Human Team Behavior and Predictability in the Massively Multiplayer Online Game WOT Blitz

FRANK EMMERT-STREIB, Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Finland
SHAILESH TRIPATHI, Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Finland, and Production and Operations Management, University of Applied Sciences Upper Austria, Austria
MATTHIAS DEHMER, Department of Computer Science, Swiss Distance University of Applied Science, Switzerland, College of Artificial Intelligence, Nankai University, China, Department for Biomedical Computer Science and Mechatronics, UMIT–Private University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Austria, and Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Finland

Massively multiplayer online games (MMOGs) played on the Web provide a new form of social, computer-mediated interactions that allow the connection of millions of players worldwide. The rules governing team-based MMOGs are typically complex and nondeterministic giving rise to an intricate dynamical behavior. However, due to the novelty and complexity of MMOGs, their behavior is understudied. In this article, we investigate the MMOG World of Tanks Blitz by using a combined approach based on data science and complex adaptive systems. We analyze data on the population level to get insights into organizational principles of the game and its game mechanics. For this reason, we study the scaling behavior and the predictability of system variables. As a result, we find a power-law behavior on the population level revealing long-range interactions between system variables. Furthermore, we identify and quantify the predictability of summary statistics of the game and its decomposition into explanatory variables. This reveals a heterogeneous progression through the tiers and identifies only a single system variable as key driver for the win rate.

CCS Concepts: • **Mathematics of computing** → **Probability and statistics**; • **Information systems** → **World Wide Web**; • **Computing methodologies** → **Machine learning**; **Modeling and simulation;**

Additional Key Words and Phrases: Massively multiplayer online games, prediction, statistical model, complex system, human behavior, computational social science

Authors' addresses: F. Emmert-Streib, Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland, 33100; e-mail: v@bio-complexity.com; S. Tripathi, Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland, 33100, and Production and Operations Management, University of Applied Sciences Upper Austria, Steyr, Austria, 4400; M. Dehmer, Department of Computer Science, Swiss Distance University of Applied Science, Brig, Switzerland, 3900, College of Artificial Intelligence, Nankai University, Tianjin, China, 300071, Department for Biomedical Computer Science and Mechatronics, UMIT–Private University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Tyrol, Austria, 6060, Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland, 33100.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2023 Copyright held by the owner/author(s).
1559-1131/2023/10-ART5 $15.00
https://doi.org/10.1145/3617509

ACM Reference format:
Frank Emmert-Streib, Shailesh Tripathi, and Matthias Dehmer. 2023. Human Team Behavior and Predictability in the Massively Multiplayer Online Game WOT Blitz. *ACM Trans. Web* 18, 1, Article 5 (October 2023), 27 pages.
https://doi.org/10.1145/3617509

1 INTRODUCTION

Advances in information technology and the World Wide Web enable online and real-time communication among users worldwide. One form of such communication is provided by **Massively Multiplayer Online Games (MMOGs)** [1, 15, 18]. MMOGs allow the connection of players who are locally separated to play a multitude of different games online. Popular examples for MMOGs with millions of active users are World of Warcraft, Guild Wars 2, Lord of the Rings Online, World of Tanks, and The Elder Scrolls [80]. From a scientific perspective, it has been realized that MMOGs allow the addressing of many different questions from a variety of fields because large amounts of data are available for a quantitative analysis [4]. Put simply, one can utilize an MMOG as a representation of a virtual world from which every aspect can be recorded and monitored. That means an MMOG is like a virtual Petri dish that can be studied to learn about social, behavioral, psychological, or economic phenomena [13, 60, 72]. Instances for previously studied questions include the network structure of gold farming or user interactivity networks [44, 70], the scaling behavior of online users (avatars), game servers or behavioral action sequences [41, 76, 78], cooperation and team formation [63, 79], and the psychology of users [8].

In this article, we study the MMOG World of Tanks (WOT) Blitz. WOT Blitz was developed by Wargaming and released in 2014. According to Wargaming, the game is played by 100 million players worldwide. The mobile game is available for tablets, smartphones, and PCs for Windows 10, Android, and iOS. WOT Blitz is a competitive strategy game between two teams, each with seven players. The game has a similar play style as dodgeball, where 'ball hits' are exchanged via 'shots' causing a certain amount of damage. Different battle types are available (regular, rating, or realistic), as well as a number of different maps that can either be played in encounter or supremacy mode. The goal of the game is essentially either to destroy all vehicles of the opponent or to capture the base(s).

For investigating WOT Blitz, we use a data science [64] and complex adaptive systems [39, 69] approach for analyzing data on the population level to gain insight into the organizational behavior of the system. Specifically, we utilize information about 403 vehicles (tanks) from 10 tiers (corresponding to difficulty levels) for which average summary statistics from 24,015 players and more than 2.62 million matches are available. Importantly, the rules of the game allow only the pairing of vehicles in matches that are either from the same tier or from adjacent tiers (e.g., from tier 4 and tier 5). That means vehicles from further apart tiers never meet in-game.

There are several reasons for selecting WOT Blitz for our study. First, despite the immense popularity of WOT Blitz, so far there are no studies investigating this game. In fact, the only study based on WOT Blitz we found investigates in-game chat and cyberbullying [58]. In contrast, in our study, we do not use information from the in-game chat at all. Second, the available data in the form of summary statistics of vehicles are quite challenging because they hide many details (e.g., about the performance of individual players or the teams). This is intriguing because it allows us to learn what can still be revealed beyond individual vehicles. Third, the available summary statistics of vehicles are robust. This is due to a feature of the game that limits the duration of a match to 7 minutes; however, a game can end even earlier. In combination with its popularity, this

allows us to gather information about the outcome of millions of games upon which the summary statistics of vehicles are based. Fourth, although WOT Blitz is not a sports game itself, it has characteristics thereof because two teams are playing against each other to win. That means the underlying mechanism that drives everything is human team behavior. All these reasons together make our study unique and provide the first steps toward deeper insights into this complex game.

In this article, we investigate the organizational principles of WOT Blitz and its game mechanics. For this, we study three major levels. First, we study the scaling behavior of the population of vehicles. This allows us to identify long-range interactions of system variables. Second, we identify and quantify the predictability of performance measures by classifying the population of vehicles. For this, we will study the hierarchical clustering of vehicles based on averaged vehicle characteristics. That means we define profile vectors of vehicles based on averaged vehicle characteristics estimated from thousands of individual matches. Performing such an analysis in dependence on the tiers allows us to reveal differences and similarities of different tier stages. This translates the abstract finding about the scaling behavior of the population of vehicles into concrete realizations with a focus on output variables. Third, we quantify the importance of covariates on the predictability by a regression framework utilizing the method LMG (Lindeman, Merenda, and Gold) [47]. LMG is a method for decomposing the coefficient of determination, R2, which can be estimated for linear regression models [32]. This will provide information about the probabilistic contribution of individual predictor variables. Hence, this analysis step focuses on the input variables on which an output variable (in our case, a performance measure) depends, and their contribution on the prediction. In general, this decomposition allows us to identify the covariates that are most important in predicting the output variable. For our analysis, we perform such a decomposition in dependence on different tiers allowing us to gain insight on the progression of the game across the offered difficulty levels.

From the description of our analysis follows a number of potential limitations. First, we would like to note that not all studies about MMOGs allow the same level of insights. For instance, in the work of Burt [11], players of the game Everquest were investigated. This allowed them to link observable data in the MMOG via the structure of their collaboration networks as they perform quests, with real-world constructs that are of general importance, including personality. To accomplish such an insightful analysis, the collaboration networks of individual players had to be analysis. In contrast, in our work, we study only data about average game statistics of vehicles that do not allow us to draw conclusions about individual players. Still, the game is played by humans and also our data are the result from human team behavior. Second, despite the generic name *massively multiplayer online games*, there are considerable differences among MMOGs. For instance, major game types one can distinguish are role-playing, real-time strategy, first-person shooter, or simulations [59], whereas WOT Blitz is of the latter type. It is clear that the game type has an impact on what can be learned, and this is also the case for our study. Third, we are not only limited to data corresponding to summary statistics of vehicles, but these data are obtained from observations, not interventions. That means the data do not result from controlled experiments specifying, for example, well-defined situations or conditions but are obtained from ordinary games.

This article is organized as follows. In Section 2, we discuss our research questions and contributions. In Section 3, we present the methods and data used in this work. In Section 4, we present findings about general characteristics of the data, scaling-behavior of system variables, detection of predictability, and quantifying the importance of predictor variables. This is followed by a discussion and interpretation of our findings in Section 5, which also places these into a broader context. Finally, we present a brief summary and concluding remarks in Section 6.

2 RESEARCH QUESTIONS AND CONTRIBUTIONS

In this article, we analyze data from the MMOG WOT Blitz. For this, we use summary statistics of vehicles. These data provide information about the average performance of 16 features of vehicles that are used by players in two teams to compete in matches. That means human team behavior is the underlying mechanism that drives everything observable in the summary statistics of vehicles. It is worth highlighting that no information about individual players including their performance is used for our analysis.

The goal of this study is to address the following five research questions based on the available data:

(1) Does WOT Blitz show signs of critical behavior?
(2) Are the vehicles among the tiers balanced?
(3) Is WOT Blitz (partially) predictable?
(4) What features are most dominant?

To 1.: The first question aims to establish a connection between WOT Blitz and a class of systems that exhibits self-organizing behavior [3]. This is important because it has been shown that self-organizing systems have a self-tuning property that allows them to reach stable working conditions. Put simply, this is an expression of a functional system. We will establish this connection by identifying power-laws between system variables as a result of local interactions. In our case, this corresponds to the competing vehicles, controlled by human players, in matches.

To 2.: The second question relates to the fact that according to the rules of the game, vehicles of adjacent tiers are selected for a match. Considering that the tiers represent the difficulty levels of the game, this requires a fair balancing among the tiers, as without such a balancing, vehicles of a higher tier would be always stronger and everyone would only want to play the highest tier to guard against sure losses.

To 3.: Since the goal of any game is to win, the win rate of vehicles provides information about the most important aspect of a game, namely the outcome of matches. The third question studies the predictability of the outcome of games by using features of vehicles. Considering the rules of the game that pairs vehicles of adjacent tiers and the need for a balancing of vehicles among tiers, we study this for each tier because the results should also be a reflection of those constraints.

To 4.: For predicting the behavior of a system, typically a high-dimensional representation is needed to achieve good results. In general, the problem with such high-dimensional representations is that they lack interpretability. For this reason, we aim to find a low-dimensional approximation that provides a good representation for outcome variables of the game. For our analysis, we use the LMG method [47], which allows a decomposition of predictor variables for quantifying their importance. From this, we identify the most dominating features of predictions.

Overall, our study makes the following contributions. First, to the best of our knowledge, the game WOT Blitz has not been studied before. Hence, we step on new ground regarding an understanding of essentially any aspect of this complex game, including the predictability of outcome variables, balancing of vehicles among tiers, and the progression through them. Given the popularity of WOT Blitz, which is played by millions of players worldwide, this is surprising because it is not a niche game but a well-known contender for MMOGs.

Second, one important aspect of our study is power-law distributions because they allow a connection to self-organizing systems. An example of another study that investigated power-law distributions is that of Chun et al. [17]. In this work, the game Aion was studied, which is a **Massively Multi-Player Online Role-Playing Game (MMORPG)**, and power-law distributions of network features were found. A similar study for various games was conducted by Kirman et al.

[46], all revealing power-law distributions of network features. In contrast, in the work of Jiang et al. [41], the MMORRG Legend of Mir was studied, and power-law distributions for the online avatar numbers revealed corresponding to the number of active players. Yet a different power-law was studied by Pittman and GauthierDickey [62]. In their work, it was shown that the spatial distribution of players in World of Warcraft follows a power-law distribution. This shows that also regarding the studied power-laws, our study is different because we investigate them neither for networks nor active players or spatial positions of players but for measured in-game features that characterize the performance of vehicles controlled by the players.

Third, in our study, we propose the combined usage of methods from data science and complex adaptive systems. This combination provides a large arsenal of available methods and allows the exploitation of complementary benefits. Specifically, in our study, we use eight different methods (descriptive statistics, linear regression, nonlinear regression, classification, hierarchical clustering, scaling behavior, resampling, and LMG) to investigate WOT Blitz from several angles. In contrast, most studies utilize only a few methods or even only one method for the analysis. For example, in the work of Chun et al. [17], network analysis was combined with the scaling behavior, and in the work of Jiang et al. [41], only a spectral analysis was used. In a wider context of MMOGs and related publications, the study by Cheng et al. [16] used descriptive statistics, correlation analysis, and classification, Szell and Thurner [74] used network analysis, and Thawonmas et al. [75] applied clustering. It is noteworthy to highlight that in the work of El-Nasr et al. [21], the importance of data science for studying MMOGs has been addressed but without a connection to complex adaptive systems, whereas in the work of Bainbridge [4], the potential of complex systems for investigating MMOGs has been discussed but without a connections to data science. This shows not only that our study is more extensive with respect to the applied methodology than most studies, even in a broader context, but also that the combined usage of data science and complex adaptive systems is so far underestimated.

The latter point is of particular importance for other MMOGs because it allows a method transfer. That means, beyond the results we obtain from analyzing WOT Blitz, our approach itself that combines methods from data science and complex adaptive systems could be of interest for other MOGGs also requiring flexibility in addressing complex problems.

3 MATERIALS AND METHODS

In this section, we discuss the data and the methods used for our analysis.

3.1 Description of WOT Blitz

The game setting consists of seven vs seven matches, whereas each team consists of seven players. Each player freely chooses a vehicle (tank) from 1 of 10 available tiers. The **Match-Making (MM)** of the games then constructs the teams according to an (unknown) algorithm based on the presence of online players, and a player cannot influence the MM. However, there are constrains for the construction of teams that are known. First, a match can only contain vehicles of the same tier or of adjacent tier pairs—that is, from tier i and tier $i + 1$ for $i \in \{1, \ldots, 9\}$. We call the latter *mixed tier matches*. Importantly, the majority of games consist of mixed tier matches. Second, two players can form a mini-team (called a *platoon*) if they follow the previous rules. In this case, the MM places both players in the same team. Third, multiple mini-teams per match are permitted.

Put simply, the game style mimics a dodgeball game, whereas 'ball hits' are exchanged via 'shots' causing a certain amount of damage (determined by an unknown algorithm). The goal of the game is either to destroy all opponent's vehicles or capture the base. A match lasts at most 7 minutes or finishes earlier if one team defeats its opponent.

Table 1. Description of the Meaning of Available Features Used for Our Analysis

Feature	Value	Meaning
tier	1–10	difficulty level of the game
battles	integer value	average number of battles for player per tank
win rate	[0, 1]	average win rate per tank
dr	positive real number	average damage ratio per tank
kdr	positive real number	average kills/deaths ratio per tank
hpb	positive real number	average hits per battle per tank
spots	positive real number	average spots per tank
wpm	[0, 1]	average win per match per tank
dpm	positive real number	average damage per minute per tank
kpm	positive real number	average kills per minute per tank
hit rate	[0, 1]	average hit rate per tank
survival	[0, 1]	average survival rate per tank
players	integer value	total number of players per tank

The scales of measurement of all features except the tier are on a ratio scale. In contrast, the tier corresponds to an interval scale.

We would like to note that the game has not been designed to simulate a particular real-world situation. This prevents any straightforward associations and identification with real-world counterparts.

3.2 Data

For our analysis, only average statistics are used. That means we do not analyze individual matches nor individual players. Instead, average statistics of vehicles are analyzed. These averages are based on the outcome of millions of individual games. Hence, a specific time point for obtaining the data is not of crucial importance.

The data for our analysis can be freely obtained from BlitzStars (https://www.blitzstars.com/toptanks). From there, we collected all data in February 2021. This repository allowed us to obtain information for all currently used tanks and provided average statistics for them. In total, information for 403 tanks was available characterized by 16 descriptive features. From these, we selected 15 (tier, battles, win rate, dr, kdr, dpb, kpb, hpb, spots, wpm, dpm, kpm, hit rate, survival, players) for further analysis because one feature (mastery) did not contain usable information.

In Table 1, we describe the meaning of the features in the used dataset.

The average damage ratio per tank is given by

$$dr = \frac{1}{N} dr_i, \tag{1}$$

where N is the total number of players for a tank and dr_i is the damage ratio for a player defined by

$$dr_i = \frac{\text{damage caused}}{\text{damage received}}. \tag{2}$$

Similarly, the kills/deaths ratio for a player is defined by

$$kdr_i = \frac{\#\text{ destroyed vehicles}}{\#\text{ being destroyed}}. \tag{3}$$

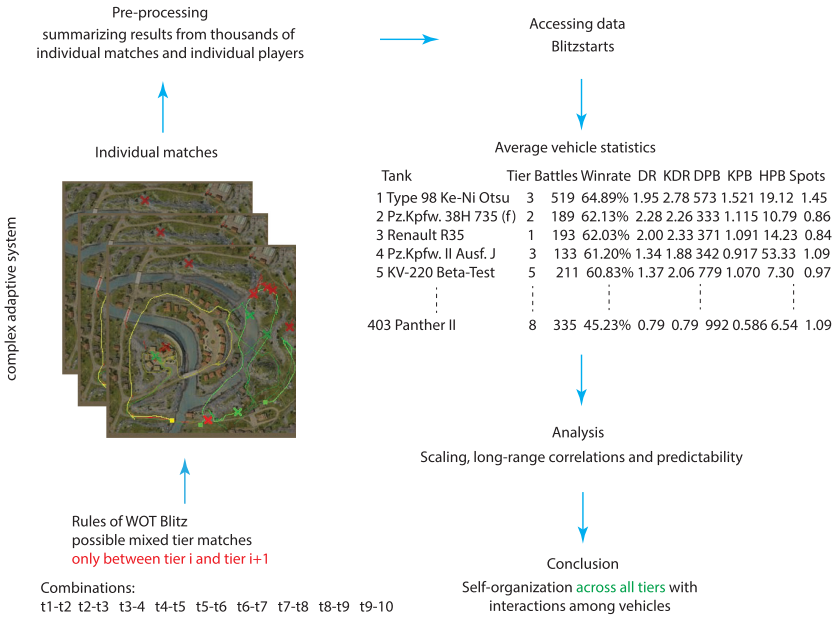

Fig. 1. Overview of our analysis. The left-hand side shows steps that cannot be influenced but are defined by the rules of the game. The right-hand side outlines the analysis and highlights the data on which our analysis is based.

We would like to emphasize that information about individual players or matches is not available for our analysis. Only the derived average vehicle statistics, as described earlier (see Table 1), are available for the analysis. That means dr_i and kdr_i for individual players are not used but

$$dr = \mathbb{E}[dr_i] = \frac{1}{P}\sum_i dr_i \tag{4}$$

$$kdr = \mathbb{E}[kdr_i] = \frac{1}{P}\sum_i kdr_i \tag{5}$$

provide the expectation values for vehicles where P is the number of players who used a particular vehicle.

Overall, information about 403 vehicles is available, representing average summary statistics from 24,015 players and more than 2.62 million matches for version 6.7 of WOT Blitz.

A general overview of our analysis is shown in Figure 1. In this figure, a snippet of the used data that are analyzed is also shown.

3.3 Hierarchical Clustering

For the hierarchical clustering, we used agglomerative clustering with Ward's minimum variance method. Distances between features vectors are measure with Euclidean distance. For the evaluation of the binary classification, we used the Matthews correlation coefficient (mcc) and the area under the receiver operating characteristic (AUROC) [25], and the Matthews correlation coefficient is defined by

$$\text{mcc} = \frac{\text{tp} \times \text{tn} - \text{fp} \times \text{fn}}{\sqrt{(\text{tp}+\text{fp})(\text{tp}+\text{fn})(\text{tn}+\text{fp})(\text{tn}+\text{fn})}}. \tag{6}$$

The Matthews correlation coefficient is bounded between −1 (for tp = 0 and tn = 0) and 1 (for fp = 0 and fn= 0). Furthermore, the Matthews correlation coefficient has the property of changing the sign when relabeling the classes—that is, by mapping the labels of class 1 and 2 by $1 \rightarrow 2$ and $2 \rightarrow 1$.

3.4 Importance of Predictor Covariates

To determine the contribution of system variables on the win rate of a vehicle, we evaluate their predictive power. For this, we formulate linear regression models using system variables as regressors for predicting the win rate of a vehicle. Put simply, the contribution of each predictor variable is assessed by the contribution for explaining R^2 (coefficient of determination) for a regression model, whereas in general R^2 is given by

$$R^2 = \frac{\text{model SS}}{\text{total SS}}. \tag{7}$$

Here, 'SS' indicates the sum of squares either of a model or in total.

A known problem of finding a decomposition of R^2 is that for correlated predictors such a decomposition depends on the ordering of the predictors. As a solution to this problem, Lindeman et al. [47] proposed a procedure, called *LMG*, to average over all possible orderings of predictors [33]. Successful examples for various applications of the LMG method can be found elsewhere [10, 22, 77].

Formally, LMG is based on a sequential decomposition of R^2,

$$seqR^2(Y|X) = R^2(Y \cup X) - R^2(X), \tag{8}$$

where X and Y are two disjoint sets of regressors. For p regressors, $x_1, \ldots, x_p$, let the vector $r = (r_1, \ldots, r_p)$ indicate the permutations of the p indices and let $X_{n+1}(r)$ denote the set of predictors ordered according to r, before the new predictor x_{n+1} is added to the model. For this situation, Equation (8) can be written as follows [34]:

$$seqR^2(\{x_{n+1}\}|X_{n+1}(r)) = R^2(\{x_{n+1}\} \cup X_{n+1}(r)) - R^2(X_{n+1}(r)). \tag{9}$$

From this, one obtains the LMG score by

$$\text{LMG}(x_{n+1}) = \frac{1}{p!} \sum_{p! \text{ permutations of } r} seqR^2(\{x_{n+1}\}|X_{n+1}(r)) \tag{10}$$

that considers all possible permutations. That means $\text{LMG}(x_{n+1})$ is an average value considering all other models not using x_{n+1} of size up to $p - 1$. In general, a computational issue of the preceding approach is caused by the averaging over all permutations. For a large number of predictor variables, this can be computationally demanding or even unfeasible. Fortunately, for our analysis, the number of predictors is sufficiently small to avoid this problem.

Evaluating Equation (10) for all predictors gives the importance for the regressors. One can prove that LMG decomposes R^2 into non-negative contributions that sum to the total R^2 [34]. Hence, normalizing LMG by the total value of R^2 one obtains a probabilistic decomposition—that is,

$$\sum_{i=1}^{p} p_i(x_i) = 1, \tag{11}$$

when the total number of predictor variables is p.

For our analysis, we perform the preceding decomposition for each tier. This allows us to identify differences in the predictability across the different tiers.

3.5 Nonparametric Nonlinear Regression

In addition to linear regression models, we use models for a nonparametric nonlinear regression. For this, we use a **Multivariate Adaptive Regression Splines (MARS)** model [28, 29]. MARS is a procedure that automatically creates a piecewise linear model. The procedure assesses each data point for each predictor as a knot and creates a linear regression model for candidate predictors. This process continues until sufficiently many knots are found, which can result in highly nonlinear prediction models.

Specifically, MARS uses a series of "basis functions" to build a model that fits the data. Each basis function is a simple equation that describes a relationship between an independent variable and the dependent variable. For example, in the following equation,

$$f(x) = \beta_0 + \sum_{i=1}^{M} \beta_i h_i(x), \tag{12}$$

the h_i are M base functions of the independent variable x. The base functions can be splines with a certain number of knots where a knot is a point where two or more piecewise basis functions join together to form a continuous curve.

These basis functions are combined to form a complex model that represents the overall relationship between the independent and dependent variables. The MARS model is built in a stepwise manner. At each step, a new basis function is added to the model, or an existing basis function is modified, to improve the fit of the model to the data. The process continues until a satisfactory level of fit is achieved or until some stopping criterion is met. One of the key features of MARS is its ability to handle nonlinear relationships between the independent and dependent variables. It does this by allowing the basis functions to have breakpoints, where the relationship between the variables changes. This allows the model to capture more complex patterns in the data.

For our numerical analysis of MARS models, we use the package "earth" [54].

4 RESULTS

In the following section, we present our results. First, we are providing a general overview of the characteristics of the data. Then we investigate the scaling behavior and the predictability of system variables.

4.1 General Characteristics of the Data

We begin our analysis by providing an overview of the characteristics of the WOT Blitz data. In total, we analyzed information about 403 vehicles representing average summary statistics from 24,015 players and more than 2.62 million matches. That means these data do not provide information about individual matches or individual players but average statistics of vehicles.

In Figure 2, we show eight system variables in dependence on the tier. Specifically, shown are the (average) win rate, survival, kills/deaths ratio (kdr), damage ratio (dr), damage per minute (dpm), spots per battle, hits per battle (hpb), and hit rate. We averaged these features over all vehicles for a given tier. In addition, results averaged over all tiers are shown on the right-hand side of each figure, indicated by 'all.' Hence, this provides the average over all tanks in all tiers.

From the results shown in these figures, one can make the following observations. First, the damage per minute is strongly increasing with the tier. This is understandable given the fact that the health points per vehicle increase steadily with the tiers. Hence, the total health points per match (summarized over all seven vehicles of a team) increases, allowing each vehicle to cause more damage. Second, despite increasing damage per minute, the damage ratio decreases with increasing tiers. This is a reflection of the dynamic interplay between causing and receiving damage and

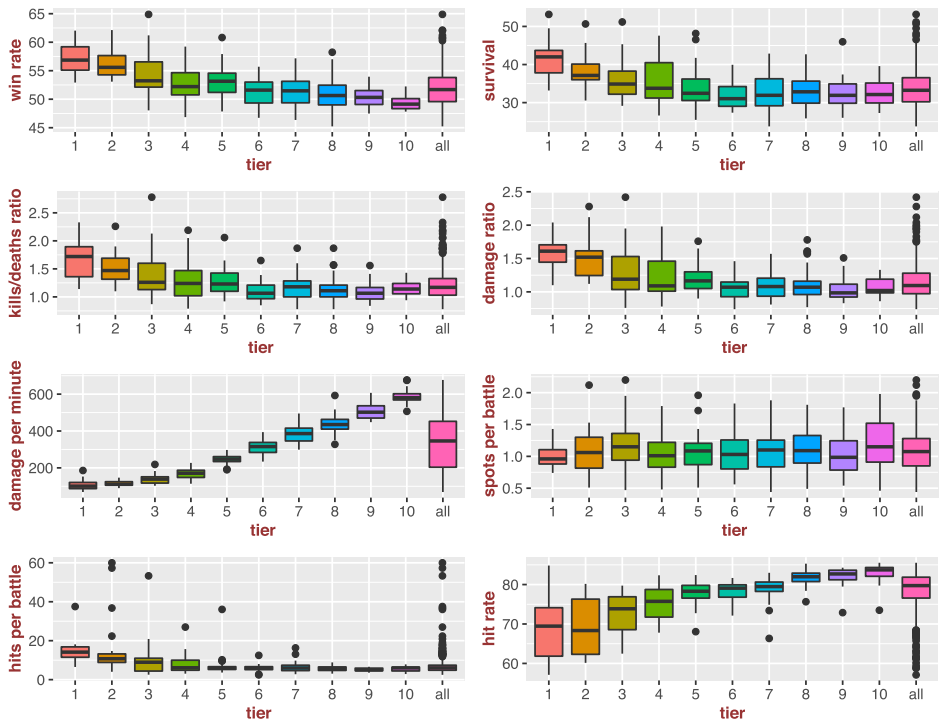

Fig. 2. Summary of different vehicle statistics in dependence on the tier. First row: Win rate and survival. Second row: Kills/deaths ratio and damage ratio. Third row: Damage per minute and sports per battle. Fourth row: Hits per battle and hit rate. Average results over all tiers are indicated by 'all.'

indicates nontrivial behavior. Third, the win rate and survival are both decreasing with increasing tiers. This indicates the increasing difficulty level of the game toward higher tiers, making it more difficult to win (respectively, survive) a match. Fourth, the hit rate also increases steadily from tier to tier. This variable is probably the best indirect reflection of the skills of individual players and shows that players on higher tiers are more skilled than on lower tiers. Hence, despite that fact that no direct assessment of individual players is provided, the hit rate allows a proxy of this conclusion on the level of summary statistics. Furthermore, comparison of the win rate with the hit rate reveals an anticorrelation. That means while the skill level of the players increases toward higher tiers, the win rate of vehicles deteriorates despite the fact that the characteristics of vehicles become stronger.

Overall, each of these figures shows the unsteady behavior of the game among the different tiers because only one of the eight features, namely spots per battle (third row, second column), shows a near constant behavior. However, this variable captures only marginal information about matches because it refers to the first spotting of vehicles (if the same vehicle is spotted multiple times, this is not recorded by this variable). In summary, there are crucial differences between the different tiers pointing to complex changes in the dynamics of the game. This is also reflected by the many outliers and the large interquartile range of the boxplots for the averaged results (shown by 'all').

4.2 Scaling Behavior

For playing any game, a key outcome variable is certainly the win rate because it indicates directly the success of playing the game. For this reason, in the following, we investigate the scaling behavior of the win rate in dependence on other system variables.

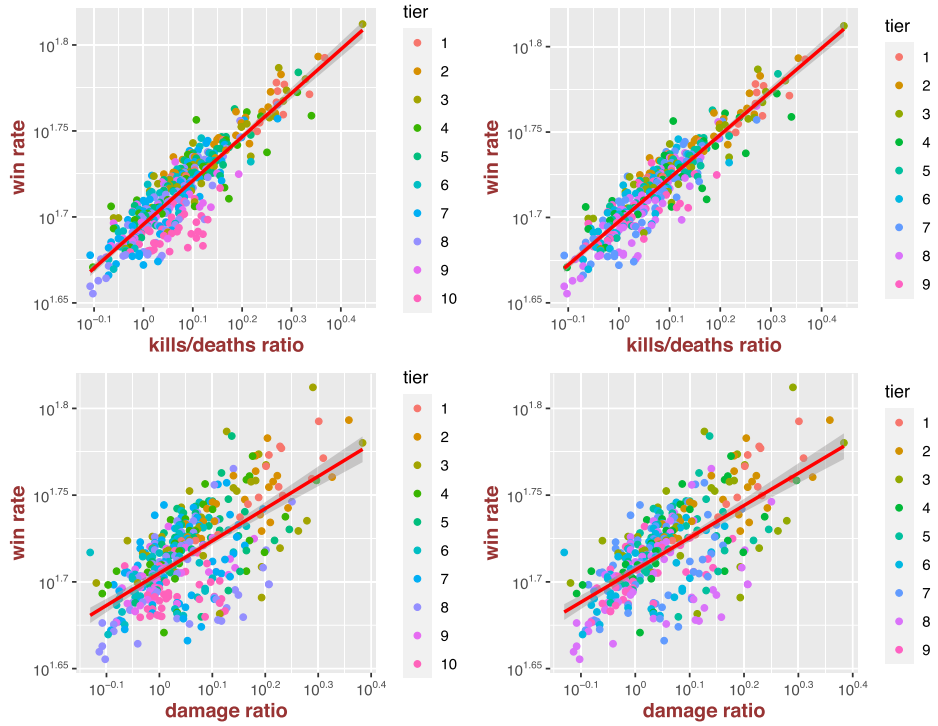

Fig. 3. Scaling behavior in double logarithmic plots for win rate vs kills/deaths ratio (top row) and win rate vs damage ratio (bottom row). The first column shows results for all tiers, whereas the second column shows reduced results because all vehicles from tier 10 have been excluded. For all figures, the legend shows the contribution of the vehicles from the corresponding tiers.

In Figure 3, we show the scaling behavior of the win rate in double logarithmic plots. Specifically, shown is the win rate vs kills/deaths ratio (top row) and the win rate vs damage ratio (bottom row). To see which data points come from what tiers, we color highlighted the vehicles of each tier by a different color (see the legend). As one can see, the shown results indicate a linear behavior between the win rate vs kills/deaths ratio and win rate vs damage ratio. That means the win rate scales like a power-law [55].

Mathematically, this behavior can be derived starting from the nonlinear relation between two variables in the original scale,

$$y \sim x^{\gamma}, \tag{13}$$

where y corresponds to the win rate and x corresponds either to the kills/deaths ratio or the damage ratio. By taking the logarithm on both sides,

$$\log(y) \sim \gamma \log(x) \Rightarrow y' \sim \gamma x', \tag{14}$$

one obtains a linear relation in the logarithmically scaled variables $y' = \log(y)$ and $x' = \log(x)$. In general, a power-law indicates a long-range behavior in the system across different scales [2].

By fitting linear regression models, we can quantify these results as shown in Table 2 (all tiers). Overall, all results are highly statistically significant because all p-values of the exponents are $< 10^{-16}$, confirming the presence of a power-law. We repeated the preceding analysis for further variables, including damage per minute, spots per battle, and hit rate; however, none of these

Table 2. Results from Linear Regression Models for Power-Laws

Model	exponent γ	p-value
win rate vs kdr, all tiers	0.25	$< 10^{-16}$
win rate vs kdr, reduced	0.25	$< 10^{-16}$
win rate vs dr, all tiers	0.18	$< 10^{-16}$
win rate vs dr, reduced	0.18	$< 10^{-16}$

Shown are the values of the exponent and the corresponding p-value for the power-laws in Figure 3. The results for 'all tiers' correspond to the first column in Figure 3, whereas the results for 'reduced' correspond to the second column in Figure 3 (where tier 10 is removed).

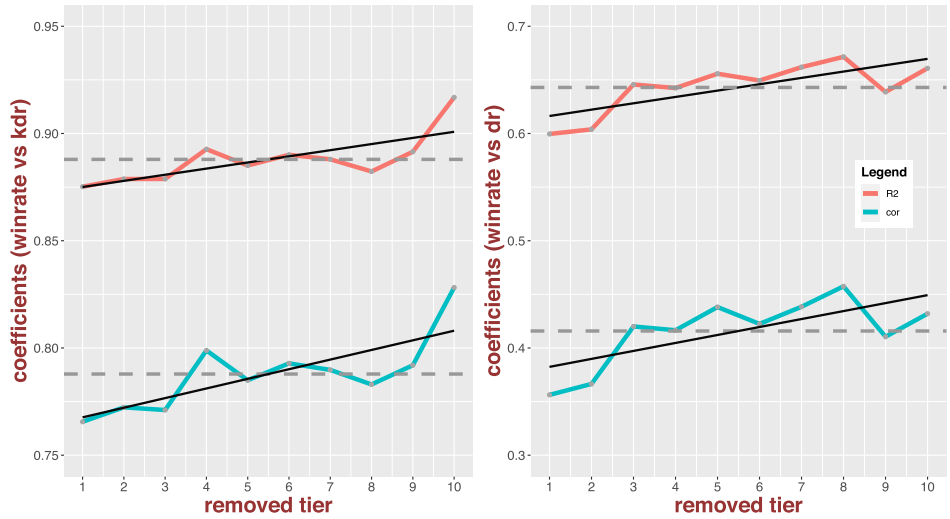

Fig. 4. Correlation coefficients (red lines) and R2 coefficients (green lines) for win rate vs kills/deaths ratio (left) and win rate vs damage ratio (right), whereas the x-axis indicates which tier has been removed from the analysis. The gray dashed lines indicate the mean values of the correlation and R2 coefficients (0.88 and 0.78 for the left and 0.64 and 0.41 for the right figure), and the black lines correspond to linear regression models.

reached correlation values as high as for the kills/deaths ratio and damage ratio, although also for damage per minute, the power-law is highly significant with a correlation coefficient of 0.74.

Next, we investigate the strength of the contribution of each tier on the power-law behavior as shown in the first column in Figure 3. For this analysis, we remove all tanks belonging to one particular tier and determine the correlation between the win rate vs kills/deaths ratio and the win rate vs damage ratio for all remaining tiers and vehicles. Furthermore, we perform a linear regression for both configurations and quantify the quality of a fit with the R-square (R2) corresponding to the coefficient of determination. The results of this analysis for all tiers are shown in Figure 4 and for tier 10 in Figure 3 (second column).

First, from Figure 4, it is clear to see that both configurations are on a different scale. Whereas the results for win rate vs kills/deaths ratio reach correlation values over 0.90 (shown in red) and R2 over 0.80 (shown in green), for win rate vs damage ratio these values drop to 0.65 and 0.45, respectively. For the correlation and tier 10, this can be confirmed by the more narrow point clouds

one can see in the top row of Figure 3 in comparison with the broader point clouds in the bottom row of Figure 3 (see the corresponding correlation values in Figure 4). Furthermore, R2 reveals that despite the statistically significance of the presence of the power-law (all p-values are significant), the quality of the fits for the kills/deaths ratio is about a factor of 2 larger than for the damage ratio. This hints also to the importance of this feature for explaining the win rate.

Second, it is interesting to note that the values for the correlations and R2 are not constant but tend to increase with higher tiers. This can be seen by comparison with the mean values of the correlation and R2 coefficients shown by the dashed lines in Figure 4 (the mean values are 0.88 and 0.78 for the left figure and 0.64 and 0.41 for the right figure). For the win rate vs kills/deaths ratio, this behavior is especially severe with a clear jump for removed tier 10 vehicles. For the win rate vs damage ratio, this behavior is more moderate and the highest values are obtained for removed tier 8 vehicles. To quantify this behavior, we performed a linear regression analysis for all four curves in Figure 4. As a result, we found that for a significance level of $\alpha = 0.05$, the values of the slope is in all cases statistically significant (regression lines in Figure 4 are shown in black).

Taken together, given the much better fit quality for the kills/deaths ratio and the largest observable effect for removal of tier 10 vehicles, one can conclude that the removal of higher tiers has on average a larger influence on the win rate.

4.3 Detecting Predictability

Next, we are extending the obtained results in the previous section to gain insights into the high values of R2 (see Figure 4) indicating a high fit quality. Specifically, so far we have been characterizing properties of the system; however, now we want to use variables for making prediction. In contrast to R2, which is only based on pairs of variables, we extend our analysis to the multivariate case.

From the preceding analysis, we learned that the three variables kills/deaths ratio, damage ratio, and damage per minute (not shown) have the strongest correlation with win rate. For this reason, we will use these four variables to perform a hierarchical clustering. That means each vehicle is described by a four-dimensional vector, $x \in \mathbb{R}^4$, whereas its components correspond to kills/deaths ratio, damage ratio, damage per minute, and win rate. We perform the hierarchical clustering of vehicles for pairs of tiers because the MM of the game also pairs adjacent tiers together in matches (from tier 1-2, tier 2-3, etc.). Hence, in total, we perform nine different hierarchical clusterings.

The results of these nine hierarchical clusterings are shown in Figures 5 through 7. For instance, for the hierarchical clustering of vehicles from tier 7-8 (see Figure 7(A)), we used a total of 137 vehicles, 63 vehicles from tier 7 and 74 vehicles from tier 8. For a better visual discrimination, we color highlighted vehicles from tier 7 in red and vehicles from tier 8 in blue. Overall, for all dendrograms, one can see two clearly discernible clusters. For a further quantification of the binary classification capability of the hierarchical clusterings, we split the dendrograms along the two main branches and quantify the observed distribution specifying the number of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn). For instance, for the hierarchical clustering of vehicles from tier 7-8 (see Figure 7(A)), we find

$$\text{left cluster} = \begin{cases} \text{tp} = 31 \\ \text{fp} = 17, \end{cases} \tag{15}$$

$$\text{right cluster} = \begin{cases} \text{tn} = 57 \\ \text{fn} = 32. \end{cases} \tag{16}$$

As an error measure to quantify this binary classification, we use Matthews correlation coefficient (mcc) because this error score is not sensitive to the labeling (class 1 vs class 2) of the clusters.

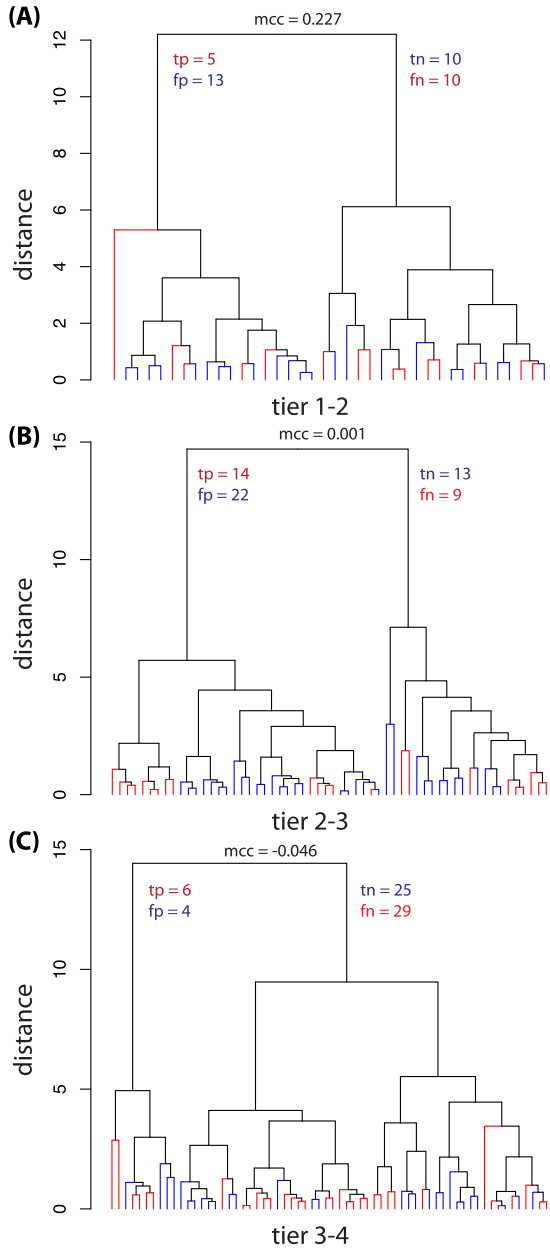

Fig. 5. Hierarchical clustering for tier 1-2 (A), tier 2-3 (B), and tier 3-4 (C) vehicles. The red branches correspond to lower tier vehicles and blue branches to the higher tier vehicles. The resulting Matthews correlation coefficient for a binary classification is $mcc = 0.227$ (A), $mcc = 0.001$ (B), and $mcc = -0.046$ (C).

For the preceding error values in Equations (15) and (16), we obtain a Matthews correlation coefficient of $mcc = 0.274$. We would like to note that, ideally, vehicles of the same tier should perform similarly. However, this would result in two main branches in a dendrogram, each containing only vehicles from the same tier corresponding to all blue and all red branches. From this, a binary classification splitting the dendrogram along the two main branches would result in a Matthews

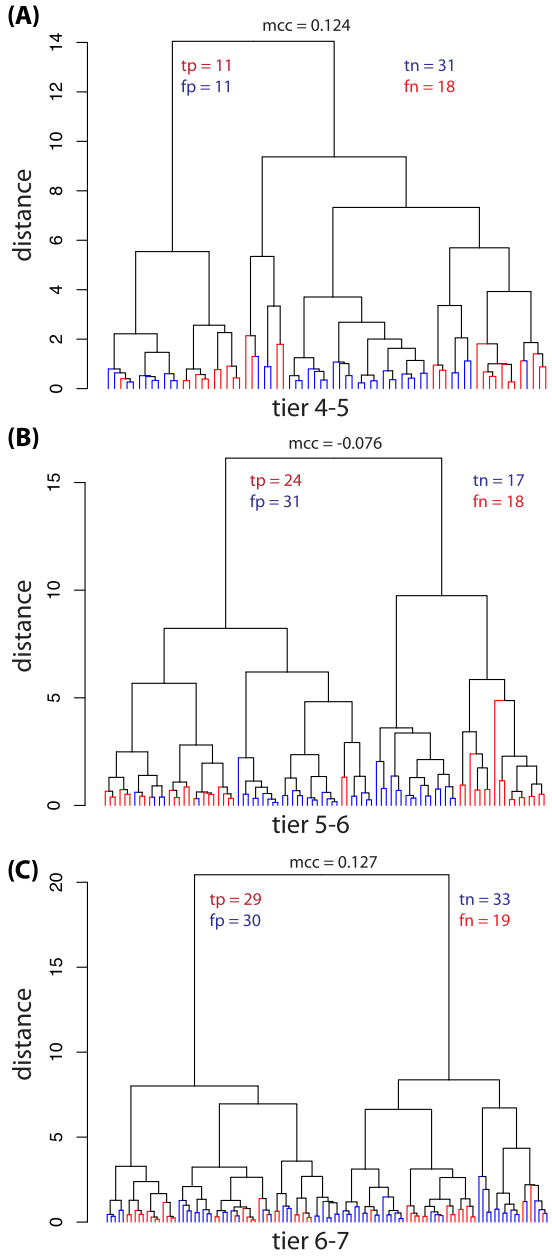

Fig. 6. Hierarchical clustering for tier 4-5 (A), tier 5-6 (B), and tier 6-7 (C) vehicles. The red branches correspond to lower tier vehicles and blue branches to the higher tier vehicles. The resulting Matthews correlation coefficient for a binary classification is $mcc = 0.124$ (A), $mcc = -0.076$ (B), and $mcc = 0.127$ (C).

correlation coefficient of ±1.0 (a relabeling of all class labels changes the sign) and an AUROC of 1.0.

A summary of all classifications from Figures 5 through 7 in the form of the corresponding Matthews correlation coefficients is shown in Figure 8. To obtain a better understanding for

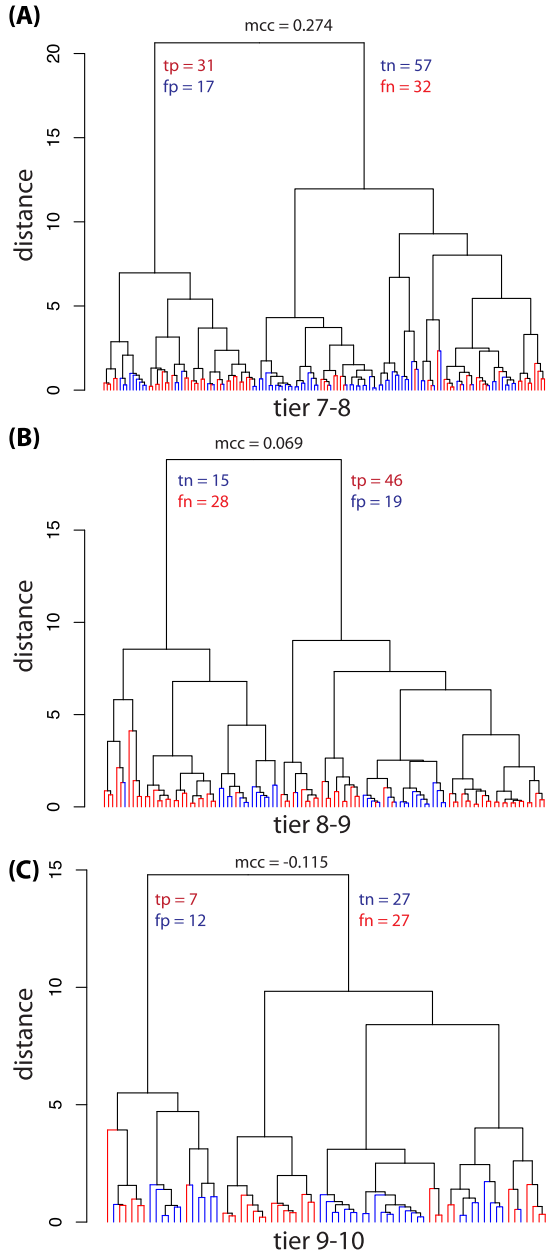

Fig. 7. Hierarchical clustering for tier 7-8 (A), tier 8-9 (B), and tier 9-10 (C) vehicles. The red branches correspond to lower tier vehicles and blue branches to the higher tier vehicles. The resulting Matthews correlation coefficient for a binary classification is $mcc = 0.274$ (A), $mcc = 0.069$ (B), and $mcc = -0.115$ (C).

the magnitude of the values of the Matthews correlation coefficient, we also included results about the randomization of the data. Specifically, we randomly assigned vehicles to either cluster maintaining the total number of vehicles of the tiers and evaluated the Matthews correlation coefficient. In this way, one can estimate the mean value of Matthews correlation coefficient and

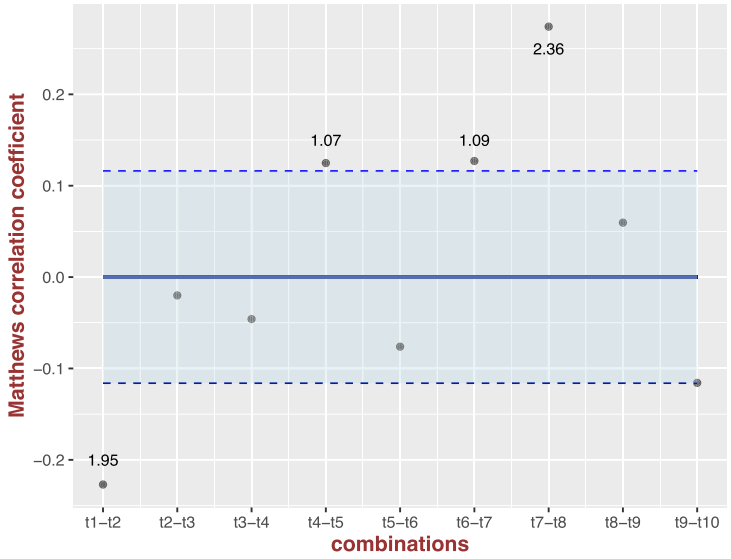

Fig. 8. Matthews correlation coefficients for the classification of all tier combinations obtained from the hierarchical clusterings in Figures 5 through 7. Shown in blue is the confidence interval estimated from randomizations of the data. The numbers outside the confidence interval indicate the distance from a Matthews correlation coefficient of zero measured by the number of standard deviations.

Table 3. Evaluation of the Hierarchical Clusterings Shown in Figures 5 through 7 by the AUROC When Splitting along the Two Main Clusters

Score	t1-2	t2-3	t3-4	t4-5	t5-6	t6-7	t7-8	t8-9	t9-10
AUROC	0.61	0.51	0.53	0.56	0.53	0.56	0.64	0.52	0.56

its 95% confidence interval. In Figure 8, the mean value of this (which is zero) is shown as a blue line and the boundaries of the 95% confidence intervals are indicated by blue dashed lines. This allows an assessment of the Matthews correlation coefficients for the hierarchical clusterings. As one can see from Figure 8, four (one is located directly on a boundary) of the nine classifications are situated outside the 95% confidence interval for randomized classifications. We would like to remark that the numbers we added to the four classifications outside the confidence interval correspond to the number of standard deviations these values are away from the mean Matthews correlation coefficient for the randomized data.

To make sure that our results do not depend on the error measure, we repeated the preceding analysis for the AUROC. The results of this are shown in Table 3. As one can see, our findings are essentially confirmed, which means that the tier pairs with higher Matthews correlation coefficients also have higher AUROC values, and similarly for the lower values of the Matthews correlation coefficients and AUROC values.

We would like to emphasize that our results are not intended to demonstrate that the behavior of the system is fully deterministic and as such can be predicted to a high degree. Instead, we demonstrated that the system is not completely random but contains predictable aspects. In our preceding analysis, we studied the mixing of vehicles from two tiers into two clusters corresponding to the two main branches in Figures 5 through 7. As one can see from either the Matthews correlation coefficients or the AUROC, this is not homogenous behavior but varies between the tier pairs.

4.4 Quantifying Importance of Predictor Variables

Finally, we complement our preceding analysis by quantifying the contribution of individual variables on the predictability. This will also allow an easier interpretation of the prediction model [26]. To study this, we define a multiple linear regression model [23],

$$y \sim \beta_1 x_1 + \beta_2 x_2 \ldots \beta_p x_p, \tag{17}$$

that maps input variables $x_1, x_2, \ldots x_p$ weighted by the regression coefficients $\beta_1, \beta_2, \ldots \beta_p$ to the output variable y. In our case, we use the win rate as the output variable, y, because it provides a clear indicator for successfully playing the game. For the input variables, we use the same variables as in our preceding analysis, namely the kills/deaths ratio (kdr), damage ratio (dr), and damage per minute (dpm) corresponding to x_1, x_2, x_3.

The regression model in Equation (17) allows us to obtain optimal values for the regression coefficients; however, the coefficients $\beta_1, \beta_2, \ldots \beta_p$ do not provide a quantification for the importance of the underlying predictor variables for explaining the outcome variable in cases of correlation between the regressors. This is a general problem with correlations between regressors. A solution to the problem has been suggested by Lindeman et al. [47]. The authors proposed a method, now called *LMG*, that provides a decomposition of R^2 (coefficient of determination). The method deals with correlations by averaging over all possible orderings of predictors [33] (see Section 3). As a result, one obtains non-negative proportions for each coefficient that sums to 1. Hence, the interpretation of the importance of the regression coefficients is straightforward due to the normalization. We would like to note that a different name for LMG is *Shapley value regression* [48], which recently became quite popular for its contributions to explainable AI (artificial intelligence) [26, 49].

In Figure 9 (top), we show the results of this analysis. As one can see, the kills/deaths ratio (kdr) dominates the damage ratio (dr) and damage per minute (dpm) in the sense that the kdr is clearly elevated above the dr and dpm across all tiers. Furthermore, all three variables display a non-smooth behavior. Interestingly, the value of R2 declines significantly toward higher tiers, indicating a decline in the quality of the fitted models. This means that for tiers 9 and 10, a linear model is no longer a suitable choice for describing the relationship between win rate and the three variables kills/deaths ratio, damage ratio, and damage per minute.

To demonstrate that the preceding behavior is similar for higher-dimensional models, we repeated the preceding analysis for six predictor variables, namely kdr, dpm, dr, hpb, hr, and spb. The results of this analysis are shown in Figure 9 (bottom). Again, the influence of kills/deaths ratio dominates all other predictors. Interestingly, the damage ratio and damage per minute are more important than the remaining predictors, which justifies their usage for the preceding three-dimensional model. In addition, the declining behavior of R2 for tiers 9 and 10 is present and even stronger than for the three-dimensional model in Figure 9 (top). Overall, these results confirm the findings in Figure 9 (top) and justify the three-dimensional modeling of the win rate, especially for lower tiers, whereas higher tiers (tiers 9 and 10) defy a linear description.

To investigate this point further, we study different forms of nonlinear regression models. Specifically, we compare models based on nonlinear transformations and nonparametric MARS. From those analyses, we found that the nonparametric nonlinear model MARS obtained the best results. The results of this are shown in Figure 10. The top figures show the R2 for a seven-dimensional MARS regression (green) and for comparison R2 for a seven-dimensional linear regression (red). Both models use the predictor variables kdr, dpm, dr, hpb, hr, spb, and surv. As one can see, for higher tiers, MARS leads to large improvements over the linear regression model, especially for tier 10 for which R2 is 0.89, whereas for the linear regression it assumes a value of 0.49.

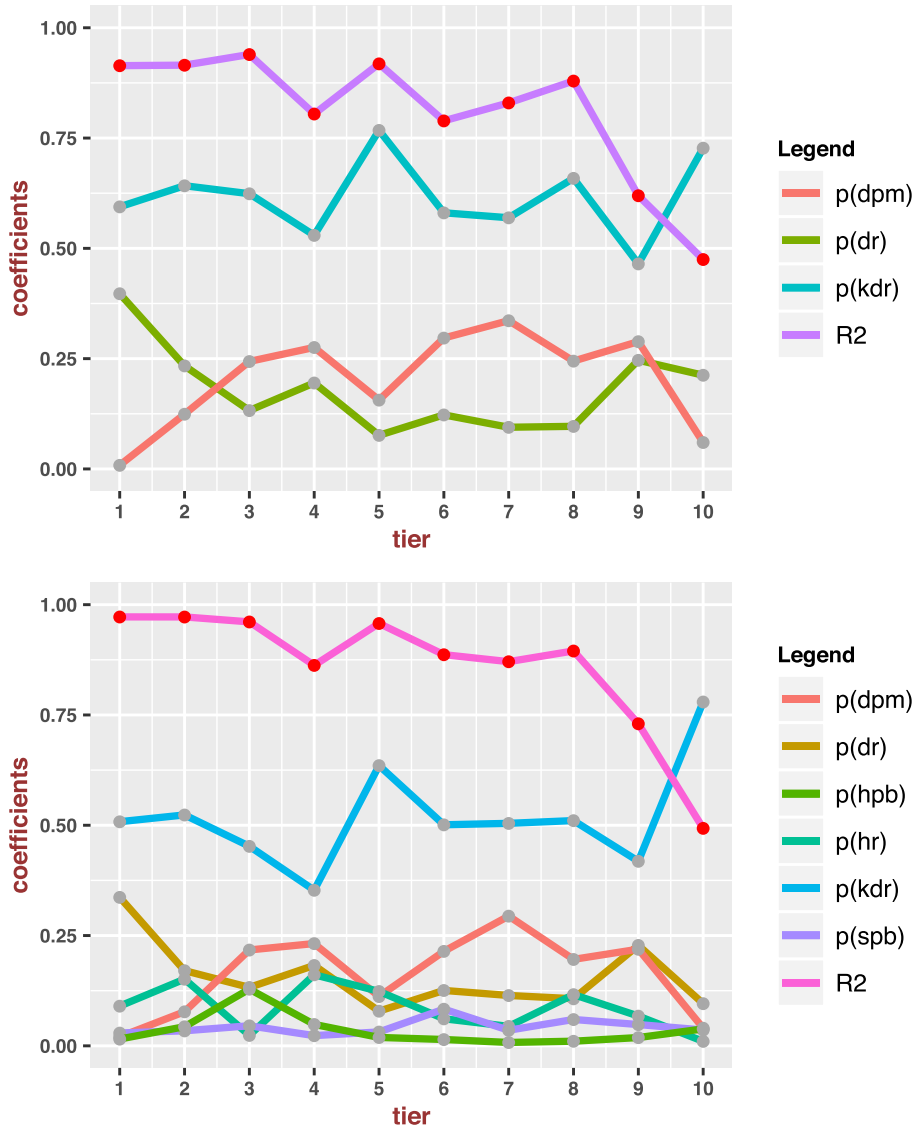

Fig. 9. Decomposition of predictive importance in multiple linear regression models. Shown are the percentages for the predictor variables. Furthermore, the quality of the linear regression models is assessed with R2. Top: Three-dimensional linear regression for the predictor variables kdr, dpm, and dr. Bottom: Six-dimensional linear regression for the predictor variables kdr, dpm, dr, hpb, hr, and spb.

For an easier comparison of the results, in the bottom figure we show the relative difference in percentage for the comparison between R2(MARS) and R2(LM). Specifically, we define this as

$$\text{relative difference} = \frac{R_i^2(MARS) - R_i^2(LM)}{R_i^2(LM)} \cdot 100\%, \tag{18}$$

where i indicates the tier. From this figure, one can see that the nonparametric nonlinear model MARS obtains an improvement of 80.7% for tier 10 and 24.2% for tier 9. Overall, this demonstrates that higher tiers require a nonlinear and higher-dimensional description.

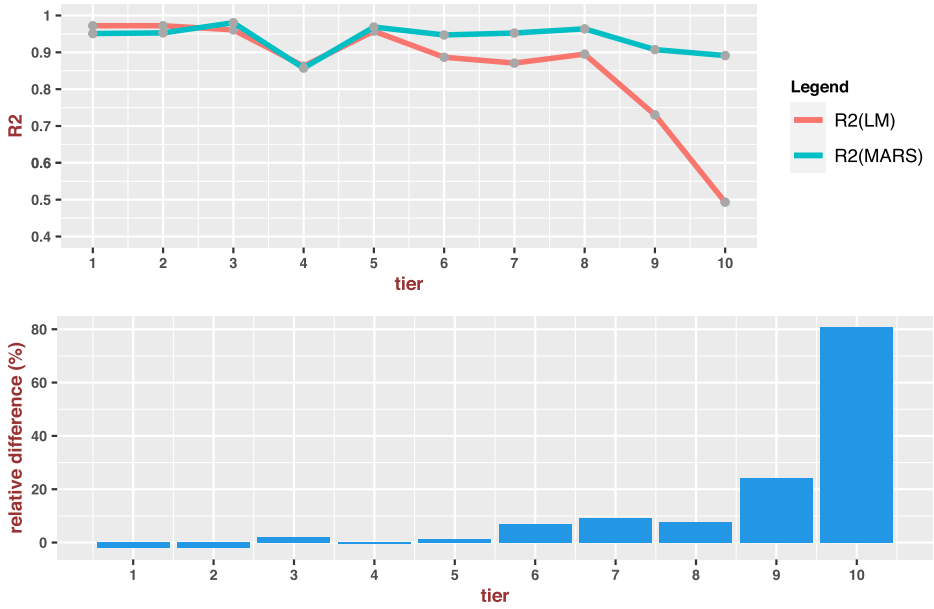

Fig. 10. Nonparametric nonlinear regression with a MARS model. Top: Seven-dimensional MARS regression (green) in comparison with a seven-dimensional linear regression model (red). The predictor variables for both models are kdr, dpm, dr, hpb, hr, spb, and surv. Bottom: Relative difference between R2 of the two models showing the improvement in percentage.

5 DISCUSSION

From our analysis, we found various power-laws (e.g., between the win rate vs kills/deaths ratio). Statistically, a power-law arises when the relation between two variables assumes a functional form with a constant power independent of the initial size and the relative changes. That means regardless of the size of the initial variable, the response of the second variable does not depend on it. Formally, one can derive this behavior from $y = ax^\gamma$ and $x \to x' = bx$ by

$$y = ax^\gamma \to y' = ax'^\gamma = a(bx)^\gamma = ab^\gamma x^\gamma = cx^\gamma . \tag{19}$$

From this behavior, one can see that regardless of the scale of x, which is controlled by b, the value of the response variable y' changes according to the same power-law exponent γ as for y. For this reason, the power-law behavior between x and y is called called *scale-free*.

The power-law behavior between the win rate and kdr is quite interesting for the following reasons. First, it reveals long-range interactions of system variables across all tiers. This is remarkable, because in each match, according to the rules of the game, at most two adjacent tiers are involved; however, the power-law involves all tiers. In general, the connection between a power-law behavior and long-range correlations has been studied extensively. Specifically, critical systems are known to exhibit a temporal and spatial scale-invariant behavior in the form of fractals and 1/f noise [5, 40]. This is a reflection of a propagation process resulting in long-range correlations based on local effects. Famously, this has been demonstrated by using a sandpile model [7] but has been subsequently found in many physical systems (e.g., ferromagnets [19], supercondutors [27], and plasma [66]). Importantly, also outside of physics, this effect has been observed for a number of different models, including the evolution of genomic DNA [53], score evolution of the game cricket [67], earthquakes [45, 61], the MMOG Pardus [74], the Minority game [14], and financial

markets [50]. For reasons of clarity, we would like to note that the interpretation of our power-law is different from conventional studies because, usually, the scaling is studied over intuitive dimensions of a system (e.g., space or time). For instance, for the sandpile model, the scaling is over the size of avalanches or their durations [7]. However, exceptions to the preceding can be found in the work of Bak and Sneppen [6], where the scaling of punctuated equilibrium has been studied over the distance of mutations changing the fitness of species, or in the work of Lux and Marchesi [50], where the scaling of the financial market has been studied over returns. Similarly, our power-laws are also observed in an abstract space provided by the system variables win rate, kills/deaths ratio, and damage ratio.

To further extend the understanding for the observed power-law behavior and its interpretation, we would like to mention that such a behavior is typical for self-organizing systems. In a general context, a "system" is a purposeful collection of inter-related components that work together to achieve an objective. A more dedicated definition of self-organization has been provided by Camazine et al. [12]:

> Self-organization is a process in which pattern at the global level of a system emerges solely from numerous interactions among the lower-level components of the system. Moreover, the rules specifying interactions among the system's components are executed using only local information, without reference to the global pattern.

In our case, the "components" are the individual players who work together in teams of seven with the "objective" to win against other teams. Furthermore, the "interrelation" between the components corresponds to the team play because if every player plays egoistically, the chances to win a game are decreased. Finally, the game is self-organizing [3, 12, 38] because there is no external "coach" who would influence or even determine the game, but the individual players make independent decisions in a way that increases the likelihood of winning. We would like to re-emphasize that the game is played by humans, not algorithms. Hence, a team play is a form of social cooperation because for each game, seven humans play against seven other humans. These observations may be obvious to some readers, but due to the technology-mediated form of the game, which masks the underlying setting abstractly, others may not see these connections immediately. To make this even more clear, we would like to add that the game is different from cyber-physical systems (CPS) [31, 56], which are systems of collaborating computational entities that interact with the surrounding physical world and its ongoing processes. Instead, WOT Blitz is in some sense an ordinary team-based game similar to soccer or basketball but with the difference that all interactions among the players, including the generation of the environment itself in which the game takes place, are computer generated and purely virtual. We believe it is interesting to note that this should allow investigations of, for example, social behavior, cooperation, and even psychological phenomena similar to other team-based games [20, 43, 51, 68].

It is also interesting to note that the scaling behavior is only observed in a double logarithmic representation of win rate and kdr. That means the inverse transformation back to the original variables (i.e., win rate and kdr) is an exponentialization. However, this means that in the original scales of win rate and kdr, the interactions are along exponential steps (not linear ones). The latter observation may provide an explanation for the highly nonlinear progression of the game moving toward higher tiers. Put simply, it is a known problem that good players on lower tiers struggle to compete on higher tiers in the sense that their performance drops considerably. The following average statistics for an individual player with a win rate of 80.1%, 73.3%, 65.0%, 57.0%, 50.9%, 53.2%, 45.1%, and 43.9% for tiers 3 through 10, respectively, exemplifies this.

An indirect effect of this may be related to abusive language and toxicity observed from analyzing in-game chats [58]. Unfortunately, this study analyzes this issue only across all tiers for World of Tanks (which is similar to WOT Blitz) and is not tier specific. For this reason, it is unclear if, for example, the use of abusive language increases toward higher tiers. However, individual players are not obligated to play on the highest tier accessible to them but can play low tier vehicles, even when they have vehicles on higher tiers. As a side note, we hypothesize that the jumping of players between tiers might spread toxicity, similar to an infection, even if toxicity would be dominating on one particular tier.

The hierarchical clusterings shown in Figures 5 through 7 allowed us to study the similarity of performance characteristics of vehicles from different tiers. Intuitively, one would expect that vehicles of the same tier perform similarly. Ideally, this would result in two branches in a dendrogram, each containing only vehicles from the same tier corresponding to all blue and all red branches. When using the dendrogram for a binary classification by splitting it along its two main branches, this would result in a Matthews correlation coefficient of ± 1.0 (a relabeling of all class labels changes the sign) and an AUROC of 1.0. This case would allow for an error-free prediction of the class to which a vehicle belongs. However, from the summary statistics shown in Figure 8 and Table 3, one can see that all tier pairs are far away from such a perfect classification, implying that there is a considerable mixing of vehicles from different tiers in the main branches of the dendrograms. This means there are many vehicles of tier $i + 1$ that are more similar to vehicles from tier i with respect to their performance. Considering the MM of the game, which allows either vehicles of the same tier or of adjacent tier pairs, this seems to be desirable to avoid the situation where low tier vehicles are always disadvantaged. Another observation one can make from the Matthews correlation coefficients in Figure 8 is that vehicles can be in two regimes. The first region is within one standard deviation of a zero Matthews correlation coefficient, whereas the second region is outside of this interval. This indicates a heterogeneous transition between the tiers because some tier pairs provide vehicles with quite similar performance characteristics (e.g., tier 2-3 or 5-6), whereas others are more different (e.g., tier 1-2 or 7-8).

From analyzing the importance of predictor variables (see Figure 9), we found that the quality of the linear regression models deteriorates toward higher tiers (i.e., R2 is declining). A possible explanation for this decline in predictability could be the underlying skill level of the players. An indirect measure of the latter is given by the hit rate, shown in Figure 2, which indeed is highest for tier 9 and tier 10. Hence, the predictability of the models could decline because the skill level of players increases making the outcome of games less predictable (skills are more important than characteristics of vehicles). Other interesting results we obtained from studying predictability are the fluctuating values of $p(\mathrm{dr})$ and $p(\mathrm{dpm})$ (see Figure 9). This indicates that the progression between tiers is quite heterogeneous and not steady (see also the Matthews correlation coefficients in Figure 8) and that the importance of system variables even changes between the tiers (see $p(\mathrm{dr})$ and $p(\mathrm{dpm})$). All these observations point to a heterogeneous progression through the tiers and to nonlinear effects.

For investigating that latter point, we also performed a nonlinear regression by using a MARS model where we used up to seven predictor variables. From a comparison between a nonlinear MARS models and a linear regression model, we found an improvement in R2 by 80.7% for tier 10 and 24.2% for tier 9 (also see Figure 10). Importantly, this was only achieved when using seven predictors, whereas lower-dimensional models still obtained an improvement, although much worse than the best ones. Overall, this demonstrates that higher tiers in WOT Blitz require a nonlinear and high-dimensional description to obtain a predictability similar to lower tiers.

When analyzing WOT Blitz, we noticed from the beginning the unconventional choice for the MM pairing vehicles of adjacent tiers together in matches (see Section 3). From our preceding

discussion, one may wonder if this is not the cause for the majority of observed issues and specifically for the heterogeneous progression through the tiers. Given the fact that vehicles from tier $i + 1$ are (on average) stronger than vehicles from tier i, this introduces a bias in the form of a competitive disadvantage. A confirmation of this may be obtained by analyzing eSports events of WOT Blitz (e.g., the Twister Cup). From this, one finds that only vehicles of tier 10 are used despite the fact that tier 9 vehicles would also be permissible. That means, voluntarily, that teams are only formed by vehicles of the same tier. Hence, neither rational scientific arguments (see the preceding) nor the behavior of the community during eSports events provide sensible arguments in favor of tier mixing in matches. However, given the fact that WOT Blitz is a commercial endeavor, one could speculate if mixed MM is supported from a marketing perspective. An argument in this favor is that players using a vehicle of tier i see in matches a vehicle of tier $i + 1$ that is (in average) better, and hence they want this vehicle to become more successful. However, this requires a progression to the next tier. Metaphorically, this corresponds to a 'carrot on a stick' strategy that seemingly leads the players through the tiers without applying force.

An important general point to emphasize is that the data upon which we based our analysis are the consequence of human team behavior. That means neither the characteristics of the vehicles nor the rules of the game alone result in an observable action but the combination of all three components. However, detailed characteristics of the human players are entirely hidden from us and only indirectly observable. One of these is the skill level of players, which can be associated with the hit rate (see Figure 2). Furthermore, this also means that technology alone is not sufficient to create a game like WOT Blitz, but there needs to be a concerted synchronization with the human actors to establish a "functional" system. Although the presence of power-laws is a good indicator of such a functioning, we also found issues for transitioning through the tiers. In fact, this progression is heterogeneous (see Figure 8). Additionally, the need for a high-dimensional, nonlinear description for predicting the win rate revealed the complexity of the game toward higher tiers (see Figure 10).

We would like to mention that there are studies that showed the impact of team composition on performance [57]. However, in our study, we neither use information about individual players nor have information about their participation in common teams in previous games. In contrast, we only use summary statistics of vehicles used by players. Furthermore, the composition of the teams is automatically provided by the MM system of the game. Hence, a player has no influence on the composition of teams at all but is automatically assigned by undisclosed rules.

Based on our findings, there are a number of future studies one could conduct. First, since most issues are more severe on higher tiers, to address these one needs to work its way down. That means starting from tier 10, characteristics of vehicles could be changed by monitoring the coefficient of determination (R2), the dimensionality, and the nonlinearity. Together, these factors need to be optimized without destroying the scaling behavior. At the moment, it is unclear if a homogeneous progression through the tiers is possible given the rules of the game allowing mixed pairing of tiers, and this analysis could bring clarity. Second, one could create special game modes that reflect particular sociological conditions. For instance, one could allow players to select their teammates potentially from a set of preferred players rather than on the individual player level. This way, the formation of teams could be studied. Furthermore, one could allow the players to determine the characteristics of vehicles, within certain boundaries and constraints and by potentially acquiring a penalty. This allows would us to study preferences of the players and resulting consequences. It is important to emphasize that all of these studies require interventions—either changes of the characteristics of the vehicles or the rules of the game. Hence, such studies could be only conducted in collaboration with Wargaming.

Finally, we would like to highlight that we analyzed WOT Blitz by utilizing methods from data science [24, 35, 64] and complex adaptive systems [39, 69], and also other MMOGs may benefit from such an approach. Interestingly, other MMOGs received very little attention from this perspective with notable exceptions of other works [30, 42, 74]. In general, the interdisciplinary nature of such studies may be a contributing factor to the challenge of developing a common language, knowledge, and values that enable equal participation and contribution from various fields. This has been extensively documented and can pose significant obstacles to collaboration and progress [9, 52]. However, real-world systems, such as the economy or society, for which games like WOT Blitz provide a virtual Petri dish, are unlikely to be fully understood through the lens of a single subject. Furthermore, the complexity of economy and human behavior necessitates the establishment of laws that can explain the observed activities. In this regard, the principles of self-organization may offer a potential framework for obtaining basic insights into these complex systems.

6 CONCLUSION

MMOGs such as WOT Blitz present wonderful opportunities for investigating virtual societies as their digital nature enables the comprehensive recording and analysis of every aspect of these communities. However, a challenge with this is to distinguish noise from information. In this article, we studied WOT Blitz with population data that combine myriad details into summary statistics.

Specifically, we studied three major levels of WOT Blitz. First, we studied the scaling behavior of the population of vehicles. This allowed us to identify long-range interactions among system variables (win rate vs kills/deaths ratio and win rate vs damage ratio). Second, we identified and quantified the predictability of system variables by classifying the population of vehicles. This translated the abstract finding about the scaling behavior of the population of vehicles into a concrete, observable behavior by assessing the predictability of the system. Third, we quantified the predictability of a performance measure of the system by a multiple regression framework utilizing LMG [47]. From this, we found a heterogeneous progression through the tiers and identified a single system variable (kills/deaths ratio) as the key driver for the win rate. Furthermore, we found that the game becomes more nonlinear for higher tiers rendering linear models suboptimal. Interestingly, this could be compensated by increasing the dimensionality of the model.

Overall, we studied WOT Blitz by combining methods from data science [24, 35, 64] and complex adaptive systems [39, 69]. We believe that, in general, such an interdisciplinary approach can be instructive also for other MMOGs beyond WOT Blitz because these virtual worlds enable different and creative views on social, behavioral, or economic problems that require flexible and powerful analysis methods [17, 18, 36, 65, 71, 73]. As we navigate the ongoing digital revolution [37], this may present an opportunity to receive the vital feedback necessary for effectively adapting to a rapidly changing world within our economy and society.

REFERENCES

[1] Leigh Achterbosch, Robyn Pierce, and Gregory Simmons. 2008. Massively multiplayer online role-playing games: The past, present, and future. *Computers in Entertainment* 5, 4 (2008), 1–33.

[2] Luis A. Nunes Amaral, Sergey V. Buldyrev, Shlomo Havlin, Michael A. Salinger, and H. Eugene Stanley. 1998. Power law scaling for a system of interacting units with complex internal structure. *Physical Review Letters* 80, 7 (1998), 1385.

[3] W. R. Ashby. 1947. Principles of the self-organizing dynamic system. *Journal of General Psychology* 37, 2 (1947), 125–128.

[4] William Sims Bainbridge. 2007. The scientific research potential of virtual worlds. *Science* 317, 5837 (2007), 472–476.

[5] Per Bak. 2013. *How Nature Works: The Science of Self-Organized Criticality*. Springer Science & Business Media.

[6] P. Bak and K. Sneppen. 1993. Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters* 71 (1993), 4083–4086.

[7] Per Bak, Chao Tang, and Kurt Wiesenfeld. 1987. Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters* 59, 4 (1987), 381.
[8] Jane Barnett and Mark Coulson. 2010. Virtually real: A psychological perspective on massively multiplayer online games. *Review of General Psychology* 14, 2 (2010), 167–179.
[9] Garry D. Brewer. 1999. The challenges of interdisciplinarity. *Policy Sciences* 32, 4 (1999), 327–337.
[10] Bénédicte Bullier, Hélène Cassoudesalle, Marie Villain, Mélanie Cogné, Clémence Mollo, Isabelle De Gabory, Patrick Dehail, Pierre-Alain Joseph, Igor Sibon, and Bertrand Glize. 2020. New factors that affect quality of life in patients with aphasia. *Annals of Physical and Rehabilitation Medicine* 63, 1 (2020), 33–37.
[11] Ronald S. Burt. 2012. Network-related personality and the agency question: Multirole evidence from a virtual world. *American Journal of Sociology* 118, 3 (2012), 543–591.
[12] Scott Camazine, Jean-Louis Deneubourg, Nigel R. Franks, James Sneyd, Guy Theraula, and Eric Bonabeau. 2003. *Self-Organization in Biological Systems*. Princeton University Press.
[13] Edward Castronova and Matthew Falk. 2009. Virtual worlds: Petri dishes, rat mazes, and supercolliders. *Games and Culture* 4, 4 (2009), 396–407.
[14] Damien Challet, Matteo Marsili, and Yi-Cheng Zhang. 2001. Stylized facts of financial markets and market crashes in minority games. *Physica A: Statistical Mechanics and Its Applications* 294, 3–4 (2001), 514–524.
[15] Elaine Chan and Peter Vorderer. 2006. *Massively Multiplayer Online Games*. Lawrence Erlbaum Associates.
[16] Meng-Tzu Cheng, Louisa Rosenheck, Chen-Yen Lin, and Eric Klopfer. 2017. Analyzing gameplay data to inform feedback loops in the Radix Endeavor. *Computers & Education* 111 (2017), 60–73.
[17] Selin Chun, Deajin Choi, Jinyoung Han, Huy Kang Kim, and Taekyoung Kwon. 2018. Unveiling a socio-economic system in a virtual world: A case study of an MMORPG. In *Proceedings of the 2018 World Wide Web Conference*. 1929–1938.
[18] Helena Cole and Mark D. Griffiths. 2007. Social interactions in massively multiplayer online role-playing gamers. *Cyberpsychology & Behavior* 10, 4 (2007), 575–583.
[19] P. J. Cote and L. V. Meisel. 1991. Self-organized criticality and the Barkhausen effect. *Physical Review Letters* 67, 10 (1991), 1334.
[20] Stewart Cotterill. 2012. *Team Psychology in Sports: Theory and Practice*. Routledge.
[21] Magy Seif El-Nasr, Truong-Huy D. Nguyen, Alessandro Canossa, and Anders Drachen. 2021. *Game Data Science*. Oxford University Press.
[22] F. Emmert-Streib and M. Dehmer. 2019. Defining data science by a data-driven quantification of the community. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 235–251.
[23] F. Emmert-Streib and M. Dehmer. 2019. High-dimensional LASSO-based computational regression models: Regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 359–383. https://doi.org/10.3390/make1010021
[24] F. Emmert-Streib, S. Moutari, and M. Dehmer. 2016. The process of analyzing data is the emergent feature of data science. *Frontiers in Genetics* 7 (2016), 12.
[25] F. Emmert-Streib, S. Moutari, and M. Dehmer. 2019. A comprehensive survey of error measures for evaluating binary decision making in data science. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 5 (2019), e1303.
[26] F. Emmert-Streib, O. Yli-Harja, and M. Dehmer. 2020. Explainable artificial intelligence and machine learning: A reality rooted perspective. *WIREs Data Mining and Knowledge Discovery* 10 (2020), e1368.
[27] Stuart Field, Jeff Witt, Franco Nori, and Xinsheng Ling. 1995. Superconducting vortex avalanches. *Physical Review Letters* 74, 7 (1995), 1206.
[28] Jerome H. Friedman. 1991. Multivariate adaptive regression splines. *Annals of Statistics* 19, 1 (1991), 1–67.
[29] Jerome H. Friedman and Charles B. Roosen. 1995. An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research* 4, 3 (1995), 197–217.
[30] Benedikt Fuchs, Didier Sornette, and Stefan Thurner. 2014. Fractal multi-level organisation of human groups in a virtual world. *Scientific Reports* 4 (2014), 6526.
[31] Carlos Gershenson. 2020. Guiding the self-organization of cyber-physical systems. *Frontiers in Robotics and AI* 7 (2020), 41.
[32] Ulrike Grömping. 2007. Estimators of relative importance in linear regression based on variance decomposition.*American Statistician* 61, 2 (2007), 139–147.
[33] Ulrike Grömping. 2009. Variable importance assessment in regression: Linear regression versus random forest. *American Statistician* 63, 4 (2009), 308–319.
[34] Ulrike Grömping. 2006. Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software* 17, 1 (2006), 1–27.

[35] Daniel D. Gutierrez. 2015. *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R.* Technics Publications.
[36] Helen Harris, Jeremy N. Bailenson, Alexia Nielsen, and Nick Yee. 2009. The evolution of social behavior over time in second life. *Presence: Teleoperators and Virtual Environments* 18, 6 (2009), 434–448.
[37] Dirk Helbing. 2015. *Thinking Ahead—Essays on Big Data, Digital Revolution, and Participatory Market Society.* Vol. 10. Springer.
[38] Francis Heylighen. 2001. The science of self-organization and adaptivity. *Encyclopedia of Life Support Systems* 5, 3 (2001), 253–280.
[39] John H. Holland. 2006. Studying complex adaptive systems. *Journal of Systems Science and Complexity* 19, 1 (2006), 1–8.
[40] H. J. Jensen. 1998. *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems.* Cambridge University Press.
[41] Zhi-Qiang Jiang, Fei Ren, Gao-Feng Gu, Qun-Zhao Tan, and Wei-Xing Zhou. 2010. Statistical properties of online avatar numbers in a massive multiplayer online role-playing game. *Physica A: Statistical Mechanics and Its Applications* 389, 4 (2010), 807–814.
[42] Zhi-Qiang Jiang, Wei-Xing Zhou, and Qun-Zhao Tan. 2009. Online-offline activities and game-playing behaviors of avatars in a massive multiplayer online role-playing game. *Europhysics Letters* 88, 4 (2009), 48007.
[43] Nancy Katz. 2001. Sports teams as a model for workplace teams: Lessons and liabilities. *Academy of Management Perspectives* 15, 3 (2001), 56–67.
[44] Brian Keegan, Muhammad Aurangzeb Ahmed, Dmitri Williams, Jaideep Srivastava, and Noshir Contractor. 2010. Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. In *Proceedings of the 2010 IEEE 2nd International Conference on Social Computing*. IEEE, Los Alamitos, CA, 201–208.
[45] Vladimir I. Keilis-Borok. 1996. Intermediate-term earthquake prediction. *Proceedings of the National Academy of Sciences* 93, 9 (1996), 3748–3755.
[46] Ben Kirman, Francesco Collovà, Fabrizio Davide, Eva Ferrari, Jonathan Freeman, Shaun Lawson, and Conor Linehan. 2011. Social architecture and the emergence of power laws in online social games. In *Proceedings of the Digital Games Research Association Conference (DIGRA'11)*
[47] R. H. Lindeman, P. F. Merenda, and R. Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis.* Scott, Foresman, Glenview, IL.
[48] Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* 17, 4 (2001), 319–330.
[49] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4768–4777.
[50] Thomas Lux and Michele Marchesi. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397, 6719 (1999), 498–500.
[51] F. Charles Mace, Joseph S. Lalli, Michael C. Shea, and John A. Nevin. 1992. Behavioral momentum in college basketball. *Journal of Applied Behavior Analysis* 25, 3 (1992), 657–663.
[52] Miles MacLeod. 2018. What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice. *Synthese* 195, 2 (2018), 697–720.
[53] Philipp W. Messer, Peter F. Arndt, and Michael Lässig. 2005. Solvable sequence evolution models and genomic correlations. *Physical Review Letters* 94, 13 (2005), 138103.
[54] Stephen Milborrow, T. Hastie, R. Tibshirani, A. Miller, and T. Lumley. 2017. earth: Multivariate adaptive regression splines. *R. Package Version* 5, 2 (2017).
[55] Michael Mitzenmacher. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1, 2 (2004), 226–251.
[56] László Monostori. 2018. *Cyber-Physical Systems.* Springer, Berlin, Germany, 1–8. https://doi.org/10.1007/978-3-642-35950-7_16790-1
[57] Satyam Mukherjee, Yun Huang, Julia Neidhardt, Brian Uzzi, and Noshir Contractor. 2019. Prior shared success predicts victory in team competitions. *Nature Human Behaviour* 3, 1 (2019), 74–81.
[58] Shane Murnion, William J. Buchanan, Adrian Smales, and Gordon Russell. 2018. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security* 76 (2018), 197–213.
[59] Katalin Nagygyörgy, Róbert Urbán, Judit Farkas, Mark D. Griffiths, Dalma Zilahy, Gyöngyi Kökönyei, Barbara Mervó, Antónia Reindl, Csilla Ágoston, Andrea Kertész, Eszter Harmath, Attila Olah, and Zsolt Demetrovics. 2013. Typology and sociodemographic characteristics of massively multiplayer online game players. *International Journal of Human-Computer Interaction* 29, 3 (2013), 192–200.
[60] Bonnie Nardi. 2010. *My Life as a Night Elf Priest: An Anthropological Account of World of Warcraft.* University of Michigan Press.

[61] Zeev Olami, Hans Jacob S. Feder, and Kim Christensen. 1992. Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Physical Review Letters* 68, 8 (1992), 1244.
[62] Daniel Pittman and Chris GauthierDickey. 2007. A measurement study of virtual populations in massively multiplayer online games. In *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*. 25–30.
[63] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, Laszlo Grad-Gyenge, and Hannes Werthner. 2013. On successful team formation: Statistical analysis of a multiplayer online game. In *Proceedings of the 2013 IEEE 15th Conference on Business Informatics*. IEEE, Los Alamitos, CA, 55–62.
[64] Foster Provost and Tom Fawcett. 2013. Data science and its relationship to big data and data-driven decision making. *Big Data* 1, 1 (2013), 51–59.
[65] Ge Qian. 2010. Psychological perspectives on social behaviors of Chinese MMORPG players. In *Proceedings of the International Conference on Technologies for E-Learning and Digital Entertainment*. 192–202.
[66] T. L. Rhodes, R. A. Moyer, R. Groebner, E. J. Doyle, R Lehmer, W. A. Peebles, and C. L. Rettig. 1999. Experimental evidence for self-organized criticality in tokamak plasma turbulence. *Physics Letters A* 253, 3–4 (1999), 181–186.
[67] Haroldo V. Ribeiro, Satyam Mukherjee, and Xiao Han T. Zeng. 2012. Anomalous diffusion and long-range correlations in the score evolution of the game of cricket. *Physical Review E* 86, 2 (2012), 022102.
[68] Gordon W. Russell. 2008. *Aggression in the Sports World: A Social Psychological Perspective*. Oxford University Press.
[69] H. G. Schuster. 2002. *Complex Adaptive Systems*. Scator Verlag.
[70] Seokshin Son, Ah Reum Kang, Hyun-Chul Kim, Taekyoung Kwon, Juyong Park, and Huy Kang Kim. 2012. Analysis of context dependence in social interaction networks of a massively multiplayer online role-playing game. *PLoS One* 7, 4 (2012), e33918.
[71] Michael Szell, Renaud Lambiotte, and Stefan Thurner. 2010. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* 107, 31 (2010), 13636–13641.
[72] Michael Szell, Roberta Sinatra, Giovanni Petri, Stefan Thurner, and Vito Latora. 2012. Understanding mobility in a social Petri dish. *Scientific Reports* 2 (2012), 457.
[73] Michael Szell and Stefan Thurner. 2010. Measuring social dynamics in a massive multiplayer online game. *Social Networks* 32, 4 (2010), 313–329.
[74] Michael Szell and Stefan Thurner. 2012. Social dynamics in a large-scale online game. *Advances in Complex Systems* 15, 6 (2012), 1250064.
[75] Ruck Thawonmas, Masayoshi Kurashige, and Kuan-Ta Chen. 2007. Detection of landmarks for clustering of online-game players. *International Journal of Virtual Reality* 6, 3 (2007), 11–16.
[76] Stefan Thurner, Michael Szell, and Roberta Sinatra. 2012. Emergence of good conduct, scaling and Zipf laws in human behavioral sequences in an online world. *PLoS One* 7, 1 (2012), e29796.
[77] Wei Tian, Yunliang Liu, Yeonsook Heo, Da Yan, Zhanyong Li, Jingjing An, and Song Yang. 2016. Relative importance of factors influencing building energy in urban environment. *Energy* 111 (2016), 237–250.
[78] Guillaume Turchini, Sebastien Monnet, and Olivier Marin. 2015. Scalability and availability for massively multiplayer online games. In *Proceedings of the 11th European Dependable Computing Conference (EDCC'15)*.
[79] Julia Crouse Waddell and Wei Peng. 2014. Does it matter with whom you slay? The effects of competition, cooperation and relationship type among video game players. *Computers in Human Behavior* 38 (2014), 331–338.
[80] Dmitri Williams, Nicolas Ducheneaut, Li Xiong, Yuanyuan Zhang, Nick Yee, and Eric Nickell. 2006. From tree house to barracks: The social life of guilds in World of Warcraft. *Games and Culture* 1, 4 (2006), 338–361.

Received 14 September 2022; revised 1 May 2023; accepted 4 August 2023