Investigation of rating systems in competitive eSports

Untersuchung von Bewertungssystememem in eSports
Bachelor-Thesis von Sebastian Sztwiertnia
Tag der Einreichung:

1. Gutachten: Johannes Fürnkranz
2. Gutachten: Tobias Joppen

TECHNISCHE
UNIVERSITÄT
DARMSTADT

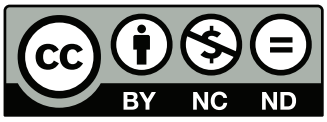
Fachbereich Informatik
Knowledge Engineering Group

Investigation of rating systems in competitive eSports
Untersuchung von Bewertungssystemem in eSports

Vorgelegte Bachelor-Thesis von Sebastian Sztwiertnia

1. Gutachten: Johannes Fürnkranz
2. Gutachten: Tobias Joppen

Tag der Einreichung:



Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 und § 23 Abs. 7 APB TU Darmstadt

Hiermit versichere ich, Sebastian Sztwiertnia, die vorliegende Bachelor-Thesis gemäß § 22 Abs. 7 APB der TU Darmstadt ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Falle eines Plagiats (§38 Abs.2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung gemäß § 23 Abs. 7 APB überein.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, den October 31, 2018

(Sebastian Sztwiertnia)

Contents

1 Introduction

This section gives a brief overview about the motivation for the thesis and outline the goals and general procedure. The scientific context and related works to this work are presented accordingly.

1.1 Motivation

The field of competitive Online Multiplayer Games grows every year in popularity among the professional players, casual players and viewers. This new rising sector attracts more commercial partners and increasingly creates more research opportunities in different scientific fields. Besides sociological and economical research, eSports offers a great opportunity for data oriented research approaches. The digital nature of a competitive eSports discipline such as Dota 2 or Counter-Strike creates finely granular and detailed data sets available for analysis. Detailed information in form of server state recordings such as replays or community driven statistic web pages are unique feature of this area.
A key point in every competitive discipline is the idea determining a participants skill compared to other participants. Usually in sports, world rankings based on won tournaments and leagues are created to make a statement about a teams or player's skill level. Qualifiers and ranking placements are used to create new competitions or tournaments like the UEFA Champions League or a World Cup. Competitive Multiplayer Games on the other hand face the challenge of matching players against each other in a fast and fair manner every minute.
Matches take less time to finish and the player come and go frequently, which makes the task of determining a players skill difficult. This thesis will investigate two widely used skill approximation systems for online match making, and one rating system specialized for sport disciplines with Offense-Defense mechanics. Motivated by the amount of data online available, those rating systems will be applied to the competitive eSports discipline of Counter-Strike Global Offensive. Those three rating systems will be trained with past competitive matches to predict the outcome of upcoming competitive matches. The rating system's characteristics and performance in terms of win prediction will be analyzed and evaluated.

1.2 Goal of this Thesis

The Goal of this thesis is to investigate the area of competitive eSports as an area of application for match outcome prediction. Competitive Counter-Strike Global Offensive is selected and presented as a suitable competitive discipline in competitive eSports. Parallels to conventional sport disciplines like football will be investigated and established as a reason to continue further evaluation. Two prominent rating systems, Elo and Microsoft TrueSkill will be explained and applied to a database of competitive match results with the goal of approximating team skills.
Further, a offense-defense based rating system presented in [7] is selected and compared to the previous two rating systems. This offense-defense based rating has been shown to perform well in the discipline of american football, which raises the question of its performance in a new discipline of competitive domain. The underlying methods of the three rating systems are explained and their implementation as a outcome predicting method discussed. Training and evaluating approaches of the respective rating systems are introduced, as well as requirements and acquisition of relevant data. Based on different training and evaluation data sets the win prediction performance of each rating system is evaluated and compared to each other.

1.3 Related Works

Nicholas Kinkade and Kyung yul Kevin Lim present two win predictors for the online game DotA 2 [6]. DotA 2 is very similar to Counter-Strike Global Offensive concerning the competitive scene and data available. One win predictor is based on post game data collected from replays of competitive matches and is used as a baseline for win prediction. Using logistic regression and a random forest classifier they achieved a prediction of 99.8% based on three features extracted from concluded matches. This shows a direct correlation between features during a game and its final outcome. Further a second predictor based only on pre match information, hero picks in this case, is presented achieving a win prediction of 73%. The predictor is based on 4 features derived from the heroes both teams picked.
Yang et al. predict the winning team in DotA 2 using information available prior to a match and information available at every minute during a match [12]. An approach with logistic regression and a Attribute Sequence Model is used to improve the win prediction accuracy before the match from 58.69% to 71.49%. Further a live predictor is presented achieving 93.7% accuracy at the 40th minute mark of a match. The predictors are based on a data set with 78.362 matches where 20.631 contained replay data.
A Elo rating system is used by Hvattum and Arntzen to predict the match results in association football [5]. The focus of the work lies on predicting the best betting opportunities in football. The Elo rating system is found to be useful in encoding past football results and has been compared to six benchmark prediction methods. Hvattum et al. concluded, that Elo performs significantly worse than two methods based on market odds, but better then the remaining benchmark methods.

The presented works indicate research interest in the field of competitive online games and offer points of reference for the approach of this thesis. The similarities between DotA2 and Counter-Strike Global Offensive concerning available data and the research done by Havattum et al. using the Elo rating system for match prediction are promising starting points for this work.

2 Background

This section's goal is to give a brief overview of the domain of online multiplayer games in general and use it to introduce the emerging competitive subset of online games in specific. It outlines the similarities between the new domain of competitive games, also known as eSports, and the conventional, well established domain of competitive sports. The specific online multiplayer game Counter-Strike:Global Offensive and its competitive scene is introduced as the application domain of this thesis. Secondly, this section provides a general overview and understanding of prominent rating algorithms used in traditional competitive domains like sports and offline games. The motivation behind developing and deploying rating algorithms in a competitive setting is outlined, as well as their position in today's online gaming market. Based on the set of introduced rating algorithms and a new competitive domain, lastly the topic of match outcome prediction is presented.

2.1 Competitive Multiplayer Online Games

The idea of developing and playing games on a private computer has its roots in the very early days of computer technology. The first games were mechanically and visually basic and were mostly inspired by real physical games, e.g. pen and paper games. One prominent example of an initial computer game is Bertie the Brain from the 1950s. The game was based on the principle of Tic Tac Toe, was interacted with using punch cards and visualized by using miniature vacuum tubes. Computer games, as every computer based application, depend strongly on technological advancements in the field of computer hardware and operating software.

Coming from a simple Tic Tac Toe arcade game, computer games have become a widely popular entertainment medium and economical market. A market study [1] conducted by the entertainment software association for the US market in 2017 concluded, that 67% of US households own a device that is used to play video games. Further, 53% of the most frequent gamers play multiplayer games. Multiplayer Games are a genre of games, where the main game mechanics evolve around playing against or with another person. Playing a multiplayer game can be done offline on the same device, e.g. a gaming console, online with friends or even with random people from around the world. Persons playing multiplayer games spend an average of 6 hours a week playing games mainly with friends and family. 29% of played multiplayer games belong to the shooter genre, followed by 28% of casual games and 27% of action games. The study outlines the impact of computer games on the modern economy and in the case of multiplayer on modern social interactions as well.

Creating a competitive environment within a game and around playing the game can be a central aspect in designing and improving online multiplayer games. It enables and encourages players to improve their ability of playing the game by playing against players with a similar skill level. This environment results in players investing more time in playing and understanding the game by having a competitive learning curve. Similar to conventional sports like football, basketball or table tennis, online multiplayer games are increasingly able to provide mechanics to create a solid competitive environment. An academic project of the Worcester Polytechnic Institute [9] concerning the design and community building of competitive gaming outlines five qualities of competitive gaming:

Game Depth The game must provide a certain depth, which comes from the interactions between entities and mechanics in the game. In order to achieve that, a game does not have to necessarily provide great design complexity. Building strategies is therefore based on a relatively simple rule set, as well as the ability of countering certain strategies.

Skill A player's skill should be the only factor affecting the outcome of a match. Therefore it is necessary to minimize the influence of luck and randomness in the game design. In general, a more skillful player should always win more often than a less skilled player.

Evolving The game design should not allow the possibility of discovering and mastering one certain strategy which has no counter strategy by design. Strategies can therefore not be organized in a hierarchical manner in terms of strength. A cyclic and ever evolving approach in this regard is desired. That way, every strategy has a set of strategies it counters and a set of strategies it gets counter by.

Lenticular Design The value and role of certain game mechanics and entities change based on the skill level of the player. This way a mechanic provided by the game can be used in different ways in the context of strategies, depending on the players understanding of the game.

Fun Even when a game creates all the competitive aspects mentioned above, it will not attract a lot of player engagement, if it simply is not fun. The definition of fun in a sport or game is very broad and strongly based on personal traits and experiences and even cultural influences. Nevertheless, a competitive oriented game should promote a fun design and characteristics, so experienced players keep enjoying to play the game and new players are attracted to the game.

By following the list of those five competitive qualities, many parallels to aspects of already established competitive sports can be found. The aspect of game depth exists in a lot of existing sport disciplines. The sport discipline soccer's depth, for example, is relatively simple, as it has a simple and easy understandable set of rules. Players are not allowed to touch the ball with their hands and arms in a defined play area and score a point by kicking the ball into the enemies goal. The depth of strategies evolves through the variation of positioning and movement of the players and interaction with the ball. Further, soccer has no real random component purposely involved in the outcome of a match. It remains a team skill based sport, which on the other side is evolving constantly in it sets of strategies, as new players and technology influence a team's strength and weaknesses. Different kicking techniques present, based on the players or even teams skill, different opportunities in creating and executing a certain strategy. A less skilled player's shot accuracy does not allow him to integrate long range passes in his play, where a skilled player can use his high accuracy for his team to build a strategy around it. Besides the lenticular design aspects of soccer, the requirements of starting to play or watch soccer games is low, which makes it an popular and enjoyable competitive discipline.

The major parallels in characteristics between competitive online games and traditional sports competition open the opportunity to apply and extend processes from the traditional domain of sports to a new domain, the competitive online multiplayer games. The scientific relevance of this domain, also called electronic sports, or eSports, increases steadily as more concepts of traditional sport and training science are being applied [11]. Electronic sports introduces, due to its electronic native platform and global distribution, new fields and opportunities for research [11]. Besides the new research opportunities created by eSports, eSports itself as a cultural phenomenon became more and more a focus of current research [3].

2.2 The eSport Discipline Counter-Strike

Counter-Strike is a first person shooter game, that has been released in 1999 as a modification of the game Half-Life by Valve Corporation. This modification, developed and distributed by two Half-Life players was built upon the Half-Life Multiplayer game mode and introduced new strategic game elements, visuals and play fields, also called maps. The game's most prominent feature was the objective-based game play with simple win conditions and easy to understand mechanics. Two opposing teams of players, labeled in the game as Counter Terrorists (CT) and Terrorists (T), play simultaneously and compete against each other to complete predefined objectives. Those game objectives are defined in a timely and local manner, such as securing and defending a specific location on the map or interacting with a game object within a certain time. Depending on the pre selected game mode and the team a player is playing on, the definition of game objectives differ and win conditions change.
Motivated by the great success of the Counter-Strike modification, the developer studio of Half-Life, Valve Corporation, acquired in 1999 the rights to further develop Counter-Strike as a standalone game. Through multiple development iterations the first stable and mature Counter-Strike "Counter-Strike 1.6" was released by Valve in the year 2000. The Counter-Strike game, including its core game mechanics, is being further improved and developed until today. The most recent standalone version of Counter-Strike today is Counter-Strike:Global Offensive (CS:GO), with its first version released in August 2012.

(a) Top-Down View **(b)** Bombsite A **(c)** Bombsite B

Figure 1: The competitive map de_mirage

Playing Counter Strike is divided in rounds of around two minutes, depending on the selected game mode. A round ends, when a team is able to achieve a win condition, resulting in winning the round and making the opposing team lose that round, respectively. At the end of a round, the team, and all players in it, receive an amount of in game currency based on their individual and team performance in the concluded round. In game currency is used to purchase better player equipment, like weapons and armor, and strategic utility items, for example flashing grenades or smoke grenades. Equipment and utility items are used to gain an advantage over the opposing team in completing the strategic objectives

in the following round. If a player's character dies during a round, the character loses every previously purchased item and the player has to wait until the end of that round to play again. The amount of in game currency a player possesses spans across rounds and is only reset under reset conditions. The punishing element of losing purchased value in form of equipment and items, results in a incentive for the player and his team to play in a manner, that maximizes won rounds and minimizes the loss of equipment and in game currency. Managing the amount of in game currency a player has over the span of multiple rounds creates an additional economical strategic aspect of the game. The strategic object in a match remains the exact same for a certain team, but can change during the course of a round.

Due to it strategic variety and good balance between playing on the CT or T team, the specific game mode "bomb defusal" emerged as the competitive game mode of Counter-Strike. This game mode, extended with a small competitive rule set, embraces every game mechanics aspect mentioned in the previous paragraph and implements the five presented qualities of a competitive discipline from section 2.1:

In the competitive game mode "bomb defusal" two teams of 5 players play against each other for a period of 30 rounds. The round time is limited to 1 minute 55 seconds. The match is played on special maps with 2 static areas on the map, defined as bombsites. Figure 1 shows the structural and visual design of a popular competitive map called *de_mirage*. In Figure 1a the areas drawn by a red line are the two bombsite, the green areas are the location where each of the two teams start. The bombsites are always placed closer to the starting area of the defending side to allow a short preparation phase for the defense. A team wins the overall match when it reaches 16 won rounds. Before the match starts a coin flip decides what team is starting on which side, on the offensive T side or the defense CT side.
After 15 rounds played, the teams change sides from CT side to T side and vice versa. If playing all 30 match rounds results in a draw, an overtime of additional six rounds will be repeatedly added until a winner is determined. With the start of the match and after changing sides, every players currency balance is reset to $800, which has a small value in terms of purchasable items. The main strategic objective for the attacking site is to reach one of the two bombsites on the map, plant an explosive device and ensure its detonation. The timer of the device is in a competitive rule set to 40 seconds. The T side, the attackers, win the round if the bomb timer reaches zero seconds. Counter Terrorists on the other hand win a round, if they are able to successfully defuse the planted device or by running down the round time. Additionally to this primary win condition, both teams are able to win the round by eliminating all players of the opposing team. This win condition is secondary and therefore applies only if the main objective of planting and defending the device on one of the two sides is not active. In other words: Both teams can win the round by eliminating all players of the other team, but only before the device has been planted on a bombsite. Therefore, if the attacker team successfully reaches and plants the device on a side, the defenders must defuse the device in the 40s time window to win the round, even when no attackers remain.
After the conclusion of a round, both teams are awarded an amount of in game currency depending of the outcome of the previous round. Winning a round with the primary win condition awards the most money ($3600), losing the round without eliminating a opposing player the least ($1400). The competitive rule set limits the maximum amount of in game money to $16000. Table 1 displays a price overview of some items and categories to put the in game currency balance into perspective. A defuse kit can be purchased by the CT side to reduce the time needed to defuse the bomb from 10 to 5 seconds.

Utility Grenade	$200-$500
Defuse Kit	$400
Kevlar & Helmet	$1000
Pistols	$200-$850
SMGs	$1050-$2350
Rifles	$2000-$5000

Table 1: Item prices in competitive CS:GO

Counter-Strike's competitive game mode provides sufficient game depth by being based on a simple conceptual game mechanic, but giving the opportunity of creating team based, dynamic strategies. The factors of positioning, real time team communication, the usage of utility items and individual player talent enables the creation of a broad variety of in depth strategies. Game components based on randomness do not exist, so a player's skill and most importantly the team's skill is the deciding factor of winning. Due to its game depth the set of strategies used by teams is constantly evolving and being adjusted to the opposing team. External influences as the acquisition of a new player or changes in the competitively played maps also have an influence on the set of strategies. Counter Strike lenticular game design follows the concept of "easy to play, hard to master". Effective movement, improving aim, reaction times, communication and utilizing all utility items in the best way provide a constant room for improvement, both for beginners and professional players.

Today, the franchise Counter-Strike has become a significant part of the eSports sector. Competitive tournaments played with the most recent version Counter-Strike:Global Offensive is attracting many of viewers online and broadcasted via television in countries around the world. CS:GO Major Championships are tournaments organized multiple times a year with a prize pool of one million dollars and held in large venues in front of a live audience. In the most recent Major Championship, the "FACEIT Major: London 2018" 24 teams participated to compete against each other in the SSE Arena at Wembley.

2.3 Rating Systems

The previously elaborated role of skill in a competitive discipline raises the question of how to measure, adjust and compare participants against each other. Tracking the skill of participants in a competitive settings allows to make statements about the level of experience and mastery a player achieved in the respective discipline. Further, it opens up the possibility to create fair encounters by matching participants of similar skill level against each other. This results in a more competitive environment by improving the experience for players as well as giving players the opportunity to test and improve their skill against players with a similar background. The approximation of a participants skill is being done in the form of assigning a numerical score to every team. A list of those numerical scores is then called a rating [7] and a rating can be used as a sorting criteria to produce a ranking. A participants score can be a summarized numerical value calculated from more granular values tracked and updated during the participants competition history. Rating systems define a score, the structure and updating process of internal numerical values based on a match outcome and a process of comparing scores to each other. This section introduces three established rating systems using a simple running example from the competitive sport soccer. The visualization of examples to support the description of Elo and TrueSkill ratings are inspired by Jeff Moser's work [8] about TrueSkill in which he collaborated with Microsoft Research Cambridge.

2.3.1 A Basic Example

To illustrate the methods of different rating systems in a more clear and understandable way, an running example based on real sport data from the german Bundesliga will be used. The idea of a having the same running example for each rating method is inspired by the structure of [7]. For the sake of simplicity only five clubs with four matches each of the season 2017/2018 are included. The results are in the form of *row-column* and therefore inverse to the diagonal.

	BSC	LEV	FRA	HAM	MUN	wper	gdif
Hertha BSC	-	2-1	1-2	2-1	2-2	0.66	1
Bayer Leverkusen	1-2	-	4-1	3-0	1-3	0.50	3
Eintracht Frankfurt	2-1	1-4	-	3-0	0-1	0.50	0
Hamburger SV	1-2	0-3	0-3	-	0-1	0.00	-8
Bayern Munich	2-2	3-1	1-0	1-0	-	1.00	4

Table 2: Subset of Bundesliga Season 17/18 results

As seen in Table 2 every team has a set of four match results. Our goal is to approximate, based on the results, the performance (skill) of every participating team and assign every team a score expressing that. The two most right columns list two intuitive approaches: The win percentage(wper) and goal difference(gdif) for every team. In the case of the win percentage score, ties are neither counted as a loss nor as a win for both teams.

	wper	gdif
BSC	0.66	1
LEV	0.50	3
FRA	0.50	0
HAM	0.00	-8
MUN	1.00	4

(a) ratings

	wper	gdif
BSC	2	3
LEV	3|4	2
FRA	3|4	4
HAM	5	5
MUN	1	1

(b) rankings

Table 3: Calculated club ratings and rankings based on winp and gdif

Derived from Table 2 two lists of *participant-to-score* pairs can be created, as shown in Table 3a, to visual approximate performance distributions. Table 3b presents, based on the preceding ratings, a ranking of the five soccer clubs, 1 being the highest and 5 the lowest rank.

Concerning the strongest and weakest club, the rankings are the same: Bayern Munich has in both ratings the highest value, Hamburger HSV the lowest. Looking at two other clubs, LEV has a higher score in terms of the goal difference than BSC. However, this is not the case in the win percentage rating, where BSC has a higher score than LEV. This result shows, that two different rating systems can result in different ranking and therefore different approximations of participants performances. It is important to keep in mind what kind performance one is aiming to approximate when choosing or creating a scoring system:
For example, if the goal is the estimation of the two most goal oriented, aggressively playing clubs in the league, the goal difference is going to be more promising as a base for a ranking than the win percentage.
Constructing scores based on multiple, more granular numerical values or even other scores is very much viable. One possibility is to use the goal difference from Table 3a to combine with win percentage, so a club with the same win percentage but higher goal difference as another club, would have higher score. This might result in an overall better ranking position of the team. In our example in particular, a combined scoring as presented, would ultimately assign LEV a better rank than FRA.

2.3.2 Elo Rating

Arpad Elo (1903-1992) was a physics professor at Marquette University in Milwaukee, Wisconsin. Besides his occupation at the university he was also a experienced chess player. His interest in chess led him to create a new way to rate and rank chess players in competitive situations. The underlying idea of Elo's rating system was to express a chess players performance as a random normally distributed variable X with the mean μ [2]. As a consequence, a player's actual performance is expressed as a derivation of the expected performance, the mean μ. Therefore, as a consequence a rating r will be adjusted based on the derivation's distance to the mean μ. Elo presented a linear form of the adjustment of r being proportional to the performance deviation from μ. Let's assume a players actual performance is S, then transforming his old rating r_{old} to r_{new} to reflect the deviation from μ is achieved with

$$r_{new} = r_{old} + K(S - \mu) \tag{1}$$

where K is a constant defining a amplification of the derivations impact on r_{old}. Elo originally has chosen $K = 10$. Figure

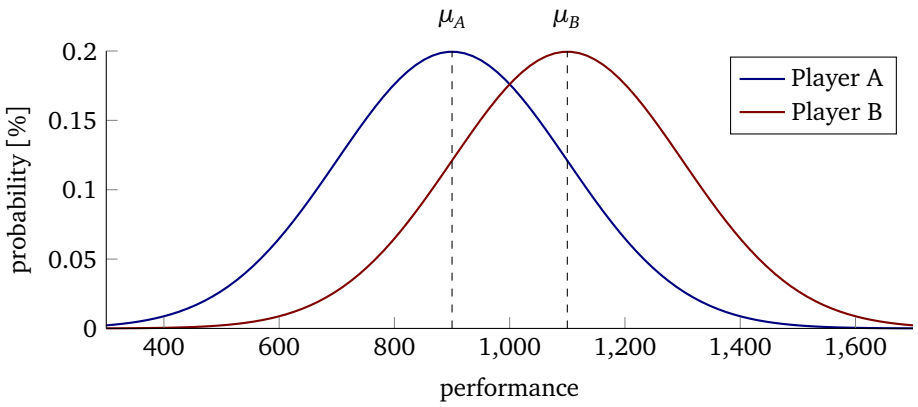


Figure 2: Expected performance of two players expressed by two normal distributions

2 illustrates two player's expected performances expressed as the probability density functions (PDF) of two normally distributed with $\mu_A = 900$, $\mu_B = 1100$ and a standard derivation $\sigma = 200$. Based on the displayed assigned Elo score, Player B is expected to win a game against Player A. In order to determine the rating adjustment amount of both players, for both possible outcomes that Player A wins or Player B wins, the calculation of the expected performance difference is required. This performance difference, resulting from the subtraction of both PDFs, answers the question of the likelihood of a specific match outcome.
Figure 3 displays the expected performance difference between previously introduced Player A and B from Player B's perspective, by subtracting Players B variable from Players A. The deciding data point in that figure is 0 on the x axis and based on that the integral from that point downwards the x axis. That integral, calculated by the cumulative density function, expresses the likelihood of Player B performing worse than Player A. Therefore the expected likelihood of Player B losing against Player A, with their prior assigned ratings, is 23.9% and since the Elo rating is a zero sum game, the likelihood of Player B loosing is 76.1%. Since the calculation of a the PDF and CDF of a normally distributed variable is only numerically solvable, modeling a players skill derivation by a logistic function was adopted. Using the logistic function proved also to be more accurate in the discipline chess and is today usually calculated with:

$$L(x) = \frac{1}{1 + 10^{-x}} \tag{2}$$

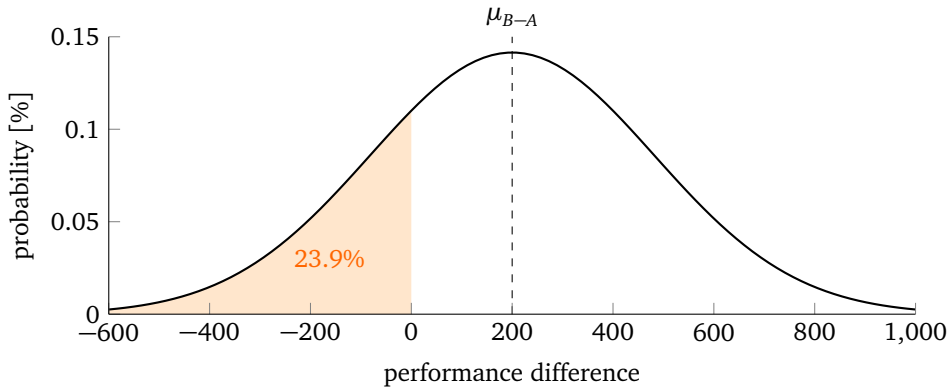


Figure 3: Expected performance difference between players

and the likelihood of a match result between Player A and B with:

$$\mu_{AB} = L(d_{AB}/400) = \frac{1}{1+10^{-d_{AB}/\xi}} \tag{3}$$

with

$$d_{AB} = r_A(old) - r_B(old) \tag{4}$$

and $\xi = 400$ serving as an additional parameter, modeling the spread of the distribution.
Based on this equations, it is possible to visualize a general dependency between the Elo rating difference and the likelihood of winning, or win probability. To understand this relation better, Figure 4 visualizes exactly that with $\xi = 400$. Adjusting the parameter $\xi = 400$ has an effect on how significant a specific Elo rating difference effects the expected win

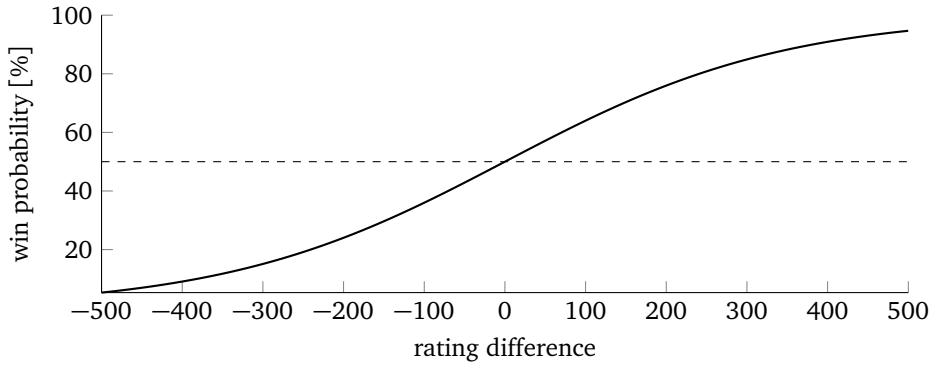


Figure 4: Generalized dependency between rating difference and win probability

probability. Returning to the running example introduced in the previous subsection, we now can use the Elo rating to calculate scores based on the results in Table 2. Choosing $\xi = 400$, $\mu = 1500$ and $k = 16$ results in the Elo rating system based ranking displayed in Table 4. The resulting ranking is similar to the win percentage based ranking of Table 3b with

Rank	Team	Elo
1	MUN	1537.06
2	BSC	1536.18
3	LEV	1499.06
4	FRA	1475.94
5	HAM	1451.75

Table 4: Bundesliga example ranking based on Elo rating

the exception of rank 3 and 4. In contrast to the win percentage based ranking, where two teams have the same score and share rank 3 and 4 respectively, the Elo rating system placed LEV clearly above FRA. Assigning ratings to teams or players using the Elo rating system turns out to be more accurate and reliable than simple approaches like win percentage or goal difference.

2.3.3 Microsoft Trueskill

The TrueSkill rating system has been developed by the Microsoft Research Lab as a solution to improve matchmaking between players on the Xbox Live platform [4]. It is designed to work for teams of arbitrary size and for matches with a arbitrary amount of teams playing against each other. The underlying idea of the TrueSkill rating system has its foundation in the previously introduced Elo rating system. A players skill is approximated by a distributed probability X around μ with a derivation of σ. TrueSkill extends the idea of the Elo rating by introducing an additional value that is adjusted parallel to μ. By tracking and adjusting σ for every participant, TrueSkill overcomes a disadvantage of the Elo rating system. Score calculated with the Elo rating system only express where the approximated skill of a player or team lies, but not how confident the system is about the skill being there. The formula of adjusting rating in the Elo system allows to model uncertainty by changing the k-value under certain circumstances. TrueSkill solves this by generally keeping track of the confidence in a players skill approximation.
The default recommended parameters for a TrueSkill environment are $\mu = 25.0$, $\sigma = \frac{\mu}{3}$, $\beta = \frac{\sigma}{2}$ and $\tau = \frac{\sigma}{10}$. The parameter τ describes a dynamic factor taking into the calculation of the σ adjustment after a match. It introduces a certain degree of randomness to σ so it does not converge to 0 with increasing amount of matches played at the same skill level. β allows to tailer the TrueSkill environment to the computer game of application by defining a distance between ratings that guarantees 80% of winning. The Microsoft Research Team refers to that parameter as the "length of one part in a skill chain". Games involving a lot of skill are usually expressed with a smaller β value than games with a larger randomness factor. Small differences in points cause in a skill based game therefore quicker to the 80% winning chance mark.
In order to visualize the core concept of TrueSkill in contrast to the Elo system, Figure 5 displays the TrueSkill based ratings of two players before a match. Player Blue is higher ranked and has been assigned $\mu = 30$ and $\sigma = 1.5$ and the

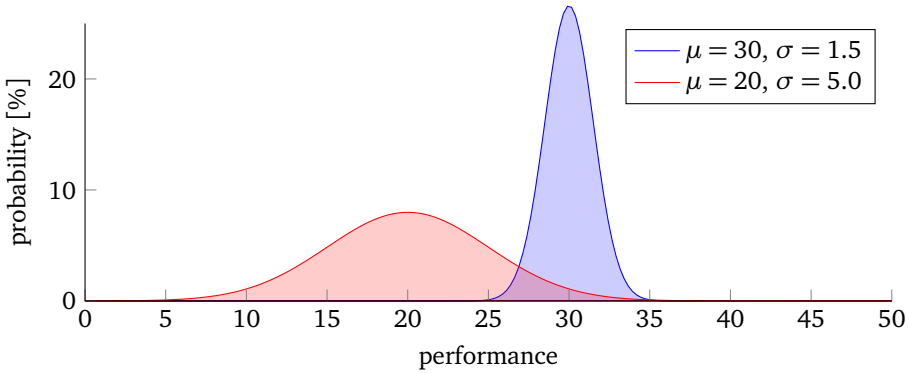


Figure 5: Two player's TrueSkill rating before the match

less skillful Player Red $\mu = 20$ and $\sigma = 5.0$ Judging from those parameters, the TruSkill system is a lot more confident in Player Blue's skill approximation than Player Red's and Player Blue is the favorite in an upcoming match. Let's assume Player Red is able to play against the odds and win the match with Player Blue, then Figure 6 reflects the adjustment μ and σ values of both players. Player Red's rapid adjustment of μ is very prominent, Player Blue's adjustment less. This is

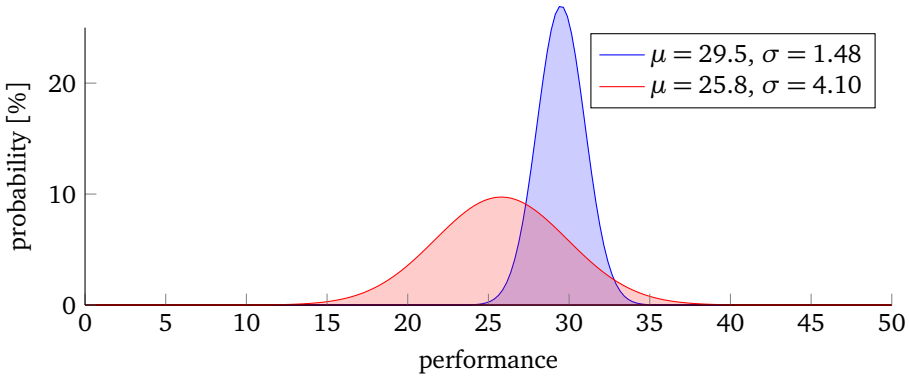


Figure 6: Two player's TrueSkill rating after the match

a desired behavior of TrueSkill and has a good effect on actual matchmaking in online games. A high skilled player, which competed frequently on his skill level does not receive a major loss in rating by losing to a worse player once. TrueSkill's

confidence value σ serves therefore as a stabilizing factor of a consistent player. This sums up the major advantage of TrueSkill compared to the Elo system, where the better player would receive a major rating loss due to having one bad game.

We use our running example once again and apply TrueSkill to the Bundesliga results. In order to rank the Bundesliga clubs based on TrueSkill we use Jeff Moser's approach [8] to derive a score from μ and σ:

$$s = \mu - 3 * \sigma \tag{5}$$

Since we use the default parameters, this will create a score starting at 0. Table 5 show the resulting ranking based on the score and the tracked parameter values for every team. Ranking the Bundesliga clubs with TrueSkill yields the same

Rank	Team	Trueskill	μ	σ
1	MUN	15.56	33.42	5.96
2	BSC	15.42	32.97	5.85
3	LEV	9.38	24.96	5.19
4	FRA	5.12	20.87	5.25
5	HAM	-1.29	15.37	5.55

Table 5: Bundesliga example ranking based on Trueskill rating

order as with Elo rating, but has one additional interesting aspect to it. TrueSkill assigned the best ranked club Bayern Munich the weakest confidence score of $\sigma = 5.96$ and Bayer Leverkusen the strongest with $\sigma = 5.19$. This information is not available with the Elo rating and can turn out to be in the scope of a larger set of results significant.

2.3.4 Offense Defense Rating

Amy N. Langville and Carl D. Meyer present in their book "Who's #1? - The Science of Rating and Ranking" a rating system for sport disciplines with offensive and defensive game aspects [7]. The rating system focuses on the respective offensive and defensive performances of competing teams and derives based on recorded results two ratings per team. The offensive rating expresses how well a team is performing in the offensive game aspect, the defensive rating the teams defensive abilities. Instead of evaluating the outcome of a match in the form of "win or lose" the proposed offense defense rating, short OD-Rating, processes solely the team scores of a match. For example, the outcome of a football match between Team A and Team B is 2:1 in goals, Team A would have gained 2 offensive points over Team B and 1 defensive point, since it let the other Team score one goal. By this definition, the higher the offensive points against another Team are, the stronger the offensive ability of a team is. In contrast to the offense points, the defense of one Team is strong, when it's defensive points of a match are low. This leads to the following definition of the OD-Rating for Team j as described in [7]:

$$o_j = \frac{a_{1j}}{d_1} + \frac{a_{2j}}{d_2} + ... + \frac{a_{mj}}{d_m} \tag{6}$$

$$d_j = \frac{a_{i1}}{o_1} + \frac{a_{i2}}{o_2} + ... + \frac{a_{im}}{o_m} \tag{7}$$

where a_{ij} stands for number of points j scored against i. For the previous example, for Team A this would result in a offensive rating defined by dividing the offensive points (2 points) it was able to score against Team B, divided by Team B's defense rating. The defense rating is calculated in a similar way and is the result of dividing the defensive points (1 point) by the offense rating of Team B. It is important to notice, that a offense rating with a high value assigned represents a strong offense, but a high value assigned to the defense rating a weak defense. By looking at the formula this leads to a desired behavior when evaluating results between two Teams. A team's offensive rating increases significantly if it was able to score a lot of offense points a_{mj} against a team with a high defense rating d_m. The same principle applies to the calculation of a Team's defense rating. If a Team is able to keep its defense score a_{im} low against another team with a strong offense o_m, the resulting impact on the teams defense rating remains low, resulting in a strong defense approximation.
The definitions of the offense and defense rating are cross referencing each other leading to the question of how to calculate the OD-Rating. A simple solution to this problem is to assume a start defense rating for every team with the value of 1. Based on this starting defense rating, the following step consists of calculating every team's offense rating. After completing the first iteration of the offense rating calculation, the defense ratings, which are still set to the start value 1, can now be calculated using the offense ratings. This process continues until all offense and defense ratings start

converging to a desired decimal delta.
The visualization of the results in our running example 2 in form of a score matrix is perfect for a OD-Rating calculation. Hertha BSC, for example, scored 2 offensive points and 1 defensive point against Bayer Leverkusen, which corresponds to the entry in the first row and second column. Applying the formulas of the OD-Rating to every match entry in 2 results for every team in a offense rating, Table 6a, and a defense rating, Table 6b. Both Tables are sorted accordingly to the approximated ratings creating a ranking of the Bundesliga clubs based on their offensive and defensive performances, respectively. Combining the offense and defense rating of one team to a single numerical value can be done in different ways as described in [7]. We choose a simple approach and express the overall rating of a team j with:

$$s_j = \frac{o_j}{d_j} \tag{8}$$

This applies the same weight to the defensive capabilities of a team as to the offensive ones. Based on the calculated overall OD-Rating per team, we can rank the five clubs from the running example as shown in Table 6c. Compared to rankings based on the Elo rating and TrueSkill rating, which produced the same order, the OD-Rating puts BSC on the fourth place instead of the second. MUN is still on the first place, as HAM remains on the last. It is interesting to notice, that the OD-Rating system assigned FRA and BSC scores, that are very close to each other, where the Elo and TrueSkill rating did not.

Rank	Team	O Rating
1	LEV	14.33
2	FRA	10.64
3	MUN	9.75
4	BSC	9.63
5	HAM	6.43

(a) O-Rating

Rank	Team	D Rating
1	MUN	0.59
2	BSC	0.94
3	FRA	1.01
4	LEV	1.13
5	HAM	1.32

(b) D-Rating

Rank	Team	OD-Rating
1	MUN	16.52
2	LEV	12.68
3	FRA	10.53
4	BSC	10.24
5	HAM	4.87

(c) OD-Rating

Table 6: Bundesliga example ranking based on OD rating

2.4 Rating Systems Comparison & Match Prediction

Table 7 summarizes the created rankings based on the three presented rating systems including the simple rating approaches. In order to evaluate how accurate the presented rankings and underlying skill approximations are, a method of comparing the rating systems to each other is required. The most intuitive choice is to evaluate the approximation accuracy of the assigned scores by predicting the outcome of future matches and comparing the prediction with the actual outcome. If a future match contains two teams with assigned ratings, the two respective scores of a rating system are compared to each other and a winner is predicted. After predicting a set of future match results, which must not have been used to establish the scores of the teams before, one can calculate a win prediction ratio for every rating system. The baseline for this evaluation metric is a simple predictor without any knowledge about the competing teams, in other words a coin flip. This baseline has the value of 50% and serves as a important point of reference for the evaluation of rating systems for win prediction.

Rank	wper	gdif	Elo	Trueskill	OD-Rating
1	MUN	MUN	MUN	MUN	MUN
2	BSC	LEV	BSC	BSC	LEV
3	LEV|FRA	BSC	LEV	LEV	FRA
4	LEV|FRA	FRA	FRA	FRA	BSC
5	HAM	HAM	HAM	HAM	HAM

Table 7: Bundesliga example ranking comparison

In the case of the running example from the Bundesliga, a valid approach would be to extend the rating calculations to process the first 70% of a season's matches. This would result in a extensive skill approximation for every participant of the Bundesliga. After the approximation of the respective skills with the selected rating systems, the training phase, one can start predicting the remaining 30% of the season's matches. The amount of matches, the right and the wrong predictions have to tracked during this evaluation phase. At the end the rating system's performance in win prediction can be evaluated by comparing the win prediction ratio to each other and most importantly to the baseline.

3 Implementation

After the introduction of the competitive discipline Counter-Strike with all its characteristics and the explanation of three rating systems, the next logical step is to apply those rating systems to professional Counter-Strike matches. Choosing the right data, like match results, acquiring and preprocessing them to build a database for the training of the rating systems Elo, Trueskill and OD is the necessary first step. Through the digital native aspect of Counter-Strike as an eSport discipline, a high volume of match data of the last three years is available. This thesis' experiment will make use of that advantage, train and evaluate the proposed rating systems. The evaluation's goal is the investigation of the rating systems qualities to predict the outcome of future competitive matches. Topic of research is to evaluate how well the rating systems perform compared to each other, under what circumstances they perform the best or worst and what the underlying trade-offs are.

This section explains the creation of the experiment, its setting and goal. First, it outlines the requirements and acquisition of the needed data of competitive Counter-Strike matches in order to train and evaluate the introduced rating systems. The sets of available and required information about a competitive match of Counter-Strike will be explained as well. The second subsection will show and explain the motivation behind the data sets created and used in the experiment. Specific qualities of those data sets will be characterized and compared to each other. In subsection three, the rating systems environment, especially the used parameters during the training phase are listed. Following that, in subsection the evaluation environment and its methods of analyzing and comparing rating systems to each other will be explained. Lastly the observed outcome of the experiment is shown in the last subsection of this section.

3.1 Data Requirements and Collection

In order to calculate scores for the three presented rating systems, data containing past match results, participating teams and the respective match scores is required. A match score should contain the information how many rounds a team won playing as the CT and T side. One team's sum of won T and CT rounds must results in 16, or in the case of added overtimes, over 16. Further, only competitive matches, meaning matches played with a competitive rule set in "bomb defusal" mode are required. Those information create the minimal requirements for the database used in this experiment. A team is treated as one entity, the players forming the respective team are not part of the observations and not required to be obtained. Further, additional to the team information, match results and score information, event information and the time and date are interesting for pre processing purposes. Match results are stored on a map basis, meaning that, even when two teams played in a best of three maps format, the results are modeled as one record per map. The name of the map played between two teams is also an interesting information for pre processing and filtering. The described data requirements are the base of the record structure for the experiment.

id: 70035/astralis-vs-natus-vincere
time: 1530972000
map: Inferno
match_url: 2324404/astralis-vs-natus-vincere-esl-one-cologne-2018
map_stats_url: 70035/astralis-vs-natus-vincere
event: ESL One Cologne 2018
team_a **id**: 6665/astralis **name**: Astralis **offense**: 8 **defense**: 5 **rounds**: 13
team_b **id**: 4608/natus-vincere **name**: Natus Vincere **offense**: 10 **defense**: 6 **rounds**: 16

Table 8: Data record structure example

The source the experiments data is the public webpage *hltv.org*. It is the leading site covering all Counter Strike major tournaments, qualifiers and leagues with detailed team, match and event information. Detailed statistics concerning a matchup between two teams are displayed and the whole match can also be downloaded in form of a file called *demo*. A *demo* contains information about the state of the match up to 120 times per second and provides therefore a vast

amount of information. Besides *demos* all required match information described in the preceding paragraph and more are publicly available. The correctness of the displayed data is maintained by a community reporting the results and the maintainer of the webpage entering them. At the time of writing this thesis, *hltv.org* contained over 37.000 match entries. This results in even more map results entries for analysis since match entries also may contain best of three and best of five map formats. Further the webpage maintains a own scoring of competitive matches, which assigns a match a score between zero and five stars. That rating reflects the competitive quality of a match based on assigning teams a internal ranking. A match is assigned, for example, one star, if one of the two teams participating in it is ranked under the 20 best teams at the time played. Two star matches are played between two Top 20 teams, three star matches between two top 10 teams and four star matches between two top 5 teams. The highest rating assigned are five stars, where both teams belong to the best three teams making those matches the highest quality. The described *hltv.org* star rating is not relevant for this thesis in the context of the three rating systems presented. We will not evaluate or compare that rating to Elo, Trueskill and Offense-Defense. It serves only for the purpose of filtering the large webpage database with the goal of creating interesting data subsets.

The entire available set of matches of *hltv.org* has been obtained to research the thesis and the star rating has been used to create data subsets. The specific data sets and their characteristics are described in the following section. Match results are split into map result records to guarantee the same level of granularity. The requirements from the previous paragraph are extended with additional information to make the resulting data structure processable. Converting time and date in a unix timestamp, assigning a unique record and team id and maintaining a webpage link to the respective *hltv.org* page result is included in the record structure. The data is fetched using a python based web crawler, encoded with the format json and imported in a MongoDB database system. Table 8 contains a sample from the map result database. The prefix of the URLs is omitted, because it does not change per entry.

3.2 Training and Evaluation Data

The goal of calculating and evaluating rating systems requires a division of the match data in a so called training and evaluation set. A training set is used as a information source for calculating ratings for every team. The evaluation set consists of records not used in the training set and serves as a testing environment for the previously calculated ratings. Rating systems are primarily evaluated in terms of map win prediction performance on the evaluation set by comparing the calculated ratings of both teams and making a prediction what team will win the map. In this thesis, all records of a training set are processed in a chronological order to maintain a realistic development of teams performances. Records of the evaluation set a also processed in a chronological order and do not influence the calculated rating of the teams. This way we have a strict distinction between a training phase and an evaluation phase, so that evaluating the ratings do not influence the ratings itself. Therefore it is not required to process the evaluation set in a certain order, since evaluating one record has no impact on the evaluation of another. This means, that choosing to evaluate records chronologically has no effect on the prediction results.

3.2.1 Data Sets

Before creating a data set for training and evaluation purposes, the crawled records have been filtered so no corrupt or inconsistent record remains. Based on all obtained matches, which we will call "ALL" data set, five more data sets has been created. "ALL" last update is the *25.10.2018* The "ONE-STAR" set contains all matches rated by *hltv.org* with one star, so it only contains results of matches in which at least of team was under the top 20 at that time. "TWO-STARS" restricts the previous data set by only containing match results of top 20 teams and "FIVE-STARS" contains only results of top 3 teams. The creation of those data sets allows to investigate the rating systems in environments of different team density.
Further, the data sets have been selected to also reflect different skill densities, with the "ALL" data set having the most spread out skill of teams and "FIVE-STARS" the least. "ALL" contains every result recorded on *hltv.org* since 2012 which results in the tracking of a lot of different teams. Since the creation and closing of a team is more rapid compared to traditional sports, this leaves us with a lot of temporary teams with very little information about their performances. Increasing the *hltv.org* stars as filter, reduces the amount of such teams and results increasingly. "ESL-NA" and "ESL-EU" contain, as the fifth and sixth data set, match results of one season of a world wide CS:GO league. More specifically, the results of the european and north american season 7 ESL Pro League, which is played out in a typical league format where every team has to play against each other in the course of a season. In this data set every participating team has the same amount of played matches, which is a special case worth looking into. Table 9 provides a overview of the five mentioned data sets with a few key characteristics.
Temporary teams are teams with less than three map results in the respective data set. "ESL-EU" has exactly 26 results per team, which is the case when 14 teams play against each other two times. The north american league data set is

not as complete as the european, since the data source contained forfeited matches. Forfeited matches do not contain a detailed score, since no map has been played and therefore those forfeited matches are excluded from the set.

	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
map results	50912	10926	4532	171	184	126
no. teams	3499	510	57	18	14	12
temporary teams	945	216	3	4	0	0
results per team[avg]	29	42	158	19	26	21
density	0.05%	0.38%	3.49%	11.1%	14.1%	16.5%

Table 9: HLTV Data Set Stats

It is interesting to notice the different characteristics of the data sets. The number of temporary teams increases with the number of total records in a data set except for the "TWO-STAR" data set, which has compared to the "ONE-STAR" set a lot less temporary teams. The requirement of both teams being in the top 20 seems to restrict the occurrence of temporary teams. As a indicator how many information a data set provides per team the average map results per team in the data set seemed promising. Since that metric, in row four, is not relative to the size of the set, a metric based on the relation between row 4 and the amount of records is used. In Table 9 it is called density and is the percentage of the average results per team in all matches played. This puts the result distribution over the participating teams into relative perspective of the data set volume and allows to better see an additional data set quality.

3.3 Rating Calculation

Every rating system is calculated in chronological order, so in the same order the records in the data sets are organized. In the following, the implementation of Elo, Trueskill and OD-Rating is explained and necessary parameter choices presented. During first observations the most promising parameters for Elo, Trueskill and OD-Rating has been chosen and therefore they will not be subject for further evaluation. In the following subsections important parameters and their values are presented for each rating systems respectively. The changes in rating for Elo and Trueskill are processed during the respective training periods. Due to its nature, the OD-Rating system is only calculated at the end of the training period. In general, no records of the data sets will be skipped or processed differently during the training of the rating systems.

3.3.1 Elo Rating

The Elo rating is calculated in the same manner for every data set presented. A starting rating value of 1500 is chosen. Based on the introduction to the rating system in Section 2, we implement the following formula to calculate the expected outcome between Team A and Team B:

$$exp_A = \frac{1}{1 + 10^{(r_B - r_A)/400}} \tag{9}$$

where r_B is the prior rating of Team B, r_A the rating of Team A and exp_A the likelihood of team a winning the match. Let's remember, that exp_A is always a value between zero and one and $exp_A + exp_B = 1$.
The rating adjustments for both teams depend on the outcome of the match. If we stay in team a's perspective, team A's score at the end of a match can be 1.0 , 0.0 or 0.5, representing a win, loss or draw respectively. For the calculation of the Elo rating offense or defense rounds and the difference in won rounds will be ignored. Team A scores 1.0 if and only if the team was able to win more rounds overall than Team B. Both teams winning 15 rounds results in a win expectation of 0.5 for Team A and B. The rating adjustment is done with the commonly used k-value of 24 and following formula:

$$delta_A = k * (score_A - exp_A) \tag{10}$$

For Team A and Team B the gain in rating has the same absolute value, but different signing and correctly grows in value anti proportional to the likelihood of the match result. Team A's new elo rating is the sum of the old rating with the calculated $delta_A$ and Team B's new rating the sum of Team B's old rating with $delta_B = -1 * delta_A$. Whenever a rating is updated, the amount it changes and the amount of times it changed is being recorded.

3.3.2 Microsoft Trueskill

The Microsoft Trueskill rating is calculated by using a python library implementing the Microsoft research paper about Trueskill [4]. It has been implemented by Heungsub Lee and is available under *trueskill.org*. The correctness of the rating calculation has been confirmed and in the case of this experiment competing teams are processed as one entity.

Processing the map results on a team base instead of players is desired to maintain the same scope across all three ratings. The Trueskill library and Trueskill itself offers the possibility of assigning every player a separate score and even create matches with N teams of M players. This functionality is not used in the course of the evaluation and the 1v1 rating method is chosen. Trueskill is calculated with the default parameters $\mu = 25.0$, $\sigma = 8.33$, $\beta = 4.166$, $\tau = 0.0833$ and a adjusted draw probability of 0.01. The library for the calculation of the cumulative distribution function Φ and the probability density function Φ^{-1} is set to use scipy.
Similar to the Elo implementation, the training of Trueskill looks at the total won rounds of a team and sets the team with most won rounds as the winning team. Calculating the expected outcome of a match is not needed for the training phase, since the provided method *rate_1v1(rating1, rating2)* returns the two new rating. The winning team is expressed by setting *rating1* to the prior rating of the winner.

3.3.3 Offense Defense Rating

The calculation of the OD-Rating presented in [7] is implemented using a OD-Matrix. Before populating the OD-Matrix, during training phase, all results are saved and validated in a MongoDB collection. If two teams played more than once in the training set, the respective scores are adjusted as the mean of the two teams scores. In contrast to Elo and Trueskill, the OD-Rating is supposed to process the respective offense and defense performances of teams. Determining what results to feed into the OD-Rating was a interesting task. After further observation and comparison of some choices to existing OD-Rating examples, the offense rounds of teams has been picked as a representation of performance. This decision may not seem intuitive since both teams have a number of offense and defense rounds recorded in the match data sets. The reasoning behind picking only the offense rounds is the idea of treating Counter-Strike matches in a similar fashion as football or soccer matches: How well a team performed in the offense is expressed by the amount of goals it was able to score. In Counter-Strike we define this as won offensive rounds, the rounds a team was able to overcome the other teams defense. Since every team in a Counter-Strike match is supposed to play 15 rounds as T and CT, the won offensive rounds equal to 15 minus the won defensive rounds. The fact that matches are closed, before all 30 rounds are played out, might add some confusion to this, so let's look at an example:

Team A wins 6 offensive rounds out of 15 rounds played in the first half. In order to win the match, Team A needs, after switching the sides from offense to defense, at least 10 defensive rounds. Let's assume Team A is able to win those 10 rounds without losing a single round, then the final map score would be 16:9, with Team A having 6 offense, 9 defense rounds and Team B having 0 offense and 9 defense rounds. This leads to the conclusion, that the winner of the match is always determined by the Team able to win the most offensive rounds. As mentioned before, this is a very similar to goals in soccer or football and therefore the perfect fit for the OD-Rating implementation. Defensive rounds are implicitly defined by the offense rounds and therefore are not used in the OD-Rating calculation. One could compare, in that regard, competitive Counter-Strike to 30 rounds of penalty kicks in soccer, where offense rounds are goals and defense rounds are defended shots by the keeper.

After adding every result of every team in the database collection, the numpy matrix is created. Calculating the OD-Rating inside the MongoDB with all results spread across different records turned out to be very inefficient, due to the complexity of the convergence calculation. The data set "ALL" contains 3449 teams, which leads to about 12 million operations per iteration. Taking the matrix based calculation of [7] and implementing it in a way that it is calculated in memory resulted in a major speed up. Convergence is guaranteed by applying the Sinkhorn-Knopp Theorem about non-negative and doubly stochastic matrices [10]. This leads us to the only parameter choice in the OD-Rating implementation, which is the value ϵ added to every matrix entry to make it strictly positive. Setting $\epsilon = 1$ resulted in a good performance of the rating system itself and converging of the matrix. Further the initial offensive rating vector is set to 1 at every entry. The implementation assumes that the OD-Matrix has converged, when the highest difference of an entry from one iteration to the next one is smaller than 0.001. This offers an rating accuracy to the third decimal digit. The OD-Matrix is converged only once, at the end of the training phase.

3.4 Rating Evaluation

Ratings are compared to each other primarily in regards of their win prediction performance. The win prediction is expressed as the percentage of correct predictions in the amount of predictions made. For example, if the evaluation set includes 4 matches and a rating system predicts the outcome of 3 matches correctly, the prediction rate is therefore 75%. Elo and Trueskill calculate the outcome of a match in the form of a win probability for one team. Internally it is a float value between 0.0 and 1.0, where 0.0 is the team having absolutely zero chance of winning. The closer the value reaches 0.5 the more balanced the match is assumed, and in the case of win prediction, the more unsure it is who will win. For this reason a parameter called "draw-delta" is introduced. The "draw-delta" or Δ_{draw} allows to define a certain absolute value around 0.5, where matches with a probability in that area are treated as uncertain to predict. Since draws are almost non existent in the presented sets, predictions within the Δ_{draw} are not treated as the prediction of a draw.

In order to keep track of how many predictions fall into this area of uncertainty, for evaluations with a set Δ_{draw}, a uncertainty value is being calculated. The uncertainty value is the percentage of uncertain matches in the evaluation set. The OD-Rating system does not predict the result via a win probability of a team, but compares two scores directly to each other. Defining a Δ_{draw} on the absolute difference of these two scores showed no significant effect on the win probability. Therefore, in the case of OD-Rating, the team with a higher calculated score is expected to win the match. Resulting from the implementation OD-Rating implementation choices presented in the previous subsection, the offense rating vector is taken as the overall score. This yielded the best performance over all subsets.
Every rating system is not only calculated on every presented data set, but also with a varying training and evaluation subsets. Three divisions have been selected to be applied for every data set: 70-30, 80-20, 90-10, where as an example, 70-30 stands for 70% of the data set is used for training and 30% for evaluation. Before dividing data sets in training and evaluation subsets, the respective set is sorted chronologically. This guarantees the training of the rating systems to be as realistic as possible. Converging the OD-Matrix of the OD-Rating is being done after entering every results [7], ergo iterating through the entire training set.

4 Evaluation

This section presents the resulting prediction rates achieved by the training and evaluation of the three selected rating systems Elo rating, TrueSkill and OD-Rating. The performance of each rating system is being evaluated and specific performance characteristics on the presented data sets investigated. A comparison of the overall rating system performances during the training and evaluation phases is discussed in the last subsection of this section.

4.1 Elo Performance

The Elo rating system is applied to each presented data set with different training phase and evaluation phase subsets. The calculated win prediction has been tracked and two additional reference values had been calculated to put the win prediction results into perspective:
In parallel to the Elo rating predictions one agent was always predicting that Team A will win, and the second agent chose randomly one team of the two. Both values converged at around 50% showing on one hand, that the data sets are not skewed towards Team A or Team B being prominent winners. Secondly it serves as a reference to a prediction agent without any knowledge.
Table 10 shows the results of the Elo rating based prediction rates with standard parameters.

set division	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	59.75	56.78	52.14	33.33	56.36	65.78
80-20	61.70	59.94	55.90	42.85	67.56	65.38
90-10	61.85	60.17	62.26	23.07	57.89	61.53

Table 10: Elo win prediction with $\Delta_{draw} = 0$

The Elo prediction performs well in data sets both with very low density and high density of played matches per team. In data sets with many matches, Elo tends to slightly improve with the amount of matches used for training. "ALL", "ONE-STAR" and "TWO-STARS", the data sets with the largest amount of matches show improvement the more training data is chosen. Looking at the data sets with the highest density, and closest team skill spread, Elo performs even better than in the previously mentioned ones. "ESL-NA" and "ESL-EU" the two league data of one season show both with the 80-20 division prediction rates above 65%.
Another result to notice is the anomaly of "FIVE-STARS". It appears Elo is in this case not reliable at all, or just very good in predicting the opposite. One possible explanation for this can be derived from the nature of the "FIVE-STARS" data set: It is a data set of matches played only between the top 3 teams ranked on the website since 2015. Overall half of the teams in the set have a relatively low amount of matches played, leaving 6 teams being the overall top teams. Matches between the top skill level of a competitive discipline are mostly very hard to predict, due to the closeness in skill and shifting strategies. Training Elo with win-lose information on a relatively small and skill wise narrow data set like "FIVE-STARS" is not enough to predict matches well. Another explanation for a bad prediction rate in the case of "FIVE-STARS" and in general can be that the approximated skills by the Elo system are very close together. In order to test this case, the evaluation process is repeated with $\Delta_{draw} = 0.01$. This will exclude matches from the win prediction statistics where the outcome prediction is closer than 0.01 to 0.5. Table 11 displays the resulting rates. It can be observed, that except for the 70-30 case where 20.8% of predicted matches have been excluded, the "FIVE-STARS" prediction rates are not positively affected. This confirms the first possible explanation, that high skill matches in the form of the "FIVE-STARS" data set vary in outcome and are difficult to predict.

win prediction	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	60.22	56.85	53.36	26.31	56.00	**66.66**
80-20	61.99	**60.75**	55.99	42.85	**64.70**	65.38
90-10	61.72	60.43	**62.46**	23.07	57.89	61.53
uncertainty	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	3.77	3.29	8.07	20.83	9.09	5.26
80-20	3.32	4.10	8.04	0.00	8.10	0.00
90-10	3.98	2.97	3.77	0.00	0.00	0.00

Table 11: Elo win prediction and uncertainty with $\Delta_{draw} = 0.01$

By further investigation of the results of the evaluation with $\Delta_{draw} = 0.01$, we can see some improvement in the prediction rate, while keeping the uncertainty below 10%. Excluding very close predictions from the prediction set seem to have a positive effect on the prediction rates. Exceptionally good results in general are highlighted in Table 11. Looking

at the prediction rates of the data sets containing match results from the ESL Pro League, "ESL-EU" and "ESL-NA", one can see a rate of almost 67%. The density of those two data sets is higher than "FIVE-STARS" and the participating teams in this league had to be qualified to participate in it. This leads to the assumption, that the true skill level of the teams playing in this league, and being in this data set, is narrow. The Elo system should perform therefore similar to "FIVE-STARS" poorly, but instead is able to predict a lot of matches correctly. One major difference between the two data sets is the time span in which the matches had been played out. The group phase of the ESL Pro League lasts little over two months, making the "ESL" data sets more dense in time. Additionally the distribution of played matches across all teams is more even compared to the "FIVE-STARS" data set. Both facts seem to have a positive effect on the prediction rate with the Elo rating system.

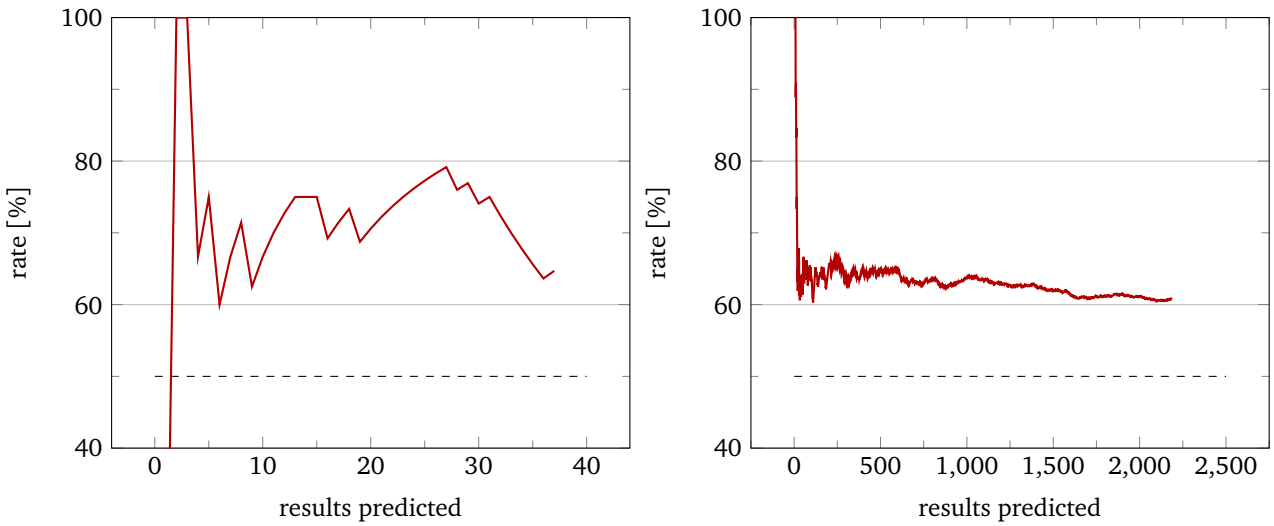


Figure 7: Elo prediction rate with 80% ESL-EU and 80% ONE-STAR, $\Delta_{draw} = 0.01$

We further investigate the Elo rating on the data sets "ESL-EU" and "ONE-STAR" with a 80% training phase and $\Delta_{draw} = 0.01$. The development of the win prediction over time during the evaluation phase for both data sets is shown in Figure 7. After predicting a match result, the win prediction rate of the Elo rating based predictor is recalculated and saved. Since the amount of match results in "ESL-EU" is rather small, rapid changes in the win prediction rate are observable. The spikes during the first few match results in both figures are caused by the adjusting phase a ratio metric has when calculated. The win prediction during the evaluation of "ESL-EU" spikes at nearly 80% at 27 results predicted, meaning that the Elo system performed very well on one specific day. Following that spike, the ratio starts dropping down to the final rate of 64.7%. A possible explanation for this behavior might be upsets or unexpected outcomes during the final days of the league, caused by teams not performing at their estimated skill.
The development of the win prediction rate on the "ONE-STAR" set is more stabilized due to the amount of results in the evaluation set. Interestingly, after reaching a peak of 65% at 1.100 matches evaluated the prediction rate slowly drops to the final value of 60.75%. This is probably caused by the time sensitivity of the prediction. Match results are being evaluated in chronological order, which can have the effect that the Elo rating misses the development of skills during evaluation. That can result, as the figure shows, in a drop of accuracy, since ratings are not adjusted in the evaluation phase.

4.2 Trueskill Performance

Parallel to the evaluation technique of the Elo rating system, the TrueSkill rating is applied to every data set previously presented. The same divisions in terms of training and evaluation set is maintained as in the Elo system. During the evaluation phases both baseline agents are calculated as in Elo to provide a point of reference. Table 12 displays the win prediction results of the TrueSkill based rating system with standard parameters. The prediction rates with TrueSkill are

set division	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	59.34	57.29	56.87	54.16	56.36	65.78
80-20	61.13	59.15	58.23	42.85	70.27	65.38
90-10	61.69	59.29	60.37	53.84	68.42	61.53

Table 12: Trueskill win prediction with $\Delta_{draw} = 0$

across all data sets, except "FIVE-STAR", at around 60% or higher. Dividing data sets to 70% training set and 30% has generally a bad influence on the TrueSkill rating performance. In the case of the "FIVE-STAR" data set TrueSkill is not able to make good prediction, independent of the set division. Especially good is the prediction rate for the high density data sets "ESL-EU" and "ESL-NA". In order to exclude match prediction with high uncertainty, where the scores being compared are very close to each other, the same evaluation phases are repeated with $\Delta_{draw} = 0.01$. Match results where the likelihood of one Team winning is at 0.5 plus minus $\Delta_{draw} = 0.01$ are therefore excluded from the win prediction ratio. Figure 13 shows the resulting prediction rates with $\Delta_{draw} = 0.01$ and the amount of matches skipped during evaluation expressed as the uncertainty in the second table. The evaluation results with $\Delta_{draw} = 0.01$ show a general

win prediction	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	60.35	57.94	56.88	35.71	58.82	65.78
80-20	62.05	**59.16**	58.19	42.85	**73.52**	66.66
90-10	63.06	**59.62**	62.46	33.33	68.42	61.53

uncertainty	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	9.23	9.72	13.31	41.66	7.27	0.00
80-20	8.57	9.52	15.30	0.00	8.10	7.69
90-10	8.49	6.82	10.24	30.76	0.00	0.00

Table 13: Trueskill uncertainty and win prediction with $\Delta_{draw} = 0.01$

trend of improvement of the win predictions. Performance on the "ALL" and "ESL-EU" data set increased and decreased on the "FIVE-STARS" data set. We can see, that by introducing the draw delta, the evaluation phase of TrueSkill excluded 41.66% of the "FIVE-STARS" data set with 70-30 division. On the other hand, when increasing the set for training to 80%, no matches are close enough in the likelihood of outcome to be excluded from evaluation. This is probably caused by the fact, that the inclusion of a few more of matches from 70-30 to 80-20 influenced the TrueSkill scores to spread more. That effect has been then reverted by extending the training set to 90% which introduced uncertainty again.
Overall are the uncertainty values inline with the Elo uncertainty values being mostly under 10%. With TrueSkill, the best prediction rate of 73.53% overall is on the "ESL-EU" data set with a 80-20 division. In General, TrueSkill seems to work good on data sets with a large amount of match results like "ALL", "ONE-STAR", and "TWO-STAR" and on data sets with high densities and evenly distributed results per team as in "ESL-EU" and "ESL-NA".

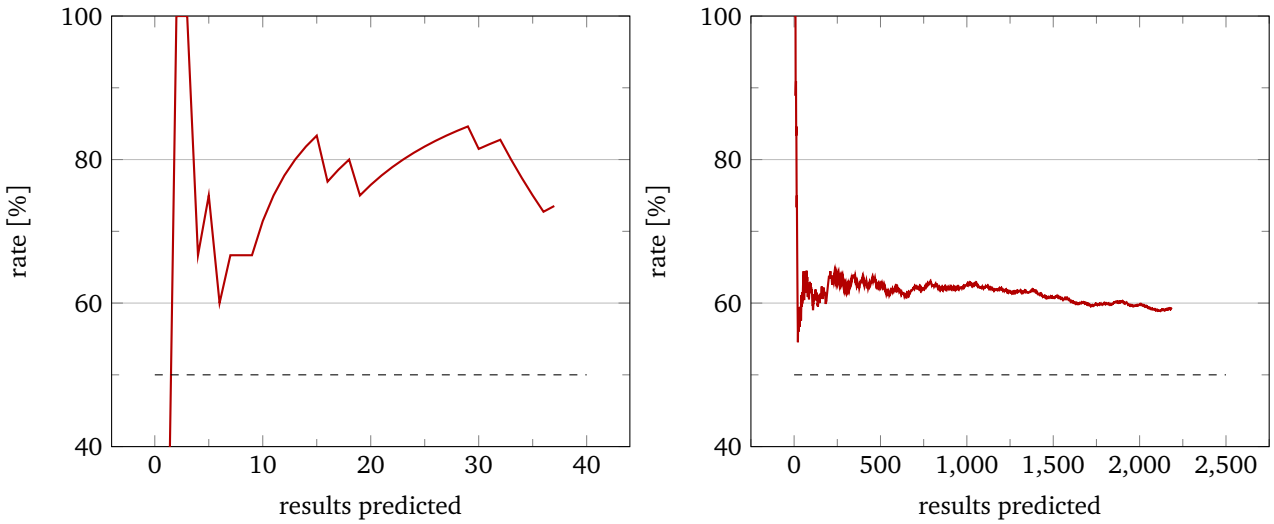


Figure 8: TrueSkill prediction rate with 80% ESL-EU and 80% ONE-STAR, $\Delta_{draw} = 0.01$

The TrueSkill rating system performs generally the best at data sets when taking 80% of the data set as the training set and 20% for evaluation.
We take a closer look at the prediction rate development for the two data sets "ESL-EU" and "ONE-STAR" where TrueSkill was able to achieve 73.52% and 59.16% respectively. The development of the win prediction, plotted by tracking the change in the win prediction rater after every evaluated match with $\Delta_{draw} = 0.01$ is displayed in Figure 8. In the case of the "ESL-EU" data set, a spike at 14 and 27 matches is observable, reaching a temporary win prediction rate of over 80%. During the further evaluation of more matches the second spike is followed by a drop to 73%. Both spikes are preceded by a constant increase, where the TrueSkill rating was able to predict a lot of matches correctly in a row. As in

the case of the Elo rating, this might be caused by team performance dynamics in a league environment, where at the beginning of a season most matches are won by the experienced teams. This assumption is a theory and would require further investigation of the league specific data sets.
The win prediction rate development during the evaluation of the "ONE-STAR" data set shows in the window of the first 500 evaluated map results a slightly better performance. After the evaluation of 500 matches the prediction rate starts to decrease until the final value of 59.16%. Therefore TrueSkill is decreasing, as Elo, in prediction accuracy the more time passes between the last match of the training set and the match being evaluated.

4.3 OD-Rating Performance

The performance of the OD-Rating, Table 14, on the selected data sets is mostly around 50% with the exception of the sets "FIVE-STARS" and "ESL-NA". In the case of those two data sets the OD-Rating is able predict the winner of evaluated matches with an accuracy significantly over 60%, in the case of 90-10 "FIVE-STARS" and 80-20 "ESL-NA" even 69%. The introduction of a Δ_{draw} to filter uncertainty in the evaluation phase is not needed for the OD-Rating due to wide distribution of the assigned team scores. Deploying a $\Delta_{draw} = 0$ yielded no significant increase in accuracy.
The OD-Rating is very well suited for approximating skill in league environments like the ESL Pro League, or environments with a very small skill distance as in "FIVE-STARS". This is easily explained with the formula of the OD-Rating being a sum of ratings of every participating team. In data sets with very low density as "ALL", "ONE-STAR" and "TWO-STAR" the created OD-Matrix contains many empty entries. An uneven distribution of those entries across the participating teams causes a rating skew towards teams who have a large history of matches played against man different teams. This is most certainly the case in low density data sets and almost non existent in the "ESL-NA" data set. By looking at the "ESL-EU" data set, the previously mentioned correlation does not seem to be a guarantee for good prediction rates. Depending on how close teams score offensive and defensive points to each other, the OD-Rating system might not be able to predict results well. This leads to the assumption, that the teams participating in the european ESL Pro league, the "ESL-EU" data set, are closer in skill to each other than the north american counterpart "ESL-NA".

set division	ALL	ONE-STAR	TWO-STARS	FIVE-STARS	ESL-EU	ESL-NA
70-30	49.23	45.69	43.81	66.66	49.09	65.78
80-20	50.75	45.49	43.19	67.85	59.45	69.23
90-10	50.98	43.28	43.39	69.23	52.63	61.53

Table 14: OD-Rating win prediction

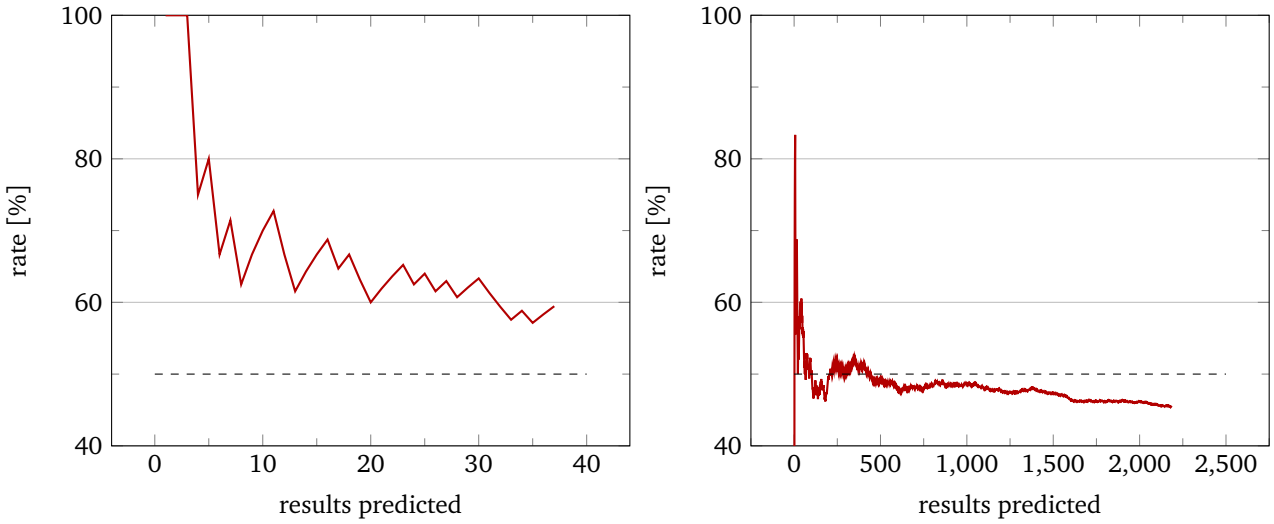


Figure 9: OD-Rating prediction rate with 80% ESL-EU and 80% ONE-STAR, $\Delta_{draw} = 0.0$

The division of data sets into 80% training and 20% evaluation is the most promising for the OD-Rating system, since both prediction rates for "ESL-EU" and "FIVE-STARS" are close to 70%. Lastly, we take a look at the prediction rate development over time for 80-20 "ESL-EU" and "ONE-STAR", data sets where the OD-Rating achieved a good rate of 59.4% and bad performance of 45.4% respectively. Figure 9 illustrates both developments.
The win prediction rate on the "ESL-EU" data set was at 73% after predicting 10 matches and decreases consequently after this. This constant trend downwards can explain why choosing a 70-30 division for the OD-Rating results in a

significantly worse prediction ratio. The OD-Rating seems to be very sensitive in regards to the time between the date of the last match in the training set and the date of the evaluating match. The more time passed between the creation of the offense defense ratings and the match being evaluated, the worse the prediction gets with the OD-Rating system. A similar trend can be seen in the win prediction development of the "ONE-STAR" data set.

4.4 Rating Systems Comparison

Based on the separate evaluations of the rating system, we can summarize a few general aspects. The Elo rating system works generally well with a lot of training data, so the ratings can stabilize slowly with the amount of matches played. Extending Elo with the tracking and adjusting of a confidence value, TrueSkill works even better when trained with a lot of data. TrueSkill does not need the same amount of matches to gain stability in rating and further, does not destabilize in the even of unlikely match results. Losing as a very good player against a newcomer brings with the Elo rating a large rating loss, but in the case of TrueSkill not necessarily, if the confidence is high enough. In the experiment, both distribution based rating systems, Elo and TrueSkill, where not able to predict the "FIVE-STAR" data set at all. The amount of results and the narrow skill gap of the teams involved did not allow both rating systems to approximate team skills well. OD-Rating was able to assign well performing team score compared to Elo and TrueSkill, but fails in larger data sets with low density. Caused by the nature of the underlying OD-Rating formula, that system is suited very well for high density data sets with evenly distributed played matches like "ESL" data sets or "FIVE-STARS". Every introduced rating system was able to show a significant improvement relative to the baseline. A common disadvantage is the observation how time sensitive skill approximation can be. Each rating system decreased significantly in the amount of correct predictions the more time passed between the last match of the training set and the evaluating match. "ONE-STAR" indicated a mark of 500 games played for the rating systems where the win prediction accuracy started decreasing. Figure 10 and Figure 11 merges the win prediction development curves from the previous sections in one figure. The similarity between Elo and TrueSkill is clear, but not without differences. At the 500 to 600 results predicted mark in Figure 11 we can see the Elo rating system better predictions for a period of time than TrueSkill.
In order to visualize the relation between Elo rating and the TrueSkill rating better, four of the top five teams after the evaluation of "ESL-EU" and "ONE-STAR" have been selected. Figure 12 shows the Elo score development for those four teams, Figure 13 the TrueSkill score development, based on the "ESL-EU" data set. To put the effect of number of matches for training to display, Figure 14 displays the Elo score development on the "ONE-STAR" data set and Figure 15 the development of the TrueSkill score.

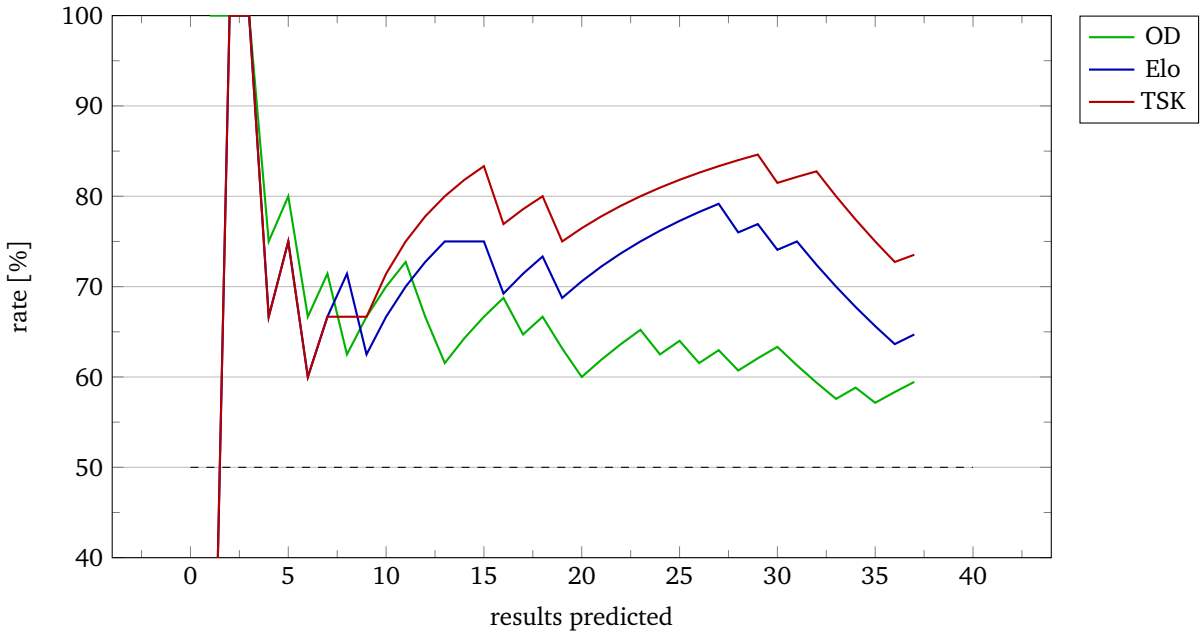


Figure 10: Win prediction development based on 80% of ESL-EU

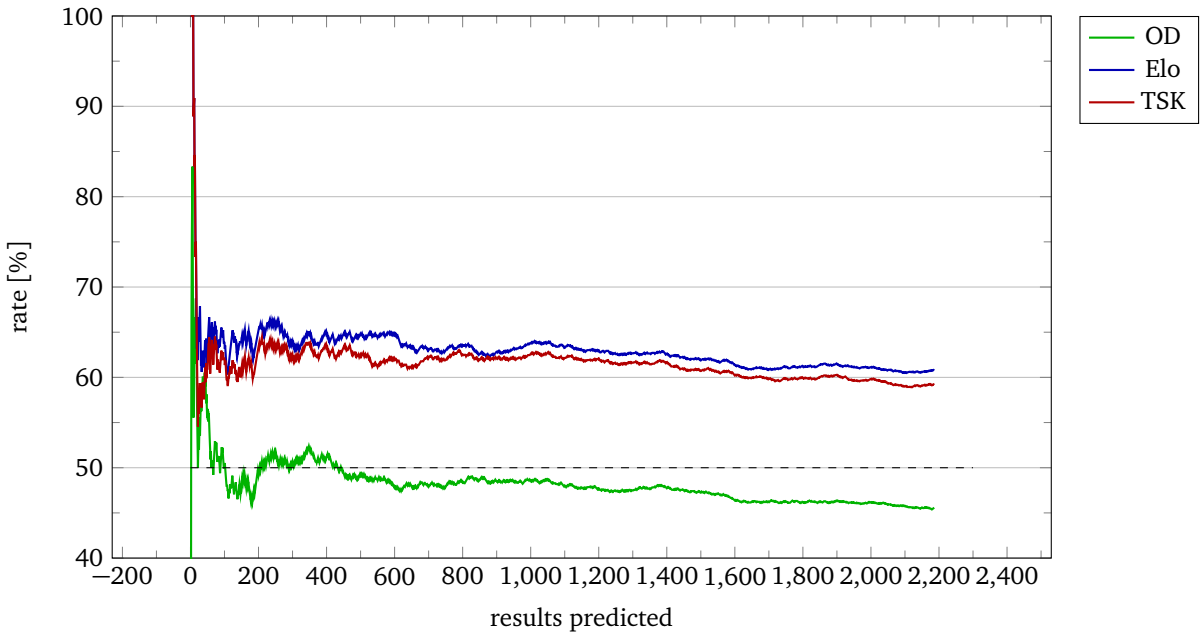


Figure 11: Win prediction development based on 80% of ONE-STAR

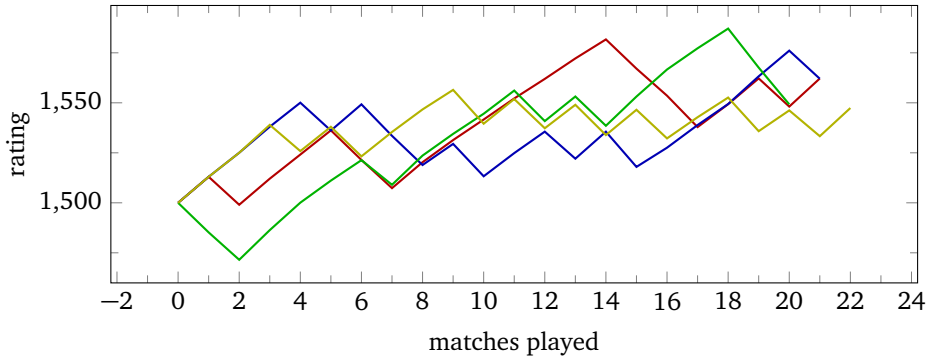


Figure 12: Elo rating development during training phase with 80% of ESL-EU

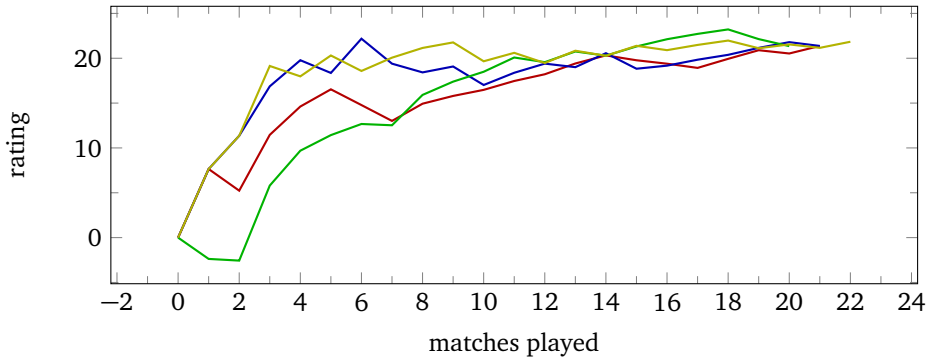


Figure 13: TrueSkill rating development during training phase with 80% of ESL-EU

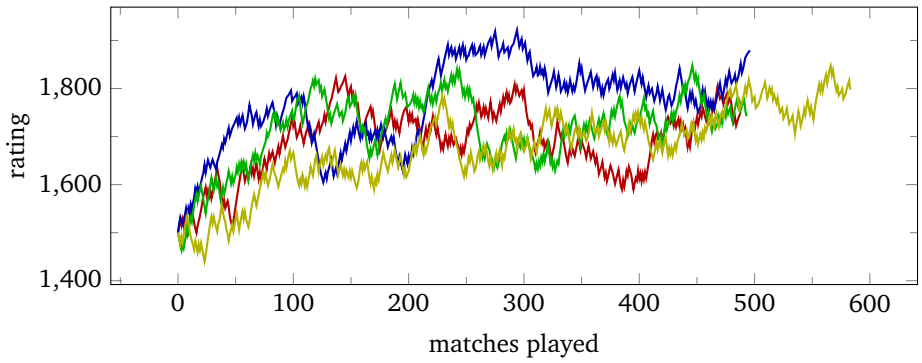


Figure 14: Elo rating development during training phase with 80% of ONE-STAR

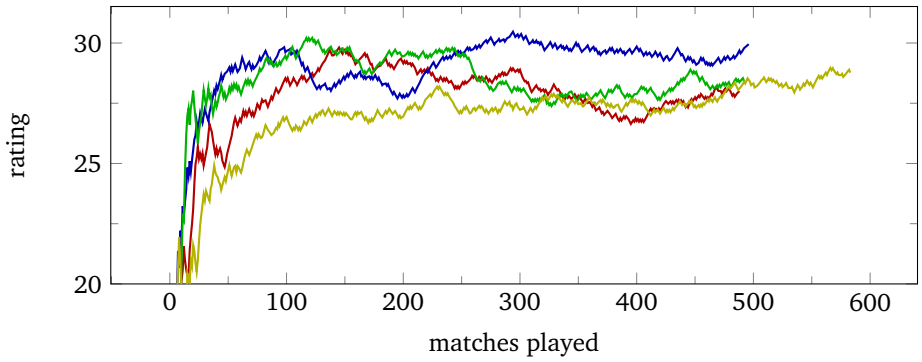


Figure 15: TrueSkill rating development during training phase with 80% of ONE-STAR

5 Conclusion

In this concluding section a brief overview of the thesis' findings are presented and secondly, opportunities for future work and improvement discussed.

5.1 Findings

The thesis' initial goal of researching, selecting and applying a set of rating system to a new application domain has been achieved. The conducted experiment offers more insight in how rating systems, Elo, TrueSkill and OD-Rating, can be applied to a data set of competitive eSports matches. Publicly available match data of the Counter-Strike Global Offensive competitive scene have been used to train three different predictors with pre-game knowledge in various data set environments.

A overall improvement of 10%-20% points compared to a baseline predictor with no knowledge has been shown across a selection of six different subsets. The Elo rating based predictor reached a maximum win prediction rate on the data set containing match results of a season in the north american ESL Pro League Season 7 with 66.66%. Microsoft's TrueSkill rating systems proves to be a general improvement to the Elo rating predictor concerning the win prediction performance on future matches. It performed 13% points better than the baseline predictor on the data set consisting of all recorded Counter-Strike Global Offensive matches since the game release in 2012. Further, the TrueSkill predictor was able to predict 73% of the 20% most recent matches of the european ESL Pro League Season 7, based on that season's results only. The OD-Rating based predictor achieved in data sets with high game density prediction rates up to 69%.
Advantages and disadvantages of the presented and implemented predictors have been evaluated based on the experiment's results. Generally, a division of 80% training set and 20% evaluation set proved to be the most promising regarding the performance in win prediction.

5.2 Future Work

Tuning the parameters of the underlying rating system of the predictors promises further improvement in predicting the outcome of future matches in competitive Counter-Strike Global Offensive. The Elo rating systems allows to tailor the formula to the competitive field it is applied to by adjusting the k-Value and ξ accordingly. TrueSkill offers more parameter tuning opportunities by having four adjustable parameters μ, σ, τ, β allowing to adjust TrueSkill to the target environment. The OD-Rating can be further improved by creating a regression model for weighting the offense and defense algorithm based on the difference in observed offensive and defensive scores. This can be used to determine how significant the offensive and defensive abilities of a team impact the resulting score and a better approach of constructing a final rating for comparison can be approximated. The automation of the parameter tuning process for each rating system is possible and would certainly yield interesting results.

Further the construction and processing of data sets can be improved by gaining deeper knowledge about the underlying data and collecting more specific pre match data. Data about the players in a team, their history and performance details have the potential to further improve the predictors. Demos of past matches can be collected and analyzed to derive performance features used to improve predictors and additionally implement a live predictor.

List of Tables

List of Figures

References

[1] Ipsos Connect. Essential facts about the computer and video game industry - sales, demographic and usage data. *Entertainment Software Association. Available online at: http://www.theesa.com/wp-content/uploads/2017/06/!EF2017_Design_FinalDigital.pdf*, 2017.

[2] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[3] Juho Hamari and Max Sjöblom. What is esports and why do people watch it? *Internet research*, 27(2):211–232, 2017.

[4] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill^{tm}: A bayesian skill rating system. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 569–576. MIT Press, 2006.

[5] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010.

[6] Nicholas Kinkade, L Jolla, and K Lim. Dota 2 win prediction. Technical report, Technical Report. tech. rep., University of California San Diego, 2015.

[7] Amy N Langville and Carl D Meyer. *Who's# 1?: the science of rating and ranking*. Princeton University Press, 2012.

[8] Jeff Moser. Computing your skill - the math behind trueskill. `http://www.moserware.com/2010/03/computing-your-skill.html`, 2010. Accessed: 2018-09-25.

[9] Maxwell Harris Perlman and Stefan Utamaru Alexander. Competitive gaming: Design and community building. 2014.

[10] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21, 05 1967.

[11] Michael G Wagner. On the scientific relevance of esports. In *International conference on internet computing*, pages 437–442, 2006.

[12] Yifan Yang, Tian Qin, and Yu-Heng Lei. Real-time esports match result prediction. *arXiv preprint arXiv:1701.03162*, 2016.