# MCM 2026 Problem C: Data With The Stars

Team #____

## Summary Sheet

- **Goal.** Infer weekly fan vote *shares* (latent) from judges' scores and elimination outcomes; compare rank vs. percent vote-combination rules; quantify drivers of judges vs. fans; propose and evaluate a better system.

- **Key modeling idea.** Fan shares are generated by a latent preference model

$$f_{i,t} = \frac{\exp(u_{i,t})}{\sum_{k \in A_t} \exp(u_{k,t})}, \quad u_{i,t} = \beta_0 + \beta_J \tilde{J}_{i,t} + \beta_P \,\text{momentum}_{i,t} + \beta_U \,\text{underdog}_{i,t} + \beta_X^\top X_i,$$

  where $\tilde{J}$ is a normalized judges score and $X_i$ are celebrity/pro covariates.

- **Estimation.** $\beta$ is fit by maximum likelihood so that the implied eliminations (and finals ordering) match observed results under the season-appropriate rule (rank vs. percent), with a "temperature" $\tau$ controlling determinism.

- **Consistency (fit to eliminations).** Using the fitted latent fan-share model, the implied eliminated contestant matches the observed in **38.2%** of percent-rule elimination weeks (76/199) and **30.3%** of rank-rule weeks (20/66), well above a random baseline of $\approx 17\%$ when $\sim 6$ contestants remain. These denominators are the *inverse-fit evaluation sample*: elimination weeks with observed elimination and fitted fan shares (265 weeks total; the full rule-comparison universe is 335 weeks). This is our primary *consistency* metric [1]. (Fit diagnostics: `reports/gonogo_report.md`.)

- **Uncertainty.** We quantify uncertainty via (i) season-bootstrap intervals for $f_{i,t}$ and (ii) a margin-to-flip "robustness radius" measuring how much $f$ must change to alter the eliminated contestant. Certainty varies strongly by week: Season 10 Week 1 radius 0 (tight) vs. Season 12 Week 3 radius $\infty$ (blowout).

- **Rule comparison.** From `season_rule_comparison.csv`: percent vs. rank disagree on who is eliminated in **0–36.4%** of elimination weeks by season (mean **13.4%**); fan-influence index is **0.33–1.0** (mean 0.92) for both percent and rank. Sensitivity sweep: at judge weight 0, 26 of 335 weeks flip; at weight 1, 0 weeks flip.

- **Controversy cases.** Seasons 2, 4, 11, 27: judge–fan disagreement is visible in inferred shares vs. judge scores; outcomes differ under rank vs. percent and under judges-save (see controversy tables and counterfactual CSVs).

- **Drivers.** From `pro_dancer_effects_top10.csv`: pro dancer and celebrity characteristics affect judges and fans differently. Pros who boost *fans* relative to judges (largest fan-minus-judge fixed effects): **Andrea Hale** (2.03), **Henry Byalikov** (1.01), Elena Grinenko (1.01), Ezra Sosa (0.87), Tyne Stecklein (0.71); judges-favoring pros appear in the bottom of the effects table.

- **Recommendation.** We recommend a **weighted percent rule with saturation** and an **optional judges-save trigger** in close weeks. The saturation softcaps *fan* share to prevent extreme fan-bloc dominance (combined score $c = w \cdot j + (1 - w) \cdot \mathrm{softcap}(f)$). The goal is not to replicate historical eliminations, but to (i) preserve fan relevance, (ii) improve robustness/predictability in close outcomes, and (iii) reduce judge–fan controversy cases (e.g., Jerry Rice, Billy Ray Cyrus, Bristol Palin, Bobby Bones). Our evaluation reports (a) how often outcomes would change under the proposed rule (scope), (b) controversy-mismatch reduction, and (c) robustness-radius improvements relative to the historical rule.

# Contents

# 1 Problem Context and Data

## 1.1 Competition and Rule Regimes

The show combines judges' scores and fan votes to eliminate the couple with the lowest combined score each week; fans vote to keep a couple (not to eliminate one), and judges score each performance [1]. In the first two U.S. seasons, the combination used *ranks*; beginning in season 3, producers switched to a *percent* method after season 2 concerns, and in response to a season 27 controversy, a judges-save modification (choose which of the bottom two to eliminate) was introduced around season 28 together with a return to the rank-based method [1]. The exact season is not known, but assuming season 28 is reasonable.

## 1.2 Assumptions and Notation

We assume: (i) season 28 is the rule-change point (judges-save and return to rank); (ii) index scaling (e.g., $V_i = f_i \times 10^7$) is a normalization for interpretability only, not true vote counts [1]; (iii) a contestant is "active" in a given week if they have not yet been eliminated and have not withdrawn; we use the active set when computing combined scores and elimination. Notation: $A_t$ = active contestants in week $t$; $J_{i,t}$ = total judges score; $f_{i,t}$ = fan vote share; $j_{i,t}$, $c_{i,t}$, $R_{i,t}$ as in the voting schemes below.

## 1.3 Dataset and Preprocessing

Describe the provided CSV structure (contestant demographics, pro dancer, week-by-week judge scores, results). Mention edge cases: varying judges, N/A weeks, no-elimination weeks, double eliminations, withdrawals, and zeros after elimination. Briefly state your preprocessing steps (wide → long, aggregate judge totals, define active set, parse elimination week and placement, normalize/standardize judges score).

# 2 Voting Schemes: Rank, Percent, and Judges-Save

Let $A_t$ be active contestants in week $t$ and $J_{i,t}$ the total judges score.

## 2.1 Percent scheme

Define judges percent $j_{i,t} = J_{i,t} / \sum_{k \in A_t} J_{k,t}$ and fan percent $f_{i,t} = V_{i,t} / \sum_{k \in A_t} V_{k,t}$. Combined score $c_{i,t} = j_{i,t} + f_{i,t}$; eliminated is $\arg\min_i c_{i,t}$.

## 2.2 Rank scheme

Let $r^J_{i,t}$ be rank of $J_{i,t}$ (best=1) and $r^F_{i,t}$ rank of $V_{i,t}$; combined $R_{i,t} = r^J_{i,t} + r^F_{i,t}$; eliminated is $\arg\max_i R_{i,t}$.

## 2.3 Judges-save modification

Among the bottom two by combined criterion, judges select which couple to eliminate; we model this probabilistically with

$$\Pr(\text{eliminate } i \mid \{i,k\}) = \sigma\big(\alpha(J_{k,t} - J_{i,t})\big),$$

and fit $\alpha$ from observed outcomes in the judges-save era. Fitted $\alpha \approx 0.029$ (from `judges_save_alpha.json`).

# 3 Model for Latent Fan Vote Shares

## 3.1 Identifiability

Under the percent rule, only *shares* $f_{i,t} = V_{i,t} / \sum_{k \in A_t} V_{k,t}$ are identified: vote totals can be scaled arbitrarily without changing combined scores $c_{i,t} = j_{i,t} + f_{i,t}$. Under the rank rule, only the *ordering* of vote totals is identified. We therefore report fan vote *shares* as primary outputs and provide index-scaled totals (e.g., $V_i = f_i \times 10^7$) only for interpretability, with a clear label that they are not true vote counts. This non-identifiability is also illustrated in the problem appendix, which notes that many hypothetical fan vote totals can reproduce the same eliminations and presents an example using an arbitrary total of 10 million votes [1].

## 3.2 Latent preference model

Let $A_t$ denote the set of active contestants in week $t$. We model fan share as a multinomial logit (softmax) over a latent utility $u_{i,t}$:

$$f_{i,t} = \frac{\exp(u_{i,t})}{\sum_{k \in A_t} \exp(u_{k,t})},$$

with

$$u_{i,t} = \beta_0 + \beta_J \, \tilde{J}_{i,t} + \beta_P \, \mathrm{momentum}_{i,t} + \beta_U \, \mathrm{underdog}_{i,t} + \beta_X^\top X_i.$$

Terms are defined as follows. **Normalized judge score** $\tilde{J}_{i,t}$: within each ($season, week$), we set $\tilde{J}_{i,t} = J_{i,t} / \max_{k \in A_t} J_{k,t}$ (and 0 if the max is 0), so that the best-scoring contestant has $\tilde{J} = 1$. This captures the extent to which judges favor contestant $i$ relative to the field. **Momentum** (denoted $p_{\mathrm{prev}}$ in code): rank by judges' score in the previous week (1 = best); for week 1 we set it to 0. Higher rank last week may attract more fan attention (momentum) or, if negative, an "underdog" effect. **Underdog**: we set $\mathrm{underdog}_{i,t} = 1$ if $J_{i,t}$ is at or below the median judges score among active contestants in that week, else 0. This allows a "rage vote" or sympathy effect for lower-scoring contestants. **Covariates** $X_i$: we include celebrity age (during the season) and an industry dummy (e.g., 1 if Actor/Actress). The coefficients $\beta = (\beta_0, \beta_J, \beta_P, \beta_U, \beta_X)$ are shared across seasons and weeks; the season-appropriate rule (percent or rank) is applied when mapping $(J, f)$ to combined scores and thus to elimination.

## 3.3 Likelihood and fitting

We fit $\beta$ and a temperature $\tau > 0$ by maximum likelihood. The likelihood has two parts.

**Elimination events.** For each week with an elimination, let $c_{i,t}$ (percent) or $R_{i,t}$ (rank) be the combined score under the rule for that season. We model the probability that contestant $i$ is eliminated as a softmax over the active set. Under percent: lower combined score $c$ implies higher elimination probability; we set

$$\Pr(\text{eliminate } i) \propto \exp(-\tau \, c_{i,t}).$$

Under rank: higher rank-sum $R$ (worse) implies higher elimination probability; we set

$$\Pr(\text{eliminate } i) \propto \exp(\tau \, R_{i,t}).$$

Thus $\tau$ controls determinism: large $\tau$ makes the lowest $c$ (or highest $R$) almost surely eliminated; small $\tau$ flattens the distribution.

**Finals ordering.** For each season's finals week, we observe the placement order (1st, 2nd, 3rd). We model this with a Plackett–Luce likelihood: the probability of the observed ordering given strengths $s_i$ (we use combined score for percent, or negative rank-sum for rank so that better rank-sum gives higher strength) is the product over positions of the probability of choosing that contestant from the remaining set, with choice probabilities proportional to $\exp(s_i)$.

The total log-likelihood is the sum of log-probabilities over all elimination events plus the sum over all finals events. We minimize the *negative* log-likelihood with respect to $(\beta, \tau)$ using L-BFGS-B, with $\tau$ bounded below by a small positive constant. Covariates are built from the contestant–week data (normalized judge score, previous-week rank, underdog, age, industry dummy); see `src/models/vote_latent.py` and `src/fit/fit_elimination.py`. Fit diagnostics (elimination match rates, finals likelihood) are reported in Section 9.

# 4 Uncertainty Quantification

We measure uncertainty in two complementary ways.

**(1) Bootstrap intervals for fan shares.** We resample *seasons* with replacement (e.g., 50 bootstrap replicates). For each replicate we refit $(\beta, \tau)$ on the resampled contestant–week data, then run the forward pass on the *full* dataset with the fitted $\beta$ to obtain fan shares $f_{i,t}$ for every $(\mathit{season}, \mathit{week}, \mathit{contestant})$. Across replicates we compute the mean, 5th percentile, and 95th percentile of $f_{i,t}$ at each cell. This yields interval estimates for $f_{i,t}$ that reflect uncertainty due to season-to-season variation in the elimination and finals data. Optionally, we can bootstrap by resampling *weeks* within each season instead of seasons; the implementation supports both (see `src/fit/uncertainty.py`).

**(2) Margin-to-flip robustness radius.** For each elimination week we ask: how much must fan shares $f_t$ change so that a *different* contestant would be eliminated under the same rule? We parameterize perturbations in log-share space: $f' = \mathrm{softmax}(\log f + z)$ with $z$ unconstrained, so that $f'$ remains on the simplex. We find the minimum L2 norm of $z$ such that the eliminated contestant under the week's rule (percent: $\arg\min_i(j_i + f'_i)$; rank: $\arg\max_i(r_i^J + r_i^F)$ with ranks from $f'$) is not the currently eliminated contestant. The optimization is a constrained nonlinear problem (minimize $\|z\|^2$ subject to the flip constraint); we use SLSQP. For interpretability we also report the L2 norm of $(f' - f)$ in share space as the "robustness radius." A *small* radius means the outcome is sensitive to small changes in fan shares (uncertain week); a *large* or infinite radius means the eliminated contestant would not change under plausible perturbations (certain week). See `src/fit/margin_to_flip.py`.

**Examples.** Tight week: Season 10 Week 1, robustness radius 0 (outcome sensitive to small changes in fan shares). Blowout week: Season 12 Week 3, robustness radius $\infty$ (no feasible perturbation changes the eliminated contestant; elimination is robust). Certainty varies by week and contestant because of the geometry of combined scores near the elimination boundary: when the bottom two are close, a small shift in $f$ can flip who is eliminated; when one contestant is clearly last, the radius is large or infinite.

Table 1 shows the distribution of margin-to-flip robustness radii under the proposed scoring rule (weighted saturation) across all 335 elimination weeks. About **8.7%** of elimination weeks are knife-edge (robustness radius 0), meaning arbitrarily small perturbations to fan shares can flip who goes home, while **14.9%** are blowouts (radius $\infty$), meaning no feasible perturbation flips the eliminated couple under the rule. Among finite-radius weeks, the median radius is **1.36** (IQR **1.00–1.77**), indicating that most weeks require a nontrivial shift in the fan-share simplex to change the outcome.

Table 1: Distribution of margin-to-flip robustness radii under the proposed scoring rule (weighted saturation). Radii are computed per (season, week) elimination event; larger means more robust.

| Quantity | Value |
|---|---|
| Number of elimination weeks | 335 |
| Zero-radius (knife-edge) weeks | 29 (8.7%) |
| Infinite-radius (blowout) weeks | 50 (14.9%) |
| Finite-radius median | 1.36 |
| Finite-radius IQR | 1.00 – 1.77 |
| Finite-radius max | 3.05 |

# 5 Rank vs Percent Across Seasons

## 5.1 Across-season disagreement rates and fan influence

Table 2 summarizes how often different elimination rules disagree about who goes home. Across 34 seasons, percent and rank rules disagree in **0.0%–36.4%** of elimination weeks (mean **13.4%**, median **10.6%**). Six seasons have *zero* disagreement (seasons 3, 4, 10, 17, 20, 26), while the largest disagreement occurs in season 32 (36.4%). Table 3 lists the highest-disagreement seasons.

Judges-save (a bottom-two selection rule) can change outcomes more frequently: percent vs. judges-save differs by **0.0%–50.0%** (mean **22.3%**), and rank vs. judges-save differs by **0.0%–54.5%** (mean **17.1%**). This is consistent with judges-save acting as a discretionary override precisely in weeks where the bottom of the leaderboard is crowded.

We also compute a *fan influence index* (FII): the fraction of elimination weeks in which the eliminated couple changes when the fan-vote component is neutralized (i.e., set equal across contestants, holding judges fixed). FII is high in most seasons: mean **0.917** under percent and **0.900** under rank, with values ranging from **0.333** to **1.000**. In other words, in a typical season, fans are pivotal in the majority of eliminations, regardless of whether ranks or percents are used.

Table 2: Across-season disagreement between elimination rules and fan influence. Each season contributes its fraction of elimination weeks where the eliminated couple differs under the two rules.

| Statistic | pct vs rank | pct vs save | rank vs save | fan infl. (pct / rank) |
|---|---|---|---|---|
| Min | 0.0% | 0.0% | 0.0% | 0.333 / 0.333 |
| Median | 10.6% | 22.2% | 19.1% | 0.909 / 0.909 |
| Mean | 13.4% | 22.3% | 17.1% | 0.917 / 0.900 |
| Max | 36.4% | 50.0% | 54.5% | 1.000 / 1.000 |

Table 3: Seasons with the largest disagreement between percent and rank elimination outcomes.

| Season | Elim. weeks | pct vs rank diff | fan infl. (pct / rank) |
|---:|---:|---:|---:|
| 32 | 11 | 36.4% | 1.000 / 1.000 |
| 11 | 10 | 30.0% | 1.000 / 1.000 |
| 19 | 11 | 27.3% | 1.000 / 0.909 |
| 21 | 11 | 27.3% | 0.909 / 0.909 |

From `sensitivity_flip_summary.csv`: varying judge weight from 0 to 1, the fraction of weeks where elimination would flip is highest at weight 0 (26/335 weeks) and 0 at weight 1; intermediate weights show modest flip rates (e.g., 2–9 flips at 0.76–0.86).

## 6 Controversy Case Studies

We examine the prompt's highlighted controversy examples [1]: Season 2 (Jerry Rice), Season 4 (Billy Ray Cyrus), Season 11 (Bristol Palin), and Season 27 (Bobby Bones). For each, we compare rank vs. percent outcomes and quantify how a judges-save rule would alter eliminations when the bottom two are close. Data: `controversy_season*_fan_shares_vs_judges.csv` (judge vs. inferred fan disagreement) and `controversy_season*_counterfactual_elimination.csv` (counterfactual eliminations under judges-save).

## 7 Drivers of Performance: Pro Dancer and Celebrity Characteristics

We estimate two parallel linear models: one for judges outcomes and one for inferred fan outcomes, using the same core covariates. Table 4 compares the common covariates side-by-side.

Two effects are strong and consistent across judges and fans: **week** has a negative coefficient in both models, and **age** is strongly negative in both models. Industry and region behave differently: **industry_dummy** is positive and highly significant for judges but near zero and not significant for fans, suggesting judges respond to industry-related factors more than the voting public does. The **region_us** indicator is modestly negative and significant for judges but not significant for fans.

(Separately, pro-dancer fixed effects provide the largest heterogeneity in the cross-sectional fit; we summarize those in `pro_dancer_effects_top10.csv` and `pro_dancer_effects_bottom10.csv`. Pros with largest positive "fan minus judge" effect: Henry Byalikov, Andrea Hale; judges-favoring: Koko Iwasaki, Ashly DelGrosso.)

Table 4: Common-covariate comparison for judges vs. fans regressions. OLS coefficients and *p*-values; stars correspond to conventional significance levels.

| Covariate | Judges coef. | *p*-value | Fans coef. | *p*-value |
|---|---:|---:|---:|---:|
| week | −0.042 *** | $1.21 \times 10^{-11}$ | −0.046 *** | $2.22 \times 10^{-15}$ |
| age | −0.033 *** | $3.38 \times 10^{-91}$ | −0.048 *** | $2.83 \times 10^{-172}$ |
| industry_dummy | 0.183 *** | $3.05 \times 10^{-6}$ | 0.010 | 0.79 |
| region_us | −0.134 ** | $1.29 \times 10^{-2}$ | −0.081 | 0.11 |

## 8 Proposed “Better” System and Evaluation

We propose a **weighted percent with saturation** and an optional trigger-based judges-save. The combined score is $c_i = w \cdot j_i + (1 - w) \cdot \mathrm{softcap}(f_i)$; the softcap is applied to *fan* share to prevent extreme fan-bloc dominance (not judge dominance). Optionally: trigger judges-save when the bottom-two margin is below a threshold. Fairness axioms: monotonicity (better combined score $\Rightarrow$ not eliminated), fan relevance bounds, robustness (margin-to-flip), transparency. Evaluation: we report (a) scope of change—how often the proposed rule would yield a different eliminated contestant than observed (see `proposed_system_eval.csv`); (b) controversy-mismatch reduction; (c) robustness radius by week (Table 1). Operationally, the robustness radius provides a transparent “closeness” signal: in our data, about 8.7% of weeks are knife-edge (radius 0), suggesting a judges-save trigger should be reserved for a small subset of genuinely close eliminations rather than used routinely.

## 9 Fit Quality and Validation

- **Elimination prediction (consistency):** Using the *fitted latent fan-share model*, the implied eliminated contestant matches the observed in 38.2% of percent-weeks (76/199) and 30.3% of rank-weeks (20/66), vs. random baseline $\approx$ 17% when $\sim$ 6 contestants. The denominator is the inverse-fit evaluation sample (265 elimination weeks with observed elimination and fitted fan shares; the full universe for rule comparison is 335 weeks). This is our primary consistency metric [1]; it is computed from the fitted model, not from the proposed scoring rule. See `reports/gonogo_report.md`.
- **Finals likelihood:** Plackett–Luce term is non-degenerate; ordering likelihood contributes to identification.
- **Robustness:** Margin-to-flip shows clear variation (Table 1): 29 knife-edge weeks (8.7%), 50 blowout weeks (14.9%), finite-radius median 1.36; see `proposed_system_robustness.csv`.

## 10 Limitations and Extensions

Identifiability: only shares (percent) or order (rank) are identified. Regime uncertainty: season 28 judges-save start is assumed. Unobserved confounders: marketing, social media, contestant visibility. Extensions: incorporate viewership or social sentiment if data become available.

## A Mechanistic Interpretation via Biased Mean-Field Voter Dynamics (Optional)

The mean-field/memory/network module (in `src/models/meanfield.py`) is used for *interpretation and robustness simulation*, not as the primary inference engine. It provides a mechanistic story for how judge scores and social influence could generate vote shares over time; we treat the $\beta$-fitted latent vote model in `main.py` as the primary estimator. Below we summarize the structure so that results from this module can be interpreted consistently with the main report.

### A.1 Role relative to the primary model

The primary model (Section 3) fits a single set of coefficients $\beta$ and temperature $\tau$ by maximum likelihood so that implied eliminations match observed outcomes. That model is *static* in the sense that fan share in week $t$ depends on covariates and judge score in week $t$ (and optionally momentum/underdog from the same or previous week), but there is no explicit dynamical law linking $p_t$ to $p_{t-1}$. The mean-field module adds a *discrete-time dynamical* layer: fan shares evolve week-to-week via a recurrence $p_{t+1} = \Phi(p_t, S_t, \ldots)$. This can be used to (i) interpret how judge signals and past popularity might combine into a plausible evolution of votes, and (ii) run robustness or scenario simulations (e.g., different memory kernels or network coupling) without re-fitting the primary $\beta$.

### A.2 Single-population mean-field update (F2)

The core recurrence is

$$p_{t+1} = (1 - \kappa)\, p_t + \kappa\, \mathrm{softmax}(u_t), \qquad u_t = \eta\, S_t + \gamma\, \log(p_t + \varepsilon) + \text{(optional terms)}.$$

Here $p_t$ is the vector of fan shares (simplex), $S_t$ is the vector of judge signals (e.g., normalized scores) in week $t$, and $\kappa \in (0, 1]$ is a mixing speed: the new share is a convex combination of the previous share and a softmax of the utility $u_t$. The term $\gamma\, \log(p_t)$ captures *incumbency* or social reinforcement (higher current share $\Rightarrow$ higher utility, all else equal). The term $\eta\, S_t$ is the judge-driven component. Optional terms include covariates $\theta^\top X$, an underdog term $\beta_U\, \mathrm{underdog}_t$, and memory terms described below.

### A.3 Switching-rate microfoundation (F1)

One can motivate the update by a *switching-rate* story: voters are drawn toward a target distribution $q_t$ that mixes current popularity and a judge-driven distribution. For example, $q_t = \rho\, p_t + (1 - \rho)\, \mathrm{softmax}(\eta\, S_t)$, normalized. Then $p_{t+1} = (1 - \kappa)\, p_t + \kappa\, q_t$ corresponds to a mean-field law where a fraction $\kappa$ of the population "switches" toward $q_t$ each period. The implementation supports this via the same recurrence with $u_t$ constructed so that $\mathrm{softmax}(u_t)$ matches the desired target (e.g., with $\gamma$ and $\eta$ playing the roles of $\rho$ and judge weight).

### A.4 Underdog / rage-vote (F3)

The underdog score is a scalar per contestant that is high when the contestant has *low* judge score but *high* current popularity (or vice versa, depending on parameterization). Formally, $\mathrm{underdog}_i = \sigma(a(p_i - \tau_p))\, \sigma(b(\tau_S - S_i))$ in smooth mode, or an indicator $1[S_i \le \tau_S]\, 1[p_i \ge \tau_p]$ in indicator mode, with thresholds $\tau_S$, $\tau_p$ (e.g., medians). This is added to $u_t$ as $\beta_U\, \mathrm{underdog}_t$, so that "underdog" contestants get a utility boost and can sustain higher share despite lower judge scores—a simple model of sympathy or rage voting.

### A.5 Memory: kernel-weighted history and Markovian state

Two types of memory are supported. (1) **Kernel-weighted history:** define $\bar{S}_t = \sum_{\tau=0}^{t} k(t-\tau)\, S_\tau$ and $\bar{p}_t$ similarly, where $k(\Delta t)$ is a kernel (e.g., exponential $k(\Delta t) = \lambda^{\Delta t}$, rectangular window, or power-law). Then $u_t$ can include $\eta_S\, \bar{S}_t$ and $\gamma_{\text{hist}}\, \log(\bar{p}_t + \varepsilon)$, so that past judge signals and past popularity influence current utility. (2) **Markovian fading state:** a single state vector $m_t$ is updated by $m_{t+1} = (1 - \lambda)\, m_t + \lambda\, S_t$, and $u_t$ includes $\eta_m\, m_t$. This gives a one-dimensional fading memory of judge signals. Both channels are optional; when omitted, the model reduces to the static-like update with only $S_t$ and $\log(p_t)$.

### A.6 Networked extension: multiple communities and small-world coupling

The code also supports $G$ "communities" (e.g., demographic or regional voter blocs), each with its own share vector $p_t^{(g)}$ on the simplex. Communities are coupled via a row-stochastic influence matrix $W$ (e.g., from a Watts–Strogatz small-world graph: ring lattice with $K$ neighbors, then each edge rewired with probability $\beta$; then $W$ is adjacency plus self-weight $\omega_{\text{self}}$, row-normalized). The utility in community $g$ is

$$u_t^{(g)} = \eta\, S_t + \gamma\, \log(p_t^{(g)} + \varepsilon) + \delta\, (W \log(p_t + \varepsilon))_g + (\text{covariates, underdog}).$$

The term $(W \log p_t)_g$ is the weighted average of $\log p_t^{(h)}$ over neighbors $h$ of $g$, so that high popularity in neighboring communities boosts utility in $g$ (social influence across blocs). The aggregate share reported for elimination can be $\bar{p}_t = \sum_g w_g\, p_t^{(g)}$ for reporting weights $w_g$. Optional logit-normal noise ($\sigma_{\text{shock}}$) can be added to $u$ before softmax for robustness checks.

### A.7 Parameters and use in the report

Key parameters: $\kappa$ (mixing speed), $\eta$ (judge weight), $\gamma$ (incumbency/social weight), $\delta$ (cross-community weight), $\beta_U$ (underdog weight), and memory parameters (kernel type, decay/window, $\eta_S$, $\gamma_{\text{hist}}$, $\lambda$, $\eta_m$). These are *not* fit by the same MLE as the primary model; they can be set for scenario or sensitivity analysis. In the main report we do not report point estimates from the mean-field module; we use it only to interpret how dynamics and memory could generate patterns consistent with the primary $\beta$-fit and to run robustness simulations (e.g., different kernels or coupling strength) when needed.

## Memo to DWTS Producers (1–2 pages)

**To:** Producers of *Dancing with the Stars*
**From:** Team #____
**Subject:** Recommended method for combining judges and fan votes

**Executive recommendation.** We recommend adopting a **weighted percent rule with saturation** and an **optional judges-save trigger** in close weeks. The saturation softcaps *fan* share to prevent extreme fan-bloc dominance (combined score $c = w \cdot j + (1-w) \cdot \text{softcap}(f)$), keeping fan votes meaningful without letting a single fan bloc dominate. The data show that percent and rank disagree on who goes home in only about 13% of elimination weeks on average (range 0–36% by season), and fan influence is already high (index 0.33–1.0, mean 0.92). The prompt's controversy examples—Season 2 (Jerry Rice), Season 4 (Billy Ray Cyrus), Season 11 (Bristol Palin), Season 27 (Bobby Bones)—show clear judge–fan disagreement; a judges-save can mitigate those cases [1].

**Evidence.**

- **Rule comparison (season_rule_comparison.csv):** Across 34 seasons, percent vs. rank disagree on who is eliminated in **0–36.4%** of elimination weeks by season (mean **13.4%**). Fan-influence index is **0.33–1.0** (mean 0.92) for both rules—fans already have substantial weight.
- **Consistency (fitted latent model):** Using our *fitted latent fan-share model* (not the proposed rule), the implied eliminated contestant matches the observed in **38.2%** of percent-rule weeks (76/199) and **30.3%** of rank-rule weeks (20/66), well above a random baseline of about 17%. Consistency is computed on the inverse-fit evaluation sample (265 elimination weeks); the full rule-comparison universe is 335 weeks [1]. See `reports/gonogo_report.md`.
- **Uncertainty and robustness:** Certainty varies by week. Season 10 Week 1 has robustness radius 0 (tight); Season 12 Week 3 has radius infinite (blowout). Our proposed rule evaluation reports robustness-radius improvements and controversy-mismatch reduction relative to the historical rule; we report the *scope* of change (how often the proposed rule would yield a different elimination) separately from consistency.
- **Controversy seasons:** In seasons 2, 4, 11, and 27 (Jerry Rice, Billy Ray Cyrus, Bristol Palin, Bobby Bones), inferred fan shares often disagree with judge rankings. A judges-save option lets producers avoid outcomes that would outrage fans when the bottom two are close.
- **Pro/celebrity effects (pro_dancer_effects_top10.csv):** Pros who boost *fans* relative to judges: **Andrea Hale** (2.03), **Henry Byalikov** (1.01). Judges-favoring pros appear in the bottom of the effects table. A stable rule keeps competition fair across pros.

**Proposed system.** Combine judge and fan contributions as $c = w \cdot j + (1 - w) \cdot \text{softcap}(f)$; the softcap is on *fan* share to prevent extreme fan-bloc dominance. Optionally: when the bottom two are within a small margin, trigger a judges-save (we fit $\alpha \approx 0.029$ from season 28+ data). This preserves fan relevance, satisfies monotonicity and transparency, and reduces controversy in close weeks.

**Expected impact.** (i) Fan influence remains high (fan-influence index 0.33–1.0, mean 0.92). (ii) Fairness improves via monotonicity and bounded fan-bloc dominance. (iii) Robustness: margin-to-flip analysis identifies which weeks are close; judges-save can be used only in those weeks. (iv) Controversy frequency can decrease when judges-save is triggered in disputed bottom-two situations.

## References

[1] COMAP, *2026 MCM Problem C: Data With The Stars*, 2026.

## AI Use Report

Describe exactly how AI was used (e.g., brainstorming model structure, code review, writing assistance), what was verified independently (tests, reproducibility checks), and what was not used (no external data unless documented).