Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/132/2024.20850

Elo in MOBA: Algorithm Comparison and Application Discussion

Jialong Dong[1,#,*], Haiyang Jiao[2,#], Yue Liu[3,#]

[1] University of California Davis, Davis, United States
[2] Minzu University of China, Beijing, China
[3] University of California Berkeley, Berkeley, United States

[1] jldong@ucdavis.edu
[2] oceanjiao_baird@163.com
[3] annika323@berkeley.edu
[#]These authors contributed equally to this work and should be considered co-first authors.
*corresponding author

Abstract. Online multiplayer games are supposed to match single players or squads with similar abilities. They depend on different kinds of rating systems to ensure a fair and competitive gaming environment. Traditional rating systems generally focus on single evaluation criteria, such as Elo scores or number of killings. However, games like *Overwatch*, which have more rating aspects, such as the usage of items and skills and the corporation of a whole team, need some more complex rating systems. Because of these, those traditional rating systems cannot satisfy the needs of modern multiplayer online games. Thus, it is urgent to develop new rating systems with new modes. In this paper, we are going to explore the newly designed Elo system with new models, especially its application in MOBA games. We will also talk about the TrueSkill and TrueSkill2 rating systems. These are two rating systems with high accuracy in win rate prediction, which can match fair games for players and teams.

Keywords: Elo, MOBA, Algorithm Comparison. Online multiplayer games

1. Introduction

Nowadays, online multiplayer games, such as *League of Legends* and *Counter-Strike 2*, are developing at full speed. At the same time, a skill rating system with higher accuracy is of great importance to the modern gaming environment. A fair and accurate rating system should consider different aspects of the performance of players and match players or squads with similar abilities. In games before, the Elo score is the only valuation criterion of each rating system. At that time, the rating systems just matched players who had similar Elo scores. However, when some players do a bad job but win the game or lose for some unexpected reasons, the traditional rating systems could be unfair. The Elo rating system was one of the traditional rating systems. It was used to match players in chess games back in the late 20th century, while it is not as good at predicting game win rates as chess. Games like *League of Legends* and *Counter-Strike 2* are much more complex than traditional chess games. So, rating systems with more complex calculations and accuracy are urgently needed. In recent years, we have

DOI: 10.54254/2755-2721/132/2024.20850

designed a kind of extension model to the standard Elo algorithm. In this way, we could make it satisfy various types of rating aspects. With those extension models, the match of players and squads is going to be more fair and competitive. At the same time, a new system applied to MOBA games is introduced. It is a kind of Elo rating system but specially designed for MOBA games. What's more, the TrueSkill and TrueSkill2 rating systems are designed for new games. They pay attention to some details of games. With the help of the Bayesian formula, the TrueSkill rating system and TrueSkill2 rating system can more accurately predict the win rate of multiplayer online games and, at the same time, match a fair and competitive game.

2. Background

2.1. The Elo Algorithm

The ELO scoring system was invented by Arpad Elo and was originally used to balance the skill levels of players in chess. The system assigns a numerical rating to the player, which is adjusted based on the results of each team's match. At the end of each match, if the player with a lower rating wins, he will receive many points from the opposing player, whereas if the player with a higher rating wins, only a few points will be transferred. However, in the event of a tie, the player with the lower rating will still receive some points from the player with the higher rating. A player's expected score is calculated based on the rating difference between the two players, and the rating adjustment is made using the constant K, which varies according to the difference between the players' ratings. Although the ELO ranking system has been superseded by its variants, it still plays an important role in the field of game ranking.

The standard Elo algorithm is a simple system with two equations as follows:

$$\hat{P}_{ijt} = \frac{1}{1+10^{-(R_{it}-R_{jt})/400}} \tag{1}$$

$$R_{i(t+1)} = R_{it} + K(W_{ijt} - \hat{P}_{ijt}) \tag{2}$$

The Elo algorithm can be considered as two steps. The Estimation Step is where the result is estimated through the system. This is the first equation. $\hat{P}_{ijt}$ is the point estimate of the probability (a value between 0-1) of player i winning the match against player j given player ratings of player i R_{it} and player j R_{it} The Update Step is where player rating is updated according to the result of the previous match. This is the second equation where the residual difference of actual win result W_{ijt} and the estimated win probability $\hat{P}_{ijt}$ is tuned by an arbitrary learning rate K and added with the player rating from before R_{it}. The learning rate is set by the user of the Elo algorithm to control by how much the player rating is changed after each match [2].

2.2. TrueSkill and TrueSkill2

In order to extend the traditional Elo system from Chess games, the TrueSkill rating system can better track uncertain- ties in player skills and player actions in games. Depending on an online learning framework, the TrueSkill2 rating system uses a Bayesian formula to analyze different aspects of the skills and performances of a player, which is more accurate than the TrueSkill system. They can both ensure fair play and increase player participation.

3. Exploring the ELO Model with Margin of Victory

The standard Elo algorithm's update method only considers the win result of previous games. However, in many cases, the Margin of Victory of the game should also be considered, i.e., by how much the players won or lost. For instance, a victory of 6:0 is different from a victory of 4:3. This paper by Stephanie Kovalchik proposes 4 different extension models to the original Elo system, including the Margin of Victory: the Linear Model, Joint Additive Model, Multiplicative Model, and

Logistic Model [2]. Through experiments, the author came to the conclusion that the Joint Additive is the best model for general purposes out of the 4 options.

To begin with, the keywords used in the experiment must be defined. Player rating is the ranking of a player's skills computed by the model. Latent ability is the player's actual skill and is the ultimate ranking that the player rating is trying to reflect. Every model presented by Kovalchik also follows the two steps of the standard algorithm: Estimation and Update.

Convergence is a key characteristic that the proposed models must fulfill in order to function as extensions of the Elo algorithm. It is the original Elo system's property that updated player ratings will eventually converge to their latent ability and remain stationary at that point. That is, after the certain number of iterations, player ratings updated according to the Elo update rules will come to reflect players' actual skills, thereby realizing the purpose of the Elo algorithm for balanced matchmaking.

There are two requirements for the extension models to adhere to the convergence property: stationarity and the Lipschitz conditions. By stationarity, Kovalchik means that the outcome of the system is only dependent on the relative player rating differences, independent of other factors. The Lipschitz conditions, on the other hand, are the constraints on the function to be uniformly continuous without abrupt changes. In the following presentation of the 4 extension models, how each model satisfies the convergence property is also analyzed.

We first explore the Linear Model, best used when assuming a linear relationship between player ratings and expected margin. Different from the standard Elo algorithm, which predicts the win result, this model estimates the Margin of Victory (MOV). The closer the MOV is to 0, the closer the match and vice versa. This model's Estimation Step for the expected margin is in proportion to the relative differences in the two players' ratings. The exact proportion has a tuning parameter to be optimized according to the specific need and purpose of using this model. The Update Step for the player rating adds the previous player rating of this same player with the residual difference between the actual margin and the estimated margin. Another tuning parameter K is used to control the magnitude of how much this residual difference is added. This Linear Model complies with the convergence property's stationarity because the only parameters are player ratings and expected margin. Additionally, it is Lipschitz continuous because the update function is continuously differentiable, thereby having no sudden fluctuations.

The second model, the Joint Additive Model, is a linear function most suitable for cases where both expected win result and margin are needed. As implied by the name, this model, in essence, is a combination of the standard Elo system and the Linear Model. In the Estimation Step, the expected win result and expected margin is the same function as the ones in the standard Elo algorithm and the Linear Model. The Update Step for player rating is the linear combination of the update function from the standard Elo system and the Linear Model. Since this update function is a linear combination, it follows that this system automatically satisfies the conditions for convergence.

The multiplicative Model, different from the prior two linear models, focuses on how the actual margin affects a player's learning rate. Previously, the learning rate was arbitrary, set, and adjusted by the users of the models to serve their objectives. In this model, however, the observed MOV determines the learning rate, marking a non-linear relationship between the winning result and the predicted win's residual difference and updated player rating. Under this system, the magnitude by which you win and lose a match will influence the gain or loss in your player rating. As for the Estimation Step, the Multiplicative Model has the exact same Estimation function as that of the standard Elo algorithm. It does not take margin into consideration since it is impossible to know the margin before the matches have started. This model also admits to the convergence condition because we can consider the margin variable in the Update function as an arbitrary K that is different in every match, essentially becoming the same Update function of the standard Elo algorithm. Hence, the convergence condition is automatically fulfilled.

The last model, the Logistic Model, is best used for a win result-focused prediction while also considering the MOV. The key to this model is the use of a generalized logistic function that takes a variable in and outputs a value between 0 and 1. The logistic function is as follows with an arbitrary base rate $\alpha > 1$:

$$L(x) = \frac{1}{1+\alpha^{-x}} \tag{3}$$

The Estimation Step for this model is simply inputting the relative difference of the two-player ratings into the logistic function to return a probability of the player winning. In the Update Step, the observed margin is also inputted into the logistic function to influence the updated player rating. Since the magnitude of MOV is now compressed into a value between 0 and 1, player ratings now is able to reflect the MOV but remain focused on the actual win results. This model also follows the convergence property because for every logistic system where the base rate $\alpha > 1$," the model is continuously differentiable and increasing in the rate difference, thereby satisfying the Lipschitz condition for convergence [2]."

After explaining how the 4 extension models of the standard Elo algorithm works and how each adheres to the convergence requirement, we can now move on to discuss testing and results.

The main experiment done in this paper is the training of the 5 models (standard Elo plus the 4 extensions) using past tennis tournament data and testing these models' accuracy in predicting win results for some out-of-sample tests. There are two metrics used to evaluate the models: accuracy percentage for estimating the correct outcome and log loss for estimating the incorrect outcome. After testing, Kovalchik found that the overall accuracy for all 4 extension models (68%) is greater than the original Elo system (66%). The experiment also considered multiple choices for the MOV (such as breakpoints won or sets won) and conducted tests on the different MOV options. When these options are averaged out, the overall accuracy for the extension models is 67%, still greater than that of the standard Elo algorithm. As for the log-loss metric, most of the models performed better (i.e. achieved a lower log-loss) than the standard Elo algorithm. Out of all the extension models, the Linear Model and the Logistic Model had a positive bias, overestimating player ranking differences by 7-8 points, while only the Joint Additive Model and the Multiplicative Model demonstrated negligible bias. Out of these two, the Join Additive Model exhibited the lowest variance in its predictions, making it the most consistent and general-purpose model out of the 4.

To sum up, this paper examines 4 possible extension models to the standard Elo algorithm and conducts experiments to evaluate the models' performances. It turns out that all the extension models outperform the standard Elo system in terms of accuracy in predicting win results while the Joint Additive Model excels out of the 4 as the most relevant model to the majority of applications.

4. Analysis of ELO Rating Scheme in MOBA Games

4.1. ELO's initial problems in MOBA games

In the MOBA game, due to the team-based nature of these games, the use of the ELO rating system has unique challenges. Unlike chess games are one vs one, MOBA games involve teamwork, so it's hard to accurately reflect an individual's contribution to the success or failure of the team. In [4], it discusses typical scenes in MOBA games. Players are distributed in different lines and roles. Each role has different contributions to the game. This complexity often leads to such a situation: a skilled player can affect the result of the game and cover the performance of other team members. For example, if a team has a large disparity in player strength, then a highly skilled player can dominate the game, creating an uneven experience. This situation can lead to player dissatisfaction. Through the record of a match, the skill of the players of one team is significantly higher than that of the other teams, thus highlighting the imbalance. This player can have an overwhelming advantage over the weaker players of the opposing team, resulting in an uneven match outcome despite the similar skill level of the overall team.

The ELO rating system was originally intended to ensure a balanced match but often resulted in an unfair experience. Players have often expressed frustration with the system because it tends to group players with different skill levels, resulting in tournament outcomes that rely more on luck than skill.

The system also penalizes highly skilled players more harshly when they lose to weaker players, further fueling discontent. When the ELO system is applied to MOBA games, it can lead to an uneven and frustrating experience for players. This results in players who perform well individually being penalized with a rating loss due to the team's overall performance. At the same time, the ELO system does not reward individual players for outstanding performance because it mainly considers the result of the match rather than individual match statistics. This can discourage players who have been playing well but are unable to win because of weaker teammates.

To address the shortcomings of the ELO system, a rating system based on players' performance in the game was proposed. The new system can incorporate matchup data into rating adjustments, more accurately rewarding players based on their performance in the game (such as kills, deaths, assists, and other relevant statistics), rather than just rewarding or punishing players based on matchup results. This rating system will provide a more balanced and fair assessment of the player's skill level and performance against the game. Make sure that players with outstanding team contributions receive adequate rewards, such as reducing or eliminating penalties when their team loses a game.

4.2. ELO's results under the new algorithm

In this paper, a simulation is conducted to compare the convergence of the ELO system and the scoring system based on player vs. game data. The results show that the system based on player match-up data converges faster and reflects the true level of players more accurately. Simulations showed that under the new system, players' ratings were more stable and representative of their individual performance over time. Compared to the ladder system, ELO ratings converge more quickly to reflect true player strength. After 100,000 games, the ELO score accurately represents the player's true game strength. However, the proposed rating scheme showed a faster rate of convergence, with players' ladder scores matching their true strength after only 10,000 matches. This shows that the new scheme is more accurate and efficient than traditional ELO systems. Compare Figure 1 and Figure 2; it's easy to find out that the new rating system showed faster convergence compared to the traditional ladder system. And it can reflect true skill levels after fewer games. Under the new system, players feel more accurately rewarded for their efforts, leading to higher satisfaction.

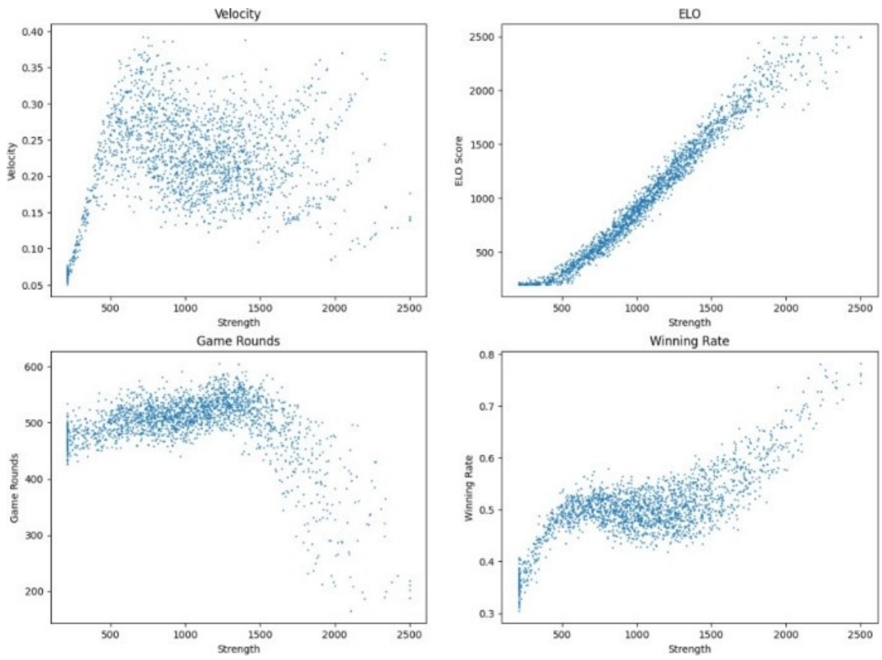

Figure 1. Experience under ELO Matching Model

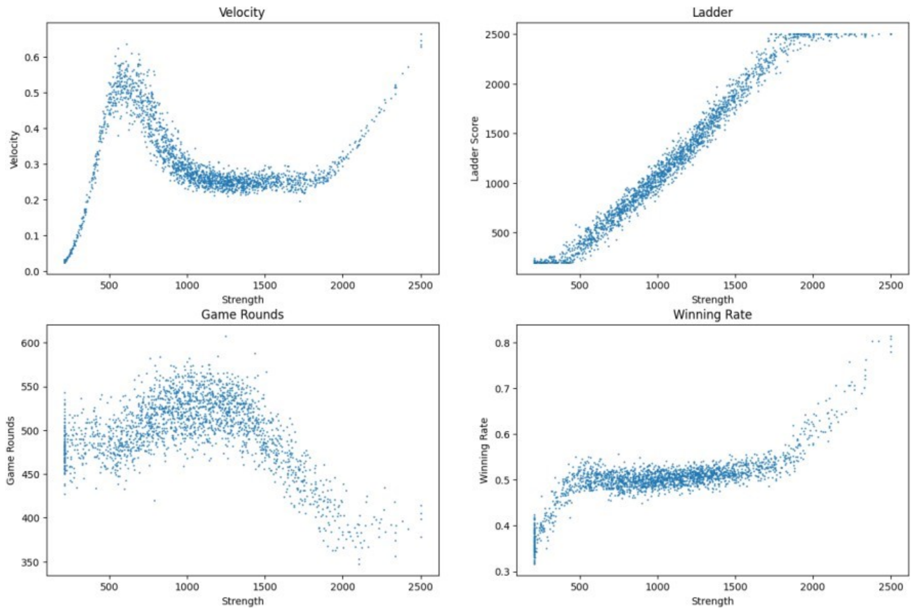

Figure 2. Experience under Proposed Matching Model.

In order to make the new rating system converge faster, author update the old equation(fig 3)to (fig 4). The different of these two equations is the K value is been deleted, and given a constant term. Using the new equation to simulate again. we can see the result between two equations. (Fig 5) is the result after 10000 matches by old equation and (fig 6) is after 10000 matches by the new one. It can be seen from the simulation results that the proposed matching and ranking scheme combined with the new formula converges faster than the ELO scheme. The position where a player's ladder score matches his real game strength. Therefore, the ranking scheme is more accurate and efficient than the ELO scheme.

While the ELO rating system is effective in one on one competitive games such as chess, its application in team based MOBA games presents significant challenges. It can effectively minimize the gap between a player's real skill and their ranking score but often inhibits the player's experience in MOBA games. The proposed new rating scheme more accurately rewards individual effort, shows better results in simulations, and provides a more satisfying player experience. Future work will focus on combining the proposed scheme with the ELO system at different stages of the ranking season to keep players motivated and ensure fairer competition. The goal is to find the best balance between accurate skill performance and an engaging player experience.

5. The Development and Application of Trueskill and Trueskill2

5.1. The TrueSkill Rating System

In order to match fair games for competitors, the TrueSkill system was developed. The TrueSkill rating system is a kind of Bayesian skills rating system. This system was designed to extend the function of the traditional Elo system, which was originally designed for Chess Games.

The TrueSkill rating system can be regarded as a generalisation of the former Elo system. It plays the role of providing matchmaking service, deciding who those players will play with. Compared to the Elo system, it can better track uncertainties in player skills [1]. With a principled Bayesian framework, the TrueSkill rating system can handle new difficulties that the traditional Elo system cannot. While the traditional Elo system focuses on solo games, the TrueSkill rating system is able to handle any number of competitors and various game modes, which satisfies the needs of modern multiplayer online games and derives individual skills from team outcomes.

So, what's the Bayesian formula? The Bayesian formula is a kind of probability formula, which was designed for calculating some specific probabilities under different kinds of conditions.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{4}$$

According to the formula shown above, if we know the details of a player's skills and performances, we can calculate the win rate accurately. In the TrueSkill rating system, we have recorded win rates in different conditions. With a Bayesian formula and large amounts of data recorded, the TrueSkill rating system can track the uncertainty about player skills well and better calculate the win rate of a player or a team.

Performance ratings and Elo scores are of great significance in very competitive games. Games' match-making systems analyze these ratings of performances to match a fair game. Performance ratings can also be a cornerstone of evaluating performances and fostering fair competition.

In newly-designed multiplayer online games, rating systems and match-making systems will face different kinds of challenges. It is difficult to distinguish personal performances from team performances. The TrueSkill rating system will record the data of both single players and the whole team. However, it focuses more on personal performance, so each player should do their best to reach the goal of the game, whether their team will win or lose. Then comes the problem that the games' outcome should not be just a winner or a loser. It should be a general evaluation of both the whole team and every single player. Players who perform well but lose the game should be comforted and encouraged in the TrueSkill rating system.

People always miss the point that match-making systems can increase players' engagement and generate players' interest. TrueSkill rating system can match players with similar abilities and interests, which makes the games balanced and enjoyable. Competing with players with similar strengths makes players more engaged.

Compared to the traditional Elo system, the TrueSkill rating system also demonstrates superior performances in predicting various aspects of predictive game datas, including the match quality, the win probabilities, and the convergence properties. The TrueSkill rating system effectively pairs players with comparable skill levels and reports skill estimates after each match. This function is achieved through an online learning framework employing a Gaussian density filter. The posterior distribution, which is approximated as Gaussian, serves as the prior for subsequent games, ensuring continuous refinement of skill assessments based on real-time performance data. However, when it comes to some specific game modes, such as "Small Teams", the accuracy of the TrueSkill system comes to be lower than the traditional Elo system. It may be because the TrueSkill rating system is not suitable for this kind of mode, where most games are Capture-the-Flag games.

5.2. The TrueSkill2 Rating System

In recent years, game studios have adopted a newer and more accurate version of the TrueSkill rating system. It is the TrueSkill2 rating system, which has been developed for making more accurate predictions as well as improving the fairness of games by improving the accuracy of the skill ratings in the match-making system. [3].

The TrueSkill2 is an expansion of the TrueSkill rating system, and it includes extra knowledge about players than the TrueSkill. For example, the TrueSkill2 system tests the membership of players in a team during the whole game, so each player should not tank in a game. The TrueSkill2 also records and predicts the tendency of a player to quit after joining into games. If a player quits in the middle of a game too many times, he will be punished or banned by the server. If the players play in a squad, their win rate will rise in some way, which is a concern for TrueSkill2. Besides, the past skill of players, the kill count, the death count, and the experiences of the player across other game modes are all included. The figure shown below tells the details of the working flow of TrueSkill2.

It should be announced that the TrueSkill2 rating system consists of various improvements to the original TrueSkill models. Compared with TrueSkill, the TrueSkill2 rating system leads to substantially more accurate skill ratings across a wide range of different quantities that are relevant to every game studio. The updated system is optimised to provide the skill value that accurately reflects a player´s performance and also respects past performance. With the improvements in accuracy in prediction, the fairness of matches will also be improved.

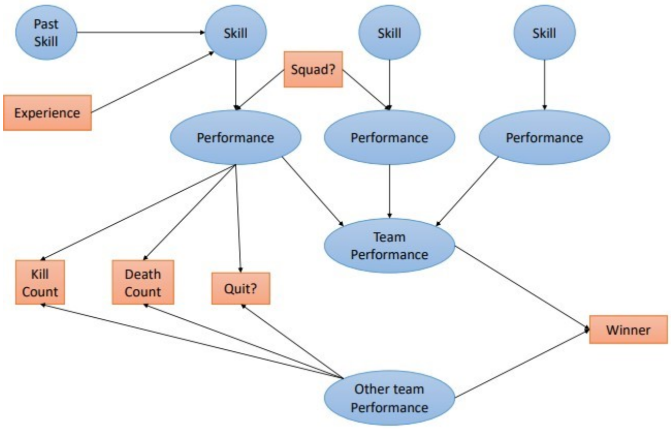

Figure 3. The generative model in TrueSkill2

There are two main modes in the TrueSkill2 rating system, helping it to make more accurate predictions. One of the two modes is a kind of online mode. It can update the ratings of the skills of each player in real-time to the server of the TrueSkill2 rating system. The other mode is a batch mode known as *TrueSkill Through Time*. The *TrueSkill Through Time* batch mode calculates the parameters and skills over the entire game time. Despite having different operational strategies for the Bayesian system, both two modes are grounded in the same probability model. Thus, the TrueSkill2 rating system is able to enhance match fairness by refining the accuracy of skill ratings integrated into the match-making service. This involves prioritizing making predictions and refining skill ratings with precise assumptions to improve accuracy.

Compared to its previous version, the TrueSkill2 rating system evaluates more aspects of performance and skills. TrueSkill. More attention is paid to enhancing the professionalism and reliability of the system in different modes of games. It considers scenarios such as solo play or squad participation while playing in a squad that will always raise the win rate. TrueSkill2 imposes penalties for players who quit mid-game and rewards those who complete matches. The player's experience, kill rate, previous outcome, and lapse will also be considered in TrueSkill2. Metric-driven modeling involves drawing inferences from data and models. The effectiveness of a metric improves with its ability to predict outcomes accurately. TrueSkill2 demonstrates superior accuracy in predicting win rates compared to the original TrueSkill metric.

For each online game, there will be thousands of beginners and new players of the game every day entering the server. Be- ginners and new players may have more difficulties matching a fair game. Aiming at ensuring those new players enjoy themselves the first time they play the game, game studios are supposed to make efforts to match fair matches for those beginners. The TrueSkill2 rating system does better in tracking the win rates of those beginners than the TrueSkill rating system, especially when those "new" players are not really new in some way. Players' performances in different modes of games are also positively correlated with their skills, meaning that we can estimate players' performance and win rate better through information from other different game modes. Balancing the win rate of beginners and new players can help them get used to the game modes faster. The TrueSkill2 rating system collects information on players' skills and performances from different modes and uses them to calculate the win rate of a game mode. In all these aspects, TrueSkill2 performs much better than TrueSkill.

DOI: 10.54254/2755-2721/132/2024.20850

5.3. Summary

From the TrueSkill rating system to its successor, the TrueSkill2 rating system, as well as those various innovations from the traditional Elo rating system, game studios, and developers are actively making joint efforts to enhance the game balance and improve the players' engagement. We are all trying to make sure that players, whether they win the game or lose it, will feel fair and enjoyable. By upgrading and refining these rating systems, we are going to create a more satisfying gaming environment.

6. Conclusion

In this literature review, we have traced the history of the mainstream matchmaking method, the Elo algorithm, exploring how it first originated as a simple 1 vs. 1 chess player rating equation and gradually developed to include more factors such as the Margin of Victory and extend to online multiplayer matchmaking. Since the appearance of the Elo algorithm, there have emerged numerous extensions of it to serve different purposes. In this paper, we first discussed 4 simple extension models that include the MOV. We then presented a proposed rating scheme for MOBA games and the TrueSkill model, which seeks to include additional in-game statistics. These models only provide a glimpse of some of the available matchmaking algorithms. Beyond these, there is still a widely available number of options. For instance, recently, there has been a rising interest in containing the fairness of the matchmaking method while also improving player engagement [5]. As long as there is a need for matchmaking, there will continue to be countless expansions of the Elo algorithm worthy for us to explore.

Acknowledgment

We would like to thank Professor William Nace for all the guidance and tips he has given us during the process. Without these assistances, it would be impossible for us to even get started. Next, we would like to thank our TAs. Diving into technical research is not an easy task. It is with their support that we managed to pull up an actual literature review paper, investigating in-depth the logic behind the algorithms.

Furthermore, Yue Liu, Jialong Dong, and Haiyang Jiao contributed equally to this work and should be considered co-first authors.

References

[1] Ralf Herbrich, Tom Minka, and Thore Graepel. "TrueSkill™: a Bayesian skill rating system". In: *Advances in neural information processing systems* 19 (2006).

[2] Stephanie Kovalchik. "Extension of the Elo rating system to margin of victory". In: *International Journal of Forecasting* 36.4 (2020), pp. 1329–1341.

[3] Tom Minka, Ryan Cleven, and Yordan Zaykov. "True skill 2: An improved Bayesian skill rating system". In: *Technical Report* (2018).

[4] Yuhan Song. "Analysis of ELO rating scheme in MOBA games". In: *arXiv preprint arXiv:2310.13719* (2023).

[5] Kai Wang et al. "EnMatch: Matchmaking for Better Player Engagement via Neural Combinatorial Optimization." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 8. 2024, pp. 9098–9106.