To What Degree Does Total Salary Impact MLS Club Performance?

(and a very gentle, non-technical introduction to Multi-Level/Hierarchical Models)

Josh MacCarty jmaccarty@gmail.com https://twitter.com/Camus_Cthulhu_1
https://www.linkedin.com/in/josh-maccarty-3133431b/

March, 2022

Contents

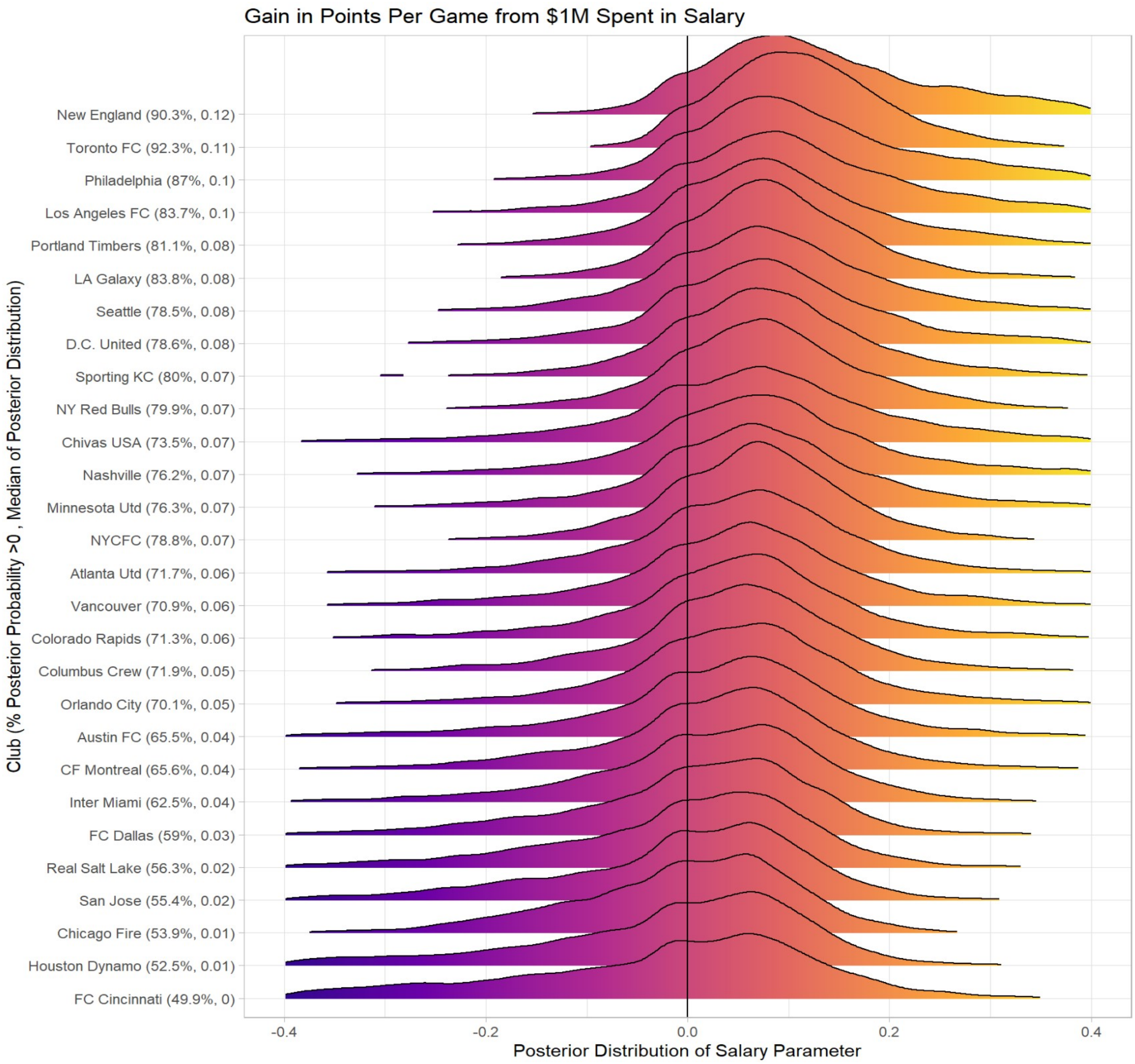

Figure 1: A preview of what's to come.

Introduction

MLS is one of the few soccer leagues in the world with a Salary cap and is perhaps the only league in the world with the DP, GAM, TAM rules. Although these rules frustrate general managers[1], they provide ample opportunity for analyzing MLS Club performance, based on salaries (both total Salary and the distribution of Salary within a Club).

The main question we explore is: *Does total Salary (taking into account the evenness of Salary distribution across players), impact a Club's performance? If so, do certain Clubs get more performance return for each dollar spent in Salary?* We source data from FBREF.com and the MLSPA, explore a causal diagram to ensure our modeling choices are logical, then fit a Bayesian multi-level varying effects regression. This sounds fancy, but it is surprisingly straight-forward, and if you are interested in soccer and/or analytics, this analysis will be a gentle introduction to multi-level/hierarchical modeling–why it probably matches your intuition and why it's often necessary and not just a "nice-to-have".[2]

If you only read the "TL;DR Findings" section without reading the rest of the report, we get it, just make sure you read the "Limitations of This Analysis" section before hitting the little X button.

TL;DR Findings

- Salary
  - Unsurprisingly, Salary is not immediately causal of better performance (PPG) for every Club in MLS, but there is some evidence that certain Clubs see a positive return for increases in Salary.
  - Evidence suggest that New England, Toronto FC, Philadelphia Union, and LAFC all see > .1 PPG increase with each $1M increase in Salary.
- Evenness of Salary Distribution
  - Club Salary Evenness is mixed in its contribution to PPG–most Clubs see no evidence of material gain from Salary Evenness increases.
  - However, Seattle and RBNY are the possible exceptions and see stronger evidence that their increases in Salary Evenness has led to increased PPG.

Outline

Due to the nature of the question at hand and the reality of the data, finding an answer (even partially) requires a good bit of parametric statistical modeling, therefore we will go into what non-analytical folks would see as highly technical and what highly technical people would see as the bare minimum to present a convincing argument that the findings are robust enough to warrant a conversation. **Non-analytical folks can skip portions noted with "Technical" precursors.**

The "paper" here covers the data used, methodology and motivation for the methodology, simulation results, and results from the actual model.

[1] https://theathletic.com/3162180/2022/03/03/2022-mls-Club-executive-anonymous-survey-candid-views-on-owners-coaches-players-and-cheating-around-the-league/
[2] https://xcelab.net/rm/statistical-rethinking/

Data, Methodology, and Simulation Results

There are many ways to approach the question of if increases in Salary or Salary Evenness “relates” to better Club performance, but fewer ways to approach this problem with techniques to attempt to establish a *causal* relationship between Salary and Salary distribution and not just an “association”, “correlation”, or any other “hedging” type language used in a cursory exploration of the data.

To that end, we begin with a Causal Diagram i.e. a Directed Acyclic Graph (DAG) to make sure there were no unaccounted for (open) backdoor paths or included colliders between the Salary and Salary distribution evenness (more on this in a second) nodes and Club’s performance (points per game, more on this as well in a second). After the DAG is constructed, we simulate data that the DAG would generate if the DAG does in fact capture the true underlying data generating process. We then fit a model to this data to make sure that the model specification we chose for the *actual data* would capture the true dynamics of the *actual data* because we *know* the simulated data generating process. This is an important step in any data analysis. After confirming model appropriateness for simulated data, we move onto model results using the actual data.

Data

1. **Salary**
    - Sourced directly from MLS Players Association’s Salary guide for the years 2010-2021.[3]
2. **Evenness of Salary Distribution** (shortened to ‘Salary Evenness’ from here)
    - Information Entropy is one way to measure the evenness of a distribution of Salary. Yes, entropy is technically a measurement uncertainty, but this is an easy conceptual extension to “evenness” of a (Salary) distribution, when you think of “given a Club’s distribution of Salary dollars to all of its players, for a given dollar, how certain are we that it came from a specific player”.
        - *E.g. For a Club with 5 players, each on $50,000 salaries, for any single dollar of salary, your uncertainty of whose dollar that is would be high because the Salary distributions are very even. In contrast, a Club of 5 with one person on a $1,000,000 and four others on a $1 Salary, for any specific dollar, you’re fairly certain that its coming from the person with the $1,000,000 Salary. Therefore the Salary Entropy (or Evenness) would be high in the first example, and low in the second example.*
    - We use Salary Entropy because it is both concise (a single number) and reflects the evenness of the salary distribution over all the players.
    - The important part to rememberhere is that as Salary Entropy increases, the Salary Evenness increases.
3. **Club performance**
    - Club performance, for the sake of this analysis, is measured by Points Per Game (PPG) from FBREF.com[4] for each individual season, for all Clubs during the time they existed in the league. PPG over an entire season, for a Club, should both reflect offensive capabilities, defensive capabilities, competitor levels, and really is the end goal of a Club for a season (maximize points) all in a single number. Why not Expected Goals or other more advanced metrics? Two reasons… First, FBREF only has expected goals for the last few seasons. Second, we wanted to avoid the Expect Goals debate in this analysis (though, rightfully so, most people have understood and are applying Expected Goals to understanding game dynamics).
4. **Date Range**
    - All data is from 2010 to 2021, though not all Clubs existed in the entire date range.

[3]https://mlsplayers.org/resources/Salary-guide
[4]https://fbref.com/en/comps/22/Major-League-Soccer-Stats

5. **Data Not Used**
   - More goes into a roster than Salary, there is guaranteed comp, bonus, non-recorded compensation, transfer prices, etc. Much of this is not recorded (though, as noted at the end of the analysis, it would be easy to repeat this modeling with listed guaranteed comp).
   - Many things impact a Clubs performance, not just Salary and how that Salary is distributed (e.g. tactics, injuries, a Clubs ability to work together, etc.) but using the analytic method we chose, much of that "unobserved/other stuff" can be at least partially accounted for. Read on to see how, but essentially, if you are still reading and are thinking **but, that doesn't take into account X, Y, Z**, just wait to see how we do in fact take those things into account.

Technical: Data

1. Salary was scaled to millions of dollars to proactively prevent any numerical overflow.
2. Evenness of Salary Distribution (Information Entropy, or Entropy of Salary distribution within each Club, each year) = $\sum_{n=1}^{20} \mathcal{P}(x_n) * log(\mathcal{P}(x_n))$ where $\mathcal{P}(x_n)$ is an individual player's Salary divided by the total Clubs Salary (for that year). We limit each Club to be their top 20 salaries to make sure the Salary Evenness is not biased by unbalanced observations per Club per year. Future analysis should include DP salaries, etc. For now, top 20 salaries on a Club for a season will still reflect the differences between Clubs and years.

Methodology

Causal Diagram/Directed Acyclic Graph (DAG)

When the goal is to establish a causal relationships between, well, anything, the first step is constructing a causal diagram – formally called a Directed Acyclic Graph (DAG). We will not go into all of the reasons, but there are wonderful sources covering this topic.[5] [6] Our first DAG below outlines the basic understanding of a Clubs performance.[7] The problem with this "reality" is that if we can't find some way to include these factors or determinants of Club Performance (row of grey nodes at the top of the visual) in the model, then we'll never get an accurate read on the causal impact of Salary and Salary Evenness on Club Performance. This DAG shows us that in our model of reality, many factors cause both Salary variables, and Club Performance in a single year, like the planned tactics, player quality, competitors and a whole host of other, unobserved factors like Club strategy, Club financial positioning, market interest, etc.

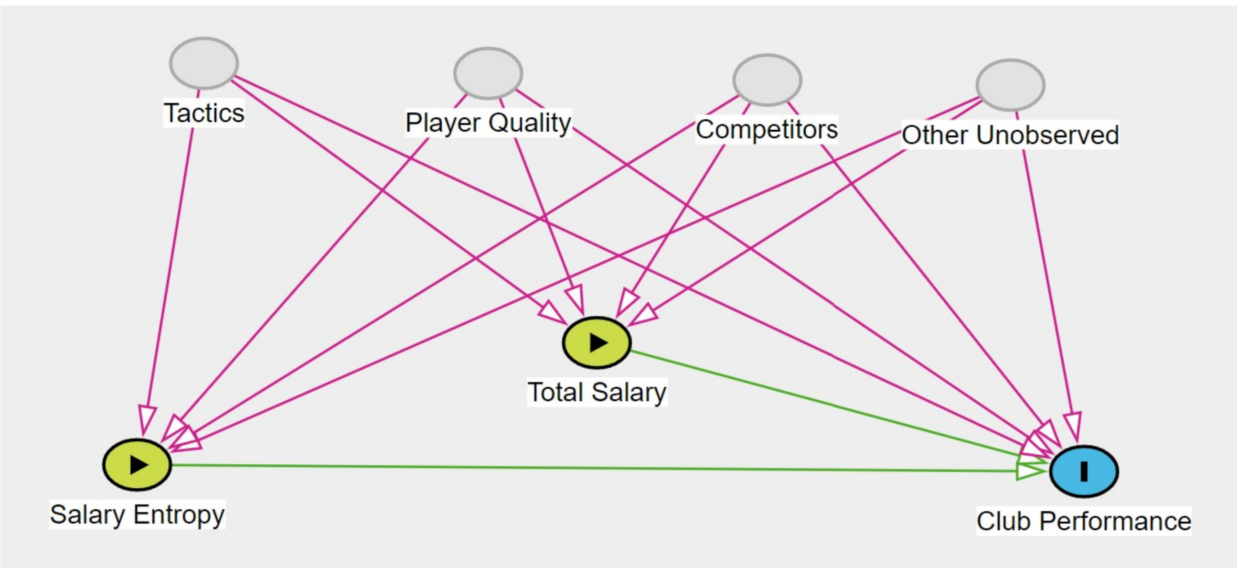

Figure 2: DAG showing an example of variables that bias our estimation of the impact of Salary and Salary Entropy on a Club's performance. The grey nodes (i.e. factors, nodes, variables), if not included somehow in our model, lead to incorrect relationship estimates between what we are interested in (Salary, Salary Entropy impact on Club performance). These are called open backdoor paths in DAG-language, and are highlighted in red.

If we were able to observe an exogenous shock to Salary or Salary Evenness (that is to say, from "outside" our mapped out DAG) then we would have a set of analytical techniques available to us like event studies, difference-in-difference, randomized control experiments, matching, synthetic control, etc. but all of this is unfeasible given the nature of an MLS Club. Luckily for us, there are other options for still quantifying the causal relationship between Salary and Salary Entropy and Club Performance (like, legitimately an entire field(s) centered around observational studies). One such technique, which we utilize here, is called a "fixed-effects" regression (we'll use something similar, but we'll save that for a "Technical" section). Multi-level/Hierarchical/"fixed-effects" regressions, effectively, combine each of the confounding, unobserved variables in grey into a single node like the DAG below and *then takes that node into account.*

[5]https://www.goodreads.com/book/show/36204378-the-book-of-why
[6]https://theeffectbook.net/
[7]dagitty.net used for all DAG graphs

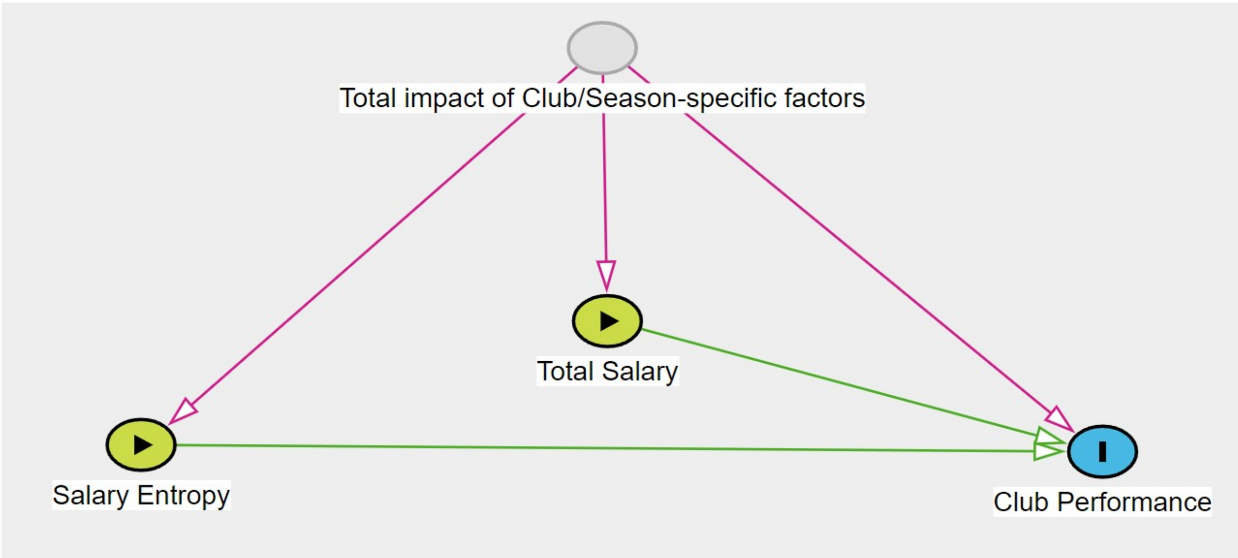

Figure 3: DAG showing the grouping of unobservable, confounding varibles into a single node.

Model Used

What a fixed-effects regression says in our scenario is that each Club in each year has a list of factors that will partially determine their Salary, Salary Evenness, and Club Performance, and those set of factors can be lumped together, and quantified by a single number that can be learned from the data via the model.

Example: Picture the Philadelphia Union having a specific year where a number of academy players all mature and hit the first team that year with enough ability to make a serious impact on their season's performance. At the same time, some veterans get injured. Similarly, in that season, Jim Curtin's 4-4-2 Diamond has some tactical innovation to make the Club more effective. For that year the "fixed-effect" learned from the data for the Club, essentially, quantifies the cumulative impact of these three things (wave of new talent, injured vets, tactical innovation).

This "fixed-effect" is learned from the data, via the model, for each Club for each year, in a way that is allowed to vary from year to year and Club to Club, but also learns across Clubs and years while still allowing that year for the Club to reflect its own idiosyncrasies (quick technical note: this is the brilliance of "Partial Pooling" and why multi-level models are necessary more often than a simple linear regression). The key here is that in allowing the "fixed-effects" for each Club/year to vary, we can isolate each specific Club's relationship between Salary and Club performance. The "fixed-effect" ideally, when chosen properly, from a DAG that reflects reality, closes the back-door path that confounds and incorrectly quantifies the relationships of interest.

Let's take a look at the visual intuition for what the "fixed-effect" is doing. Imagine you took all of the Clubs and years and plotted Salary vs. PPG, and ended up with something like the figure below. There is a strong negative relationship between Salary and PPG–which really makes no sense. But if your condition on Club and re-plot the scatterplot by coloring and trending for each individual Club, with the exact same data, you see something much more sensible: mostly positive relationships between Salary and PPG. The model that we built finds these varying intercepts and varying slopes for each Club in the relationship between Salary (and Salary Evenness) and PPG. This is nothing new or fancy, but simply a visual example of why modeling all Clubs together in a standard linear regression could yield incorrect results. This phenomenon goes by many names like "Bias of Aggregation", "Fallacy of Ecology", and "Yule/Simpson Paradox".

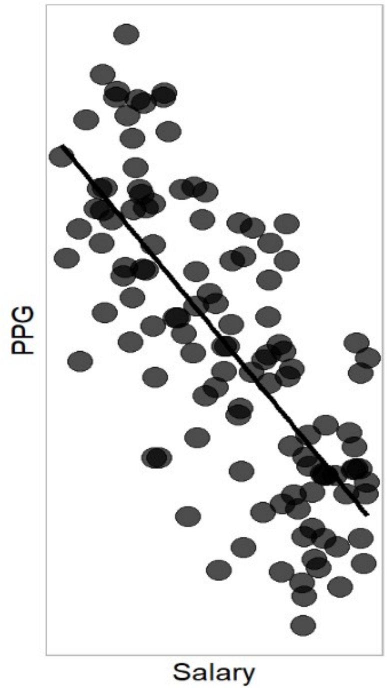

Figure 4: Salary vs. PPG for all Clubs over all years, in the simulated data set shows a strong negative relationship

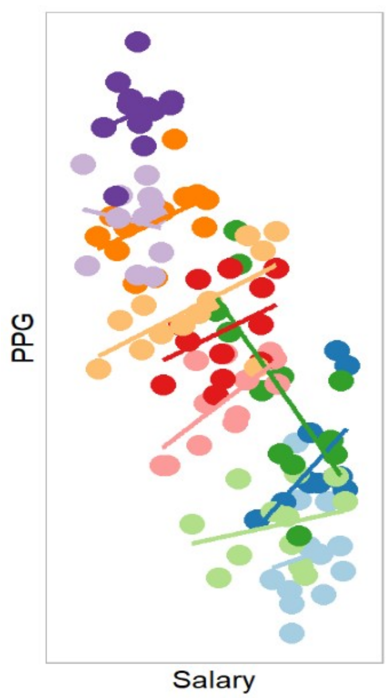

Figure 5: Salary vs. PPG showing the exact same data as above but conditioned/colored by Club, in the simulated data set shows Club-specific intercepts and slopes.

This is the most obvious reason why multi-level/hierarchical regressions are needed (when possible). Just throwing the data into a linear regression and "hitting run" would yield poor results–even if they *looked* correct. However, if we specify our model to capture these dynamics like varying slopes and intercepts, then we have a hope in accurately estimating the relationship between things like Salary and Salary Evenness and PPG. The next (technical) sections outlines how we specifically try to do just that.

Technical: Model used

We use a straight-forward Bayesian multi-level varying effects regression with varying intercepts for each Club/year, and a random effect for each Club's Salary and Salary Entropy. We also include an dummy variable for a Club's first season with no Club-specific index.

$$PPG_{i,j} \sim Normal(\mu_{i,j}, \sigma)$$
$$\mu_{i,j} = \alpha_{i,j} + \beta_i Salary_i + \gamma_i SalaryEntropy_i + \psi FirstYearFlag$$

The $\alpha_{i,j}$ is a matrix of α parameters for each i Club and j year − this is a cross-classified varying intercept, while β_i and γ_i are varying slopes of the Salary and Salary Entropy based only on the Club (and specifically not by year). This was a modeling choice to help answer the question of interest, viz. if Clubs Salary and Salary Distribution are driving performance. This model structure allows for each Club to have a varying intercept by year–the main mechanism by which we are accounting for the unobserved node in the DAG, and thus, ideally, closing the backdoor path and blocking the confounding effect of Club/year-specifics on the causal path from Salary and Salary Entropy to Club Performance.

$$\alpha_{i,j} \sim Normal(\bar{\alpha}, \sigma_a)$$

$\alpha_{i,j}$ are the Club/year intercepts whose prior is normally distributed with $\bar{\alpha}$ (the mean) being an adaptive prior to be learned from the data, and a standard deviation of σ_α. These are not fixed effects like mentioned above, but are random − again, allowing partial pooling to work its magic-not-magic.

$$\begin{bmatrix}\beta_i \\ \gamma_i\end{bmatrix} \sim MultiVariateNormal\left(\begin{bmatrix}\mu_\beta \\ \mu_\gamma\end{bmatrix}, \mathbf{S}\right)$$

is a matrix of varying slopes with a prior distribution that is multi-variate Gaussian with means μ_β and μ_γ and a covariance matrix $\mathbf{S}$ that can be factored into a matrix of standard deviations and correlations (below). We model both the impact of Salary and Salary Evenness together due to potential error correlation–essentially allowing both of the parameters priors to be a joint normal can help "learn more efficiently".

$$\mathbf{S} = \begin{pmatrix}\sigma_\beta & 0 \\ 0 & \sigma_\gamma\end{pmatrix} \mathbf{R} \begin{pmatrix}\sigma_\beta & 0 \\ 0 & \sigma_\gamma\end{pmatrix}$$

$$\mathbf{R} \sim \textit{LewandowskiKurowickaJoe}\left(2\right)$$

$$\bar{\alpha} \sim Normal(1.4, .2)$$

is a hyper-prior for the mean of the $\alpha_{i,j}$, with a wide variance.

$$\beta \sim Normal(0, \sigma_\beta)$$
$$\gamma \sim Normal(0, \sigma_\gamma)$$
$$\sigma_\alpha \sim \textit{Exponential}(1)$$

$$\sigma_\beta \sim Exponential(1)$$
$$\sigma_\gamma \sim Exponential(1)$$
$$\sigma \sim Exponential(1)$$

Then finally the fixed priors, including an informative prior for first season at the Club:

$$\psi \sim Normal(-.2, 0.05)$$

Arguments Opposing this Model (not just strawmen)

One could argue that some sort of indexing or centering the Salary for each Club within each year tell us more than raw Salary since inflation and league/club resources have increased over time. We think there is validity in this idea and should be explored in the future, but also expect the varying effect over Club/Year to partially take this into account.

Technical: Arguments Opposing this Model (not just strawmen)

One thing that will jump out to technically-savvy readers will be the inclusion of both Salary and Salary Evenness (Entropy) in the same model. Wouldn't these two variables be highly correlated (leading to issues of multi-collinearity in the model)? Theoretically, one would imagine there is a strong negative correlation between Salary and Salary Entropy because, well, entropy–that is to say there are many ways to increase a Club's Salary that would decrease entropy (e.g. increasing any one players Salary would do this), but only in the case where increasing Salary such that every players Salary are more similar to every other player would increasing Salary lead to increased entropy. The figure below shows just this relationship: increases in Salary (x-axis) are correlated with a decrease in the Salary Entropy (Y-axis), but that's not the end of the story.

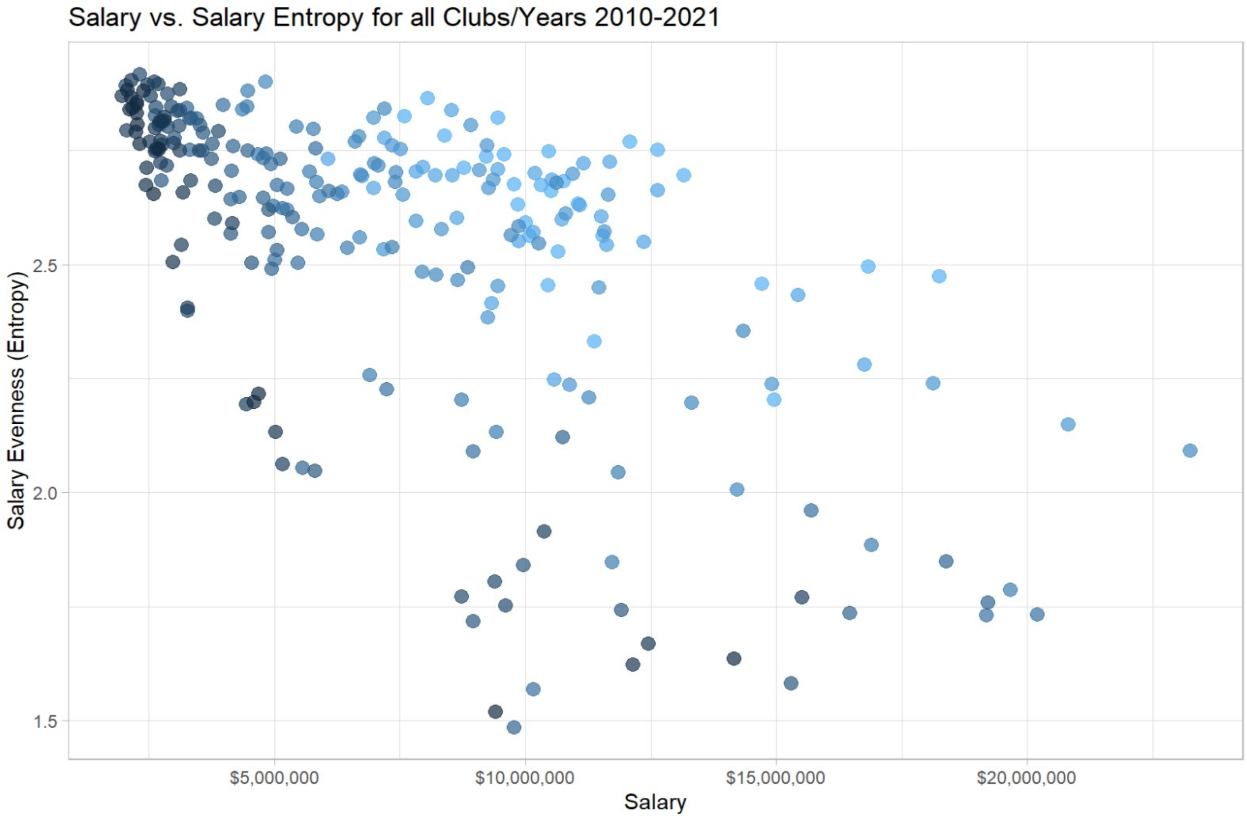

Figure 6: Salary vs. Salary Entropy for every Club/Year. The lighter the season, the more recent the data point. This shows a generally negative relationship between Salary and Salary Entropy, as expected. Also note, in more recent years, Salary Entropy (Evenness) is higher.

However, it's *not* the unconditional correlation that matters when assessing if multi-collinearity is an issue for a model, but the correlation of the variables after conditioning on other variables (or structure) of the model. Our model uses Club/Year random effects, but we can at least look at the Salary and Salary Entropy distribution after conditioning on the Club (figure below)–you can see that in many places the multi-collinearity disappears, and this is without even conditioning on year as well.

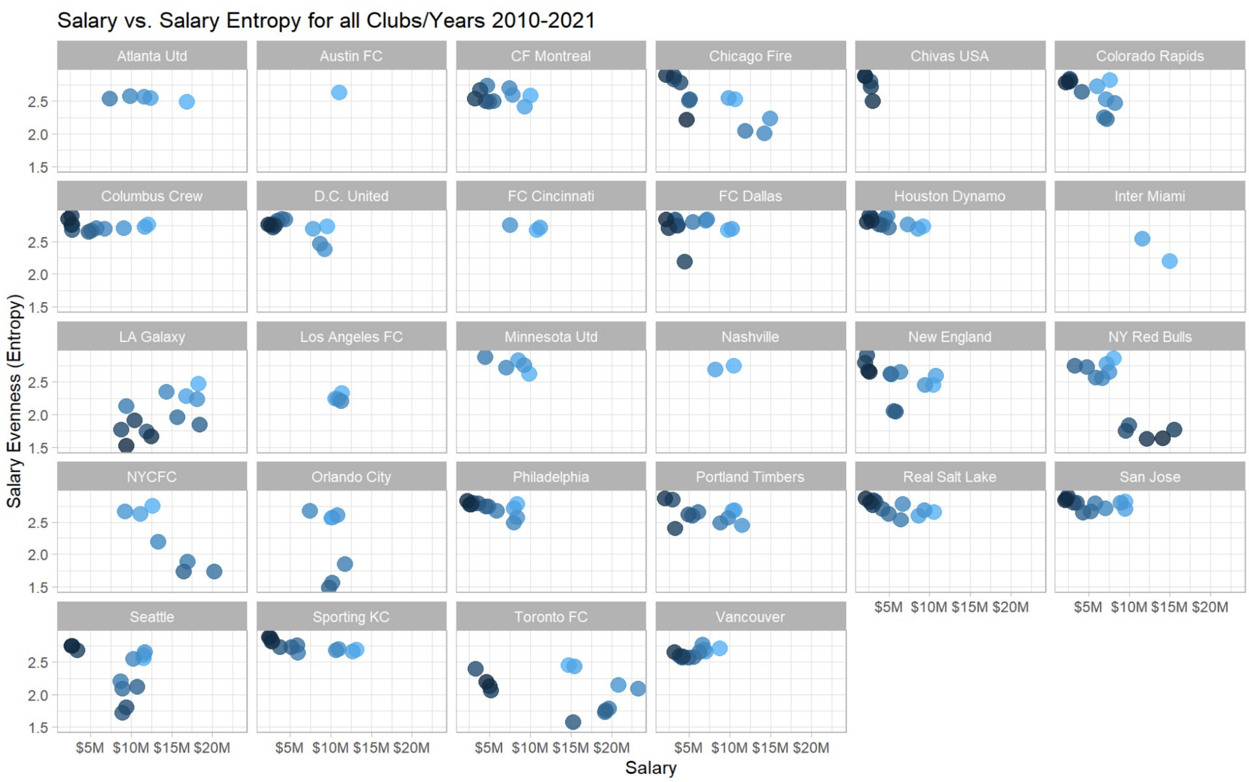

Figure 7: Salary vs. Salary Entropy for every Club/Year conditioning on Club. The lighter the season, the more recent the data point. Interesting to note (outside the scope of this analysis), the differences between Clubs over time – some are very consistent like Philadlephia or Sporting KC, and some are widely varying like Toronto FC.

Technical: Simulation

Not much detail here but for the sake of keeping the report as short as possible. But we did want to emphasize the necessity to simulate data to make sure that the model structure will accurately estimate the causal effects when we know the real effects (data generating process) because we simulated it. To that end, many data sets were simulated with different levels of relationship between Salary and Salary Entropy with PPG and modeled using the structure outlined above. At risk of sounding too repetitive, this is absolutely necessary when modeling. The idea is if we can force a relationship between Salary and Salary Entropy with PPG in a simulated data set, then our model should recover this relationship *and we can verify this* because we know the relationship. This is a way to stress test our model specification to make sure that the model will capture a real relationship, and that if our model shows no relationship between the variables of interest, we know it's not simply because the model is specified correctly (assuming our simulation is correct). Details on the simulation can be shared if interested. TL;DR of this section, sensitivity and robustness were tested.

Results

One way to test our model validity is to compare our models prediction PPG (technical: posterior predictive distribution) vs. the actual data's PPG. One important thing to keep in mind is that just because we can *predict* well, doesn't mean we've captured anything relevant to the *causal structure*. However, some degree of predictability is expected–which is what we want to check.

As always, our parameter estimates are conditional on the data and the model. If you specify the model differently, or change scales or centering of the data, the results will differ. If you don't believe the DAG, or that the model structure (varying effects) closes the back door confounding path, then the estimates from the model are irrelevant to you and you can stop reading. However, if you believe the DAG structure, and agree with the model structure, then interpreting the relationships between Salary and Salary Entropy can be achieved by looking at the median and variance of the distribution.

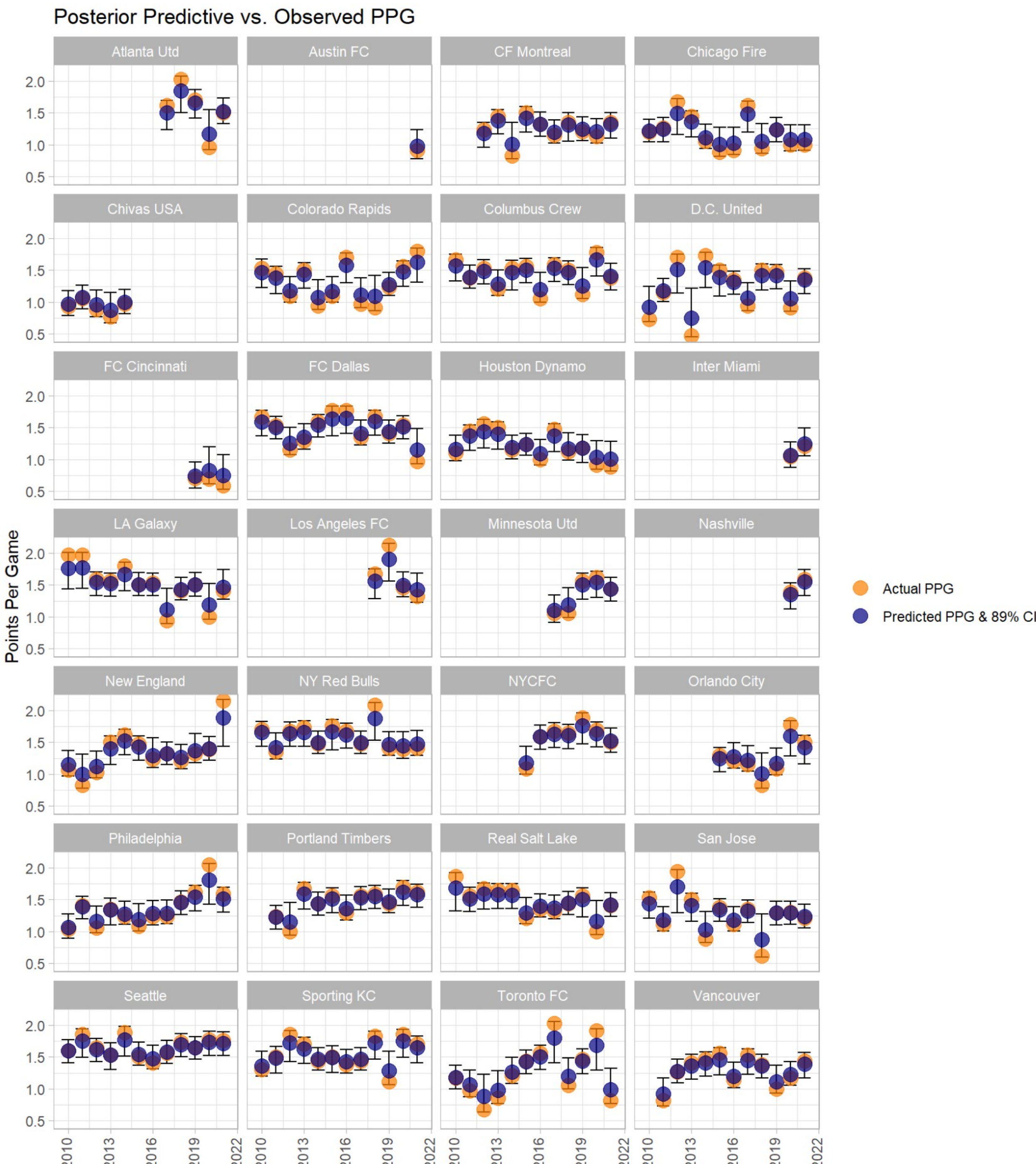

Figure 8: Actual PPG is in orange and the posterior prediction mean and standard errors are in blue/black. It's clear the model predicts well–for each year the prediction is close to the actual PPG. You'll notice that when the PPG is especially high (low), then the prediction is slightly lower (higher)–this is not a bug, this is an example of Bayesian Shrinkage–and is a strength of predicting using multi-level Bayesian models–and is a direct effect of partial pooling.

Salary

Below is a visual of the Club-specific $\beta's$ from the model, ordered from greatest to least probability greater than 0 (in some statistical settings this is used to "calculate" statistical significance but that term is forbbiden here ;)).

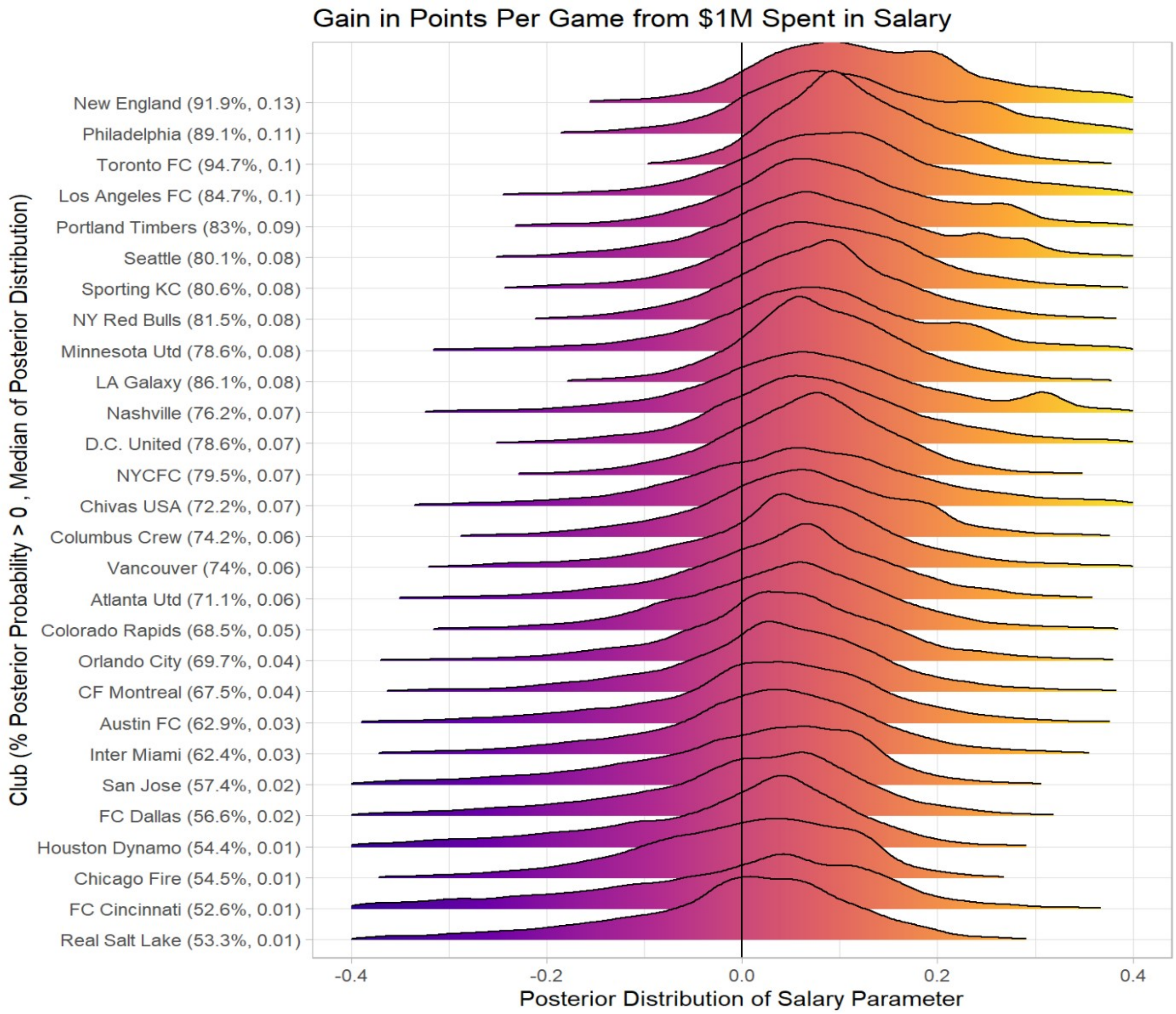

Figure 9: Distributions show the β (the PPG return for $1M spent) for each Club. Along the Y axis you can read the Club name, the median of the posterior distribution, and the amount of the distribution that is > 0. Clubs are ordered from most posterior β distribution greater than 0. If you are new to analytics, the distribution of the parameter is not a deficiency of the model but a strength of the approach–it helps us understand not just "our best guess" (e.g. median), but also the uncertainty around that guess (variance, or "width" of distribution)

Notes on Salary's impact on PPG:

- Since the data was not pre-centered, we would imagine PPG would be 0 or positive for each Club–the findings are consistent with this (i.e. distributions are centered around 0 or centered to the right of 0).
- PPG increase driven by Salary (again, conditioned on our DAG being correct, and model specification being sufficient) differs from Club to Club with New England, Philadelphia, Toronto FC, and LAFC all seeing at least 0.1 PPG increase for each $1M in Salary.

- Clubs with more variance in Salary and have existed longer have less variance in the parameter distributions (again, consistent with expectations)
- The only reason we have even reasonable distributions around newer Clubs like FC Cincinnati, Nashville, Inter Miami, Austin FC, etc. is due to the partial pooling (these Club estimates lean more heavily on the information "borrowed" from all Clubs)
- Remember, the Salaries prior was centered at 0–there was no strong/informed prior constraining non-negative values.
- As an MLS fan, is there anything surprising when looking at this figure?

Salary Evenness (Entropy)

The model intentionally included a metric of Salary distribution evenness (Salary Entropy) in a (simplistic) attempt to not assume just raw Salary impacts PPG but also how that Salary is spent. Of course, Entropy is not a complete picture of Salary Distribution but just one metric for Evenness of Salary distribution–that said, it is still an interesting metric to include and even identifies a few Clubs that (given the data, and the model) benefit from increasing the Evenness of their Salary distribution.

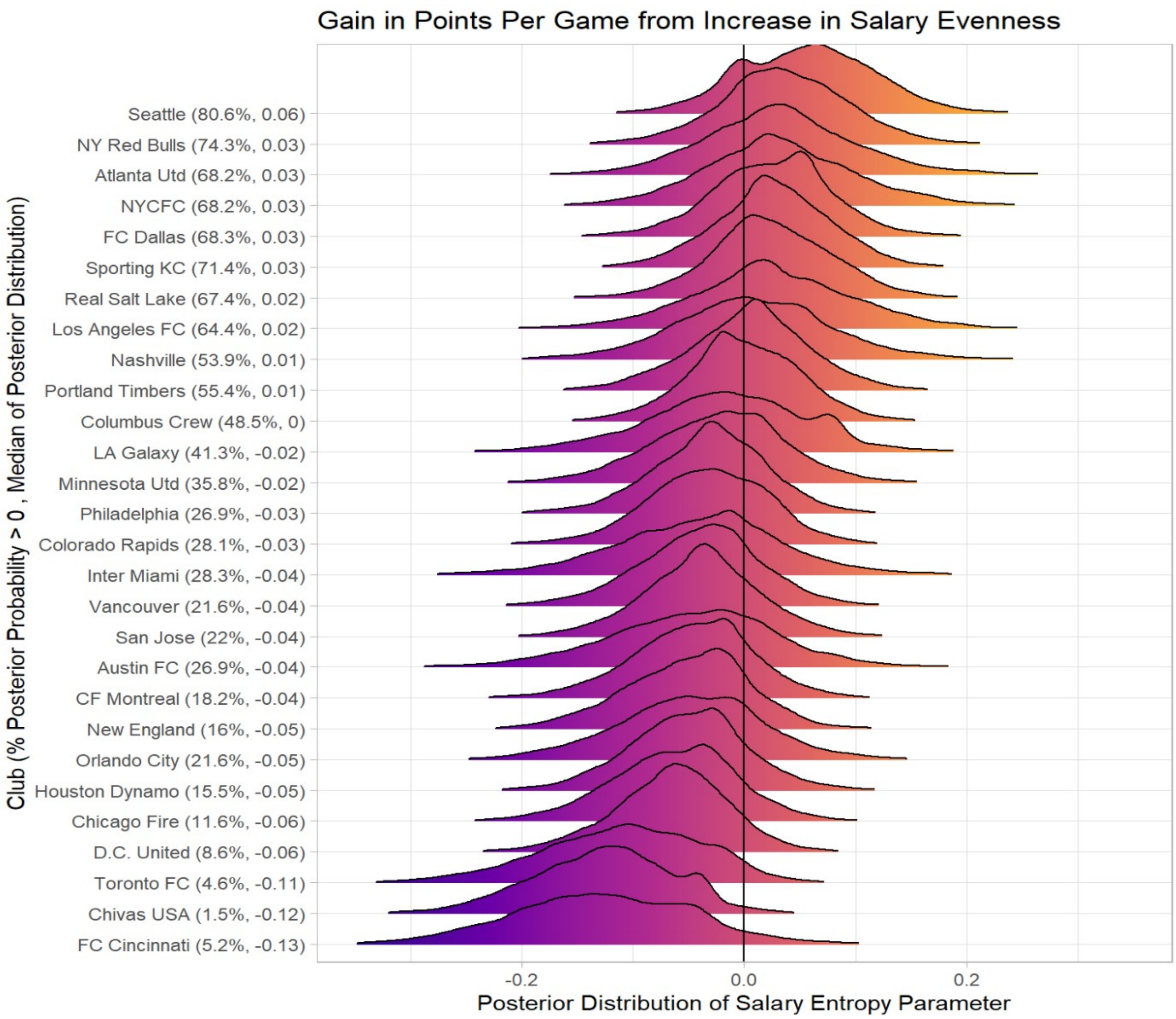

Figure 10: What does "Evenness" even mean from a unit perspective–Entropy is not a scale easily interpreted, so let's leave it at interpreting probability greater than 0.

Notes on Salary Evenness impact on PPG:

- Salary Evenness, like Salary, was not centered, so seeing posterior parameter distributions for Clubs on both sides tells us that overall, increasing the evenness of Salary differs drastically from Club to Club.
- Seattle is the clear standout at seeing a return on distributing their Salary more evenly, but RBNY also see evidence that evenly distributing Salary increases PPG.
- Like with Salary, Clubs with little change in Salary Evenness over time, or that have not existed long, have much more uncertainty around estimating their relationship between Salary Evenness and PPG.

- Salary Evenness was originally included as a variable to help better estimate the Salary expenditures (i.e. shrink the standard errors), so over-interpretation here is cautioned against.

Limitations of this Analysis & Future Improvements

- Entropy as a measure of Salary Evenness is a good metric, but evenness of a Salary's distribution across players for a Club in a season does not capture all the dynamics that go into Salary strategies at a Club. GAM, TAM, DP, etc. in MLS rules provide other interesting factors that could and should be integrated into an analysis in the future.
- With all statistical models, results are conditional on the data and the model specification–if you agree with the DAG, the model outlined, that the Club/Year varying effect appropriately closes the backdoor biasing/confounding relationship from Salary (and Salary Evenness) through the confounding factors to Points Per Game, and the data and data transformations used in this analysis, the results are meaningless.
- However if you agree the DAG is reasonable, the model is sufficient, and the data is real, then the observations may be relevant.
- Game-level data could provide (many) more observations and more opportunity.
- Surely PPG is not the only measure of Club performance–XG or difference between XG and XGA seems like a good dependent variable to explore.
- Many Clubs see little Salary Entropy variance–this should be reflected in the standard errors of the posterior distributions, but it is still important to note.
- Similarly, a few Clubs have only existed a few seasons, and any strong conclusions drawn from their results is also cautioned against.
- This analysis was just one author exploring some freely available data during their free time, there are (surely) many ways to improve this analysis beyond the few notes above–this was not peer-reviewed.[8]

[8] Throughout this write-up, the author uses "we" instead of "I" for the sake of attempting to sound less egotistical