An Example of a fully Bayesian Analysis of Sport

How often Does the Best Team Win? Understanding Randomness Across sports

Q Can we understand differences in team strength, win probabilities, and home field advantage across the MLB, NBA, NFL, and NHL?

- estimate team strengths
- estimate the between-season, within-season, and game-to-game variability of team strengths
- estimate each team's home field advantage?
- compare competitiveness across leagues?

* Bayesian Model to provide a unifying framework for contrasting the 4 major North American sport leagues!

Indices sports league $q \in \{MLB, NBA, NFL, NHL\}$

season s, week k

consider the matchup of team i vs. j

city i^* since some teams Relocated, so home field advantage is over the cities, not teams

$s = 1, \ldots, S_q$ S_q = # seasons of data from sport q

$k = 1, \ldots, K_q$ K_q = # weeks in sport q

$i, j = 1, \ldots, t_q$ t_q = # teams in sport q

$i^* = 1, \ldots, t_q^*$

Outcome variable the probability that team i beats team j in season s during week k of league q

$$p_{(q,s,k)ij}$$

→ Assume p is known, implied by casino odds

Home Advantage Parameters (unobserved)

α_{q0} = league wide Home Advantage in sport q

$\alpha_{(q)i^*}$ = team-specific effect for team i among games played in city i^*

center Home Advantage about 0 for identifiability,

$$\sum_{i^*=1}^{t_{q^*}} \alpha_{(q)i^*} = 0$$

Team Strength Parameters (unobserved)

$\theta_{(q,s,k)i}$ and $\theta_{(q,s,k)j}$ season-week team strength parameters for teams i and j

Translate into each team's probability of beating a league-average team.

Center team strength about 0 for identifiability,

$$\sum_{i=1}^{t_q} \theta_{(q,s,k)i} = 0$$

Fully Bayesian Model

* Win probability as a function of team strength and Home Advantage:

$$\text{logit}(P_{(q,s,k)i}) \sim N\left(\theta_{(q,s,k)i} - \theta_{(q,s,k)j} + \alpha_{q0} + \alpha_{(q)i^*},\ \sigma^2_{q\text{-game}}\right)$$

* Allow the strength parameters to vary auto-regressively from season-to-season and week to week:

$$\theta_{(q,s+1,1)i} \sim N\left(\gamma_{q\text{-szn}} \cdot \theta_{(q,s,K_q)i},\ \sigma^2_{q\text{-szn}}\right)$$

$$\theta_{(q,s,k+1)i} \sim N\left(\gamma_{q\text{-week}} \cdot \theta_{(q,s,k)i},\ \sigma^2_{q\text{-week}}\right)$$

→ team strength params are shrunk towards the league average over time in expectation.

* PRIORS:

$$\theta_{(q,1,1)i} \sim N(0,\ \sigma^2_{q\text{-szn}})$$

$$\alpha_{(q)i^*} \sim N(0,\ \sigma^2_{q\text{-}\alpha})$$

Let $\tau^2_{q\text{-game}} = \frac{1}{\sigma^2_{q\text{-game}}}$, $\tau^2_{q\text{-szn}}$, $\tau^2_{q\text{-week}}$, $\tau^2_{q\text{-}\alpha}$ similar

each $\tau^2 \sim \text{uniform}(0,\ 1000)$ $\quad \gamma_{q\text{-szn}} \sim \text{Unif}(0,1)$

$\alpha_{q0} \sim N(0,\ 10000)$ $\quad \gamma_{q\text{-week}} \sim \text{Unif}(0,\ 1.5)$

Fitting the Model

- write the model in matrix-vector form
- get data from sportsinsights.com (includes moneylines) and create data matrices
- estimate the full posterior distribution of each parameter using MCMC methods, specifically Gibbs sampling via the rjags package. Separate posterior dists for each league q.

Results

- for each league q,
  use the posterior distributions of
  α/data to learn about Home Advantage,
  θ/data to learn about team strength,
  γ/data to learn about how team strength changes across weeks and seasons
  σ^2/data to learn about the variability of team strengths
  and win prob. as a function of these to learn about competitiveness within each sport.

Team Strength coefficients over time

- more competitive balance in NBA, NFL: lower between-team differences in quality

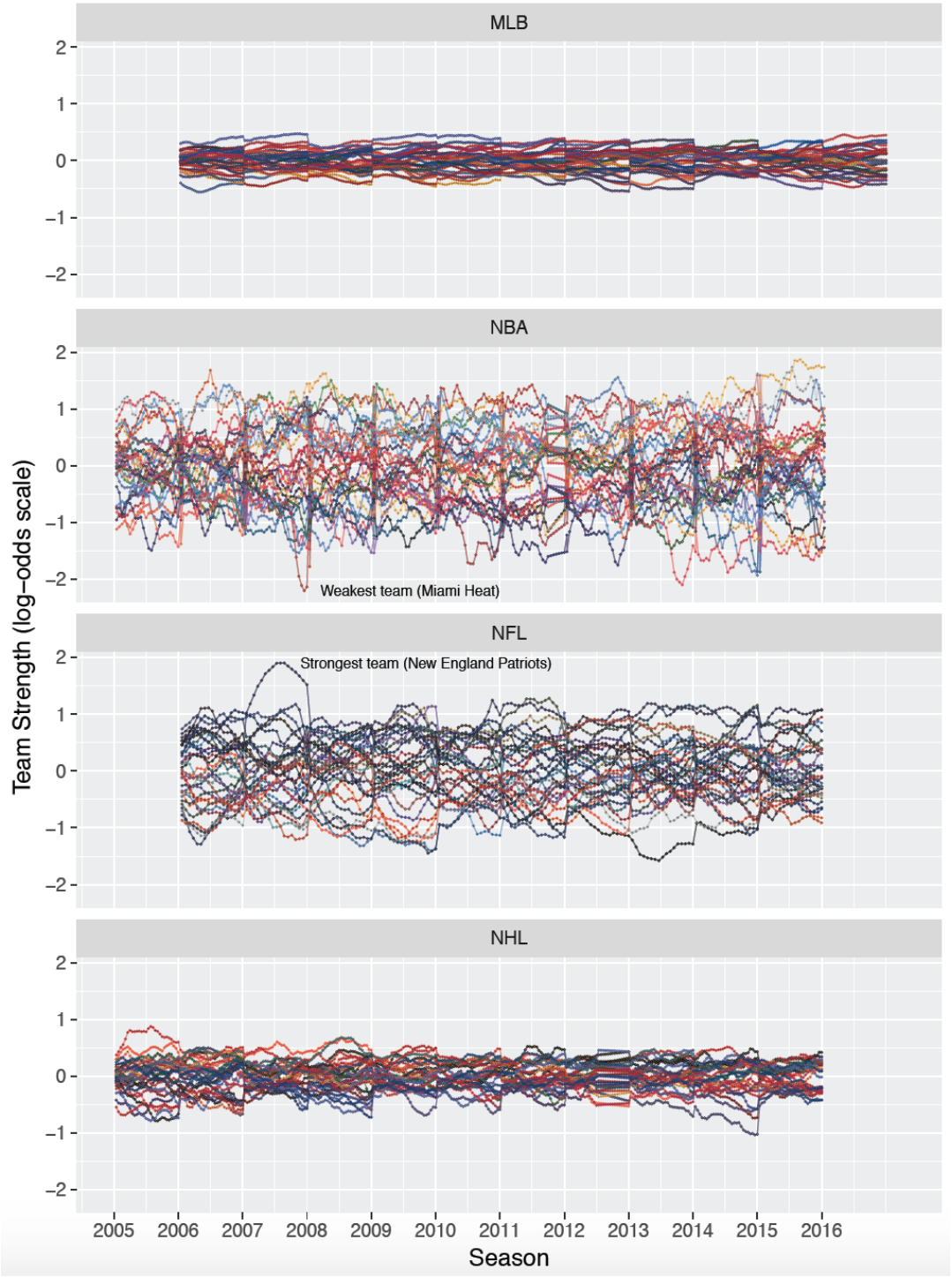

FIG 4. *Mean team strength parameters over time for all four sports leagues. MLB and NFL seasons follow each yearly tick mark on the x-axis, while NBA and NHL seasons begin during years labeled by the preceding tick marks.*

League (q)	$\gamma_{q,season}$	$\gamma_{q,week}$	$\sigma_{q,game}$	$\sigma_{q,season}$	$\sigma_{q,week}$
MLB	0.618 (0.031)	1.002 (0.002)	0.201 (0.001)	0.093 (0.005)	0.027 (0.001)
NBA	0.618 (0.04)	0.977 (0.003)	0.274 (0.002)	0.44 (0.02)	0.166 (0.003)
NFL	0.69 (0.042)	0.978 (0.005)	0.233 (0.008)	0.331 (0.019)	0.147 (0.006)
NHL	0.542 (0.027)	0.993 (0.003)	0.105 (0.001)	0.121 (0.006)	0.053 (0.001)

TABLE 4

Mean posterior draw (standard deviation) by league.

Careful: each γ, σ^2 is tied to each sport's relative logit scale

$\mathbb{E}\gamma_{NBA\text{-}szn} = \mathbb{E}\gamma_{MLB\text{-}szn} = 0.62$

but larger gaps in NBA team strength, so larger reversions in season-level strength in NBA on an absolute scale

$\sigma^2_{q\text{-}game}$: NBA highest game level errors

$\sigma^2_{q\text{-}week}$: NBA highest between-week uncertainty

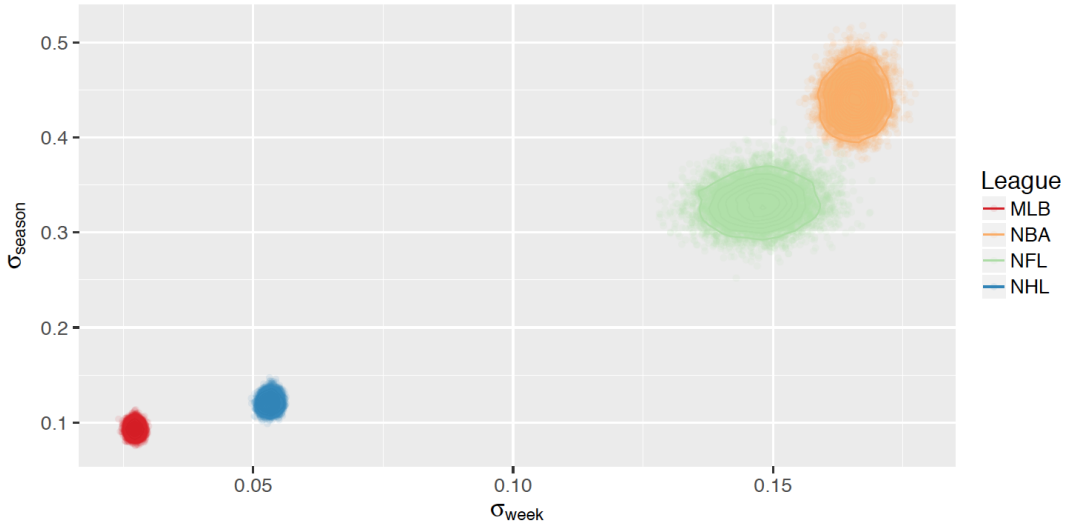

FIG 18. *Contour plot of the estimated season-to-season and week-to-week variability across all four major sports leagues. By both measures, uncertainty is lowest in MLB and highest in the NBA.*

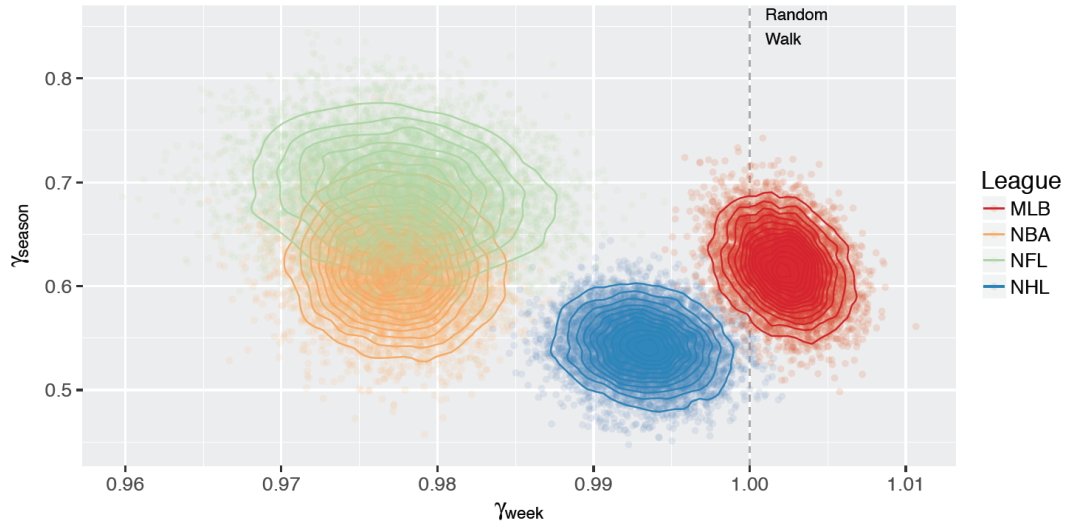

FIG 19. *Contour plot of the estimated season-to-season and week-to-week autoregressive parameters across all four major sports leagues.*

Home Advantage

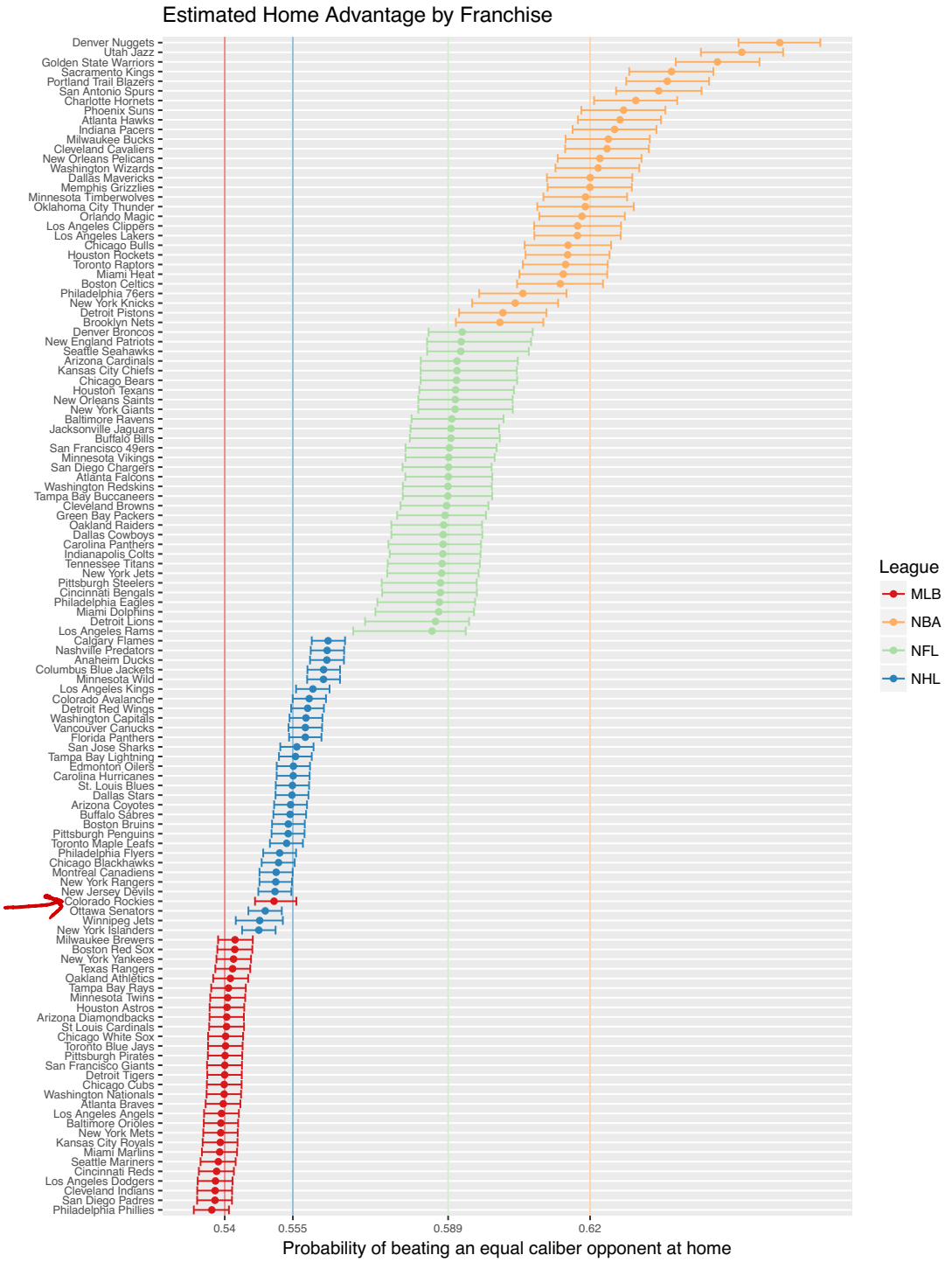

FIG 5. *Median posterior draw (with 2.5th, 97.5th quantiles) of each franchise's home advantage intercept, on the probability scale. We note that the magnitude of home advantages are strongly segregated by sport, with only one exception (the Colorado Rockies). We also note that no NFL team, nor any MLB team other than the Rockies, has a home advantage whose 95% credible interval does not contain the league median.*

Regular Season Parity

* Simulate $n_{sim} = 1000$ draws of $\hat{p}_{q,sim} = \hat{p}_{(q,\hat{s},\hat{E})\hat{i}\hat{j}}$ where $(\hat{s},\hat{E},\hat{i},\hat{j})$ are sampled from observed schedules and $\hat{p}$ sampled from posterior dist.

* $\text{RegParity}_q = 2\int_{1/2}^{1} \mathbb{P}(\tilde{p}_q \le x)\,dx$

deterministic 1
MLB 0.79
NHL 0.73
NFL 0.55
NBA 0.47
fair coin flip 0

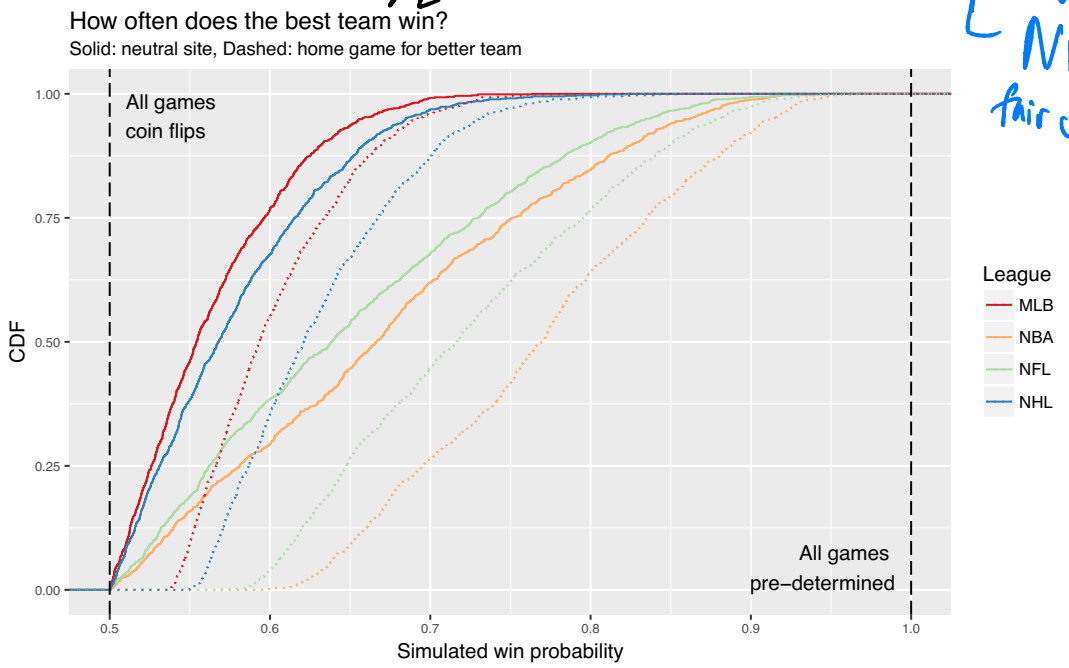

FIG 7. *Cumulative distribution function (CDF) of 1000 simulated game-level probabilities in each league, for both neutral site and home games, with the better team (on average) used as the reference and given the home advantage.*

Postseason Parity

- Collect $z \in \{8, 16\}$ teams with highest average team strength estimates over last 4 weeks of szn in each sport
- Simulate 1000 postseason tournaments

$F = (F_1, \ldots, F_z)$ $\quad F_d$ = the round of tournament finish of the d^{th} seed

pure chalk: $F_1 = 1, F_2 = 2, F_3 = F_4 = 3, F_5 = F_6 = F_7 = F_8 = 4, \ldots$

$F_d = \lceil \log_2 d + 1 \rceil \Rightarrow \mathbb{E}F_d = \lceil \log_2 d + 1 \rceil$

pure randomness: $\mathbb{E}F_d = \sum_{d=1}^{z} \frac{1}{z} \cdot \lceil \log_2 d + 1 \rceil = f_z$ some constant

$$\text{PostParity}_z = 1 - \frac{(\mathbb{E}F - f_z 1_z)^T (\mathbb{E}F - f_z 1_z)}{\sum_{d=1}^{z} \left(\lceil \log_2 d + 1 \rceil - f_z \right)^2}$$

$PP_z = 0 \Rightarrow$ higher seed always wins

$PP_z = 1 \Rightarrow$ all seeds have same expected finish

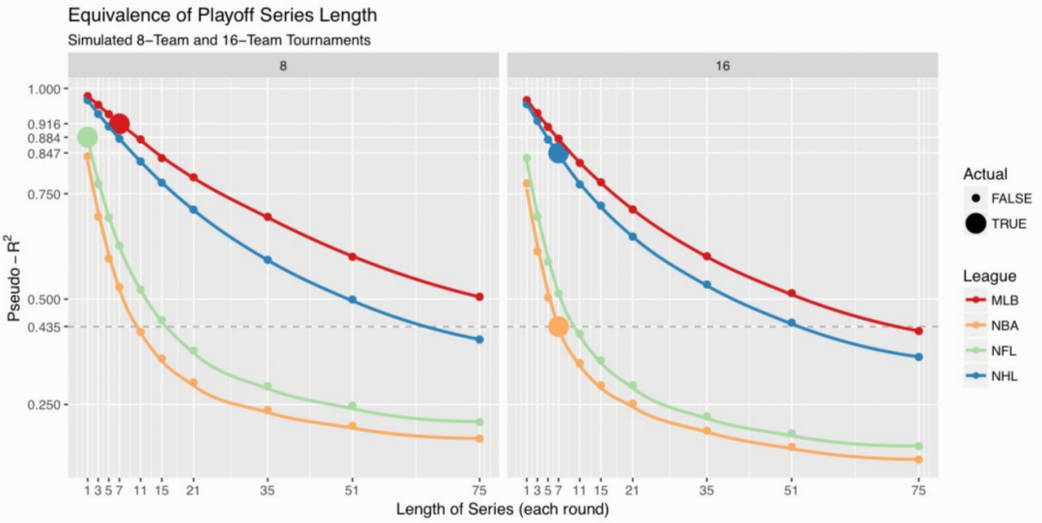

FIG 8. *Parity measures for simulated playoff tournaments. Each line shows how our pseudo-R^2 parity metric changes as a function of tournament series length for both 8- and 16-team tournaments in each sport. We note that in order for MLB to achieve the same lack of parity as the NBA, it would have to play 75-game series in a 16-team tournament. Conversely, the NBA would have to switch to an 8-team, single-game tournament to match the parity of the other three sports.*