

The Playtime Principle: Large-scale Cross-games Interest Modeling

Rafet Sifa*, Christian Bauckhage*[†] and Anders Drachen[‡]

*Fraunhofer IAIS, St. Augustin, Germany

[†]B-IT, University of Bonn, Germany

[‡]Aalborg University, Aalborg, Denmark

Abstract—The collection and analysis of behavioral telemetry in digital games has in the past five years become an integral part of game development. One of the key challenges in game analytics is the development of methods for characterizing and predicting player behavior as it evolves over time. Characterizing behavior is necessary for monitoring player populations and gradually improve game design and the playing experience. Predicting behavior is necessary to describe player engagement and prevent future player churn. In this paper, methods and theory from kernel archetype analysis and random process models are utilized to evaluate the playtime behavior, i.e. time spent playing specific games as a function of time, of over 6 million players, across more than 3000 PC and console games from the Steam platform, covering a combined playtime of more than 5 billion hours. A number of conclusions can be derived from this large-scale analysis, notably that playtime as a function of time, across the thousands of games in the dataset, and irrespective of local differences in the playtime frequency distribution, can be modeled using the same model: the Weibull distribution. This suggests that there are fundamental properties governing player engagement as it evolves over time, which we here refer to as the Playtime Principle. Additionally, the analysis shows that there are distinct clusters, or archetypes, in the playtime frequency distributions of the investigated games. These archetypal groups correspond to specific playtime distributions. Finally, the analysis reveals information about player behavior across a very large dataset, showing for example that the vast majority of games are players for less than 10 hours, and very few players spend more than 30-35 hours on any specific game.

I. INTRODUCTION

Recent years has seen a deluge of behavioral data from players becoming available to the game industry. The reasons for the data surge are many, including the introduction of new business models, technological innovations, the popularity of online games and the increasing persistence of games. Irrespective of the causes, the proliferation of behavioral data leads to the problem of how to derive and implement insights from them. Game analytics as a domain of inquiry and practice has grown to meet that need. While game analytics is applied broadly across game development similar to business analytics in other industries, the major focus remains the *players*, or *users*. Players are alpha and omega to the success of digital games, and ensuring an engaging user experience is central to game design. It is therefore not surprising that the main line of investigation in game analytics is player-focused [1]. Analyzing behavioral telemetry, whether through simple descriptive methods or advanced data mining techniques, allows game developers and -publishers to uncover patterns of behavior among the user base, which assists with testing and evaluating design-

and business decisions in a variety of situations [2]–[4]. In concert, game development companies obtain greater certainty about design decisions and potentially increase revenue [1].

The availability of behavioral telemetry from games enables investigations that just a decade ago would be unfeasible. Importantly, telemetry-based research has enabled game user research to move out of the laboratory and into the wild. Telemetry datasets contain information about the populations playing games (potentially dozens of millions of users for major commercial titles), rather than samples. Behavioral telemetry can be very granular (precise), detailed and potentially longitudinal, covering hours or years of gameplay, enabling an analytical precision it can be hard to achieve with other methods for user research [1], [5], [6]. One of the key challenges in game analytics is the development of methods for characterizing and predicting player behavior as it evolves over time. This is a topic that has seen increasing interest in the past few years [2], [4]. This topic is of broad interest in the games industry as it is in any other sector that has users interacting with products over shorter or longer durations. For example, being able to identify people who will use a particular product or play a game for a long time vs. those who will quickly stop, informs about the value of these different customers [7]. The temporal aspects of player behavior permits the description and monitoring of player engagement, assist with preventing future player churn and informs decision making regarding marketing- and design strategies [2]. From an academic perspective, the topic is of interest because research in the domain informs about the nature of human-game interaction.

In this paper, methods and theory from kernel archetype analysis and random process models are utilized to evaluate the playtime behavior, i.e. time spent playing specific games as a function of time, with the goal of investigating how human interest in playing a game evolves over time, and building profiles of clustered models of interest.

II. CONTRIBUTION AND MAIN RESULTS

This paper presents a large-scale analysis of the temporal patterns in player behavior across more than 3000 digital games hosted on the Steam distribution platform. The dataset we analyzed covers more than 6 million players and a combined total of over 5 billion hours of playtime. Patterns in playtime behavior for digital games have been indicated previously, and documented formally for a sample of five games by Bauckhage et al. in [2] (a precursor study to the work presented here), however, to the best knowledge of the

authors, it is the first time that behavioral telemetry from games has been analyzed at this scale. Previous publications on game analytics - whether focused on playtime or not - have focused on individual, or in a very few cases two to a handful of games [3]. The one known exception is Chambers et al. [8] who reportedly worked with over 550 games. Datasets covering thousands of games have however become accessible to researchers thanks to the ability to extract data from the Steam platform client, or for smaller numbers of games using APIs released in connection with other games (e.g. Dota2, League of Legends) or thanks to services which collect player statistics such as the p-stats network [4]. The first main contribution of the research presented is that the playtime frequency distributions across the sampled games can be modeled using the Weibull modeling [2]. This suggests that there are fundamental properties governing playtime, and by extension player engagement or *interest* in playing games, and how it evolves as a function of time, irrespective of underlying specific properties of the game in question. We refer to this as the playtime principle. The second main contribution is the application of Kernel Archetype Analysis to the built models of the playtime frequency data from the sampled games and users. The analysis shows that there are distinct clusters, or archetypes, in the playtime frequency distributions of the investigated games. These archetypal groups correspond to specific playtime distributions, which (while all following the same underlying distribution model) exhibit different initial playtimes and varying rapidity in the decay of player interest (i.e. playtime). Finally, the analysis reveals information about player behavior in games carried on the Steam platform, across a very large dataset, showing for example that the vast majority of games are played for less than 10 hours, and very few players spend more than 30-35 hours on any specific game.

III. STEAM AND DATA PRE-PROCESSING

For the work presented here, behavioral telemetry (specifically playtime per time unit) from the online game distribution and game hosting platform Steam¹ are used. Steam, developed by Valve Corporation², is one of the largest online distribution platform for games in terms of the number of active users. According to Steampowered³, the reporting service of the Steam platform, the number of concurrent users on Steam varies from around 3-7 million. The number varies as a function of time of the day, week and year. Steampowered does not inform how many games are available on the platform, but Valve reported⁴ 75 million active users in January 2014 across over 3000 games. Valve reports a growth of 30% during 2013, which marks the 10 year anniversary of the online game service.

Steam not only hosts games from various genres, but it also supports multiple platforms; those include the most widely used operating systems for desktops (including Linux) and the newly announced dedicated operating system only for the Steam platform called Steam OS *Steam OS*⁵. Although it has

TABLE I: General Statistics About the Used Dataset

Number of Analyzed Players	6,049,520
Number of Games	3,007
Total Gameplay [hours]	5,068,434,399

started to be used as an alternative operating system for the self-assembled game consoles⁶⁷, the Steam OS will be part of the future game consoles called *Steammachines*, that will be produced by numerous third party manufacturers. Steam also supports the mobile platforms which allow the players to socialize, communicate with the other players and purchase game items. In summary, Steam is an ideal platform for cross-games analytics due to the large user base, cross-platform reach and number of games hosted across most genres. The platform does not, however, contain games aimed at mobile platforms.

The dataset we use in this study - following pre-processing as detailed below - contains the total playtime distributions of games based on the game play experience of 6,049,520 Steam players that are members of the most populous 3500 communities and have public profiles. Using the by Valve provided web Application Programming Interface we have extracted player data and anonymized the player IDs by random hashing to make them untraceable. A combined total playtime of around 5 billion hours (around 580,000 years) is included in the dataset (see Tbl. I). Nowadays working with such large amount of gameplay dataset is rapidly becoming a usual practice for the most popular online games, e.g. a reported 3.3 billion hours of playtime through 2013 for the WWII tank shooter World of Tanks⁸. Similarly, Orland⁹ reported a total playtime for Dota2 on Steam of 430,000 years. According to Steam, the game is played roughly 1140 years per day. The script used to extract data from the Valve API provide readily formatted playtime telemetry that we can use to build playtime models for games. It is important to note that the dataset only covers playtime on the Steam platform, not the time spent playing games outside of that framework. The dataset appears to cover any game genre that exists on the supported platforms, including action, arcade, First Person Shooter (FPS) and strategy games. The dataset also supports games ranging from major commercial AAA-level titles to indie titles, including Antichamber, Tidalis, Dota2, and Tom Clancy's Ghost Recon Online.

A series of decisions were made during pre-processing regarding what to retain in the final dataset, which brought the initial sample of about 3,200 games down to the final 3,007 games. The description of the data set and the applied pre-processing methods are grouped under four steps that we present as follows. Firstly, the data were harvested during the Spring 2014, and there was no distinguishing between new and old Steam accounts. This means that for some players, the API will have provided complete play histories for all games played, whereas for others only part of their play histories,

¹<http://store.steampowered.com>

²<http://www.valvesoftware.com/>

³<http://store.steampowered.com/stats/>

⁴<http://www.joystiq.com/2014/01/15/steam-has-75-million-active-users-valve-announces-at-dev-days/>

⁵<http://store.steampowered.com/livingroom/SteamOS>

⁶<http://store.steampowered.com/steamos/buildyourown>

⁷<http://www.pcworld.com/article/2027390/how-to-build-your-own-steam-box-today.html>

⁸<http://n4g.com/news/1434313/world-of-tanks-780-million-players-wrecked-22-billion-tanks-in-2013-worldwide>

⁹<http://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/>

as they might be playing the game after our harvesting. This may add a bias towards making the playtime durations for individual games shorter than they actually are. To be specific, the dataset informs how much a player played a particular game from the beginning data he/she started to play the game, till the day of the data extraction. For example, with an API call for player *A*, we obtain playtime profiles detailing how much time player *A* spent on games *alpha*, *beta* and *omega*, up until the point of data extraction. Secondly, the Steam ecosystem includes tools (such as software development kits) and game demos that are free previews of the upcoming or published games to let the user have a taste of the particular game. These were removed from the dataset because they do not represent *complete* games. However, the same set of experiments were run on the dataset including and excluding demos, and there was no major difference in the results in terms of Weibull fitting. In the cluster analysis (detailed below), the demo games would generate their own archetype. Since we are interested in the global *gameplay* behavior of the players we left the analysis of the behavior of the game demos as feature work. Thirdly, games were removed where the total playtime for the games formed less than 3 hours bins, for example if a game has been played by only 10 people for {1,1,1,2,2,2,3,3,3,3} hours respectively we have not considered this game in our analysis. This is a reasonable assumption as some initial playtime information is necessary to fit the Weibull distribution to histograms. These games are thus not significant to the analysis presented. Furthermore, there was a small set of games with no playtime information, i.e. games that do not save the information about whether it has been downloaded and not played. We also eliminated such games. Fourthly, games that were played by very few people were also removed. Games with records of less than 25 players were removed. For example, Battlefield: Bad Company 2 Vietnam had only 5 recorded players in the dataset (this does not mean the games does not have other players on Steam and outside Steam at all).

Finally, it is important to note a few uncertainties about the data obtained from Steam. For any of these, estimating the potential effect is not possible with the current data access. 1) Tracking of playtime apparently did not begin until March 2009¹⁰. This can bias the results presented here if the same game is played before and after March 2009. 2) It has not been possible to verify if games that are running in minimized mode or paused are still registered as being actively played or not. It is also uncertain if Steam tracks playtime for games launched in offline mode. 3) There are informal reports of discrepancies between the playtime reported by the Steam client and the impressions of players. Potential reasons include the same game having multiple versions, platform differences, or mode variations. 4) An unknown fraction of Steam users have actively selected to have a private profile. This does not impact directly on the results presented here, but if users with a private profile have a substantially different playtime distribution than those with a public profile, the generalizability of the results maybe impacted. This is impossible to estimate without a comparative analysis.

IV. RELATED WORK

Cross-games analysis is an approach that remains comparatively rare in game analytics, in part due to the relative young age of the domain for research and development, and in part due to the lack of access to behavioral datasets covering multiple games. That being said, in recent years there have been publications from variety of research studies, indicating that behavioral features such as playtime, session time and inter-session time follow distinct patterns within games. With a few exceptions, research published in game analytics contain only data from one or a small number of games as done notably in [2], [3], [8]. While major game companies do have access to cross-game telemetry datasets, these are generally considered confidential [5]. Understandably, industry reports, white papers and presentations in the area therefore tend to be high-level rather than specific. Some reports by analytics companies contain cross-games information, but generally contain only descriptive measures of user behavior and often focus on monetization and marketing aspects of user behavior. However, over the past years a number of studies have been published that take advantage of behavioral telemetry that can be harvested from online services such as Gamespy¹¹ [8], Steam [9] and the P-stats network¹² [3], or by directly logging the server-client data stream in massively multi-player online games (MMOGs) [10].

One of the earliest publications on playtime in games is done by Chambers et al. [8], who focused on network support for online games and quality of service. Among the results of interest here is that session time distributions for a Counter-Strike server, *cs.mshmo.com*, were reported to follow a Weibull distribution. The authors noted the difference with this behavioral pattern against the more heavy-tailed distributions reported for internet traffic at the time. They also investigated game popularity, but only for the top 50 games in now-defunct GameSpy service, using average number of players per day and noted a power law distribution. Bauckhage et al. [2] specifically focused on playtime distributions across five major commercial titles, covering 250,000 players. The authors examined a number of random process models in order to investigate which distribution model fits playtime frequency data (not session data), reporting the Weibull as the overall good fit across the investigated games. Bauckhage et al. [2], targeting the specific question of modeling player interest, or engagement, as a function of playtime, noted that the psychological drivers behind human-game interaction are abstract and cannot be measured from raw data, e.g. user experience, which is a mental construct that can only be indirectly measured using elicitation techniques such as psycho-physiological measures, surveys or interviews. A central conclusion that can be derived from this is that, even in situations where reliable temporal behavioral patterns can be found, explaining why they occur from telemetry alone is not always a simple task [11], [12].

Within network analysis, a number of studies have focused on online games. For example, Tarnig et al. [13] focused on player departure and the ability to predict player churn. Using support vector machines, players were classified, and their playing pattern modeled within the MMOG Shen-Zhou Online.

¹⁰<http://forums.steampowered.com/forums/showthread.php?p=10247483>

¹¹<http://www.gamespy.com/>

¹²<http://p-stats.com/>

Suznjevic et al. [10] broke up the player actions for a sample of data from the MMORPG World of Warcraft, in order to investigate perceived quality and network support. While the authors do report on playtime data, they include information about time spent on different types of actions, including questing, trading and raiding. Furthermore, they report a cumulative distribution function (CDF) of session durations which bears similarity to Tarng et al. [13]. Pittman and GauthierDickey [14] investigated player distribution in two large-scale MMORPGs, World of Warcraft and Warhammer Online. The authors fit session length data to a Weibull distribution, reporting similar good fits as those reported by Chambers et al. [8]. Feng et al. [15] reported that the distribution of the number of sessions that a person plays before quitting the sci-fi themed MMORPG Eve Online fits a Weibull distribution closely. This means that most players do not stay long in the game, but that the distribution has a long tail with a small fraction of people that play for a very long time. Weber et al. [16] modeled player retention as a regression problem, specifically looking at game features and their effect on retention. Lim [17] reported on an analysis of “several dozen freemium games”, that player behavior is better approximated by a power law than a normal distribution. The report focused not only on playtime, but also e.g. time to conversion from non-paying to paying user. Lim [17] highlights that doubling the player base does not necessarily double revenue, which was one of the early indications of the importance of differentiating between different types of users when considering user acquisition in the game industry. In a similar vein, Nozhnin [11] reports on a case study focusing on predicting player churn in the game Aion, noting that playtime frequency distributions are useful for predicting player departure.

V. KERNEL ARCHETYPAL ANALYSIS

Archetypal Analysis is a clustering technique based on selecting extreme basis vectors, that are called *archetypes*, with convexity constraints on belongingness values. That is, the main objective of archetypal analysis is to find descriptive and prototypical extreme values to summarize and reconstruct the dataset. In game mining research the variants of the method have been used to cluster individual game play behavior [3], [4], [18], group behavior [19], [20] and generate human-like motion [21]. To our knowledge, in the context of game data mining this is the first study that uses kernel archetypal analysis to analyze the interest of the players. In this section we give a general introduction to Archetypal Analysis and explain how it can be kernelized.

Given a column data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ with n data entries and m dimensional features and an integer k which is smaller than n (i.e. $k \ll n$), archetypal analysis finds k archetypes $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k] \in \mathbb{R}^{m \times k}$ and a stochastic coefficient matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k] \in \mathbb{R}^{k \times n}$ for soft clustering.

Casting the problem as a matrix factorization problem, we reconstruct the data matrix as

$$\mathbf{X} \approx \mathbf{Z}\mathbf{A} = \mathbf{X}\mathbf{B}\mathbf{A} \quad (1)$$

where $\mathbf{B} \in \mathbb{R}^{n \times k}$ is a stochastic column matrix. Hence the main problem of archetypal analysis becomes finding appropriate matrices for \mathbf{B} and \mathbf{A} . Formally, this can be shown

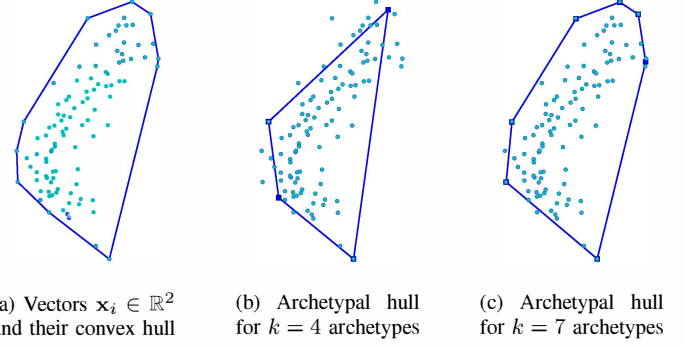


Fig. 1: Archetypal analysis approximates the convex hull of a set of multivariate data. Increasing the number k of archetypes improves the approximation.

as a constrained quadratic optimization problem to minimize the residual sum of squares (RSS) as

$$\min_{\mathbf{A}, \mathbf{B}} RSS = \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}\|^2 \quad (2)$$

with the following constraints

$$b_{ij} \geq 0 \wedge \sum_{i=1}^n b_{ij} = 1 \wedge a_{ji} \geq 0 \wedge \sum_{j=1}^k a_{ji} = 1. \quad (3)$$

Therefore, having found appropriate matrices \mathbf{A} and \mathbf{B} we can approximate each data element as

$$\mathbf{x}_i \approx \mathbf{Z}\mathbf{a}_i = \sum_{j=1}^k \mathbf{z}_j a_{ji} \quad (4)$$

where we define the archetypes as

$$\mathbf{z}_j = \mathbf{X}\mathbf{b}_j = \sum_{i=1}^n \mathbf{x}_i b_{ij}. \quad (5)$$

That is, archetypal analysis represents the data elements as convex combinations of the archetypes which are themselves convex combinations of the data elements. That gives us easily interpretable and prototypical vectors to describe the whole dataset.

Having identified the possible values of k , Cutler and Breiman [22] show for $k = 1$ the minimizer of (2) is the data mean, where as for $k = n$ the data elements as archetypes are the global minimizers of the RSS, where its value is zero. For values in between (i.e. $1 < k < n$), the archetypes reside in the data convex hull and increasing the number of archetypes approximates the the data convex hull (see Fig. 1).

As mentioned earlier finding archetypes in general is based on minimizing the RSS in (2). Cutler and Breiman [22] proposed an alternating least squares algorithm that solves convex least squares problems to find archetypes and the coefficients. The complexity of this method increases cubically with respect to the number of data elements. In order to achieve speed-up, Thureau et al. [23] introduced Simplex Volume Maxi-

Algorithm 1 Kernel Archetypal Analysis

Calculate the kernel matrix: $\mathbf{K} = \mathcal{K}(\mathbf{X}^T \mathbf{X})$

Initialize \mathbf{A} and \mathbf{B}

while Stopping condition is not satisfied **do**

$$\mathbf{A}_{t+1} \leftarrow \mathbf{A}_t - \eta_A \left(2 \left[\mathbf{B}_t^T \mathbf{K} \mathbf{B}_t \mathbf{A}_t - \mathbf{B}_t^T \mathbf{K} \right] \right)$$

$$\mathbf{B}_{t+1} \leftarrow \mathbf{B}_t - \eta_B \left(2 \left[\mathbf{K} \mathbf{B}_t \mathbf{A}_{t+1} \mathbf{A}_{t+1}^T - \mathbf{K} \mathbf{A}_{t+1} \mathbf{A}_{t+1}^T \right] \right)$$

end while

mization which is a linear time variant of Archetypal Analysis relaxing the constraints of archetypes by finding archetypes from data elements that maximize the data simplex. Similar distance geometry based approaches have been introduced for archetypal analysis to handle very large data sets [20], [24].

Analyzing another approach, Morup and Hansen [25] approached the problem of finding archetypes as a gradient descent based search problem which also allows kernelization of the method. They introduced an alternating projected gradient descent approach to find the archetypes. Namely, using the equality

$$\begin{aligned} RSS &= \left\| \mathbf{X} - \mathbf{XBA} \right\|^2 \\ &= \text{tr} \left[\mathbf{X}^T \mathbf{X} - 2 \mathbf{X}^T \mathbf{XBA} + \mathbf{A}^T \mathbf{B}^T \mathbf{X}^T \mathbf{XBA} \right] \end{aligned} \quad (6)$$

the gradients can be calculated as

$$\frac{\partial RSS}{\partial \mathbf{A}} = 2 \left[\mathbf{B}^T \mathbf{X}^T \mathbf{XBA} - \mathbf{B}^T \mathbf{X}^T \mathbf{X} \right] \quad (7)$$

and

$$\frac{\partial RSS}{\partial \mathbf{B}} = 2 \left[\mathbf{X}^T \mathbf{XBA} \mathbf{A}^T - \mathbf{X}^T \mathbf{XA} \mathbf{A}^T \right]. \quad (8)$$

Having initialized matrices \mathbf{A} and \mathbf{B} randomly and defined the step sizes η_A and η_B , the algorithm finds local optimal coefficients by updating the matrices in an alternating fashion as

$$\mathbf{A} \leftarrow \mathbf{A} - \eta_A \frac{\partial RSS}{\partial \mathbf{A}} \quad \text{and} \quad \mathbf{B} \leftarrow \mathbf{B} - \eta_B \frac{\partial RSS}{\partial \mathbf{B}}. \quad (9)$$

We can show that the above method can be kernelized by observing the occurrences of Gram matrix $\mathbf{G} = \mathbf{X}^T \mathbf{X}$ in equations (7), and (8). We can use the kernel trick by replacing every occurrence of the inner products $\mathbf{x}_i^T \mathbf{x}_j$ by a kernel function $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ of our choice. This gives us the power of using kernel methods to find the ideal coefficient values for the archetypes. The essential steps of kernel archetypal analysis are demonstrated in Alg. 1.

VI. PLAYER INTEREST MODELING THROUGH WEIBULL DISTRIBUTION FOR CLUSTERING GAMES

Analyzing the total interest (or engagement) of the players within the lifetime of games through total spent time (as a proxy measure of interest) gives a valuable insight about the behavior of the players. One of the challenges for such analysis is the varying length of the total playtime per game. Namely, each game is planned and produced to be played differently

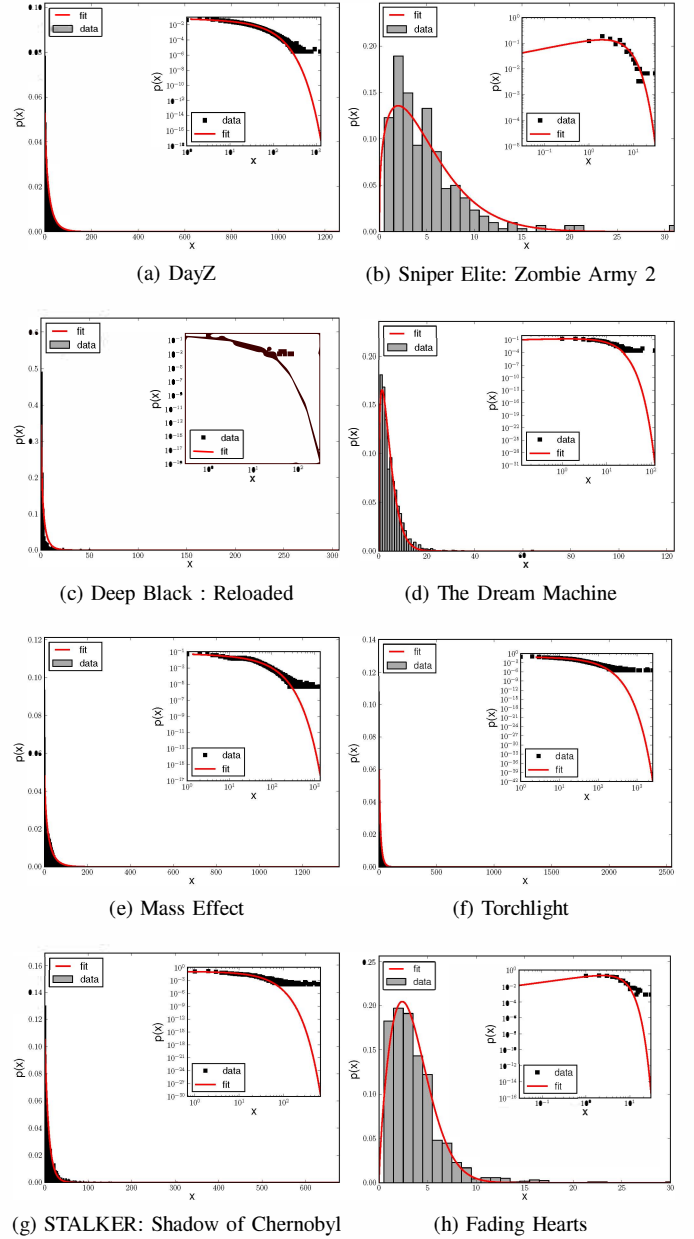


Fig. 2: Examples of playtime frequency distributions and their corresponding parametrized Weibull modeling for Steam games. Weibull distribution provides a general approach to model the interest in games from variety of genres and different time ranges.

resulting in different lifetime profiles. In this section we show how our theoretically founded interest modeling approach can be used to model the playtime distributions for each game which will at the end allow us to group the individual models to come up with prototypical playtime profiles using kernel archetypal analysis.

We have previously presented the theoretical foundations of modeling the gameplay interest using Weibull distribution and showed empirical results in [2]. The probability density function of the Weibull distribution is defined for $t \in [0, \infty)$

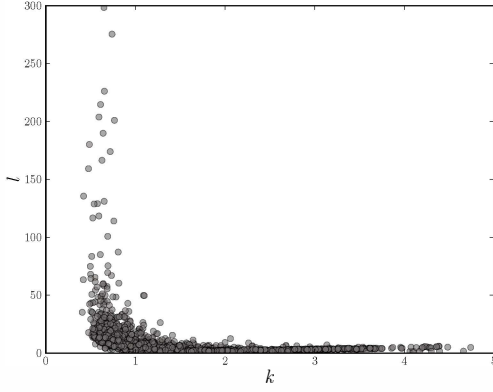


Fig. 3: Reducing the dimensionality of the playtime frequency distributions of the analyzed 3007 Steam games to the non-linear Weibull parameter space. In this way we have a non-linear embedding of the game distributions.

and given by

$$f(t | k, l) = \frac{k}{l} \left(\frac{t}{l}\right)^{k-1} \exp \left[-\left(\frac{t}{l}\right)^k \right] \quad (10)$$

where $k > 0$ and $l > 0$ are shape and scale parameters respectively. Following that analogy, we fitted Weibull distribution to the total playtime histograms we obtained from the Steam games using Maximum Likelihood Estimation as shown in [2] and obtained an average Kolmogorov Smirnov test statistic value of 0.15 (1.0 is the possible maximum value) with standard deviation of 0.06 for all of the 3007 total game distributions we have analyzed. Fig. 2 shows example games with their Weibull models.

At this point, modeling the player interest using the Weibull distribution has two advantages: *generality* and *dimensionality reduction*. Firstly, as inherently the playtime frequency distributions for games are of various lengths (see Fig. 2 for example distributions), common techniques fail to provide a way to have a comparable representation of the distributions. Whereas, using our model we can have a framework to compare players' interest for every game. Secondly, our model provides a non-linear embedding through the shape and scale parameter of the fitted models. This can lead to computationally efficient techniques to identify groups of games and prototypical gameplay profiles in this space.

We show the 2 dimensional embedded game models in Fig. 3, in which every point represents a Weibull fit to the total playtime frequency distribution. Here every model parameter represents the characteristics of the playtime experience of its corresponding game. In this way, we can use divergence measures to compare two given game models. It is important to note that this space is not truly metric, that is, the use of Euclidean norms to compare the games does not always yield true results (see the probability density function in (10)). In order to compensate that, we used a *Kullback-Leibler* (KL) divergence based kernel function [26] that uses KL divergence for measuring distinction among the data elements. KL divergence between two continuous distributions p and q

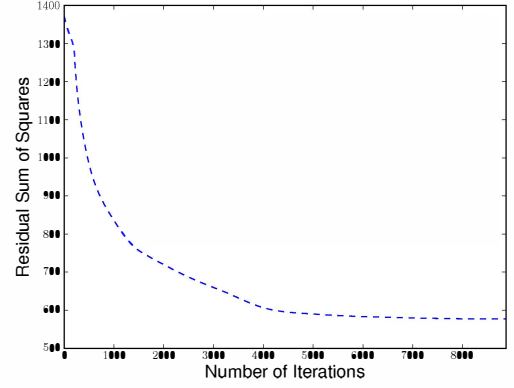


Fig. 4: The change in the values of the Residual Sum of Squares to find the archetypes using kernel archetypal analysis for this study. For our case the method converges after the 8866th iteration.

is defined as

$$D_{KL}(p, q) = \int p(x) \log[p(x)/q(x)] dx. \quad (11)$$

As shown in [27], the KL divergence between two Weibull distributions $f_1(x|k_1, l_1)$ and $f_2(x|k_2, l_2)$ is defined as

$$D_{KL}(f_1, f_2) = \log \frac{k_1 l_2^{k_2}}{l_1^{k_1} k_2} + (k_1 - k_2) \left[\log l_1 - \frac{\gamma}{k_1} \right] + \left(\frac{l_1}{l_2} \right)^{k_2} \Gamma \left(\frac{k_2}{k_1} + 1 \right) - 1. \quad (12)$$

So as to compare the probability density functions from the learned parameters, using the Symmetric Kullback-Leibler (SKL)

$$D_{SKL}(p, q) = D_{KL}(f_1, f_2) + D_{KL}(f_2, f_1) \quad (13)$$

we define a kernel [26]

$$\mathcal{K}_{SKL}(f_1, f_2) = e^{-V D_{SKL}(f_1, f_2) + W} \quad (14)$$

where V and W are the scale and shift parameters respectively. Therefore, for each occurrence of the Gram matrix $\mathbf{X}^T \mathbf{X}$ in the projected gradient descent algorithm (see Alg. 1) we replace the values with the ones obtained from the Symmetric Kullback-Leibler kernel \mathcal{K}_{SKL} .

In the next section we show the results of our experiments where we modeled the playtime distributions of 3007 games and applied archetypal analysis to obtain clusters and prototypical gameplay models.

VII. RESULTS AND DISCUSSION

We have run the kernel archetype analysis in Alg. 1 to the results obtained from the Weibull modeling of the 3007 games. Having tested the method with different numbers of basis vectors, we found the results with 4 archetypes presented a reasonable balance between reconstruction error difference and explanatory strength. Running the algorithm with $k = 4$, we obtained convergence after the 8866th iteration (see Fig. 4 for the decrease of the RSS values). We obtained 4 different archetypal game play profiles that we show in Fig. 5 and

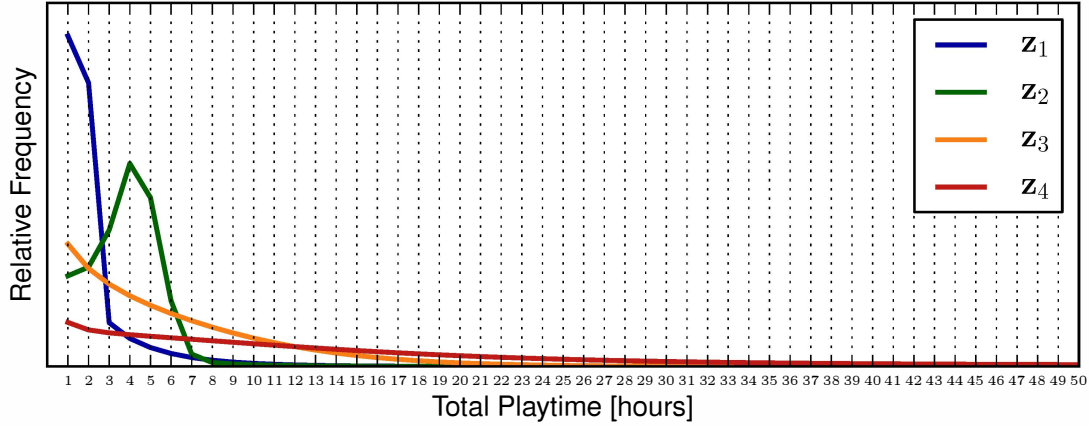


Fig. 5: Calculated archetypes from interest models we created using the total playtime frequency distributions of 3007 games that have been hosted on the Steam platform. The found archetypes represent a set of prototypical game play behavior and hence they show an approximate view of interest evolution in games. The archetype shown by z_1 indicates the short-lived games while the archetype shown by z_4 shows the gameplay behavior of AAA games. In between we observe the archetype z_3 that has a slowly decaying profile compared to z_1 but the likelihood of gameplay hits zero earlier than z_4 . The archetype z_2 draws a totally different picture than the other three archetypes that it does not have a monotonic decrease and a peak at 4 hours. Across all four archetypes, there are very few people who play a game for more than 30-35 hours.

present their content and the hard clustering results in Tbl. II. The profiles we show give us a broad overview of how the general interest of more than 6 million players is distributed among the games we analyzed.

TABLE II: Kernel Archetypal Analysis Results

Archetype	Common Game Types	Representation
z_1	FPS, Action and Indie	44.1%
z_2	F2P RPG and Adventure	10.2%
z_3	Adventure and Point & Click	22.4%
z_4	AAA Games	23.3%

Furthermore, it is important to note that, the optimal coefficient matrix A can be used to segment the dataset in a soft clustering manner. Particularly, every data element x_j in our dataset has corresponding probability vector which is a stochastic vector and contains the belongingness distribution to the found archetypes. As these belongingness vectors are stochastic, we can show them in the archetypal simplex which allow us to visually see the belongingness distribution for the data elements as shown in Fig. 6.

While a variety of games are contained within each archetype cluster, there are some similarities which we describe in the following. Starting with the first archetype z_1 , which represents 44.1 % of the dataset, we observe a very fast decrease in the total amount of time spent playing, with a total *life time* of a couple of hours for the games that form this cluster. The games contained within this archetype are primarily action-oriented titles, indie (minor commercial) titles and First Person Shooter (FPS) games. The second archetype z_2 represents 10.2% of the data and it is primarily comprised of free-to-play real time strategy- and casual games. This archetype differs from the others as it exhibits a later peak rather than having a monotonically decreasing profile

showing specifically a rising interest in gameplay till 4 hours and a fast decrease afterwards, where the likelihood after 15 hours is significantly low. Additionally, comparing the relative frequencies, this profile can be interpreted as a gameplay behavior that requires the players to invest a considerable amount of time (till 4 hours) before they decide to quit.

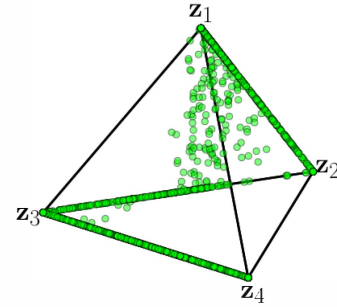


Fig. 6: Simplex Plot shows the distribution of the belongingness values for every game we have analyzed. The models are represented as convex combinations of the found archetypes which allows soft clustering.

The third archetype z_3 represents 22.4% of the games and it consists of adventure and point & click games that have a rather slower decay than the above archetypes. Finally, z_4 (23.3%) primarily contains major commercial, AAA, titles. These titles exhibit the slowest playtime decay, and cover a variety of genres including FPS games, Third Person Shooter games and sports games. It is important to note that even though we observe a slow decrease across total playtime for some archetypes, the majority of players across any game in the sample play for less than 30-35 hours and quit playing afterwards.

VIII. CONCLUSION AND FUTURE WORK

In this paper, a large-scale analysis of patterns in playtime frequency distributions, as well as archetypal patterns in these distributions, has been presented for a dataset covering more than 3000 PC and console games, over 6 million users, encompassing over 5 billion hours of playtime. The presented research is to the best knowledge of the authors the first to cover thousands of games, and the first to introduce kernel archetypal analysis in the context of mining player behavior data. The work presented substantially advances the state-of-the-art of knowledge about player behavior in games, providing empirical evidence for the presence of a specific pattern in playtime frequency distributions, which is here tentatively referred to as the playtime principle. Following fitting of the Weibull random distribution model to the games in the dataset, kernel archetypal analysis was employed to identify four prototypical playtime profiles. Each cluster of games show different playtime frequency distribution patterns, with varying *tail thickness* and decay rates.

Future work will provide a detailed analysis of each cluster and the features that separate games with a large segment of extended-play users from those with rapid decays in the user base across just a few hours of playtime. Additionally, we are aiming to investigate how the playtime behavior of the demo files differ from the playtime behavior of the games and if correlations in between can be found for future prediction of success or failure of the games. The choice of the Weibull distribution was inspired by the precursor study to the one presented here by Bauckhage et al. [2], who tested the fit of the first passage time distributions of different random process models and detailed the theoretical basis for why the Weibull model provides a good fit for the psychological processes of engagement in games.

It was also observed that in general the overall interest of the players does not extend beyond a 30-35 hours threshold of total playtime, with many games showing massive churn rates after just a few hours of gameplay. While the work presented here did not evaluate the frequency patterns of session times, as has been the focus of previous research, e.g. [8], [14], the work of these previous studies collectively lend support to the idea that session time frequencies follow a similar pattern as playtime frequencies. Future work on the dataset used here will investigate this hypothesis.

ACKNOWLEDGMENTS

The work reported in this paper was carried out within the Fraunhofer / University of Southampton research project *SoFWIREd*. The authors gratefully acknowledge this support.

REFERENCES

- [1] M. Seif El-Nasr, A. Drachen, and A. Canossa, *Game Analytics: Maximizing the Value of Player Data*. Springer, 2013.
- [2] C. Bauckhage, K. Kersting, R. Sifa, C. Thureau, A. Drachen, and A. Canossa, "How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing Times," in *Proc. IEEE CIG*, 2012.
- [3] A. Drachen, R. Sifa, C. Bauckhage, and C. Thureau, "Guns, swords and data: Clustering of player behavior in computer games in the wild," in *Proc. IEEE CIG*, 2012.
- [4] R. Sifa, A. Drachen, C. Bauckhage, C. Thureau, and A. Canossa, "Behavior Evolution in Tomb Raider Underworld," in *Proc. IEEE CIG*, 2013.
- [5] J. H. Kim, D. V. Gunn, E. Schuh, B. Phillips, R. J. Pagulayan, and D. Wixon, "Tracking Real-time User Experience (TRUE): A Comprehensive Instrumentation Solution For Complex Systems," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, 2008.
- [6] A. Drachen and M. Schubert, "Spatial Game Analytics and Visualization," in *Proc. IEEE CIG*, 2013.
- [7] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky, "Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry," *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 690–696, 2000.
- [8] C. Chambers, W. Feng, S. Sahu, and D. Saha, "Measurement-based Characterization of a Collection of On-line Games," in *Proc. of ACM SIGCOMM Conf. on Internet Measurement*, 2005.
- [9] C. Lim and F. Harrell, "Modeling Player Preferences In Avatar Customization Using Social Network Data: A Case-study Using Virtual Items in Team Fortress 2," in *Proc. IEEE CIG*, 2013.
- [10] M. Suznjevic, O. Dobrijevic, and M. Matijasevic, "MMORPG Player Actions: Network Performance, Session Patterns and Latency Requirements Analysis," *Multimedia Tools and Applications*, vol. 45, no. 1-3, pp. 191–214, 2009.
- [11] D. Nozhnin, "Predicting Churn: Data-Mining Your Game," 2012. [Online]. Available: http://gamasutra.com/view/feature/170472/predicting_churn_datamining_your_php/
- [12] A. Drachen and A. Canossa, "Evaluating Motion: Spatial User Behaviour in Virtual Environments," *International Journal of Arts and Technology*, vol. 4, no. 3, pp. 294–314, 2011.
- [13] P. Tarnag, K. Chen, and P. Huang, "On Prophesying Online Gamer Departure," in *Proc. of IEEE NetGames*, 2009.
- [14] D. Pittman and C. GauthierDickey, "Characterizing Virtual Populations in Massively Multiplayer Online Role-playing Games," in *Proc. of Int. Conf. on Advances in Multimedia Modeling*, 2010.
- [15] W. Feng, D. Brandt, and D. Saha, "A Long-term Study of a Popular MMORPG," in *Proc. of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*, 2007.
- [16] B. Weber, M. John, M. Mateas, and A. Jhala, "Modeling Player Retention in Madden NFL 11," in *Proc. Innovative Applications of Artificial Intelligence*, 2011.
- [17] N. Lim, "Freemium Games Are Not Normal," 2012. [Online]. Available: http://www.gamasutra.com/blogs/NickLim/20120626/173051/Freemium_games_are_not_normal.php/
- [18] A. Drachen, R. Sifa, C. Thureau, and C. Bauckhage, "A Comparison of Methods for Player Clustering via Behavioral Telemetry," in *Proc. SASDG FDG*, 2013.
- [19] C. Thureau and C. Bauckhage, "Analyzing the Evolution of Social Groups in World of Warcraft," in *Proc. IEEE CIG*, 2010.
- [20] C. Thureau, K. Kersting, and C. Bauckhage, "Convex Non-negative Matrix Factorization in the Wild," in *Proc. ICDM*, 2009.
- [21] R. Sifa and C. Bauckhage, "Archetypal Motion: Supervised Behavior Learning Using Archetypal Analysis," in *Proc. IEEE CIG*, 2013.
- [22] A. Cutler and L. Breiman, "Archetypal Analysis," *Technometrics*, vol. 36, no. 4, 1994.
- [23] C. Thureau, K. Kersting, and C. Bauckhage, "Yes We Can: Simplex Volume Maximization for Descriptive Web-Scale Matrix Factorization," in *Proc. ACM CIKM*, 2010.
- [24] K. Kersting, M. Wahabzada, C. Thureau, and C. Bauckhage, "Hierarchical convex NMF for clustering massive data," in *Proc. Asian Conf. on Machine Learning*, 2010.
- [25] M. Morup and L. Hansen, "Archetypal Analysis for Machine Learning and Data Mining," *Neurocomputing*, vol. 80, pp. 54–63, 2012.
- [26] P. Moreno, P. Ho, and N. Vasconcelos, "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications," in *Proc. NIPS*, 2003.
- [27] C. Bauckhage, "Computing the Kullback-Leibler Divergence between two Weibull Distributions," *arXiv:1310.3713 [cs.IT]*, 2013.