Behavior Evolution in Tomb Raider Underworld

Rafet Sifa*†, Anders Drachen*§, Christian Bauckhage†‡, Christian Thurau* and Alessandro Canossa§
*Game Analytics Copenhagen, Denmark
†University of Bonn,Bonn, Germany
‡Fraunhofer IAIS, Sangt Agusitin, Germany
§Northeastern University, Boston, USA

***Abstract*—Behavioral datasets from major commercial game titles of the "AAA" grade generally feature high dimensionality and large sample sizes, from tens of thousands to millions, covering time scales stretching into several years of real-time, and evolving user populations. This makes dimensionality-reduction methods such as clustering and classification useful for discovering and defining patterns in player behavior. The goal from the perspective of game development is the formation of behavioral profiles that provide actionable insights into how a game is being played, and enables the detection of e.g. problems hindering player progression. Due to its unsupervised nature, clustering is notably useful in cases where no prior-defined classes exist. Previous research in this area has successfully applied clustering algorithms to behavioral datasets from different games. In this paper, the focus is on examining the behavior of 62,000 players from the major commercial game Tomb Raider: Underworld, as it unfolds from the beginning of the game and throughout the seven main levels of the game. Where previous research has focused on aggregated behavioral datasets spanning an entire game, or conversely a limited slice or snapshot viewed in isolation, this is to the best knowledge of the authors the first study to examine the application of clustering methods to player behavior as it evolves throughout an entire game.**

I. Introduction

Over the past decade, principles from business intelligence have gained substantial traction in commercial game development [1], [2]. This for several reasons, for example to inform decision making across operational and strategic levels. The advantages of data-driven development in the game industry as applied to project management and game user research are not new however, but have had a more limited applicability given the traditional retail-based business model, and the limited capacity for data collection on user behavior, performance and processes. With the increasing availability of data on all aspects of game development and economics, combined with the introduction of business models such as Free-to-Play (F2P) which break with the retail-based paradigm and fundamentally require analytics to operate, business intelligence methods have become a topic of substantial interest in commercial game development [1]–[3].

The interest in business intelligence in game development has notably been described and debated within the area of user behavior analytics [1]–[6]. Over the past decade, principles from fields such as user research, web analytics [7] and geomarketing have been adopted and adapted for game development, e.g. for analyzing user-game interaction, monetization/purchasing behavior and social behavior. These analyses operate on increasingly larger datasets, obtained from client-side or server-side logging, i.e. operating in "the wild". A central challenge is dimensionality: Games range in complexity from the relatively simple to very complex information systems supporting millions of users and thousands of potential user actions and system responses [8]–[10]. Given the complexity of some game forms in terms of the mechanics of the underlying systems, data mining methods which are able to reduce the complexity of the behavioral datasets, and provide actionable insights driving game design, are of interest [1], [11], [12]. Interpretability and reliability of results is vital, as decisions based on them affect game design and thus ultimately the revenue.

II. Contribution and main result

In this paper, a step is taken towards addressing the challenge of obtaining actionable insights from unsupervised behavioral clustering in "AAA" level major commercial computer games, based on a 62,000 player behavior dataset from the AAA-level Tomb Raider: Underworld (TRU). The research presented focuses on developing and evaluating behavioral clusters as they evolve during a game, thereby advancing the current state-of-the-art, which is focused on running behavioral analysis based on aggregate datasets covering entire games, or smaller segments viewed in isolation from the rest of the content of a game.

The main contribution is the presentation and evaluation of a method for discovering behavioral clusters as they evolve during a game, and discussion of how to describe cluster results in a way that is meaningful to game designers. The method presented is based on Simplex Volume Maximization (SIVM) [13], an adaptation of Archetype Analysis (AA) that is applicable to large-scale datasets [14]. SIVM was applied to playtime distribution of the players per level and to each of the seven main levels of Tomb Raider: Underworld, resulting in behavioral clusters or groups, which can be described in terms of design language [5], [11], [15], and which describe how player behavior groups evolve during the game, for example in response to level design changes and learning effects.

III. Tomb Raider

In Tomb Raider: Underworld, as any of the previous and later games in the series, the player takes control of the games main protagonist, Lara Croft. Croft is thematically a combination of an action heroine and Indiana Jones. The games are a combination of adventure games and 3D platformer. Historically the emphasis has been on navigation and exploration, but with Underworld (8th game) the series took on a greater emphasis on fighting various types of enemies using a range of different weapons. The Tomb Raider series

978-1-4673-5311-3/13/$31.00 ©2013 IEEE

of games have been published on multiple different platforms, including mobile devices. In Underworld, Lara Croft travels to a number of exotic locations such as Thailand and the Jan Van Mayen islands.

The gameplay centers around entering forgotten tombs, gradually exploring a linear storyline, and solving about two hundred puzzles along the way. The game is played via a third-person camera to facilitate the 3D platformer gameplay. The game consists of seven game levels plus a (skippable) prologue. Challenges in the game stem from solving navigational puzzles, strategic route planning and fighting enemies using limited resources. The primary and virtually ever-present danger in the game is falling, but the game also provides different types of environmental hazards (e.g. traps) and different types of mobile AI-controlled enemies.

IV. Related work

Behavior analysis in computer games is a topic of interest in both game development and game-related research, e.g. AI [4]–[6], analytics, experience modeling, learning/serious games and game psychology [1], [2]. Segmentation analysis, cohort analysis, funnel analysis, clustering and classification are methods that see widespread use due to their ability to dissect a population of players according to their behavior, in order to drive design decisions or inform agent behavior, adaptive games etc. [11], [16]. The state-of-the-art (SOTA) of clustering and classification techniques in game industry and game research was recently provided by Drachen et al. [14], and will therefore only be covered briefly here, focusing on the industry side.

Industry: With regards to the state-of-the-art of behavior analysis in the game industry, this is as noted by [14] an area that is difficult to evaluate due to both data and the methods employed to analyze them being considered proprietary. In essence, analytics practices have become a means for gaining a competitive edge in an already competitive marketplace for interactive entertainment, and this discourages knowledge sharing [2]. Early work in adopting analytics methods for game user behavior analysis was championed by Microsoft Studios Research and applied to e.g. the Halo series of games [10], [17]. In the past years behavioral analytics has spread to the rest of the industry and all major publishers have dedicated analytics teams, although the details of the methods used are kept confidential. More recently, game industry events such as the Game Developers Conference [15], industry magazines, blogs and news sites, as well as middleware analytics tool providers, provides some insights into the general SOTA but not specific algorithms used. For example, most middleware providers in the field provide functionality for segmentation, funnel analysis and cohort analysis, but not clustering or classification. Services such as Playnomics, Game Analytics and Games Analytics offer what appear to be more advanced forms of player segmentation and behavioral prediction, but the services are black boxes, and therefore not possible to evaluate. Most middleware segmentation tools use pre-defined classes, generally linked with monetization [1], [2]. This approach can be useful but has the inherent problem of fitting data to classes that may not exist in the dataset. This is especially problematic in games of a persistent nature, where the population of players change over time. Using pre-defined features prevent dynamic exploration of the dataset, and thus risks missing patterns of behavior. The use of pre-defined classes for segmentation should at the least be checked against unsupervised cluster analysis. Recently, some major game publishers have teamed up with academic research teams to investigate game data mining, including e.g. Ubisoft, EA, Sony and Square Enix [2], and research publications are emerging which are investigating e.g. server load/network effects, gameplay, social systems etc. [18]. Some of these research publications, which rests on industry data, are described below. In recent years, the emergence of new business models such as F2P has increased the requirements for data-driven input in decision making processes [2], [19].

Academia: Within research fields focused on game analytics, game AI, agent modeling, adaptive games, social network analysis, communication studies and player experience research, the use of behavioral telemetry has been in use for over a decade. Research focusing on clustering and classification remains infrequent. Categorizing players into behavioral types has been an important topic in game research for decades, and since the seminal essay by Bartle [20], which divided players into four types based on the authors personal experience, has generated a number of attempts to develop player behavior categories, initially from survey data but increasingly from in-game behavioral telemetry, e.g. Harrison and Roberts [21] working with achievement data from World of Warcraft, or Weber and Mateas [12], who employed a series of classification algorithms for recognizing player strategy in StarCraft. Thurau and Bauckhage [22], applied Convex-Hull Non Negative Matrix factorization to categorize player guilds in World of Warcraft. Using player experience level distribution of the guilds the authors extracted 8 different types of group behavior. Drachen [14] used k-means and Simplex Volume Maximization to create behavioral profiles from telemetry data of two commercial Massively Multiplayer Online Role Playing Games (MMORPG) and First Person Shooter (FPS) game. Ducheneaut and Moore [23], examined group player behaviors in the MMORPG Star Wars Galaxies via action frequency analysis. Finally, Drachen et al. [11] used Self-Organizing Networks to identify four clusters of player behavior for 1365 players in Tomb Raider: Underworld. These were converted into behavioral profiles for the developers of the game.

In summary, there is as yet no substantial body of knowledge freely available, to guide the application of game data mining to behavioral telemetry. This includes clustering and other forms of user grouping techniques. A key concern is interpretability, which is important given the varied nature of the stakeholders who are on the receiving end of game analytics, including designers, producers, managers, programmers, QA, marketing and user research [2]. Ideally, an expressive label should be assignable to groups, however, there is no objective criterion available which defines what a descriptive representation is. Here the operational definition is that results are interpretable when they embed data whose basis vectors correspond to actual data points [11], [24].

V. Mining Player Behavior Using Archetypal Analysis

There are numerous supervised and unsupervised techniques from machine learning and pattern recognition to analyze the player behavior from the player telemetry. Clustering algorithms provide a way to analyze such data in an unsupervised manner to yield hidden patterns. Introduced by Cutler and Breiman [25], Archetypal Analysis is a soft clustering method that allows us to describe the data entities using convex combination of extreme entities called archetypes. Given a dataset with d dimensions and n samples as a column matrix $V^{d\times n}$, Archetypal Analysis aims to find a set of archetypes $W^{d\times k}$ where k is a non negative integer with $k \ll n$ and a set of non-negative coefficients vectors $H^{k\times n}$ that contain the stochastic belongingness values of each point to the archetypes with the property $1^T h_j = 1$. Interpreting this as a matrix factorization problem, we aim to find matrices W and H to minimize the Frobenius norm $\|V - WH\|$ which quantifies how well the dataset is approximated.

Algorithm 1 Simplex Volume Maximization

Select x_i randomly from X
Choose the first basis vector:
$w_1 = \text{argmax}_l\, dist(x_l, \text{argmax}_z\, dist(x_i, x_z))$
for index $i \in [2, k]$ **do**
  Let S_{i-1} be the current simplex with $i-1$ vertices.
  Find the vertex that maximizes:
  $w_i = \text{argmax}_q\, Vol(S \cup x_q)$
  Update $S_i = S_{i-1} \cup w_i$.
end for

Simplex Volume Maximization Algorithm (SIVM) is a highly scalable linear time approach to find archetypes proposed by Thurau et. al. [13]. Restricting the archetypes to be data entities, SIVM iteratively finds the archetypes by fitting a simplex to the data with maximum volume using Cayley-Menger Determinant. Namely, given a dataset the main intention of the algorithm is to find data entities that maximizes the volume of the data-simplex rather than minimizing the Frobenius norm. The main steps of the algorithm is shown in Algorithm 1.

The archetypes found by SIVM reside in edges of the multidimensional space giving an attractive way of analyzing player behavior telemetry data. That is, rather than concentrating on central behavior classes, SIVM provides a compact way of describing the player behavior through extreme behavior. Another advantage of using SIVM for game telemetry analysis is the easiness of interpretation as the found basis vectors are actual players.

VI. Dataset

For the study presented here, a dataset containing player behavior telemetry from Tomb Raider: Underworld, was analyzed. The dataset is a sample drawn from the Square Enix metrics servers, covering all data collected for the game during a two month period (1st Dec 2008 - 31st Jan 2009). The dataset includes records from approximately 203,000 players, and includes 706 total features from each player. The game was launched in November 2008, so the data represent a time period where the game was recently released to the public. The behavioral features extracted from the dataset were originally chosen by Drachen et al. [11] and Mahlman et al. [16], who analyzed smaller portions of the dataset (1365 and 10,000 players respectively). The earliest of these studies focused on a smaller set of features and aggregated data across the entire game. The latter used the full range but was focused on classification, not clustering. The rules guiding feature selection were described by Drachen et al. [11] and fundamentally state that the first features to be selected in an exploratory behavioral analysis should be those relating to the primary game mechanics as these are the most descriptive of the way a game is played and how players can interact with the game system. TRU is a 3D platformer, with navigation being a major part of the gameplay, as is solving puzzles and fighting enemies (fighting was emphasized in the Underworld game to a higher degree than in previous iterations of the Tomb Raider game series). The features used for the current analysis all relate to the core mechanics of the game. This places some limits on the behavioral features that can be developed, and the same technique applied in other contexts and for different purposes may require a different rationale for feature selection to be applied, for example if the purpose is to evaluate purchasing behavior [1].

The behavioral features were described in detail in Mahlman et al. [16] and are therefore only briefly outlined below:

Player death: The total number of deaths for each player. There are 4.47 million instances of death registered, across all levels/MUs and death causes (μ = 71.04, varying from 0-939 death events; σ = 63.86). The death count is dependent on e.g. how much of TRU that a player has played, and the skill and playstyle of the player.

Help-on-Demand: The number of times help was requested from the Help-on-Demand (HOD) system integrated in TRU. The HOD provides help in the form of either hints or answers on how to handle the puzzles in the game. Overwhelmingly players request both hints and instruction and answers jointly if using the HOD system. These two values were therefore aggregated. A total of 926,734 HOD-requests are recorded (varying from 0-717, μ = 14.72, σ = 28.8).

Causes of death: The various causes of death in TRU can be grouped into the following four types (note that death events caused by game bugs, for example players dying during cinematic encounters, were not included): Enemies (melee). Deaths caused by melee enemies, including sharks and tigers, comprising 3.03% of the total number of death events. Enemies (ranged). Deaths caused by enemies using ranged weapons, e.g. mercenary snipers. Comprises 4.14% of the total number of death events. Environment: Deaths caused by environmental factors, e.g. fire or traps. Comprises 29.9% of the total number of death events. Falling: Deaths caused by the player falling. This cause of death comprises the 62.92% of all death events making it the dominating way to die in TRU - as would be expected from the game design.

These numbers vary from those reported in [11], and it is important to note that the samples are different: in the study of Drachen et al. [11], a sample of 1365 players was used who completed the game, whereas the current sample is comprised

of players who completed the first level in the game.

Adrenalin: The number of times the adrenalin feature was used. The adrenalin feature is an advanced gameplay feature, which when activated slows down relative execution time, allowing the player better time to perform special attacks etc. When activated, a cursor has to be moved to the head area of the target, which will trigger a headshot event. Adrenalin use requires a certain amount of skill (and interest in using the feature). This is reflected in that only 53.3% of the players use the feature. Activation of the adrenalin feature is recorded 291,370 times in the dataset, varying from 0-187 times (μ = 4.63, σ = 10.5).

Rewards: The number of rewards collected. TRU contains substantial numbers of ancient artifacts, shards and other forms or relics which players can collect, notably by exploring. A total of 1,120,708 artifacts/shards were located by the players in the game (μ =112.08, σ = 86.9).

Playing time: The time that each player spent playing the game. A total of 97.1 years of playtime were included in the dataset (including the game prologue) (μ = 13.52 hours).

Setting changes: In TRU, players can change different parameters of the game. These include four that affect the core game-play, notably in terms of difficulty. They are therefore of interest when evaluating player behavior. A total of 15,317 settings changes were made (max 104, μ = 1.53). Only 1740 of the players used this feature of the game, with an average frequency of μ = 8.8. Settings changes were vastly more common in the first two levels than the latter five, possibly reflecting the players adjusting the difficulty parameters of the game early on, until they are satisfied. The four features are as follows: Ammo adjustment: Adjusts how much ammunition Lara Croft is able to carry. Changing this setting comprises 29.6% of the total number of settings changes. Enemy hit points: Changes the amount of hit points that AI enemies have, either in a positive or negative direction. 31.5% of the setting changes are of this type. Player hit points: Adjusting how many hit point Lara Croft (the player character) has, effectively increasing or decreasing how much damage she can take before dying. Changing this setting comprises 19.5% of the total. Saving grab adjustment: The player can change the recovery time when performing jumps in the game, increasing the time available to gain a handhold. 19.4% of the settings changes are included here.

VII. Data preparation and analysis

The dataset was initially cleaned so that records with missing information were removed. Furthermore, for players who had played the game more than once, only the first play through was used. Finally, only players who completed level 1 were included, as the churn rate in the game means that only roughly 30% of the players who started the game make it through the (skippable) tutorial and the first level of the game. Of the players who complete level 1, only about 1:6 actually complete the game. For the current study, the goal was to evaluate player behavior as they progress through a game, and therefore it was decided to remove all player who churn out early in the game. This reduces the 203,000 player sample to 62,000 players. These roughly correspond to 4.2% of the total number of players who played the game, meaning those whose installation and starting of the game was logged by the Square Enix metrics servers (roughly 1.5 million).

A second typical problem in behavioral analysis in games is data type mixing and the existence of the outliers. Data type mixing requires the adoption of normalization strategies such as min-max and variance (or zero mean, ZMN) normalizations [26]. ZMN normalizes field values according to mean (μ) and the standard deviation (σ). The ZMN algorithm subtracts the values from μ and divides the result by σ. On the other hand min-max normalization transforms the data into a defined range using the minimum and maximum values of the field and the particular range [26]. While data type mixing would not appear to be an issue in the current study, ZMN and min-max normalization were both applied to the datasets to estimate the effect of the choice of normalization strategy, and the corresponding SIVM results found to be similar. However, it is important to note that the similar results are likely the result of the lack of outliers in the two datasets, i.e. the 1% peeling of the convex hull using the *Fastmap Algorithm* [27], that we also applied here to remove the outliers as described in [14]. As the results from both normalization techniques were similar, here we present the results from min-max normalization. Following data preprocessing (cleaning and outlier removal) and normalization evaluations, SIVM was applied to the seven sub-samples from TRU, corresponding to data aggregated across each of the seven levels in the game.

VIII. Results and discussion

In this section we describe how we used Archetypal Analysis to analyze the progression of the behavior of players from two perspectives. Firstly we analyze how the players that completed the game (16% of the analyzed players) spent time in levels. The aim here is to show how the players are grouped according their completion times per level. Secondly, incorporating the above mentioned features together with the total completion time, e.g. number of rewards and number of HOD requests, we aim to analyze how the playing behavior progressively changes across levels. For both of our applications we presented the results with 6 archetypes. The decision to use six archetypes rests on two considerations: 1) Representation error difference values indicate that for most of the levels in TRU, six archetypes forms a good balance between explanatory strength and representation value difference, i.e. using six archetypes to seven respectively does not decrease the representation error value substantially. 2) The exception is level 2 where error indicates that 5 archetypes might work better, however in the interest of visualization, it was decided to keep the number of clusters constant throughout the analysis.

A. Playtime Profiles

Initially, the play (level completion) time data were analyzed using SIVM. Using level completing times for players as features we identified six archetypes showing different patterns of completion time across the seven levels of TRU. Figure 1 shows the bar-plots of the completing time per level for the founds 6 archetypes. In terms of completion time, the majority of the players (93%) fall into the same archetype, which is characterized by having a total completion time across all seven levels of about 15-20 hours. This can be

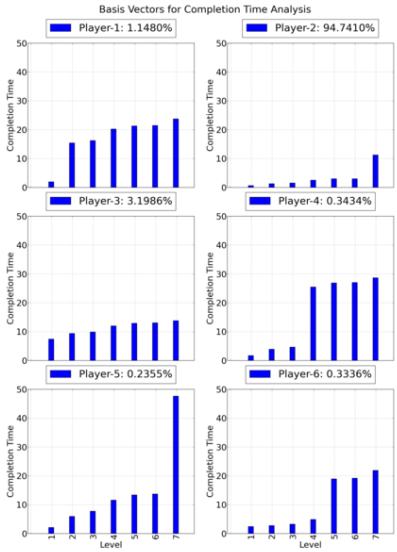

Fig. 1: Across-level completion time patterns for Tomb Raider: Underworld, based on six archetypes developed using SIVM.

visually observed by the 2 dimensional simplex projection of the belongingness values for each player, as shown in Figure 1, where the majority of the players are close to the 2nd basis vector. This aligns with the intended duration of TRU. About 7% of the players fall into one of five archetypes showing markedly higher completion time profiles, but with all but one showing a tendency towards longer level completion times the further into the game we are. This also aligns with the design of TRUs levels, which get progressively longer and more difficult throughout the game.

B. Behavioral Clusters

Based on the application of SIVM, histograms can be generated for each of the six archetypes across the seven levels of TRU. An example is shown in Figure 4, for level 4 of the game, including the hard-clustering results together with the normalized feature values. From the histograms, descriptive profiles can be built of each of the clusters. From the perspective of game design, profiles should be descriptive of the major behaviors of the clusters, backed by magnitudes along each feature, which allows for easier interpretation than numerical descriptives only. In the current case, rather than using descriptive but symbolic terminology, as in e.g. [14], cluster profile descriptions were based on the actual behavioral features, focusing on features that had a high respective value. For example, a cluster could be characterized by TIME-REWARDS if it has high respective values for these two features. This method for describing cluster profiles reduces the amount of information present in the analysis (in the example here from 13 to 2 behavioral features), but serves to draw attention to the highest-value features for each cluster. In the discussion presented here, low respective feature values (e.g. rapid completion rate) are also integrated, as low values are equally important to describe player behavior. Different

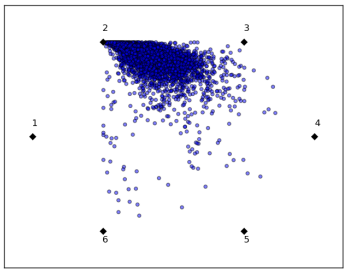

Fig. 2: Simplex projection of the mixing coefficients for players who completed the game, for playtime data only. The majority of the players fall into one archetype (group), which is characterized by exhibiting rapid completion times (see Player 2 in Figure 1).

strategies can be leveraged in describing multivariate clusters [26], the approach suggested here is just one of them.

The archetype profiles developed across each of the seven levels of TRU varies substantially, across all the measured aspects of player behavior. The variations notably relate to playtime, and the completion time for each level consistently plays a significant role in driving the separation of the archetypes.

Four behavioral profiles are more or less consistent throughout all levels, with some exceptions. These four profiles comprise the vast majority of the players in game, across all seven levels. However, the fraction of the total players for each level that is included in each of the four clusters varies to a degree where no multivariate statistical test is needed to verify this (e.g. the TIME-REWARDS cluster varies from 0.2-74.6%). The reason cannot be inferred from the dataset but can relate to: 1)The way the input data are normalized and the frequency ranges within each feature can inflate the weight of e.g. the absence or presence of a variable with a small range; 2) The interpretation of the clusters. For example, for level 4, the TIME-REWARDs cluster is very infrequent at 0.2% of the total players who completed that level, and it is possible to argue for similarity between other clusters from level 4 and the TIME-REWARD type profile. However, no other cluster for level 4 exhibits the dramatically high completion time values. 3) Finally, and importantly, the reason could relate to design restrictions in TRU imposing limits on play-style variance, and at the same time players varying their play-styles through the game in a response to the changes in the design of the different TRU levels (which are intended to vary to pose ongoing variation and new challenges to the player). That the distribution of the behavioral clusters in level 4 is different than any other level could support this hypothesis as this is an unusual level in terms of length, frequency of ranged enemies and number of death events [16]. The strength of the forcing from the design onto the behavior of the user is difficult to measure quantitatively, but forms an interesting subject for future research.

On an additional note, when Drachen et al. [11] analyzed a small subsample of the current dataset (1365 players), aggregating a smaller set of behaviors across the entire game, four

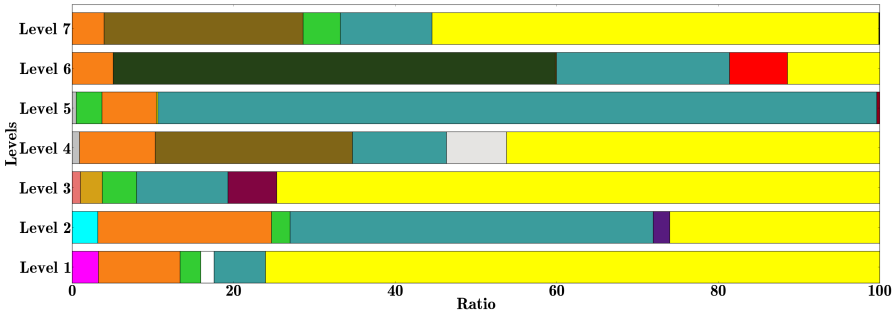

Fig. 3: Stacked bar chart showing the distribution of different behavior profiles across the seven levels of Tomb Raider: Underworld (2008, Crystal Dynamics/Square Enix). Each profile labeled according to the features that are most frequent within each profile (an alternative would be to focus on features with less frequency. For example, a low playtime score indicates players that relatively rapidly complete the level. Labels could also be assigned based on a mixture of features, for example based on key game mechanics. Cluster profiles and their corresponding web colors are as follows: Recurrent clusters: DEATH-REWARD-CAUSES OF DEATH (Pacifists): dark orange; ADRENALIN-REWARD (Veterans): light green; HOD-REWARD-ENVIRONMENT: dark turquoise; TIME-REWARDS (Solvers): yellow. Clusters occurring in a single or max. two levels, by level: (1) MELEE-ADJUSTMENTS: magenta, DEATH-FALL-SAVINGS GRAB ADJUSTMENT: white. (2) ENVIRONMENT-SAVINGS GRAB ADJUSTMENT: cyan; REWARDS-ENEMY HITPOINT ADJUSTMENT: purple. (3) ENVIRONMENT-ADJUSTMENTS: light coral; DEATH-MELEE: gold; DEATH-REWARDS-RANGED: maroon. (4) DEATH-FALL-ADJUSTMENTS: silver; ADJUSTMENTS: oak brown; DEATH-ENVIRONMENT-REWARDS: platinum. (5) DEATH-REWARD-FALL: burgundy. (6) SAVINGS GRAB ADJUSTMENT: dark forest green; TIME-DEATH-FALL: red. (7) DEATH-REWARD-ENVIRONMENT: black. Best viewed in color.

behavioral clusters were found which contained over 90% of the players. The four profiles were Runners, Pacifists, Solvers and Veterans. The latter three correspond to three of the consistent profiles found in the current dataset. This should not be taken as meaning that aggregate-level dimensionality reduction analysis is as informative as progression-based (evolutionary in the current terminology), however, as the current analysis shows a high degree of variance in the percentage distribution of these behavioral clusters across the levels of TRU. The four (more or less) consistent clusters are characterized as follows:

DEATH-REWARD-ENVIRONMENT (all levels except level 3): this cluster is characterized by having the highest death values of any cluster, across all six levels it is present in. It is also characterized by commonly having high death values for melee and ranged enemies, depending on the level. For example, in level 1 in TRU there are only few movable enemies, this cluster is characterized by death by environmental causes. The cluster averages 10.3% of the total players, but with a SD of 6.7 showing amble variation between the levels. The behavioral pattern is reminiscent of the Pacifist profile of Drachen et al. [11], which is also characterized by dying often, especially from enemies, average completion times but finding many rewards.

ADRENALIN-REWARD (all levels except level 4,6): This cluster is characterized by using the adrenalin feature of TRU, finding many rewards and in general dying rates in the low-middle range, completing the game fast and making few adjustments to the game. This cluster profile is reminiscent of the Veteran profile of Drachen et al. [11]. The profile on average comprises 3.4% of the players with a SD of 0.96.

TIME-REWARD (all levels): This is generally the most frequent cluster, comprising 41.4% of the players on average, but with a high variance (SD = 29.9, mainly driven by level 6 where this cluster only comprises 0.2% of the players). The players in this cluster overall do well, but are very slow to complete the game and collect a high number of rewards, suggesting explorative behavior and taking time to go through the puzzles in the game to get at even hidden treasures. This cluster bears resemblance to the Solver profile of Drachen et al. [11].

HOD-REWARD (all levels): This cluster is characterized simply by using the HOD system order of a magnitude more than any other cluster, presumably because the players do not care for or have problems with the puzzles in TRU, and collecting relatively many rewards. They feature fast completion times and in several levels tend to die often from a variety of causes. This profile is thus reminiscent of the Runner profile of Drachen et al. [11], with the difference that the authors found HOD values ranging widely for their Runner profile, while in the current case this cluster consistently has the highest HOD values of any cluster at any level. The cluster comprises 29.9% of the players on average, but with SD = 29.9 (high variance notably due to level 6 forming a clear outlier where this cluster comprises 89% of the total. Removing level 6 reduces average percentage to 17.8 and SD to 14.2)).

The remaining clusters in the seven levels are confined to a single or two levels, and usually comprise very small fractions of the overall number of players (from a fraction of a percent to a few percent, with one exception a 24.4% cluster [ADJUSTMENTS] on level 4 that is characterized by a high value for settings adjustments). The remaining features of this cluster show some similarity with the TIME-REWARDS cluster and could be interpreted as forming part of the same group. Figure 3 depicts the hard clustering results focusing on the distribution of clustered behavior groups across levels.

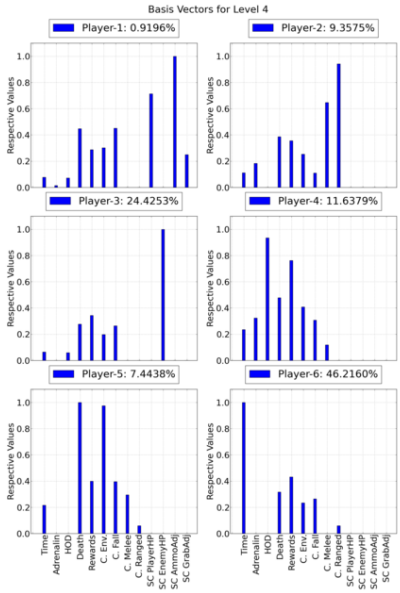

Fig. 4: Example of independently analyzing the behavioral clusters for Level 4. Each cluster illustrates a different type of behavior varying in one or many features.

C. Behavior Evolution Across Levels

In concert with describing semi-persistent or persistent behavioral profiles, each level can be viewed independently in order to evaluate behavior within that level. This approach is useful for evaluating level design, e.g. to examine if the clusters in for a given game level conform to the expected player behaviors. To take an example, a description for level 4 (see Figure 4) could read: Players here fall into three main clusters. 46% are characterized by very slow completion times but high relative reward scores. 24% are characterized by rapid completion times, fairly high death counts, mainly from ranged enemies, and is reminiscent of one of the smallest clusters in level 3 that exhibits the same profile (2.7%). 11.6% are characterized by a high use of the HOD system, and very high scores in terms of finding rewards, as well as death events being frequent, although mainly from environmental causes and falling. Approximately 9% are characterized by having similar rapid completion times as the five smallest clusters, but showing a weakness to melee and ranged enemies. While these players handle the navigational- and puzzle-based challenges of TRU well, they are challenged by the movable enemies in the game. Descriptions such as these should in a real-life context be more detailed to provide e.g. level designers with actionable feedback on the overall behaviors exhibited by players (for example during a beta test prior to launch).

The results described above indicate that a few consistent or almost consistent behavioral profiles form the majority of the players across the seven levels of TRU. In effect, the majority of the players follow one of four highly varied play-styles. The analysis indicates that the same players do not necessarily adopt the same play-style throughout the game, as the clusters vary substantially in relative size across the levels. In the background of this development is the steadily decreasing number of players, where at the latter levels of TRU only a fraction of the original 62,000 players are retained (16% completed the final level). A possible next step is to analyze the path of specific players across the different through the game (discussed further below).

D. Discussion

There are several conclusions that can be drawn from this analysis towards informing game design. For example, a consistent cluster of users rely heavily on the HOD system, averaging 29.9% of the players. This indicates that either these players had problems with the puzzles in TRU, or just did not care about them and used the HOD system to bypass these challenges. Irrespective, that almost a quarter of the player base potentially has a problem, or a feature they ignore, is worth investigating further, e.g. via drill-down analysis to obtain more detail on this behavioral pattern [2]. Another example is formed by the behavioral clusters present in some levels only, which provides information specific to those levels. For example, 3% of the players in level 2 die commonly from environmental causes and use the savings grab adjustment feature, which could indicate that they have trouble handling the jumps in that level. This could indicate a level design issue, and drill-down analysis will be able to inform if it is for example specific jumps in the level that cause this pattern to emerge. Other clusters indicate high rates of dying from mobile enemies, which could lead to player frustration. Examples like these indicate the usefulness of behavioral clustering across levels (or segments of a game). Clustering is a high-level early process analysis that can be used to obtain an idea about the general patterns of behavior (notably for a centroid seeking algorithm), and in the case of SIVM and similar convex-hull methods, the extreme behaviors specifically.

A game analyst, producer, designer or similar who in a practical context work with profile results such as those outlined above, should be able to drill down through these higher-order profile descriptions to investigate the distribution across all 13 features. Simple descriptions of behavioral clusters are useful to facilitate that users can orient themselves in the results [2], [11]. In the context of cluster analysis in potentially high-dimensionality situations, as is common in game analytics [1], [2], a human interpretation element is present in an analysis, notably to make decisions about how to describe behavioral clusters, i.e. to make them actionable to the target audience. Errors or bias may be introduced during this interpretation process, and it is therefore necessary in practice in data mining to allow users of such descriptions to backtrack (drill down) and view the data informing the descriptive profiles, in case any results appear suspicious. Current research is investigating how to improve the visualization of the results, using the D3.js visualization library for building web-based interactive visualization of the flow of players between clusters. The ulterior aim is to enable a general audience to interact with and explore results as well as associated information, e.g. general statistics for each TRU level in relation to the behavioral clustering results. A central limitation of the current analysis is that it did not consider tracking the migration of individual players or groups of players along or between clusters. While evaluating how player behavior breaks down across a level (or other section of a game) is useful in its own right, e.g. for evaluating difficulty or if the desired behaviors are actually manifesting [10], being able to follow the flow

of players between clusters allows for evaluation of the route players take in an out of a game. For example, evaluating how players move from novice to expert levels of competence, or the paths that lead a non-paying user to become a paying user [1]. This approach is notably interesting in the context of persistent games such as Massively Multi-player Online Games (MMOGs) and online games that rely on the Free-to-Play (F2P) or similar revenue models.

IX. Conclusion

The ability to condense high-dimensionality datasets of player behavior, across millions of players and extended durations of real-time or playtime, is important to the game industry as it informs about the overall ways in which users play specific games. An important requirement is that clusters need to be interpretable and actionable; and furthermore methods are needed to address a variety of needs, e.g. finding extreme behaviors vs. general behaviors [1], [5], [6], [10], [11], [14]–[16], [28]. In this paper, an approach is presented for defining and describing behavioral clusters which allows for examining the ongoing evolution in the behavior of groups of players throughout a game. The method was applied to a 62,000 player dataset from the AAA-level commercial game Tomb Raider: Underworld, and behavioral profiles for clusters of players described across the seven main levels of the game, using Simplex Volume Maximization to define archetypes of behavior [13], [24], [25], describable using game design language. SIVM has been applied to datasets from other games [14], and the progressive or evolutionary perspective applied here would appear to be broadly applicable to behavioral clustering in games in general. The research presented contributes to the field of game analytics by developing multiple consecutive clusters, across the playing duration of a game, rather than focusing on game-level aggregate data or slices of a game viewed in isolation. In principle, the fundamental approach suggested in this work can be applied to any type of player behavior analysis. Current research is investigating how to improve the visualization of the results, notably towards being able to present cluster result interactively, with the aim of giving a non-expert audience the ability to explore the results and associated information, e.g. general statistics for each TRU level in relation to the behavioral clustering results.

X. Acknowledgements

The authors extend their warmest thanks to Georgios Yannakakis, University of Malta and IT University of Copenhagen, Julian Togelius, IT University of Copenhagen, Tobias Mahlmann, IT University of Copenhagen, and Hector Perez, University of Malta, for help preprocessing and evaluating the Tomb Raider: Underworld dataset. The work presented here would not be possible without the collaboration of these colleagues. The authors also extend their warmest gratitude to Square Enix, Crystal Dynamics and IO Interactive for sharing telemetry data, feedback and advice on developing actionable behavioral profiles - we would like to direct a special thanks to Janus Rau Soerensen, Tim Ward and Jim Blackhurst.

References

[1] T. Fields and B. Cotton, *Social Game Design: Monetization Methods and Mechanics*. Morgan Kaufmann, 2011.

[2] M. Seif El-Nasr, A. Drachen, and A. Canossa, *Game Analytics: Maximizing the Value of Player Data*. Springer Publishers, 2013.

[3] I. Ahmed, A. Mahapatra, M. S. Poole, S. J., and B. C., "Identifying PLayer Typology Based on Longitudinal Game Data," in *Proc. Social-Computing*, 2012.

[4] G. N. Yannakakis and J. Hallam, "Real-time game adaptation for optimizing player satisfaction," *IEEE Trans. on Computational Intelligence and AI in Games*, vol. 1, no. 2, pp. 121–133, 2009.

[5] B. Weber, M. John, M. Mateas, and A. Jhala, "Modeling player retention in madden nfl 11," in *Proc. Innovative Applications of Artificial Intelligence*, 2011.

[6] G. Yannakakis, "Game AI Revisited," in *Proc. ACM Conf. on Computing Frontiers*, 2012.

[7] B. Jansen, *Understanding User-Web Interactions via Web Analytics*. Morgan & Claypool Publishers, 2009.

[8] J. Bohannon, "Game-miners Grapple with Massive Data," *Science*, vol. 330, no. 6000, pp. 30–31, 2010.

[9] 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.

[10] 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.

[11] A. Drachen, A. Canossa, and G. Yannakakis, "Player Modeling using Self-Organization in Tomb Raider: Underworld," in *Proc. IEEE CIG*, 2009.

[12] B. G. Weber and M. Mateas, "A data mining approach to strategy prediction," in *Proc. IEEE CIG*, 2009.

[13] C. Thurau, K. Kersting, and C. Bauckhage, "Yes We Can: Simplex Volume Maximization for Descriptive Web-Scale Matrix Factorization," in *Proc. ACM CIKM*, 2010.

[14] A. Drachen, R. Sifa, C. Bauckhage, and C. Thurau, "Guns, swords and data: Clustering of player behavior in computer games in the wild," in *Proc. IEEE CIG*, 2012.

[15] G. Zoeller, "Development Telemetry in Video Games Projects," in *Game Developers Conference*, 2010.

[16] T. Mahlmann, A. Drachen, J. Togelius, A. Canossa, and G. N. Yannakakis, "Predicting Player Behavior in Tomb Raider: Underworld," in *Proc. IEEE CIG*, 2010.

[17] C. Thompson, "Halo 3: How Microsoft Labs Invented a New Science of Play," *Wired Magazine*, vol. 15, no. 9, 2007.

[18] Y.-T. Lee, K.-T. Chen, Y.-M. Cheng, and C.-L. Lei, "World of Warcraft Avatar History Dataset," in *Proc. ACM MMS*, 2011.

[19] L. Mellon, "Applying Metrics Driven Development to MMO Costs and Risks," *Versant Corporation*, 2009.

[20] R. Bartle, *Designing Virtual Worlds*. New Riders, 2004.

[21] B. Harrison and D. L. Roberts, "Using Sequential Observations to Model and Predict Player Behavior," in *Proc. ACM FDG*, 2011.

[22] C. Thurau and C. Bauckhage, "Analyzing the Evolution of Social Groups in World of Warcraft," in *Proc. IEEE CIG*, 2010.

[23] N. Ducheneaut and R. J. Moore, "The Social Side of Gaming: a Study of Interaction Patterns in a Massively Multiplayer Online Game," in *Proc. ACM Conf. on Computer Supported Cooperative Work*, 2004.

[24] C. Thurau, K. Kersting, and C. Bauckhage, "Convex Non-negative Matrix Factorization in the Wild," in *Proc. IEEE ICDM*, 2009.

[25] A. Cutler and L. Breiman, "Archetypal Analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[26] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan kaufmann, 2006.

[27] G. Ostrouchov and N. F. Samatova, "On FastMap and the Convex Hull of Multivariate Data: Toward Fast and Robust Dimension Reduction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1340–1343, 2005.

[28] D. King and S. Chen, "Metrics for Social Games." Presentation at the Social Games Summit, 2009.