The Swiss Tournament Model

Christopher Hua

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

WHARTON RESEARCH SCHOLARS

Faculty Advisor:

Linda Zhao
Professor of Statistics

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MAY 2017

Abstract

The Swiss tournament structure is well-known and commonly used, particularly in chess, because of its seemingly simple heuristic of preferring matchups with teams of similar win count. However, it has proven difficult to simulate. This paper presents a framework for simulation and investigation of Swiss-style tournaments and other similar tournaments. This paper considers the desirability of this tournament structure in different contexts, using simulation and empirical results, with particular emphasis on recovering partial and full rankings. The results show that while the Swiss structure performs well under certain criteria, under many circumstances it does not outperform a simple random pairing of teams.

Keywords: pairwise comparisons, tournaments, ranking, matching

Contents

Introduction

An important goal of tournaments is to find an overall winner, but it is often important to find the top-k contestants as well. This paper considers the effectiveness of Swiss-style tournaments in finding a partial ranking over the top-k contestants, and compares those results to fully random tournaments with the same number of rounds. We find that in the case of American high school policy debate, Swiss tournaments do not offer much advantage versus randomly paired tournaments. However, we identify several conditions under which Swiss tournaments outperform the random pairing method.

Swiss-style tournaments are typically credited to J. Muller of Switzerland, who first used the tournament structure to run a chess tournament at Zurich in 1895 (Federation 2003). There are many variations on the general structure of the Swiss tournament, but the key idea is that the first few rounds are randomly paired, and the remaining rounds are power-matched. Powermatching means teams are paired with teams that have similar records, i.e. same number of wins or at most 1 difference. These are subject to the constraints that teams cannot debate teams from the same school, and they cannot debate teams who they have been paired with in earlier rounds.

A simple example of where this is useful is a preliminary tournament used for seeding purposes, prior to an elimination tournament. In some cases, knowing an exact ranking within this top subgroup is important, such as when a tournament will pay out monetary rewards based on finishing place; in other cases, knowing the exact ranking is not as important. A further advantage is that the Swiss tournament can be run with an arbitrary number of rounds, making the structure flexible to tournament constraints on time and space.

Understanding the performance of the Swiss tournament, and other tournament structures, can lead to important decisions. Tournaments are often used in high-stakes settings, including Swiss tournaments, and understanding the theoretical performance of tournaments can improve the parity for all involved actors. Additionally, these results can be used to inform further research direction in pairwise comparisons. One major reason that Swiss tournaments have not been studied before is the complexity in determining pairings and performing the simulation in an efficient way. A major contribution of this thesis, in addition to its analysis, is the tournament simulation framework which we implement and make available in Python.

This paper presents the case of American high school policy debate, in which teams compete in "regular-season" tournaments throughout the year in order to win 'bids' to the Tournament of Championships, the de facto culminating championship. Each round has two teams of two debaters, one "affirmative" (aff) and one "negative" (neg), and a judge. The affirmative side argues a policy-based plan which affirms that year's debate resolution, and the negative argues against the affirmative. For example, the 2012-13 resolution was "The United States federal government should substantially increase its transportation infrastructure investment in the United States."

All tournaments are structured in two parts, with a preliminary Swiss-system tournament and then a knockout/single-elimination tournament. The ultimate goal of "regular season" tournaments is to earn a bid to the "championship" tournament, the Tournament of Champions. These are allocated to tournaments roughly on the basis of tournament size and strength; the effect is that tournaments with more bids attract stronger teams. The bids are set up so that teams who make it to a given round of the tournament get the bid, e.g. octafinals means 16 bids, semifinals is 4 bids, etc. A perverse result of this bid system is that rounds after the bid round, containing the best teams, are treated as unimportant - teams routinely run less serious arguments or simply forfeit rounds - but the bid round and earlier rounds have enormous strategic investment.

This setup leads us to consider the efficacy of the Swiss-tournament design, in finding this top-k ranking.

Literature Review

Previous research has been done into creating tournament structures, and considering the various desirable aspects to optimize for.

Part of the simulation process involves finding pairings of competitors for each round, which are directly analogous to a matching, a set of n disjoint pairs of participants. In particular, the Swiss tournament structure is considered in Ólafsson (1990). Ólafsson considers the Swiss tournament structure and various chess-specific considerations; however, he focuses on an algorithm to create pairings which fulfill chess's requirements. He presents a method using maximum weight perfect matching to perform the pairing matching. Under this method, the tournament staff employ a graph structure to represent the teams (nodes) and the possible pairings (edges). The graph is initialized as a complete graph with equal weights; that is, all possible pairings are equally desirable, which fits the structure of a random initial pairing. The graph is complete because any team *could* play each other, but we can represent desirability via edge weights. At the conclusion of each round, the edges are reweighted to fit the desirability of the pairing. These weights are functions of various competitive factors, including the difference in wins, how many rounds they have played on white/black, whether the pairing has already occurred, and others. The general idea is that a higher weight represents a higher preference. Then, the maximum weight perfect matching algorithm finds a matching among the possible pairs.

The weighted perfect matching algorithm is a well-studied problem in computer science and graph theory. The first polynomial time algorithm for the problem was found by Edmonds (1965), known as Edmond's blossom algorithm, and is still largely in use, though improved on by numerous others, including Cook and Rohe (1999) and most recently by Kolmogorov (2009). The implementation used in this paper follows most directly the process given by Galil (1986). This implementation runs with time complexity $O(nm \log n)$, where n is the number of nodes and m is the number of edges in the graph. Exact details on the method can be found in any of these papers, or from examining the source code of the open-source programs used for simulation.

In a similar vein, Kujansuu, Lindberg, and Erkki (1999) present a method for pairing players as an extension of the stable roommates problem. The canonical reference for the stable roommates problem is McVitie and Wilson (1971). In the stable roommates problem, each "roommate" creates a full preference ranking of the others. Then, the matching is stable if there are no potential roommates i and j who prefer each other to their matched roommate. In the tournament context, after each round, each team has a preference list constructed for them of the teams available to play. The weights can be assigned in a similar manner to Olafsson and have a relatively analogous meaning, representing how preferable a possible pairing between two teams is.

To my knowledge, very few other studies have considered the effectiveness of the Swiss tournament structure, though previous researchers have investigated the knockout and the round robin tournament structures, e.g. Isabelle et al. (2013) or Mcgarry, Schutz, and Schutz (1997).

In particular, Mcgarry, Schutz, and Schutz (1997) chooses not to investigate the Swiss tournament because of the computational complexities in the Swiss pairing procedure. With more modern technology, we choose to tackle this problem.

Research by Glickman and Jensen (2005) has considered alternative tournament formulations from a more theoretical basis. Specifically, Glickman and Jensen present a tournament structure where rounds are matched by maximizing expected Kullback-Leibler distance, a measure of difference between distributions. The pairings are picked such that they maximize the expected Kullback-Leibler distance between the prior and posterior distributions of θ, the distribution of player strengths. This model is heavily influenced by Bayesian optimal design. Notably, Swiss tournaments out-perform their model for small numbers of rounds.

Hanes (2015) researched the effect of power matching in policy debate tournaments, comparing the outcomes from the win rankings with the speaker points assigned to teams. He finds a disparity between the two rankings, and argues that we should prefer the results given by speaker point rankings instead, or at minimum a combination of wins and speaker points. It is worth noting that he considers the implications across a full season, while we focus on the effects on a tournament level.

Solution Approach

Tournaments are repeated sets of paired comparisons, and we employ what is known as the Bradley-Terry model to understand the comparisons (Bradley and Terry 1952). The Bradley-Terry model belongs to a family of models known as linear paired comparison models, where win probabilities are only affected by player strengths in terms of the delta between the pairs. For several reasons, it is one of the most, if not the most, popular models for analyzing pairwise comparisons. It is given by:

$$\Pr(Y_{i,j} = 1) = \frac{\theta_i}{\theta_i + \theta_j}$$

Here, $Y_{i,j}$ is an indicator for the outcome of the pairwise comparison between competitors i and j, and θ_i and θ_j represent the underlying strength of competitors i and j. These θ values are relatively unconstrained, though under the traditional B-T assumptions they are positive numbers.

We draw these strength parameters from empirically estimated Bradley-Terry strength parameters drawn from a dataset containing the results of the full 2009-2010 and 2010-2011 debate seasons. The net effect is that we can closely see the results of tournaments under real and simulated conditions. We also test the model performance under different probability distributions, because we do not assume that the debate parameters are representative of all results.

Tournament design and procedure

As described above, teams compete in six rounds of competition, with the first two rounds randomly paired and the following rounds power-matched. We implement this procedure using an adaption of the maximum weight perfect matching technique (Ólafsson 1990).

Our process is as follows:

1. Team strengths are generated according to the given distribution, parameters, and random seed. In the case of empirical data, we pick a number of teams n and then draw the n strengths without replacement. Teams are represented as nodes in a symmetric directed graph, and edges are possible pairings.
2. A first round is paired randomly.
3. Results for the round are simulated and recorded, following the Bradley-Terry model for pairwise comparisons.
4. All rounds after the second are paired using maximum weight perfect matching.
5. After the second round, we reweight the graph.

The maximum weight perfect matching procedure is an ingenious method to guarantee good pairings. We have several desirable characteristics in pairings: first, that teams which play each other should not meet in further rounds, and that teams should prefer teams which have the same win total, but if necessary, play teams with a difference of 1 win. We can represent these characteristics within our graph model of a tournament by assigning weights to edges which reflect the desirability of the pairing. Our exact formula for weighting a possible pairing between teams i and j is as follows:

$$W_{i,j} = \alpha - (\beta * |s_i - s_j|)^2$$

Here, α and β are constants which can be thought of as a location and scale parameter, respectively. We also present a delta value, $|s_i - s_j|$, which is the absolute value of the the difference between the two teams' wins. To make computation easier, we avoid negative weights by first checking the win delta and setting the pairing to a weight of 1 if the difference is greater than 1 win. When a particular pairing is done, we assign the pairing a weight of 0. This method lends itself to a maximum weight method because the larger a weight is on a particular pairing the more desirable it is in a pairing.

Weights are rebalanced at the end of each round, i.e. when all pairings are simulated. All edges that have not been picked are rebalanced, since even if a pairing is undesirable after k rounds, it could be desirable for the $k + 1$ round. Picked edges are assigned fixed weights of 0 so that they are not picked. We then develop a pairing for the next round, which is represented as a maximum weight perfect matching. We use Edmond's blossom algorithm, as implemented in **Python** by NetworkX (Hagberg, Schult, and Swart 2008).

Although the algorithm which we use runs in $O(nm \log n)$ time, since our graph is fully connected, we have $m = n(n-1)/2$, which means that the algorithm runs in $O(n^3)$ time, where n is the number of teams competing. This becomes computationally intensive for relatively large tournaments but relatively manageable.

Note that the algorithm is used to find pairings for round 2, since the round is intended to be randomly paired. At this point the graph is initialized with equal weights for every pairing except those which have occurred, which have a 0 weighting. Then, since we have no other constraints, the maximum weight perfect matching returns an acceptable pairing which conveniently guarantees no repeat matches.

If we wish to create random pairings, we only need to modify the weighting procedure. Here, we draw a random integer value for any pairing which hasn't yet occurred and assign that as the weight. Then, the maximum-weight perfect matching procedure will come up with a set of pairings which have not occurred, where any pairing is equally likely to be picked.

Variables tested

We consider the effect of several different variables on the results from the tournament. For each of these variables, we create a simulation setup with those varied variables, and then run 500 simulations of the modeled tournaments.

We consider several different tournament configurations, and run 500 simulated tournaments for each of them. We repeated these experiments using empirical and simulated data.

First, we consider the effect of the number of teams in the tournament. We use configurations with 32, 64, 128, and 256 teams, with 6 rounds each. Second, we consider the effects of the number of rounds. We use a tournament with 64 teams to find the top-16 teams, and use rounds between 3 and 9.

Metrics reported

This paper reports several metrics of success, which are described below. In general, these are measures of how well the given tournament determines the best team, how well the tournament captures the top-K partial ranking, and how well the tournament captures the full ranking.

A number of metrics come from the search ranking literature, which generally focuses on the idea of "relevant" results. Consider the case of a search engine, which wants to return useful results to a given user query. For each user, the "best" result might be subjective, but all the results on the first page should be at least relevant. In our context, we define a relevant team as a team within the top-k teams by underlying strength rank. The metrics we specify are well-defined and commonly accepted; see e.g. Agichtein, Brill, and Dumais (2006).

- **Undefeated champion**: First, the experiments indicate if the top-rated player went undefeated throughout the tournament.
- **Copeland champion**: Similarly, we also measure if the top-rated player is at least tied for first place. Note that this condition should be strictly more common than the undefeated champion condition, since an undefeated player must be at minimum tied for first. This is known as the Copeland winning condition, which is any player with a maximum score (Saari and Merlin 1996).

- **Top-K percent**: Each of these tournaments has a particular K associated with them. These hark back to the goal of finding the K teams who will earn a bid for the tournament; here, we show the percent of teams in the top-K by strength who also place that highly by win rank. This can also be thought of as a partial ranking measure, because we test for group membership of the top-k teams but not the actual placement among those teams.
- **Precision at K**: the percent of top-K teams by wins, which are truly top-K teams as measured by underlying strength metric.
- **(Normalized) discounted cumulative gain at K**: The NDCG is given by $N = M \sum_{i=1}^{K} (2^{r(j)} - 1)/\log(1+i)$. This metric is a function of the relevance of the team ranked at position i, with additional priority applied to the top teams by wins. The term M is a normalization constant, which adjusts the maximum gain to 1, which is obtained by a perfect ranking.
- **Kendall's τ**: Kendall tau-b is given by $\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$ (Kendall 1945). We use the tau-b implementation as it is robust to ties in rank, which happen quite often in this context, since there are a discrete number of rounds played.
- **Spearman's ρ**: Spearman's rho is given by $r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \dfrac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$ (Zwillinger and Kokoska 2001). This metric, along with Kendall's tau, is a measure of full ranking accuracy.

We do not report the p-values of the Kendall or Spearman coefficients because these values are all 0.001 or lower, and very highly significant.

Validation

Data

We scraped multiple sources of pairwise comparison data, including 2 year-long datasets, covering multiple policy debate tournaments in the 2009-2010 and 2010-2011 seasons. The 2009-2010 dataset consists of 13310 debated rounds by 1424 teams, in 67 tournaments. The 2010-2011 dataset consists of 12915 debated rounds by 1521 teams, in 71 tournaments. We choose to use these datasets because they represent a large sample of teams with repeated comparisons in the data. This should yield better results in terms of finding underlying strength parameters.

Using these datasets and their final results, we can estimate Bradley-Terry parameters for the teams participating in those tournaments. Our maximum-likelihood estimation (MLE) is done using the **R** language (R Core Team 2016). In particular, we use the **BradleyTerryScalable** package (Kaye and Firth 2017). This package follows the procedure laid out in Caron and Doucet (2010) for maximum likelihood estimation of Bradley-Terry parameters when Ford's assumption does not hold. Ford's assumption is: in every possible partition of players into two non-empty subsets, some individual in the second set beats some individual in the first set at least once (Ford 1957). Our datasets are very sparse and cover a wide range of teams, meaning that Ford's assumption does not hold; in particular, this means that the more traditional MLE estimation methods of minorization-maximization (Hunter 2004) and Iterative-Luce Spectral Ranking (Maystre and Grossglauser 2015) cannot be used.

In particular, for each year of data, we typically find 2 or 3 such disjoint groups. The parameters estimates should hold for in-group analysis, but cannot be used in comparing members of these disjoint groups (Huang, Lin, and Weng 2005).[1] However, we have for each year a single dominant group, which contains the vast majority of members. For example, the 2009-2010 dataset has a group with 1376 of 1424 teams, or 96.6% of the teams). It is thus a reasonable assumption that these groups are representative of the whole dataset, so we run the same experiments as above using these empirical results.

For these simulations, we run 500 simulations in each configuration, similar to our process with the simulated data, and examine the difference in each.

Below, we show a histogram of the empirically determined parameters, with facets for the 2009-2010 season, 2010-2011 season, and a lognormal distribution whose logarithm has mean equal to 0 and standard deviation equal to 1. Note that the x-scales are different, but the shapes are very similar - this is acceptable because the Bradley-Terry model is scale-invariant.[^invariant]

[^invariant] By this, we mean that we could multiply the parameter vector θ by some scale constant Λ and the likelihood estimates will not vary. See e.g. Caron and Doucet (2010).

This distribution of strength makes sense in general. The distribution implies a large number of bad teams, with most teams being somewhat mediocre, and a small number of extremely strong teams. Given certain aspects of the debate community structure, this is likely a reason-

---

[1] A simple illustration of this would be to take two groups of people and make pairwise comparisons within each group, but without involving the other group. Then, if we recorded these results in a dataset and attempted to estimate parameters, we would necessarily need two groups of results.

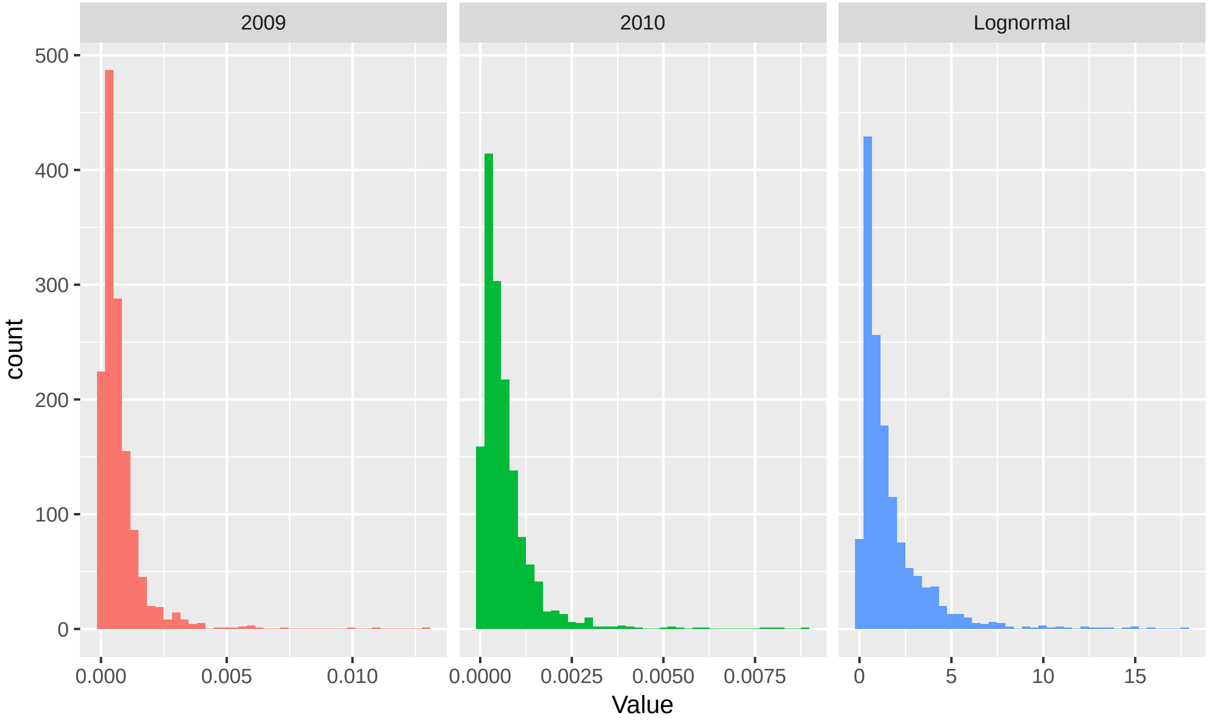


Figure 1: Empirically determined Bradley-Terry strengths from actual tournaments

able estimate of the true parameters. The debate community is relatively small but with a high level of inequality, especially at the high school level. Competition is traditionally dominated by a small number of elite schools in wealthy areas (e.g. North Shore of Chicago, suburbs of Dallas), who can afford to hire coaches and pay for expensive specialized summer camps. These inequalities are compounded by the strongly traditional culture and history of the activity, though this has begun to change in recent years (Reid-Brinkley 2008).

Results

As noted above, we ran simulations using the empirically determined strength parameters and tested the effects of changing certain variables.

Varying rounds

For this configuration, we used a tournament structure with 64 teams and 6 rounds, where we are interested in finding the top-16 teams. We show the effect of changing the number of rounds here, and illustrating the effects between 3 and 9 rounds. First, we consider the metrics which we would consider measures of full ranking performance.

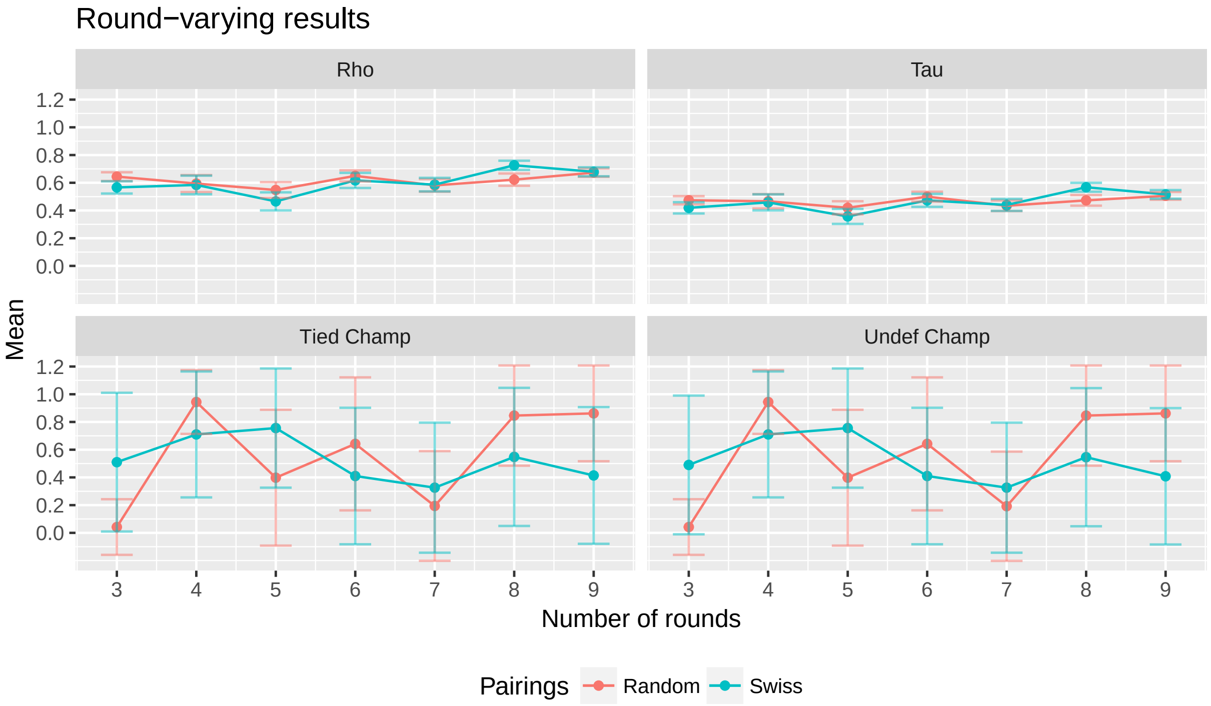


Figure 2: Full ranking metrics

Here, there is no clear overriding pattern for Swiss versus random pairings in their performance. Most metrics appear to move in step with each other under both frameworks. One noticeable trend is that the "undefeated champ" metric decreases in the Swiss tournament with increased numbers of rounds, while that metric stays higher under random pairings. This is because the Swiss tournament should be pairing the top team with stronger teams, while the random tournament should give easier matchups. The larger a tournament is, with random pairings, the likelier that the top team is undefeated. This is reasonable because the top team should get easier placements in the random framework than in the power-matched framework, which is the goal of the Swiss-style tournament. While this finding has little effect on the debate tournaments, which place relatively little emphasis on preliminary winners, it is important for other contexts in which Swiss tournaments are used to pick winners, such as chess.

Next, we examine metrics more closely tied to partial-rankings.

The results here look substantially similar, again. We see the same general trends in the metrics between both pairing methods. Interestingly, there appears to be the best performance at $m = 4$ rounds, in terms of the top-K metric. For both random and Swiss pairings, around 97% of the top-16 teams are picked correctly here. For small numbers of rounds (less than 7) the Swiss

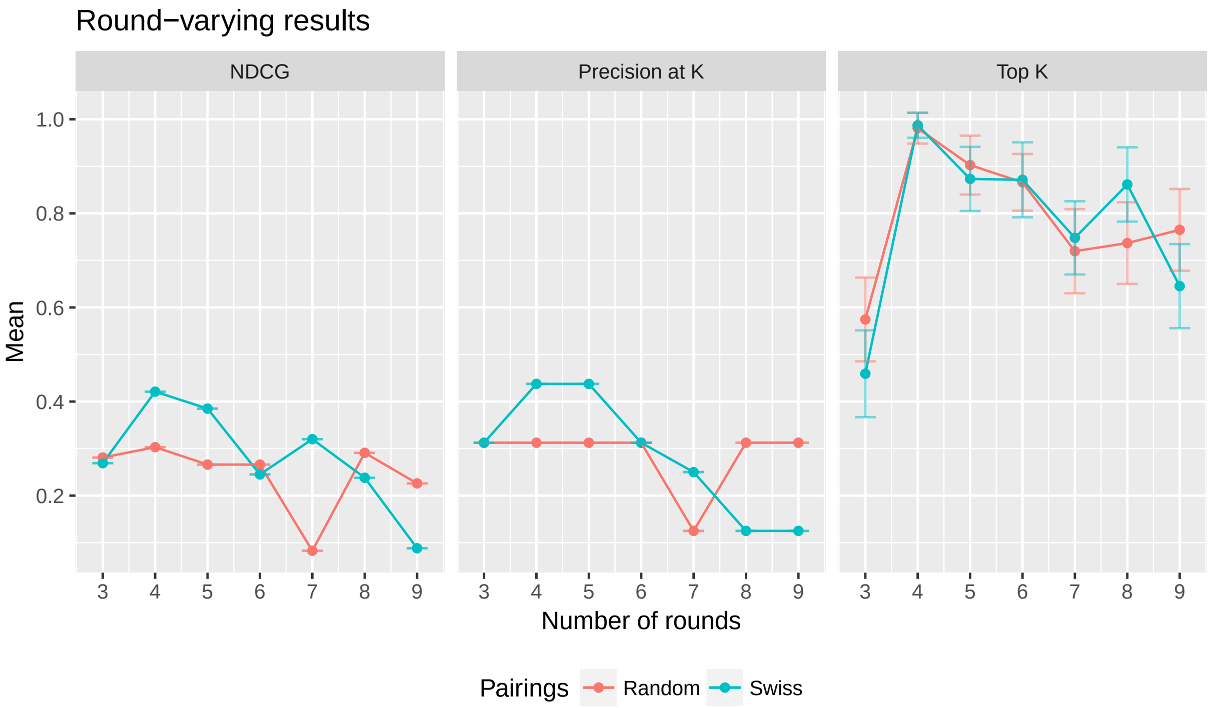


Figure 3: Partial ranking metrics

tournament performs better than the randomly paired tournament under the search ranking metrics.

This same characteristic was noted by Glickman and Jensen (2005), where their tournament model underperformed Swiss tournaments in sum of squared deviations in ranks (SSDR) for tournaments of 4 and 8 rounds, but was better in 16 round specifications. One possible explanation is that the random pairing rounds employed by Swiss tournaments in the first 2 rounds actually contribute the most to the full rankings, and the power-matched rounds reduce the accuracy. Further research could test the effect of varying the number of random rounds used in the Swiss tournament, to balance the goals of picking a top-k group of teams as well as yielding fair rankings to all participants.

Varying teams

We show the effect of varying the number of teams. Here, we varied the numbers of teams, with 32, 64, 128, and 256 teams. For our partial ranking measures, we fixed k at $n/4$, that is, where $k = 8, 16, 32, 64$. Essentially, we found the top quarter of teams.

First we graphically represent the full ranking measures.

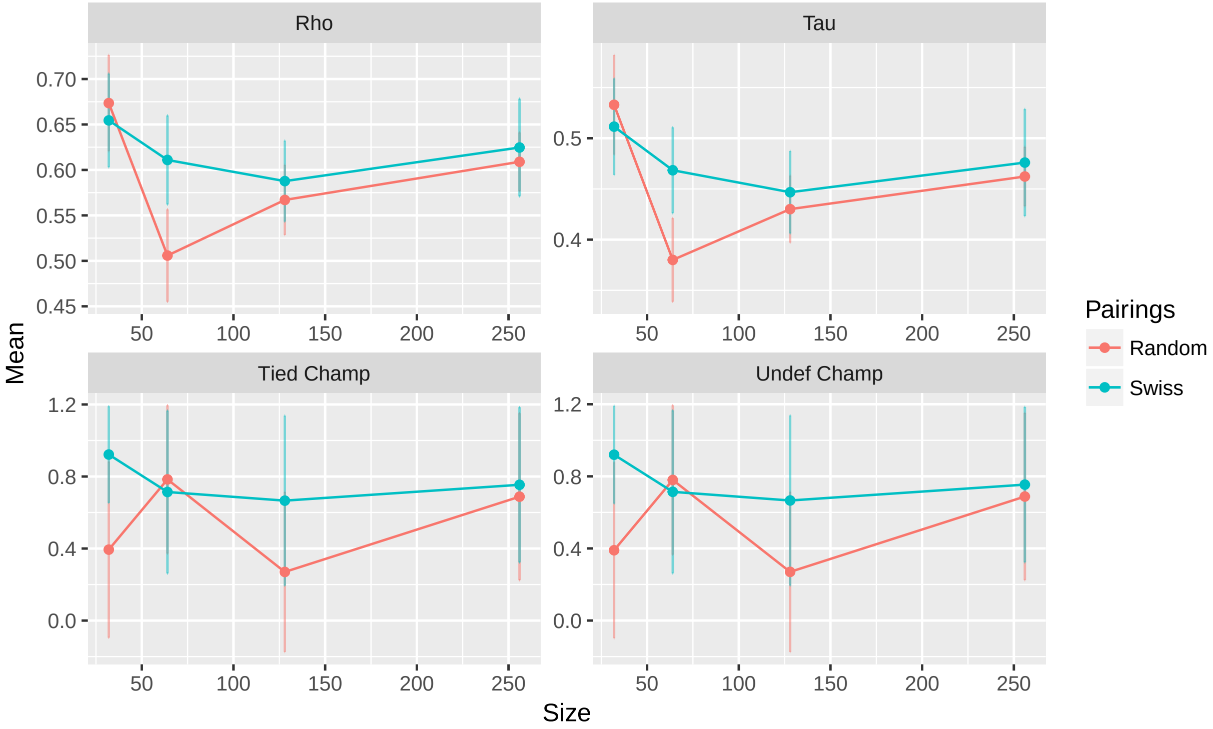


Figure 4: Full ranking metrics

It appears that Swiss pairings are somewhat better under these metrics than the random pairings. Interestingly, random pairings in the 64 team case perform poorly on the correlation measures but perform well in picking the top team by underlying strength.

We also show the results for the partial ranking metrics.

Under the search ranking metrics of net cumulative discounted gain and precision at K, the random pairing significantly outperforms the Swiss tournament. For both pairings, the best results are obtained with a 64 team tournament, which is interesting given the poorer performance on the full-ranking metrics seen above. It is possible that there is a tradeoff between performance in identifying a top-K set of teams and identifying the full ranking, and that's being seen in these results. The percent of top-K teams identified, however, seems to perform worst with 64 teams and using the random pairing.

With the random pairings framework, for the large and extra large tournaments, the ranking measures defined over all of the teams actually show better results than for the Swiss-style pairings. This is a pretty surprising result, especially because the Swiss-style tournament performs noticeably better than the random pairing tournament in picking the top-k teams. This can be thought of as a discrepancy between partial ranking and full-ranking measures. Under-

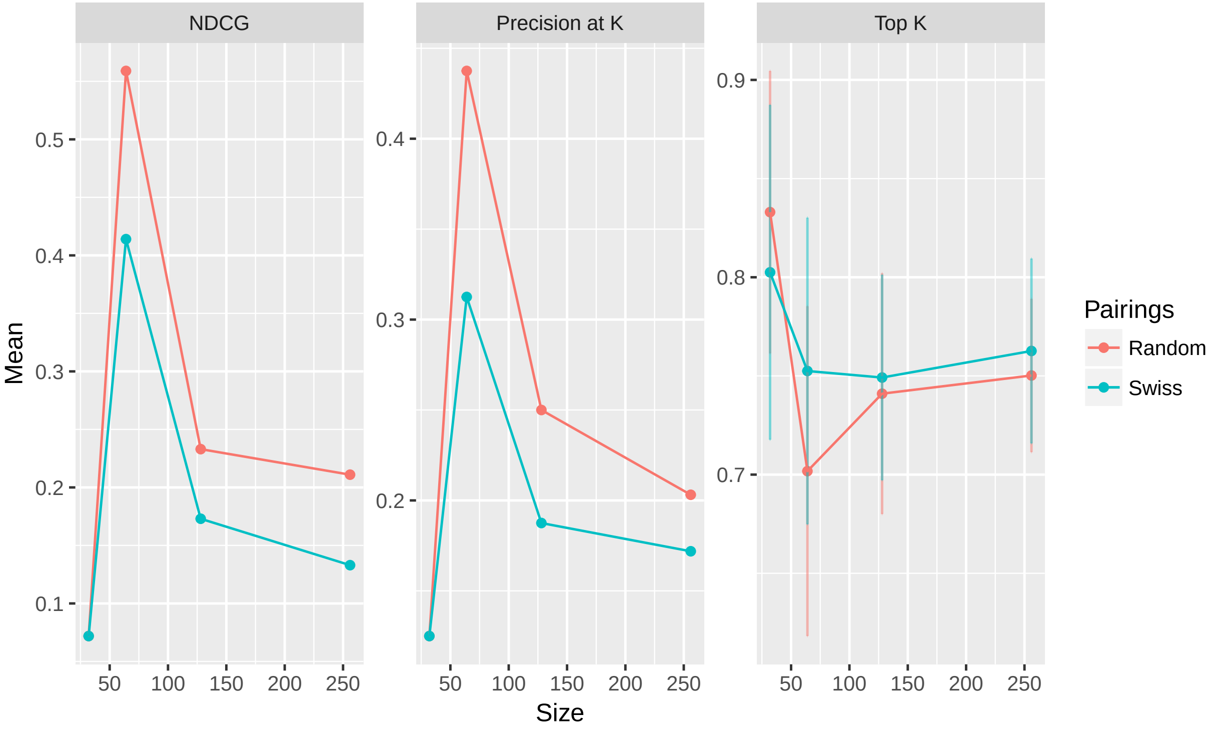


Figure 5: Partial ranking metrics

standing why this discrepancy exists would be a worthwhile research direction.

Overall, we find limited support for the hypothesis that the Swiss tournament leads to better empirical results than the random pairing model.

Results under exponential distribution

While in the above sections we used empirical data drawn from the distribution of debate data, we consider here an alternate setting, using Bradley-Terry strength parameters drawn from the exponential distribution.

The probability density function of the exponential distribution is given by:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

We also show the PDF of an exponential distribution with scale parameter λ of 3. Although the shape is superficially similar to that of the lognormal distribution, the exponential is less uniform, that is, with larger kurtosis. This means that teams are less similar in strength.

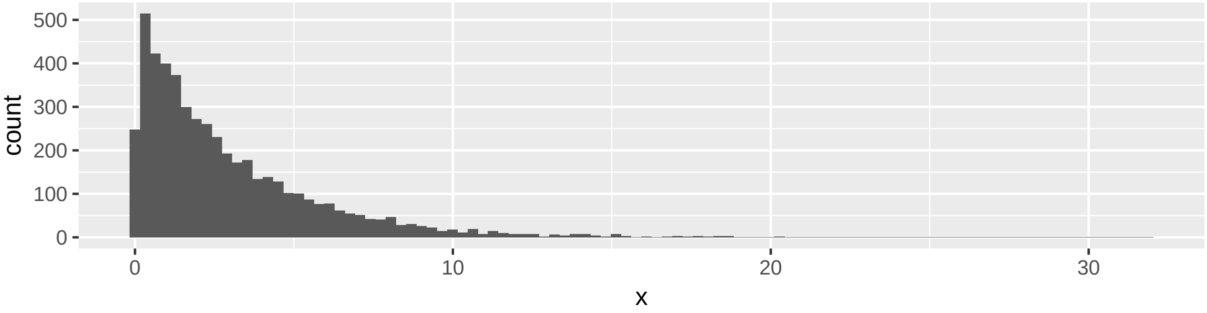


Figure 6: Exponential distribution

While this distribution of parameters doesn't necessarily fit the results of the group of debaters, it might fit the distribution of strength in other settings, such as in a chess tournament. In a championship chess tournament, certain chess players may be simply dominant versus others by skill, whereas in a debate, a well-prepared but weaker team may be able to come up with certain The implication is that the strengths are less widely dispersed.

Varying rounds

We show the results of varying the number of rounds for a 64 team tournament here, under the exponential distribution of strengths. First, we consider the effect on the full ranking measures.

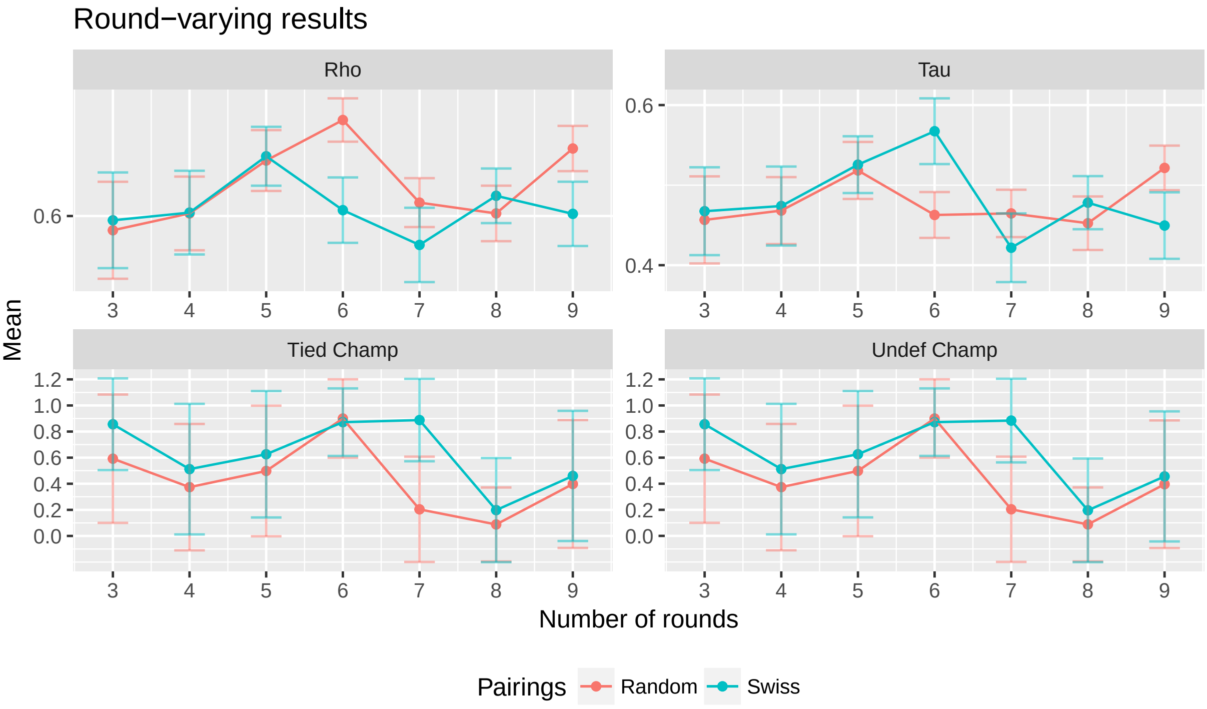


Figure 7: Full ranking results - varying rounds

Similarly to the empirical experiments, we do not see much difference in these full ranking metrics, between the Swiss and random pairing methods.

Next, we consider the partial ranking measures.

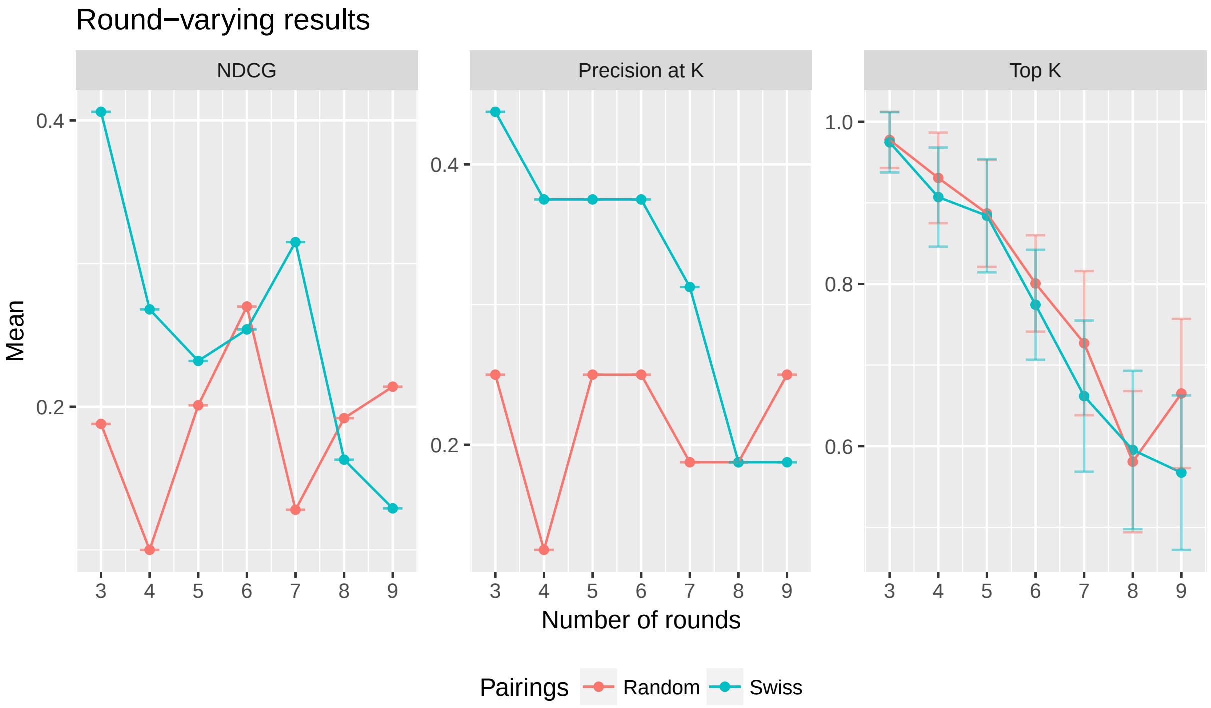
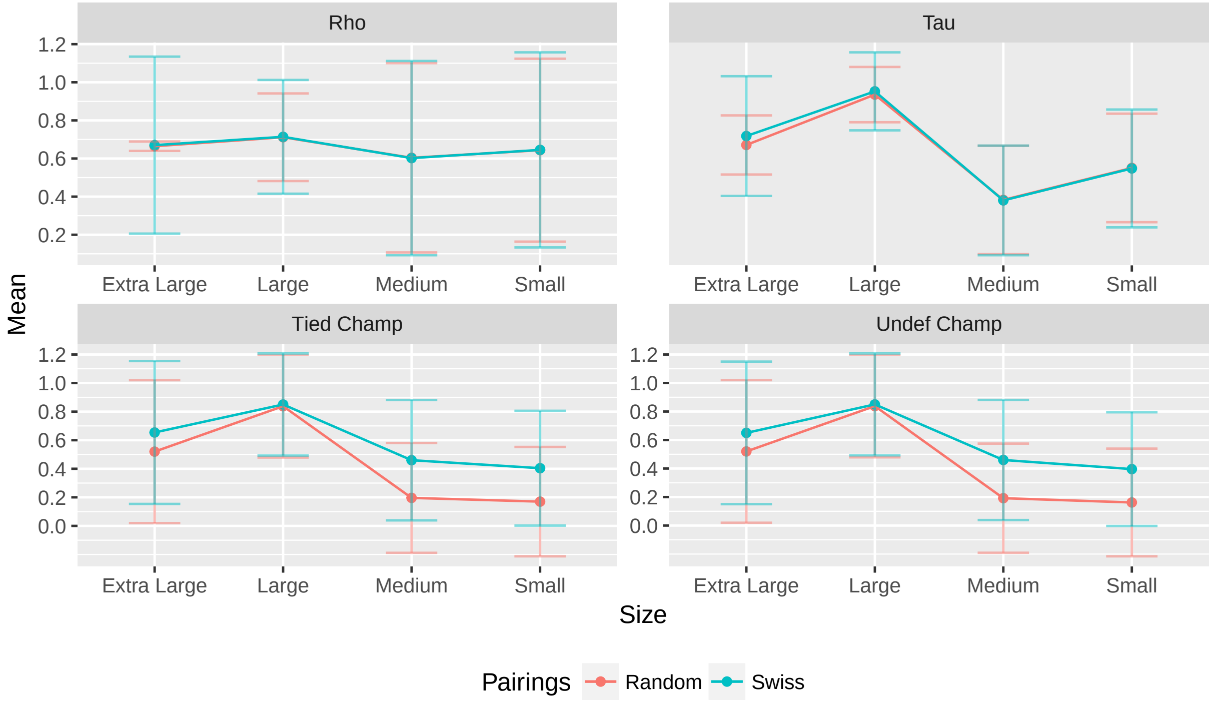


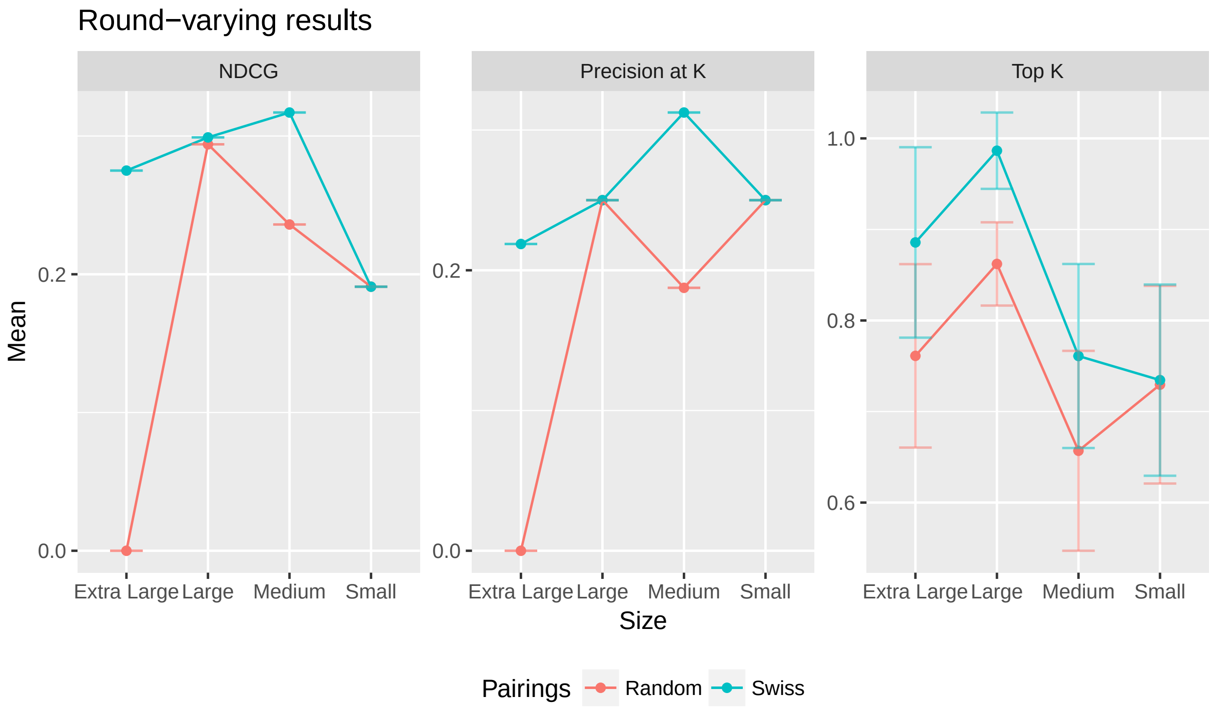
Figure 8: Partial ranking measures - varying rounds

Here, the Swiss tournament appears to clearly outperform the random pairings, for all but 8 or 9 round tournaments. Interestingly, the top-K results seem to monotically decrease for both methods; however, the search ranking results clearly favor the Swiss tournament across each round configuration.

Varying teams

We ran in an identical setup the





Discussion and Conclusion

In this paper, we have developed a flexible framework for working with Swiss tournaments and understanding the implications of their results. Previous researchers have not simulated and studied Swiss tournaments because of the computational complexities in performing the pairing procedure, and this paper represents a major step forward as well as a synthesis of the existing work.

A common argument against Swiss-style tournaments is that they provide poor results for top debate teams, because of possible disparities in schedule strength. We do not find proof for this hypothesis, when comparing the Swiss-style tournament to a randomly paired tournament with similar settings. Furthermore, we do not find evidence that, in general, Swiss tournaments yield better top-K partial or full rankings of competitors.

A key limitation of our analysis is that we focused on the case of American high school policy debate, which is a relatively small edge case in terms of tournaments. We examined the case of exponentially distributed strength parameters, under which the Swiss tournament does indeed perform better than randomly paired teams. Here, the results are similar among the full ranking measures but clearly show better performance among the partial ranking metrics. This suggests that Swiss tournaments can perform better than random pairings, under certain conditions, and one major condition is the overdispersion of strength parameters.

While our analysis takes into account empirical outcomes from the tournament, these are not the only considerations that tournament organizers must consider. There are numerous practical and operational considerations which might inform the pairing procedure.

An example of a structure made inpractical by these considerations would be a round-robin tournament. In a large tournament, we would not reasonably expect teams to winningly participate in hundreds of rounds, especially in debate, where each round is expected to take around 2 hours.

A particularly curious question is why the Swiss tournament performs well for small tournaments (both in terms of rounds and teams involved) but does not perform well in larger settings, when compared to a random pairing. Further work could investigate the differences in strengths that are created in the Swiss model vs a random model.

Additionally, work has been done to create alternative tournament pairing methods, such as in Glickman and Jensen (2005). Given our particular question of finding top-k teams, it would be useful to test these other models. One note is that while the Glickman model performs well in theory, the computational difficulty and explanation difficulty would likely make it difficult to implement in practice. Our study considered models which are feasible to implement in the real world.

In the end, these operational concerns may override the theoretical considerations for choosing any particular structure. Swiss tournaments are relatively easy to comprehend (even if they are difficult to implement), keep participants engaged by minimizing rounds with disparate ability, and are flexible enough to increase or decrease the number of rounds with minimal disruption.

Appendix

Code

Code and all other resources used in writing this paper can be found at the author's Github.

A major issue in the simulation of Swiss tournaments for analysis is the difficulty of creating pairings. A key contribution here is a comprehensive review of and the simplification of that problem.

The software which we used to implement the model and perform simulations is open-source and intended to be extensible. Within our paper, we take advantage of this, by utilizing a common framework for testing different numbers of teams and rounds, as well as creating summary statistics. Furthermore, we implement a random pairing model in the model for comparison testing.

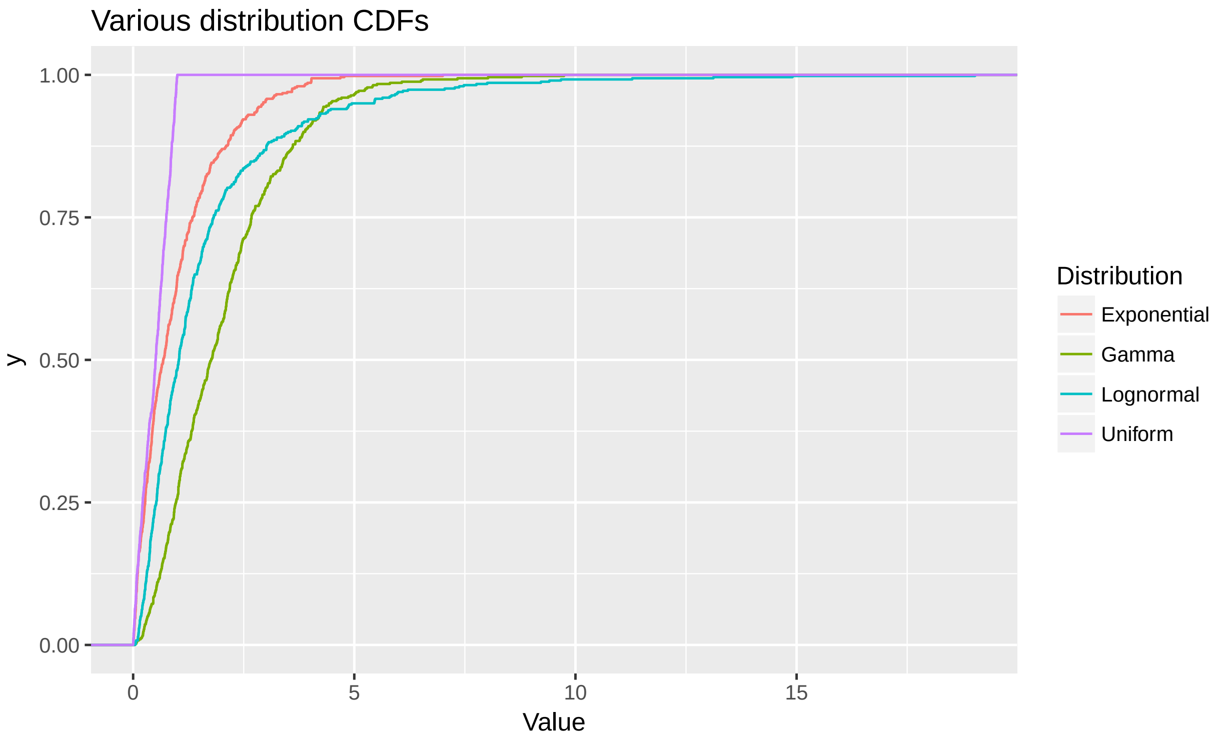
In the paper, we drew our theoretical strengths from a lognormal distribution with a mean $\mu = 0$, and a standard deviation $\sigma = 1$. Our framework includes support for the following distributions:

- Exponential distribution
- Uniform distribution
- Lognormal distribution
- Beta distribution
- Gamma distribution

Each of the above can be specified, along with optional shape parameters in the tournament and simulation keyword arguments. See the author's simulation code for examples on how to make these configurations.

Distributional effects

Here, we show four distributions which we used to model team strength, as well as the default parameters that we used.



Expected results

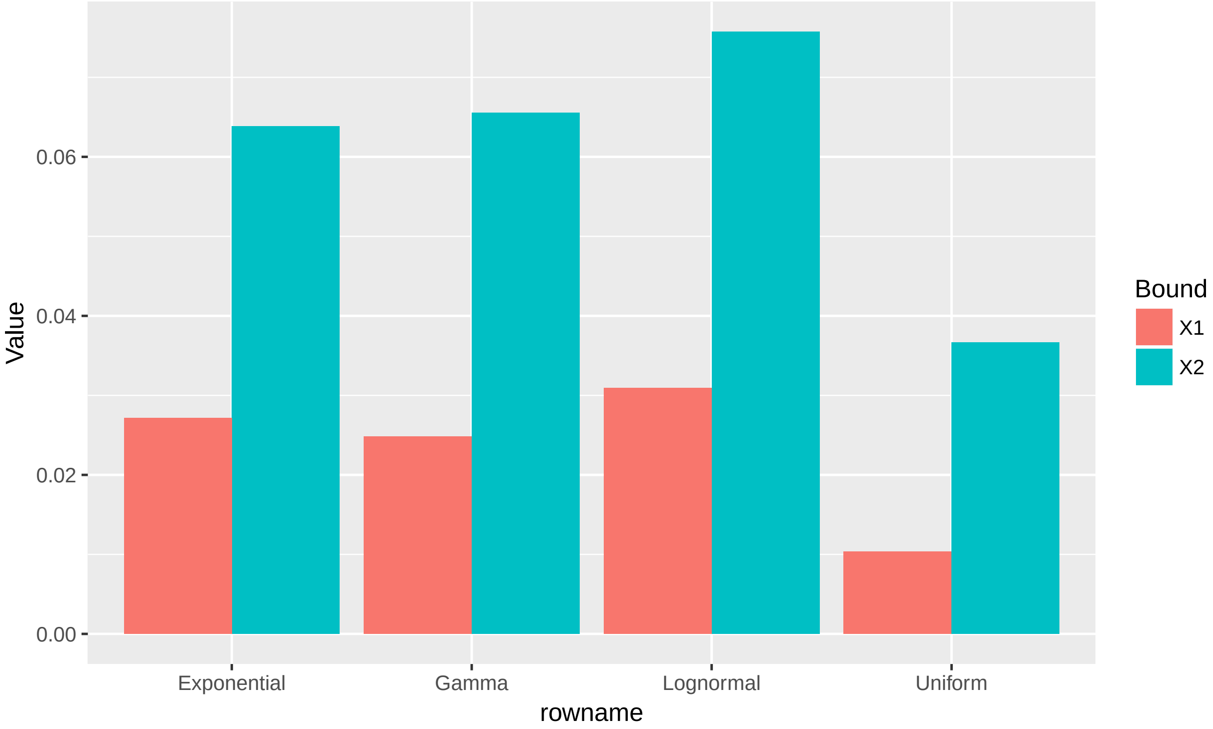
Some of the calculated statistics should be very simulation dependent, particularly the results which measure interaction with other teams. However, we can easily calculate a lower-bound on the top-team going undefeated, which is a particular win condition.

Under the Bradley-Terry model, we have fixed win probabilities for each possible pairings of teams. We can construct a win probability matrix M, that is, each cell $M_{i,j}$ represents the probability that team i beats team j.

We illustrate here a simple such matrix; given four teams with Bradley-Terry parameters of $1, 2, 3, 4$, we obtain the following probability matrix:

$$\begin{pmatrix} 0.50 & 0.33 & 0.25 & 0.20 \\ 0.67 & 0.50 & 0.40 & 0.33 \\ 0.75 & 0.60 & 0.50 & 0.43 \\ 0.80 & 0.67 & 0.57 & 0.50 \end{pmatrix}$$

Then, we can estimate a lower-bound on the chance of the top-team going undefeated, which happens if the team faces the strongest possible teams for the entire tournament. We calculate the lower bound for two scenarios - if the top team is paired versus the top-6 teams and versus the top-4 teams, essentially excluding the randomly paired rounds.



However, note that in tournaments with no more than $\log_2 n$ rounds, where n is the number of teams, there must be an undefeated team. Under that condition, we would still expect the top-rated team to have the highest chance of finishing undefeated and on top.

Tables for graphs

Rounds	Pairings	Undef Champ	Tied Champ	Top K	Precision at K	NDCG	Tau	Rho
3	Swiss	0.49	0.51	0.46	0.31	0.27	0.42	0.57
4	Swiss	0.71	0.71	0.99	0.44	0.42	0.46	0.58
5	Swiss	0.76	0.76	0.87	0.44	0.39	0.36	0.47
6	Swiss	0.41	0.41	0.87	0.31	0.24	0.47	0.62
7	Swiss	0.33	0.33	0.75	0.25	0.32	0.44	0.59
8	Swiss	0.55	0.55	0.86	0.12	0.24	0.57	0.73
9	Swiss	0.41	0.41	0.65	0.12	0.09	0.52	0.68
3	Random	0.04	0.04	0.57	0.31	0.28	0.47	0.64
4	Random	0.94	0.94	0.98	0.31	0.30	0.47	0.59
5	Random	0.40	0.40	0.90	0.31	0.27	0.42	0.55
6	Random	0.64	0.64	0.87	0.31	0.27	0.50	0.65
7	Random	0.19	0.19	0.72	0.12	0.08	0.43	0.58
8	Random	0.85	0.85	0.74	0.31	0.29	0.47	0.62
9	Random	0.86	0.86	0.77	0.31	0.23	0.51	0.67

Table 1: Empirical - varying rounds, means

Size	Pairings	Undef Champ	Tied Champ	Top K	Precision at K	NDCG	Tau	Rho
32	Random	0.39	0.39	0.83	0.12	0.07	0.53	0.67
64	Random	0.78	0.78	0.70	0.44	0.56	0.38	0.51
128	Random	0.27	0.27	0.74	0.25	0.23	0.43	0.57
256	Random	0.69	0.69	0.75	0.20	0.21	0.46	0.61
32	Swiss	0.92	0.92	0.80	0.12	0.07	0.51	0.65
64	Swiss	0.71	0.71	0.75	0.31	0.41	0.47	0.61
128	Swiss	0.67	0.67	0.75	0.19	0.17	0.45	0.59
256	Swiss	0.75	0.75	0.76	0.17	0.13	0.48	0.62

Table 2: Empirical - varying teams, means

Rounds	Pairings	Undef Champ	Tied Champ	Top K	Precision at K	NDCG	Tau	Rho
3	Swiss	0.86	0.86	0.97	0.44	0.41	0.47	0.59
4	Swiss	0.51	0.51	0.91	0.38	0.27	0.47	0.60
5	Swiss	0.63	0.63	0.88	0.38	0.23	0.53	0.68
6	Swiss	0.87	0.87	0.77	0.38	0.25	0.57	0.61
7	Swiss	0.88	0.89	0.66	0.31	0.32	0.42	0.56
8	Swiss	0.20	0.20	0.60	0.19	0.16	0.48	0.63
9	Swiss	0.46	0.46	0.57	0.19	0.13	0.45	0.60
3	Random	0.59	0.59	0.98	0.25	0.19	0.46	0.58
4	Random	0.37	0.37	0.93	0.12	0.10	0.47	0.60
5	Random	0.50	0.50	0.89	0.25	0.20	0.52	0.67
6	Random	0.90	0.90	0.80	0.25	0.27	0.46	0.73
7	Random	0.20	0.20	0.73	0.19	0.13	0.46	0.62
8	Random	0.09	0.09	0.58	0.19	0.19	0.45	0.60
9	Random	0.40	0.40	0.67	0.25	0.21	0.52	0.69

Table 3: Exponential - varying rounds, means

Acknowledgements

Special thanks go to Makena Finger and Kevin Huo for their contributions in reading drafts of this paper. Additional thanks are due to my professors, especially Linda Zhao and Shivani Agrawal, for their advice, encouragement, and support.

All remaining errors are mine.

Size	Pairings	Undef Champ	Tied Champ	Top K	Precision at K	NDCG	Tau	Rho
Small	Swiss	0.40	0.40	0.73	0.25	0.19	0.49	0.65
Medium	Swiss	0.46	0.46	0.76	0.31	0.32	0.47	0.60
Large	Swiss	0.85	0.85	0.99	0.25	0.30	0.55	0.71
Extra Large	Swiss	0.65	0.65	0.89	0.22	0.28	0.52	0.67
Small	Random	0.16	0.17	0.73	0.25	0.19	0.49	0.64
Medium	Random	0.19	0.20	0.66	0.19	0.24	0.47	0.60
Large	Random	0.84	0.84	0.86	0.25	0.29	0.55	0.71
Extra Large	Random	0.52	0.52	0.76	0.00	0.00	0.51	0.66

Table 4: Exponential - varying teams, means

Bibliography

Agichtein, Eugene, Eric Brill, and Susan Dumais. 2006. "Improving Web Search Ranking by Incorporating User Behavior Information." *SIGIR Forum* 39: 19–26. http://citeseerx.ist.psu.edu/viewdoc/dow http://search.ebscohost.com/login.aspx?direct=true{\&}db=lih{\&}AN=22879675{\&}lang=pt-br{\&}site=ehost-live{\&}scope=site.

Bradley, R, and M Terry. 1952. "Rank analysis of incomplete block designs. I. The method of paired comparisons." *Biometrika* 39: 324–45.

Caron, Francois, and Arnaud Doucet. 2010. "Efficient Bayesian Inference for Generalized Bradley-Terry Models." *Journal of Computational and …* 21 (2004). Taylor & Francis Group: 1–28. doi:10.1080/10618600.2012.638220.

Cook, Williams, and Andre Rohe. 1999. "Computing Minimum-Weight Perfect Matchings." *INFORMS Journal on Computing* 11 (2): 138–48. doi:10.1287/ijoc.11.2.138.

Edmonds, Jack. 1965. "Paths, trees, and flowers." doi:10.4153/CJM-1965-045-4.

Federation, U.S. Chess. 2003. "United States Chess Federation's Official Rules of Chess, Fifth Edition." Random House. https://books.google.com/books?id=syxPAgAAQBAJ{\&}pg=PT31{\&}lpg=PT3 YRr-{_}xLlaQsxBVys{\&}hl=en{\&}sa=X{\&}ved=0ahUKEwjqycaxrM{_}TAhUG4SYKHfzSCNQQ6AEISD muller zurich 1895{\&}f=false http://www.amazon.com/dp/0812935594.

Ford, L. R. 1957. "Solution of a Ranking Problem from Binary Comparisons." *The American Mathematical Monthly* 64 (8): 28. doi:10.2307/2308513.

Galil, Zvi. 1986. "Efficient algorithms for finding maximum matching in graphs." *ACM Com-*

puting Surveys 18 (1). ACM: 23–38. doi:10.1145/6462.6502.

Glickman, Mark E., and Shane T. Jensen. 2005. “Adaptive Paired Comparison Design.” *JOURNAL OF STATISTICAL PLANNING AND INFERENCE* 127. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.7306.

Hagberg, Aric A, Daniel A Schult, and Pieter J Swart. 2008. “Exploring network structure, dynamics, and function using {NetworkX}.” In *Proceedings of the 7th Python in Science Conference (Scipy2008)*, 11–15. Pasadena, CA USA.

Hanes, Russel. 2015. “Study of speaker points and power-matching for 2006-7.” http://art-of-logic.blogspot.com/2015/07/study-of-speaker-points-and-power.html.

Huang, Tzu-Kuo, Chih-Jen Lin, and Ruby C Weng. 2005. “A Generalized Bradley-Terry Model: From Group Competition to Individual Skill.” *Advances in Neural Information Processing Systems 17*, no. 3: 601–8. https://papers.nips.cc/paper/2705-a-generalized-bradley-terry-model-from-group-compet
pdf.

Hunter, David R. 2004. “MM algorithms for generalized Bradley-Terry models.” *Annals of Statistics* 32 (1). Institute of Mathematical Statistics: 384–406. doi:10.1214/aos/1079120141.

Isabelle, Stanton, Williams Virginia Vassilevska, Stanton Isabelle, and Williams Virginia Vassilevska. 2013. “The structure, efficacy, and manipulation of double-elimination tournaments.” *Journal of Quantitative Analysis in Sports* 9 (4). De Gruyter: 319–35. http://econpapers.repec.org/article/bpjjqsp
335{_}3an{_}3a3.htm.

Kaye, Ella, and David Firth. 2017. *BradleyTerryScalable: Fits the Bradley-Terry Model to Potentially Large and Sparse Networks of Comparison Data*. https://github.com/EllaKaye/BradleyTerryScalable.

Kendall, M. G. 1945. “The treatment of ties in ranking problems.” *Biometrika Trust* 33 (3): 239–51. doi:10.1093/biomet/33.3.239.

Kolmogorov, Vladimir. 2009. “Blossom V: A new implementation of a minimum cost perfect matching algorithm.” *Mathematical Programming Computation* 1 (1). Springer-Verlag: 43–67. doi:10.1007/s12532-009-0002-8.

Kujansuu, Eija, Tuukka Lindberg, and M Erkki. 1999. “The Stable Roommates Problem and Chess Tournament Pairings.” *Divulgaciones Mathematicas* 7 (1): 19–28. http://emis.ams.org/

journals/DM/v71/art3.pdf.

Maystre, Lucas, and Matthias Grossglauser. 2015. “Fast and Accurate Inference of Plackett – Luce Models.” *Advances in Neural Information Processing Systems* 28: 1–9. doi:no DOI. URL correct.

Mcgarry, T, R. W. Schutz, and Rw Schutz. 1997. “Efficacy of Traditional Sport Tournament Structures.” *Source: The Journal of the Operational Research Society Journal of the Operational Research Sodety* 48 (48): 65–74. doi:10.1057/palgrave.jors.2600330.

McVitie, D. G., and L. B. Wilson. 1971. “The stable marriage problem.” *Communications of the ACM* 14 (7). MIT Press: 486–90. doi:10.1145/362619.362632.

Ólafsson, S. 1990. “Weighted matching in chess tournaments.” *Journal of the Operational Research Society* 41 (1): 17–24. doi:10.1038/sj/jors/0410103.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Reid-Brinkley, Shanara Rose. 2008. “THE HARSH REALITIES OF ‘ACTING BLACK’: HOW AFRICAN-AMERICAN POLICY DEBATERS NEGOTIATE REPRESENTATION THROUGH RACIAL PERFORMANCE AND STYLE.” PhD thesis, University of Georgia. https://getd.libs.uga.edu/pdfs/ brinkley{_}shanara{_}r{_}200805{_}phd.pdf.

Saari, Donald G., and Vincent R. Merlin. 1996. “The Copeland method.” *Economic Theory* 8 (1). Springer-Verlag: 51–76. doi:10.1007/BF01212012.

Zwillinger, Daniel, and Stephen. Kokoska. 2001. *Standard Probability and Statistics Tables and Formulae*. Vol. 43. 2. Chapman & Hall/CRC. doi:10.1198/tech.2001.s620.