# An Example of a fully Bayesian Analysis of Sport

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

Q Can we understand/compare differences in

- competitiveness (e.g. parity, win Prob.)
- home advantage
- variability of team strength
  - within a season
  - between seasons
  - game-to-game

across sports?

Approach fully Bayesian model to provide a unifying framework for contrasting the 4 major North American sport leagues.

Outcome: wins/win probability

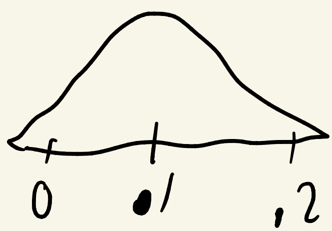
<u>Bayesian</u>: treat parameters as having a distribution

prior → belief abt the dist of the parameter before seeing data

then you see data

posterior → updated belief (dist.) of the parameter

$\beta$




<u>ex</u> Beta-Binomial

Prior: $p \sim \text{Beta}(\alpha, \beta)$

data $W \sim \text{Binomial}(n, p)$

→ posterior $p|W$

**outcome variable** the probability that team $i$ beats team $j$ in season $s$ during week $k$ of league $q \in \{NBA, NFL, MLB, NHL\}$

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

→ assume this is known and given by casino implied WPs.

**Home Advantage Parameters** (unobserved)

$\alpha_{q0}$ = league wide home advantage in sport $q$

$\alpha_{(q)i}^{*}$ = team-specific home advantage effect for team $i$ at game played in city $i^*$

center home advantages around 0 for identifiability $\sum_{i^*} \alpha_{(q)i^*} = 0$

$$\begin{cases} \alpha_{q0} = .5 \quad \alpha_{qi^*} = .1 \\ \alpha_{q0} + \alpha_{qi^*} = .6 \\ \alpha_{q0} = .4 \quad \alpha_{qi^*} = .2 \end{cases}$$

## Team Strength (unobserved)

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

→ can be translated into each team's probability of beating a league-average team

$$\sum_i \theta_{(q,s,k)i} = 0$$

# Fully Bayesian Model

* win prob. as a function of team strength & Home Adv:

$\alpha_{q0}$, $\alpha_{qi^*}$, $\theta_{(q,s,k)i}$, $\theta_{(q,s,k)j}$

$$\mathbb{E}\, P_{(q,s,k)ij} = \text{Logistic}\left(\alpha_{q0} + \alpha_{qi^*} + \theta_{(q,s,k)i} - \theta_{(q,s,k)j}\right)$$

$$\downarrow$$

$$\text{Logit}(z) = \text{Logistic}^{-1}(z)$$

$$\text{Logistic}(z) = \frac{1}{1+e^{-z}}$$

$$\text{Logit}(z) = \log\left(\frac{z}{1-z}\right)$$

$$\mathbb{E}\, \text{Logit}\left(P_{(q,s,k)ij}\right) = \alpha_{q0} + \alpha_{qi^*} + \theta_{(q,s,k)i} - \theta_{(q,s,k)j}$$

$$\downarrow \text{Bayesian}$$

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

→ Likelihood: given params, what is the likelihood you see the data?

$$\begin{cases} \text{before: } \hat{p} \\ \text{now: entire dist. of } p \rightarrow \text{approx. methods} \end{cases}$$

Need prior distributions on our parameters:

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

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

$1$ → week 1

$\gamma_{q\text{-}szn} < 1$

$K(q)$ → final week of prev. szn

shrinking the strength params towards 0 at the start of next season

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

shrinks the strength params towards 0 (albeit slightly) from week to week

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

* Home Advantage Prior

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

$$\alpha_{q0} \sim \mathcal{N}(0, 10000)$$ → we don't know before seeing the data how large the Home Adv. effect should be.

* Prior for the auto-regressive params

$$\gamma_{q\text{-}szn} \sim \text{Unif}(0, 1.5)$$

$$\gamma_{q\text{-}week} \sim \text{Unif}(0, 1.5)$$

* Priors for variance params

Let $\tau^2_{q\text{-}game} = \dfrac{1}{\sigma^2_{q\text{-}game}}$, $\tau^2_{q\text{-}stn}$, $\tau^2_{q\text{-}week}$, $\tau^2_{q\text{-}\alpha}$

$$\tau^2 \sim \text{Uniform}(0, 1000)$$

* But how to fit the model?
How to actually estimate the posterior distribution of all these parameters?

↓

Use MCMC methods
(Markov Chain Monte Carlo)

- Gibbs Sampling
- Hamiltonian Monte Carlo → STAN
- NUTS (no U-turn sampling)

Shane's class

Then MCMC methods

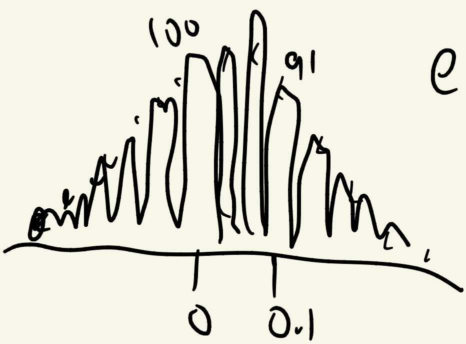
take the data
do a shitload of sampling

↓

out pops a full posterior dist. on all the parameters

↓

posterior samples




e.g. 8000 post. samples of each parameter

Output of MCMC:

for each league $q$,
get posterior dists

$$\begin{cases} p(\alpha \mid \text{data}) \\ p(\theta \mid \text{data}) \\ p(\gamma \mid \text{data}) \\ p(\sigma^2 \mid \text{data}) \end{cases}$$

finally, can go back to the Wins scale
via

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

estimation $\rightarrow$ estimate the params $\alpha, \theta, \gamma, \sigma^2$

attribution $\rightarrow$ what do these params imply about the nature of sports?

prediction

Team Strength coefficients over time



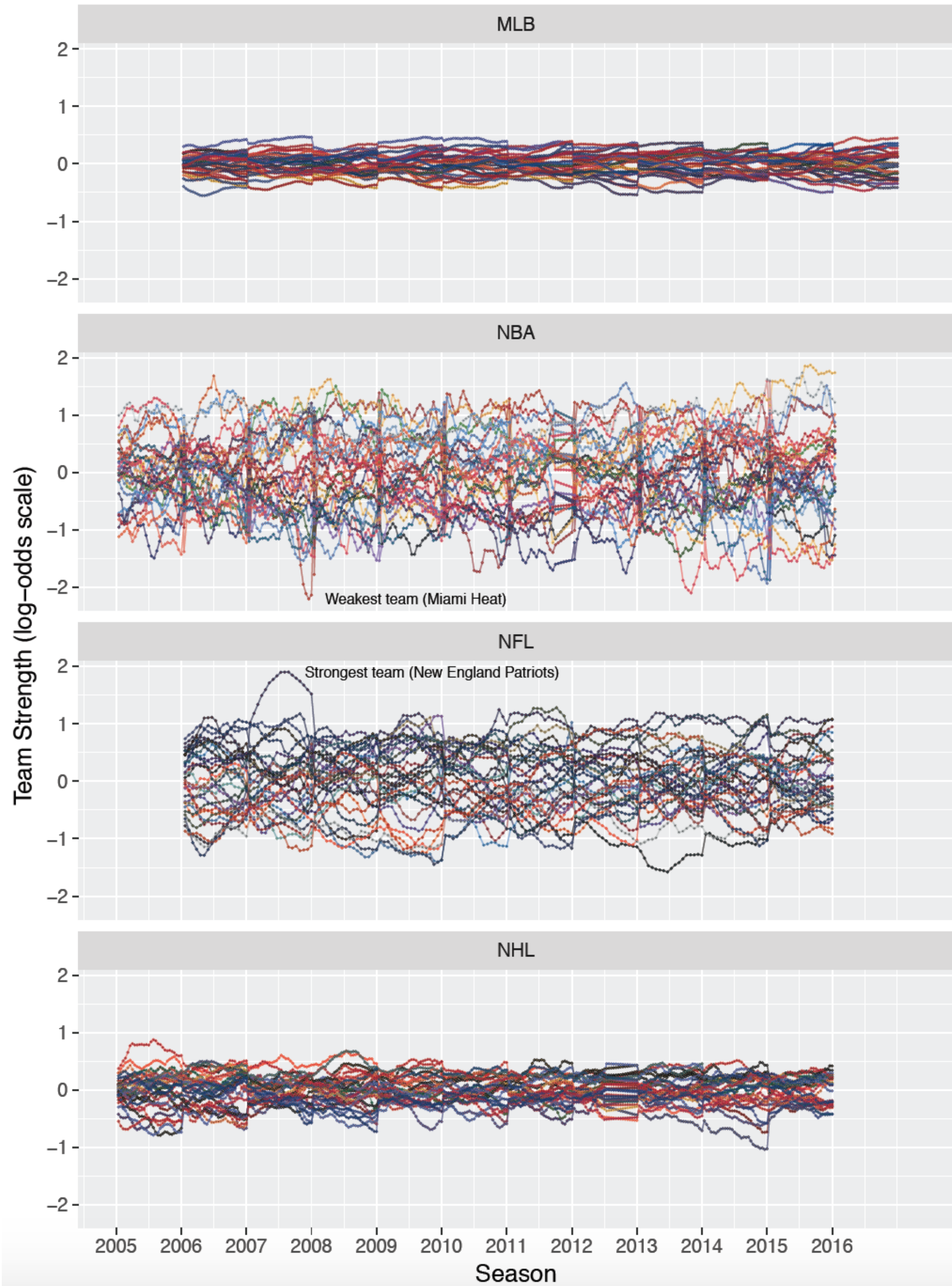



FIG 4. *Mean team strength parameters over time for all four sports leagues. MLB and NFL seasons follow each yearly tick mark on the x-axis, while NBA and NHL seasons begin during years labeled by the preceding tick marks.*

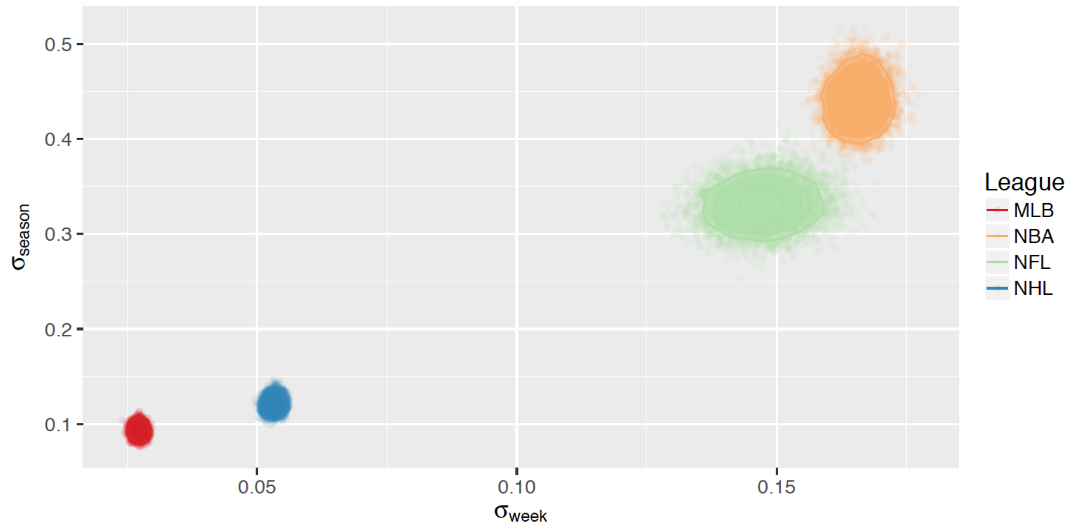



FIG 18. *Contour plot of the estimated season-to-season and week-to-week variability across all four major sports leagues. By both measures, uncertainty is lowest in MLB and highest in the NBA.*

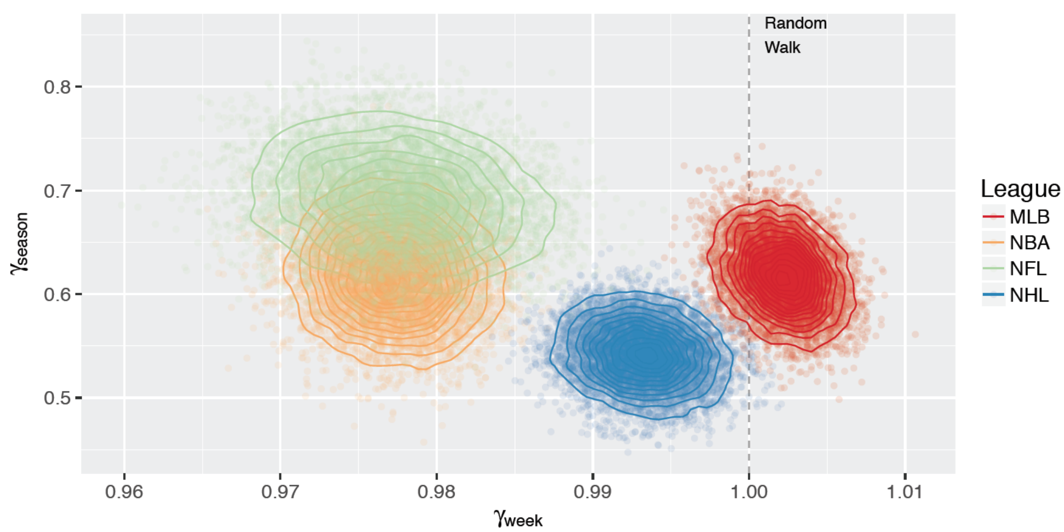
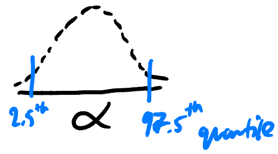



FIG 19. *Contour plot of the estimated season-to-season and week-to-week autoregressive parameters across all four major sports leagues.*

# Home Advantage





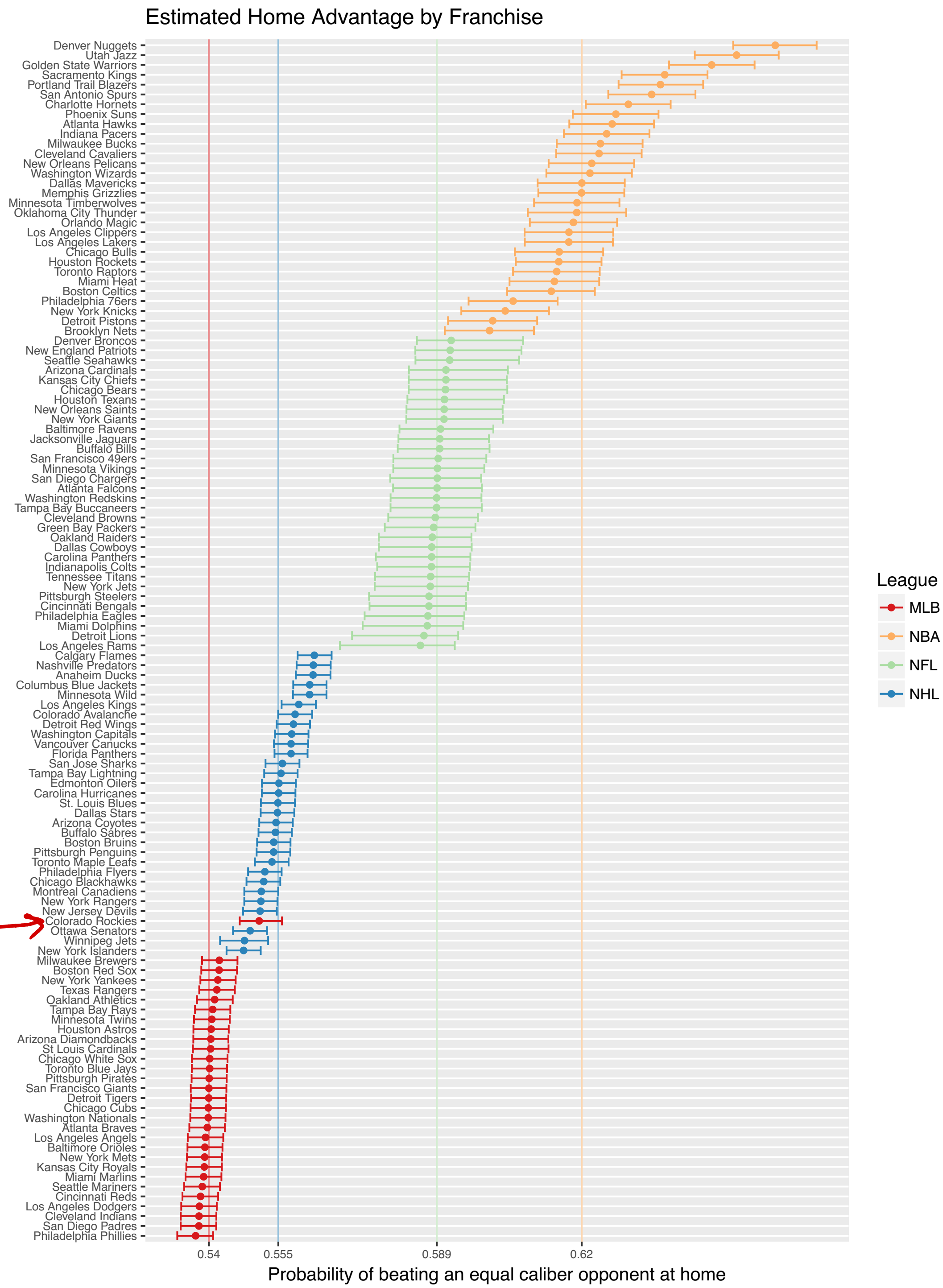




FIG 5. *Median posterior draw (with 2.5th, 97.5th quantiles) of each franchise's home advantage intercept, on the probability scale. We note that the magnitude of home advantages are strongly segregated by sport, with only one exception (the Colorado Rockies). We also note that no NFL team, nor any MLB 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 $\hat{\tilde{p}}_{q,sim} = \hat{\tilde{p}}_{(q,\hat{s},\tilde{k})\hat{i}\hat{j}}$ where $(\tilde{s}, \tilde{k}, \hat{i}, \hat{j})$ are sampled from observed schedules and $\tilde{p}$ sampled from posterior dist.

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

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

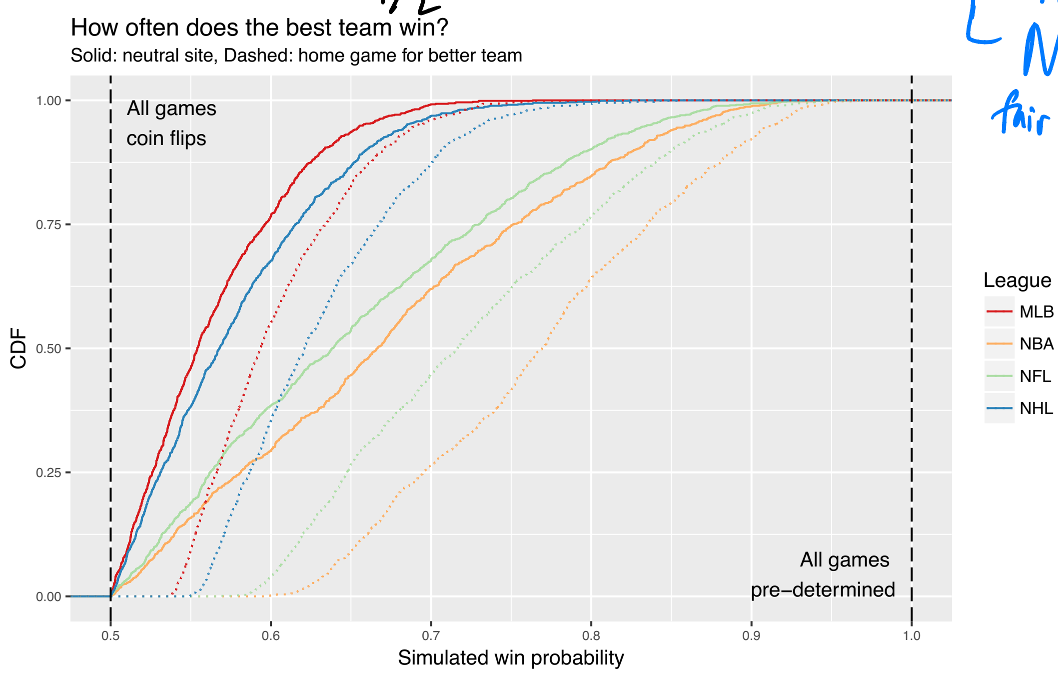



FIG 7. *Cumulative distribution function (CDF) of 1000 simulated game-level probabilities in each league, for both neutral site and home games, with the better team (on average) used as the reference and given the home advantage.*

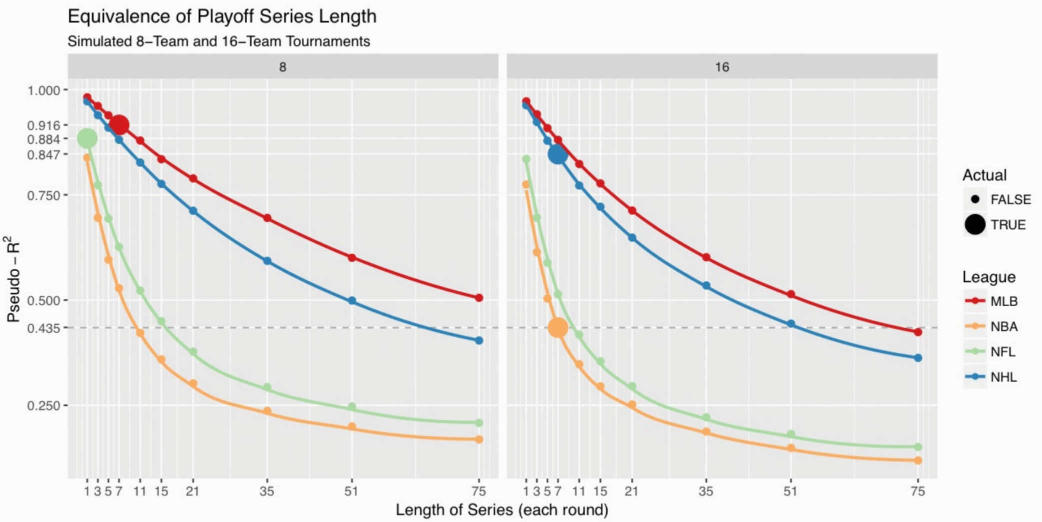



Fig 8. *Parity measures for simulated playoff tournaments. Each line shows how our pseudo-$R^2$ parity metric changes as a function of tournament series length for both 8- and 16-team tournaments in each sport. We note that in order for MLB to achieve the same lack of parity as the NBA, it would have 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.*