Modern Competitive Matchmaking Systems

Cyrus Sowdaey[#1]

[#]*Electrical, Computer & Energy Engineering, University of Colorado Boulder, Boulder, CO 80309*
United States of America

[1]cyrus.sowdaey@colorado.edu

***Abstract*— This document provides a brief investigation and exploration into modern Game Theory, relating to Matchmaking Systems, and specifically, the basic Elo Rating Algorithm. This document covers the computational complexity, and also provides quantitative and qualitative analysis of the Match and the algorithms involved in constructing an encompassing Matchmaking System that attempts to create the most fair competitive match between two teams, including teams whose size exceeds a 1-on-1 match, such as 5v5 player matches popular in games such as modern MOBAs.**

***Keywords*—**
MOBA - Multiplayer Online Battle Arena
Skill Rating - Estimated measure of an individual's performance relative to another individual.

I. Introduction to matchmaking systems

Firstly, it is important to understand the purpose and function of a Matchmaking System, and its applicability in modern cases.

A. How should opponents and teammates be chosen?

In many modern video games, both competitive and non-competitive, game developers and designers need some sort of way to have players be matched with, and against each other so that a match can take place. In an ideal world, one would want to have the odds of either team winning, to be exactly equal. You also need a way to keep track of a player's skill over time, based on the matches they have played, how they have performed, who they have played against, and other significant factors. Further, developers must also ask themselves how these two elements should interact with each other, such as how the probability of one team winning should affect how much their skill should change, depending on if they win or lose.

B. How much should a match count for?

For example, logically, if you were to assign a "skill rating" to a player, and if they were to win a game, their rating should go up. However, how do you decide how much it should go up by, and what factors should affect it? Certainly, if you took two players who had an equal skill rating, and they both played in two separate matches, one facing extremely difficult opponents, and the other facing very easy opponents, they should not gain the same amount of rating for beating vastly different opponents, as the player who faced very difficult opponents and managed to beat them should be rewarded greatly, compared to a someone who faced an easy opponent and was heavily favored to win.

C. What does "skill rating" mean?

This process is further complicated by the fact that the "skill rating" for players must be accurately tracked, updated, and maintained over periods of time, including innumerable outside factors that may also affect the "skill rating" a player may be assigned. This "skill rating" should attempt to most closely determine the "true skill" of a player, based on all the information that has been gathered.

II. What is considered a solution to the problem?

It is important to consider that these systems can only attempt to model and determine a player's "skill rating", as there is no way to objectively determine someone's skill in regards to these games.

Additionally, one should understand that skill is generally considered to be a relative number, based on how one compares to others, not to some specific standard or basis. After all, the entire purpose of a skill rating is so that one can understand how they would be expected to perform against someone else.

This is the core problem that a Matchmaking System seeks to solve. In general, an effective Matchmaking System will provide players with matches that are as fair as possible, in that neither team is distinctly advantaged over another, as well as that changes in players' skill are continually updated, based on prior results, in order to determine future matches. Meaning, the system is iterative fundamentally. Over time, this system should be able to evolve without outside interaction.

However, in practice, these systems may often require maintenance and changes, as one flaw in the system may be continuously compounded over time as the systems take past input in order to determine future output, where a flawed past input will cause further damage.

III. Scope of investigation

Considering all the moving parts that are involved within Matchmaking Systems (setting up matches, accurately assigning changes to ratings as a result of matches, maintaining an accurate ladder to rank players by skill), and how significantly they affect one another, the experimental part of this investigation will be limited to analyzing and understanding specifically how the Elo Rating algorithm may estimate the probability of a player winning in a specific game, as well as how leaderboards may be constructed to rank players. We can also look into how much a player's skill rating should change as a result of that match.

However, there are numerous other components in Matchmaking Systems that would be fascinating to study, such as determining how teams should be set up for matches given a set of players, instead of manually picking teams, as well as more complicated approaches to considering factors that may affect match outcomes and player skill. As an example, in team based games players contribute a different amount of work toward winning or losing, compared to their teammates. In that way, for an effective matchmaking system these factors are key.

In a more thorough investigation, it would be desirable to implement a complete matchmaking system that can consider all these external factors, as well as create a system that can completely handle all aspects of matchmaking, between player rating, setting up fair matches, and creating an appropriate skill distribution in a large population of players. Analyzing these algorithms and the interactions between the components in these systems opens the gate for very complicated ideas.

IV. Understanding the Elo Rating Algorithm

The *Elo Rating System,* was a rating system designed in 1960 by Arpad Elo, generally known as the father of modern matchmaking systems. This system was designed for Chess, one of the oldest competitive games, whose need for a way to accurately model a player's skill became important as its popularity increased as a game and Chess Tournaments became a significant spectacle, and previous methods to rank and match Chess Players simply became ineffective.

However, the Elo System has many limitations due to its simplicity and the fact that chess is a two player game, while many video games involve teams of individual players who may only play together for one match, then be matched with new teammates to play with in the future. In video games, one needs an effective way to evaluate a team's rating, and an individual player's rating, then update the individual's rating based on results of a team match.

A. Elo's Design - Core function of Matchmaking Systems

Arpad's system was designed in such a way that by modeling the probability a player wins a match, based on their individual skill ratings, you could determine appropriately by how much each player should gain, or lose rating, based on the outcome of the match.

$$P(c_1 > c_2 | r_1, r_2) = \Phi\left(\frac{c_1 - c_2}{\sqrt{2}\beta}\right)$$

$$\text{where } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Above, the probability of an individual with a performance, c_1, beating an opponent's performance, c_2, whose skill ratings are given as inputs, r_1 and r_2 is modeled in Elo's original system. β^2 is used to represent a fixed variance in these players' skill.

By applying the cumulative distribution function of the standard normal distribution, one can then attempt to update the players' ratings, r_1 and r_2.

$$\Delta r = \alpha\beta\sqrt{\pi}\left(\frac{w+1}{2} - \Phi\left(\frac{c_1 - c_2}{\sqrt{2}\beta}\right)\right)$$

Here, the change in a player's rating as a result of the match can then be determined, where w represents the actual outcome of a match, $w = 1$ being p_1 as the winner, and $w = -1$ for p_2 as the winner. α is used to represent the weighting of the results compared to the estimate of the match outcome. Specifically, this is a constant that can be tuned as a result of experimentation and observation. Having a higher value for α correlates to more volatile changes in skill in the system both upward and downward.

As such, the change in players skill as a result of the match outcome is then

$$r_1 = r_1 + w\Delta \text{ and } r_2 = r_2 - w\Delta$$

Knowing all this information, the Elo system originally was based on a Normal Distribution in order to estimate the odds that a player will beat another player, which would then apply those odds in conjunction with the actual outcome of the match in order to come up with the amount of rating that should change, based on the outcome of the match and who was favored.

By doing this, one can approach a hallmark problem of matchmaking, where certain matches should matter more or less based on how you are favored to win or lose.

B. Elo System in consideration to Matched Games

In reference to the introduction, this helps to consider that a player who is perhaps twice as good as another, and thus heavily weighted to win , should barely gain any rating if they were to win, since it is completely expected. On the other hand, if they were to lose, they would lose a significant amount of rating because of how heavily stacked the odds are, against the worse player.

C. Normal Distribution in Elo's Design

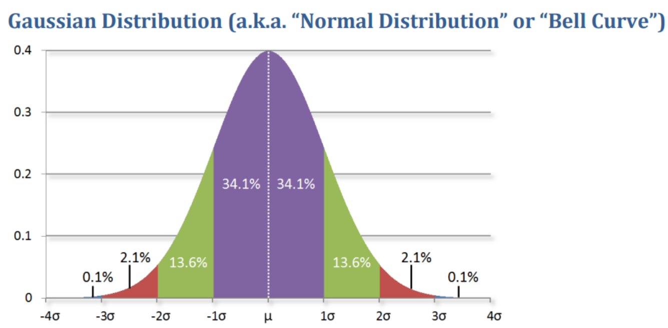
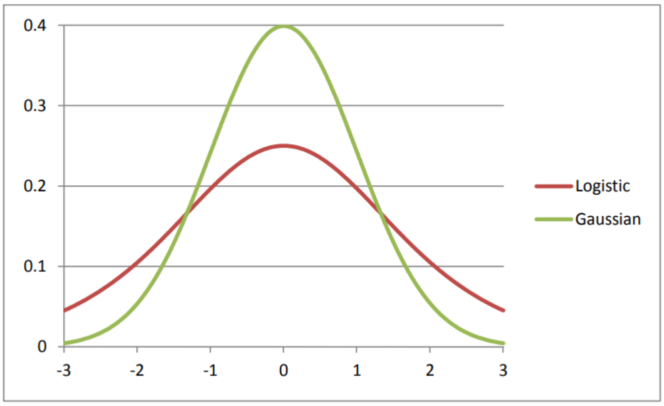


Fig. 1. Original distribution for player skill and chance to win in Arpad Elo's design

Shown above, is the fairly well known Normal Distribution, also called a bell curve for its bell-shaped nature. This how skill is thought to be distributed in the original system, with $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ being the cumulative distribution function. One major discovery, however, is that over time, data scientists have found that a logistic curve actually tends to better encapsulate distribution of skill, especially in chess specifically. As such, that method has taken prevalence in more modern matchmaking systems and implementations of the Elo Rating Algorithm. This investigation will apply the logistic version.



Comparing a logistic curve to the Normal Distribution, it looks significantly less extreme at both edges, and overall the majority of players are more evenly distributed in the range. This translates to the Algorithm being more forgiving with relation to how it estimates the probability that a player will beat another, depending on their skill difference. Realistically, a player within two standard deviations of skill is much more likely to beat another player than the Gaussian curve would predict.

V. Modern Approaches and Improvements to Matchmaking

The TrueSkill system is a modern matchmaking system, developed by Microsoft in 2005 to address a need for matchmaking in online competitive games hosted through the Xbox Live service. The TrueSkill system is fundamentally the same idea as the much simpler Elo System, however, it makes a significant amount of improvements and adjustments, many of which are beyond the scope of this project, but are very fascinating. Some of these concepts, such as Bayesian Inferencing are discussed briefly to help provide contextual understanding, but are not the mathematical focus of this project.

A. Belief and Certainty in calculated skill

One of the more significant adjustments that the TrueSkill system seeks to address is representing uncertainty in a player's skill. Logically, a new player who has only played a few matches has a much less certain rating on their skill than someone who has played multiple hundred matches. Yet, we still need to be able to match both players with each other, against each other, and update their skill rating as accurately as possible.

In order to do this, the TrueSkill system employs Bayesian Inference, with two significant elements. σ is measured as the "belief" or confidence in a rating, and μ is the mean skill of a player. By using previous evidence to update this belief, one can more properly model someone's skill in a game over time, and help to immediately provide a good experience to new players.

B. Handling Teams of Players

Additionally, there were modifications made in order to consider the fact that instead of having two individuals face each other, you may now have teams of multiple individuals with their own rating, facing other teams of multiple individuals. In general, one should also understand that some players on a team may do a greater job in winning the game than others. By adding more variables to the algorithm, such as evaluating tasks that a player may do and their individual performance, the algorithm can more effectively estimate skill, as well as more properly match players together.

VI. Using a modified Elo Rating Algorithm in world championships

A. *Beginning with a game*

League of Legends is known to be one of the most popular video games currently. Specifically, it is a 5v5 MOBA, where victory is determined by the first team to destroy the opposing team's base (referred to as the Nexus).

League of Legends uses its own matchmaking system, for players, which is not publicly available.However, at its core, just like it is an adaptation of the Elo System's fundamental ideas.

B. *Determining a population*

In League of Legends, players are matched with random teammates for ranked matches, which tends to complicate Matchmaking Systems, as you must now factor in a team's combined rating as well as individual ratings. One must also realize the performance of individual players and how much that should affect their individual rating changes.

In this investigation, we will evaluate the most recent World Championship in League of Legends, as organized teams stay together as the same players for the entire tournament and we can consider the overall performance of each team as if it were an individual player, with a single skill rating.

C. *Understanding the Purpose of the experiment*

The goal with this experiment is to apply Elo's algorithm to the 2021 World Championship, consisting of the best teams from multiple regions around the world. By applying the algorithm, we can construct a leaderboard system for the teams in the tournament, as well as evaluate how their performance against different teams would change their measured skill rating as a team.

Since the World Championship has already concluded and all the match standings and results of matches are known, we can sequentially run all the matches that occurred through the Elo Algorithm, and then construct an overall ranking system based on the performance of individual teams.

With more time and data, it would be ideal to track each team's performance throughout the competitive season, which begins early in the spring and continues through the year up until the World Championship in mid October. By having that additional data, teams could have a skill rating already calibrated to an extent, going into the Championship.

D. *Setting up the matchmaking system*

In this case, we will begin the tournament with all teams starting at an equal skill rating. The number in this case is somewhat arbitrary, and depends more on its value relative to the certain constants involved in the algorithm more than anything else. Chess, by convention, uses 1500.

After setting the initial skill of the 22 teams that played in the world championship, we can import the data through the entire tournament into an excel spreadsheet. 100 Matches were tracked over the course of the tournament. This includes matches that were best of 1, as well as best of 5 matches. If a team won 3 matches and lost 2 in a best of 5, each individual match resulted in a change in Elo Rating, instead of only the winner of the entire best of 5 winning one match.

E. *Considerations about implementation*

Wins and Losses by each team were tracked throughout running the algorithm through all 100 matches. It is important to note that because every team started with the same skill rating at the beginning of the tournament, despite some teams being heavily favored and considered far better than other teams from the beginning.

In that way, a team known to be bad, beating a team known to be good early in the tournament would not have as much of an affect on rating, compared to later in the tournament when skill ratings for teams are more cemented. This is partly why having a long history of data for teams throughout a season could provide better insight for the end of year tournament.

F. *Tournament Format*

With the format of the tournament, even though 100 matches were played, the amount of games played by each team varied heavily based on if they moved on to further stages.

The championship itself began with a Play-In round, with the rules as following, taken from Liquipedia's Worlds 2021 Page:

- Ten teams are divided into two groups where they play a Single Round Robin format.
- All matches are played in a Bo1.
- Top teams in each group advance to Group Stage.
- 2nd, 3rd and 4th teams in each group advance to Play-In: Round 2.
- Bottom team in each group is eliminated.

The second Play-In round followed,

- Six teams from Play-In: Round 1.
- The 3rd and 4th placed teams of the same group of Round 1 face each other, winners compete against the 2nd placed team of the other group.
- All matches are played in a Bo5.
- Two winners will advance to the Group Stage.
- Four losers will be eliminated.

After the second Play-In Round, the Group Stage began.

- Four teams from the Play-In Stage join twelve teams with direct entry from China, South Korea, Europe, Southeast Asia and North America.
- All sixteen teams are divided into four groups where they play a Double Round Robin format.
- All matches are played in a Bo1.
- Top two teams in each group advance to the Playoffs.
- Bottom two teams in each group are eliminated.

When the amount of Teams remaining was narrowed to 8 teams, the final stage began, involving purely best of 5 matches with single elimination.

- Eight teams play in a single elimination bracket where the 1st placed team of each group faces the 2nd placed team from another group.
- All matches are Bo5.

G. *Running the Algorithm and Code on 100 Matches*

By extracting data from an xlsx spreadsheet, we can gather the order that matches occurred in, the two teams that played against each other, as well as which team won. This was then kept track of in a list of classes, where each team had their name, rating, wins, and losses stored.

The code begins at the first row and first column of the spreadsheet, which is then iterated until the last column is reached. The first column contains the first team, the second column contains if that team won, and the third and fourth columns contain if the second team won, and which team they were. The third column was not needed, as teams cannot tie a match and there can only be one winner, so if the first team did not win, the second certainly did.

After reaching the end of the first column, the Elo Rating Algorithm was applied to the match, which updated ratings for the teams depending on which team won. Then, the next row was iterated and it began again at the first column in that row to determine the first team to play in the following match.

After all matches were run and the final ratings for every team were completed, the list was re-sorted by ascending rating, then printed.

H. *Analyzing the Skill Ratings with consideration to the results of the tournament*

The narrative throughout the tournament, as well as the skill ratings determined by the algorithm align fairly closely. However, the team that won the tournament was not actually the team that was rated to be best by the rating system, nor were they thought to be the best by expectations of analysts or the overall community going into the tournament. In fact, they were considered quite weak and the skill rating algorithm also rated them as the third best team, despite being the eventual winners of the tournament.

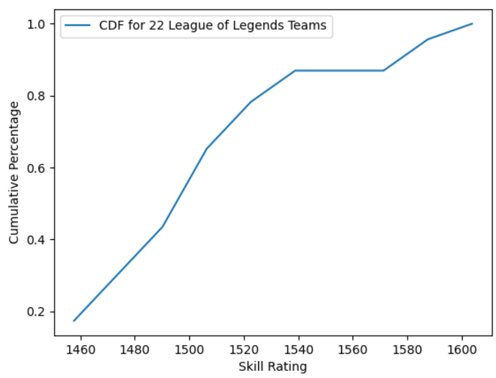
```
Total Matches Analyzed:  100
DWG KIA 1603.74 W: 14 L: 5
T1 1585.55 W: 10 L: 4
EDG 1579.83 W: 12 L: 9
RNG 1530.27 W: 7 L: 4
LNG 1527.13 W: 7 L: 4
Gen.G 1518.01 W: 9 L: 7
Hanwha 1512.54 W: 7 L: 6
100 Thieves 1506.51 W: 3 L: 3
MAD Lions 1505.05 W: 5 L: 6
PEACE 1501.25 W: 2 L: 2
PSG Talon 1500.35 W: 3 L: 3
MEAN RATING 1500.00 W: 0 L: 0
Galatasaray 1497.93 W: 2 L: 2
Team Liquid 1489.12 W: 3 L: 4
Unicorns of Love 1474.33 W: 1 L: 3
Rogue 1474.21 W: 2 L: 4
Beyond 1472.14 W: 1 L: 3
RED Canids 1469.41 W: 1 L: 3
FunPlus Phoenix 1468.91 W: 2 L: 4
CloudNine 1455.49 W: 5 L: 8
Fnatic 1447.42 W: 1 L: 5
DetonatioN FocusMe 1443.87 W: 3 L: 7
Infinity 1441.46 W: 0 L: 4
```

Fig. 2. Formatted output by running Python Code, displaying in descending order of rated skill. Team names, Skill Rating, Wins, and Losses are shown

Damwon Gaming Kia, denoted by DWG KIA, was considered to be the best team going into the tournament, and up until the final set of matches they also performed the best, being essentially undefeated going into the finals. However, they lost the final best of 5 series to Edward Gaming, or EDG in a narrow 3-2 ending. When we look at the results of running the code, DWG Kia ended the tournament 14-5, 3 of those losses being the final series. Additionally, their Elo rating was at 1603.74 by the end of the tournament, while EDG was at 1579.83 with a W-L record of 12-9. Between these two finalists was the semifinalist, T1, also thought to be one of the best teams in the world. Rated at 1585.55 with a record of 10W-4L, three of those coming from the semi-final best of five against DWG, who would go on to the finals.

It is interesting to note that these ratings do correlate to the expectation throughout the tournament, and measure fairly well the "skill" of each team. Keeping in mind that the Elo rating system does not indicate that a team with higher rating WILL win, simply that their expectation to win is higher.

One can see, from this cumulative distribution of the teams and their calculated ratings, at the end of the tournament, about 60% of teams remain at or below the initial rating of 1500, and only about 10% of teams rank above 1540 skill.



Contextually, These teams are DWG KIA, T1, and EDG, all of whom were rated much higher than other teams due to their overall performance throughout the tournament. Beyond those three teams, ratings were much closer for teams that did not perform as well.

On the opposite end of the spectrum, the worst performing teams had a skill rating that decreased from the initial rating of 1500 down to, at minimum, 1441.46. Infinity Esports, INF, were knocked out early in the tournament and had a record of 0 wins, 4 losses. Between this range of 1441.46 and 1603.74, the teams were distributed in accordance with their performance and the teams they faced and games they played, as the system intends to calculate.

One should keep in mind that since some teams did not play many games, such as Galatasaray, their skill rating may not be very accurate, especially since their games also all occurred in the beginning of the tournament, where the Elo Algorithm could not yet accurately determine team placements and ratings. In this case, they had a record of 2W-2L, ending with a rating of 1497.9, nearly where the team started.

I. *Time Complexity of the Algorithm and Constructing a Leaderboard*

The time complexity for the basic Elo Rating Algorithm is constant,O(1), as the algorithm takes only a few inputs and performs the same amount of operations regardless of the skill ratings involved, and teams. Specifically, it takes the two ratings as input which are then converted to probabilities for each team to win, that are then converted into an adjustment in rating based on the third input of which team won the match.

In constructing the leaderboard, one needs to compute sequentially all the matches that occurred, since a team playing one match will have a change in rating before playing their next match against a different opponent, for example. The time taken to complete calculations for all matches is thus dependent on the total amount of matches. In that way, it is linearly increasing, where one match will require one operation and one hundred matches require one hundred operations. Thus, the total complexity of creating a leaderboard is of complexity O(n).

J. *Analysis of Logarithmic Elo Rating Algorithm*

Initially, the Elo Rating Algorithm began with using a Normal Distribution in order to rank players and determine how likely players are to beat one another, depending on their skill. However, over time it has been found that a Logistic Curve better fits skill distribution and likelihood of winner. Specifically, a Normal Distribution is too extreme in determining whether a player will win or lose against another, as it over-estimates the chance that a player of lower skill will be unable to beat a higher skill player. If a Normal Distribution were applied to the same results and matches that were analyzed for this tournament, one would see the estimated probability that EDG would win would be significantly lower compared to using a logistic curve.

The formula for determining the expected probability for two teams is implemented as follows:

$$P(t_1 > t_2 | r_1, r_2) = \frac{1}{1+10^{(r_1-r_2)/400}}$$

The probability that team two, t_2 wins is simply

$$1 - P(t_1 > t_2 | r_1, r_2)$$

This probability, is then used to adjust rating when results for a match are determined,

$$r_1 = r_1 + k(w - P(t_1 > t_2 | r_1, r_2))$$

$$r_2 = r_2 + k(w' - P(t_2 > t_1 | r_1, r_2))$$

w and w' is used to represent the outcome of the match, where the winner will have w=1 and the loser will have w' =0. The value for k represents the K-factor, a constant that is set in accordance with the desired volatility of the algorithm. A higher K-factor relative to the skill ratings r_1, r_2 will result in skill rating changing in larger quantities as a result of matches being played. In some cases, this is desirable especially when player skill is not known exactly. However, in cases where skill has been closely determined, such as a player playing many matches in a time period, a high K-factor will change the perceived skill of the player far too easily depending on the most recent results that are analyzed. The value, 400, in the formula for calculating probability that a player will beat another player, contextually means that if two players have a skill difference at or beyond 400 points of skill (with the arbitrary mean set at 1500), the likelihood of the worse opponent winning is ten times lower. In this case, 400 points is about 26% of the mean score.

Adjusting the K-factor, the value of 400, and changing the mean rating of 1500 to start players at, will all affect each other and the results that the algorithm will determine. K=32, 400, and a mean rating of 1500 were used in this investigation for simplicity and as a reference for a good starting point, based on the implementation in Chess.

K. *Flaws of implementation*

Some flaws and weaknesses of the implementation of this algorithm into the case of the League of Legends World Championship have already been discussed, however, there is still more to mention still.

Matchmaking Systems are concerned with attempting to model relationships between competitive matches and players. As such, they cannot objectively determine skill or who will win. However, they are meant to determine as closely as possible the true skill of players and winning or losing.

Because of this, their applicability and preciseness is heavily variant on what population and case they are being applied to. In League of Legends, modeling player and team skill often has a lot more to do with factors beyond winning or losing, such as individual performance and what actually happens during the course of the game.

Additionally, unlike Chess, League of Legends involves two teams of five players picking “Champions”, or characters that have uniquely different abilities. In some cases, certain characters may counteract other characters and allow a team who is worse in terms of skill, to beat an opponent of greater skill when they should not have. Some argue, however, that this component is certainly a different aspect of skill, in choosing the best character for a situation.

L. *Further Investigation*

For further research, expanding this project to tracking a competitive season’s worth of matches would be a great way to practice and become more familiar with concepts discussed in this paper. Additionally, implementing additional components to the skill rating algorithm would provide a chance to further refine the accuracy as well as develop understanding of the significant amount of labor that goes into developing Matchmaking Systems.

I. Conclusions

Matchmaking Systems are an extremely valuable tool and are continually evolving to fit the needs of new games. As technology and research continues, these systems are becoming significantly more complicated and nuanced, and machine learning applied to matchmaking is also becoming more common. As these systems evolve, they more closely accomplish the purpose of providing accurate information regarding skill and components that determine skill of players in games.

Being able to design an efficient algorithm that is capable of accurately accomplishing all the components needed in a matchmaking system is a challenging task that requires a significant amount of time investment and trial and error, but the satisfaction gained from seeing it work and perform in a real world setting would be immeasurable. It’s very interesting to see mathematical concepts and ideas applied in a relevant way to something that I am personally invested in.

Even just applying one of the simplest rating algorithms to the League of Legends World Championship was able to provide contextual insight on the results of the tournament and allows one to support the qualitative evaluation that many had during the tournament, seeing the matches, by providing a mathematical and solidly quantitative rating to teams. It is important to keep in mind that both of these components are important, as the numbers do not always tell the full story, neither does qualitative analysis. However, the combination of these is an extremely valuable tool.

References

[1] “2021 World Championship.” *Liquipedia League of Legends Wiki*, Team Liquid, https://liquipedia.net/leagueoflegends/World_Championship/2021

[2] Heath, Jerome. “Worlds 2021: Overall Scores, Standings, and Results.” *Dot Esports*, Dot Esports, 6 Nov. 2021,

https://dotesports.com/league-of-legends/news/worlds-2021-overall-scores-standings-results

[3] Lee, Heungsub. "Trueskill." *TrueSkill*, https://trueskill.org/#what-s-trueskill

[4] Minka, Tom, et al. "TrueSkill™ Ranking System." *Microsoft Research*, 15 Nov. 2021, https://www.microsoft.com/en-us/research/project/trueskill-ranking-system/#:~:text=The%20TrueSkill%20ranking%20system%20assumes,is%20determined%20by%20their%20performance

[5] Moser, Jeff. "Computing Your Skill." *Computing Your Skill*, 8 Mar. 2010, http://www.moserware.com/2010/03/computing-your-skill.html

[6] "Ryan Madden." *Adapting Elo for Relative Player Ranking in Team-Based Games*, 18 Apr. 2017, https://ryanmadden.net/posts/Adapting-Elo