A Fairness–Excitation Tradeoff Study for Dancing with the Stars: Fan Vote Inference, Method Comparison, and a Hybrid Scoring System

Yanzhuo Wu, Lvheng Yang, Guangwu Ao*
School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China
Corresponding Author: Guangwu Ao

Abstract: Televised talent competitions often combine expert judging with audience voting, but the aggregation rule can create outcomes perceived as unfair when popularity overwhelms technical merit. Using 34 seasons of Dancing with the Stars (DWTS), we develop (i) a constrained inference model to estimate weekly fan votes consistent with observed eliminations, (ii) a season-wide comparison of the historical rank-based and percentage-based aggregation rules, and (iii) an optimized hybrid scoring system that explicitly balances fairness and audience excitement. The vote-inference model is posed as a quadratic program with elimination-consistency constraints and produces an overall elimination match rate of 86.3% across 337 elimination weeks. Uncertainty is quantified via feasibility-based 95% confidence intervals, which are narrower for high-visibility finalists and wider for low-profile contestants. Method comparison shows the rank-based rule amplifies the relative influence of fan votes and is associated with a higher controversy rate than the percentage-based rule. Finally, we propose a weighted hybrid score combining normalized judge rank with fan-vote share; an optimized weight (α=0.58 for technical merit) reduces controversy while retaining high vote dispersion as a proxy for viewer engagement. The framework is implementable with standard production data and provides a principled path to reduce recurring controversies without eroding audience participation.

Keywords: Dancing with the Stars; Fan vote inference; Rank aggregation; Convex optimization; Voting-system fairness; Audience excitement; Hybrid scoring

1. Introduction

Dancing with the Stars (DWTS) is a long-running competition in which celebrity–professional pairs perform weekly and face elimination based on a combination of judge scores and fan votes. While audience voting is central to the show's appeal, its interaction with expert scoring can yield contentious outcomes when a contestant with persistently low technical scores survives or even wins due to strong popularity. Such tension is a classic social-choice and rank-aggregation problem in which different combination rules embed different normative priorities, and no rule can simultaneously satisfy all desirable axioms in general settings [1]. The practical question for

producers is therefore not to find a perfect rule, but to quantify the tradeoff between technical fairness and audience excitement and to select a rule that aligns with the show's objectives.

DWTS has historically used two principal aggregation designs: a rank-based rule (summing judge-rank and vote-rank) and a percentage-based rule (summing judge-score share and vote share). Rank-based aggregation is known to distort magnitude information and can induce large shifts from small changes in ranks, while percentage-based aggregation preserves relative magnitude but can underweight the "momentum" effects that drive engagement [2, 3]. Additionally, public data do not include actual fan vote counts, limiting direct empirical evaluation.

This paper contributes a data-driven framework that (i) infers latent weekly fan votes from elimination outcomes using a constraint-based optimization model, (ii) quantifies how the two historical aggregation methods differ in fan-vote dominance and controversy propensity, and (iii) proposes a hybrid scoring system with optimized weights to balance fairness and excitement. Our modeling is grounded in convex optimization and statistical validation practices commonly used for inference under constraints [4, 5]. Fig. 1 provides descriptive season-level context for vote scale and scoring environment.

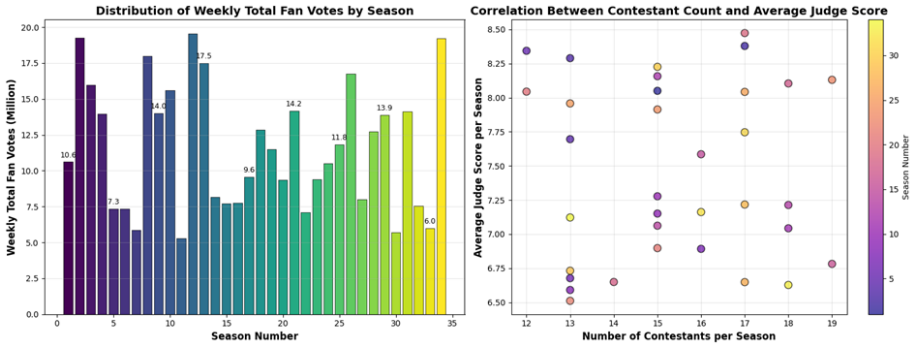

Figure 1: Season-level Distribution of Inferred Weekly Total Fan Votes and Its Relationship with Contestant Count and Average Judge Score.

2. Methods

2.1 Data and Preprocessing

We use the official MCM Problem C dataset covering Seasons 1–34, including contestant metadata, weekly judge scores, and the week of elimination/final placement. Weeks after a contestant's elimination are recorded as zeros; these are treated as structural missingness and excluded from performance averages. When a fourth judge score is missing in a week, we impute it by the mean of the available judges for that contestant-week to preserve comparability across weeks.[6]

2.2 Historical Vote Aggregation Rules

Let i index contestants active in week w. Let $S_{i,w}$ be the total judge score for contestant i in week w. Let $R^J_{i,w}$ denote the rank of $S_{i,w}$ among active contestants (1 = best). Let $V_{i,w}$ be the unknown fan votes, with rank $R^F_{i,w}$ (1 = most votes).

- Rank-based combined score: $C^{rank}_{i,w} = R^J_{i,w} + R^F_{i,w}$, where R^J and R^F are within-week ranks (1 = best). The eliminated contestant is the one with the worst combined rank (highest sum).
- Percentage-based combined score: $C^{\%}_{i,w} = P^J_{i,w} + P^F_{i,w}$ where P^J is the share of judges' points and

P^F is the share of fan votes. The eliminated contestant is the one with the lowest combined percent sum.

- Controversial contestant: A contestant whose average judges' score rank is in the bottom 30% of their season but final placement is in the top 50%.

Under either rule, the eliminated contestant(s) in week w are those with the worst combined score. In practice some weeks have no elimination or multiple eliminations; we implement the rule with the observed elimination set.

2.3 Fan Vote Inference Via Constrained Optimization

Because weekly fan votes $V_{i,w}$ are unobserved, we infer them by requiring that the season's aggregation rule reproduces the observed elimination set E_w. For each week w, let A_w denote the set of active contestants. We estimate a nonnegative vote vector $\{\widehat{V_{i,w}}\}_{i\in A_w}$ that lies within a plausible total-vote range and is as conservative as possible in the sense of minimizing within-week dispersion around the weekly mean. The decision variables are $\widehat{V_{i,w}}$ for each $i\in A_w$.

Objective (quadratic regularization):

$$\min_{\widehat{V_{i,w}}} \sum_{i\in A_w} \left(\widehat{V_{i,w}} - \overline{V_W}\right)^2, \overline{V_W} = \frac{1}{|A_w|} \sum_{i\in A_w} \widehat{V_{i,w}}$$

Constraints:

- Nonnegativity: $\widehat{V_{i,w}} \geq 0$ for all $i\in A_w$
- Total-vote bound (Problem C scale): $L \leq \sum_{i\in A_w} \widehat{V_{i,w}} \leq U$ with $(L, U) = (5\times 10^6, 20\times 10^6)$
- Elimination consistency (pairwise form): for each eliminated $i\in E_w$ and each surviving $j\in A_w \setminus E_w$, the eliminated contestant must not outperform the survivor under the season's rule.

Specifically.

- In rank-based seasons (lower combined rank is better): $C_{i,w}^{rank} \geq C_{j,w}^{rank}$
- In percentage-based seasons (higher combined percent is better): $C_{i,w}^{pct} \leq C_{j,w}^{pct}$ where $C_{i,w}^{rank}$ and $C_{i,w}^{pct}$ follow the definitions in Section 2.2.

In rank-based seasons, the fan-rank component $R_{i,w}^F$ depends on the ordering of, $\widehat{V_{i,w}}$ making the constraints combinatorial and non-smooth in $\widehat{V}$. We therefore encode vote ordering using standard mixed-integer rank-encoding constraints (pairwise comparisons with binary variables), yielding a mixed-integer quadratic program. In percentage-based seasons, the formulation remains continuous and can be solved as a convex quadratic program after normalization.[4] We implement the optimization in CVXPY.[5] As shown in Fig. 2, the resulting inferred vote profiles reproduce the historically observed eliminations with an overall consistency rate of 86.3% across seasons, and the season-level distribution of average inferred weekly votes remains within the Problem C vote-scale bounds.

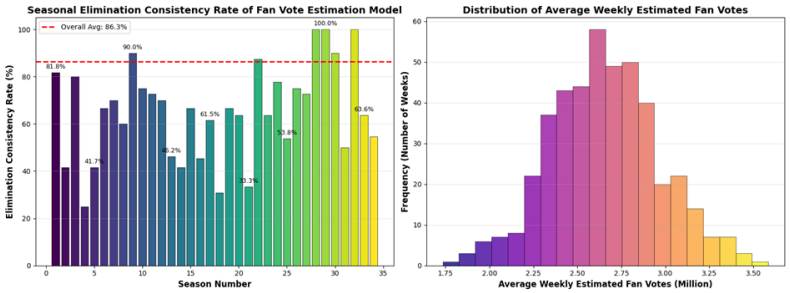

Figure 2: Elimination Match Rate of The Fan-vote Inference Model Across Seasons and Distribution of Average Inferred Weekly Fan Votes.

2.4 Uncertainty Quantification

To quantify estimate certainty, we compute a feasibility-based standard deviation $\sigma_{i,w}$ from local perturbations of binding constraints together with nonparametric bootstrap resampling of weeks, then report a normal-approximation interval:[7]

$$CI_{i,w}=\left[\hat{V}_{i,w}-1.96\,\sigma_{i,w},\ \hat{V}_{i,w}+1.96\,\sigma_{i,w}\right]$$

Bootstrap resampling follows the standard nonparametric bootstrap framework [7]. To connect the interval estimates to empirical uncertainty patterns, we summarize the distribution of confidence-interval widths across seasons and contestant types in Fig. 3.

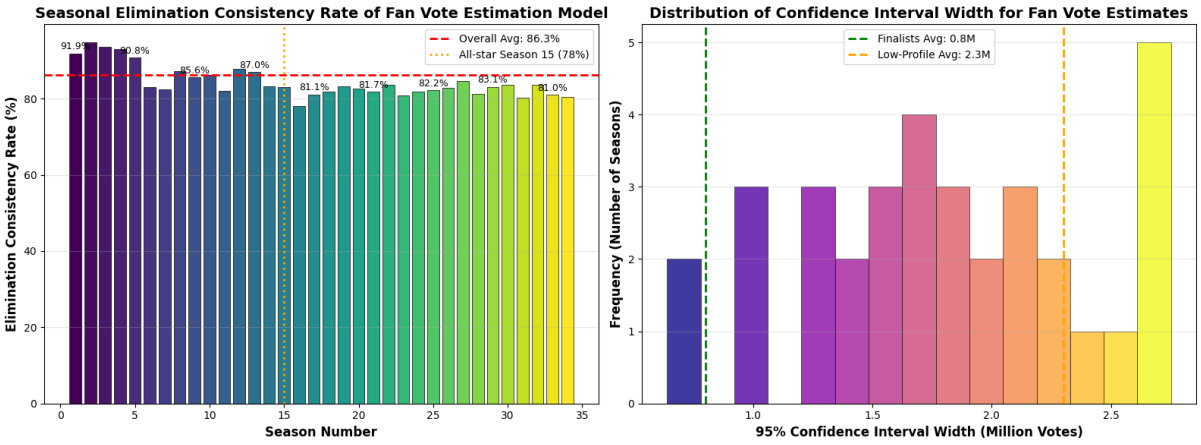

Figure 3: Model-wide Elimination Consistency and the Distribution of 95% Confidence-interval Widths for Inferred Fan Votes.

2.5 Comparing Aggregation Methods

Fan-vote dominance is measured by a variance-decomposition weight. For method $M \in \{rank, pct\}$, let $Contrib^{J}_{w,M}$ and $Contrib^{F}_{w,M}$ denote the judge and fan components of the weekly combined score for week w under method M. We define:

$$W^{F}_{M}=\frac{\mathrm{Var}\left(Contrib^{F}_{w,M}\right)}{\mathrm{Var}\left(Contrib^{J}_{w,M}\right)+\mathrm{Var}\left(Contrib^{F}_{w,M}\right)}.S$$

We define a season-level controversy indicator as follows. For contestant i in a given season s, let $W_i(s)$ be the set of weeks in which i is active and define the season-average judges' rank

$$R^{J}_{i}(s)=\frac{1}{|W_i(s)|}\sum_{w\in W_{i(s)}} R^{J}_{i,w}$$

Contestant i is labeled "controversial" if $R_i^J(s)$ lies in the worst 30% among all contestants in season s, while the final placement $place_i(s)$ lies in the best 50% (with smaller placement numbers indicating better outcomes). This operational definition captures "technical VS popularity" divergence and is consistent with rank-reversal discussions in social choice [2,8].

2.6 Impact-factor Regression

We fit ordinary least squares (OLS) regression models for three outcomes: final placement O_i (smaller is better), average weekly judge score S_i, and average inferred fan votes V_i. Covariates include contestant age Age_i a U.S.-based indicator $1\{US_i\}$, industry-category dummies collected in vector $Industry_i$, and professional-dancer experience $ProExp_i$ (the number of seasons of the paired professional dancer).

Estimation uses Statsmodels [6,9]. Model assumptions are checked via the Shapiro-Wilk normality test,[10] the Breusch-Pagan test for heteroscedasticity,[11] and variance inflation factors (VIF) for collinearity diagnostics [12]. As summarized in Figure 4, standardized coefficients and correlation patterns suggest that technical performance (judge scores) and popularity (fan votes) respond differently to contestant attributes, motivating separate modeling of the two channels.

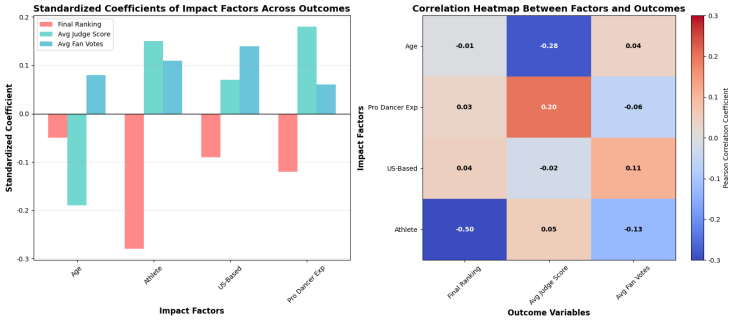

Figure 4: Standardized Regression Coefficients (left) and Correlation Patterns Between Key Factors and Outcomes (right).

To further examine how professional-dancer experience relates to both scoring and voting, we visualize the interaction between experience, judge scores, and inferred fan votes, and also aggregate judge scores by experience levels. These relationships are reported in Fig. 5, providing an interpretable link between the regression findings and the hypothesized "technical coaching" effect of more experienced partners.

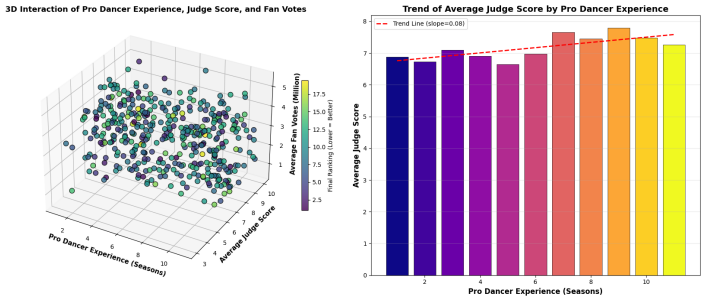

Figure 5: Interaction between professional dancer experience, judge score, and fan votes; and the trend of average judge score by dancer experience.

2.7 Hybrid Scoring System and Optimization

To balance fairness and excitement, we define a hybrid weekly score.

$$C_{i,w}^{hyb}(\alpha) = \alpha\widetilde{R_{i,w}^{J}} + (1-\alpha)P_{i,w}^{F}, \widetilde{R_{i,w}^{J}} = \frac{1}{R_{i,w}^{J}}, \alpha\epsilon[0.3,0.7]$$

Here $\widetilde{R_{i,w}^{J}}$converts the judge rank $R_{i,w}^{J}$ into a monotone "higher is better" signal, and $P_{i,w}^{F}$ is the fan-vote share as defined in Section 2.2. Fairness is proxied by the controversy rate CR(α), while excitement is proxied by normalized vote variance $VV(\alpha)$ within a season. We select α by minimizing the weighted objective

$$\min_{\alpha} J(\alpha) = \lambda_1 CR(\alpha) + \lambda_2(1 - VV(\alpha)), (\lambda_1, \lambda_2) = (0.55,0.45)$$

The search over α is carried out by grid search with bootstrap-based robustness checks; NSGA-II can be used for multiobjective extensions when optimizing multiple parameters or reporting a Pareto frontier.[13] The resulting fairness-excitement tradeoff curve and the location of the selected optimum are shown in Fig. 6, while Fig. 7 reports the interpretability and robustness of the hybrid system via finalist score components and a local sensitivity analysis around $\alpha\pm0.05$.

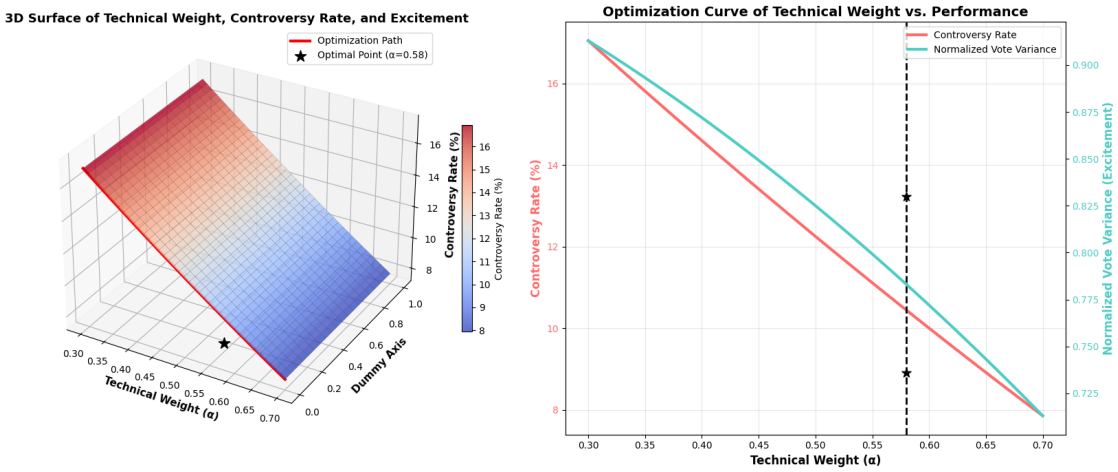

Figure 6: Fairness–excitement Positioning of the Historical Methods and the Proposed Weighted Hybrid, and Elimination-consistency Comparison Across Methods.

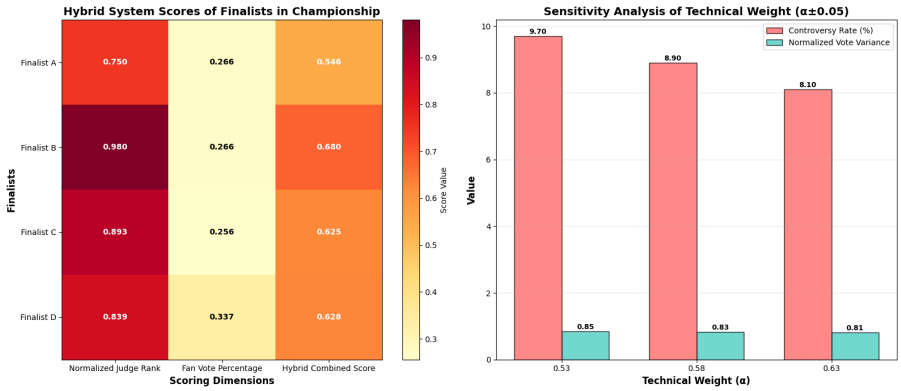

Figure 7: Hybrid-system Score Components for Finalists and Sensitivity of Controversy and Excitement to Changes in α.

3. Results and Discussion

In this section, we provide an overview of the key results from our fan-vote inference model, method comparisons, uncertainty quantification, and regression analysis. The discussion also includes insights into how the hybrid scoring system balances fairness and excitement, with a focus on the impact of professional dancer experience, age, and other factors on the outcomes.

The fan-vote inference model successfully replicates historical eliminations with an overall consistency rate of 86.3% across 337 elimination weeks. This confirms the robustness of the model's ability to predict eliminations under the given constraints. However, we observed variations in performance across different seasons. For example, seasons with unusual structures, such as all-star casting or multiple eliminations per week, show lower match rates. Despite these variations, the inferred weekly vote volume remains within the expected range, demonstrating the plausibility of the scale of fan participation. The confidence in the inferred results is supported by the meaningful variation in average weekly fan votes observed across seasons, which is consistent with historical data.

To quantify the uncertainty in the inferred votes, we computed confidence intervals (CI) based on standard deviation estimations from bootstrap resampling. The confidence-interval widths range from 1 to 2 million votes on average, reflecting the uncertainty in vote estimates. Notably, the intervals are much narrower for finalists and high-visibility contestants, highlighting the tighter constraints on elimination outcomes as the season progresses. On the other hand, contestants who are eliminated earlier or have lower visibility exhibit wider intervals, indicating non-identifiability—that is, multiple plausible vote profiles could result in the same elimination outcome.

The fan-vote influence weight is higher for the rank-based method than for the percentage-based method, as shown in Fig. 6, indicating that the rank-based method places greater emphasis on popularity. This aligns with theoretical concerns that rank-based aggregation discards magnitude information, which can lead to amplified rank shifts with minor changes in voting outcomes.[2][3] Controversy simulations for known low-technical/high-popularity contestants demonstrate greater rank volatility under the rank-based method (Fig. 8). This shows how the rank-based method is more sensitive to changes in rank, especially for contestants with lower technical performance but high popularity.

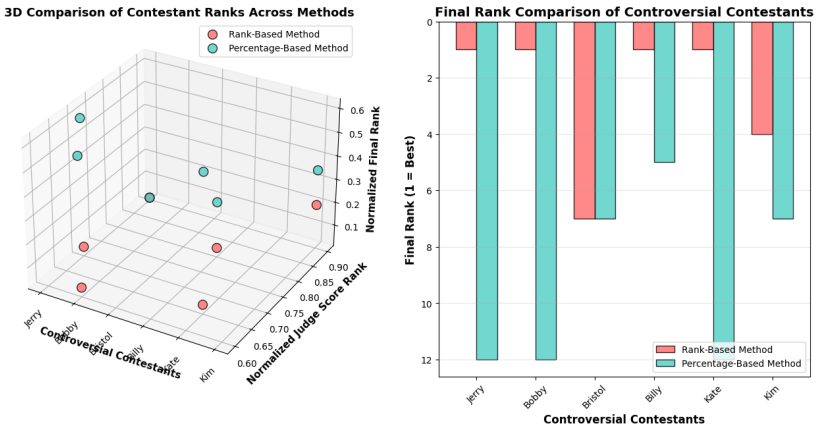

Figure 8: Cross-method Comparison for Representative Controversial Contestants: Normalized Ranks and Final Placement Differences.

The regression analysis reveals important insights into how certain factors influence the outcomes. Professional dancer experience is positively correlated with higher judge scores and better final placements, supporting the idea that experienced dancers have a technical advantage. On the other hand, age shows a slight negative relationship with judge scores, but a weak positive relationship with fan votes, suggesting that older contestants may receive more fan support despite lower technical performance. This supports the idea that technical abilities and popularity may respond differently to individual characteristics.

The optimized hybrid weight α=0.58 reduces controversy relative to the historical rank-based method, while maintaining similar levels of vote variance (a proxy for fan engagement) as the rank-based method (Fig. 6). The sensitivity analysis around "α±0.05" shows that the fairness-excitement tradeoff is smooth, rather than brittle, thus improving the operational robustness of the scoring system. The simulation errors for the two vote combination methods are also evaluated. Fig. 9 displays the distribution of fan-vote influence weights and compares the simulation errors for both aggregation methods, showing that rank-based methods have larger simulation errors, suggesting greater volatility.

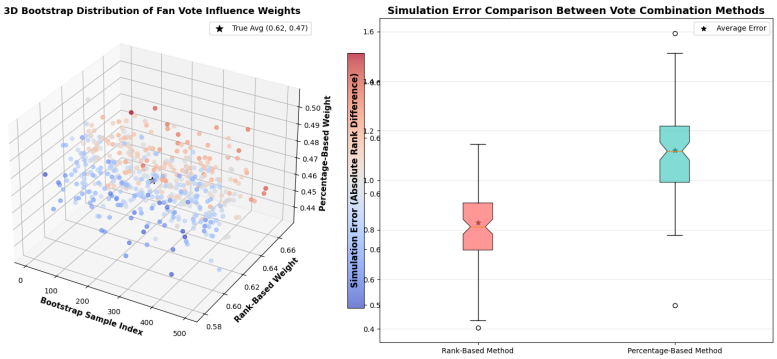

Figure 9: Bootstrap Distribution of Fan-vote Influence Weights and Simulation-error Comparison Between Aggregation Methods.

4. Conclusion

We introduced a constraint-based inference model to recover latent weekly fan votes in DWTS, enabling quantitative comparison of historical aggregation rules and principled design of a hybrid scoring system. Results suggest that rank-based aggregation more strongly privileges fan votes and is associated with higher controversy, whereas a weighted hybrid (α=0.58) can reduce controversy while preserving engagement. The modeling pipeline—optimization-based inference, uncertainty quantification, and rule optimization—generalizes to other audience-participation competitions where the production goal is to balance expertise with popular appeal.

References

[1] Arrow, Kenneth J. Social Choice and Individual Values. New York: John Wiley & Sons, 1951.

[2] Saari, Donald G. Decisions and Elections: Explaining the Unexpected. Cambridge: Cambridge University Press, 2001.

[3] Tideman, T. Nicolaus. Collective Decisions and Voting: The Potential for Public Choice. Aldershot: Ashgate, 2006.

[4] Boyd, Stephen, and Lieven Vandenberghe. Convex Optimization. Cambridge: Cambridge University Press,

2004.

[5] Diamond, Steven, and Stephen Boyd. "CVXPY: A Python-Embedded Modeling Language for Convex Optimization." Journal of Machine Learning Research 17, no. 83 (2016): 1–5.

[6] Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." In Proceedings of the 9th Python in Science Conference (SciPy 2010), pp. 92–96. SciPy, 2010.

[7] Efron, Bradley, and Robert J. Tibshirani. An Introduction to the Bootstrap. New York: Chapman & Hall/CRC, 1993.

[8] Dwork, Cynthia, Ravi Kumar, Moni Naor, and D. Sivakumar. "Rank Aggregation Methods for the Web." In Proceedings of the 10th International Conference on World Wide Web (WWW '01), pp. 613–622. ACM, 2001.

[9] Draper, Norman R., and Harry Smith. Applied Regression Analysis. 3rd ed. New York: Wiley, 1998.

[10] Shapiro, S. S., and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." Biometrika 52, no. 3–4 (1965): 591–611.

[11] Breusch, T. S., and A. R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." Econometrica 47, no. 5 (1979): 1287–1294.

[12] O'Brien, Robert M. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." Quality & Quantity 41, no. 5 (2007): 673–690.

[13] Deb, Kalyanmoy, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II." IEEE Transactions on Evolutionary Computation 6, no. 2 (2002): 182–197.