Key Factors to Dota2 Win Rate

Yang Song
New York University
New York, United States
ys3603@nyu.edu

William Deng
New York University
New York, United States
wnd211@nyu.edu

Abstract—

DOTA 2 is the most popular multiple-player online battle arena game. It is very meaningful for players and teams to understand the key factors contribute to the win rate of a DOTA 2 match. In this paper, we present analytics on two large datasets of matches records. We used Hadoop MapReduce and Pig to clean and profile the dataset, and use Hive queries to explore the factors we proposed. In our exploration, we found that team composition and counter picking are the most important factors to the win rate. We further analyze the factors and provide several actionable insights that enable new strategies for picking/banning heroes that improves overall win rate. These new strategies are shown to be consistent with those from game guides.

Keywords—analytics, Dota2

I. Introduction

Defense of the Ancient 2 (DOTA 2) is the most popular multiplayer online battle arena (MOBA) video game. A match is played by two teams of five players, with each team occupying and defending their own separate base on the map. Each of the ten players independently controls a powerful character, known as a "hero", who all have unique abilities and differing styles of play. During a match, players collect experience points and items for their heroes to successfully defeat the opposing team's heroes in player versus player combat. A team wins by being the first to destroy the other team's "Ancient", a large structure located within their base [1]. Players select their hero during a pre-game drafting phase, where they can also discuss potential strategies and hero matchups with their teammates [2].

In this paper, we analyze 2 large datasets on DOTA 2, and identify key factors that contribute to win rate. In section 2, we explain the motivation of this analytic and review 4 publications related to DOTA 2 win rate prediction and analytics. In the design detail section, we discuss our approach to this problem using a design diagram. In section 3, we introduce the 2 datasets we used in this project in detail. By exploring the factors, we found the win rate of a DOTA 2 is highly related to the composition of heroes. Good composition taking consideration of both teammates and the enemy team's hero selection is more likely to win the match. In section 4, we report our experiment process using Big data tools (Hive). We show that the insights retrieved from the datasets are consistent with our hypothesis. In section 5, we briefly discuss how we improve/expand our analytic in the future. Finally, in section 6, we prove the goodness/validity of our result using online analysis from [9] and show that our insights can be substantiated using game-based reasoning from [10][11] with a specific example.

II. Motivation

DOTA 2 is known for having a very steep learning curve. It takes hundreds of games to learn just most of the mechanics as well as the usage of each hero. Since on average each game takes about 35-40 minutes, hundreds of games is a substantial time commitment, for just getting used to the game. As a result, many people lose interest before getting to that point, due to losing too much, running out of patience to learn everything, etc.

One way to remedy this problem is to lessen the steep learning curve, by providing ways for new players to improve their win rate, without extensive knowledge of the game. To tackle this, we decided to apply analytic tools to explore possible factors that affect the win rate and extract useful insight that can be used to guide new players to improve their win rate. These insights can be used to create recommender systems to help new players make better game choices, and improve their game experience, without going through the entire learning curve.

III. Related Work

Some rather interesting researches have been done in this area. In Dota2 Win Prediction [3], Kinkade and Lim analysis post-match data and hero selection data of 62,000 matches. In the exploratory analysis, they report that some of the heroes can have a win rate of 60% while some of the others only have win rate of 36%. Furthermore, Kinkade and Lim implemented two predictors with logistic regression and random forest classifier with the two datasets. They claim that the accuracy of their prediction can reach over 99% on the post-match data. However, the feature they use for the predictors includes a plethora of key information at the end of matches (advantage in gold and experience is almost sure to result in winning). In our analytics, we avoid the use of this kind of post-match data and explore other interesting information extracted in the drafting phase/early stage of the matches. For the hero selection data, Kinkade and Lim state that Logistic regression can prevent the predictor from overfitting thus gives a better result for the test set. They report that their predictor can reach the accuracy of 73% given information only from the drafting phase. Their results proved that the result of the match is highly related to drafting, and thus provides good indicators on which factors might be important to win rate.

In Outcome Prediction Using Machine Learning Methods [4], Wang et al. introduce new machine learning methods, including Logistic Regression, SVM and Random Forest to predict the outcome of the match with both drafting information and real-time data collected as the match progressed. In the paper, they report in detail the progress of collecting data of high-level matches and approaches they used to filter the dataset in order to make the dataset more representative. After that, Wang et al. discuss the 17 features they extract from the dataset, and the 3 classifiers they implemented based on these features. They claim that their best classifier can reach the accuracy of 72.9% with all 17 features. This result is very close to the one made by Kinkade et al.

Another machine learning method used for win prediction is an adopted Elo rating system [6]. In this paper [6], data from about 1.1 million matches, collected using Valve's API, is used to train an Elo rating system to predict win rate. The system assigns an "Elo Rating" to every player, and uses the collective Elo Ratings from both teams to predict the win rate. The Elo rating for each player is updated according to the games participated, using a logistic regression model. However, the results showed that this system was unable to predict the win rate (AUC from ROC curve was only 0.58, which is very close to the baseline of 0.5). This suggests that using players as the only covariate in determining Elo Rating is insufficient in determining win rate. To address this, the authors suggest using hero selection and other in-game variables as other covariates to improve the prediction.

Non-machine learning methods have also been used to try to predict win rate for specific teams. One such method is by using an Analytical Hierarchy Method (AHP) [5], which assigns an "importance" rating to possible factors in the game and use them to predict win rate. The factors explored in Prediction of DOTA 2 Match Result by Using Analytical Hierarchy Process Method [5] are experience per minute (XPM), gold per minute (GPM), matchmaking ratio (MMR, a rating for how "good" a player is), result for 3 matches against the opposing team, and the result of the last 10 matches. They used data mining techniques to extract the required features from online data sets from DotaBuff, a popular DOTA 2 analytic website, and from DOTA 2's official website. Using AHP, they were able to achieve an accuracy of 75%, which is consistent with other related works.

IV. Design and Implementation

A. Design Details

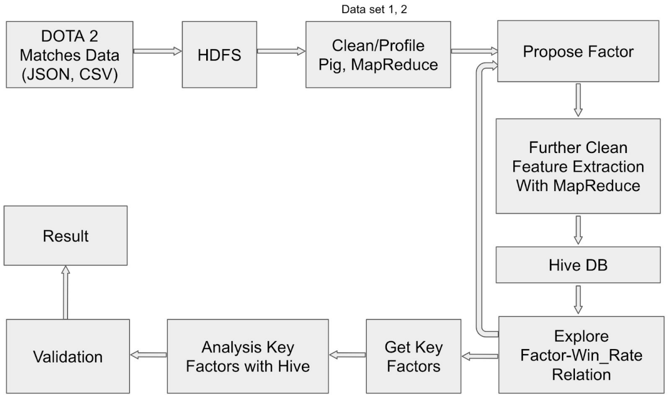

Data set 1, 2
DOTA 2 Matches Data (JSON, CSV)
HDFS
Clean/Profile Pig, MapReduce
Propose Factor
Further Clean Feature Extraction With MapReduce
Hive DB
Explore Factor-Win_Rate Relation
Get Key Factors
Analysis Key Factors with Hive
Validation
Result

Figure 1: Design diagram

For this analytic, we used two data sets for DOTA 2 match data, one in zipped JSON format and one in CSV format. The 2 datasets are downloaded from their respective sources and loaded into HDFS on Dumbo cluster. Data cleaning and profiling for both datasets are done via MapReduce, to create comma-delimited output files, one line per match. Then we propose the factor we think may have a relation with win rate, further clean (remove outliers) and extract the feature we need from the cleaned data. To do this, we parse the nested JSON objects and only save used features. The output file is then loaded into Hive for analysis. In order to make sure we cleaned the data properly, we did preliminary profiling on the data, and compare it with other DOTA 2 data sources [9]. Once we verified that the data is consistent with respect to our analytic, we continue onto a detailed analysis of each factor. To do that. we explore the relationship between each factor and win rate through visualization, and attempt to extract meaningful insights from data pattern. We repeat this process to identify the key factors and further analyze them with Hive to get actionable insights.

In order to validate our results, we will run our analytic on a separate data source, to make sure the numbers we calculated on both data sets are consistent. We then check other online resources [9] as a secondary validation. Once we verify that our results are consistent with other data, we use our results to extract useful insights. The insights will then be cross-checked with empirical evidence (guides by experienced players), to make sure that there exist concrete game-based explanations for them.

V. Datasets

A. DOTA2 Data Dump [7]

The dataset has a size of 3.5GB and is in CSV format. The dataset is a sample of the Dota2 matches dump from March 2011 through March 2016. Each row of the file represents a match, each row has fields such as match_id, radiant_win, duration, lobby_type, human_players, and a json object contains basic information (player_id, hero_id, and slot number) about each of the players in the 10 slots. For both datasets, there is a JSON file mapping from hero_id to hero_name.

B. YASP 3.5 Million Data Dump [8]

The YASP Data Dump is a JSON file, and has size 99.8 GB (zipped) and size 458 GB (unzipped). The data consists of a huge JSON array, with each element being a JSON Object representing data from a match. The entire data set contains data for matches from yasp.co as of December 2015 (about 3.5 million matches). Each match JSON Object contains data of that match, including hero selection, gameplay events, gold/experience, player data, duration, and game result. The schema is very complex, so for this analytic, we reduced the schema to better fit our scope of analysis. The schema we use for this data set includes game mode, duration, gold/experience advantage at 10 minutes, and the game result (for the match and for each player). The game result for each player includes their hero selection, gold per minute, experience per minute, total damage, kills, and assists.

VI. Results

Our data analytics is meant for normal games in DOTA 2's public matchmaking. This eliminates "fun" modes and other game modes that operate under a different set of assumptions. To do this, we restricted "game_mode" attribute to all pick, ranked, and captain's mode, which is all normal/competitive game modes. However, the preliminary analysis led to unexpected results which deviated dramatically from other sources of DOTA 2 analytic [9]. Thus, we investigated the data further and found out that two other attributes: "lobby_type" and "leaver_status" were required to eliminate anomalies from the data set. This is because some of the games happened in private lobbies, which had Developer Console enabled, and is used for testing purposes. "leaver_status" is used to indicate whether a player has left the game. Because having a player leave leads to drastic changes to that game, we decided to further remove all games where any of the 10 players have left before the game ended. After this additional layer of filtering, we were able to verify our results with other sources [9].

To explore early game gold and experience advantage for each team, we explore the range and split the possible values into 10 bins, from -4000 to 4000, with a difference of 1000.

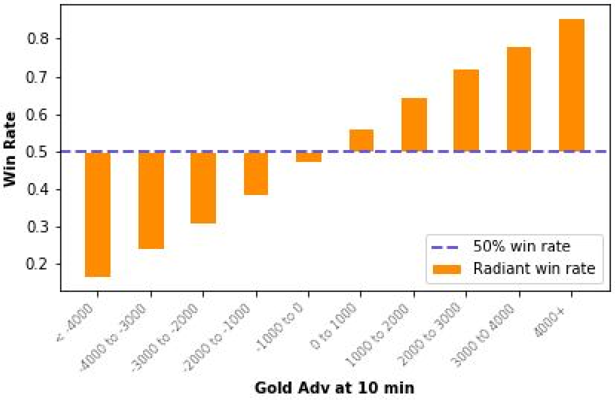


Figure 2: Gold Advantage vs. Win Rate

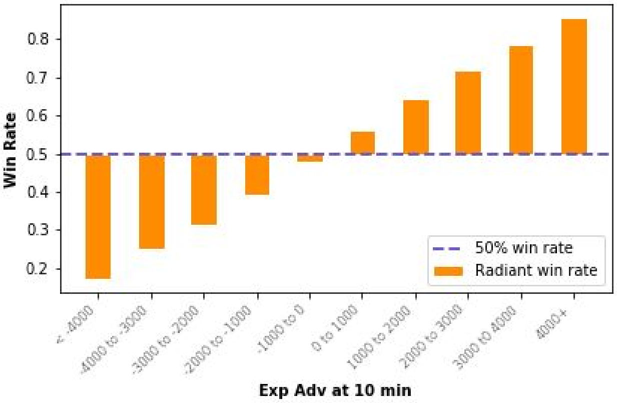


Figure 3: Experience Advantage vs. Win Rate

The results are shown in Fig 2 and 3. Based on the figures, we can conclude that a positive gold and experience advantage is required to have a higher than 50% win rate. The more gold and experience advantage the team have in the early stage, the more likely it can win the match at the end. As shown in the figures, the win rate is almost linear to the advantage value for both gold experience. The result suggests the player obtain more resources in order to win the game. However, the result is too trivial to find more meaningful patterns. And we were unable to extract more insight from analyzing this factor.

To explore a hero's popularity with respect to win rate, we queried the Hive table with hero pick information and plotted for each hero its pick rate (percentage of games the hero was selected by either team), against its overall win rate. The result is shown in Fig. 4. In the figure, we can tell that most of the heroes have win rate between 45% to 55%. The most popular hero is Windranger. With a pick rate of 29.3%, she has a win rate of 48.6%. However, the graph showed no viable correlation between pick rate and win rate.

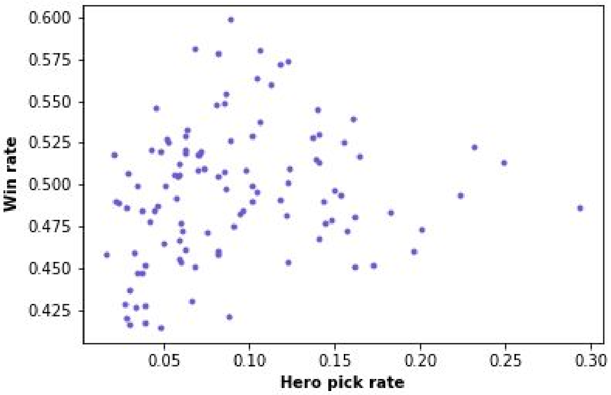


Figure 4: Hero Pick Rate vs. Win Rate

Team composition is a complicated factor, and thus we divided it into two parts: teammate composition and enemy composition. Teammate composition analyzes how having certain heroes on the same team affects win rate, and enemy composition analyzes how having certain heroes as enemies affect win rate. To simplify the analysis, we explored each composition as hero pairings.

The primary focus of this analytic is to investigate how "drafting", or hero selection, impacts the game's outcome. To investigate hero composition, we queried data for win rates of different hero combinations in pairs, as well as their individual win rates. Some of the best pairings are shown in Fig. 2. The data from the figure shows that specific hero pairings have high "compatibility". This means they have a much higher win rate when they are on the same team vs. when they are by themselves. On the contrary, there are also bad hero pairings, some of which are shown in Fig. 8. These heroes have bad compatibility, so when they are on the same team their win rate is worse than by themselves.

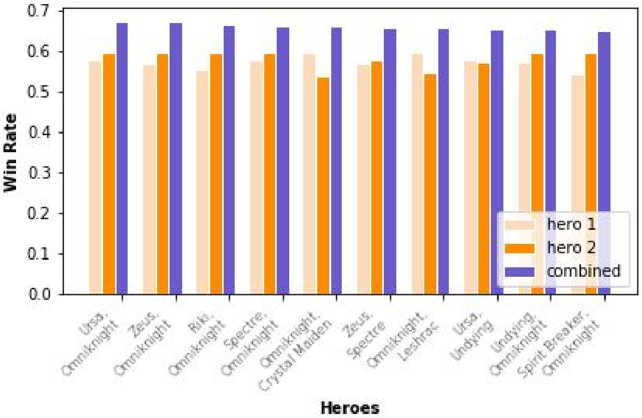


Figure 7: Heroes vs. Win Rate (Highest 10)

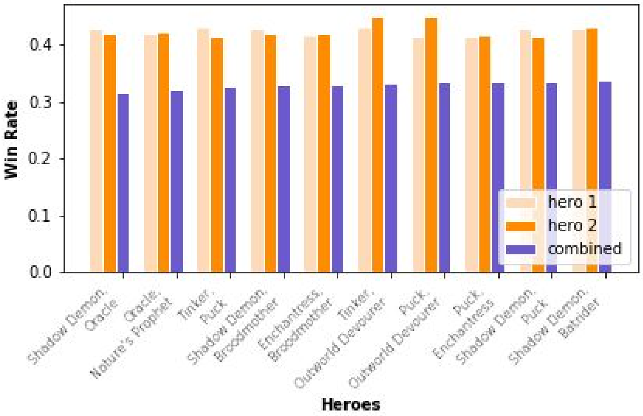


Figure 8: Heroes vs. Win Rate (Lowest 10)

Fig. 7 and Fig. 8 show that depending on which hero pairs are together, the win rate can vary drastically, by up to +/-10%. This result verifies our hypothesis that team composition is vital to the game outcome. These results provide insight on methods to improve win rate by identifying heroes that work well together, which enables drafting strategies.

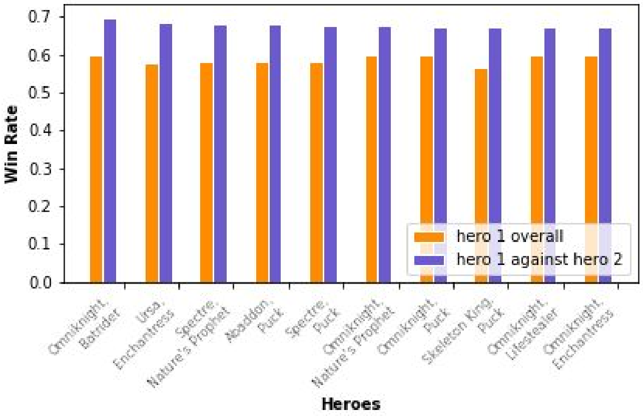


Figure 9 Heroes (Different teams) vs. Win Rate

A hero's win rate may also depend on the enemy team picking. Fig. 9 shows the highest 10 win rates of "hero 1" when "hero 2" is in the enemy team, some of the hero's win rate is up by 10%. This means that hero 1's skills and characteristics can best restrain hero 2's performance. In contrast, when hero 1 is in the enemy team, hero 1's win rate would be lower. The result proves that enemy composition is also an important factor to a game's outcome.

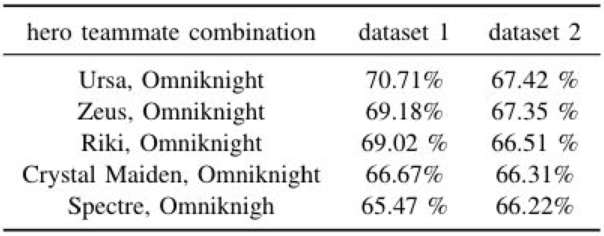
Actionable insights can be made from this result in the form of drafting strategies. Our analytic provides win rates of combinations for both one's team and one's enemy. Thus, we can use our data to select/narrow down hero choices that statistically lead to the better game outcome. In DOTA 2, drafting is done in stages. Thus, during a player's pick phase, our analytic can proceed to calculate a pool of heroes that work well with the heroes already picked on this team, and another pool of heroes that counter those picked on the enemy team. Furthermore, we can also create pools of heroes that work badly with the current team and are countered by the enemy team. Thus, our analytic provide new insight on which heroes are likely to be good, and which heroes are probably not very good for winning. This is useful for new players, since it allows them to make good choices, without having prior knowledge on all the heroes.

VII. Future Work

Due to time constraints, we had to select just a couple of features, based on preliminary analysis. However, there are a lot of other possible features in match data that could be important. One example is item usage. There are over a hundred items in the game, and it takes a lot of time to learn their mechanics, when to buy them, and which hero should buy which items. In the future, we would isolate the item data for each hero for each match, and then compute win rate with respect to each hero-item combination. This can enable new insights into which items on which heroes leads to higher win rate, and be used in recommendation systems to recommend the best items for their picked hero to new players.

VIII. Conclusion

The results show that team composition is vital to win rate, and our analytic provides insight on how to pair teammates/enemies in order to improve win rate. To verify our results, we ran our analytic on data set 1 and achieved similar results. Some sample data are shown in Table 1.



hero teammate combination	dataset 1	dataset 2
Ursa, Omniknight	70.71%	67.42 %
Zeus, Omniknight	69.18%	67.35 %
Riki, Omniknight	69.02 %	66.51 %
Crystal Maiden, Omniknight	66.67%	66.31%
Spectre, Omniknigh	65.47 %	66.22%

Table 1 Comparison Between Results from 2 Datasets

Furthermore, we compared our results to other DOTA 2 analytic sources [9] and found similar numbers. In order to verify that our analytic produces useful insights, we decided to apply our analytics to a random hero to figure out what are the best/worst teammate/enemy heroes. The top 5 in each category are shown in Fig. 10, 11, 12, 13. We then look for empirical evidence from popular game guides to support these insights.

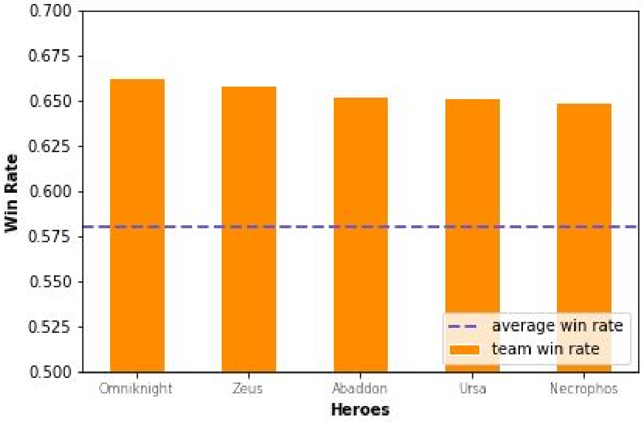


Figure 10: Heroes (Same Team) vs. Win Rate (Highest 5)

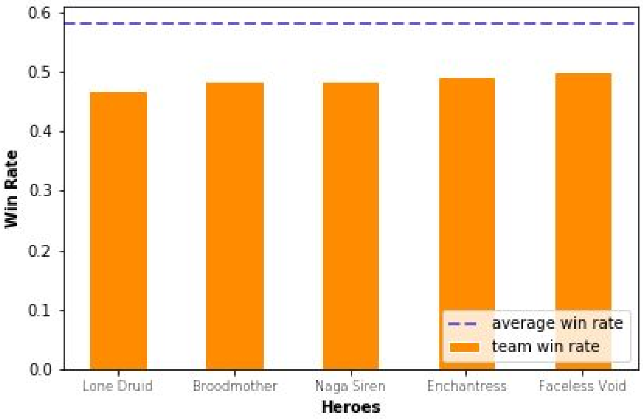


Figure 11: Heroes (Same Team) vs. Win Rate (Lowest 5)

The hero we chose is "Spectre". Based on guide from the official wiki [10][11], this hero is very weak early game. However, in the late game, it is one of the strongest tank (hard to kill) heroes. It is good at catching motility heroes (hard to pin down), and damage heroes (heroes who deal high damage, but has low health). It is weak against heroes who are strong in the early game. The hero works well with good supporting heroes (heroes who can help her during the early game), and bad with other heroes who need resources (these heroes would be contesting with Spectre for the limited resources in game).

Fig. 12 shows which heroes Spectre would team up with for highest win rate, based on our analytics. Out of these heroes, Necrophos, Abaddon, and Omniknight are strong early game supports. Zeus and Ursa are strong mid-game heroes but do not need a lot of resources. This shows that our analytic was able to provide good insight into best teammate combinations for Spectre.

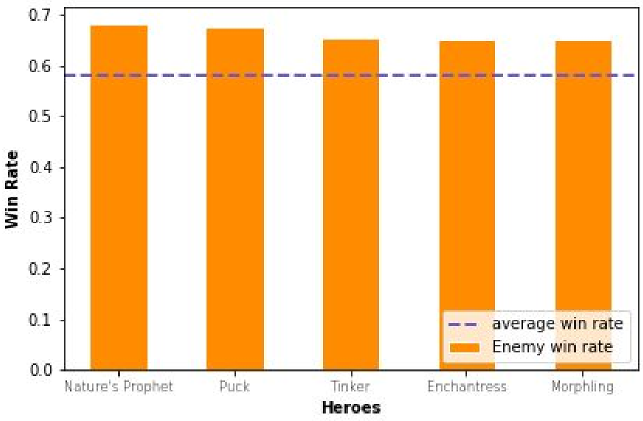


Figure 12: Heroes (Different Team) vs. Win Rate (Highest 5)

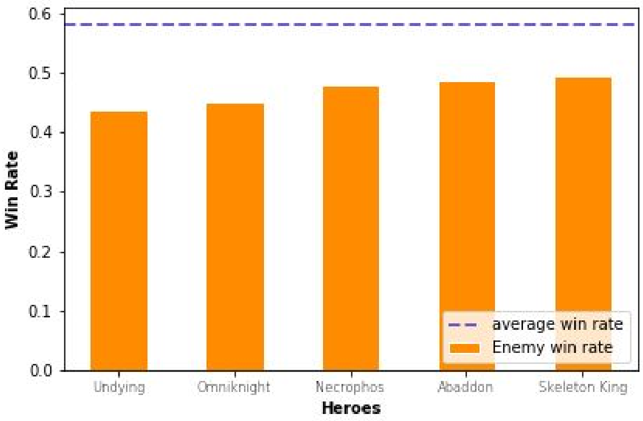


Figure 13: Heroes (Different Team) vs. Win Rate (Lowest 5)

Fig. 13 shows which heroes Spectre should not team up with. Out of these heroes, Lone Druid, Faceless Void, Naga Siren, and Broodmother are all resource-hungry heroes, who will contest with Spectre for these resources. Enchantress is a support but does not have good synergy with Spectre. Again, our results are consistent with the game guide.

Fig. 14 shows which heroes Spectre are good against. All 5 of these heroes are damage heroes, with low health. Again our result is consistent with the game guide.

Fig. 15 shows which heroes Spectre are bad against. All 5 of these heroes are strong early game heroes. Our result is consistent with the game guide.

This example analysis shows that our results are valid, and provides true and actionable insights for hero composition, that is consistent with empirical evidence from game guides [9][10].

Thus, our analytics can be used to provide insights on hero selection for new players, so they can select heroes that can maximize their win rate, without having to learn the characteristics of all the heroes. This reduces the learning curve and can improve the game experience for new players. Furthermore, our results are shown to be consistent with game guides, which means they have strong game-based reasoning to support them. The procedure of our analytic can be used to build real-time recommendation systems for new players, and generate good hero recommendations without the need to understand all the game mechanics.

Acknowledgment

Thanks to Gabin for DOTA 2, YASP/OpenDota for parsing match data from Valve's API, and NYU HPC for providing Big Data Tools for analyzing the data.

References

1. Dota 2 wikipedia, https://en.wikipedia.org/wiki/Dota_2
2. Kim, Ben. "A comprehensive comparison of Dota 2 and League of Legends". PC Gamer. Archived from the original on August 1, 2016. Retrieved August 3, 2016.
3. N. Kinkade and K. Lim. Dota2 Win Prediction
4. N. Wang, L. Li, L. Xiao, G. Yang, Y. Zhou, Outcome Prediction of DOTA2 using Machine Learning Methods.
5. G. Aryanata, P. Rahadi, Y. Sudarmojo, Prediction of DOTA 2 Match Result by Using Analytical Hierarchy Process Method, IJEET, Vol. 2 No. 1, June 2017
6. P. Kostic, M. Berensten, Predicting the outcome of Dota 2 Matches using Elo
7. The OpenDota Project, https://blog.opendota.com/2017/03/24/datadump2/
8. yasp.co data dump, http://academictorrents.com/details/5c5deeb6cfe1c94404 4367d2e7465fd8bd2f4acf
9. DotaBuff Hero Win Rate, https://www.dotabuff.com/heroes/winning
10. Dota Wiki Spectre guide, https://dota2.gamepedia.com/Spectre/Guide
11. Dota Wiki Spectre guide, https://www.dotabuff.com/heroes/spectre