Bayesian Ranking of Quizbowl Teams

Kuleen Sasse
Department of Applied Math and Statistics
Johns Hopkins University
Baltimore, MD 21218
ksasse1@jh.edu

Abstract

Most approaches to quizbowl rankings use hand crafted features and models. However, this process allows for implicit biases to be embedded within a ranking framework, possibly leading to weaker models. In this report, we introduce rigorous statistical inquiry to quizbowl team rankings through the use of the Bradley-Terry model. We scrape over 300,000 matches from National Academic Quiz Tournaments (NAQT) as our dataset. We compare the differences between the Bayesian and frequentist formulations of the Bradley-Terry model. We find that the Bayesian formulation of the Bradley-Terry model is vastly superior to its frequentist version, ranking teams more in line with the current consensus. [1]

1 Purpose

Quizbowl is an academic trivia game that is extremely popular in the United States especially amongst middle schools, high schools, and colleges. As it is a game with many teams, one might ask a simple question: Which teams are the best? Many ranking methods have been proposed throughout the years. However, most use personal intuition about the game and hand crafted models. The purpose of this report is to begin the process of introducing rigorous statistical modeling techniques to quizbowl rankings.

2 Background and Terminology

As quizbowl is a fairly niche hobby, we will spend this part of the paper explaining the structure of a quizbowl match, how tournaments work, and the important statistics recorded during a match.

2.1 Quizbowl Match

Each quizbowl match consists of two teams of 4 players each. During the course of a match, a moderator reads out twenty toss-up questions. These toss-up questions can be about any academic subject like science, fine arts, literature, etc.. During these toss-up questions, any player can interrupt the moderator with a buzzer as soon as they know the answer. Players cannot consult with their team during these questions. If a player gets the toss-up question correct, they are awarded 10 points. They can get an additional 5 points if they answer it early enough. This is called "Powering" a toss-up question. In addition, getting the toss-up question entitles the player and their team to a set of three bonus questions each worth 10 points for a total of 30 points. These bonuses are read one at a time and are related by a single academic theme. During bonuses, the team can talk to one another. If the player who interrupted the moderator gets the toss-up question wrong, they receive a "Neg" which is a deduction of 5 points, and your team cannot buzz until a new toss-up question is read. The game is decided by who gets the most points at the end of the 20 toss-up and bonus pairs [1].

[1] Code and data is available at https://github.com/KuleenS/BayesianBowl

1st Conference of Bayesian Statistics Final Projects (BSFP 2023), Baltimore, Maryland.

2.2 Quizbowl Tournaments

Quizbowl matches are often conducted through a tournament format. Tournaments are set up in a round robin with pools format. Each tournament is assigned what is called a "question set" or "set" for short. Multiple tournaments can be assigned the same question set as long as no one leaks the questions. A set is about 12-13 packets of 20 toss-up questions and bonuses. Each set has its own difficulty level. Within each difficulty level, the difficulty of the set can vary due to the questions contained in them. Teams are read most if not all packets in the set during the course of the tournament.

2.3 Statistics Explained

During match there are six important statistics. **Score** is the total amount of points a team received from tossups, powers, and bonuses and subtracting negs. **Powers** are the number of tossup questions the team got early and received extra points for. **Tossups** are the number of tossup questions the team got regularly. **Interrupts/Negs** are number of tossup questions the team got incorrect. **Bonus Points** are the total number of points the team received from bonuses. **Points per Bonus** is the number of bonus points the team received divided by the number of bonus rounds heard by that team [2].

3 Literature Review

3.1 Paired Comparison

We specifically focus on a paradigm of ranking called paired comparison. Paired comparison is a type of data where one has a list of preferences between sets of two items [8]. One can have only two items like Soda A or Soda B or multiple items that are partitioned into sets of two like all the matches between American chess players. These preferences can be subjective like an opinion in the soda example or objective like a win or a loss in the chess example.

This paradigm can be applied to almost any field with this type of data. For example, it can be used by marketers to find what beverage people prefer over others based on their choices at a restaurant, or it can be used by chess federations to determine rankings of players based on their wins and losses.

3.2 Statistical Models for Paired Comparison

The statistical treatment of paired comparisons was introduced by Thurstone through the Thurstone-Mosteller model. It represented the probability of item i beating item j as $\Phi(\lambda_i - \lambda_j)$ where Φ is the CDF of a normal and λ_i and λ_j are the the preference scores (estimated from data) of those items [19].

A couple years later, Zermelo introduced a different paired comparisons model [21]. This model was later discovered by Bradley and Terry to develop the Bradley-Terry model. This model used a logistic function instead of the CDF of a normal to normalize to a probability [6]. It was found later that these two models yield the same solutions, but Bradley-Terry is easier to compute and therefore became more popular.

Bayesian versions of Bradley-Terry models started to develop right after. Davidson and Solomon introduced the first Bayesian extension of Bradley-Terry [10] with a fairly complicated system of conjugate priors. Chen and Smith [9] quickly followed up by changing the prior to a Dirichlet distribution. Finally, Leonard [13] found success using non-conjugate priors like the Normal Distribution.

Bayesian Bradley Terry has been applied many times to ranking many professional teams and players. [20] used a modified version to rank Hockey Teams. [18] used a hierarchical version to rank MLB teams. [7] used their model to rank both NASCAR racers and chess players.

3.3 Ranking of Quizbowl Teams

There are only partially statistical methods that have been used to rank quizbowl teams: HSQBRank [15] and GrogerRanks [11]. Both of these methods look at the statistics recorded during the game rather than the win-loss records as they believe that the single win or loss loses a lot of the nuance in close matches.

3.3.1 HSQBRank

HSQBRank is a defunct ranking service made by a very passionate community member. It focused on ranking teams using a special derived statistic called Adjusted Points Per Bonus (aPPB) and then breaking ties by hand using a combination of other statistics.

To calculate aPPB, they choose a question set during the season to serve as a baseline. Then, after all tournaments who were assigned a different question set finish playing that question set, they collect all the teams who played both the baseline set and the newly finished set. They then compute the difference in all the teams' PPB between those two sets and then average all those differences to get a difficulty adjustment.

They then add this adjustment to all the teams original PPB for that newly finished set to get the aPPB for all teams for that set. To rank the teams, they take a team's highest aPPB and compare to other team's highest aPPB [15].

3.3.2 GrogerRanks

GrogerRanks is the successor to HSQBRank which incorporates four new changes. First, they calculate the adjustment for aPPB differently as they use a least squares regression to find the optimal adjustments. Second, they introduce Adjusted Powers per Game (aP/G) and Negs per Game (N/G). N/G is a fairly easy statistic to calculate as it is often recorded on the scoresheet. aP/G is calculated like how HSQBRank calculates aPPB but using a team's number of Powers. Thirdly, they factor all of these statistics into this hand crafted equation below to create a score for a team on a question set.

$$Score = \frac{100}{32}\left(\left(\sqrt[3]{\text{P/G}} + \text{P/Gadj}\right)^3 + (\text{PPB} + \text{PPBadj})\right) - \text{N/G} \tag{1}$$

Fourthly, they take a weighted average of a team's top three scores to get an overall score [14].

4 Method

4.1 Questions

We will be answering one fundamental question: Who is the current best high school/middle school quizbowl team?

4.2 Cleaning and Collecting Data

National Academic Quiz Tournaments (NAQT) is one of the premier and longest running tournament organization companies. They host all their stats on their websites spanning back 20 years [3]. Using a webscraper, we collected 10,121 tournaments with around 398,827 matches with 16,912 unique teams. Each datapoint in our dataset was a match with a winning team, a losing team, and possibly some recorded statistics about the match.

We quickly cleaned the data. First, we removed any teams that were not associated with a middle/high school or were not in the United States. Second, we removed all the stats as we just wanted the pair of winning and losing teams for each match.

4.3 Bradley-Terry Model

We can model this match up data using the Bradley Terry model [6].
Our likelihood looks like:

$$\mathbb{P}[i \text{ beats } j] = \frac{\exp \lambda_i}{\exp \lambda_i + \exp \lambda_j}$$
$$y_{i,j} \sim \text{Bernoulli}(\mathbb{P}[i \text{ beats } j])$$

Our prior will look like:

$$\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2)$$
$$\sigma_\lambda^2 = 1$$

The objective of the Bradley-Terry model is to rank n teams using pairwise matches between those teams. It estimates the probability that a team i beats team j as a comparison between their latent strengths λ_i and λ_j and treats that match as Bernoulli random variable. These latent strengths are inferred through either maximizing log-likelihood or MCMC.

Similarly to [13], we choose to sample the strengths from a normal distribution as strengths can span the entire real line. We choose a mean of 0 as any other number would just shift the latent strengths by that number.

4.4 Implementation

4.4.1 Bayesian Model

The Bayesian model will be implemented using PyMC [16], a probabilistic programming framework for building Bayesian models in Python.

4.4.2 Sampling

We use the No-U-Turn Sampler (NUTS) [12] that is implemented in the PyMC package to sample our Bayesian model 1000 times. NUTS is part of the Hamiltonian Monte Carlo family which is slightly different than Metropolis-Hastings. We will provide a brief explanation below of both.

4.4.3 Hamiltonian Monte Carlo (HMC)

Hamiltonian Monte Carlo (HMC) improves the convergence of sampling in higher dimensions and acceptance rates significantly by using the gradient of the target distribution.

We turn to physics for motivation. Imagine we are a marble on an infinite surface. This surface is our posterior distribution. It has many different hills and valleys that we need to explore. At different points on our surface, like in the real world, we have some implicit energy due to gravity from our height.

To explore and traverse this surface, we are given a certain amount of energy that we can expend in a direction. Given that we have our energy from our position and some given energy in a direction, we can simulate travel on the surface using physics specifically, Hamiltonian mechanics.

Hamiltonian mechanics is a formulation of physics based on an object called the Hamiltonian of a system. The Hamiltonian is defined as the sum of the potential energy of a system (our position in the landscape) and the kinetic energy of our system (given energy). We use this object in conjunction with Hamilton's equations (3) to simulate our travel on the surface.

Let us now formally define our Hamiltonian of our system. Let $f(\boldsymbol{\theta})$ be the posterior distribution that we are trying to sample from. We are given energy through a momentum vector $\boldsymbol{p}$ which chooses our direction. We also have a matrix M that acts as our mass and adjusts the speed that we sample our posterior. It is often chosen to be a live estimate of the covariance of the posterior.

Let $H(\boldsymbol{\theta}, \boldsymbol{p})$ be the Hamiltonian of our system at sample point $\boldsymbol{\theta}$ with momentum vector $\boldsymbol{p}$. Therefore, our Hamiltonian will be defined as:

$$H(\boldsymbol{\theta}, \boldsymbol{p}) = -ln(f(\boldsymbol{\theta})) + \frac{1}{2}\boldsymbol{p}M^{-1}\boldsymbol{p} \tag{2}$$

We can break the left hand side down into two parts. $-ln(f(\boldsymbol{\theta}))$ will be our potential energy. As we can see, our potential energy will be our posterior distribution. However, we take the negative natural log of it before plugging it in so that we can interpret the Hamiltonian in terms of probabilities. And the right hand term ($\frac{1}{2}\boldsymbol{p}M^{-1}\boldsymbol{p}$) will be our given energy, which the multi-dimensional version of kinetic energy. To simulate our travelling, we use Hamilton's equations:

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial \theta_i} \tag{3}$$

These equations allow us the ability to predict changes in momentum and position over time which is perfect for simulating travel across our surface. We cannot solve these equations exactly for all distributions so we take a discrete approximation using leapfrog integration. Leapfrog integration numerically solves these equations for L discrete steps of size Δt to get our next position [5].

Leveraging our newfound understanding of Hamiltonian dynamics, we will now introduce the Hamiltonian Monte Carlo Algorithm. Suppose, we are currently at sample point $\boldsymbol{\theta_n}$, then

Step 1: We sample a random momentum vector which gives us a random direction and energy, $\boldsymbol{p_n} \sim \mathcal{N}(0, M)$

Step 2: Two, we run our dynamics using the leapfrog integration for L steps of size Δt

Step 3: We obtain our new position $\boldsymbol{p^*}$, $\boldsymbol{\theta^*}$

Step 4: We now run an acceptance check like Metropolis-Hastings where $\boldsymbol{\theta_{n+1}} = \boldsymbol{\theta^*}$ if $u \sim Uniform(0,1) < \alpha(\boldsymbol{\theta_n}, \boldsymbol{\theta^*})$ else $\boldsymbol{\theta_n}$

Our threshold parameter is very similar to Metropolis-Hastings but uses the energy of our system.

$$\alpha(\boldsymbol{\theta_n}, \boldsymbol{\theta^*}) = min\left(1, \frac{exp(-H(\boldsymbol{\theta^*}, \boldsymbol{p^*}))}{exp(-H(\boldsymbol{\theta_n}, \boldsymbol{p_n}))}\right) \tag{4}$$

4.4.4 No-U-Turn Sampler (NUTS)

NUTS improves on HMC by choosing L dynamically as too many or too few steps lead to dependent samples. First, it runs Hamiltonian dynamics both forwards and backwards until the "No-U-Turn" condition is satisfied. The "No-U-Turn" condition is met when two paths seem to be going in the opposite direction of the initial momentum. At this point, we then choose our sample point uniformly from the combined forward backwards path.

Second, there is a change in the acceptance check. Instead of doing a ratio test like HMC, at the beginning of our leapfrog integration, we sample a threshold $u \sim Uniform(0, exp(-H(\boldsymbol{\theta_n}, \boldsymbol{p_n}))$. After sampling our proposal point, we then check if $u < exp(-H(\boldsymbol{\theta^*}, \boldsymbol{p^*}))$ [5].

4.4.5 Frequentist Models

The frequentist (MLE) versions will be implemented using PyTorch [17] and will be fit using gradient descent on the log-likelihood for 4300 steps. To calculate the standard errors for the MLE, we will be using the Fisher Information Matrix. Due to memory and time constraints, we will only be calculating the frequentist confidence intervals for the top 50 teams.

5 Data Analysis

To give a little picture of what the data looks like, we have provided a couple visualizations and summary statistics. As said before, we had 398,827 matches between 16,912 teams from 10,121 tournaments from about 20 years of games. In Table 1, we provide a list of summary statistics about the number of wins, losses, games played, and the win loss ratio. It looks like most teams play around 50 matches total and win about half of them. As we can see in Figure 1, the distributions of wins, losses, and games played are extremely peaked around their means. The summary statistics and distributions show we have a good amount of data to rank most teams as most teams will have around 50 points of data associated with them. We provide a second type of visualization in Figure 2. The

Statistic	Mean	Min	Max	Median	SD	Lower 95% Quantile	Upper 95% Quantile
Win Loss Ratio	0.45	0.017	1.0	0.43	0.22	0.1	1.0
# of Wins	26.02	1	975	8.0	50.08	1.0	167.0
# of Losses	24.18	1	411	11	35.09	1.0	125.0
# of Games	51.40	1	1085	21	80.28	4.0	278.0

Table 1: Summary Statistics for Dataset

second type of visualization is a directed graph of all matches in the dataset built in Cosmograph [4]. The nodes represent the teams and the edges go from the winning team to the losing team. We generated two graphs: one of all the matches and one of just a single team's set of matches (DCC A). As we can see in Figure 2a, there is a large cluster of teams in the middle that play often together

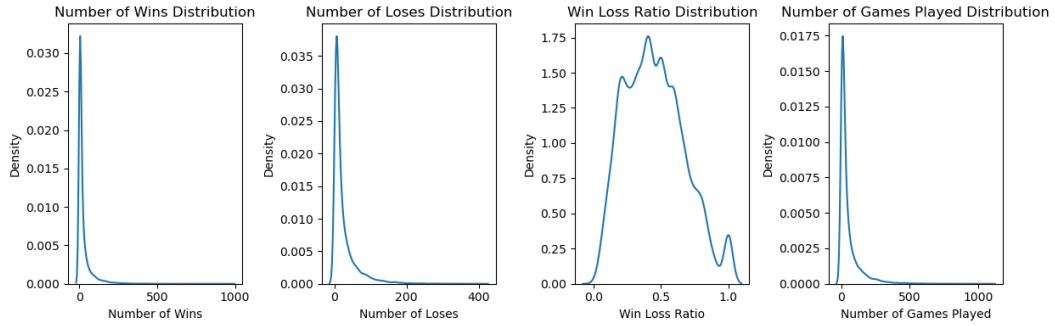
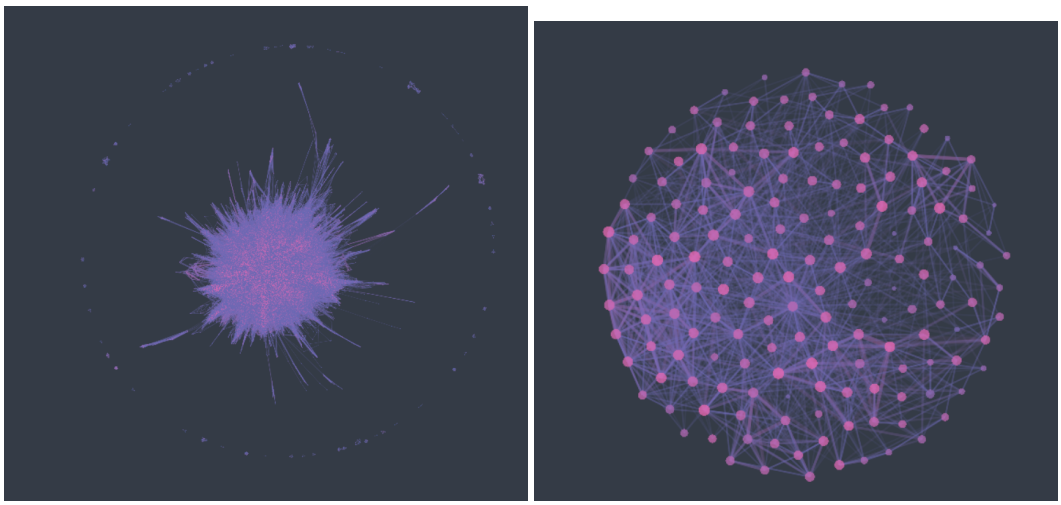


Figure 1: Distributions of Wins, Losses, Win Loss Ratio, and Games Played

which means we have enough interconnected and related matches data to be able to infer strengths. Zooming into an individual school, as we can see in Figure 2b, teams often have very dense graphs which leads to a better ability to rank these teams.



(a) All Matches

(b) Matches for Single School

Figure 2: Visualization of Quizbowl Matches as Graphs (nodes are teams, directed edges are winner to loser)

6 Sampling and Training Results

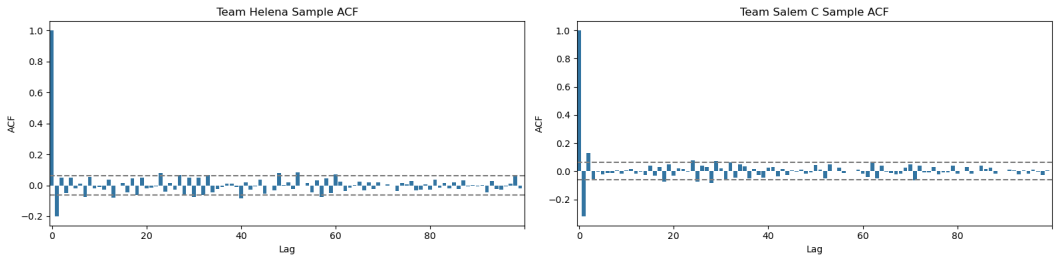


Figure 3: ACF Plots for McDonald B and Helena Middle

We ran our Bayesian model for 1000 samples. As we see in Figure 3, we have almost independent samples as we have a steep drop off in ACF past the first and second lag. Our effective sample size

(ESS) backs this up. Our mean effective sample size over all parameters was 1608. Our median ESS was 1600 and the 95% quantiles were 914 to 2315. These numbers are way higher than the number of real samples we had because we have that extremely negative autocorrelation between our samples. Having negative autocorrelation indicates extremely fast convergence of our chain. As for

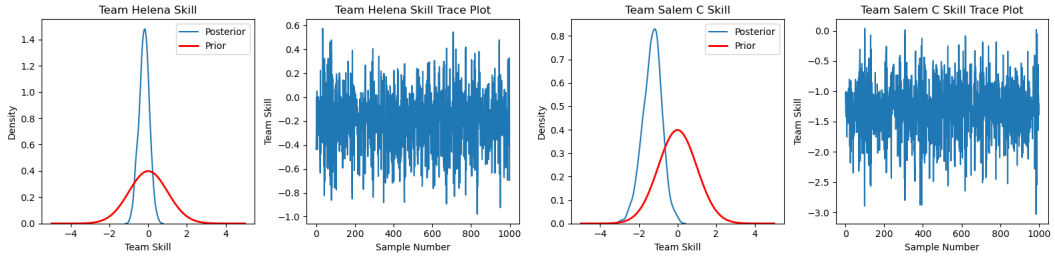


Figure 4: Traceplots and Prior and Posteriors for Helena and Salem C

the mixing of our chain, we can look towards Figure 4. In Figure 4, our chain looks well mixed as it hops between many parts of the distribution. In addition, we do see to be learning the strength of the teams as the variances of our distribution have shrunk and or our means have changed. As for our

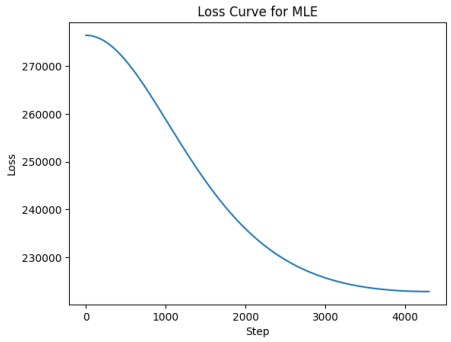


Figure 5: Loss Curve for MLE Model

frequentist model, we can see in Figure 5 our MLE model converges very nicely to a local minimum within a reasonable amount of steps.

7 Ranking Results

We show our predictions for the Top 10 Teams in Table 2, their predicted strengths, and 95% CIs. We rank the teams by their predicted strength by each model. For the Bayesian model, we take the mean of the samples as the predicted rating.

The Bayesian model out performs the MLE in many different factors. As we can see, the Bayesian model has more useful information in its CIs as the MLE CIs are extremely high and cover too large an area.

In addition, talking with invested community members, the predictions of the Bayesian model line up more with what the community thinks rather than the MLE model. They believed that teams like Eden Prairie, Dorman, and Solon do not seem to be teams deserving to be in the Top 10. They also believed that DCC A, Hunter A, and Beavercreek A make sense as a top 3 predictions.

Looking at the total predictions, it seems that the Bayesian model has a better fit in its rankings. In the left hand side of Figures 6 and 7, we can see that the MLE has a flatter middle section of the total rankings. This flatter section means that the associated strengths for middle of the pack teams are not as stratified as the Bayesian model, leading to rankings being decided by small margins. In addition, the MLE only seems to have bigger differences its rankings near the tails which means the Bayesian model is better at ranking teams are in the middle.

Rank	Bayes Team	Mean Bayes Str.	Bayes 95% CI	MLE Team	MLE Str.	MLE 95% CI
10	IMSA A	3.61	3.24 , 3.99	TJHSST A	1.96	-142.51 , 146.43
9	Middlesex A	3.66	3.26 , 4.10	Rockford Auburn	2.01	-142.46 , 146.48
8	Ladue	3.68	3.20 , 4.19	St. Marks A	2.13	-142.34 , 146.60
7	Richard Mont. A	3.70	3.43 , 4.01	Hunter A	2.29	-142.18 , 146.76
6	Montgomery Bl. A	3.75	3.30 , 4.20	Dorman A	2.36	-142.11 , 146.83
5	Ladue A	3.75	3.18 , 4.41	Solon A	2.42	-142.05 , 146.89
4	Challenger-Alm.	3.78	3.01 , 4.54	DCD A	2.45	-142.02 , 146.92
3	Beavercreek A	3.84	3.23 , 4.54	Eden Prairie A	3.28	-141.19 , 147.75
2	Hunter A	4.08	3.74 , 4.41	DCC A	3.58	-140.89 , 148.05
1	DCC A	4.09	3.83 , 4.37	Wayzata A	3.94	-140.53 , 148.41

Table 2: Top 10 Teams and Their Predicted Strengths

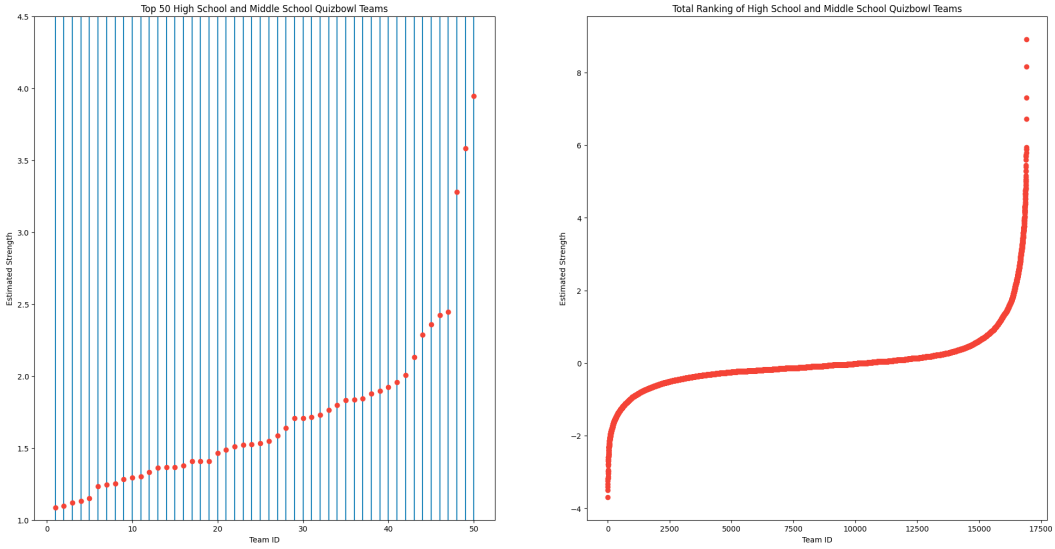


Figure 6: MLE Rankings: Predicted strength (red), 95% CI (blue); Top 50 Teams (left) Total Rankings (right)

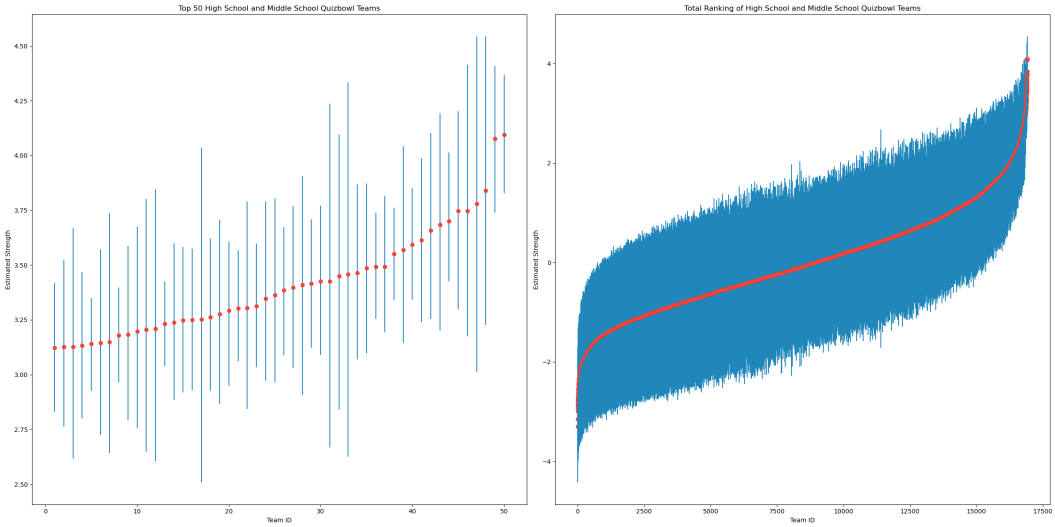


Figure 7: Bayesian Rankings: Mean predicted strength (red), 95% CI (blue); Top 50 Teams (left) Total Rankings (right)

References

[1] URL: https://www.naqt.com/nationals/press-guide.jsp.

[2] URL: https://www.naqt.com/stats/explanation.jsp.

[3] URL: https://www.naqt.com/.

[4] URL: https://cosmograph.app/.

[5] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2018. arXiv: 1701.02434 [stat.ME].

[6] Ralph Allan Bradley and Milton E. Terry. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3/4 (1952). Publisher: [Oxford University Press, Biometrika Trust], pp. 324–345. ISSN: 00063444. DOI: 10.2307/2334029. URL: http://www.jstor.org/stable/2334029 (visited on 12/15/2023).

[7] Francois Caron and Arnaud Doucet. *Efficient Bayesian Inference for Generalized Bradley-Terry Models*. 2010. arXiv: 1011.1761 [stat.ME].

[8] Manuela Cattelan. "Models for Paired Comparison Data: A Review with Emphasis on Dependent Data". In: *Statistical Science* 27.3 (2012), pp. 412–433. DOI: 10.1214/12-STS396. URL: https://doi.org/10.1214/12-STS396.

[9] C. Chen and Theodore M. Smith. "A Bayes-type estimator for the Bradley-Terry model for paired comparison". In: *Journal of Statistical Planning and Inference* 10 (1984), pp. 9–14. URL: https://api.semanticscholar.org/CorpusID:122446976.

[10] Roger R. Davidson and Daniel L. Solomon. "A Bayesian Approach to Paired Comparison Experimentation". In: *Biometrika* 60.3 (1973), pp. 477–487. ISSN: 00063444. URL: http://www.jstor.org/stable/2334996 (visited on 12/15/2023).

[11] John John Groger et al. *Groger Ranks*. Nov. 2023. URL: https://grogerranks.com/.

[12] Matthew D. Hoffman and Andrew Gelman. "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *J. Mach. Learn. Res.* 15 (2011), pp. 1593–1623. URL: https://api.semanticscholar.org/CorpusID:12948548.

[13] Tom Leonard. "An Alternative Bayesian Approach to the Bradley-Terry Model for Paired Comparisons". In: *Biometrics* 33.1 (1977), pp. 121–132. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2529308 (visited on 12/16/2023).

[14] Steven Liu. *Groger Ranks 2019-20 Methodology Changes*. Accessed: 12-15-2023. 2019.

[15] Fred Morlan. *FAQ*. Mar. 2013. URL: https://hsqbrank.com/faq/.

[16] Abril-Pla Oriol et al. "PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python". In: *PeerJ Computer Science* 9 (2023), e1516. DOI: 10.7717/peerj-cs.1516.

[17] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: 1912.01703 [cs.LG].

[18] Gabriel C. Phelan and John T. Whelan. *Hierarchical Bayesian Bradley-Terry for Applications in Major League Baseball*. 2017. arXiv: 1712.05879 [stat.AP].

[19] L. L. Thurstone. "The method of paired comparisons for social values." In: *The Journal of Abnormal and Social Psychology* 21.4 (1927). Place: US Publisher: American Psychological Association, pp. 384–400. ISSN: 0096-851X(Print). DOI: 10.1037/h0065439.

[20] John T. Whelan and Jacob E. Klein. "Bradley–Terry modeling with multiple game outcomes with applications to College Hockey". In: *Mathematics for Application* 10.2 (Feb. 2022), pp. 157–177. ISSN: 1805-3629. DOI: 10.13164/ma.2021.13. URL: http://dx.doi.org/10.13164/ma.2021.13.

[21] E. Zermelo. "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung". In: *Mathematische Zeitschrift* 29.1 (Dec. 1929), pp. 436–460. ISSN: 1432-1823. DOI: 10.1007/BF01180541. URL: https://doi.org/10.1007/BF01180541.