Random Utility Models (RUMs). By formalizing the utility of discrete options as random variables, RUMs effectively accommodate the stochastic noise and heterogeneous nature of agent behavior. Within this domain, Maximum Likelihood Estimation (MLE), specifically when integrated with the Bradley-Terry model, has emerged as a cornerstone methodology for synthesizing global rankings from pairwise comparison datasets. Its capacity to extrapolate a comprehensive preference hierarchy from sparse, head-to-head observations renders it an indispensable tool for large-scale decision-making infrastructures.

However, the inherent sensitivity of MLE-based estimators introduces profound adversarial vulnerabilities. The integrity of both democratic institutions and organizational decision-making frameworks rests on the presupposition that input data authentically reflects agent preferences. Yet, the non-linear dynamics of the MLE objective function imply that strategic agents—acting either individually or in collusion—can perturb their inputs to disproportionately distort the estimated utility parameters. Such strategic exploitation fundamentally undermines the core tenets of representativeness and distributive justice in social choice.

In this paper, we conduct a rigorous vulnerability analysis of MLE-based ranking frameworks, contextualizing our work within the established literature of axiomatic social choice and pairwise statistical inference [1]. We formalize strategic preference interference as a combinatorial subset selection problem and introduce the Adaptive Subset Selection Attack (ASSA). ASSA is an innovative heuristic algorithm designed to identify high-impact data perturbations. By leveraging an adaptive mechanism to isolate high-influence manipulation subsets, ASSA achieves superior alignment between the synthesized global ranking and adversarial targets, surpassing the efficacy of traditional greedy or stochastic baseline strategies.

Our empirical evaluation, substantiated by extensive synthetic simulations and validated through real-world electoral data from Berkeley and the UK Labor Party, reveals a critical structural fragility in these systems. Specifically, we identify a distinct phase-transition phenomenon: beyond a modest perturbation budget, the global ranking undergoes a precipitous shift toward the adversarial target. This research elucidates the disproportion-

ate influence of a small, informed cohort in sparse electorate environments, providing pivotal theoretical and empirical foundations for the development of the next generation of robust, manipulation-resistant voting protocols. Beyond algorithmic considerations, our findings have broader implications for the integrity of collective decision-making systems. The results suggest that widely used statistical aggregation methods may exhibit systemic vulnerabilities under coordinated strategic behavior, raising concerns about fairness, robustness, and trust in computational social systems.

2 System Model and Problem Formulation

This section provides a rigorous mathematical formalization of the Bradley-Terry ranking framework and defines the strategic manipulation task as a constrained combinatorial optimization problem. We analyze the system's vulnerability by investigating how a targeted perturbation of input data influences the resulting Maximum Likelihood Estimation (MLE) of latent parameters.

2.1 Preliminaries and Notations

We consider a set of m candidates denoted by $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ and a set of n voters $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$. The observed preferences are represented as a pairwise comparison dataset D, where each entry $d \in D$ is a triplet (v, c_i, c_j) indicating that voter v prefers c_i over c_j.

The key notations used throughout this paper are summarized in Table 1.

2.2 The Bradley-Terry Framework

Following the statistical decision-theoretic framework for social choice [2], we model the pairwise comparison between candidates c_i and c_j using the Bradley-Terry model. Each candidate c_i is associated with a positive strength parameter $p_i > 0$. The probability that c_i is preferred over c_j is given by the logistic-link function:

$$P(c_i > c_j) = \frac{p_i}{p_i + p_j}. \tag{1}$$

Given the dataset D, we define n_{ij} as the aggregate number of times c_i was preferred over c_j across all voters. The goal of the ranking system is to

Table 1: Summary of Key Notations

Symbol	Description
$\mathcal{C}, \mathcal{V}$	Sets of candidates and voters, respectively
p_i	Latent strength parameter of candidate c_i
$\mathbf{p}$	Vector of strength parameters $[p_1, \ldots, p_m]^T$
n_{ij}	Total counts of c_i defeating c_j in dataset D
$\tau(\mathbf{p})$	Global ranking derived from the sorted vector $\mathbf{p}$
Δ	Set of adversarial perturbations (flipped pairs)
k	Perturbation budget (maximum allowed flips)
$K_d(\cdot, \cdot)$	Kendall Tau distance metric
$\mathcal{L}(\mathbf{p}\|D)$	Log-likelihood function of the Bradley-Terry model

find the parameter vector $\mathbf{p}^*$ that maximizes the log-likelihood function:

$$\mathcal{L}(\mathbf{p}|D) = \sum_{1 \le i < j \le m} \left[n_{ij} \ln \left(\frac{p_i}{p_i + p_j} \right) + n_{ji} \ln \left(\frac{p_j}{p_i + p_j} \right) \right]. \tag{2}$$

To ensure the identifiability of the parameters, a common constraint is applied, such as $\sum p_i = 1$. The resulting global ranking $\tau(\mathbf{p}^*)$ is obtained by sorting the candidates in descending order of their estimated strengths. This MLE-based approach is known to be sensitive to the structural properties (e.g., spectral gap) of the comparison graph [**?**].

2.3 Adversarial Manipulation as an Optimization Problem

We formalize the strategic manipulation task from a white-box perspective. An adversary seeks to modify a subset of the dataset D to produce a manipulated ranking τ_m that minimizes the distance to a target preference τ_t.

2.3.1 Objective Function

Let $\Delta \subset D$ be the set of pairwise comparisons that the adversary decides to "flip". The modified dataset is $D' = D \oplus \Delta$. The adversarial objective is:

$$\min_{\Delta} \quad K_d(\tau_t, \tau(\mathbf{p}^*(D \oplus \Delta))) \tag{3}$$

2.3.2 Constraints

The optimization problem in (3) is subject to:

1. **Budget Constraint:** $|\Delta| \leq k$.
2. **Data Integrity:** Δ must only involve the target voter's or the coalition group's existing votes.
3. **Model Consistency:** The ranking τ must be derived from the standard MLE process [3].

This formulation characterizes the manipulation as a **constrained combinatorial subset selection problem**. Due to the discrete nature of flips and the non-linearity of MLE, the objective (3) is non-differentiable, necessitating heuristic strategies [4].

3 Strategic Manipulation Framework

We formalize preference manipulation as a **constrained combinatorial subset selection problem**. This framing aligns with contemporary research in adversarial preference learning, which highlights the structural vulnerability of pairwise comparison models to strategic data injection [4].

3.1 Baseline Heuristics

To establish a performance floor, we define two fundamental heuristics:

- **Random Flip (RF):** A stochastic control where a random subset of comparisons is inverted, establishing a baseline for unstructured interference.
- **Greedy Flip (GF):** A local search strategy that iteratively selects flips providing the maximum immediate reduction in K_d. While intuitive, GF often converges to sub-optimal local extrema due to the non-convex nature of the ranking landscape.

3.2 Randomized Subset Attack (RSA)

The Randomized Subset Attack (RSA) serves as a stochastic search precursor to our main contribution. RSA explores the perturbation space by partitioning the adversary's available comparisons into b random subsets. Unlike a pure brute-force approach, RSA iteratively evaluates the collective

impact of flipping these subsets, retaining only those perturbations that yield a monotonic reduction in the Kendall Tau distance relative to the adversarial target.

Algorithm 1 Randomized Subset Attack (RSA)

procedure RSA($\mathcal{D}, \tau_t, b, n$)
  $\mathcal{D}^* \leftarrow \mathcal{D}$
  $cur_dist \leftarrow K_d(\tau_t, \text{MLE}(\mathcal{D}^*))$
  for $i = 1$ **to** n **do**
    $\mathcal{S} \leftarrow \{d \in \mathcal{D}^* \mid \text{voter } v \in \text{Coalition}\}$
    $\mathcal{P} \leftarrow$ Shuffle and divide $\mathcal{S}$ into b subsets $\{\mathcal{S}_1, \ldots, \mathcal{S}_b\}$
    for each subset $\mathcal{S}_j \in \mathcal{P}$ **do**
      $\mathcal{D}_{cand} \leftarrow \mathcal{D}^* \oplus \text{flip}(\mathcal{S}_j)$
      $new_dist \leftarrow K_d(\tau_t, \text{MLE}(\mathcal{D}_{cand}))$
      if $new_dist < cur_dist$ **then**
        $cur_dist \leftarrow new_dist$
        $\mathcal{D}^* \leftarrow \mathcal{D}_{cand}$
      end if
    end for
  end for
  return $\mathcal{D}^*$
end procedure

3.3 Adaptive Subset Selection Attack (ASSA)

To transcend the stochastic inefficiencies of RSA, we propose the **Adaptive Subset Selection Attack (ASSA)**. While RSA treats every iteration as an independent shuffle, ASSA introduces an evolutionary refinement process. By classifying subsets into an "Effective Pool" (C_{good}) and an "Ineffective Pool" (C_{bad}), the algorithm dynamically prunes the search space over successive iterations.

This mechanism exploits the non-linear sensitivity of MLE parameters to specific data clusters. By iteratively narrowing the focus to high-influence manifolds, ASSA demonstrates that even a sparse electorate can be disproportionately swayed by a small cohort of strategic voters using targeted perturbations.

Algorithm 2 Adaptive Subset Selection Attack (ASSA)

Input: Dataset $\mathcal{D}$, target ranking τ_t, budget k, iterations T, subset factor b
Output: Manipulated dataset $\mathcal{D}^*$
Initialize $\mathcal{D}^{(0)} \leftarrow \mathcal{D}$, $\mathcal{S} \leftarrow$ Perturbation candidates
for $t = 1$ **to** T **do**
    Partition $\mathcal{S}$ into b disjoint subsets $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_b\}$
    $C_{good} \leftarrow \emptyset$
    for each subset $\mathcal{S}_j$ **do**
        $\mathcal{D}_{tmp} \leftarrow \mathcal{D}^{(t-1)} \oplus \text{flip}(\mathcal{S}_j)$
        $\hat{\mathbf{p}}^{(t)} \leftarrow \arg\max_{\mathbf{p}} \mathcal{L}(\mathbf{p} \mid \mathcal{D}_{tmp})$ ▷ MLE Estimation
        $\Delta\mathcal{L}_j \leftarrow K_d(\tau_t, \tau(\hat{\mathbf{p}}^{(t)})) - K_d(\tau_t, \tau(\hat{\mathbf{p}}^{(t-1)}))$
        if $\Delta\mathcal{L}_j < 0$ **then**
            $\mathcal{D}^{(t)} \leftarrow \mathcal{D}_{tmp}$
            $C_{good} \leftarrow C_{good} \cup \mathcal{S}_j$
        end if
    end for
    $\mathcal{S} \leftarrow C_{good}$ ▷ Adaptive search space pruning
    if $|\Delta\mathcal{D}| \geq k$ **then break**
end for
return $\mathcal{D}^* = \mathcal{D}^{(T)}$

3.4 Theoretical Justification via Implicit Function Theorem

The efficacy of ASSA can be analyzed through the lens of the Implicit Function Theorem (IFT). Let $\mathcal{L}(\mathbf{p}, \mathcal{D})$ be the log-likelihood function of the Bradley-Terry model. The MLE estimate $\hat{\mathbf{p}}$ satisfies the first-order optimality condition:

$$\nabla_{\mathbf{p}} \mathcal{L}(\hat{\mathbf{p}}, \mathcal{D}) = 0 \tag{4}$$

When a subset of data Δ is perturbed, the change in the estimated parameters $\delta\mathbf{p}$ can be approximated by considering the total derivative of the optimality condition:

$$\nabla^2_{\mathbf{p}} \mathcal{L} \cdot \delta\mathbf{p} + \frac{\partial \nabla_{\mathbf{p}} \mathcal{L}}{\partial \mathcal{D}} \cdot \delta\mathcal{D} \approx 0 \tag{5}$$

Rearranging, we obtain the influence of a perturbation on the parameter vector:

$$\delta\mathbf{p} \approx -[\mathbf{H}(\hat{\mathbf{p}})]^{-1} \left(\frac{\partial \nabla_{\mathbf{p}} \mathcal{L}}{\partial \mathcal{D}} \delta\mathcal{D} \right) \tag{6}$$

where $\mathbf{H}(\hat{\mathbf{p}})$ is the Hessian matrix of the log-likelihood.

ASSA effectively approximates this gradient-based influence by iteratively selecting subsets $\delta\mathcal{D}$ that align $\delta\mathbf{p}$ with the target direction. The observed "phase-transition" in ranking stability corresponds to the regime where the perturbation $\delta\mathcal{D}$ is large enough to shift $\hat{\mathbf{p}}$ across a decision boundary in the permutahedron (the space of all possible rankings).

3.5 Computational Complexity and Scalability

A critical advantage of the Bradley-Terry MLE framework in large-scale systems is its data-invariant parameter estimation. For a set of m candidates and N pairwise comparisons, the log-likelihood $\mathcal{L}(\mathbf{p} \mid \mathcal{D})$ depends only on the aggregate counts n_{ij}, forming an $m \times m$ adjacency matrix.

Consequently, the complexity of each iteration in ASSA is $O(b \cdot \text{Cost}_{\text{MLE}})$, where Cost_{MLE} is the cost of solving the m-dimensional optimization problem (typically $O(m^3)$ or $O(I \cdot m^2)$ using iterative methods). Notably, the computation primarily depends on aggregated pairwise counts rather than the raw number of comparisons, allowing ASSA to scale seamlessly to datasets involving millions of pairwise comparisons, provided the number of candidates m remains computationally tractable.

4 Experimental Results and Analysis

This section evaluates the performance of the proposed RSA and ASSA algorithms against baseline manipulation strategies. We analyze the sensitivity of MLE-based rankings to adversarial budgets and investigate the impact of hyperparameter configurations on attack efficacy.

4.1 Impact of Adversarial Budget on Ranking Stability

The robustness of the Bradley-Terry MLE framework is assessed by varying the perturbation budget k, defined as the fraction of modifiable pairwise comparisons. We evaluate four strategies: Random Flip, Greedy Flip, RSA, and ASSA across $k \in \{1\%, 5\%, 10\%, 20\%\}$.

As illustrated in Fig. 1, the **Random Flip** strategy yields negligible improvements, occasionally even increasing the Kendall tau distance due to unstructured noise. In contrast, the proposed **ASSA** algorithm consistently outperforms all baselines, particularly under constrained budgets ($k \leq 5\%$).

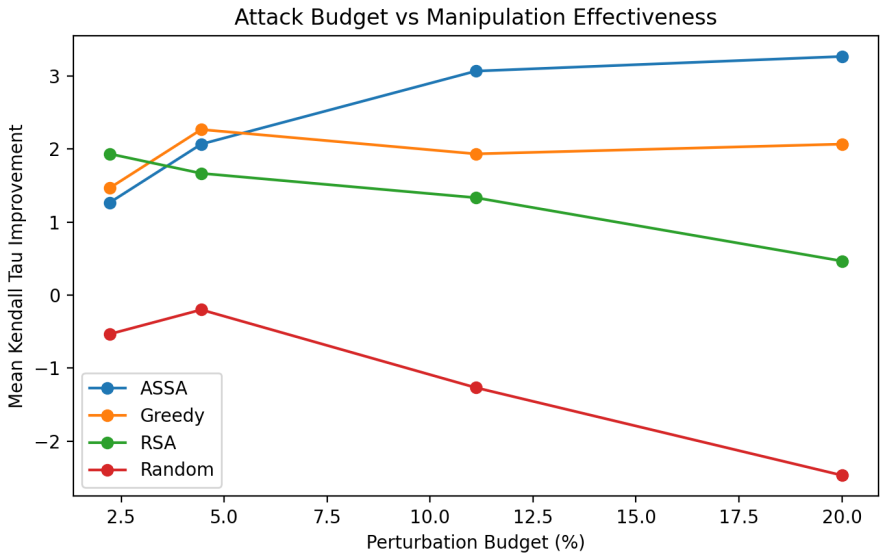


Figure 1: Effect of adversarial budget on manipulation effectiveness (Kendall Tau reduction). Structured attacks (RSA and ASSA) achieve significantly higher improvement compared to random and greedy baselines.

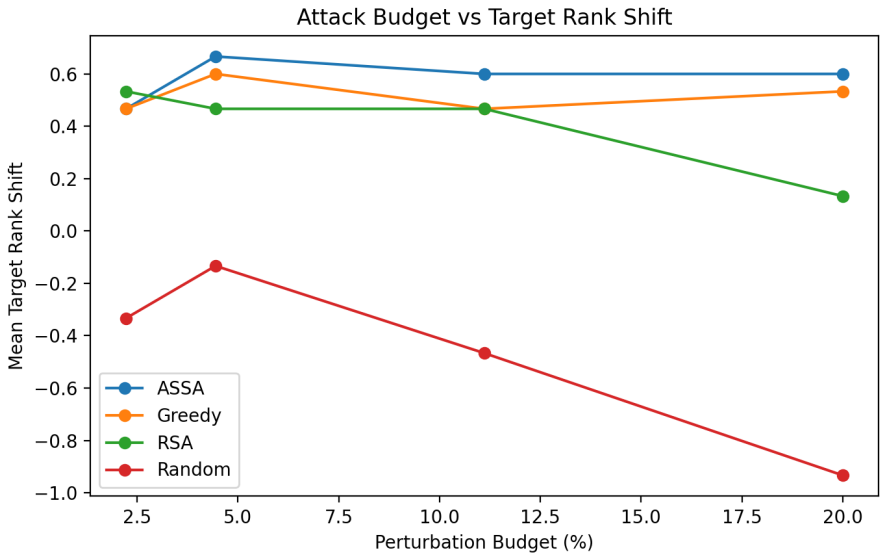


Figure 2: Target candidate rank shift under varying perturbation budgets. Higher values indicate more successful manipulation toward the desired ranking.

Fig. 2 depicts the rank shift of the target candidate. We observe a distinct **non-linear response** to the perturbation budget. Specifically, a "phase-transition" phenomenon occurs: beyond a critical threshold of modified pairs, the target candidate's rank ascends precipitously. This indicates a structural fragility in the Bradley-Terry model.

4.2 Hyperparameter Sensitivity Analysis

We further investigate the influence of the number of subsets b and iterations n on the manipulation success rate.

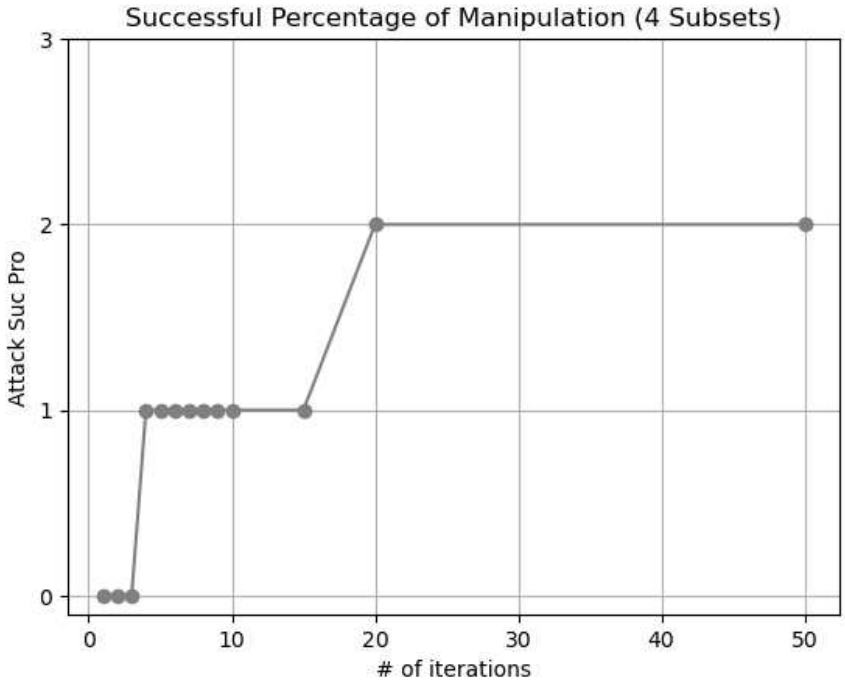


Figure 3: Impact of iteration count n on manipulation success rate under ASSA.

4.2.1 Convergence and Iteration Impact

Fig. 3 illustrates the relationship between the number of iterations n and the success rate. We observe that both the success rate and the reduction in Kendall tau distance exhibit diminishing marginal returns. The algorithm

typically reaches an optimal state within a relatively low iteration threshold (e.g., $n \approx 50$), highlighting the efficiency of ASSA.

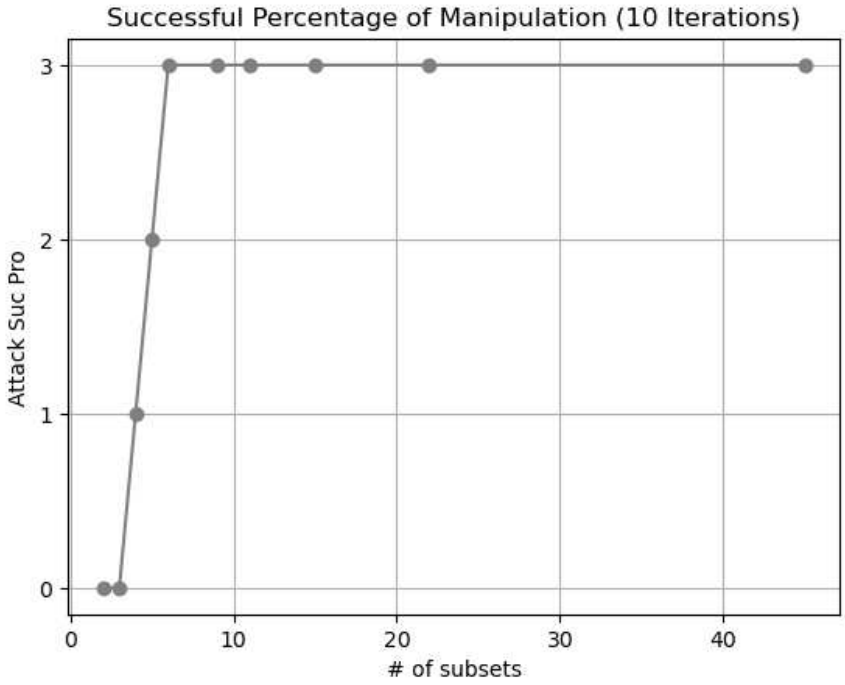


Figure 4: Impact of subset count b on manipulation success rate.

4.2.2 Subset Granularity

The impact of subset count b is shown in Fig. 4. While increasing b allows for more granular control, excessively large values lead to a "fragmentation effect," where individual subsets lack sufficient collective influence to alter the global MLE landscape.

4.3 Optimization for Optimal KT Distance

In Fig. 5 and Fig. 6, we further examine the algorithm's ability to reach the theoretically minimum Kendall Tau distance (the optimal manipulation). Our findings indicate that both parameters reach a stability threshold relatively early, suggesting that the ASSA heuristic is robust to parameter selection within a wide operational range.

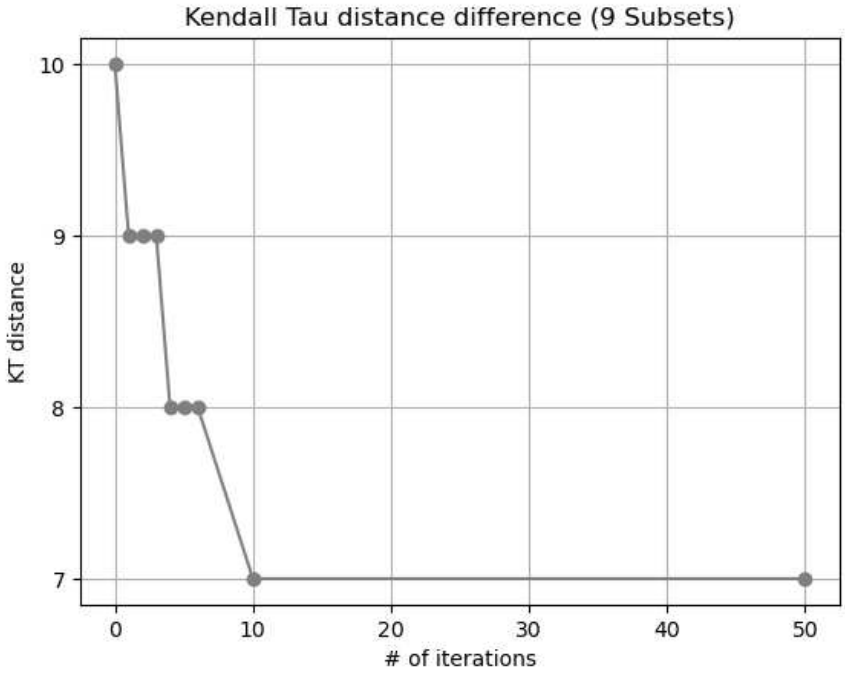


Figure 5: Impact of iteration count n on achieving optimal Kendall Tau distance.

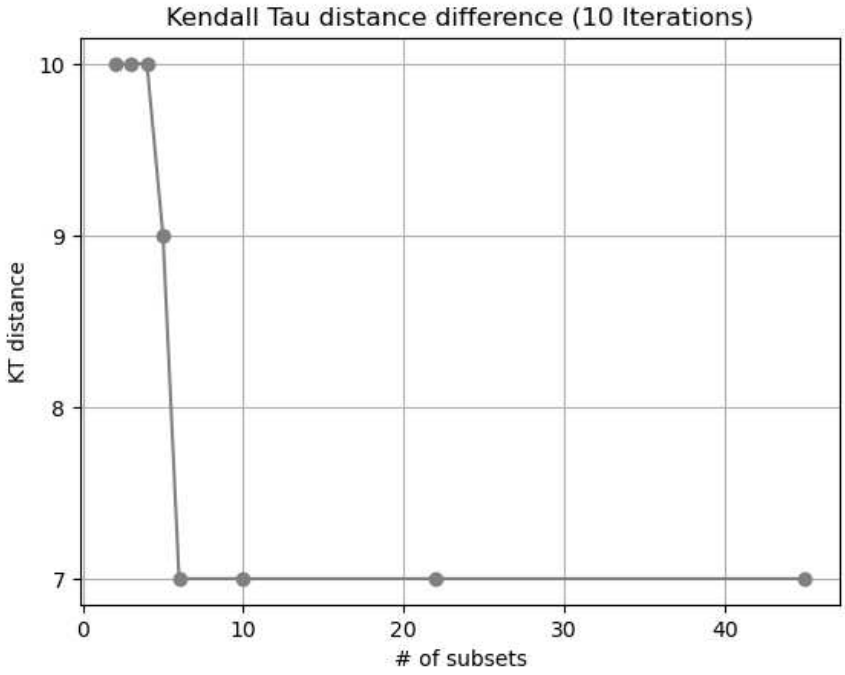


Figure 6: Impact of subset count b on achieving optimal Kendall Tau distance.

4.4 Real-World Case Studies

4.4.1 UK Labor Party Dataset

The UK Labor Party leadership vote dataset provides a unique opportunity to evaluate the algorithm's performance in an environment characterized by incomplete preference rankings. We observed that incomplete ballots inherently increase the system's susceptibility to manipulation, as the MLE estimator operates under fewer constraints to maintain the global ranking's stability.

By applying the ASSA framework with our proposed complementary ranking strategy, we successfully identified critical preference gaps. Our results indicate that strategic voters can exploit these gaps to disproportionately influence the final parameter estimation of the Bradley-Terry model. This finding underscores the necessity for more robust data-imputation protocols in digital voting systems to mitigate the risk of adversarial interference derived from sparse data [5].

4.4.2 Berkeley Election Dataset

In the District 7 Berkeley City Council election dataset ($|V| = 4,189$ voters), we simulated the impact of collusive voter behavior. The initial MLE prediction aligns with the historical outcome: $Kriss\,Worthington > George\,Beier > Ces\,Rosales > Jesse\,Arreguin$.

Our experimental analysis quantified the **collusion threshold** required to subvert the election outcome. Specifically, for a voter group targeting a preference shift with an initial Kendall distance of $K_d = 2$ (e.g., Target: $[3, 2, 1, 4]$), we determined that a coalition of 353 strategic voters—approximately 8.4% of the total electorate—is sufficient to flip the global ranking in favor of the adversarial target. When the manipulation objective is more aggressive (initial $K_d = 3$), the required coalition size increases to 465 voters (11.1%). These results demonstrate that even in large-scale democratic processes, a relatively small, informed cohort can strategically sway the collective outcome, highlighting a fundamental structural fragility in standard MLE-based ranking mechanisms.

5 Related Work

5.1 Pairwise Comparison and Probabilistic Ranking

The interpretation of pairwise comparison data is a foundational challenge in computational social choice. Noothigattu et al. [1] investigated the axiomatic properties of machine learning models in this domain, demonstrating that while Maximum Likelihood Estimators (MLE) like the Thurstone-Mosteller model achieve Pareto efficiency, they may diverge from majority consensus under specific noise distributions. Further axiomatic analysis by González-Díaz et al. validated the Bradley-Terry model as a robust framework for general tournaments, emphasizing its computational viability in practical voting systems. Meanwhile, recent advances in Bradley-Terry and pairwise comparison models have focused on efficient estimation and generalization [6, 7], but their robustness under adversarial manipulation remains underexplored.

5.2 MLE and Statistical Social Choice

The conceptualization of voting rules as maximum likelihood estimators, popularized by Conitzer and Sandholm , provides a probabilistic lens to assess the "correctness" of electoral outcomes. By deriving optimal estimators from hypothesized noise models, this line of inquiry illuminates how statistical assumptions can either mirror or distort the collective will. However, the reliance on MLE for parameter estimation introduces significant vulnerabilities. Recent research by Liu and Mueller suggests that MLE frameworks are particularly fragile in sparse preference graphs, where limited information allows for disproportionate parameter shifts. Our work builds upon this statistical foundation, focusing on the deliberate manipulation of these parameter estimation processes.

5.3 Robustness and Manipulation in Voting Systems

The vulnerability of voting mechanisms to strategic manipulation has led to the development of robust voting rules through the lens of algorithmic statistics [8]. Expanding on this, Zhao and Xia [5] explored the learning of mixtures of Random Utility Models (RUMs) from incomplete preferences, offering methodologies to secure preference models against strategic misrepresentation. Despite these defensive strides, the computational challenges of implementing social choice rules under complex models like Mallows [2] persist. Contemporary studies in 2024 and 2025 have begun to investigate

the role of AI-driven strategic agents in identifying these computational bottlenecks to facilitate large-scale manipulation. Recent work has also demonstrated the vulnerability of ranking systems under adversarial perturbations, particularly in neural ranking models and preference learning settings [**?**, 9, 10].

5.4 Adversarial Machine Learning in Ranking

While significant attention has been paid to defensive robustness, the algorithmic construction of adversarial attacks against MLE-based ranking systems has received relatively little scrutiny. Wang et al. [4] pioneered adversarial preference learning, focusing on the robustness of pairwise comparisons against data injection. In the context of networked systems, strategic perturbations have been shown to impact collective decision-making significantly. Our research complements this body of work by introducing the Adaptive Subset Selection Attack (ASSA), which addresses the gap between theoretical vulnerability and practical, scalable adversarial strategies in both synthetic and real-world electoral landscapes.

6 Conclusion

This research provides a comprehensive vulnerability analysis of pairwise comparison models under the Maximum Likelihood Estimation (MLE) framework. By introducing the Adaptive Subset Selection Attack (ASSA), we have demonstrated that strategic interventions can significantly distort collective preference hierarchies, even within constrained adversarial budgets. Our empirical findings reveal a critical non-linear response in ranking stability: manipulation effectiveness does not scale linearly with perturbation size, but rather exhibits distinct threshold effects and diminishing returns. This phase-transition behavior underscores a fundamental structural fragility in the Bradley-Terry model when faced with coordinated data perturbations. Ultimately, this work highlights the imperative need for integrating robust statistical safeguards into digital voting infrastructures to ensure that collective decisions authentically reflect the electorate's will.

7 Limitations

Despite the efficacy of the ASSA framework, several constraints bound the scope of this study. First, the heuristic nature of our subset selection mech-

anism, while computationally efficient, does not guarantee a theoretically optimal manipulation path in all preference landscapes. The stochastic elements involved in subset partitioning suggest that the algorithm's performance may vary under extreme data sparsity.

Furthermore, while validated against real-world datasets from Berkeley and the UK Labor Party, a significant portion of our granular analysis relies on synthetic preference models. These models assume a degree of voter transitivity and rationality that may not fully capture the idiosyncratic or noisy behaviors prevalent in large-scale, real-world social choice scenarios. The assumption of a white-box threat model, while essential for assessing upper-bound vulnerabilities, also overlooks the practical challenges an adversary would face in acquiring complete real-time preference data.

8 Future Work

The findings of this study open several promising avenues for future inquiry. To address current limitations in data representation, our subsequent research will focus on the acquisition of high-fidelity, longitudinal electoral datasets. We aim to model more complex voter dynamics, including dynamic preference shifts and the impact of misinformation on the initial preference formation process.

Moreover, we intend to expand our adversarial analysis beyond the Bradley-Terry framework to include alternative social choice mechanisms such as Approval Voting, Preferential Voting, and Cumulative Voting. Investigating the cross-model resilience of these systems will facilitate the development of a unified theory of algorithmic influence.

The ultimate objective of this trajectory is the design of certified robust ranking estimators. By leveraging adversarial training and robust optimization, we aim to engineer next-generation voting protocols that maintain fairness and representativeness even in the presence of sophisticated strategic actors, thereby upholding the integrity of democratic processes in the digital age.

We declare the following potential conflicts of interest:

9 Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] R. Noothigattu, D. Peters, and A. D. Procaccia, "Axioms for learning from pairwise comparisons," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 745–17 754.

[2] H. A. Soufiani, D. C. Parkes, and L. Xia, "A statistical decision-theoretic framework for social choice," in *Advances in Neural Information Processing Systems*, 2014, pp. 3185–3193.

[3] D. R. Hunter, "Mm algorithms for generalized bradley-terry models," *The Annals of Statistics*, vol. 32, no. 1, pp. 384–406, 2004.

[4] Z. Wang, Q. Xu, K. Ma, Y. Jiang, X. Cao, and Q. Huang, "Adversarial preference learning with pairwise comparisons," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 656–664.

[5] Z. Zhao and L. Xia, "Learning mixtures of plackett-luce models with features from top-l orders," *arXiv preprint arXiv:2006.03869*, 2020.

[6] M. E. J. Newman, "Efficient computation of rankings from pairwise comparisons," *Journal of Machine Learning Research*, vol. 24, 2023.

[7] J. Fageot *et al.*, "Generalized bradley-terry models for score estimation from paired comparisons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[8] A. Liu and A. Moitra, "Robust voting rules from algorithmic robust statistics," *arXiv preprint arXiv:2112.06380*, 2021.

[9] X. Chen, B. He *et al.*, "Towards imperceptible document manipulations against neural ranking models," *arXiv preprint arXiv:2305.01860*, 2023.

[10] A. Bigdeli *et al.*, "Empra: Embedding perturbation rank attack against neural ranking models," *arXiv preprint arXiv:2412.16382*, 2024.