Proceedings, The Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-15)

Large-Scale Cross-Game Player Behavior Analysis on Steam

Rafet Sifa
Fraunhofer IAIS
Sankt Augustin, Germany
rafet.sifa@iais.fraunhofer.de

Anders Drachen
Aalborg University
Aalborg, Denmark
drachen@hum.aau.dk

Christian Bauckhage
Fraunhofer IAIS
Sankt Augustin, Germany
christian.bauckhage@iais.fraunhofer.de

Abstract

Behavioral game analytics has predominantly been confined to work on single games, which means that the cross-game applicability of current knowledge remains largely unknown. Here four experiments are presented focusing on the relationship between game ownership, time invested in playing games, and the players themselves, across more than 3000 games distributed by the Steam platform and over 6 million players, covering a total playtime of over 5 billion hours. Experiments are targeted at uncovering high-level patterns in the behavior of players focusing on playtime, using frequent itemset mining on game ownership, cluster analysis to develop playtime-dependent player profiles, correlation between user game rankings and, review scores, playtime and game ownership, as well as cluster analysis on Steam games. Within the context of playtime, the analyses presented provide unique insights into the behavior of game players as they occur across games, for example in how players distribute their time across games.

Introduction and Contribution

Game companies today are able to collect behavioral telemetry data from entire populations of players, and using cloud based storage technologies, it is possible to collect and process every single user event from games. Furthermore, with the help of global game platforms; such as Steam, Good Old Games, or console-based services, as well as social networking platforms like Facebook or Tango, increasingly larger and broader audiences can be reached. However, despite a remarkable growth of interest, fueled by new business models (notably Free-to-Play, F2P) and mobile technologies, publicly available behavioral analytics in digital games has as yet been predominantly confined to single games. Unlike other sectors such as e-commerce, there have been no large-scale cross-game behavioral studies, in part due to the recent, if highly accelerated, introduction of analytics practices in the game industry, but perhaps more importantly due to the confidentiality associated with behavioral telemetry data. This means that while dozens of telemetry-based studies and hundreds of observational studies of behavior in games have been published or presented, there is very little knowledge available in the public domain about how these translate across games (exceptions including (Chambers and Saha 2005; Bauckhage et al. 2012)). This places challenges in the way of establishing behavioral patterns that operate across some or all games, also for applied purposes such as informing game design, for example via improving retention and engagement (Bauckhage et al. 2012; Pittman and GauthierDickey 2010; Feng and Saha 2007). It also limits the ability to develop techniques used in e.g. e-commerce and behavioral economics for understanding and modeling user behavior (Resnick and Varian 1997; Ricci et al. 2011; Bogers 2009). The importance of cross-games behavioral analysis is emphasized when considering the increasing number of available platforms that offer games, and that the same players tend to own multiple games. Understanding how games are played is not a trivial task considering that multiple gameplay profiles can be observed from individual players.

In this paper we present four experiments performed on a 6 million player dataset, covering a total playtime of over 5 billion hours of play across more than 3000 games distributed via the Steam platform. Additional data was collected covering game rankings and review scores, as well as information on the genre, type and key game mechanics. The results provide insights into the patterns around playtime in the games bought and played by Steam users, as well as patterns about the users themselves. Playtime is the focus of the experiments conducted because this feature is an indication of player interest or engagement with a game. In a highly competitive global marketplace for games, understanding the connections between the games played by a user, not just within any one game, is vital e.g. for tasks such as cross-game promotions, migrating players between games (Sifa, Ojeda, and Bauckhage 2015) or game recommender systems (Sifa, Bauckhage, and Drachen 2014a). Summarizing the results: 1) **Playtime distribution - players**: Cluster analysis shows that the majority of players are more or less dedicated to one or a few games. Only about a third of the players put similar amounts of time into a variety of games (given a $k = 11$ solution). Playtime distribution is highly skewed. The number of owned games is distributed in a similar way. The average number of owned games is 22.1 (standard deviation = 35.5). 2) **Playtime distribution - games**: Cluster results run on aggregate playtime data from the Steam games rather than the players reveal the existence of four specific archetypes of games, which are differentiated by having different retention profiles. 3) **Game ownership**: Frequent Itemset Mining (FIM) and Association Rule Mining (ARM) (Agrawal and Srikant 1994; Han, Pei, and Yin 2000; Gow et al. 2012) show that there are distinct patterns game ownership. Some of these combinations of game ownerships (itemsets) are very frequent and the rules have high confidence values. 4) **Ranks and reviews vs. ownership and playtime**: Analysis of player-generated rankings and aggregate review scores show no strong correlations with game ownership or playtime, questioning the notion that review scores correlate with sales.

Related Work

Due to space constraints this section will focus on key related work in game analytics. For an extended overview of behavioral analytics for games, see for example (Seif El-Nasr, Drachen, and Canossa 2013; Bauckhage et al. 2012). Cross-games analytics is a rare occurrence, in part due to the recent rapid emergence of the practice in the industry, the confidentiality associated with behavioral data and the lack of public datasets. Exceptions exist, such as the aforementioned (Bauckhage et al. 2012) and (Chambers and Saha 2005; Feng and Saha 2007; Drachen et al. 2012; Sifa et al. 2013). Some industry white papers, notably from analytics companies, contain high-level descriptive measures, but methods are not specified and the underlying data are kept confidential. Recently a few studies have been presented which take advantage of data that can be accessed via player stats tracking services or distribution platforms as in the current case. The alternative approach has been to mine the server-client connection stream in online games (e.g. (Pittman and GauthierDickey 2010)). Similarly, (Chambers and Saha 2005) used data from the GameSpy service to model player session frequency in a First-Person Shooter (FPS) game, noting that games popularity follows a power law distribution. (Bauckhage et al. 2012) observed the same pattern across five game titles, and examined a range of random process models. The authors also presented an explanation for why these models provide good fits on various aspects of player behavior (playtime, session frequency, session length, inter-session time). (Lim 2012) reported from "several dozen freemium games", that player behavior is better approximated by a power law than a normal distribution. The author highlighted that doubling the player base does not necessarily double revenue, indicating the importance of differentiating between users when considering acquisition strategies, a topic also covered by (Seufert 2014; Fields and Cotton 2011; Seif El-Nasr, Drachen, and Canossa 2013). Related to Steam, the work of (Orland 2014) was conducted in parallel with the research presented here. Orland (Orland 2014) mined a smaller (1/24th in size) sample of 250,000 Steam player profiles, providing descriptive statistics only, e.g. on which games that are played the most. (Orland 2014) reported on examples where the sampled data were extrapolated to the full range of approximately 172 million Steam accounts, showing good correlation between sales data estimated from the sample and spot tests against actual sales data as reported by game development companies. Other relevant studies include a number of publications in network analysis, where network balancing for online games form a topic of interest. Two relevant examples are (Pittman and GauthierDickey 2010) who investigated player distribution in World of Warcraft and Warhammer Online. The authors fit session length data to a Weibull distribution, similar to (Chambers and Saha 2005). Feng et al. (Feng and Saha 2007), working with Eve Online reported that the distribution of the number of sessions that a person plays before quitting fit a Weibull distribution. This means that most players do not stay long in the game, as denoted by the long-tail distribution.

Game	Total Playtime (hours)
DOTA 2	887,701,351
Team Fortress 2	638,489,137
Counter-Strike	505,944,559
Counter-Strike: Source	482,431,858
Garry's Mod	159,561,947
Call of Duty: Modern Warfare 2 - Multiplayer	146,445,499
Left 4 Dead 2	114,134,730
Counter-Strike: Global Offensive	103,571,160
The Elder Scrolls V: Skyrim	94,895,353
Call of Duty: Modern Warfare 3 - Multiplayer	63,203,811
Sid Meier's Civilization V	60,567,442
Terraria	59,127,631
Call of Duty: Black Ops - Multiplayer	49,774,569
Borderlands 2	46,378,311
Left 4 Dead	41,985,976
Counter-Strike: Condition Zero	40,483,935
Killing Floor	34,528,886
Call of Duty: Black Ops II - Multiplayer	31,780,194
Day of Defeat: Source	29,760,902
Battlefield: Bad Company 2	27,128,032
Fallout: New Vegas	26,223,025
Mount & Blade: Warband	22,555,784
Warframe	22,288,693
Portal 2	20,291,456
Borderlands	19,333,583

Table 1: The 25 most played games on Steam, and total amount of playtime spent in the dataset.

Data and Pre-processing

Steam is the largest online game distribution platform for PC games, with around 75 million active users and roughly 172 million accounts in total, with 3-7 million concurrent users according to Valve[1]. A distinctive feature of Steam is that it is cross-platform, supporting multiple gaming environments, including the current operating systems and the upcoming Steammachines[2], Valve's new consoles. The dataset was harvested from public Steam profiles using the web API provided by Valve, and contains records from over 3200 games and applications, but after running through the pre-processing steps detailed below, the dataset was constrained to 3,007 full games and 6,049,520 Steam players, covering 5,068,434,399 hours of game-play. The players are selected form the most populous 3500 communities and their IDs are anonymized by random hashing.

The data was harvested in 2014 and contains the total

[1]http://www.joystiq.com/2014/01/15/steam-has-75-million-active-usersvalve-announces-at-dev-days/

[2]http://store.steampowered.com/livingroom/SteamMachines

playtime of the players until the time of the retrieval. This means that for some players, the dataset may not cover their full player histories (i.e. still active players), and may bias results towards showing shorter playtimes than they actually are. It is also important to note that the tracking in the Steam platform started after March 2009, which eliminates the playtime of the players before this time. A series of preprocessing steps have been performed: any demos have been removed, Software Development Kits (SDKs) and games that are not played by at least 25 people. 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. These games were eliminated. The dataset only covers playtime on the Steam platform, not time spent playing the same games outside of that platform. Tbl. 1 shows the most played 25 games.

Playtime Distribution - Players

In order to investigate how much the players invest their time on particular games, we have conducted cluster analysis based on the players' relative spent time on the 3007 Steam games. We use k-means here, as this is a very well established approach, makes it possible to benchmark against other analysis, and it builds on previous work in games, e.g. (Drachen et al. 2012; Sifa et al. 2013). Clustering provides a compact way of representing and summarizing the key features and elements in large datasets. Our aim in this section to observe how the players are grouped according their playtime behavior. The main goal of clustering can be casted as factorizing the given data matrix into lower rank matrices that gives us the flexibility to explicitly define constraints. Specifically, given a matrix $D_{m \times n}$, clustering aims to factorize this matrix into two matrices $P_{n \times k}$ and $C_{k \times m}$ to minimize the Frobenius norm:

$$E = ||D - PC||^2. \quad (1)$$

When clustering players the matrix P contains prototypical players representing the behavior of the cluster and matrix C contains the belongingness coefficients. Every clustering method imposes different constraints to the factor matrices P and C. For a more thorough discussion of clustering player behavior telemetry data please refer to (Drachen et al. 2012; 2013; Bauckhage, Drachen, and Sifa 2015). Fig. 1 shows a heat-map generated based on the prototypical players found running k-means clustering algorithm with 11 basis vectors.

The results show how people spend their time across the different games on the Steam platform, and emphasizes the skewed playtime vs. games distribution mentioned above. The clustering results are based on normalized player vectors indicating to what game or games the players invest their time. Having run k-means with 5-fold cross validation over 5 different combinations of equally chunked test data sets for $k \in [2, 3, ..., 15]$, $k = 11$ yielded the highest separability based on Silhouettes whereas $k = 5$ as the best in terms of the gap-statistic. Considering the solution with $k = 5$, four clusters (comparable to number 2, 5, 6 and 10 in the list below) occur in the resulting basis matrix, which each contain players dedicated to one game. The rest

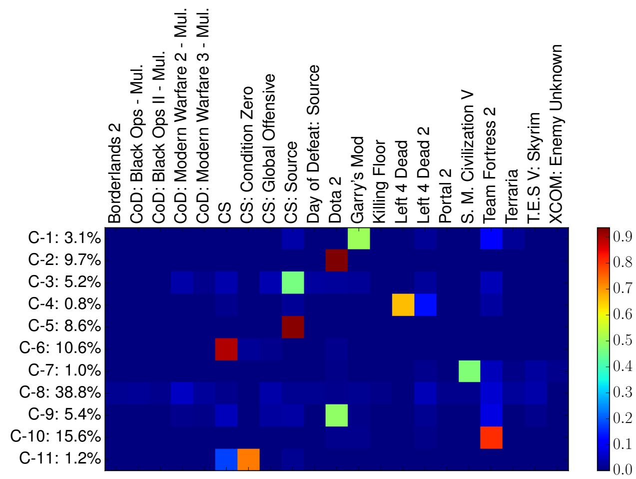

Figure 1: Clustering of 6 million Steam players based on their time spent playing 3007 games. Each cluster (rows in the heatmap matrix) represents a prototypical player profile. The types of game play varies from single game users that spent most of their time playing a single game, as C-{2,5,6,10}, to those primarily focused on one to a few games, C-{1,3,4,7,8,9,11} to C-8, which contains players that distribute their playtime across a variety of games. Best seen in color.

of the clusters in the below list are packed into one cluster comprised of nearly half the players in the dataset. The immediate implication of this result is that almost half the Steam players are focused on one of four major Valve titles (DOTA 2, Team Fortress 2 [TF 2], Counter-Strike [CS] and CS: Source), and the rest distribute their playtime across multiple games. The result is visualized in Fig. 1.

However, as k increases above 5, clusters split off which contain players focused primarily on one game (although with a minor component in a few other games), until k=11, where further splits provide clusters that are hard to meaningfully separate from each other, leaving about a third of the players in this cluster. As a solution, k = 11 is more interpretable and yields a better separation in term of Silhouettes (see Fig. 2), it is therefore included here. The resulting 11 cluster solution (Fig. 1) shows that for 10 clusters, players primarily play one game, each one of the most played games on Steam (e.g. TF 2, DOTA 2, CS-versions, Garry's Mod and the Left 4 Dead series). The final cluster (roughly 38% of the players) shows a comparatively more varied interest among the players. For k = 11 the characteristics of the clusters are as follows:

1. **Customizers' Cluster, C-1**: Representing 3.1% of the dataset, contains players that played Valve's flagship customization game TF 2 and Garry's Mod for most of their playtime.
2. **DOTA 2 Cluster, C-2**: Representing 9.7% of the players, this group of players is the typical *DOTA-only* players, that they only played DOTA 2.
3. **FPS Cluster, C-3**: Representing 5.2% of the players,

Figure 2: Values of separation coefficient for different numbers of clusters that calculated based on Silhouettes.

this group contains players that mostly played the famous First Person Shooter (FPS) games that include CS: Source, CoD and TF 2.

4. **Left 4 Dead 2 Cluster, C-4**: A minority cluster (with 0.8% belongingness) of FPS players with a heavy emphasis on Left 4 Dead 1 and 2.
5. **CS: Source Cluster, C-5**: Representing 8.6% of the players, this group heavily contains players playing the CS Source version.
6. **Counter-Strike Original Cluster, C-6**: Representing 10.6% of the data this cluster is formed by the players that play the original CS game, released in the year 2000, most of the time (more than 89%).
7. **Civilization V Cluster, C-7**: Representing 1% of the player this group's player mostly prefer Sid Meier's Civilization V and also spend small amount of their time on Steam's other flagship games such as DOTA 2 and Left 4 Dead.
8. **Active Steam Players, C-8**: The most populated cluster (38.8%), that play variety of games across different genres nearly equal amount of time. Unlike the other clusters, players are not dedicated to a single game but rather distributed their time to many different games that include for example all the games in the CS-series, TF 2, DOTA 2, Borderlands 2, Portal 2, Left 4 Dead 1 and 2, and CoD.
9. **Balanced DOTA 2, C-9**, 5.4% , players forming this cluster play mostly DOTA 2. Unlike the DOTA 2 Cluster, player's here are more inclined to play other games including TF 2, The Elder Scrolls V: Skyrim, CS-series and Left 4 Dead, CoD-series etc.
10. **Team Fortress 2 Cluster, C-10**: Represents 15.6% of players that almost only played the free to play shooter game TF 2 (they spent nearly 80% of their time playing this particular game and the rest to Valve's flagship games such as DOTA 2 and Garry's Mod).
11. **Counter-Strike Original Cluster, C-11**: Representing 1.2% of the data, this group is a Counter Strike cluster that contains players that mostly play CS Condition Zero (73%) followed by the original 2000 version of the game (17%).

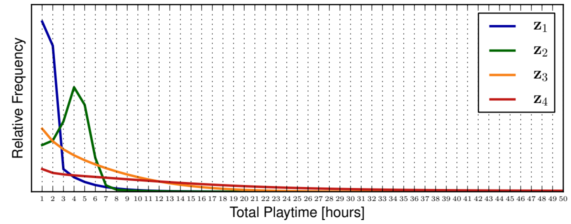

Figure 3: Results of cluster analysis of 3007 playtime frequency distributions, revising those of (Sifa, Bauckhage, and Drachen 2014b). The profile shown by $\mathbf{z}_1$ represents games with short playtime. $\mathbf{z}_2$, peaks at 4 hours and shows a quick drop afterwards, $\mathbf{z}_3$ and $\mathbf{z}_4$ represent slower decaying games, but $\mathbf{z}_4$ is dominated by *AAA* games. Note that the four profiles all indicate that the global interest in a game is limited to 30-35 hours.

Playtime Distribution - Games

Similar to clustering players based on their playtime behavior, games can be clustered in the same way. In this we follow in the the method outlined by (Sifa, Bauckhage, and Drachen 2014b) who used Archetypal Analysis (Cutler and Breiman 1994) on aggregate playtime curves from Steam players to identify archetypes of games. (Sifa, Bauckhage, and Drachen 2014b) describe four clusters of games, each exhibiting a different prototypical playtime profile, noting that the aggregate playtime patterns follow a Weibull distribution. Here the analysis is rerun with the more heavily pre-processed dataset used here (please refer to (Cutler and Breiman 1994) for a detailed breakdown of how Archetype Analysis operates), and additional information was collected on the genre, type and key features of the games in the sample, using Steam's own denominator system. This in order to investigate if there are any high-level patterns in the distribution of the games across the four archetypes i.e. if particular types or genres of games are typical of specific playtime patterns. This analysis shows which games that have good retention profiles, either in the short or long term. This is important knowledge for e.g. game design and benchmarking analysis, and there is little knowledge publicly available on this topic. The analysis also reveals specific patterns such as the similar playtime profiles of games in the same series, and the difference in playtime across recent indie and major commercial titles. Fig. 3 shows the calculated prototypical playtime distributions for the dataset, labelled $\mathbf{z}_1$-$\mathbf{z}_4$.

For three of the archetypes (profiles), a declining pattern is observed, while one of the profiles exhibit a peak at 4 hours of playtime (see Fig. 3), following which a sharp decline happens, with a likelihood of players still playing at 15 hours being lower than for any of the other profiles at 0.0006. This profile, $\mathbf{z}_2$ represents 10.3% of the Steam games. The games in the profile comprise a mixture of genres but are predominantly smaller commercial titles (e.g. Octodad: Dadliest Catch), with many F2P titles included, with a few older AAA-level titles such as EverQuest II and TitanQuest. There are no immediate clues as to why these games peak at 4 hours of play, but it is clear that they on average manage to get a large fraction of the players to at least stay engaged for a few hours. $\mathbf{z}_4$, which comprise 23.7% of the Steam games, exhibit the slowest decay rate in playtime, with a flat distribution of aggregate playtimes. The profile contains primarily major commercial titles, which exhibit the slowest rate of playtime decay. A variety of genres are included, from FPS/shooters (e.g. F.E.A.R., TF 2), RPGs (Darksiders), strategy games (e.g. Empire: Total War), survival games (e.g. DayZ), MMOGs (e.g. Guild Wars) to sports games (e.g. Football Manager) and smaller but highly popular titles such as Dungeon Defenders and Tropico. The majority of the games in the dataset that are played the most, including DOTA 2, TF 2, CS: Source, the CoD-series, Left 4 Dead 1 and 2, Portal 1 and 2, etc. are found in this cluster. A number of these are developed by Valve, the company who owns and manages the Steam platform (Orland, 2014). Games in the same series tend to have all games from that series placed in the same cluster, indicating that games within the same series exhibit similar playtime profiles (this is the case for all four clusters). Also, it is generally the case in the Steam data that smaller commercial (indie) titles are played fewer hours than major commercial titles, with exceptions including Terraria and Garrys Mod. The first profile $\mathbf{z}_1$, which represents 43.8 % of the dataset, we observe a fast decrease in playtime: by 3 hours, less than 10% remain. The games in this cluster are older major or minor titles, in some cases re-releases (e.g. Earthworm Jim, Hexen II). Many are shooters/action titles, with a few RPG titles such as the Avernum series, and puzzle game such as Bejewelled Deluxe and Crazy Machines. The vast majority of the players who try these games play them for a very short time. The third archetype $\mathbf{z}_3$ represents 22.2% of the games and exhibits the 2nd-most slow decline profile. A higher fraction of the players stop playing these games are a few hours as compared to $\mathbf{z}_4$, and the curve for $\mathbf{z}_3$ crosses the curve for $\mathbf{z}_4$ at 13 hours of playtime. Different genres are included, with a predominance of adventure games (e.g. Dreamfall: The Longest Journey), action games (e.g. Super Meat Boy) and point & click games (e.g. the Nancy Drew series). Some older AAA-titles are included, e.g. the Far Cry-series, Max Payne 1 and 2, and Doom 3.

Game Ownership Patterns

In order to figure out what games to market to players, it is crucial to know which games they have already played, and their relative engagement with these games. Understanding which games that are played together and by who is thus important for running effective marketing campaigns and combating the growth in User Acquisition Costs (UAC) (Hadiji et al. 2014; Runge et al. 2014; Rothenbuehler et al. 2015). Platforms such as Steam provide a tool for investigating ownership patterns (limited to games delivered via Steam). We used Frequent Itemset- and Association Rule Mining (FIM and ARM respectively) to observe most frequently played set of games and their associations (Agrawal and Srikant 1994; Gow et al. 2012). Originally, FIM and the follow-up ARM have been introduced in the beginning of 1990s as business intelligence techniques to have implicit recommendations to the customers of the products by introducing discounts or positioning the most commonly sold items together. Since then it has been widely used in the data-mining community to generate first insights in massive datasets (Han, Pei, and Yin 2000). Given a finite number of items $I = \{i_1, i_2, ..., i_m\}$ and a set of transactions $T = \{t_1, t_2, ..., t_n | t_j \subset I\}$, the main aim of FIM is to find the single items and frequently occurred items with frequency less than given minimum *support* value $s \in N$. The main aim of ARM is to find interesting associations between the frequent itemsets by finding the probability, called *confidence*, of an occurrence of an itemset given another disjoint set. Namely, having found the frequent itemsets and given a minimum probability threshold, ARM finds association rules that have confidence values over a specified threshold. Finding itemsets and association rules is a challenging problem due to the combinatorial complexity of the required settings. Casting the problem as a classical search problem, the goal becomes finding the appropriate combination of items or itemsets that satisfy the acceptance condition.

Game(s)	Ratio
TF 2	60.06%
DOTA 2	40.4%
CS: Source	35.05%
Left 4 Dead 2	34.4%
DOTA 2 and TF 2	28.33%
TF 2 and Left 4 Dead 2	27.96%
CS	24.08%
CS: Global Offensive	23.72%
TF 2 and CS: Source	22.1%
Garry's Mod	22.11%
Garry's Mod and TF 2	20.24%
DOTA 2 and Left 4 Dead 2	19.91%
Portal	19.42%
Portal 2	19.03%
Alien Swarm	18.11%
CS: Global Offensive and TF 2	18.02%
DOTA 2, Left 4 Dead 2 and TF 2	17.15%
TF 2 and Portal	16.96%
Terraria	16.94%
Portal 2 and TF 2	16.8%

Table 2: Top 20 frequent itemset mining results (based on ratio scores) for the mostly owned/downloaded games on Steam

While there are many methods to discover the itemsets and association rules, the analysis of the algorithms is beyond the scope of this paper and we refer to (Borgelt 2012). Having extracted the played games for each player, we used the FP-Growth algorithm (Han, Pei, and Yin 2000) to extract the frequent item sets. The FP-Growth algorithm (Han, Pei, and Yin 2000) is an efficient algorithm that represents the transactional database in a prefix tree and discovers frequent itemsets in a depth-first search manner. After obtaining the itemsets we extracted the association rules as in (Agrawal and Srikant 1994). Based on gameplay histories, Tbl. 2 and Tbl. 3 show the results with the highest support. Relating ARM and FIM findings with those of total playtime, a range of patterns become evident, of which a few are discussed here. The results in Tbl. 2 indicate that the games that are played the most consists of nearly only Valve's flagship games. It is important to note that some of the games occurring together, such as DOTA 2 & TF 2 or TF 2 & CS: Source are in aggregate played more than other games. Among the

Game(s)	Confidence
CoD: Modern Warfare 2 → CoD: Modern Warfare 2 - Multiplayer	0.96
CS: Condition Zero → CS	0.94
Garry's Mod and DOTA 2 → TF 2	0.94
Garry's Mod and Left 4 Dead 2 → TF 2	0.94
Terraria and Left 4 Dead 2 → TF 2	0.93
Spiral Knights → TF 2	0.92
DOTA 2 and Terraria → TF 2	0.92
Left 4 Dead 2 and Half-Life 2 → TF 2	0.92
DOTA 2 and Portal 2 → TF 2	0.92
Left 4 Dead 2 and Portal → TF 2	0.92
Garry's Mod and CS: Source → TF 2	0.92
Alien Swarm and Left 4 Dead 2 → TF 2	0.92
Portal 2 and Left 4 Dead 2 → TF 2	0.91
Garry's Mod → TF 2	0.91
CoD: Modern Warfare 2 - Multiplayer → CoD: Modern Warfare 2	0.90

Table 3: Top 15 association rules, note TF 2's prevalence.

top 25 rules the F2P TF 2 occurs very frequently, however, the player-based clustering analysis shows that not so many of the TF 2 players are actually devoting their time on this game. A total of 223 association rules with a support over 50% were found. The majority involve Valve's TF 2 (90% of the rules with 85% or better confidence) and some other shooter game/games. This highlights that the vast majority of the players in the sample has TF 2 installed, more than DOTA 2, but the latter is played more. About 28.33% of the people in the sample played both DOTA 2 and TF 2. Furthermore, the different variants of CoD and CS also exhibit association rules. CS: Source is played by 35.05% of the players, with CS: Global Offensive reaching 23.72%. If specifically looking at associations not involving TF 2; Portal 2 and Left 4 Dead 2 provide a confidence of 73%. Similarly, there are multiple associations involving Left 4 Dead 2, Alien Swarm and DOTA 2. At confidence levels of around 60 - 70%, there are numerous rules involving games in different genres, e.g. The Elder Scrolls V: Skyrim and Left 4 Dead 2 (67%), supporting the results of C-8. The indie game Terraria is associated with all of Valve's own shooter games. Other results include e.g. that 90% of the players of CoD: Modern Warfare 2 played both the multi-player and single-player version, but only 4% played single player only. 95% of the people who played CS: Condition Zero also played the original CS. Garry's Mod is involved in multiple association rules, with 92% of the players also playing TF 2, and 94% of them also play DOTA 2 and/or Left 4 Dead 2, indicating a strong synergy between these titles. In summary, over 3000 games are included but only about a dozen are involved in the association rules with confidence above 50%. Despite the offering on Steam, a small number of games are not only the most popular but also the most highly associated.

Ranks/Reviews vs. Ownership/Playtime

The question of the relationship between reviews and game sales forms an ongoing debate, notably because there is a widespread practice in the game industry of assigning bonuses and payment to development companies depending on how well a game does in terms of review scores. The work presented here extends previous attempts to correlate reviews with sales (e.g. (Orland 2014)) by including player-generated rankings, and including the actual time spent playing. We harvested review scores from MetaCritic.com for 1426 games, and player ranking scores from SteamGauge.com for 1213 games. Running the Pearson correlation analysis against the scores from these two sites and playtime and game ownership individually revealed no strong correlations. For game ownership there is a statistically significant correlation at $r = 0.22$ for MetaCritic and $r = 0.25$ for SteamGauge, but neither of these explain a lot of the variance in the data (low r-squared values). The correlation between the total playtime of the games weighted by the total number of players and the two sets of scores. For SteamGauge we observe values $r = 0.22$, for MetaCritic $r = 0.06$ indicating a lack of strong relationships. Reviews may serve a purpose beyond scores, and games intended for short play duration may add noise, but the results nonetheless emphasize that sales and playtime, has little to no correlation with aggregate review scores or player rankings.

Conclusion and Discussion

Here analyses have been presented focusing on playtime-related, cross-games behavior of users of the digital game distribution platform Steam, covering a sample of 6 million players and over 3000 games. Results reveal high-frequency itsemset and rules with high confidence for groups of games that are bought/owned together. Cluster analysis of players show that the majority are more or less dedicated to one game, although roughly a third distribute playtime equally among multiple games. Games are also shown to fall into four playtime-based clusters showing some relationship with genre/types. Most of the games we analyzed are generally played for a few hours or less. Some have persistent followings, and about a dozen dominate in terms of players and playtime. Additionally, results indicate that there is no-minimal correlation between review scores/ranks and playtime/game ownership. The methods used are established statistical or machine learning algorithms and can be applied in other cross-application situations, e.g. to profile players for the purpose of migrating them between games via targeted advertisement. In addition to playtime, specific types of games share features (e.g. game mechanics) which could be used to improve profiles and fed into predictive models. Future work will focus on even more detailed analyses of player behavior, focusing on time-series analysis with an overall goal of mapping temporal patterns. Adding to the discussion in the beginning of this paper, it is worth noting that the experiments are also of interest for game recommender systems as each target one of four main dimensions in these: people and products (players/games); implicit and explicit feedback (playtime/rankings). The potential uses for recommender systems in games is obvious, given the tens of thousands of games published each year and the associated discoverability problems, and indeed Steam already features a recommendation function (principles are unpublished). Recommender systems can however also be used to help developers identify e.g. high-value users and inform on how to migrate them between games toward mitigating the cost of user acquisition, which forms another venue for future work in cross-games and large-scale analytics.

References

Agrawal, R., and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. In *Proc. Int. Conf. Very Large Data Bases, VLDB*, volume 1215.

Bauckhage, C.; Kersting, K.; Sifa, R.; Thurau, C.; Drachen, A.; and Canossa, A. 2012. How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing Times. In *Proc. IEEE CIG*.

Bauckhage, C.; Drachen, A.; and Sifa, R. 2015. Clustering Game Behavior Data. *IEEE Transactions on Computational Intelligence and AI in Games*.

Bogers, T. 2009. *Recommender Systems for Social Bookmarking*. Ph.D. Dissertation, Tilburg University.

Borgelt, C. 2012. Frequent Item Set Mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(6):437–456.

Chambers, C., F. W. S. S., and Saha, D. 2005. Measurement-based Characterization of a Collection of On-line Games. In *Proc. of ACM SIGCOMM Conf. on Internet Measurement*.

Cutler, A., and Breiman, L. 1994. Archetypal Analysis. *Technometrics* 36(4).

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

Drachen, A.; Thurau, C.; Sifa, R.; and Bauckhage, C. 2013. A Comparison of Methods for Player Clustering via Behavioral Telemetry. In *Proc. SASDG FDG*.

Feng, J., B. D., and Saha, D. 2007. A Long-term Study of a Popular MMORPG. In *Proc. ACM SIGCOMM WNSSG*.

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

Gow, J.; S., C.; P., C.; and Miller, P. 2012. Mining Rules from Player Experience and Activity Data. In *Proceedings of the 8th AAAI AIIDE conference*, 12–33.

Hadiji, F.; Sifa, R.; Drachen, A.; Thurau, C.; Kersting, K.; and Bauckhage, C. 2014. Predicting Player Churn in the Wild. In *Proceedings of IEEE CIG*.

Han, J.; Pei, J.; and Yin, Y. 2000. Mining Frequent Patterns Without Candidate Generation. In *ACM SIGMOD Record*, volume 29. ACM.

Lim, N. 2012. Freemium Games Are Not Normal. In *Gamasutra*.

Orland, K. 2014. Introducing Steam Gauge: Ars reveals Steams most popular games. In *Ars Technica*.

Pittman, D., and GauthierDickey, C. 2010. Characterizing Virtual Populations in Massively Multiplayer Oline Role-playing Games. In *Proc. of Int. Conf. on Advances in Multimedia Modeling*, 87–97.

Resnick, P., and Varian, H. R. 1997. Recommender systems. *Communications of the ACM* 40(3):56–58.

Ricci, F.; Rokach, L.; Shapira, B.; and Kantor, P. 2011. *Recommender Systems Handbook*. Springer.

Rothenbuehler, P.; Runge, J.; Garcin, F.; and Faltings, B. 2015. Hidden markov models for churn prediction. In *Proceedings of IEEE IntelliSys*.

Runge, J.; Gao, P.; Garcin, F.; and Faltings, B. 2014. Churn Prediction for High-value Players in Casual Social Games. In *Proceedings of IEEE CIG*.

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

Seufert, E. 2014. *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue*. Morgan Kaufmann.

Sifa, R.; Bauckhage, C.; and Drachen, A. 2014a. Archetypal Game Recommender Systems. In *Proceedings of the Learning, Knowledge, Adaption Conference*.

Sifa, R.; Bauckhage, C.; and Drachen, A. 2014b. The Playtime Principle: Large-scale Cross-games Interest Modeling. In *Proc. IEEE CIG*.

Sifa, R.; Drachen, A.; Bauckhage, C.; Thurau, C.; and Canossa, A. 2013. Behavior Evolution in Tomb Raider Underworld. In *Proc. IEEE CIG*.

Sifa, R.; Ojeda, C.; and Bauckhage, C. 2015. User Churn Migration Analysis with DEDICOM. In *Proc. ACM RecSys*.