



Predicting Winning Team and Probabilistic Ratings in “Dota 2” and “Counter-Strike: Global Offensive” Video Games

Ilya Makarov^(✉), Dmitry Savostyanov, Boris Litvyakov,
and Dmitry I. Ignatov

National Research University Higher School of Economics, Moscow, Russia
iamakarov@hse.ru

Abstract. In this paper, we present novel winning team predicting models and compare the accuracy of the obtained prediction with TrueSkill model of ranking individual players impact based on their impact in team victory for the two most popular online games: “Dota 2” and “Counter-Strike: Global Offensive”. In both cases, we present game analytics for predicting winning team based on game statistics and TrueSkill.

Keywords: Game analytics · Rating systems · TrueSkill
Machine learning · Data mining · Counter-Strike · Dota 2

1 Introduction

eSport is a rapidly growing direction having the advantage over traditional sports [1]. While some people still don’t take it seriously, the viewership count records as well as prize pool records are being updated regularly during the biggest tournaments, reaching millions watching Dota 2 or Counter-Strike:GO. Both of the games are most popular games with over 500 000 humans playing simultaneously [2]. Such a popularity becomes a base for many betting companies oriented on eSport events [3].

It is worth mentioning, that eSport has a great advantage for game analytics over classic sport games providing structured information on the matches held online [4]. Many studies compared forecasting market with several rankings that appear not to be of high relevance to the real results [5, 6] due to limitations of information aggregation on the whole teams and bias of ratings with respect to different artificial measures. We aim to study on the quality of winning team prediction based on in-game analytics and individual rating systems. There are two aspects of sport analytics that we wanted to take into account in the current research: evaluation of players’ ratings system for match making in online games and predicting the match outcome.

The task of assessing the level of the game of individual players in team eSport online games is of high practical importance. This work is focused on the popular eSport multiplayer online battle arena (MOBA) discipline Dota 2.

The aim of the work is to develop a method for ranking players of one team on the basis of personal contribution to the victory in the match. The Bayes formula and logistic regression are used as the core idea of the ranking system. The ratio of the estimated probabilities of team victory in the match based on the information about each single player of this team and on the information about the team as a whole. The result of the work is a model that allows estimating the player's contribution to the team victory using the basic game indicators and their dynamics throughout the match. In addition, the model makes it possible to compare the importance of factors influencing the victory, and can be used for match making system [7].

We then focus on a dynamic match result prediction based on the large dataset of demorecords for the core championships in multiplayer first-person shooter (FPS) called "Counter-Strike: Global Offensive". In fact, Counter-Strike is one the most popular shooters in the world for more than 10 years. It was originally developed in 1999 by Minh Le and Jess Cliffe as a Half-Life modification before the title rights moved to Valve. This paper describes a data-driven approach to identify game actions that lead to winning or losing in a game round after the bomb was planted on defense maps in Counter Strike.

Moreover, Bayesian rating model called TrueSkill is evaluated for both, "Dota 2" and "Counter Strike: GO" video games, in order to compare specific aspects of game analytics for different game genres similar to the comparison in [8].

2 Related Work

Several attempts were made to understand the key features of successful playing multiplayer first-person shooter online games [9,10]. Most of the aspects under consideration were devoted to individual characteristics of human players, which sometimes are hard to measure [11]; moreover, these features may change over time. Several researchers try to evaluate statistic-based approaches of mining human behavior in FPS games [12,13] and MOBA games [14–16].

In practical applications for online games, it is important to create a rating system for the problem of matchmaking, when the game should adapt team members in order to have the prior probability of a certain team winning as 50%. Fairness of such a system in Dota 2 game was evaluated in [17,18], in general. For Counter-Strike we are the first to verify it. In what follows, we describe the individual and team ratings used in sport and eSport competitions.

2.1 Individual Ratings and Team Ratings

In many competitions, the organizers should compare players or teams while the players should be ranked in according to their results in the whole tournament. One of the first Bayesian rating system was the Elo rating developed for rating Chess tournament players [19]. The Elo rating was designed to provide unified ratings when there were no player who did not lose any match, which is the usual case in Chess tournaments. The idea for rating system could also be used

for matchmaking when we could see a battle of players with almost the same skill inspiring entertainment component of holding a competition.

Basically, the first ratings evolve under paradigm that one could compare several elements with respect to a certain number of simple properties, but could not compare all elements precisely and at once. For example, Bradley–Terry models [20] can be used for classification to multiple labels based on binary classification [21–24]. The comparison of the mentioned above Elo and BT ratings was presented in [25], while certain improvements of Elo system was published in [26, 27]. The application to sport ratings was presented in [28], in which the problem of learning rate over time was improved. Since the Elo rating invention, probabilistic rating systems have been generalized to handle team competitions with different team members between matches.

2.2 TrueSkill

The TrueSkill matching system was presented in [29]. TrueSkill is a Bayesian skill rating system which generalize the Elo rating used in Chess game [30, 31]. The presented system deals with an arbitrary number of competing teams and players. The main advantage of this system is that it can deduce individual skills from team results only. Despite it discards individual skills, the player is rated by the number of his impact on winning of his respective team. The system was evaluated in the “Halo 2” video game made by Microsoft [32].

The idea of this rating comes from ELO system that was adopted by many sports organizations around the world, including World Chess Federation FIDE [33]. The basic assumption of both rating systems is that players’ skill is a normally distributed random value. So players’ performance could slightly differ on different days but basic belief of the model is that it would be concentrated around some mean value. Two numbers, μ and σ , are used to describe the skill of each player: $skill_{team} \sim N(\mu, \sigma^2)$.

What differs TrueSkill from the ELO rating is that the former is adopted to work with any number of players in teams, including unequal teams. Another difference is that the ELO rating has a fixed value for the σ parameter while the TrueSkill algorithm generalizes the Elo rating by keeping track of two variables: the average skill μ and the system’s uncertainty about that estimate σ^2 [34]. These changes to the ELO system make TrueSkill more flexible, according to original research by Microsoft [29]. In matchmaking application for “HALO 2” video game, the lower bound $\mu - 3 \cdot \sigma$ was taken in order to stabilize skill learning policy for matching the players to opposing teams. Useful thing about both the ELO and TrueSkill rating models is that they allow to get winning probability for any two given teams in a direct way by simply subtracting two Gaussian random variables that stand for the opposing teams’ skill. They use Gaussian property that the sum/difference of two Gaussian random values is a Gaussian one: $P_1 \sim N(\mu_1, \sigma_1^2)$, $P_2 \sim N(\mu_2, \sigma_2^2)$, $P_{1-2} \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.

Basic idea of modeling opposing teams’ performance is to see what are the chances that performance of one team would be better than the others. That is exactly what P_{1-2} here describes. So, in case its value is greater than zero,

the model assumption is that team1 will win. In other case the winner is team2, according to the model. It means that $P(P_{1-2} > 0) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$. Note that because there are always two opposing teams and no draws in both, Dota 2 and Counter-Strike games, two-teams case only is considered in this paper. However, it should be noted that TrueSkill could also provide draw chances and account any number of participants both in terms of teams and players.

The TrueSkill model was improved to handle ties (originally, omitted) [35] and different team sizes [36] with a training algorithm based on Expectation Propagation [37]. Later work experimented with additional information on tournament results, taking into account for the match-score differences [38, 39].

2.3 Machine Learning Ranking and Applications

A machine learning approach was used in order to evaluate player skill based on mouse-tracking technologies [40]. While such an approach seems promising, it could not be easily generalized for different games and require significant amount of time in order to catch the dynamic of game statistics during real eSport competitions. Neural networks were also used for predicting Bradley–Terry rating [41]. The rating systems are quite useful when considering the application to football ratings [42] and multi-label classification problems [43, 44]. The application of TrueSkill algorithm was suggested for context ads click prediction on the Web [45]. In order to test predicting winning team based on individual impact of team players, we use the implementation of Trueskill rating made in [46, 47].

2.4 TeamSkill

The research on individual and team performance in FPS games was made in [48]. In [49], the authors presented a prediction for winning team using the Elo, Glicko, and TrueSkill ratings, called TeamSkill. In what follows, the authors introduced several weighted approaches and included previously inseparable game-specific features improving prediction quality for the teams [50]. Their new TeamSkill-EVMixed classification method based on the threshold of the prior probability of defeat outperforms all previous approaches in tournament environments, even when a number of matches is small. The suggested approach was evaluated on Halo 3 video game history and NBA records [51]. However, in our work, we follow the standard definition of team score as a sum of team players skills, which is sufficient to get high accuracy of team ratings.

3 Probabilistic Rating Systems for “Dota 2” Video Game

3.1 “Dota 2” Overview

“Dota” is a multiplayer online battle arena video game in which two teams of five players try to destroy the opponent “Ancient” construction while defending their

own. The defense of the Ancients uses real-time strategy controls, represented on a single map with shadowed regions in 3D-isometric perspective.

Each player choose one of the 113 playable characters, known as “heroes”, each having their own advantages and drawbacks. Heroes are divided into two primary roles, known as the “carry” and “support”. Carries are the weakest ones on the start, but they are able to become the powerful ones, thus leading their team to the victory. Support heroes lack the abilities of making heavy damage; their purpose is to utilize the available resources to provide assistance for their respective carries. The two teams — called Radiant and Dire — appear at the fortified bases places in the opposite corners of the map, which itself is divided in half by a crossable river and connected by three paths, called “lanes” [52]. The lanes are guarded by defensive towers constantly attacking opposing units within their range. A small group of weak computer-controlled creatures called “creeps” goes through the predefined paths along the lanes and tries to attack any opposing units met on its way. Creeps spawn with some period from the two buildings, called the “barracks”, that exist in each lane and are located within on the team bases. The map is permanently covered for both teams in the “fog of war”, which does not allow a team from seeing the opposing team’s actions and units if they are not directly in sight of a player’s team unit.

4 Predicting Winning Team for “Dota 2”

The first stage in Dota 2 video game is also known as a pick stage or draft. Players from different teams alternately pick 10 heroes and ban 10 heroes which can be useful to the enemy side, or they do not want to play. Draft stage prediction was considered when building heroes recommender system made by Kevin Conley and Daniel Perry. Their k-Nearest Neighbors (kNN) model results in 70% of accuracy on test data which consists of 50,000 matches and 67.43% accuracy on overall data cross-validation; the logistic regression classifier trained 18,000 matches and got 69.8% accuracy on the test set [53]. The improvements of predictions on draft stage was made in [54].

It appears that some heroes in Dota 2 become more useful in tandem with some other heroes. In [55], the authors built the logistic regression model obtaining 62% prediction accuracy on the test set. It was caused by high dimensionality of model’s input vectors combined with the complexity of using hero statistics represented by principal components. In [56], another group of authors studied the tandem idea: they worked with history of 6,000 matches, representing feature space as 50 vectors of 2 hero interactions and obtained overfitted model with the accuracy on the training set 72%, and only 55% on the test data. The authors of [57] included interactions among the heroes and pairwise winning rate for Radiant and Dire teams. Despite of the overfitting problem the Random Forest and Logistic Regression algorithm demonstrated 67–73% accuracy on the test set. In [58], the authors made a comparison of heroes statistics via Logistic Regression, Support Vector Machines, Gradient Boosting, and Random Forest. The last one showed the best result and after some parameters tuning it was

used as a final classifier with 88.8% test accuracy showing that in-game process information makes a great boost of an estimator's accuracy.

In order to properly distribute the impact of each role in winning team, we should make a proper mapping between team players and their positions [59]. Understanding positions in a "Dota 2" team was solved by machine learning algorithms [60], and it is of great importance for predicting winning team [61].

4.1 Problem Setting

Consider a set of matches $M = \{m_1, \dots, m_n\}$ and a set of teams $T = \{t_1, \dots, t_k\}$. The match involves two teams and there are only two possible outcomes, i.e. one of the teams won. Each team contains 5 players. Let $P_t = \{p_1, \dots, p_5\}$ be a set of players of team $t \in T$. Teams do not change the composition of players, and the player can not be in several teams. Players in teams are assigned to roles $R = \{r_1, \dots, r_5\}$, also denoted as $R = \{\text{Carry, Mid-Lane Solo, Hard-Lane Solo, Semi-Support, Full Support}\}$. Each member of a team performs one role throughout a match, moreover, he performs the same role in all matches. Thus, there is a one-to-one correspondence between the set of players of the given team and the roles $\forall t \in T : P_t \longleftrightarrow R$. The problem is to rank the players of the team on the basis of personal contribution to the victory in this match.

4.2 Contribution Function

Because of one-to-one correspondence $P_t \longleftrightarrow R$ there is no difference between ranking of players or roles of the team in the match. Let us denote the probability of winning of a team t in a match m as $P_{m,t}(w)$. Note also that the roles forms a set of pairwise disjoint events whose union is the entire sample space. As a result, the total probability law can be used in the following way:

$$P_{m,t}(w) = \sum_{i=1}^5 P_{m,t}(r_i) \cdot P_{m,t}(w|r_i).$$

By definition, all roles are equally probable $\forall i, j : P_{m,t}(r_i) = P_{m,t}(r_j) = \frac{1}{5}$. So,

$$P_{m,t}(w) = \frac{1}{5} \cdot \sum_{i=1}^5 P_{m,t}(w|r_i), \text{ and } \sum_{i=1}^5 \frac{P_{m,t}(w|r_i)}{P_{m,t}(w)} = 5.$$

The roles' contribution to the victory in the current match is determined as

$$C_{m,t}(r_i) = \frac{P_{m,t}(w|r_i)}{P_{m,t}(w)}, \text{ with mean value } \sum_{i=1}^5 \frac{C_{m,t}(r_i)}{5} = 1$$

This function allows us to compare the players of a team t in a match m based on the contribution to the victory of the roles they perform.

4.3 Role-Based Model Evaluation

The probability $P_{m,t}(w|r_i)$ can be estimated by using logistic regression model. Every single role r_i in a match m for a team t can be represented as a vector $\mathbf{x} = \mathbf{x}(m, t, r_i) = (x_1, \dots, x_l)$. The vector consists of components which reflects such in-game information as “Gold Earned”, “Damage Dealt”, “Used Hero”, etc. It is allowed to estimate the probability of win based only on information about current role/player in the following way:

$$P_{m,t}(w|r_i) = \sigma(\langle \beta, \mathbf{x} \rangle),$$

where β is a vector of parameters, $\sigma(t) = \frac{1}{1+e^{-x}}$ is a logistic sigmoid.

4.4 Experiments

We consider professional competitions for “Dota 2” during March–April 2017 from Opendota resource [62], containing participant of Kiev Major grand challenge. We choose the following features from the open stats table: amount of gold and experience earned by each player on the end of the match, logs for kills, purchases, and ward placements; we have collected over 5000 records and 150 features. All the data is split into five roles, for which individual models are trained separately. 5-fold cross-validation quality metrics are represented in Table 1.

Table 1. Quality metrics for role-based models in “Dota 2”

Role	AUC	F1	Recall	Precision	Accuracy
r_1 - Carry	0.98	0.93	0.92	0.93	0.93
r_2 - Mid-Lane Solo	0.97	0.93	0.93	0.93	0.93
r_3 - Hard-Lane Solo	0.96	0.90	0.91	0.90	0.90
r_4 - Semi-Support	0.97	0.90	0.91	0.91	0.90
r_5 - Full Support	0.96	0.90	0.92	0.89	0.90

In addition, to evaluate predictions quality, aggregated team skill are computed as a sum of individual role impacts in winning team $P_{m,t}(w) = \sum_{i=1}^5 P_{m,t}(w|r_i)$. For every match m between teams t_1 and t_2 the following rule is used: if $P_{m,t_1}(w) > P_{m,t_2}(w)$ then t_1 won, else t_2 won. On the other hand, TrueSkill can be used as a baseline for predicting winning team in the middle of the match bounding from below 0.92 accuracy of prediction model by the 0.72 accuracy based on TrueSkill only.

4.5 Discussion on “Dota 2” Results

Without corresponding model of roles impact in Dota 2, we try to evaluate our approach not only by predicting a winning team, but also using an expert opinion represented below. Let us consider the match between OG and EG teams during Kiev Major tournament [63]. The role distribution is given by EG and OG teams.

In Table 2, the information on the end of the match for EG team is shown. We could see the players in positions 4 and 5 farm greater amount of gold than expected, while the relations between kills, deaths, and assists is better than for positions 1 and 3, meaning the lack of performance of the latter players. Player 2 performs well in terms of all the characteristics except damage to the opponent buildings, which should be one of the main priorities for positions 1 and 2, thus reducing performance of the second player. The described in-game analysis is well predicted, which is shown in Table 2 with a small drawback of decreased influence of role 1. As for OG team, players 4 and 5 performed well, but spent much gold to buy together 37 sentry wards, which reduces their impact for supporting computer-controlled players with lack of efficiency. Player 2 performed well in all the aspects except he died many times. Player 3 has less damage than player 5 while player 5 should be support class, thus reducing self-performance (Table 2). The lack of impact for player 1 who made 40% damage of the hole team can be described by the great number of deaths and the lack of damage to the opponent buildings.

Table 2. Estimated contribution to the victory

Team	Role	Player	Contribution	Team	Role	Player	Contribution
EG	Cr1t	5	0.35	OG	2	ana	0.28
	zai	4	0.34		4	JerAx	0.26
	Suma1L	2	0.26		5	Fly	0.23
	Universe	3	0.04		3	s4	0.15
	Arteezy	1	0.01		1	N0tail	0.08

5 Predicting Winning Team in Counter-Strike

5.1 Counter-Strike Overview

Counter-strike is a type of a game that is called tactical first-person shooter. The gameplay is based on shooting your opponents and trying to kill them looking from a first-person perspective. However, the game also provides a great diversity on its strategical and tactical parts because the competitive matches during world tournaments are played between two teams of five players each.

What differs tactical shooter from a DeathMatch or “Free-for-all” gaming mode is that a gaming location (called a map) has some specific goal. Achieving that goal leads to a victory in a round. Killing all of your opponents usually leads

to a win too, but sometimes the map goal can be reached even if everyone from a team was killed in a round. In competitive Counter-Strike maps, the goal of “Terrorists” team is to plant the bomb at some specific location called BombSite and prevent the Bomb from being defused. After the bomb is planted there is a 35–40 s countdown. The opposing team called “Counter-terrorists” aims to prevent bomb planting or defuse it in a limited amount of time after plant, otherwise they lose a round after the bomb explodes. In both cases, the way of winning by killing all the players from the opposite team is also a way to win the round. Teams play rounds repetitively until the end of the 15th round when the teams switch sides and continue playing until one of the teams would have 16 rounds won in total. In case of 15–15 score a few additional rounds can be played to define a winner if needed. We consider the model of predicting a winning team based on after-plant in-game situations. After the bomb is planted, the 40-s countdown is started. All of the situations are split into 1-s time intervals, for which we try to measure the winning chances.

In order to build prediction model we, first, have downloaded game replays, which are further used for loading game attributes with the self-made parser based on OpenSource project by StatsHelix [64] and have extracted raw game data into a .csv file. Using Google’s Protocol Buffers as a message/object serialization language we parse the original .dem files for the game records from the world changes in a sequential way [65]. The dataset covers the last four years of demorecords from the storage [66]. A C# wrapper is built over the demoinfo tool to get the raw data regarding game events, such as: kills, shots, movements, player coordinates and view directions, and several descriptive statistics over them. Most of the chosen features are related to after-plant situations with respect to the round goal, i.e. to explode/defuse the planted bomb, depending on a team: the number of players alive, difference in the current team sizes, the total and average equipment cost, the number of damaged and healthy players, the smoke cover on bomb plant, the total number of different grenades, the total TS of a team, and TS prediction on a winner. 162 demos in total have been harvested to feed the after-plants model. It was noticed that most of these games were played during the period October 2016–January 2017.

6 Experiments

6.1 Metrics Used

We use TrueSkill judged based on accuracy and Log-Loss, while metrics for prediction of winning team with after-plant feature analysis using Decision Trees and Logistic Regression were taken as accuracy and Log-Loss. In case of binary classification problem one could imagine a naive prediction model of a fair coin tossed for every prediction, which should be worse than god prediction model. For a fair coin toss, Log-Loss on average is close to 0.7, while average accuracy is exactly 50%.

Table 3. Comparing ratings

data_train	data_test	acc	logloss
all games	all games	0.62	0.675
dust2 only	dust2 only	0.59	0.69
all games	dust2 only	0.57	0.75

6.2 TrueSkill for Winning Team Prediction in Counter-Strike

Each player is given with a personal rating. We go through the list of played games and refresh the ratings after each game. Different pre-learn periods are tested on a test sample of 6 different months. 9 months period is chosen for maximizing accuracy or the prediction models, which stabilizes on 0.68 value with 0.61 Log-Loss change. We choose learning TrueSkill from 2016-02 until 2016-09, with just 3 games from a set played in September, placing a gap before the test dataset. The prediction period chosen is in between 2016-09-28 and 2017-03-11. Three combinations of train/test datasets based on all the games or de_dust_2 only games were used during evaluation. The results are shown in the Table 3.

7 Discussion and Future Work

We have considered the application of machine learning techniques for predicting winning team for the two most popular online games. The results obtained show that the quality of prediction is higher when we check the in-game parameters closer to the round end. We use TrueSkill rating system to measure baseline for the prediction model when we want to step back from the end of the match and have a prediction on game features that should not have less accuracy than prediction based on a Bayesian probabilistic rating system. We are looking forward to compare our system with other rating systems and improve TrueSkill model [39] in order to take into account the role impact distribution during the game round for both video games.

Acknowledgments. The work was supported by the Russian Science Foundation under grant 17-11-01294 and performed at National Research University Higher School of Economics, Russia. We would like to thank Alexander Semenov and Petr Romov for their piece of advice.

References

1. Taylor, T.: Raising the Stakes: E-sports and the Professionalization of Computer Gaming. MIT Press, New York (2012)
2. Powered by Steam: Steamcharts. An ongoing analysis of steam’s concurrent players (2017). <http://steamcharts.com/>. Accessed 09 May 2017

3. Kaytoue, M., et al.: Watch me playing, i am a professional: a first study on video game live streaming. In: Proceedings of the 21st International Conference on WWW, NY, USA, pp. 1181–1188. ACM (2012)
4. Wagner, M.G.: On the scientific relevance of eSports. In: International Conference on Internet Computing, pp. 437–442 (2006)
5. Luckner, S., Schröder, J., Slamka, C.: On the forecast accuracy of sports prediction markets. In: Gimpel, H., Jennings, N.R., Kersten, G.E., Ockenfels, A., Weinhardt, C. (eds.) Negotiation, Auctions, and Market Engineering. LNBP, vol. 2, pp. 227–234. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-77554-6_17
6. Tsai, M.: Fantasy (e)Sports: the future prospect of fantasy sports betting amongst organized multiplayer video game competitions. UNLV Gaming LJ **6**, 393 (2015)
7. Zhang, L., et al.: A factor-based model for context-sensitive skill rating systems. In: 2010 22nd IEEE International Symposium on TAI, vol. 2, pp. 249–255 (2010)
8. Coulom, R.: Whole-history rating: a Bayesian rating system for players of time-varying strength. In: van den Herik, H.J., Xu, X., Ma, Z., Winands, M.H.M. (eds.) CG 2008. LNCS, vol. 5131, pp. 113–124. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87608-3_11
9. Dick, M., Wellnitz, O., Wolf, L.: Analysis of factors affecting players' performance and perception in multiplayer games. In: Proceedings of 4th ACM SIGCOMM IW on Network and System Support for Games, NY, USA, pp. 1–7. ACM (2005)
10. Wright, T., Boria, E., Breidenbach, P.: Creative player actions in FPS online video games: playing counter-strike. Game Stud. **2**(2), 103–123 (2002)
11. Rioult, F., Métivier, J.P., Helleu, B., Scelles, N., Durand, C.: Mining tracks of competitive video games. AASRI Procedia **8**, 82–87 (2014)
12. Hladky, S., Bulitko, V.: An evaluation of models for predicting opponent positions in first-person shooter video games. In: 2008 IEEE International Symposium on CIG, pp. 39–46 (2008)
13. Bird, A.M.: Development of a model for predicting team performance. Am. Alliance Health Phys. Educ. Recreat. **48**(1), 24–32 (1977)
14. Drachen, A., et al.: Skill-based differences in spatio-temporal team behaviour in defence of the ancients 2 (dota 2). In: 2014 IEEE GME, pp. 1–8 (2014)
15. Pobiedina, N., et al.: On successful team formation: Statistical analysis of a multiplayer online game. In: 2013 IEEE 15th International Conference on Business Informatics, pp. 55–62 (2013)
16. Yang, P., Roberts, D.L.: Knowledge discovery for characterizing team success or failure in (A)RTS games. In: 2013 IEEE International Conference on CIG, pp. 1–8, August 2013
17. Wu, M., Xiong, S., Iida, H.: Fairness mechanism in multiplayer online battle arena games. In: Proceedings of 3rd International Conference on SAI (ICSAI), pp. 387–392, November 2016
18. Myślak, M., Deja, D.: Developing game-structure sensitive matchmaking system for massive-multiplayer online games. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8852, pp. 200–208. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15168-7_25
19. Elo, A.: The Rating of Chessplayers, Past and Present. Arco Pub., New York (1978)
20. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
21. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res. **5**, 975–1005 (2004)
22. Huang, T.K., et al.: Generalized Bradley-Terry models and multi-class probability estimates. J. Mach. Learn. Res. **7**, 85–115 (2006)

23. Fujimoto, Y., Hino, H., Murata, N.: An estimation of generalized Bradley-Terry models based on the em algorithm. *Neural Comput.* **23**(6), 1623–1659 (2011)
24. Matsumoto, I., et al.: Online density estimation of Bradley-Terry models. In: *Proceedings of International Conference on Learning Theory*, Paris, France, pp. 1343–1359. PMLR (2015)
25. Király, F.J., Qian, Z.: Modelling Competitive Sports: Bradley-Terry-Elo Models for Supervised and On-Line Learning of Paired Competition Outcomes. *arXiv preprint arXiv:1701.08055* (2017)
26. Glickman, M.E.: *The Qlicko System*. Boston University, Boston (1995)
27. Glickman, M.E.: *Example of the Qlicko-2 System*. Boston University, Boston (2012)
28. Glickman, M.E., Hennessy, J., Bent, A.: A comparison of rating systems for competitive women's beach volleyball. <http://www.glicko.net/>
29. Herbrich, R., Minka, T., Graepel, T.: TrueskillTM: a Bayesian skill rating system. In: *Proceedings of the 19th International Conference on NIPS*, MA, USA, pp. 569–576. MIT Press (2006)
30. Graepel, T., Herbrich, R.: Ranking and matchmaking. *Game Dev. Mag.* **25**, 34 (2006)
31. Dangauthier, P., Herbrich, R., Minka, T., Graepel, T., et al.: Trueskill through time: revisiting the history of chess. In: *NIPS*, pp. 337–344 (2007)
32. Huang, J., et al.: Mastering the art of war: how patterns of gameplay influence skill in halo. In: *Proceedings of the SIGCHI International Conference*, NY, USA, pp. 695–704. ACM (2013)
33. Wikipedia: Fide world rankings - wikipedia, the free encyclopedia (2017). https://en.wikipedia.org/w/index.php?title=FIDE_World_Rankings&oldid=776755738. Accessed 5 May 2017
34. Moser, J.: Computing your skill (2010). <http://www.moserware.com/2010/03/computing-your-skill.html>. Accessed 9 May 2017
35. Nikolenko, S., Sirotkin, A.: A new Bayesian rating system for team competitions. In: *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pp. 601–608 (2011)
36. Nikolenko, S.I., Sirotkin, A.V.: Extensions of the trueskilltm rating system. In: *Proceedings of the 9th International Conference on AFSSC*, pp. 151–160. Citeseer (2010)
37. Bishop, C.M.: Pattern recognition. *Mach. Learn.* **128**, 1–58 (2006)
38. Nikolenko, S.I., Serdyuk, D.V., Sirotkin, A.V.: Bayesian rating systems with additional information on tournament results. *Trudy SPIIRAN* **22**, 189–204 (2012)
39. Nikolenko, S.: A probabilistic rating system for team competitions with individual contributions. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) *AIST 2015. CCIS*, vol. 542, pp. 3–13. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_1
40. Buckley, D., Chen, K., Knowles, J.: Rapid skill capture in a first-person shooter. *IEEE Trans. Comput. Intell. AI Games* **9**(1), 63–75 (2017)
41. Menke, J.E., Martinez, T.R.: A Bradley-Terry artificial neural network model for individual ratings in group competitions. *Neural Comput Appl.* **17**(2), 175–186 (2008)
42. Tarlow, D., Graepel, T., Minka, T.: Knowing what we don't know in NCAA football ratings: understanding and using structured uncertainty. In: *Proceedings of the 2014 MIT Sloan Sports Analytics Conference (SSAC 2014)*, pp. 1–8. Citeseer (2014)
43. Lee, J.-S.: TrueSkill-Based pairwise coupling for multi-class classification. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) *ICANN 2012. LNCS*,

- vol. 7553, pp. 213–220. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33266-1_27
44. Naik, N., et al.: Streetscore-predicting the perceived safety of one million streetscapes. In: Proceedings of the IEEE International Conference on CVPR Workshops, pp. 779–785 (2014)
45. Graepel, T., et al.: Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft’s bing search engine. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 13–20 (2010)
46. Hamilton, S.: Pythonskills: implementation of the trueskill, glicko and elo ranking algorithms (2012)
47. Lee, H.: Python implementation of trueskill: the video game rating system (2013)
48. Shim, K.J., et al.: An exploratory study of player and team performance in multiplayer first-person-shooter games. In: 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 3rd International Conference on Social Computing, pp. 617–620, October 2011
49. DeLong, C., Pathak, N., Erickson, K., Perrino, E., Shim, K., Srivastava, J.: TeamSkill: modeling team chemistry in online multi-player games. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011. LNCS (LNAI), vol. 6635, pp. 519–531. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20847-8_43
50. DeLong, C., Srivastava, J.: TeamSkill evolved: mixed classification schemes for team-based multi-player games. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012. LNCS (LNAI), vol. 7301, pp. 26–37. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30217-6_3
51. DeLong, C., Terveen, L., Srivastava, J.: TeamSkill and the NBA: applying lessons from virtual worlds to the real-world. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in SNA and Mining, NY, USA, pp. 156–161. ACM (2013)
52. McDonald, T.: A beginner’s guide to dota 2: Part one - the basics (2013). <https://www.pcinvasion.com/a-beginners-guide-to-dota-2-part-one-the-basics>. Accessed 25 July 2013
53. Conley, K., Perry, D.: How does he saw me? A recommendation engine for picking heroes in dota 2. Np, nd Web 7 (2013)
54. Semenov, A., Romov, P., Korolev, S., Yashkov, D., Neklyudov, K.: Performance of machine learning algorithms in predicting game outcome from drafts in dota 2. In: Ignatov, D.I., Khachay, M.Y., Labunets, V.G., Loukachevitch, N., Nikolenko, S.I., Panchenko, A., Savchenko, A.V., Vorontsov, K. (eds.) AIST 2016. CCIS, vol. 661, pp. 26–37. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52920-2_3
55. Agarwala, A., Pearce, M.: Learning dota 2 team compositions. Technical report, Stanford University (2014)
56. Song, K., Zhang, T., Ma, C.: Predicting the winning side of dota2. Technical report, Stanford University (2015)
57. Yang, Y., Qin, T., Lei, Y.H.: Real-time esports match result prediction. arXiv preprint [arXiv:1701.03162](https://arxiv.org/abs/1701.03162) (2016)
58. Johansson, F., Wikström, J.: Result prediction by mining replays in dota 2 (2015)
59. Inkarnate: Dota 1-to-5 system (2012). <http://www.liquiddota.com/forum/dota-2-strategy/454943-dota-1-to-5-system>. Accessed 05 Sep 2011
60. Eggert, C., Herrlich, M., Smeddinck, J., Malaka, R.: Classification of player roles in the team-based multi-player game dota 2. In: Chorianopoulos, K., Divitini, M., Hauge, J.B., Jaccheri, L., Malaka, R. (eds.) ICEC 2015. LNCS, vol. 9353, pp. 112–125. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24589-8_9

61. Pobiedina, N., et al.: Ranking factors of team success. In: Proceedings of the 22 International Conference on World Wide Web, NY, USA, pp. 1185–1194. ACM (2013)
62. Powered by Steam: Stats from professional dota 2 matches (2017). <https://www.opendota.com/explorer>. Accessed 01 May 2017
63. DotaBuff: Kiev major: team eg vs. team og (2017). <https://www.dotabuff.com/matches/3148721353>. Accessed 16 May 2017
64. StatsHelix: Cs:go demos parser by statshelix (2014). <https://github.com/StatsHelix/demoinfo>. Accessed 9 May 2017
65. Valve: csgo-demoinfo (2014). <https://github.com/ValveSoftware/csgo-demoinfo/tree/master/demoinfogo>. Accessed 5 May 2017
66. HLTV.org: Hltv.org demos section (2017). hltv.org/?pageid=28. Accessed 9 May 2017