SMU Data Science Review

Volume 6
Number 2 *Summer 2022*

Article 2

2022

Application of Probabilistic Ranking Systems on Women's Junior Division Beach Volleyball

Cameron Stewart
Southern Methodist University, thecameronstewart@gmail.com

Michael Mazel
Southern Methodist University, mmazel13@gmail.com

Bivin Sadler
Southern Methodist University, bsadler@smu.edu

Application of Probabilistic Ranking Systems on Women's Junior Division Beach Volleyball

Cameron Stewart[1], Michael Mazel[1], Bivin Sadler, PhD[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] SMU Technical Assistant Professor and Course Lead Faculty,
Southern Methodist University,
Dallas, TX 75275 USA
{cameronstewart, mmazel, bsadler}@smu.edu

Abstract. Women's beach volleyball is one of the fastest growing collegiate sports today. The increase in popularity has come with an increase in valuable scholarship opportunities across the country. With thousands of athletes to sort through, college scouts depend on websites that aggregate tournament results and rank players nationally. This project partnered with the company Volleyball Life, who is the current market leader in the ranking space of junior beach volleyball players. Utilizing the tournament information provided by Volleyball Life, this study explored replacements to the current ranking systems, which are designed to aggregate player points from recent tournament placements. Three probabilistic/modern ranking techniques were tested, specifically an Elo variant, TrueSkill, and a random walker graph network. This study found that Elo could predict match outcomes with a 13% higher accuracy than the preexisting systems and TrueSkill with an 11% higher accuracy.

1 Introduction

Women's beach volleyball is the fastest growing National Collegiate Athletic Association (NCAA) Division 1 sport over the last five years (*Beach,* 2022). The rise in popularity is being seen at Division 2 and Division 3 levels as well. To put this in perspective, the number of college programs that offer the sport at a varsity level across all divisions has increased from 15 to 173 over the past ten years (*Total # of College Beach Teams*, 2022). This explosive growth has also been seen in the junior divisions (Ages 10-18) and has created logistical challenges for college scouts to sort through thousands of potential recruits. With the number of scholarships available in women's beach volleyball increasing from 35.5 to 188.5 (531% increase) from 2012-2018, the financial implications of these decisions are significant to both the school and the player (DeBoer, 2019). For this reason, the college scouts have an ethical responsibility to utilize methods that limit bias as much as possible when evaluating potential recruits. Currently, scouts must rely on websites that can aggregate national and local tournament results to systematically assess the skills of the large number of players.

There is a challenge for websites that collect beach volleyball tournament results to clearly communicate a player's ranking to a scout, due to organizational structures and varying point systems. The current ranking system used by websites to rank junior beach volleyball is based on the system developed by the Fédération Internationale de Volleyball (FIVB) for professional volleyball players (Glickman et al., 2018). This system has been slightly modified for juniors to allow a player to cumulatively gather points over the trailing 365 days at each tournament based on three criteria: finish position, tournament size, and age division (*Volleyball Life Point Systems*, 2022). There are multiple organizations that host tournaments for junior players, and each has defined their own point allocations for the three key criteria. The rankings for each of these organizations are unique, and there is no single comprehensive ranking system available. A player also may play in one or multiple of these organizations as well, so a single player may have multiple rankings. This also means the group of players that a player is ranked against within each organization is different. These inconsistencies can present a murky picture to a scout when rating a player and introduce bias in how these results will be interpreted.

This research will focus on exploring how to transform the existing ranking methodology in women's junior beach volleyball into a comprehensive single ranking by applying modern, probabilistic approaches. Research in ranking methodologies has been a relevant area of study in competitive environments for many decades. In the 1960s, the Elo rating system was a breakthrough for chess and is still used to rank players/teams in many different competitive leagues today (Glickman, 1995). Building on the Elo ranking system, both the Glicko (1995) and Microsoft's TrueSkill (2005) further progressed the model to capture the complexities around dynamic uncertainty in a player's ranking. These three models have been the foundation of competitive rankings research, and initial studies have shown that all three are more effective than the existing system at ranking professional beach volleyball (Glickman et al., 2018).

In addition to unifying the organizational ranking systems, this research has identified two potential gaps in the existing scoring methodology. The first gap is the level of granularity used when ranking players. By focusing rankings only on tournament finish position, tournament size, and age division, there is no inclusion of valuable game level information. This can dramatically increase the amount of information available and more easily allow for cross-organizational ranking comparison by including all game results. The second gap is that the rankings are based on cumulative points obtained in tournaments over the trailing 365 days which will reward activity over the substance of the outcomes. For example, a player would get the same ranking score improvement for winning two different tournaments with 20 teams regardless of the ranking of those other teams. The current system is not able to factor in the quality of opponent. These gaps give this research a significant opportunity to bring a more sophisticated approach to this domain.

To capitalize on these gaps and create a unified ranking, this research needs access to historical game level information for tournaments in this domain. For this reason, the research team has partnered with the company Volleyball Life (https://volleyballlife.com/) who has access to one of the most extensive junior beach volleyball databases in the United States. Volleyball Life is also the market leader for users who are looking to track player rankings for junior beach volleyball. Based on

data availability, this research limited the scope to unifying the rankings of the four most popular organizations in competitive junior beach volleyball according to Volleyball Life which are the Amateur Athletic Union (AAU), Association of Volleyball Professionals (AVP), Beach Volleyball National Events (BVNE), and P1440.

The aim of this research is to develop a methodology to unify and enhance the player rankings of the four most popular competitive organizations in women's junior beach volleyball by applying modern, probabilistic ranking techniques. By creating a methodology to present a clear centralized rankings that more accurately evaluates players, scouts will have better tools available to reduce bias during the recruiting process.

2 Literature Review

2.1 Cumulative Point Models vs. Probabilistic Models

Probabilistic models gained notoriety in the 1960s with the creation of the Elo rating system (Glickman, 1995). Players are assigned scores that are updated based off the outcome of a game along with the projected chance of a particular player winning that game (Albers and Vries, 2001). Over the years variations of Elo were created, including the widely used Glicko rating system (Glickman, 1995). This algorithm added an additional parameter for the standard deviation of each player's score. While these algorithms were initially designed for chess, they can easily be applied to other one versus one player games (e.g., Scrabble, table tennis). In 2005, Microsoft created the TrueSkill algorithm for ranking players in video games, which evolved to become an incredibly flexible algorithm that could be applied to most team games (Herbrich et al., 2007). Originally, this was designed for two versus two scenarios but then was extended to include any number of team members.

While these probabilistic methods have seen tremendous growth in popularity, they are not yet adopted across all competitive landscapes. The current, professional beach volleyball ranking system, developed by the FIVB, sums points collected over the most recent year per player (Glickman et al., 2018). Players are awarded more points for placing higher in tournaments, and bonus weights are applied to larger-scale tournaments. However, a player cannot lose points from a poor performance. Glickman, Hennessey, and Bent conducted a study testing the predictive power of FIVB against four probabilistic models (i.e., Elo, Glicko, Glicko-2, and Stephenson). They found that FIVB generated the worst misclassification rate for match outcomes at 35%. The Stephenson model performed the best with a 31% misclassification rate (Glickman et al., 2018).

Many other beach volleyball leagues structured their ranking system in a similar format to FIVB. Association of Volleyball Professionals (AVP) is an association that focuses on United States professional beach volleyball competitions (AVP, 2021). Similar to FIVB, they also use a cumulative summation system in which they typically analyze player performance based on their best five results from the last 365

days (AVP, 2021; Glickman et al., 2018). Tournaments with more competitors also have a higher ceiling for potential points earned. Volleyball Life is a similar association to AVP that manages beach volleyball tournaments for kids and young adults. Their Amateur Athletic Union (AAU) point system totals player points from all competitions from the past 365 days. The higher someone places, as well as the more prestigious the event, the more points awarded.

2.2 Algorithm Advancements

While FIVB performed worse than each probabilistic model, it still provided a unique advantage. Specifically, tournament prestige is built directly into the formula by using a weighting criterion (Glickman et al., 2018). Prestigious events are inherently a larger focus for players and spectators, and these are reflected in the number of potential points awarded in the FIVB system. The probabilistic models do not discriminate between the perceived importance of tournaments. There has been only a small collection of studies that explored probabilistic models accounting for tournament prestige. Beumer (2021) analyzed Judo competitions using the default Elo algorithm against two variants with a larger K-factor for more prestigious tournaments or for later rounds within a tournament. The K-factor in Elo determines the speed at which a player's rating can rise or fall, so the larger this parameter, the greater a player's score will adjust after a match (Albers and Vries, 2001). Between the baseline and two variants, players finished with significantly different ratings. The exact effects that these methods had on the rating of an individual athlete was still unclear, however. In various situations, the fluctuations of a player's rating were challenging to interpret (Beumer, 2021). Due to the limited literature of the prestige feature, the impact of this on beach volleyball is challenging to speculate.

The traditional probabilistic models were designed for competitions which naturally are then analyzed temporally; however, they do not effectively consider the possible evolution of players' skill. Utilizing a time series method allows for two major advantages: 1) A player's rating will be less reliant on the random order of opponents causing skill updates along with their random skill at the time, 2) If a previous player's skill adjusts sharply over a short period of time, a time series model can reflect that change in recent previous opponent's rankings (Herbrich et al., 2008). This method not only can improve match outcome prediction, but also put historical players in better perspective compared to modern players. In 2008, a research team explored the idea of expanding TrueSkill with the use of time series analysis to infer a skill curve for each player (Herbrich et al., 2008). To analyze this, the research team studied match results and skill ratings of top chess players over the last 150 years. The methodology proved to be computationally taxing but provided a significant advantage over traditional TrueSkill. A major drawback of this algorithm is that TrueSkill is already challenging to implement compared to its predecessors. These time series additions further exacerbate this issue.

Elo is perhaps the most interpretable probabilistic method due to its limited number of parameters (Albers and Vries, 2001; Glickman, 1995; Herbrich et al., 2007). Various studies have found ways to tweak this algorithm to improve its performance while maintaining its simplicity. Ingram (2021) explored methods of

extending the Elo algorithm with the inclusion of margin of victory, correlated skills (surface of match), while also accounting for differences in format (number of games in a set). The model used data from 2010-2019 from the "OnCourt" dataset. After comparing many combinations of adding and removing these new factors with Elo and Glicko, the result showed the best model was the Elo model with all the additional features. Compared to many other applications where Elo does not hold up, this shows that Elo can be more effective than advanced models in certain domains. Similar to Ingram, Sullivan and Cronin also found that Elo's predictive power can be enhanced by using margin of victory, except their focus was within the English Premier League (Sullivan and Cronin, 2015). Algorithm improvements that were considered include accounting for home field advantage, consecutive win/lose streaks, and adjusted K-factors. All hypothesized improvements ended up outperforming the original Elo system. When all four parameters were optimized, the new Elo was found to outperform the previous by 20%. The most impactful feature of this improvement was the inclusion of home field advantage.

With the expansion of parameters added to improve predictive power, there has also been a response in which players intentionally perform worse in the short term to maximize their long-term rating. Ebtekar and Lieu (2021) set out to make a model that was robust to these situations. Specifically, TopCoder and Glicko-2 were found to be prone to being exploited by players purposely losing matches to increase their uncertainty score. With a higher uncertainty score, future consecutive wins would then be rewarded at a greater level. Other critiques of popular rating systems were TrueSkill's tendency to over-respond to select matches if the performance was particularly unique and unexpected. The study's proposed algorithm was found to be more favorable compared to existing models in both predictive power and computational speed.

A major challenge in rating systems is how to accurately rate new players. Some algorithms provide the same starting rating for every unranked player (Albers and Vries, 2001). In other rating algorithms, players are not provided scores immediately after their first game, and instead, a larger sample must be collected beforehand (Herbrich et al., 2007). USA Table Tennis implemented a unique solution which considered game scores instead of just game outcomes (Marcus, 2001). This additional data was only considered for unranked players and helped to provide further insight between opponents. A major limitation of this method is in practice, many scores are not reliably recorded and are handled by the players themselves, instead of officials. Other variants allowed for larger ranking shifts for players with limited tournament appearances as well as for those who have not participated in recent tournaments (Glickman, 1995; Marcus, 2001).

2.3 Network Methods

Network methodologies for understanding complex systems have been around for several decades. These methods differ from the previous algorithms discussed by conveying information through a set of nodes in a system inter-connected by a set of edges. The recent surge in popularity in network methodologies has come from the success of Google's PageRank algorithm in 1998 (Brin and Page, 1998).

Perhaps, the fairest method to rank players in a competition network would be to have every team play every other team an equal number of times. This is what is called a complete network. When a ranking is derived from a complete network the ranking is called the "Natural Ranking" (Park and Yook, 2014). Unfortunately, this is often not feasible for most competitions due to time and health constraints. Any network that is not complete is in-turn defined as an incomplete network which typically means you have at least two players who have never interacted (Park and Yook, 2014). This is similar to ranking beach volleyball players because many of the players will have never played and need to be compared. The collective set of assumptions not proved through direct competition is called the hidden network. Research from Park and Yook (2014) showed how to use Bayesian Inference to approximate the hidden network to create a pseudo-Natural Ranking (Park and Yook, 2014). This Bayesian Inference method has an expected value and variance similar to TrueSkill. Using English Premier League and American College Football as a reference, this study demonstrates this Network Bayesian Inference method is as effective (if not more) than traditional probabilistic models such as Elo.

Many of the prominent network methodologies were not designed to rank players based on competitive outcomes, but modifications allowed some to become useful with incomplete networks. Research from Beggs et al. (2017) adjusted the PageRank algorithm to use an iterative system where the loser votes for the winner (to create an edge) which was collected in an adjacency matrix (Beggs et al., 2017). Although the network can become so complex that it is not visually interpretable, the modified PageRank method can naturally rank based on node centrality which can remove the bias that often occurs in traditional statistical ranking algorithms. When tested on ranking track athletes from 2016, the modified PageRank outperformed the existing points-based ranking system as well as other traditional, algorithmic approaches.

Novel network algorithms were also created to approach ranking entities in competitive environments. A study by Park and Newman (2005) created the Park-Newman Network Ranking Method (also known as the Win-Lose Score) (Park and Newman, 2005). This algorithm sets up a network that focuses on the efficient calculation of direct and indirect wins. A direct win example is Team A beats Team B. An indirect win occurs when Team A beats Team B and Team B beats Team C. Team A has now indirectly beaten Team C. The ability to capture indirect wins is essential in incomplete competitive networks where many teams do not play. The algorithm uses a linear combination of direct wins and down-weighted indirect wins to rank the nodes in the network. When applied to team rankings in the 2004 College Football Bowl Championship Series (BCS), the Park-Newman Ranking Method was able to outperform the existing BCS composite computer ranking when compared to the official rankings.

The weakness of many network designs used for ranking, such as the Park-Newman Network Ranking Method and the modified PageRank, is that they assume the skill level is static across the entire timeframe provided. In competition, the skill of players will likely fluctuate over time. When a network accounts for changes in data based on time, it is called a temporal network. To create a temporal network focused on competitive ranking, Motegi and Masuda (2012) applied a time-based dynamic centrality measure to the Park-Newman Method called the Dynamic Win-Lose Score (Motegi and Masuda, 2012). This network methodology updated the

network after every game and stored the updates sequentially in time. This method added two unique additions to the existing algorithm. First, it added a time decay to players' rankings. Second, it retro-actively captured the actual skill of players when an indirect win occurs. This new method was applied to predict solo tennis matches from 1972-2010. When compared to the actual player rankings, the new Dynamic Win-Lose Score outperformed the original Park-Newman Method in larger samples.

The discussed advancements in network methods have primarily been focused on the construction of the network and less on the ideal evaluation method. Research from Shin et al. (2014) showed that using a Random Walk methodology (similar to Google's Page Rank) can be effective in other common network constructions such as the Park-Newman Method (Shin et al., 2014). A Random Walk is a simulation where an entity starts in a random location and traverses the network. The Random Walker most commonly follows the gradient towards the stronger nodes. When you repeat the simulation many times, you develop a natural ranking based on how often the Random Walker ended in each node. Using the English Premiere League and National Football League, the study showed that the Random Walk methodology outperformed other commonly used evaluation techniques such as node centrality, linear combination, and node connectivity.

Network methodologies have advanced recently in the competition ranking domain and offer a unique alternative to traditional statistical methods. It is anticipated that capturing varying skill over time would benefit ranking junior beach volleyball players. Most network structures do not inherently capture this temporal difference, which in turn neglects player improvements.

2.4 Team Dynamics

Ranking players within a team environment can cause two primary additional complexities compared to a one-versus-one competition. First, a model must determine how the strength (or weakness) of a teammate will impact the expected outcome of a competition. The expected outcome has a direct impact on the ranking update after the match. The second complexity one can additionally consider is whether the level of cohesion in a team can impact an outcome. For example, a team who has played together for years will likely outperform another team of the same rankings who have never played together.

Developing a team strength from the individual teammate rankings can be done with simplistic algorithms. A relatively naive approach to handling multi-team games would be to recognize each opponent as an independent match (Williams, 2013). For example, if there are two teams each with two people, a game could end in either the first team winning or losing. If team one won then the ranking of person A on team one would be updated as if they won in a one-versus-one against person A on team two, as well as a one-versus-one against person B on team two. The primary limitation of this method is that team average ratings are not considered, so a player will see an especially high change in their rating after facing a player of a much different rating. To overcome this attribute, another approach considers using the mean rating within a team and across all opponents faced. A major advantage of these approaches is it can transform any one versus one algorithm (e.g., Elo) to one that can

be utilized for team games. In a series of simulated games with simulated players, evidence has shown that averaging the opponents' rating statistics resulted in a better overall predictive accuracy than using individual updates.

Team rating differential can also be factored out by applying more sophisticated statistical methods as shown in a study from Clarke and Leister (2019). Given sufficient sample size over the course of a season, an additive model created with regression and exponential smoothing was used to rate non-elite tennis players. This model was employed to remove the partner effect across doubles games (Clarke and Leister, 2019). This method leveraged the fact that the dataset had significant overlap between players within teams and opponents. Most tournaments involved teams of three pairs of two players and numerous games would be played with different team pairing permutations. This allowed the researchers to remove the partner effect and then fit an additive model. The method used to evaluate the ratings considered the subjective feelings of players. Most of the players felt their rankings from the additive model were reasonable for their most recent season.

Markov Chains offer a different approach to discerning how to understand within-team contributions by analyzing play by play data. A Markov Chain works by estimating the transitional probabilities of the current state to each possible other state. This method was tested in a study from Strauss and Arnold (1987) by analyzing each individual rally within a racquetball game and the rally outcomes. This could then be repeated to develop match outcome predictions while simultaneously differentiating between teammate abilities. The study's methodology could also apply to other doubles sports where someone serves the ball, such as volleyball, squash, or badminton (Strauss and Arnold, 1987). The main drawback to this method and the other previously discussed team strength methodologies is that they demand a very structured partner system and/or incredibly precise documentation of data (Clarke and Leister, 2019, (Marcus, 2001).

Network methodologies like those discussed in the previous section allow you to understand teammate contribution. A study from Quint (2007) used a network methodology in Contract Bridge to overcome the drawbacks of the previous methods (Quint, 2007). Contract Bridge is a two versus two card game which has similar teammate complexities to beach volleyball. The proposed ranking system for bridge players centered around the idea of overcoming the "nonuniqueness problem" (Quint, 2007). That is, the issue of determining who is the better/worse player on a team and by what magnitude. A network analysis which used diagonally dominant matrices was employed that reasonably ranked the players. The primary issue when constructing the matrices was developing a way to ensure players had met the model criteria requiring an observation with a matching teammate for the player within the team. For example, say player A and player B are a team. To compute their skill, they both need to have also played with any other common individual. To compare player A to player B: player A and player W as well as player B and player W need to have played before. Another network analysis from Gill and Swartz (2019) investigated how to define if the outcome of a game is more dependent on the strong or weak link in pickleball (Gill and Swartz, 2019). A strong or weak link sport is determined by whether the stronger players on a team or the weaker players on a team have a larger impact on the outcome. The research ranked pickleball players using a network analysis to create linear models based on the team and opponent rankings. Using a

normalized parameter between zero (completely weak link) and one (completely strong link), the model showed pickleball relies more heavily on the higher ranked player with a link weight of 0.87. Using a similar calculation, this research could develop a better estimation of teammate contribution than the mean from the simplistic model.

To address the second complexity of ranking players in multi-team competitions, a study from DeLong et al. (2011) attempted to consider team cohesion in the ranking process (DeLong et al., 2011). Using professional team gaming data from 2008 and 2009, the study attempted to modify traditional methodologies (i.e., Elo, Glicko, and TrueSkill) by including team cohesion in the model to predict match outcomes. The challenge with this methodology of using every unique two-player team as a feature can inevitably lead to low sample sizes with a subset of teams. While being cautious of applying this methodology to teams with low sample size, the result of this study showed a significant improvement to Glicko and TrueSkill prediction accuracy by including the team cohesion metric (particularly in closely matched games).

There are many approaches that can be used to assess the complexities of ranking players in multi-team competitions. Individual teammate contribution can be estimated with simplistic models, but when you have sufficient data to meet the needs of a network analysis, the reliability of these results can be improved. Including team cohesion in traditional models can boost predictive accuracy when the teams have sufficient sample size.

3 Methods

The data for this research was provided by Volleyball Life. They provided de-identified game level and player data for women's junior beach volleyball. The data included 783 tournaments, 61,000 games, and 11,000 unique players, along with game level details such as game scores. Data was cleaned through removal of game records containing irregularities such as incorrect number of players and incorrect dates.

The methods employed include two traditional ranking algorithms, Elo and TrueSkill. Various parameter values were explored to optimize their performance to the beach volleyball dataset. Additionally, a graph network ranking algorithm was created that paralleled the structure of social media network algorithms. Nodes within the network in this study represented each player, while edges represented the average proportion of points lost to each connected player. The network design was structured to reflect the modifications of traditional network methods to account for incomplete network structures, competitive ranking through node centrality, and dynamic network structures. Using a random walker that moves in a random direction based on a weighted average of the edges, ensured the walker typically progresses toward the better player. The number of steps for the random walker was controlled by random walker iterations. By adding a set probability to regenerate the walker randomly to a new node called restart probability, allowed the network to overcome the incomplete structure while identifying node centrality. To account for the dynamic network structure over time, the network dropped any information added to the network beyond a specified training period. The players were ranked in descending order according to

the quantity of times the random walker landed on the player. The outcome of each game was predicted using an average of the rankings of the players on each team. The accuracy of these predictions was then evaluated.

The traditional methods that were employed (i.e., Elo and TrueSkill) followed the format of Glickman (2018), which ranked professional beach volleyball players. The rating systems were trained on the first n period of data, then the remaining one minus n period of data were used as the test set. This translated to the first 52,000 games being used as the train set and the following 9,000 as the evaluation set. Within the evaluation set, players' ratings were updated immediately after each match. In other words, the trained model with player ratings would be used to predict the outcome of the immediate upcoming match. The predicted outcome was recorded for that particular match and then the true game outcome was used to update the players' ratings. This process would then be repeated for each following match.

Accuracy of the traditional methods was measured according to the model's predicted outcome and the actual outcome. All outcomes were binary, so the model simply predicted win or loss. Each rating algorithm would predict the winner based on which team had the higher average rating. In addition, if two teams had the exact same rating, then the models would be constrained to only being able to predict a tie. Matches that met this criterion in the validation set were skipped and did not contribute to the final performance metrics.

Each of the traditional models were optimized based off each algorithm's unique hyperparameters. The hyperparameters trained on TrueSkill included sigma and beta. Sigma represents the rate at which the variation in expected rating score changes. Beta represents the expected variation in performance that naturally occurs in competition. The parameters were used to update player ratings using the TrueSkill Update Algorithm (Herbrich et al., 2007). TrueSkill's optimization process would begin first with sigma, followed by beta. A range of values were considered for both hyperparameters. The best performing sigma value, according to training accuracy, was then used for the following beta tuning process.

Elo was optimized using a random search approach and considered five hyperparameters. These include the K-factor for new players, maximum K-factor, inertia, rating decay rate, and rating restore rate. Three of these five hyperparameters influenced the K-factor depending on how many matches a player participated in. Inertia was the rate at which the K-factor for new players would approach the maximum K-factor modifier, after each game. The formula for inertia can be found in Figure 1. Static K-factors (K-factors that did not evolve depending on players' samples) were also considered. The decay rate and restore rate controlled if and how quickly players' ratings would be updated towards the initialized player Elo of 1200. After each day, the decay rate would reduce any players' rating by a small fraction, if they were above 1200. The same was tested for restore rate, except this would increase players' Elo by a small fraction, if they were below 1200.

Figure 1

$$\text{Adaptive K-Factor} = K_0 + \frac{K_\infty}{n^{1/\text{inertia}}}$$

Where:

K_0 = K-factor for new players

K_∞ = Maximum K-factor modifier
n = Number of games played
inertia = Resistance to K-factor change

Since Elo was designed for one versus one competition, another model consideration was how to convert the system to support a two versus two structure. Two methods from Williams (2013) were applied to the random search. The first method was performed by averaging the Elo of each team pairing. This essentially created new, artificial players to then calculate probabilities and points won/loss. These aggregated team amounts were applied to each player individually. That allowed for each player to have their individual Elo scores as well, instead of simply having a team score. For the second method, only the opponent parameters were averaged together to form a new, artificial opponent. This opponent was then used to calculate and update the outcome for each individual player of interest.

In addition to the three ranking algorithms, data for three of the preexisting junior beach volleyball point systems were provided. Specifically, BVNE, P1440, and AAU player points were available on a weekly basis across the evaluation set. Accuracy for these point systems were calculated in the same manner as the hypothesized ranking systems. Final model performance and evaluations were based in accordance with a given model's runtime, interpretability, and effectiveness as the methodology to be implemented at Volleyball Life after the completion of the research. The primary effectiveness metric used was accuracy, with log loss also considered as a supplementary metric. Accuracy was selected as the main metric over log loss because log loss could not be produced for the three preexisting point systems or the graph model. Elo and TrueSkill naturally produce a probability, which allowed for log loss to be calculated, unlike the graph model.

4 Results

The hypothesized models were tuned according to the accuracy score on the train set. The best performing TrueSkill model had a sigma of 2.5 and beta of .5. TrueSkill resulted in a 75.21% average accuracy on the evaluation set. This meant, nearly three quarters of the time the model accurately predicted the winner of the match. The team with the higher average rating (mu) was always selected as the predicted winner of any given match. The log loss associated with this TrueSkill model was 0.284 per match on average.

The best performing Elo model, according to accuracy, had a non-adaptive K value of 100, and consequently used the same K value for all players, whether they had zero or hundreds of games previously played. The best restore rate discovered was 0.3% after each day and a decay rate of 0%. In addition, averaging the points together from both players per team was selected over averaging only the opponents. This Elo model was found to outperform TrueSkill having an average accuracy of 76.81%. The log loss associated with this Elo model was 0.492 per match on average.

Due to the natural volatility in competitions, stronger players losing to weaker players is a common occurrence. For example, a 1500 rated player, according to Elo, would be expected to win most, but not every game over the 1400 rated player.

In fact, according to the Elo algorithm, the 1500 rated player would be estimated to win exactly 64% of the time on average. TrueSkill and Elo can not only be measured by their accuracy, but also based on the win rate probability alignments between actual and projected probabilities. Since a 1500 rated player is expected to beat a 1400 player 64% of the time, the predicted outcomes can be evaluated and aggregated for all matches that occurred between a 1500 and 1400 player. If the projected winner of the Elo algorithm was correct 64% of the time between a 1500 and 1400 player, then the model is performing exactly as expected. This provides additional evidence that the 1500 and 1400 rated players were appropriately rated.

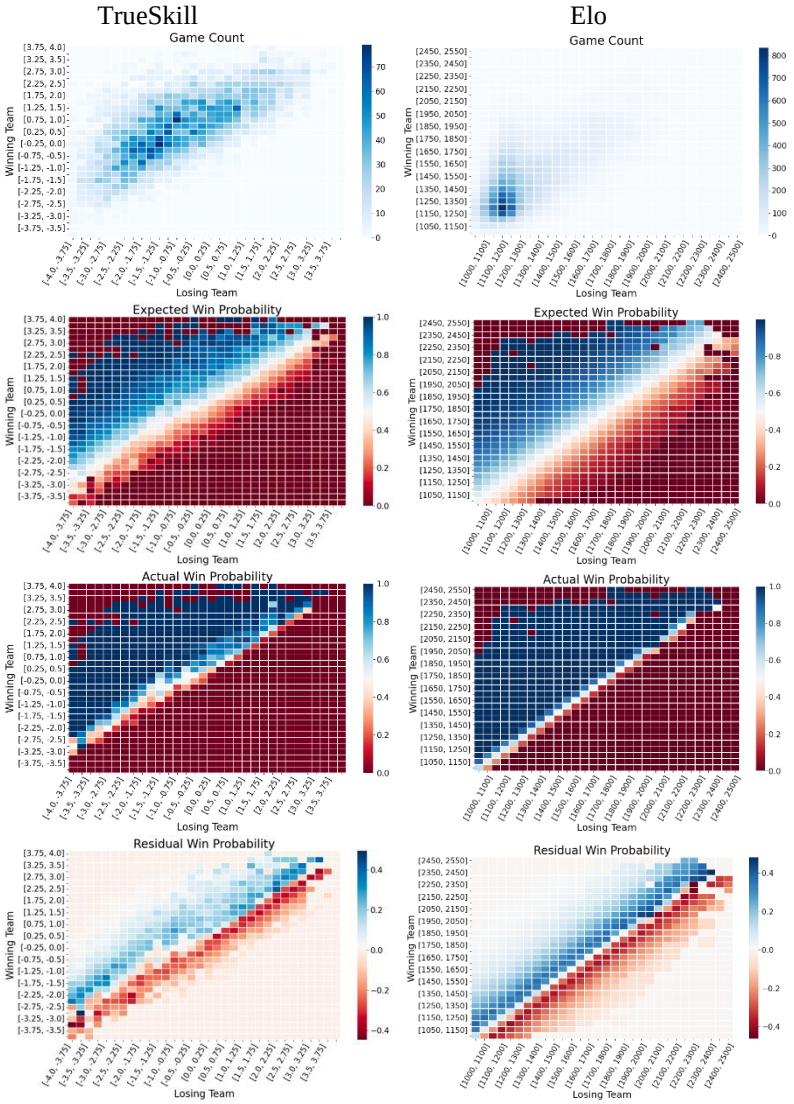
Figure 2 illustrates the performance of TrueSkill and Elo using four visualizations that are key to understanding the relationship between expected win probability against the actual win probability. Figure 2 row 1 shows the frequency of samples collected based on the winning team (WT) and the losing teams' (LT) ratings using a color scale where the darker blue values indicate higher frequencies. The diagonal from the bottom left to the top right of the plot indicates when the WT and LT had the same (or nearly the same) ratings. Any observations above this diagonal indicate the WT had a higher rating than the LT and vice versa for below the diagonal. In more than half of the samples for both Elo and TrueSkill, the WT falls above the diagonal. This provides evidence that each model is correctly discriminating between better and worse teams, particularly compared to a random guess. If the accuracy results were 50%, or as good as a random guess, then the samples would be randomly distributed around the diagonal. This plot also reveals the distribution of ratings across each of the algorithms. TrueSkill has a relatively normal distribution centered at a rating of zero, while Elo has a right skew distribution with a peak frequency at a rating of 1200. Figure 2 row 1 produces additional insight when cross referenced with the plots in rows 2-4. It indicates which rating intersections have the largest weight and where frequencies are so low that anomalies may appear due to lack of sample size. These anomalies appear, for example, in the top left corner of row 2 and 3 where sample size was exceptionally low, and the area is depicted as dark red (near zero win probability).

Figure 2 row 2 indicates the expected win probability of both Elo and TrueSkill based on the WT and LT ratings. This plot uses the expected probabilities from the respective model across the recorded games to understand how the model is predicting the probability of winning across different rating matchups. In Elo, the update formula is constant based on the relative ratings and parameter values. Therefore, the predicted win probability is symmetric across the diagonal. For TrueSkill, the probability of winning is dependent on the number of games played and relative consistency of player performance. The TrueSkill plot in row 2 is non-symmetric over the diagonal due to differences in the dependencies across the matches played between teams with the specified ratings. If dependencies were held constant, then TrueSkill would have symmetric expected win probabilities over the diagonal.

Figure 2 row 3 depicts the actual win probability of both Elo and TrueSkill based on the WT and LT ratings. The plot for each algorithm describes the outcomes of the same sets of games in the evaluation set but appear slightly different due to differences in the rating distribution between the players. The visualization describes

the ideal plot for the expected win probability. Any difference between the expected and actual win probability plots is captured in Figure 2 row 4, which represents the residual win probability based on WT and LT rating. The higher the residual win probability the worse the model is at predicting the outcome of games with teams at the specified range of ratings. TrueSkill has a tighter distribution of high residual win probabilities compared to Elo. This captures the same idea produced by the log loss metric, with TrueSkill producing the lower of the two models.

Figure 2



It should be noted that because the graph network model does not naturally create a probability for each match outcome, the visualizations in Figure 2 could not be replicated for this model. The training and testing methodology for the graph network was also slightly differed compared to Elo and TrueSkill. Instead of updating the model after each prediction, the graph network was updated after 30 days of consecutive matches across all players. This immediately put the model at a disadvantage because it was not updated as frequently as Elo and TrueSkill. This monthly prediction window was chosen due to the significantly longer runtime for the graph network. The best performing graph network model had an accuracy of 63.75%. The optimal parameters that were associated with this score had a training period of the prior 90 days of matches, random walk iterations of 70,000, and a random restart probability of 0.05%.

The average runtime for a full train and test passthrough for the graph network model took approximately eight minutes when a random walk iteration of 70,000 was used. The time increased proportionally to the number of iterations. Meanwhile, Elo and TrueSkill were considerably faster, taking about four seconds for Elo and 2.6 seconds for TrueSkill. This also allowed for a larger range of possible values for each parameter to be searched, since it took only a fraction of the graph network's runtime.

The three preexisting point systems evaluated were BVNE, P1440, and AAU. These point systems did not need to be trained on the training set. This is because at any point in time, these point systems would have points allocated to players according to their past year of tournament placements. Instead of data being updated daily like Elo and TrueSkill, the data was updated each Monday. In the test set, BVNE, P1440, and AAU had accuracy scores of 61.06%, 61.50%, and 63.96%, respectively.

5 Discussion

5.1 Application of Results

5.11 Interpretations

The three models considered were TrueSkill, Elo, and a graph network. Each one developed their own player ratings, and then these ratings were used to predict the match outcomes on the evaluation set of data. Whichever team had the higher average ratings were predicted by the model to win the match. The results showed that TrueSkill had an accuracy of 75.21%, Elo had an accuracy of 76.81%, and the graph network had an accuracy of 63.75%. Game outcome prediction accuracy was used as a proxy for the best rating algorithm. The results suggest that Elo would perform the best, followed by TrueSkill, followed by the graph network. The secondary metric considered was log loss with Elo producing a 0.492 and TrueSkill a 0.274. Unlike accuracy, this supports TrueSkill as the preferred model. This would suggest that TrueSkill produced more precise probabilities assigned to each match outcome. Although Elo is correct more often in the match outcome, it is predicting incorrect outcomes with larger residuals.

This study provides evidence that algorithmic ranking systems, specifically Elo and TrueSkill, would provide a more precise ranking method for junior beach volleyball players, compared to the preexisting, cumulative systems. TrueSkill outperformed the best cumulative system with an 11% improvement in match outcome accuracy, and Elo outperformed the best cumulative system with a 13% improvement.

5.12 Implications

The advantage of any of these three models over the current, cumulative ranking system is the reduction in the human bias component. The cumulative point systems (e.g., AAU or BVNE), were constructed to reward players with points according to tournament placement and tournament size. Therefore, those systems exhibit a bias towards players that participate in many tournaments, even if those players had relatively poor performance. In contrast, Elo, TrueSkill, and the graph network were designed to reflect player ability.

These models were trained and evaluated on junior beach volleyball players with varying amounts of skill and levels of experience. It would be expected that the models would perform similarly if extended to new data or put into production. It is less certain if these models would generalize to leagues such as college or professional level beach volleyball. The algorithms may also need their parameters reoptimized if the scope of competition changes. If implemented, any of these models would allow for a straightforward way to monitor their accuracy over time. Ideally, one of these could replace the many cumulative rankings in the sport and provide for a consistent ranking.

5.13 Recommendations

Not only should the performance be considered towards future research or real-world deployment, but also the logistics of the models. Specifically, the graph network model took several minutes to train on the machine used for this study, compared to TrueSkill and Elo which only took several seconds. Depending on the frequency of when scores need to be updated, as well as the total number of players within the database, this could change the practicality of the graph model.

Due to the significantly higher accuracy of Elo, it would be recommended as the ideal algorithm based off the results from this study. However, if match probabilities were more important to an individual, then TrueSkill would be better due to its lower log loss. Accuracy tends to also be more interpretable than log loss, so that is another reason as to why the Elo model would be deemed superior in this particular environment.

5.2 Limitations of the Study

The results suggest that rating junior beach volleyball players using a probabilistic method would provide a significant improvement over the preexisting, cumulative models. The key characteristic of the probabilistic methods is that they are

less susceptible to human influence and inherent scoring bias, and therefore should serve as a more reliable system. This may act as a limitation, however, in scenarios when the ranking designer wants to reward players based on match attributes, like the prestige level of a tournament. Conversely, many of the cumulative rating system components may be deemed arbitrary and biased. This may result in compromised statistical validity or degraded perception of fairness. The probabilistic models also have an inherent advantage of having an expected probability for each game which allows for the continuous evaluation according to prediction accuracy as well as the residual for actual versus expected outcomes.

Another potential concern of the Elo model is the extremely reactive K-factor. As mentioned earlier, most K-factors fall around 10 to 50, with larger K-factors resulting in quicker rating adjustments. When Elo was tuned for the best K-factor, 100 was found to be the best performing. While this number was supported from the performance in this specific data, the unusually large value presents suspicion. This may be a byproduct of Elo being used in a two verse two setting, while all the K-factors from the cited literature were only one versus one. This high K-value was likely a contributor to Elo having a worse log loss than TrueSkill. A high K-value results in strong probabilities predicted for each match, so when Elo is incorrect, it sees a huge penalty. This is also supported by the large residuals in either positive or negative directions shown in Figure 4.

Traditionally, domains that applied network algorithms used much larger and more connected datasets than the data used in this study, such as social networks. Also, computational demand is relatively high to produce consistent ranking recalculations. This is vital due to the dynamic adjustment of player abilities over time. If the data utilized had a greater average player participation rate and the computational workload barrier could be overcome, the network methods discussed may be more practical and could achieve much higher effectiveness.

5.3 Future Work

Previous research has demonstrated that including additional predictor variable can increase model performance (Ingram 2021). Specifically, margin of victory and home court advantage have been found to lead to better accuracy scores. One trend that was discovered in the dataset is many competitions being clustered in Florida and California. Geographic location could potentially be used as a predictor variable, with the expectation that better players are more likely to either live in or travel to the most popular beach volleyball states. Future work could explore predictor variables such as these.

Another area of research that could build on this study is the use of neighborhoods in network algorithms as features in graph neural networks. Neighborhoods capture information about nodes and edges within a specified adjacency distance from the player. This could provide additional information to the model about the surrounding opponents of each player.

5.4 Ethics

Replacing the preexisting, cumulative model with these probabilistic methods could result in drastically different futures of junior beach volleyball participants. A considerable number of players eventually continue the sport into college or even professionally. The rating system could potentially influence the perception of these junior players. This could subsequently trickle down into having major financial impact or influencing the opportunity for juniors to eventually participate in beach volleyball at a higher level. These situations and adjacent ethical situations were considered thoroughly throughout the process of this study.

All players were de-identified at the beginning of the data intake process to ensure no player biases could leak into the following stages. Since only players' historical performance was used in this process, various demographic data had no opportunity to provide unintended influence. On the contrary, excluding these variables, especially age, could also result in adverse effects. A potential shortcoming could occur if a player started participating in the league at a young age, played poorly for many games, but years later was significantly better and only participated in a few games. This situation could potentially be quite common, because young players have not had enough time to develop their skills or bodies and will naturally see quite a low rating. This rating lingers throughout their junior beach volleyball run and could mask recent success. This hypothetical provides reason as to why age brackets may be beneficial, instead of having just a singular rating ecosystem.

Another important area that has rarely been explored in similar literature is the impact of player injuries. A player with an injury could be expected to perform much worse than their typical self during their period of recovery. None of the algorithms used in the study or found in related literature have accounted for cases such as this. Some may argue that poor performance during an injury period should not be weighted equally, and an algorithm should be designed to provide the fairest output as possible. This, however, would bring many logistical complications to implementing a "healthy player component" to the algorithm due to the varying degrees of injury severity. Additionally, not everyone would agree that it would be a fair component that should be included as it could be exploited to hide non-injury-related drops in performance.

6 Conclusion

This study demonstrated that probabilistic methods would likely improve the ranking performance of women's junior beach volleyball. Various algorithms were identified that were found to be effective in outside domains, but in similar applications. These algorithms included Elo, TrueSkill, and graph networks. Using real game data from Volleyball Life, the performance of these models was evaluated and compared. Findings showed that model performance was negatively related to the level of complexity in the algorithm. The recommended Elo model was found to achieve game level prediction accuracy of 76.81% and a log loss of 0.492. This model

provided an interpretable method to create a comprehensive ranking of all players across organizations that did not previously exist. Additionally, it only required game outcomes, and unlike cumulative systems, it excluded potentially bias inducing features. This new ranking methodology has the ability to provide college scouts with a more accurate ranking of players based on performance. This may promote recruiting best practices by removing ambiguity created by multiple organizational ranking systems.

This research built upon previous literature in the beach volleyball domain by exploring more complex algorithms such as graph networks and variants to the Elo algorithm. A holistic view of the hypothesized models was provided through visualizations, comparing the residual probability based on player ratings. Accuracy was leveraged as the primary metric, and log loss was added as a secondary metric to further explain the models. The methods contained in this study along with the model comparison techniques can be extended to future research of other competitive volleyball areas, as well as to competitive ranking research in other domains.

References

Albers, P. C. H., & de Vries, H. (2001). Elo-rating as a tool in the sequential estimation of dominance strengths. *Animal Behaviour*, *61*(2), 489–495. https://doi.org/10.1006/anbe.2000.1571

AVP. (2021). *Athlete Rankings*. AVP Beach Volleyball. https://avp.com/ranking-listing/

Beach. American Volleyball Coaches Association, LLC. (2022). https://www.avca.org/Groups/Beach-Volleyball#:~:text=The%20popularity%20of%20beach%20volleyball,five%20years%20in%20Division%20I.

Beggs, Shepherd, S. J., Emmonds, S., & Jones, B. (2017). A novel application of PageRank and user preference algorithms for assessing the relative performance of track athletes in competition. *PloS One*, *12*(6), e0178458–e0178458. https://doi.org/10.1371/journal.pone.0178458

Beumer, K. (2020). Measuring and Comparing the Performance of Elite Judokas. 75.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117.

Clarke, S., & Leister, B. (2019). Rating Non Elite Doubles Tennis Players. https://doi.org/10.21276/jaspe.2019.2.4.3

DeBoer, K. (2019). *2019 College Beach Volleyball*. Avca.org. https://www.avca.org/res/uploads/media/BeachVB-College-Summary-Report-8-19-_2.pdf.

DeLong, Pathak, N., Erickson, K., Perrino, E., Shim, K., & Srivastava, J. (2011). TeamSkill: Modeling Team Chemistry in Online Multi-player Games. In *Advances in Knowledge Discovery and Data Mining* (pp. 519–531). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20847-8_43

Ebtekar, A., & Liu, P. (2021). An Elo-like System for Massive Multiplayer Competitions. *ArXiv:2101.00400 [Cs, Stat]*, 12.

Gill, & Swartz, T. B. (2019). A characterization of the degree of weak and strong links in doubles sports. Journal of Quantitative Analysis in Sports, 15(2), 155–162. https://doi.org/10.1515/jqas-2018-0080

Glickman, M. E. (1995). The glicko system. *Boston University*, *16*, 16-17.

Glickman, M. E., Hennessy, J., & Bent, A. (2018). A comparison of rating systems for competitive women's beach volleyball. *Statistica Applicata*, *30*(2), 233–254. https://doi.org/10.26398/IJAS.0030-010

Herbrich, R., Minka, T., & Graepel, T. (2007, January). TrueSkill (TM): A Bayesian Skill Rating System. *Advances in Neural Information Processing Systems 20*, 569–576. Opgehaal van https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/

Herbrich, R., Minka, T., & Graepel, T. (2008, January). TrueSkill Through Time: Revisiting the History of Chess. *Advances in Neural Information Processing Systems 20*, 931–938. Opgehaal van https://www.microsoft.com/en-us/research/publication/ trueskill-through-time-revisiting-the-history-of-chess/

Ingram. (2021). How to extend Elo: a Bayesian perspective. *Journal of Quantitative Analysis in Sports*, *17*(3), 203–219. https://doi.org/10.1515/jqas-2020-0066

Marcus, D. J. (2001). New Table-Tennis Rating System. Journal of the Royal Statistical Society: Series D (The Statistician), 50(2), 191–208. https://doi.org/10.1111/1467-9884.00271

Motegi, & Masuda, N. (2012). A network-based dynamical ranking system for competitive sports. *Scientific Reports*, *2*(1), 904–904. https://doi.org/10.1038/srep00904

Park, & Newman, M. E. J. (2005). A network-based ranking system for US college football. *Journal of Statistical Mechanics*, *2005*(10), P10014–P10014. https://doi.org/10.1088/1742-5468/2005/10/P10014

Park, & Yook, S.-H. (2014). Bayesian Inference of Natural Rankings in Incomplete Competition Networks. *Scientific Reports*, *4*(1), 6212–6212. https://doi.org/10.1038/srep06212

Quint, T. (2007). A new rating system for duplicate bridge. Linear Algebra and Its Applications, 422(1), 236–249. https://doi.org/10.1016/j.laa.2006.09.025

Shin, Ahnert, S. E., & Park, J. (2014). Ranking Competitors Using Degree-Neutralized Random Walks. PloS One, 9(12), e113685–e113685. https://doi.org/10.1371/journal.pone.0113685

Strauss, D., & Arnold, B. C. (1987). The Rating of Players in Racquetball Tournaments. Journal of the Royal Statistical Society. Series C (Applied Statistics), 36(2), 163–173. https://doi.org/10.2307/2347548

Sullivan, C., & Cronin, C. (2015). *Improving Elo Rankings For Sports Experimenting on the English Premier League*. 9.

Tarlow, D., Graepel, T., & Minka, T. (2014, January). Knowing what we don't know in NCAA Football ratings: Understanding and using structured uncertainty. *MIT Sloan Sports Analytics Conference*. Opgehaal van https://www.microsoft.com/en-us/research/publication/knowing-what-we- dont-know-in-ncaa-football-ratings-understanding-and-using-structured-uncertainty/

Total # of College Beach Teams. Avca.org. (2022). https://www.avca.org/res/uploads/media/College-Beach-Program-Growth-2011-2022-7-21-.pdf.

Volleyball Life Point Systems. (2022). Volleyball Life. Retrieved March 21, 2022, from https://volleyballlife.com/points

Williams, G. J. (2013). *Abstracting Glicko-2 for Team Games* [Thesis]. University of Cincinnati.