André Rattinger, BSc

Performance and Engagement Analysis in Multi-User Networks

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingeneur

submitted to
Graz University of Technology

Supervisor
Dip.-Ing. Dr.techn. Univ.-Doz. Christian Gütl

Institute for Interactive Systems and Data Science

Co-Supervisor
Dip.-Ing. Johanna Pirker, BSc

Graz, January 2017

André Rattinger, BSc

Analyse des Erfolges und der Bindung zu Online-Systemen anhand von Netzwerken

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingeneur

eingereicht an der
Technischen Universität Graz

Betreuer
Dip.-Ing. Dr.techn. Univ.-Doz. Christian Gütl

Institute for Interactive Systems and Data Science

Zweitbetreuerin
Dip.-Ing. Johanna Pirker, BSc

Graz, Januar 2017

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present masters thesis.

Graz, ____________________ ____________________

Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Graz, am ____________________ ____________________

Datum Unterschrift

Abstract

It is common nowadays for online systems to feature some kind of social component. In most of those services, users can form implicit connections, even if there are no features that support building a traditional social network like defined friendships, or different types of relationships that indicate direct connections between the users.

In this thesis a framework is proposed to extract implicit social connections from distinct online systems, apply metrics to measure factors like success and engagement and take a look at user behavior. For this purpose two datasets are selected that provide social exchange possibilities and also have clear definitions of what qualifies as success and good performance. The examined datasets are *Destiny*, a hybrid online shooter with a massive number of players developed by Bungie, and *Galileo*, which is a collection of massive open online courses from the University of Galileo.

The datasets are explored to identify the metrics and the possible implicit relationships that are then used to create social networks. The generated social networks are the basis for further analysis. In addition to graph metrics, behavioral information is explored, and the soft clustering method archetypal analysis is employed to infer details on how users behave in the systems. Social influences are examined to show if those have an effect on users in the context of the systems, whether this involves reaching a high score or getting better grades.

Network properties such as weight and degree are found to be good indicators for performance in both of the examined datasets. Users who display more interactive behavior or have stronger ties are also more engaged and exhibit more frequent use of the provided tools in the system. This also applies to communities within the datasets: members of those outperform average users, and are more engaged. The results are influenced by the kind of implicit connection captured and media use like voice-chat, e-mail communication and other ways users exchange information are not included in this approach.

Kurzfassung

Systeme im Internet haben heutzutage oft eine soziale Komponente. In vielen von diesen Systemen können Benutzer implizite Verbindungen mit anderen formen, auch wenn keine direkte Form von zu schlißenden Freundschaften existiert, oder auch keine direkten Indikatoren einer Beziehung vorhanden sind.

In dieser Arbeit wird eine Vorgehensweise präsentiert, die implizite soziale Verbindunden von verschiedenen Online Systemen extrahiert, um den Erfolg und die Bindung zu dem System zu messen. Dies wird durch die Untersuchung des Benutzerverhaltens unterstützt. Zwei Datensätze, die ihren Benutzern die Möglichkeit bieten sich sozial auszutauschen, wurden für diesen Zweck ausgewählt. Zusätzlich müssen die Datensätze klare Definitionen davon beinhalten, was in diesen Systemen erfolgreiches Verhalten ausmacht. Die zwei untersuchten Datensätze sind: *Destiny*, ein hybrider Online-Shooter, der von Bungie entwickelt wurde und *Galileo*, eine Sammlung von Massive Open Online Courses von der Galileo Universität.

Die Datensätze werden untersucht, um Metriken und impliziten Beziehungen zu finden, die benutzt werden, um soziale Netzwerke zu erzeugen. Die erzeugten Netzwerke dienen als Basis für die weitere Analyse. Zusätzlich zu den benutzten Graphenmetriken werden Verhaltensweisen untersucht und die Soft-Clustering Methode Archetypal Analysis wird verwendet, um weitere Details über das Benutzerverhalten zu finden. Soziale Einflüsse auf den Erfolg und auf das Verhalten werden untersucht, und es wird gezeigt, ob diese Auswirkungen auf die Benutzer haben, sei es bei dem Erzielen eines neuen Rekordes oder einer besseren Note.

Netzwerkeigenschaften wie Gewichte oder die Anzahl der Verbindungen sind gute Indikatoren für den Erfolg in beiden untersuchten Datensätzen. Benutzer die stärkere Beziehungen oder interaktivere Verhaltensweisen haben, sind auch stärker an das System gebunden. Dies wird auch durch die vermehrte Benutzung von den bereitgestellten Funktionen des Systems gezeigt. Einzelne Untergruppen in dem Netzwerk sind auch erfolgreicher und aktiver in den Systemen. Die Resultate werden von anderen Medien die nicht festgehalten sind beeinflusst. Kommunikation oder soziale Interaktion die außerhalb der Datensätze stattfindet wird nicht berücksichtigt.

Acknowledgments

After an extraordinary year of research and writing, I can finally add this page to the thesis and thank everyone that helped me along the way.

Firstly, I would like to express my sincere gratitude to my supervisor Christian Gütl, for his useful comments, remarks and engagement through the learning process of this master thesis.

I would also like to express my deep gratitude to Johanna Pirker for the guidance she provided throughout this thesis. The door to her office was always open when I had one the many questions about my research or writing, and the thesis wouldn't be the same without her. She is definitely responsible for part of my motivation to proceed with research and academia.

In addition I would like to thank Anders Drachen, Rafet Sifa and Günter Wallner, for providing their expertise and insights when working together, and giving me the opportunity to write my first scientific publication. Furthermore I would like to thank Bungie for making detailed behavioral telemetry from Destiny available.

I want to gratefully thank the Erasmus Plus project MOOC Maker, in particular Rocael Hernández and his team from Galileo University in Guatemala, for providing the MOOC data sets.

Finally, I would like to thank all of my friends, family and especially my parents who encouraged and supported me throughout the last years and always stayed patient. Without them, none of this would have been possible.

Graz, January 2017 André Rattinger

Contents

Chapter 1

Introduction

Interaction in online systems between users can exist in many different forms and are often encouraged by the designers of the medium. Explicit relationships between people are only a small part of what defines the actual day-to-day behavior of users and how they are in contact with each other. The traditional way to build social networks is to use the explicit relationships the users define themselves in the system. To only use explicit connections to draw meaningful conclusions about the underlying social structure has been questioned in the past (Wilson et al., 2009), and might not always be the optimal way to arrive at a good interpretation of the underlying connections. A good example for the different kinds of relationships is as follows: when users add someone as a friend in a social network, they might not interact with each other after the initial action. A more meaningful way to detect those friendships would be to take messages they exchange or events they attend together as an indicator of their relationship, with a much more expressive model as an outcome (van de Bovenkamp et al., 2014). Other online systems might also not provide explicit relationships, and implicit connections are the only available information source to model the underlying structure. Compared to the systems analyzed in this work, traditional social networks usually do not have any conditions that indicate the performance of a person in the system: There is no definition of a goal that makes someone a better performer in a normal social network. This is required for the analysis conducted in this thesis, and both of the examined systems provide features that enable comparison between the users.

The main tool used in this thesis is social network analysis. Social network analysis is a widely used technique for analyzing user behavior in online systems, and uses network and graph theory techniques to build the networks and investigate the social structure in them (Otte & Rousseau, 2002). Visualization of those networks can enhance the perception of the underlying system, and lead to more insightful observations. To be applicable for this approach, the observed systems only have to fulfill two prerequisites and do not have to have much in common otherwise: 1) Comparable performance and engagement metrics must exist for every user. 2) Users can interact with each other in an implicit form. In addition to the work presented here, two publications were created and released past year. Some of the material has been incorporated (Rattinger et al., 2016; Pirker et al., 2017). In this work factors are explored that influence performance and engagement, especially those related to the relations that emerge from the structure of the social networks.

1.1 Objectives and Motivation

The main objective of this work is to examine social networks of online systems and to investigate them to find influences that represent the impacts on user performance and engagement. The following questions are the focus of this thesis:

1. Do user relationships, interactions, community or group membership relate to better performance?
2. Do user relationships, interactions, community or group membership relate to better engagement?
3. Can behavioral traits be identified that improve performance and engagement?

The implementation of the objective is done by building networks out of datasets that feature implicit relationships and leverage the networks for further analysis of the users performance and engagement. Doing this entails the exploration and analysis of the dataset to identify features that are relevant for mapping the underlying structure. Employing common social network measures is another objective that is used to showcase how the networks are organized and if the building step was successful. Furthermore all of the results generated by the previous step are explored in the context of user behavior. The analysis of two datasets conducted should provide a basis and framework for similar approaches with other types of data, and demonstrate that building of implicit social networks is a viable approach to analysis of performance and engagement metrics related to the users social behavior.

1.2 Methodology and Structure

This thesis is split into three main parts. The first part consists of the theoretical background involved in the work (Chapter 2). The second part deals with the data exploration and processing needed to run further experiments (Chapter 3 & 4). The third part showcases the results and takes a look at further network aspects (Chapter 5). Fig. 1.1 shows an overview of the presented work. The theoretical background and the dataset domain knowledge form the basis for the analysis process. This is followed by an exploration and cleanup step, which provides the essential information needed to create the networks. After the networks are build, each of them is examined with typical network metrics to show the viability of the different network approaches. This supports the final step of the analysis, which evaluates and compares the networks and extracted dataset metrics.

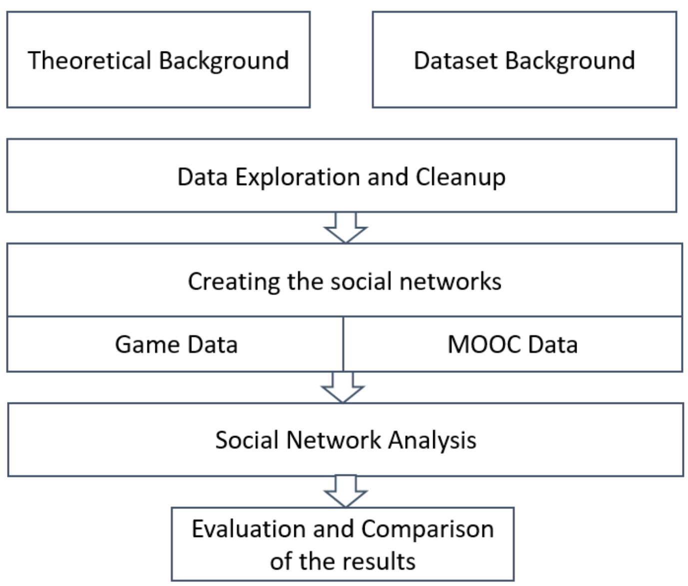


Figure 1.1: Structure of the presented work: Information about the theoretical background and the datasets build the basis for the further stepwise implementation and analysis.

Chapter 2 takes a look at the theoretical background of the social network and behavioral analysis techniques used and discusses related work that has been done in the area.

Chapter 3 showcases and explores the datasets used in further experiments. Detailed information is provided on how the datasets are structured, and what types of data can be found in them. In addition, it also explains what kind of pre-processing and cleanup techniques were used, and why those techniques differ for each of the datasets. The metrics that define the implicit relationships are defined and features used in later analysis are extracted. The chapter also identifies certain peculiarities each of the datasets has and analyses what they entail for later analysis.

Chapter 4 explores how social networks can be built from arbitrary datasets, and shows the methods used for defining the relationships in those networks. The approach of calculating a different weight for each network type is evaluated. Additionally several networks for each of the datasets are showcased. This is done to compare those alternatives and choose the most viable option for further analysis. The networks discussed in this section are all created from the arbitrary datasets outlined in Chapter Three, and therefore display major differences in some of their properties and how connections and weights are defined in them. Common network metrics are used to draw conclusion about the networks and compare the different versions and datasets with each other. Moreover small exerts of the networks are visualized that serve as a good approximation of the actual larger networks.

Chapter 5 discusses the data analysis performed on the datasets and the networks and also shows the experiments run on different metrics that define success in the underlying domains. In addition it displays the differences between the social networks created in the thesis and compares their success and engagement factors. The metrics defined in this chapter are used to make the networks more comparable to each other. Besides the performance and engagement analysis, archetypal analysis is used for further behavioral analysis. Furthermore different methods for visualization of user archetypes are explored.

Chapter 6 outlines the lessons learned from the exploration of the datasets and the usage of the applied methods. Chapter 7 provides suggestions for further work that could be conducted on the datasets and on the networks. Chapter 8 summarizes the results and gives a final conclusion on the work done.

Chapter 2

Background and Related Work

The work presented here, focuses on key aspects in two different domains: 1) Social Network Analysis (SNA) and 2) Behavioral Profiling (BP). This chapter outlines the background of those two domains and highlights work that is related to the further analysis.

2.1 Background

This section provides some background knowledge about the most important aspects discussed in this work. All of the addressed domains are broad areas with many different subtopics. For this reason the section will focus on the key aspects important for the areas.

2.1.1 Social Network Analysis

With emergent new technologies, SNA has become an important tool and research field. SNA has been employed on many of the big datasets that emerge from large online platforms such as *Facebook* and *Twitter*, as well on less obvious platforms and datasets (Wilson et al., 2009; Huberman et al., 2008). As a strategy for investigating social structure, SNA emerged from sociology and has been used in a many other fields (Otte & Rousseau, 2002).

Social Networks

SNA can be used to formalize the relationships build in systems between users, and therefore can be used to analyze the interactions taking place and is often used to examine the properties and dynamics of social networks (Scott, 2012). Social structures and networks build the core of SNA and are defined by the individuals who make them up. The individuals are usually called "nodes" or "vertices" in the context of social networks. The nodes in a social network can represent other entities like organizations, groups of people or similar items. The entities have to be connected with each other in order to form a network. Those connections are usually called "edges" or "links". The edges between the nodes are used to symbolize friendships, interests or other properties they have in common. Compared to other approaches, SNA is a non-individualistic technique that considers the actions of a combination of actors instead of only a single actor. The relationships between those actors, whether it is a trait they share or an activity they perform together, is the main concern (Otte & Rousseau, 2002).

Graph Theory

Graph theory is the foundation of SNA and provides many of the common techniques to analyze social networks, as well as a way to formally represent them. A graph G consists of a set of nodes N, and a set of edges E. The set E contains ordered pairs (i, j), where i and j are the connection from the nodes i and j. The connections in the graph can have properties that indicate the flow of information, and are divided into undirected, directed and mixed or multi graphs. Undirected graphs are used for relationships that are symmetrical and information flows both ways. If a person writes an article together, they are both co-authors, and one person can't be a co-author without the other. In an undirected graph, the information only flows one way. This can indicate information flow, but is also useful for other applications like mapping of processes. Directed connections can go both ways, and are called symmetrical when they do. Mixed graphs can contain a mixture of directed and undirected connections. Multigraphs are networks that can contain multiple connections between nodes. Compared to multigraphs, symmetrical directed graphs are limited to two connections. Graphs and especially social networks for that matter can grow quite big and there are approaches to draw and visualize graphs to gain insights about the structure (Di Battista et al., 1994). Different parameters in graph visualization and other algorithms create different arrangements of the nodes and edges, which have different insights attached to them. An important aspect in graph theory to gain better insights about the graphs build are community detection and other algorithms. Communities are sets of nodes that build a densely connected substructure of a graph, and can be independent inner structures that expresses different properties (Fortunato, 2010).

Network Properties and Measures

Network Measures are an essential tool to gain insight about the inner workings of a social network. This section presents a short overview of the network measures and properties used in this thesis.

Average Degree / Degree: The average degree of a graph measures how many edges the graph contains compared to the number of nodes. Because each edge of a graph is connected to two nodes, the expression is as following:

$$k\_avg = \frac{2 * E}{V}$$

Largest connected component (LCC): The largest connected component of a graph is the largest part of a graph, where all nodes have a connection within. A component is any structure that has at least one path between all of its nodes, but a node without any edges is also a component.

Average Clustering Coefficient: The clustering coefficient is a measure of the degree of cluster-forming between nodes. This is done by calculating the average of all local clustering coefficients in the network. The local clustering coefficient quantifies how complete the substructures are to being complete graphs.

Network Diameter: The network diameter is the shortest path that can be build between the nodes in a graph that are the furthest from each other.

There are a lot more approaches to understanding underlying graph structures, but are not discussed here. Many of them are based on simple properties such as distances between nodes, number of connections or how the ratio of nodes and connections is.

2.1.2 Behavioral Profiling & Archetypal Analysis

In online systems, especially if they are competitive, many strategies and styles to navigate those environments emerge. Grouping the users by their behavioral patterns is a solution that lets designers of those systems and the users themselves get a better understanding and an overview about the underlying structures.

Behavioral Profiling

As online systems contain vast amounts of telemetry data of users actions, extracting the important features that relate to behavior is a common source of data. Behavioral profiling is a technique that is employed in growing virtual environments as many different circumstance need to be controlled, such as guaranteed stable economics and prevention of fraudulent behavior. Therefore it has been adopted especially in game environments, but a lot of it is applicable to other areas, as it can be carried out in many different ways (Drachen et al., 2012). Behavioral datasets from such systems are big multi-dimensional datasets, as they have to capture all of the user interaction at all times. This leads to the complex task of filtering the information to identify and extract the relevant pieces. An approach that helps with the vast amount of multi-dimensional data is clustering, as it is an unsupervised method, that helps to identify certain important traits automatically (Bauckhage et al., 2015).

Archetypal Analysis

This section presents Archetypal Analysis as a technique for behavioral profiling. Archetypal analysis is a technique that seeks extremal points in multidimensional data (Cutler & Breiman, 1994). Each of the individuals found are represented as their probability to belong to each archetype, and archetypes are a a representation of minimizing the error that a user belongs to an archetype. When given a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ factors need to be found to minimize the matrix norms:

$$\left\| \mathbf{X} - \mathbf{XBA} \right\|^2 = \left\| \mathbf{X} - \mathbf{ZA} \right\|^2$$

$\mathbf{Z} \in \mathbb{R}^{m \times k}$ represents the archetypes and $\mathbf{A} \in \mathbb{R}^{k \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times k}$ represent the coefficients respectively (Rattinger et al., 2016). As with other clustering methods, the number of archetypes have to be specified beforehand, and output selection is up to the user.

2.2 Related Work

In this section, different approaches to data analysis grounded in SNA and BA are discussed. The work discussed here is the basis for the further analysis and the comparison of the two examined datasets.

2.2.1 Game Data Analytics

Games analyses through telemetry data has been plentiful in the last couple of years. Compared to other mediums, online games are oftentimes designed with the specific goal to be engaging. To reach this peculiar goal designers employ interactive game elements with a combination of narrative structures (Dickey, 2005). Compared to users of serious games, those groups are motivated through the entertainment games provide, and the term “gamification” is often used for this purpose (Deterding et al., 2011). Game analytics help to gain better insights on why players act in certain ways and on how their social structures are build. Due to limited amount of data on hybrid shooter games like *Destiny* little directly comparable work exists on the combination of performance and engagement factors in a social network structure. The following sections highlight some of the approaches separately.

Social Network Approaches

Social network analysis has become a commonly applied tool in many areas outside of online social networks. Due to that is also has been applied to many big Multiplayer Online Games (MOG), such as *Starcraft* or *Dota* that feature massive amounts of users. These studies focus on the factors on how the socal networks can be build, and while friendship relationships might be an indicator on how players interact, and how this might not be the best indicator for “real” relationships (van de Bovenkamp et al., 2014; Jia et al., 2015). Jia et al. (2015) also demonstrate how does relationships can be build out of online games, and showcases strategies on how to reveal them. When interacting with a online system, users form implicit relationships, that are different from traditional friendship relationships that are common in online social networks. Building networks from implicit relationships can help with discovering new ways to look at the underlying social structure. Similar to van de Bovenkamp et al. (2014), those relationships are analyzed and mapped to a social graph, but the network from a massive open online course is also looked at and the same methods are applied to form a implicit network of course participants. The strength of the relationships in those networks can be defined in many ways and by exploiting many parameters, and they will vary with the type of network build. To map those correctly usually some kind of domain knowledge is needed, that properly matches the strength of those to the actual revealed structure in the background (Xiang et al., 2010).

A lot of the previous work in SNA and Games focused on Games played in social networks or Massively Multiplayer Online Role-Playing Game which are mostly designed to have a big social component (Shin & Shin, 2011; Wohn et al., 2010; Simon & Apt, 2015). The social component can have a big influence in player retention and their general motivation to keep playing the game: Exchanging ideas, forming relationship, defeating challenges together are all influencing factors on how players experience the game world and keep experiencing it (Shin & Shin, 2011; Yee, 2007).

Behavioral Approaches

Behavioral traits of games are highly dependent on the possibilities game designers provide to the users, and the examination of those influences are a specific area of game analytics (Drachen et al., 2013; Bauckhage et al., 2015; Drachen et al., 2012). A manifold of approaches of behavioral analytics in games have been explored, which range from purchasing decisions to optimizing many experience factors (El-Nasr et al., 2013; Drachen et al., 2012; Sifa, Hadiji et al., 2015; Ducheneaut et al., 2006). Clustering methods as unsupervised learning techniques are commonly explored and fit well into research about how players interact in different environments, and demonstrate what kind of behavior they exhibit compared to fellow players. The main purpose of clustering is to identify certain traits that represent different styles of behavior (Bauckhage et al., 2015; Sifa, Drachen & Bauckhage, 2015). Those traits are closely linked to the players social experience, and approaches to detect patterns in the behavior are widespread: Ducheneaut et al. (2006) explored social dynamics by using longitudinal data of the successful MMORPG World of Warcraft and analyzing their grouping and guild structures. Thurau & Bauckhage (2010) also analyzed data on World of Warcraft, but examined the evolution of guild structures over time. Another approach to behavioral analysis through clustering is applying AA to the game datasets which has been done for game data before, but not for the unique environment that *Destiny* provides (Sifa et al., 2014; Lim & Harrell, 2015). Exploring this technique provides a basis for further approaches in classifying user data (Sifa & Bauckhage, 2013).

2.2.2 Massive Open Online Courses

There are a multitude of approaches to analyze MOOC participant data according to win further insights or improve on them. It has been shown that performance in MOOCs is linked to active engagement, with more engaged student outperforming others (Phan et al., 2016). Examining performance and engagement factors often is done in terms of the completion rates of the participants. Engagement poses a particular problem in MOOCs, caused by the mentioned high dropout rates. Proposals exist to battle those dropout rates by providing the users with time management and social interaction tools (Nawrot & Doucet, 2014). It is also reasoned that some course participants might have differing goals to a course completion, and might only engage in the parts that interest them. To include such participants in a potentially successful group Kizilcec et al. (2013) identifies four engagement trajectories, which are used to compare between the engagement and success between courses. The following sections discuss some of the approaches in regard to the analysis in this work. It especially takes a look at the SNA and BA approaches that emerged from the problems MOOCs face.

Social Network Approaches

Multiple formal approaches to online learning forms like E-Learning have been explored, and methods that examine the social structure as well as the interaction have been defined in the past (Haythornthwaite, 2005; Goggins et al., 2010; Breslow et al., 2013). Many of the forms have been questioned in terms of their effectiveness and other forms of distance and e-learning have been proposed (Kop & Hill, 2008). In addition, SNA is not only a tried method in general, but has also found application in studying various social factors that influence the outcomes for the students. Compared to other online systems, MOOCs oftentimes have an absence of direct connections between users, that would provide the basis for building a traditional social network based on explicit relationships (Haythornthwaite, 2005).

A main objective of the SNA approaches is usually to identify users that have a high risk not to successfully pass the course. When employing SNA, especially users who are very active are easily identifiable. The approach covers finding students influence in the system and importance as initiators of discussions and their posting patters (Sinha, 2014). Course discussions are the most prominent way social networks are build, because provide one of the few instances where actual social exchange is happening, but the approaches on how to use this information are various: this reaches from methods who assign the users to multiple subcommunities via partitioning, Survival analysis or even identifying detailed information about the comment structures and substructures within those discussion forums (Rosé et al., 2014; Sinha, 2014; Anderson et al., 2014). Generally it can be said that leveraging discussion forum behavior is a popular strategy to find opportunities in supporting the participants of a MOOC.

Behavioral Approaches

Compared to games, MOOCs exhibit other behavioral traits, especially when it comes to dropout or churn rates. SNA is used here as a tool to improve attrition and to generate models of the network through different supervised and unsupervised methods, and build sub-communities that way (Rosé et al., 2014). One of the main issues found with MOOCs concerning the high attrition rates, have been addressed many times, and different models have been proposed to formalize the problem and to divide the participants into three classes: persistence learners, healthy attrition and unhealthy attrition (Clow, 2013). Further metrics, that are not directly based on interaction are also a approach outside of SNA that seems to provide promising results.

Vitiello et al. (2016) discusses how dropouts can be recognized early on through the interaction patterns of the users. Those interaction patterns stem from the access behavior of different tools on the MOOC website, and users are classified according to their tool usage. This results into early recognition of users from the unhealthy attrition category. An important aspect in this is how users are identfied and what features are used to accomplish this task.

While many studies have analyzed the emergent social structure of games or online course systems, most focus on the peculiarities of a particular system, and do not apply the won insights outside of their domain. In addition only a few behavioral analysis aspects are used in combination with SNA. This work tries to improve on those aspects, while applying insights won out of the analysis of one dataset to the other and drawing parallels between their properties for further comparison.

2.3 Summary

Each online system that stems from a different domain has its own definition of performance. MOOC Performance is usually linked to passing a course with good grades and at the same time is linked heavily to engagement. The definition of good performance in games is usually straightforward because they naturally provide different scoring mechanics. Investigating social structures is an important tool and research field, and has been employed on many large datasets. The examination process that is used for this is social network analysis, and it is a common approach to create insights on how individuals interact. There are many ways to identify relations between users that do not always correspond to classical friendships. Those types of connections are called implicit relationships. The structure of social networks formed between users have been explored in the domains of the two datasets before: there are multiple examples on how players of a game can form a network, as well as MOOC participants. Compared to previous work in those fields, this thesis combines the social network approach with aspects of behavioral analysis. Due to a great amount of telemetry data available in online system, it is possible to identify many behavioral traits of players. Many of the approaches to analyze traits of players are linked to behavioral profiles. Extracting these profiles is often done with clustering techniques. Compared to games, MOOC research often focuses on the behavior of participants in terms of dropout, and how to improve passing rates. This thesis applies SNA insights with the analysis of behavioral traits.

Chapter 3

Datasets and Preprocessing

Data exploration and preprocessing are important first steps when starting an analysis. It is primarily used to get a look at the characteristics of the data and get a better understanding of the underlying structure. Data exploration is an especially important step that can help when selecting the right tools for further processing and recognizing patterns in the data early on. When looking for patterns, one of the first steps that can be taken is to look at how the data is distributed and where potential anomalies might be found. A quick and helpful way to detect anomalies is to use visualizations that show the statistical properties of the data. Preprocessing is used to prepare the data for further analysis by filling missing data with values, removing unwanted anomalies and performing various transformations on the data to create a consistent dataset for later use. Not using preprocessing techniques on the data can lead to misleading results, caused by possible irrelevant, missing or noisy information in the dataset.

This chapter discusses the used datasets and their preprocessing in separate sections. This is done because the underlying structure for each of the datasets is fundamentally different starting from file formats to general organization of the data.

3.1 Destiny

Destiny[1] is a first person hybrid online shooter game, that includes elements of other game genres like adventure games and massively multiplayer online RPGs (MMORPG). The player can perform many actions typical of those genres, like creating their own characters, examine different environments or equipping their characters with different weapon load-outs. Compared to other MMORPGs,

[1] https://www.destinythegame.com/

there is an absence of some features that traditionally define those genres, but it exhibits other features like a leveling system typical of RPGs. This is why it is described as a shared-world shooter by Bungie.

In *Destiny* the player has to defend Earth from various Alien races that threaten to destroy it. The role of the player is that of a guardian, a group of soldiers that defend earths last cities, by using a power called “Light“. Gaining levels and therefore improving a character is done by gaining experience points by playing the main quests of the game (*Destiny Overview*, 2016). The game lets the player choose between three classes (Titan, Warlock, Hunter) which all have their own unique abilities and upgrades. Each of the classes also have three subclasses, giving a player more room to customize their characters to their playstyle.

Destiny was developed by Bungie and published and released worldwide by Activision on September 9, 2014. After the initial launch, Bungie released four expansion packs which added new content, missions and story elements. The third expansion called the *The Taken King*, which was released in September 2015 and changed a lot of the core gameplay elements, like switching to a model of limited time events.

The dataset presented here does not include any games that were played after the release of the fourth expansion in September 2016.

Figure 3.1: *Destiny* gameplay example. (c) Bungie, Inc, Destiny, the Destiny logo, Bungie and the Bungie logo are registered trademarks of Bungie, Inc. All rights reserved. Image by courtesy of Bungie Inc.

3.1.1 Gameplay

Destinys gameplay generally can be compared to other first-person action shooter environments. In contrast to the other games in the loot-based shooter genre, it departs from a few typical conventions. It features on-the-fly matchmaking, which matches the players with other groups without requiring their action while they play. This requires the players to always be online while playing the game. The game offers two game-modes: Player-vs-Environment (PvE) and Player-vs-Player (PvP). PvE Missions can be played alone or in cooperation with other players, and make up the majority of the content a player can experience. The PvE mode is mainly used to tell the story of the game, and the main story missions can be played with teams that have up to three players. The main zone, where people are preparing for their Missions is called the *Tower*. The *Tower* is the starting point and home base for players. It is also a neutral zone where players can socialize and form parties without experiencing combat. At the tower and other similar locations, players can also experience the main RPG elements *Destiny* provides. It is possible to buy items, collect challenges known as bounties or to build reputation among certain factions. To leave the *Tower* and go on a mission, players have to take a vehicle called a “Jumpship“, which carries them to mission locations on earth or the rest of the solar system.

The PvP mode, which is called “Crucible“ in *Destiny* encompasses missions played against other players in teams, called “Fireteams“. Crucible matches range from three to six players per fireteam, and come in several game types. These game types include scenarios where a team has to take control of a zone or area and hold it, a classical death-match mode between teams, a capture-the-flag mode, and a few more variations on those with varying tactical objectives.

3.1.2 Dataset

The *Destiny* dataset observed here, consists of the collected data of more than 3.5 million *Crucible* matches. To generate the *Destiny* dataset, a random sample of 10000 Players which have played for at least 2 hours was selected. This was done to exclude players from the analysis which just installed the game and did not play the game afterwards, or players who primarily focused on the PvE aspects of the game. The *Destiny* dataset consists of two parts. The first part is detailed player information about the 10000 players used as basis for the random sample. This includes data about what characters a player created, their characters statistics, character equipment and a multitude of customization information. The second part contains data about PvP *Crucible* matches with a variety of gameplay metrics collected by Bungie. The dataset consists out of

Table 3.1: Statistics of the Destiny dataset

Players	3,450,622
Matches	930,720
Clans	318,007
Classes	3

a collection of 930720 lines of individual json files, each representing a *Crucible* match, with various information that is used in further analysis.

The recorded matches span from September 2014 to January 2016. The files offer information about the participating players, a big amount of performance metrics (Kill, Deaths, Damage Done), weapon load-outs and different scoring mechanisms. The amount of data provided is shown in Table 3.1 (Pirker et al., 2017).

There is a varying degree on how much time players spend with the game and how much experience they gathered. Fig. 3.2 (Pirker et al., 2017) shows the Level Distribution in the dataset. It includes markers for when the game was expanded through new Downloadable Content (DLC).

The spike that is displayed in the referenced figure can be explained by two occurrences: When a new expansion is released the maximum level is increased, and players either buy the expansion and keep playing or move on. Especially noticeable is the lack of players between level 35 and 39. This was caused by a fundamental change of the game mechanics with the release of the expansion *The Taken King*. The expansion changed the leveling mechanic, so that players can now increase their characters level to the maximum level of 40 in a relatively short time compared to previous expansions and the basic game. Beyond the normal levels, *Destiny* has a system called “Lightlevels“. A Players Lightlevel compared to the normal leveling system is affected by the armor, weaponry and artifacts a Player has equipped to his guardian, and can be improved by gaining better versions of those. Some of the events and missions in the game can only be entered when the player reached a certain Lightlevel.

Table 3.2 (Pirker et al., 2017) shows how many matches the players in the dataset played against each other. 97.93% of all players in our dataset played less than 11 games.

Similar to many other online games with role playing elements, players can be part of a group other than their normal friends. This group called a “clan” in *Destiny* and is used socially and for creating teams in events that require more than just a few players. The dataset contains 31807 clans with 1.3 million players who belong to the clans. Clans in *Destiny* tend to be relatively small groups of

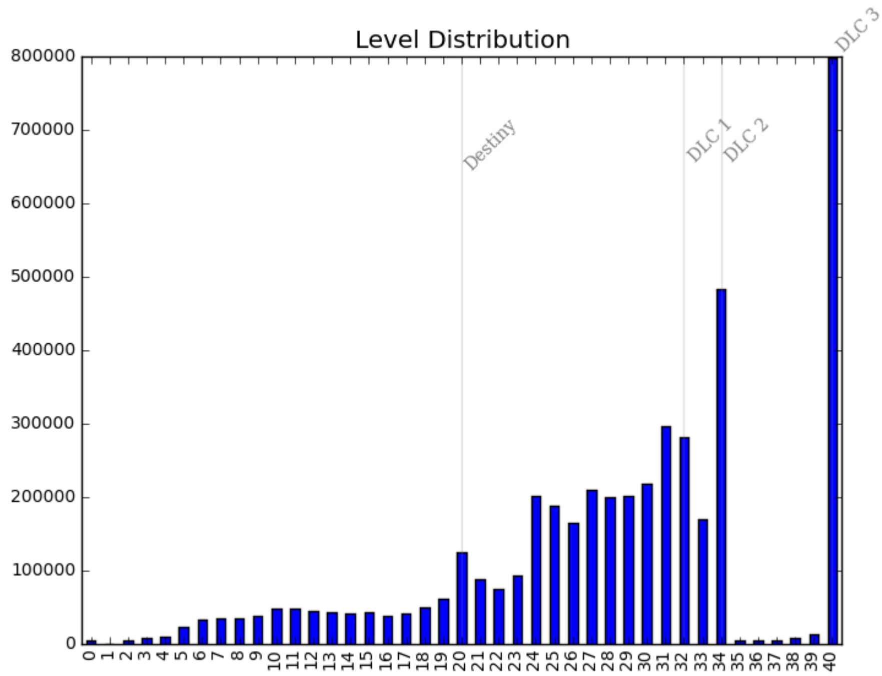


Figure 3.2: Level distribution in Destiny (Pirker et al., 2017)

Table 3.2: Number of matches played by players

Games	Players
1-10	3,293,187
11-20	54,836
21-50	8,758
51-100	2,660
101-200	1,674
201-300	610
301-500	469
501-1000	333
1000+	109

people with a range between 20 and 60 Players. The distribution of clan sizes is shown in Fig. 3.3.

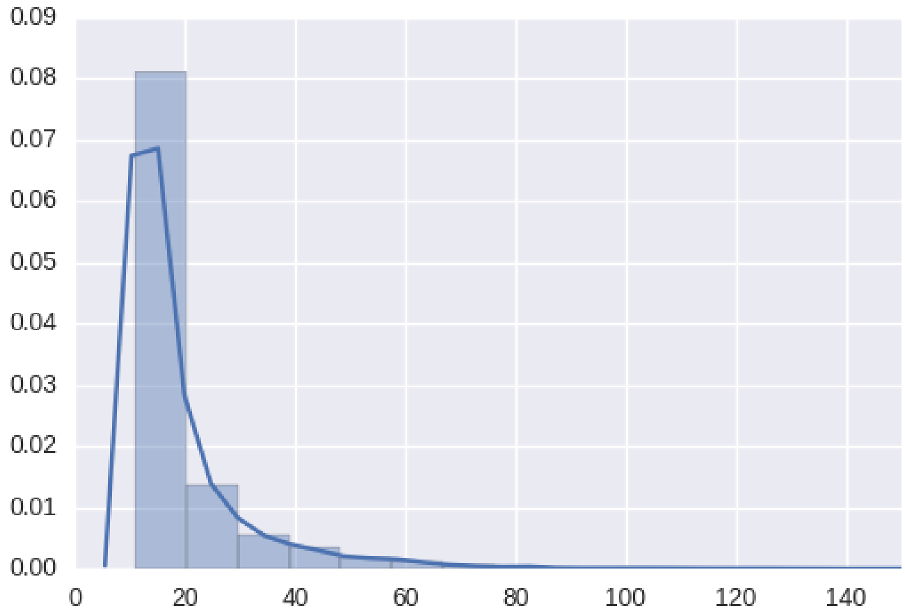


Figure 3.3: Clan Size Distribution in the Destiny Dataset.

Player Preferences

In the *Destiny* dataset not every class a player can choose from is equally popular: 38.64% are playing as the class Hunter, 29.20% as Titans, and 32.15% as Warlocks. Fig. 3.4 (Pirker et al., 2017) visualizes that distribution.

3.1.3 Preprocessing and Feature Extraction

The first step in preprocessing the dataset was excluding unwanted match types. The dataset includes data from two special events that can have a lot of players, that do not qualify for what we call PvP Matches. The excluded modes are the Free-for-All and Mayhem Rumble, which can include over 100 players in a single json file that should only represent a single match. To conduct this, all game modes needed to be identified from their assigned ID in the dataset, and removed accordingly. These match types were removed for comparability with other games and the second dataset, which does not display comparable social structures. The second step was to identify the wanted features for further analysis. The dataset contains a huge amount of features of each single match. The most important ones are the players team, the duration of a game, and

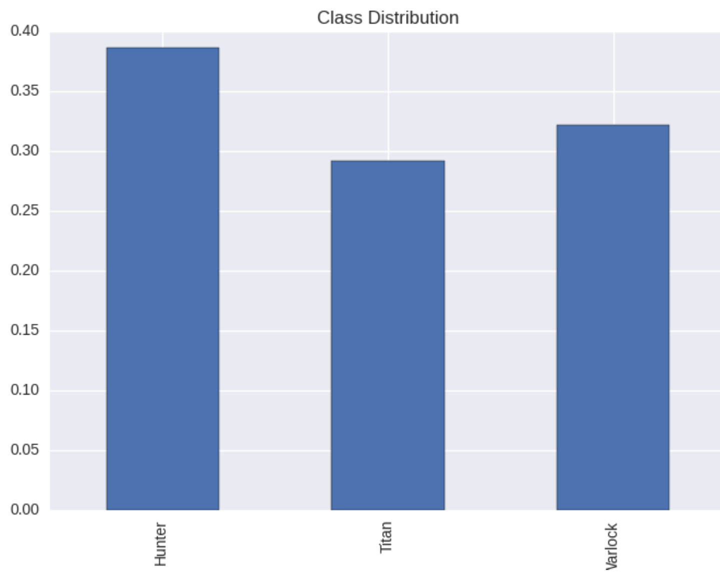


Figure 3.4: Class distribution of players' first-choice character (Pirker et al., 2017)

Table 3.3: Features of the Destiny dataset

Performance	Kills, Deaths, Wins, Losses, Accuracy
Engagement	Duration of Playtime, Total Playtime, Retention
Social	Clan membership, Team membership

the teammates. The data provided with the files is split into three categories for the next steps: Performance, Engagement and Social Features. Performance features are all features that provide information about how well a user performs in the game. Common features in this categories are kills and deaths, but they also provide information about the overall behavior of the player. When looking at the kills a player achieved with a certain weapon or a certain range it is possible to extract behavioral patterns. Engagement features are all features that look at the time spend playing the game. Important features in this category are the duration of a single game session, the total playtime the user spends with the game, the length of the matches and how often a player comes back to the game. Social features are everything that indicates a relationship or a social interaction. Important social features for further analysis in this work are games played together and clan membership. An overview of some of the other important features in the *Destiny* dataset can be seen in Table 3.3.

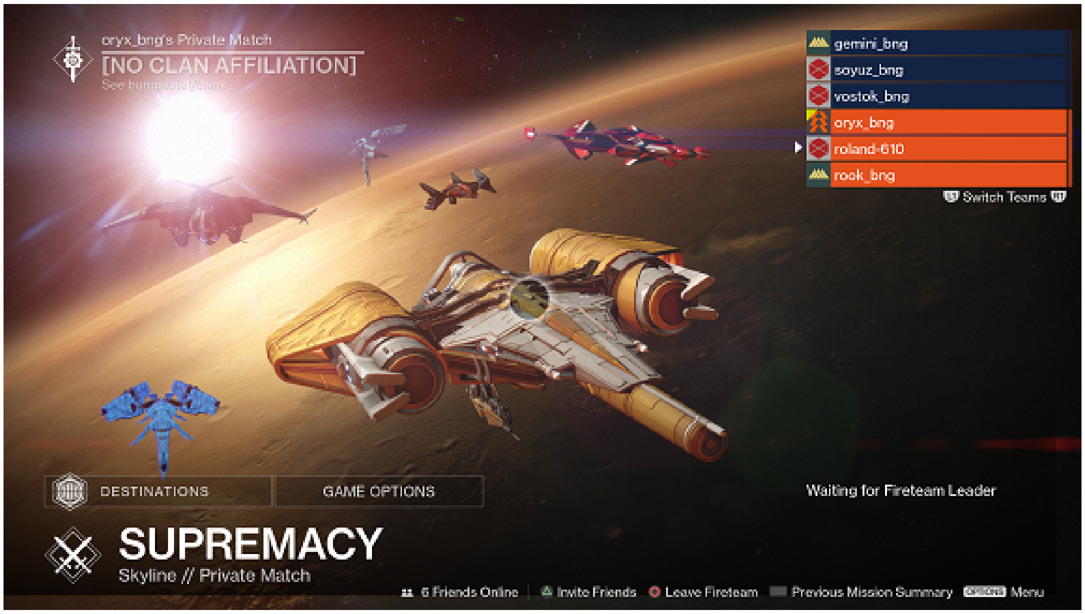


Figure 3.5: Overview when setting up a private match in *Destiny*. (c) Bungie, Inc, Destiny, the Destiny logo, Bungie and the Bungie logo are registeret trademarks of Bungie, Inc. All rights reserved. Image by courtesy of Bungie Inc.

The next step in preprocessing was to map player measurements to actual values, and normalize them. Some values like who won a game are mapped to zero for the losing team and one for the winner.

3.2 Massive Open Online Courses (University of Galileo)

The *University of Galileo*[2] is a University located in Guatemala City, which was founded in October 2000, and offers a multitude of Massive Open Online Courses (MOOC),

3.2.1 Dataset

The MOOC Dataset consists of eleven courses offered by the University of Galileo in Guatemala. Every single one of the courses was running for 8 weeks. The data for each MOOC consists of all requests conducted by the user, their forum interaction, as well as information on how they scored on the different assignments in

[2]http://www.galileo.edu/

the courses. This dataset is quite different from the *Destiny* dataset discussed above, as the general forms of interaction and the size of the datasets differ quite a bit.

The dataset displays a high dropout rate, making for only a few successful participants of each of the courses. Fig. 3.6 visualizes the general distribution of final points the participants reached in the courses. Each of the visualizations has been cleared of any participants that signed up to the course but never scored any points. Therefore it displays the results of all users that put in some amount of work, and excludes everyone who signed up for the course, but did not do any of the course work. Even with the cleanup step, each of the courses displays a big spike where participants dropped out of the courses early, and did not turn in any of the later assignments. Table 3.4 shows the completion and dropout rates of the course. Another important aspect of the dataset are the forum entries. Forum entries are the main way users interact with each other. Not every course has the same amount of interaction, some courses encourage exchange between the users more than others. This might be caused through harder courses needing more discussion, or different levels of clarity of the assigned matter.

Table 4.7 shows the amount of interaction that took place in the forums. Especially noticeable is the amount of forum threads and replies created in the *Community Manager* course. The course deals with social media, and might encourage social interaction more than other courses looked at in this thesis. The third part of the dataset consists of all the requests conducted by users. This can be helpful to track the periods where users performed different actions on the MOOC site, fulfilled assignments or even interacted with each other.

3.2.2 Preprocessing and Feature Extraction

The *Galileo* dataset is in the form of csv files, which are easy to handle with data science libraries in the form of tables, but needed some amount of cleaning.

The first step in cleaning the data is to normalize the point scales the course participants were graded with. The maximum scores that could be reached in a course varied between 80 and 120 points. This was normalized to a scale from 0 to 100. The data was labeled with the maximum scores included in the title and all labels had to be replaced to be unanimous.

Forum entries were referencing the previous post made, and not the overall thread. A column has been added to display to which thread they belong. This proves to be helpful when creating the actual networks. Similar to *Destiny*, the features where divided into Performance, Engagement and Social Features.

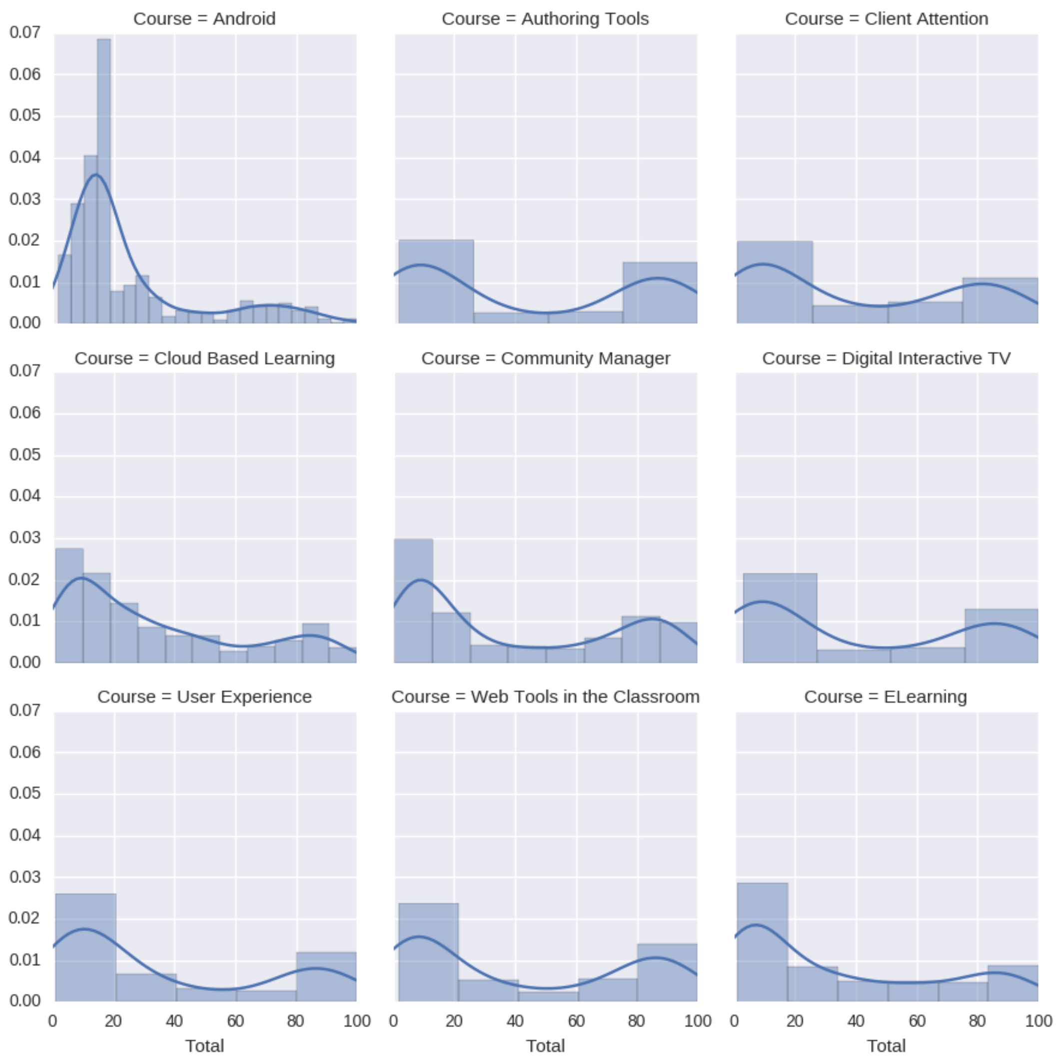


Figure 3.6: Grade Distribution

Table 3.4: Participants and Completion Rates of the Online Courses

Course Name	Participants	Max. Score	Successful	Dropout (%)
Android	509	120	78	84.68
Authoring Tools	239	105	106	55.65
Client Attention	169	100	68	59.76
Cloud Based Learning	559	100	138	75.31
Community Manager	918	100	344	62.52
Digital Interactive TV	167	105	68	59.28
Medical Emergencies	651	100	67	89.71
User Experience	204	100	63	69.12
Web Tools, Classroom	333	105	137	58.86
Web Tools, Educational	206	105	103	50.00
ELearning	316	100	97	69.30

Table 3.5: Forum Entries and Threads in the MOOC

Course Name	Forum Threads	Forum Entries
Android	531	2754
Authoring Tools	163	1067
Client Attention	53	612
Cloud Based Learning	1080	7037
Community Manager	1920	15213
Digital Interactive TV	94	655
Medical Emergencies	184	1153
User Experience	85	618
Web Tools in the Classroom	262	1322
Web Tools and Educational Applications	118	953
ELearning	547	4374

Performance Features are the grades the participants of the courses achieved. Engagement Features are based on the number of interactions, session times and general time spend interacting with the website. Social features involve everything that let users interact with each other. In this dataset, the main form of interaction between users take place in the form of forum entries.

3.3 Summary

Cleaning and preprocessing the datasets is an important step for further analysis. In the exploration phase that takes place before the actual preprocessing, many of the factors that can have a negative influence on the experiments and the building of the graphs can be prevented. Some of those factors can be well hidden especially in big datasets. An approach that can help with such problems is plotting the distribution of various factors before starting further analysis. Preprocessing is a phase that can be very distinct between datasets, and the principal cleanup methods used can vary. Altough the datasets provided were relatively clean, meaning they did not have much false or missing data, some operations had to be performed.

The data from *Destiny* included information about unwanted game-modes, that do not fit into the metrics used to build the social network structures discussed in the next chapter. Progress Players made in the game and the amount of games they played in our dataset varies. A big part of the players have played less than ten games. The levels players is also a notable factor: Over 22% of all Players in the dataset have reached the maximum level in the game.

In comparison, the *Galileo* dataset, has a great number of participants, who did not score any points in a single assignment or test. Even after removing users who did not submit anything, the dataset still has a high dropout rate with a maximum of 89.71% of participants not completing a course successfully. Another interesting aspect of the *Galileo* data is, that the course *Community Manager* created a huge amount of forum interaction between members compared to other courses.

Chapter 4

Network Structures and Characteristics

In this chapter we are taking a look at the general structure of the social networks, and how this relates to the user behavior in the system. Specifically we are taking a look at how the social networks were build, and how the connections between the nodes were formed. All networks discussed in this section are undirected with weighted edges. Complex networks exhibit features which can be important to consider in later analysis. Relationships between the users of the systems can have different strengths, a user who interacted with someone else frequently will have a greater weight than someone that just interacted with another player once.

4.1 Network Building

There are several approaches to build a social network out of a dataset, with multiple ways to map different kinds of relationships to each other. Nodes in a graph can represent different entities depending on the requirements of the further analysis. The following sections discuss the choices used in constructing the graph structure.

4.1.1 Definition of Relationships

The usual way to define relationships in social networks is based on the definition of friendships or connections by the underlying system, but relationships exist and are expressed in many other ways. With the two datasets we compare, we analyze implicit social structure and map interaction users have with each other to replicate the underlying social system (van de Bovenkamp et al., 2014).

A simple example where many implicit interactions occur is a social network: a user might have a stronger connection with the person the user exchanges messages with on a regular basis, compared to a person that was added as a friend once and never interacted with.

The *Destiny* dataset we are building the social networks from does not contain any traditional friendships, but does contain information about the ways user interact with each other. This information is used to define the strength of the relationships in the social networks. To analyze the behavior of the players, we take a look at how often they are involved with each other in the games they play.

With the Galileo Dataset we look at relationships in an online course in a similar way. People who interact with each other, are more likely to form bonds. The main interaction type we use here are implicit connections that form between users and their interactions.

4.1.2 Destiny Networks

In the *Destiny* network, the players are represented by graph nodes. Based on the interactions observed from *Crucible* matches, we identify three ways users can interact with each other. These represent the weighted edges of the network, where the weight is the number of occurrences of the event.

- Players playing on the same team (*Teammates*, T): Players who play on the same side in a *Crucible* match might have joined a match together on purpose, or might start to recognize each other from previous games after having been assigned to the same team a few times be the on-the-fly matchmaking system.

- Players playing on opposing sites of a team (*Opponents*, O): Players who play each other often on opposing sites might develop a adversarial relationship with each other, hoping to get revenge for the last loss, or simply defeating a past enemy.

- Players who are part of the same match on either side (*Matchmates*, M): This way to look at relationships provides a combined metric out of the two previously mentioned types. Increases in the connections here other than the ones created by random matchmaking could for example stem from clans training with each other on multiple teams. Fig. 4.1 shows a typical clan network in the dataset that could be the basis for that. Otherwise it should be just a combination of the previous two networks.

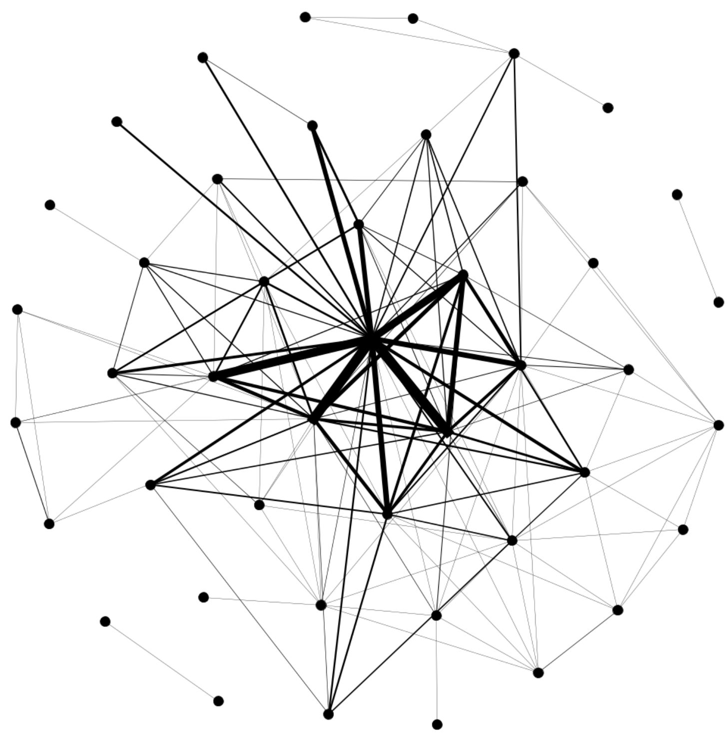


Figure 4.1: Destiny Clan: This figure shows a typical clan structure extracted through the implicit relationships in the dataset.

Table 4.1: Network relationships (Pirker et al., 2017)

M	Players in the same match (Matchmates: M)
T	Players playing together in the same team (Teammates: T)
O	Players playing against each other as opponents (Opponents: O)

Those observations build the foundation on how the connections in the destiny networks are defined. Table 4.1 provides an overview of the created networks and their abbreviations. Table 4.2 shows how players of those different networks are connected with each other. A high number of games played together implies that the players know each other from many games. The marginal difference of players that played more than ten games together on a team compared to players who played in the same match shows that the underlying relationships do not benefit when building the network of players in the same match (M). The high number of meetings in the compliment network (O) when only one to five

Table 4.2: Number of matches played together between different players (Pirker et al., 2017)

Games	Same Team	Opposite Team	Complement
1-5	22,582,015	27,491,957	47,382,583
6-10	32,816	2,561	46,308
11-20	12,851	201	13,386
21-50	7,025	20	7,168
51-100	2,140	1	2,179
101-200	873	0	900
201-300	207	0	214
301+	135	0	140

matches were played is caused by random match-making of the system (Pirker et al., 2017). The three *Destiny* networks are all weighted graphs, where the weights signify how strong the connection between those players is. Therefore the weights correspond to the number of matches two players participated together in.

4.1.3 Galileo Networks

The *Galileo* dataset consists out of eleven courses with barely any overlap between the members of those. Based on the interaction and the knowledge that a lot of forum interaction took place in most of them, we build a social network for each one of them. This results in eleven small networks. Course participants are represented as graph nodes, and the forum interaction with each other is used as basis for the edges. The forum data is used as a sole source for mapping implicit relationships because it provides the best overall information on how users interact in the dataset. Another possible approach is to infer more social structure from unconventional interactions like using the same times users accessed a certain site or tool (Jia et al., 2015). Table 4.3 shows the names of the courses and the number of nodes and links of the resulting networks.

The main way to create implicit connections between the *Galileo* course members is to look at their interaction in the communication platforms provided. Most of the interaction in the dataset is captured by the actual conversation in forum entries. A single thread in a MOOC Forum was taken, and connections were added for all users who participated in the thread. Participants who answered at the beginning of a thread, compared to someone who replied much

Table 4.3: Overview of Galileo Course Networks

Network Course	Nodes	Links
Android (A)	437	7209
Authoring Tools (AT)	454	12221
Client Attention (CA)	178	6447
Cloud Based Learning (CBL)	445	21549
Community Manager (CM)	2114	93947
Digital Interactive TV (DITV)	232	6190
Medical Emergencies (ME)	208	6245
User Experience (UE)	331	8859
Web Tools in the Classroom (WTC)	549	14020
Web Tools and Educational Applications (WTEA)	273	6305
ELearning (EL)	237	9885

later do not have the same relationship as someone that is directly replying to someone else though. That is why the distance between posts and a logistic function to create balanced weights on the interaction are used:

$$weight = \frac{e^{-distance}}{(1 + e^{-distance})^2}$$

With many users communicating with each other in every single thread, this leads to more pronounced connections, with a more realistic weight distribution.

4.2 General Structure

This section discusses the general structure of the networks and takes a closer look at the properties they exhibit. In addition to that, differences between the networks are discussed and analyzed, and peculiarities of the networks are shown.

4.2.1 Destiny

A part of the networks that were generated out of *Destiny* match-up data can be seen in Fig. 4.2 . The figure is part of the generated team (T) network, and contains 1000 nodes and 2356 edges. All of the *Destiny* networks are far too big

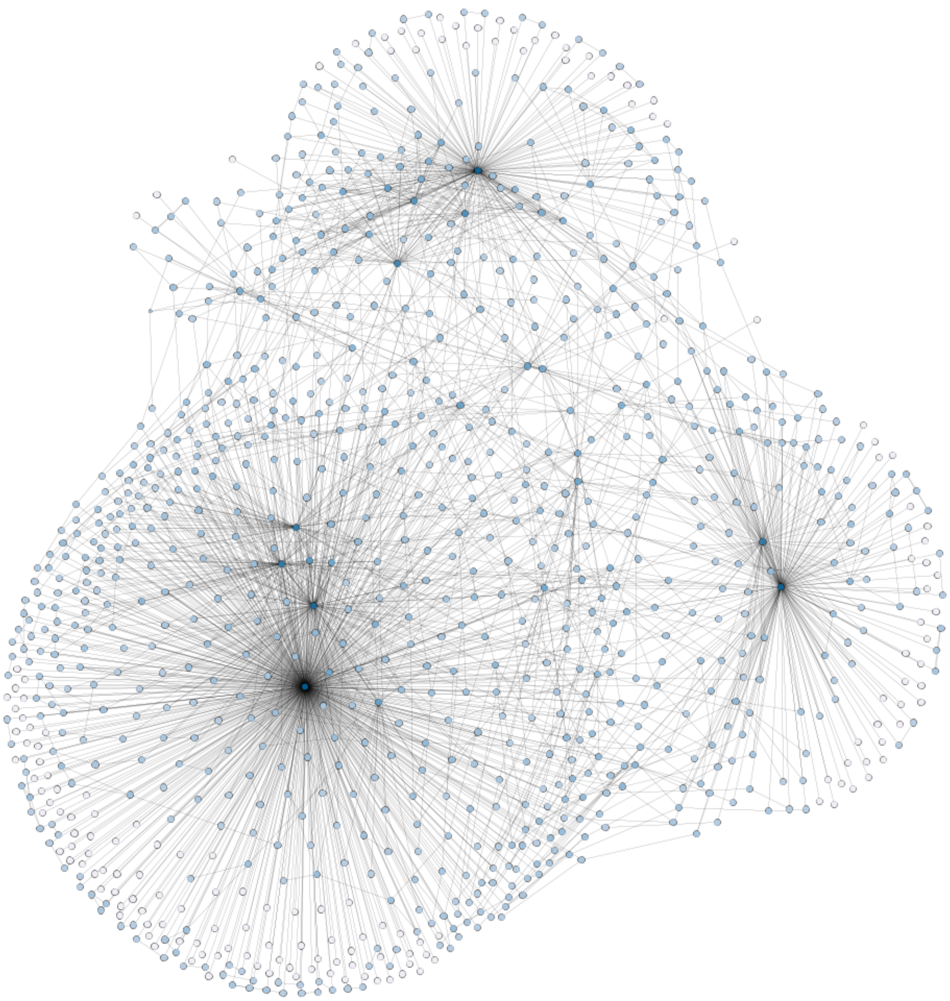


Figure 4.2: Destiny Network Example: This figure shows a part of the Destiny Network that was extracted by choosing a few well connected players as starting points.

Table 4.4: Overview of Destiny Networks (Threshold minimum games played - 3) (Pirker et al., 2017)

	Same Team (T)	Opposite Team (O)	Same Match (M)
Nodes	725,704	725,704	725,704
Links	6,729,257	8,682,726	14,048,455

Table 4.5: Overview of the threshold behavior (Pirker et al., 2017)

Min Games	Nodes remaining % (Rel)	Edges remaining % (Rel)
1	55.46 (55.46)	68.53 (68.53)
2	33.68 (60.72)	45.21 (65.97)
3	21.58 (64.09)	29.73 (65.74)
4	14.35 (66.47)	19.64 (66.08)

to visualize, with all of them having over 725 thousand nodes. The figure was generated by searching for a few players who have strong connections with each other and doing breadth first search for players who are connected to them. This results in a representation of network slice that is close to the actual network structure.

Table 4.4 shows an overview of the networks used for further analysis. All of the players who did not play a minimum of three games are removed from the initial networks to only include people with a certain amount of experience, and players where we have enough data for. This thresholded version of the networks is used in all analysis in future chapters. Figure 4.3 shows how the resulting network changes when certain thresholds are set. When deleting every player that did not play a certain amount of games, the network shrinks greatly, having less than a third of its size when everyone is removed that did not play at least the three games. Table 4.5 shows how many nodes and edges remain in the network after removing players and their connections. After removing every player that played less than fives games, only 14.35 % of the nodes and 19.64 % of the connections are still existent. This is caused by the mode of data collection discussed in Chapter 3. Table 4.6 shows the number of games the players in the dataset played with each other. 97.93% of all players have played less than 11 games (Pirker et al., 2017).

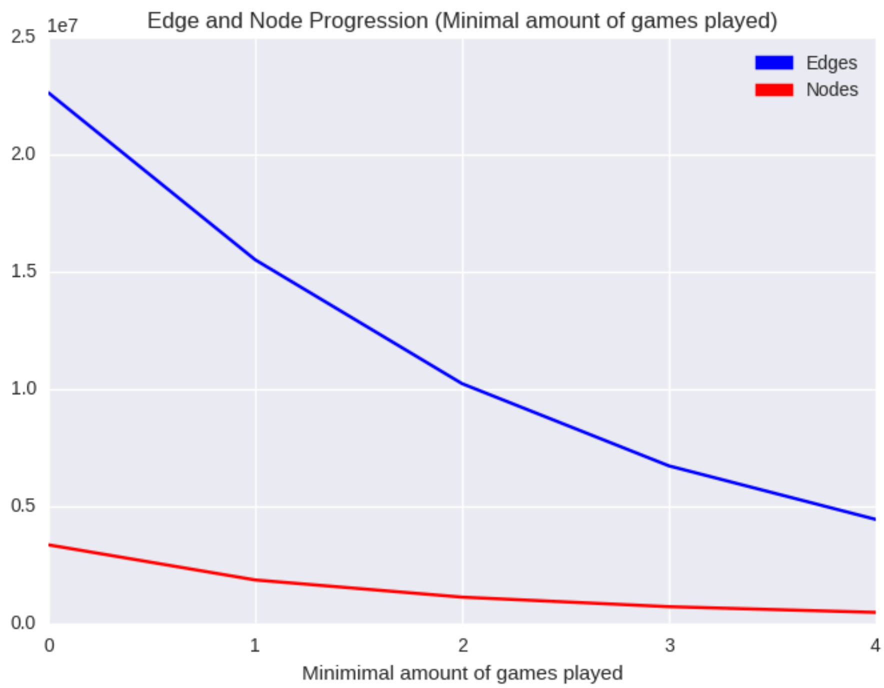


Figure 4.3: Deletion of nodes - After removing players who have not played together at least four times many connections are removed (Pirker et al., 2017)

Table 4.6: Number of matches played by players (Pirker et al., 2017)

Games	Players
1-10	3,293,187
11-20	54,836
21-50	8,758
51-100	2,660
101-200	1,674
201-300	610
301-500	469
501-1000	333
1000+	109

4.2.2 Galileo

Fig. 4.4 shows the network for the course "authoring tools". All course data results in similar networks size and number of connections. The nature of forum threads to have more than one reply can be observed when looking at the figure. None of the visible nodes that can be observed are endpoints. Table 4.7 illustrates the amount of forum posts users have composed. The majority of the user-base has written less than 6 posts.

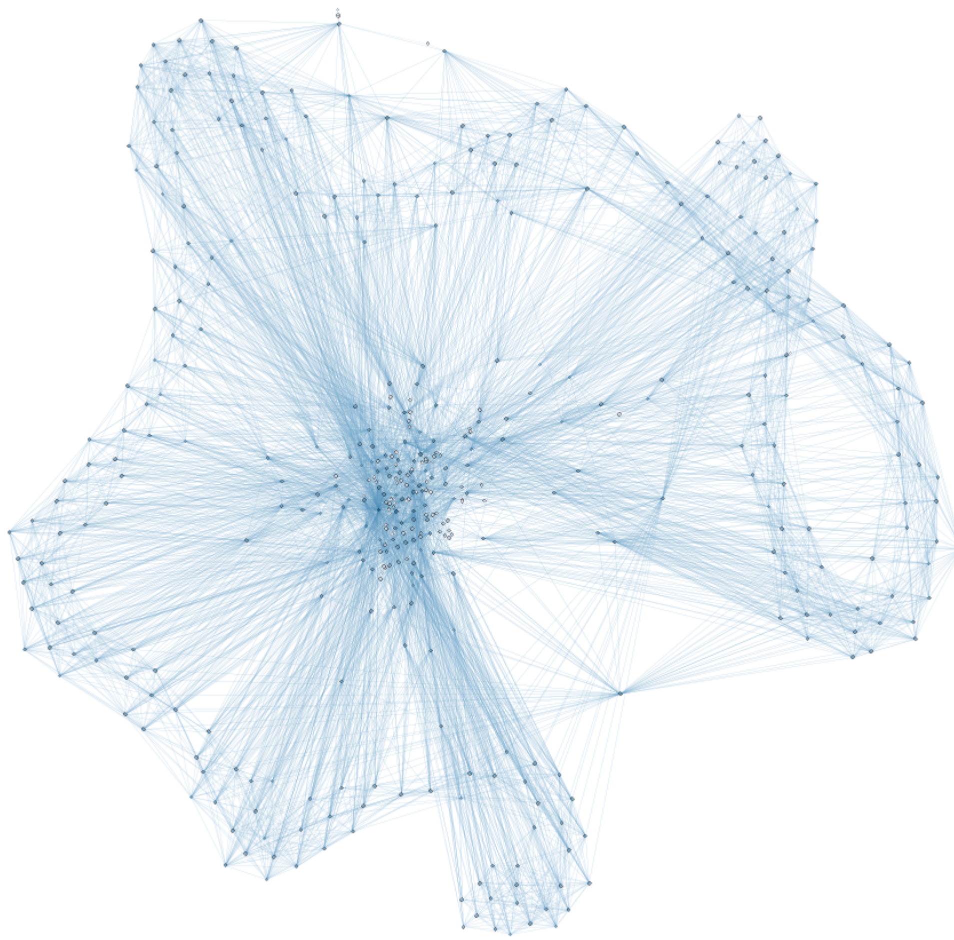


Figure 4.4: Galileo Network Example: The network shown here is one of the eleven courses in the dataset, "Authoring Tools". Im comparison to the Destiny Network, the figure shows the entire network.

Table 4.7: Number of forum posts by users

Forum Posts	Users
1-5	4087
6-10	441
11-15	158
16-20	80
21-25	48
26-30	39
31-35	21
35-40	14

4.3 Network Measures and Characteristics

This section takes a closer look at common network measures and characteristics, providing deeper insights about the general structure of the networks. Big or complex networks are oftentimes hard to compare to each other, and visual comparison is unfeasible. The metrics presented here are designed to give an overview on the structure and composition, and to help drawing conclusions about them. Table 4.8 shows an overview of common network measures for the three generated *Destiny* networks. Table 4.9 shows an overview of the network measures for all of the *Galileo* networks.

Table 4.8: Methodological comparison of the three networks (Threshold minimum games played - 3)

	Same Team (T)	Opposite Team (O)	Same Match (M)
Nodes	725,704	725,704	725,704
Nodes in LCC	725,599	725,693	725,703
Avg. Degree (k_avg)	18.55	23.93	38.72
Links	6,729,257	8,682,726	14,048,455
Links in LCC	6,729,190	8,682,726	14,048,455
Diameter (D)	13	11	9
Avg. Cl. Coeff. (C_avg)	0.024	0.0082	0.026

Table 4.9: Methodological comparison of the Galileo Networks: A = Android, AT = Authoring Tools, CA = Client Attention, CBL = Cloud Based Learning, CM = Community Manager, DITV = Digital Interactive TV, ME = Medical Emergencies, UE = User Experience, WTC = Web Tools in the Classroom, WTEA = Web Tools and Educational Applications, EL = ELearning

	A	AT	CA	CBL	CM	DITV
Nodes	437	454	178	445	2114	232
Nodes in LCC	389	439	177	442	2106	226
Avg. Degree (k_avg)	32.99	53.84	72.44	96.85	88.88	53.36
Links	7209	12221	6447	21549	93947	6190
Links in LCC	7201	12217	6446	21549	93942	6190
Diameter (D)	5	5	4	4	6	4
Avg. Clust. Coeff. (C_avg)	0.34	0.516	0.556	0.379	0.272	0.606
	ME	UE	WTC	WTEA	EL	-
Nodes	208	331	549	273	237	
Nodes in LCC	205	321	526	270	236	
Avg. Degree (k_avg)	60.048	53.528	51.074	46.19	83.417	
Links	6245	8859	14020	6305	9885	
Links in LCC	205	319	522	268	236	
Diameter (D)	4	5	6	4	4	
Avg. Clust. Coeff. (C_avg)	0.476	0.630	0.517	0.511	0.478	

Edge Weight Distribution: The distribution shows how the weights in the player-base are distributed. The weights in the *Destiny* graphs are based on how often users play with each other. Fig. 4.5 shows the distribution of the weights in when players play together or against each other. In Fig. 4.5b it can be observed that the weights of the adversarial network are not well defined, because adversarial relationships between players are not well defined or even non-existent in the *Destiny* dataset. The weights in the *Galileo* networks are defined in a different way as the *Destiny* networks, and are kept relatively low. This is caused by method of calculating the weight, and makes the networks hard to compare in terms of weight. There are some outliers that receive really big weights up to 50 though. These members can be seen as central figures in the community. The weight distribution of the combined *Galileo* networks can be seen in Fig. 4.6 .

Largest connected component (LCC): The largest connected component is the biggest subgraph of a network, where a subgraph is a graph where all nodes have at least a single connection to another node of the graph. Table 4.8 shows that the largest component almost spans the whole graph, with only a few nodes that are not part of the LCC. The *Galileo* networks show similar behavior in this regard. There is one big component in all of them, with only a few nodes that are not part of it. Compared to other systems, it is expected to have connections between almost all of the nodes in the network. This is due to the used method of constructing the graphs.

Average Degree (k_avg): The average degree of a graph expresses how many edges are in the graph compared to the nodes:

$$k\_avg = \frac{2 * E}{V}$$

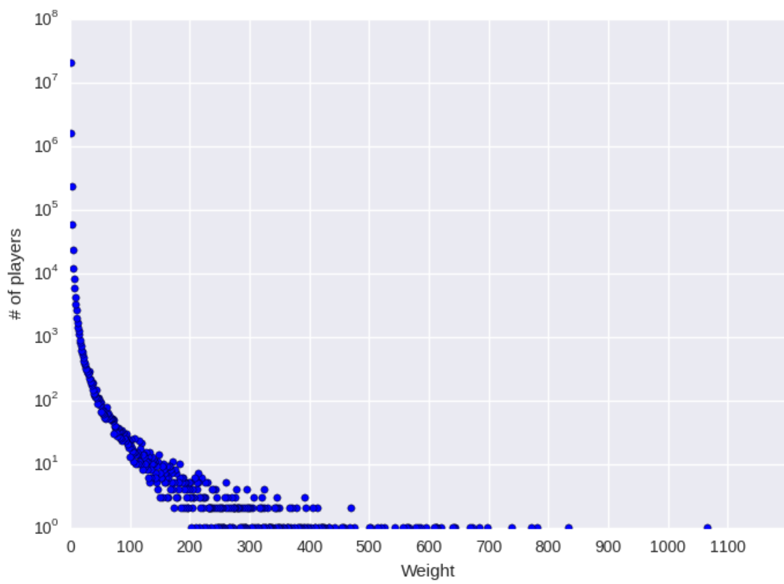
The average degree of the *Galileo* networks is generally greater than the average degree of the *Destiny* networks, implying that the nodes are better connected with each other. This implies that there is more implicit interaction to be found in the *Galileo* dataset, but could also mean that we have not found and mapped unknown relationships in the *Destiny* dataset. It is likely that there are more relationships to be found and explored when more data vectors are collected.

Average Clustering Coefficient: The clustering coefficient is a measure on how much nodes tend to build groups together. It can be calculated by the following equation:

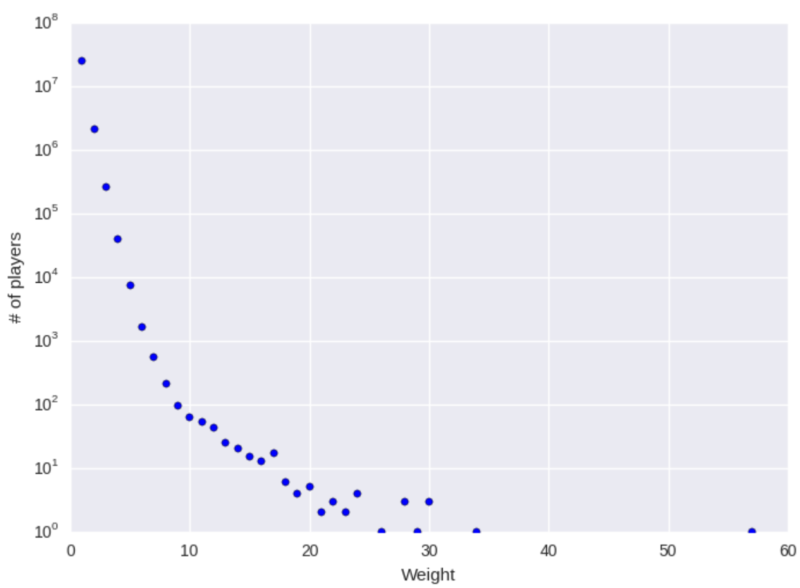
$$C(v) = \frac{E(v)}{k_v(k_v - 1)}$$

The average clustering coefficient is the average of all coefficient over all clusters, and can be used to tell if the implicit connections from the build networks correspond to real groups. Due to the smaller nature of the course networks, they also exhibit a greater clustering coefficient. Notable in the *Destiny* networks is that the opposite team network (O) exhibits a clustering coefficient that is much lower compared to the other networks, furthering the evidence that adversarial behavior can not be observed from the dataset.

Network Diameter: The diameter is the "longest shortest path" between two nodes in the network. The more connections a graph usually has, the smaller



(a) Playing on the same team



(b) Playing against other players

Figure 4.5: Edge weight distribution (Pirker et al., 2017)

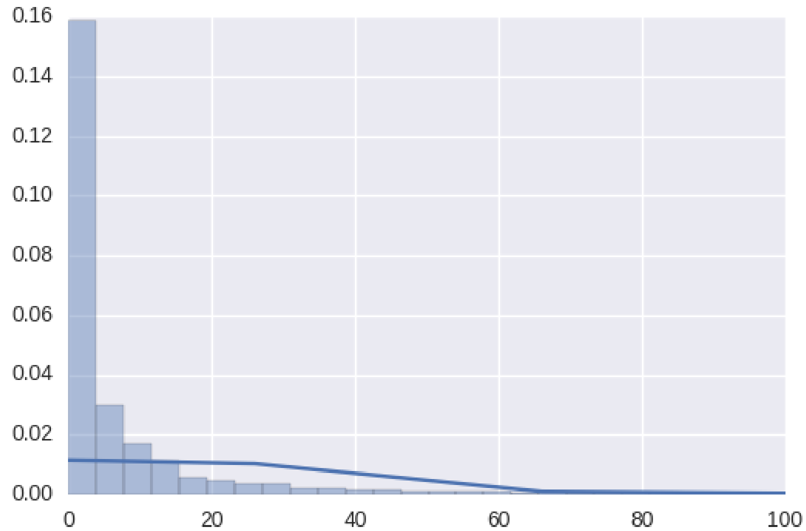


Figure 4.6: Galileo Weight Distribution

the degree usually gets. This can be seen in all of the networks created. The match-mate network (M) in *Destiny* is a combination of the other two networks and therefore exhibits the smallest diameter. In the *Galileo* networks, the ones that are connected best have the smallest diameter. The difference between the datasets stems from the sheer size discrepancy and from the fact that the *Galileo* networks have more connections.

Degree Distribution: Table 4.10 shows how degrees of the *Destiny* networks are distributed. Most of the players have over six connections and less than 21 connections. This is caused because the minimum amount of players for most events is three, creating at least two connections for each player. Fig. 4.7 shows the combined degree distribution of all *Galileo* course networks. When looking at the figure, two peaks can be observed: The first one is below the 10 connections mark and the second one is around the 30 user mark.

4.4 Summary

When creating social networks based on datasets, it can be a good idea to look at implicit relationships. Implicit relationships are all relationships that do not express themselves in an explicit form, like friendships. The two ways interaction is mapped on the discussed datasets, are how and how often users play together and

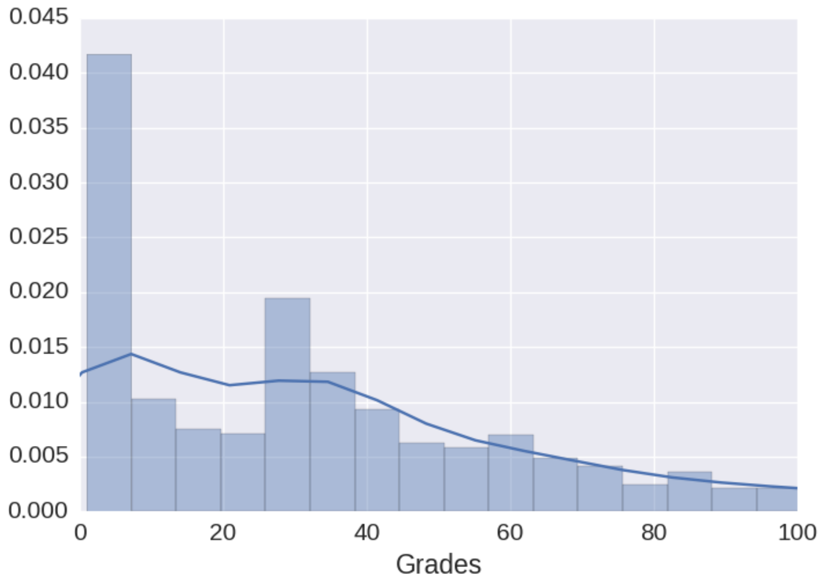


Figure 4.7: Galileo Degree Distribution

Table 4.10: Comparison of Network Node Degrees

Degree	Same Team (T)	Opposite Team (O)	Same Match (M)
0 - 2	1,477	1,990	12
3 - 5	54,812	128,146	1,747
6 - 10	1,627,084	1,502,516	145,872
11 - 20	1,004,600	991,962	1,703,801
21 - 30	322,135	377,496	617,112
31 - 40	129,651	170,993	318,234
41 - 50	56,783	82,109	193,247
51 - 60	26,379	41,892	123,064
61 - 70	12,987	22,646	80,429
71 - 80	6,766	12,535	52,356
81 - 90	3,726	7,152	35,120
91 - 100	2,160	4,479	23,848

how they interact with each other using forum messages and threads. For later comparability, a few variations on the networks were created. *Destiny* features a network for players who played with each other, players who played against each other and players who were in the same match. All of the *Destiny* networks used in the chapter and in further analysis, are thresholded to a minimum of three games players had to play. Although the *Destiny* dataset is much bigger than the *Galileo* networks, it has a few limitations: A great amount of players participated only in a small amount of games. When removing those players, some information gets lost in terms of how connected the remaining players are and some communities might get smaller or vanish.

For the *Galileo* dataset, a network was build for each course. The networks that were build are all undirected weighted graphs, where the users or players are the nodes and the implicit relationships are the graph edges. Forum interaction is the main way the edges are mapped in the *Galileo* networks. The size of the networks compared to *Destiny* is so small that they still can be visualized using common graph visualization techniques. Network measures are a way to get insights about networks and their characteristics that are too big or complex for visual interpretation. Measures can also help to make networks more comparable to each other. Although the *Destiny* networks are bigger than the *Galileo* networks, certain parallels on their structure can be drawn. Specifically faulty networks that do not contribute anything in further analysis can be identified and removed like the *Destiny* adversarial network.

Chapter 5

Data Analysis and Results

In this chapter several experiments for each dataset are performed to get a overview of the networks in terms of their performance and engagement. Each of the observed datasets has different values related to the users success, and we define metrics that have as much in common as possible. In addition to the performance and engagement experiments, several behavioral features of the datasets are examined.

5.1 Destiny

Some of the main measures that relate to success in the *Destiny* dataset are represented as how their success in a particular game was. We define this as combat performance and use the kill/death ratio as a metric. To determine the overall success of a player for all the games played, we use the win/loss ratio. In addition to that we use another performance metric which gives more insight on how efficient players are. The metric used for that is time spend per match. All of the mentioned experiments on the metrics are in relation to the actual player behavior, which will be defined in more detail in the next section.

Table 5.1 gives an overview of the metrics used for the different categories.

To measure how different user groups behave, we rank players according to how they choose their teammates: Players who play and interact with the same group of people a lot (*Player Group 1: Focused Players*) reach a higher score in our metric. This is shown in equation 5.1 (Pirker et al., 2017).

$$FocusedPlayer = \frac{Sum\,of\,weights}{degree} \cdot \frac{\#matches\,played}{\#matches}$$

Players who prefer to use the random match-making (*Player Group 2: Open Players*), will naturally score lower. In addition to the first part of the equa-

Table 5.1: Destiny Experiment Features (Pirker et al., 2017)

Performance & Engagement Metrics	
Performance	win/loss ratio kill/death ratio time/match ratio
Engagement	Number of matches played Total Playtime
Other	Clan Membership

tion which calculates the group interaction, the second part adds a factor to the equation to remove the disadvantage players that played a lot of games have otherwise. *Number of matches played* is the sum of all matches a player participated in. *Number of matches* is the sum of all matches recorded in the dataset.

5.1.1 Performance & Engagement

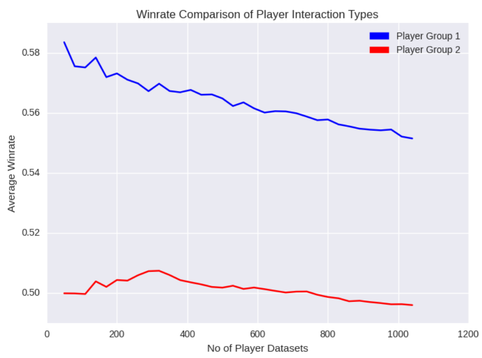
To measure the performance of the players we perform the following experiments with the metrics listed in Table 5.1:

1. How do the relationships in the game influence the win/loss ratio?
2. How do the relationships in the game influence the kill/death ratio?
3. How much time do the different Player Groups spend playing games compared to each other?
4. Do relationships in the game have an impact on player engagement?
5. Has clan membership an impact on performance and engagement factors?

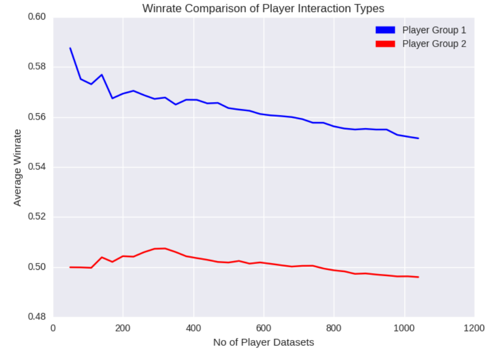
The combined experiment results are intended to provide a clearer picture about the general performance and engagement of *Destiny* players.

How do the relationships in the game influence the win/loss ratio?

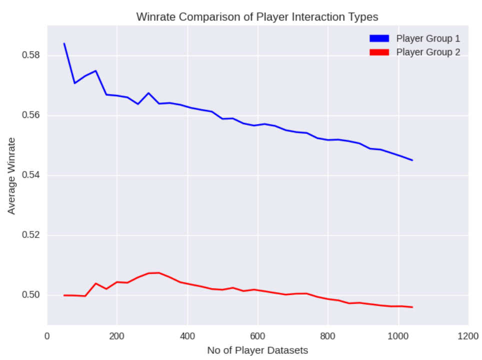
Fig. 5.1 shows how the average win/loss ratio for the two Player Groups. The x-axis represents the number of player datasets used to calculate the average win/loss ratio. When using 500 player datasets the figure shows the average win/loss ratio of the top 500 players ranked by equation 5.1. The subfigures



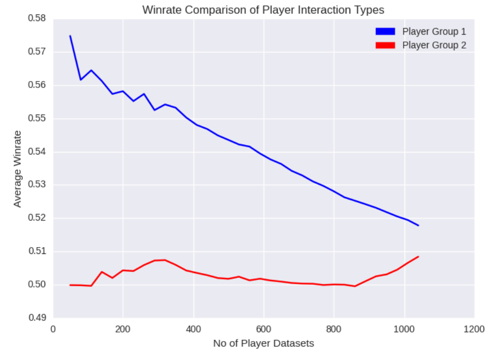
(a) 25 Games



(b) 50 Games



(c) 100 Games



(d) 250 Games

Figure 5.1: Win/Loss Ratio Comparison of Player Groups (Pirker et al., 2017)

Comparison of players that played at least a) 24 b) 50 c) 100 d) 250 games.

show the different thresholds applied to the player-base, and the impact it has on the average win/loss ratios. The thresholds are the minimum of games a player had to play to be included in the metric, to exclude players who had just a few bad or lucky games. The figure indicates two things: The win/loss ratio of users playing with random people or using the matchmaking system are almost 50%, which means that their performance is average and *Player Group 1* outperforms *Player Group 2* by up to 8%, which indicates that playing in similar teams might yield better performance in the game.

Fig. 5.2 visualizes the win/loss ratio compared to the degrees of the individual classes. The class *Hunter* scores slightly above the average and outperforms the classes *Warlock* and *Titan*. The relationship between the degree and the win/loss ratio is further visualized in Fig. 5.3 . Due to the prevalence of the *Hunter* class it also exhibits the biggest average degree of all classes (Pirker et al., 2017).

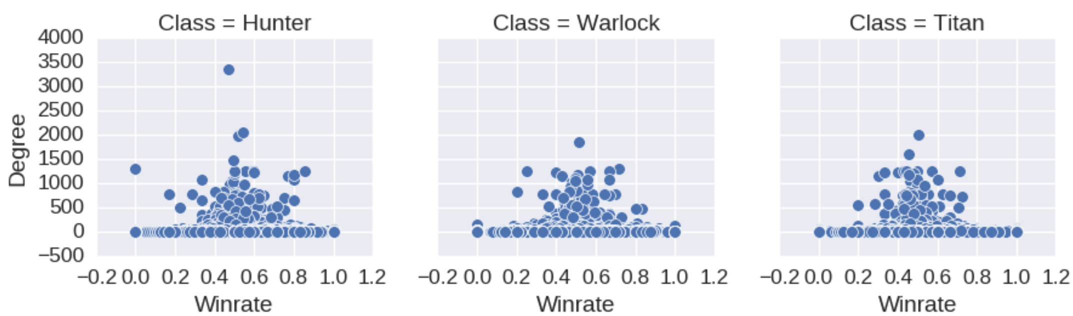


Figure 5.2: Winrate of Destiny Classes - threshold: 5 Games

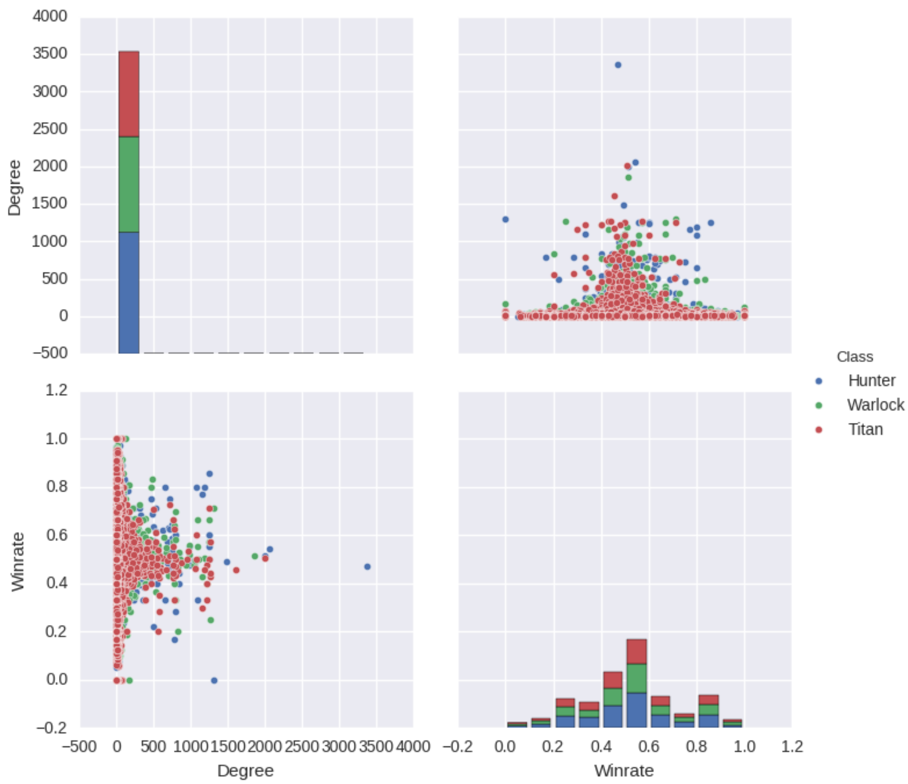
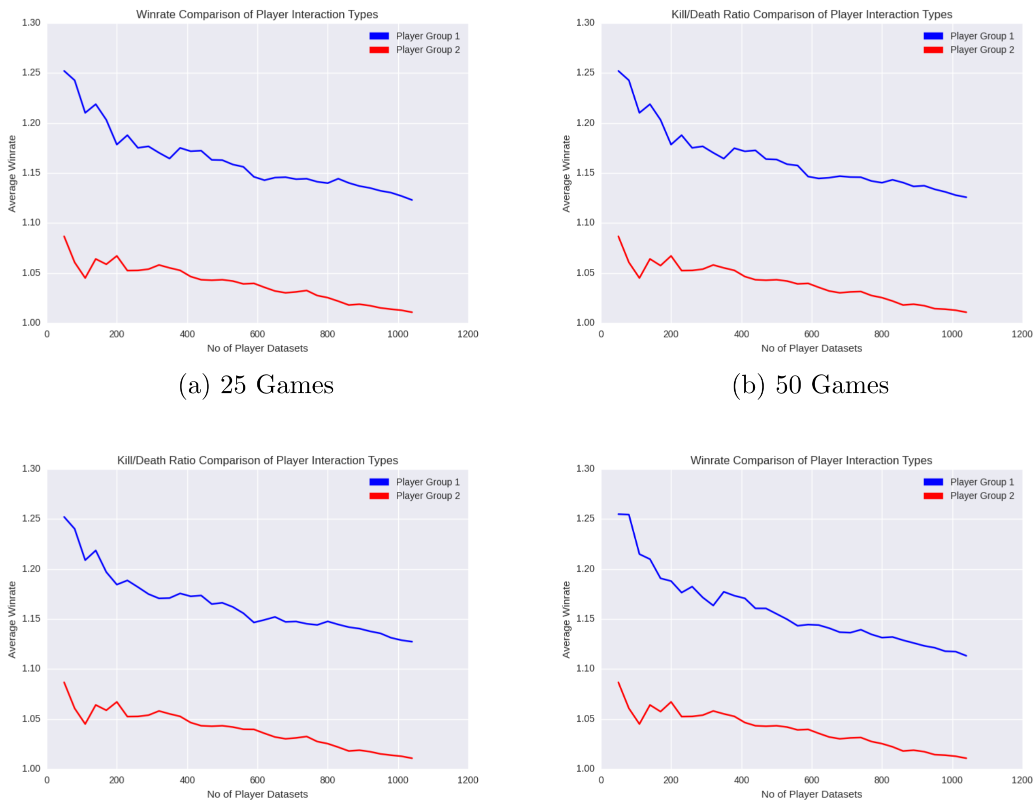


Figure 5.3: Pairplot of the Destiny Degree and Winrate - threshold: 5 Games

How do the relationships in the game influence the kill/death ratio?

Compared to the win/loss ratio, the baseline for the kill/death ratio is a value of 1, where everything greater than 1 is a positive above average outcome and better performance. Fig. 5.4 shows the average kill/death ratio with the respective thresholds applied. Notable compared to the win/loss ratio is that *Player Group*

2 performs better than the baseline, but is still outperformed by *Player Group 1* by a relatively big margin. Fig. 5.5 shows the general distribution of the kill/death ratio data, and also visualizes the same observation made in Fig. 5.4, that the general kill/death ratio in the dataset is slightly higher than the expected average of 1.



(a) 25 Games

(b) 50 Games

(c) 100 Games

(d) 250 Games

Figure 5.4: Kill/Death Ratio Comparison of Player Groups (Pirker et al., 2017)

Comparison of players that played at least a) 24 b) 50 c) 100 d) 250 Games.

How much time do the different Player Groups spend playing games compared to each other?

Fig. 5.6 visualizes how the playtime of successful players compared to average players. The success is measured by the win/loss ratio of the player. It is notable that only the players on the very top of the metric have shorter games compared to average players, and after considering the first 1200 datasets the average time

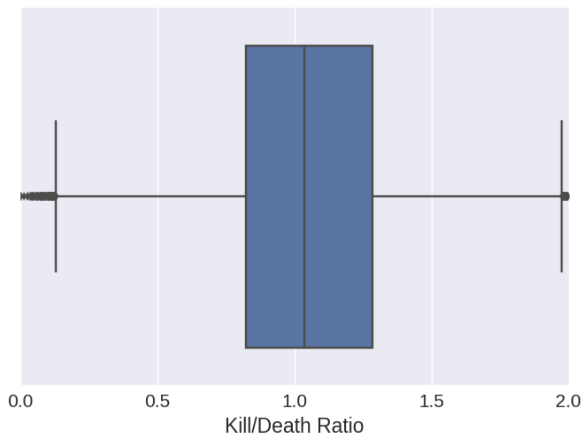


Figure 5.5: Kill/Death Ratio Distribution - threshold: 10 Games

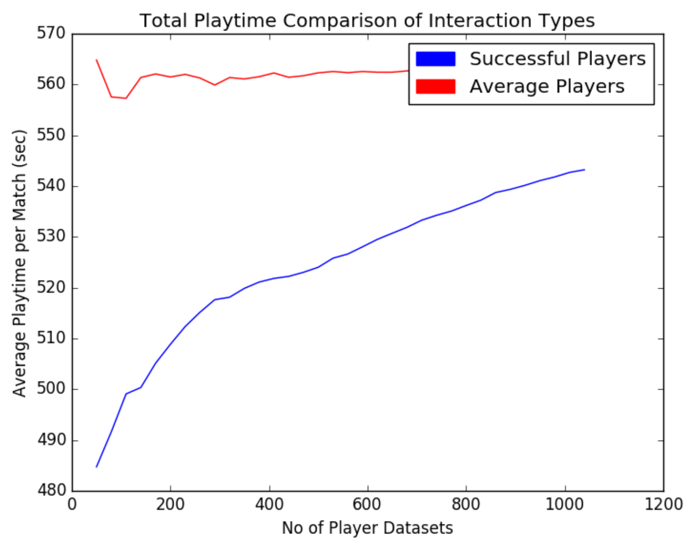
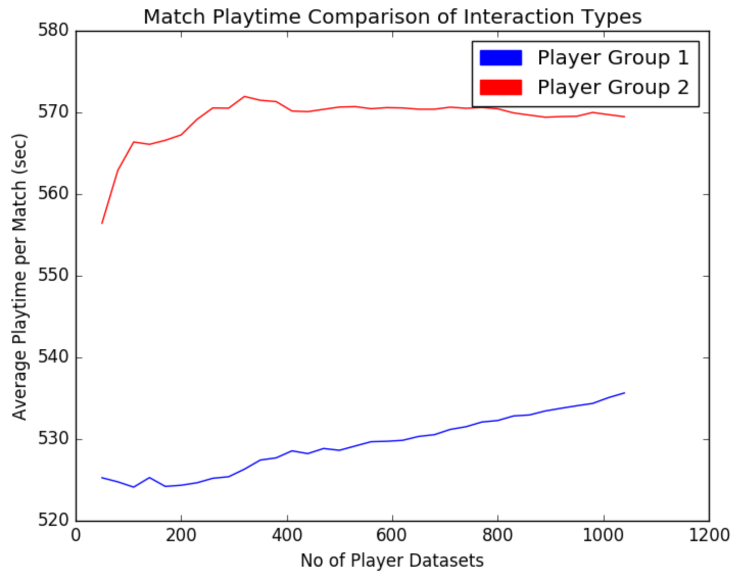


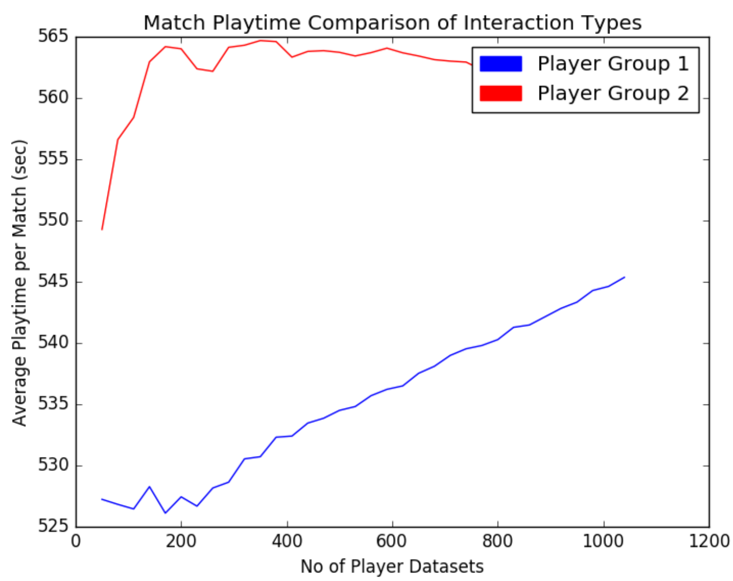
Figure 5.6: Average Playtime in seconds of different Player Groups - threshold: 100 Games (Pirker et al., 2017)

per match spend is almost as long as the time an average player spends. Possible explanations for that are that the successful players achieve their objectives faster than other players, or that they surrender sooner in games where they are likely to loose, maximizing their time spend in matches.

Fig. 5.7 shows the average playtime of the player groups. The average time of a game of *Player Group 2* is close to the length of an average crucible match.



(a) 100



(b) 200

Figure 5.7: Average Playtime in seconds of two different interaction groups with a minimum of a) 100 b) 200 Games played (Pirker et al., 2017)

The highest scoring users in *Player Group 1* have similar average game lengths as the successful players in Fig. 5.6. Compared to the successful player groups the average game time is more consistent up to 1200 players. This might possibly be explained by the fact that users who tend to play with the same players more often also behave similar to these groups (Pirker et al., 2017).

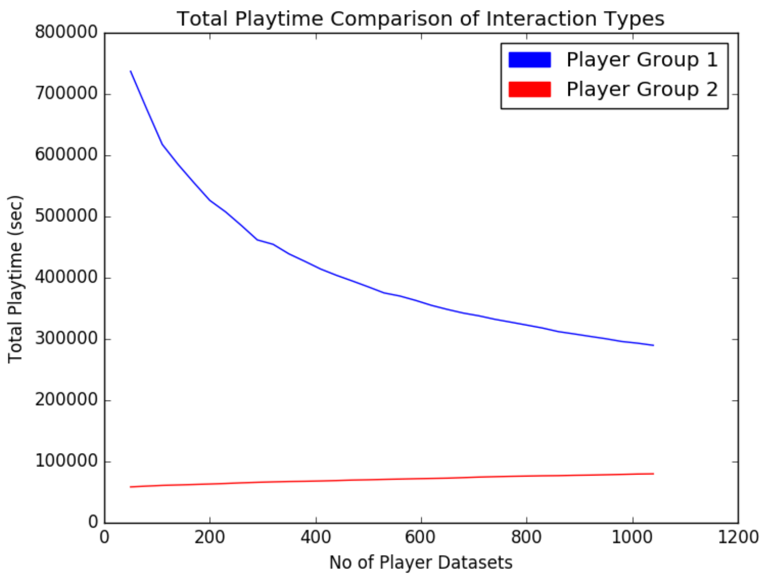
Do relationships in the game have an impact on player engagement?

Fig. 5.8 shows how the total time spend withing games of the different player groups and the average number of matches they participate in. Even when considering the shorter length of games of *Player Group 1*, they still spend more time in total with the game. This is mainly caused by the higher number of games they play in general. Answers from the previous two experiments are probably related to the increased engagement and therefore greater experience *Player Group 1* has with the game. The curves in the figures 5.8a and 5.8b progress in similar ways, hinting that shorter games in *Player Group 1* as shown in figure 5.7 have limited influence on the overall game length.

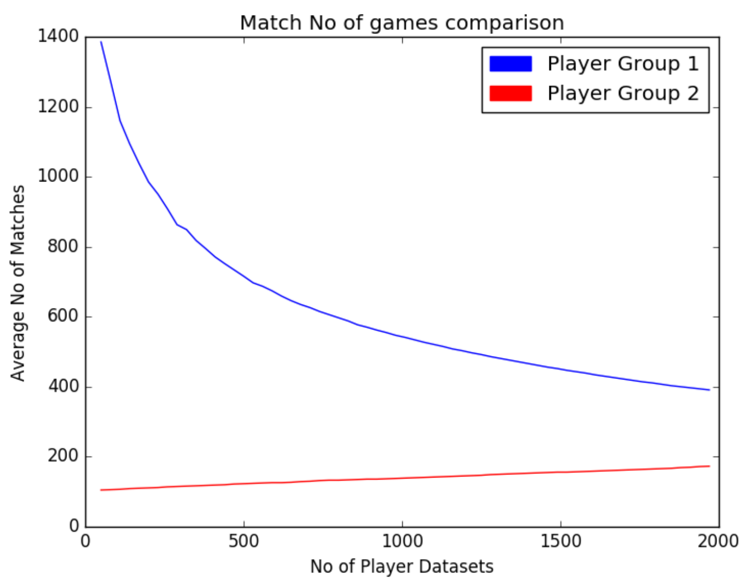
Has clan membership an impact on performance and engagement factors?

To answer this question we look at similar metrics as discussed in the previous section but in relation to clan membership. The win/loss ratio comparison for clan members is shown in Fig. 5.9. The groups displayed here are the top players of the dataset that belong to a clan and the top players without a clan. Similar to *Player Group 1*, clan members exhibit a better win/loss ratio than their counterparts.

Fig. 5.10 visualizes the distribution of the kill/death ratio of clan members. Similar to the win/loss ratio clan member also perform better than player without a clan. This demonstrates similar behavior of clan members to *Player group 1*, and might stem from the similar social nature intricate to both. Fig. 5.11 displays the time clan members spend playing a single game on average. Clan members can end their games a few seconds sooner on average than other players, which might mean that they are able to play more games over the same period of time, hinting that they also gain experience quicker and are generally more successful. Fig. 5.12 shows how many games the top players with a clan played compared to players without a clan. The top 200 players of both groups played almost the exact same amount of games, and does not exhibit the difference in experience *Player Group 1* and *Player Group 2* have (Pirker et al., 2017).



(a) 100



(b) 100

Figure 5.8: Total Playtime Comparison of Interaction Types and Average Number of Matches played (Pirker et al., 2017)

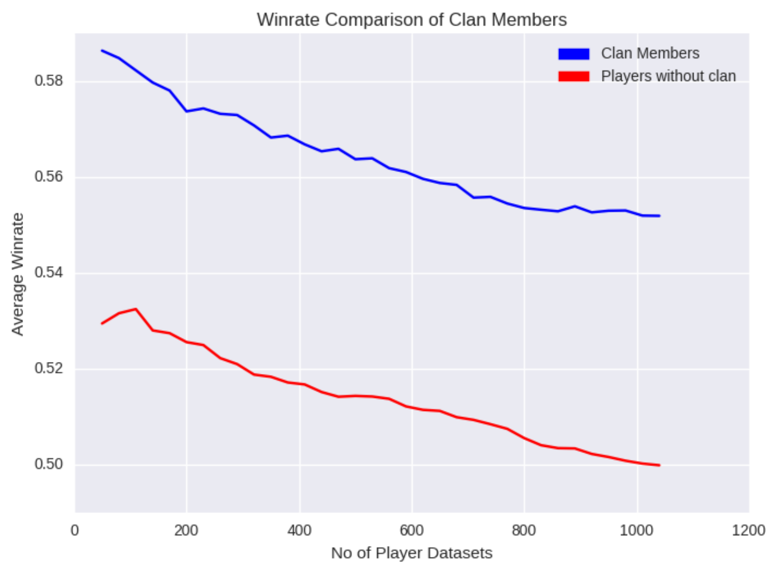


Figure 5.9: Win/Loss Ratio of Clan Members that played at least 100 Games (Pirker et al., 2017)

5.1.2 Behavioural Profiling

Archetype Analysis

Another behavioral experiment that was performed on the *Destiny* dataset is Archetypal Analysis (AA) (Cutler & Breiman, 1994). Similar to clustering methods, AA takes the number of Archetypes as an input, which are used to determine in how many archetypes or clusters the data is split. It was found that 5 Archetypes created a clear differentiation between the inputs seen in Table 5.2. For the operation, 15 distinguished features were used. Eight of those features are weapon related, and the other 7 describe the success and the behavior of the user in the game. The weapon related features were specifically chosen because *Destiny* is a First-Person Shooter, and different playstyles might emerge in the different archetypes.

Fig. 5.13 shows the detailed results of the analysis. The x-axis shows the features, and the importance of the feature for each archetype. Features that score low, have been found to have lesser importance than other features for clustering purposes. The Figure further illustrates how different archetypes use different weapons. Notable is that although archetypes 2 and 4 both prefer mid-range weapons, their weapon choice still varies.

AA is a probability based soft clustering method. This entails that AA,

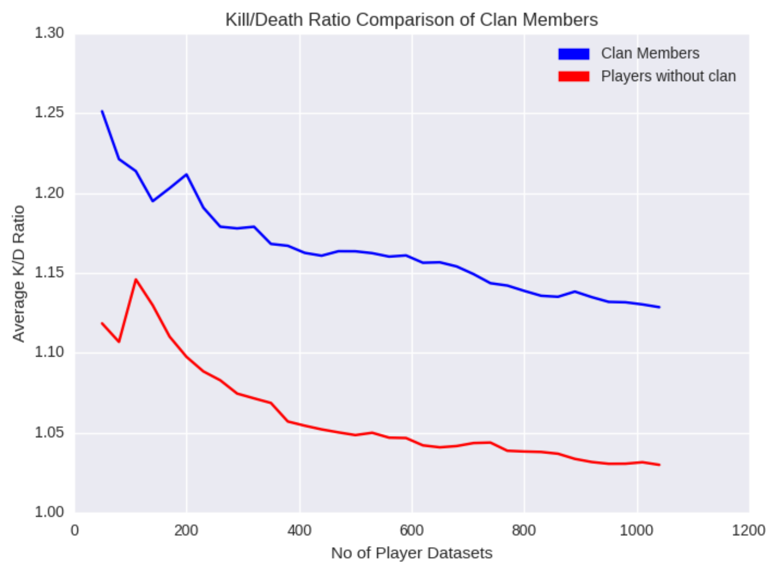


Figure 5.10: Kill/Death Ratio of Clan Members that played at least 100 Games (Pirker et al., 2017)

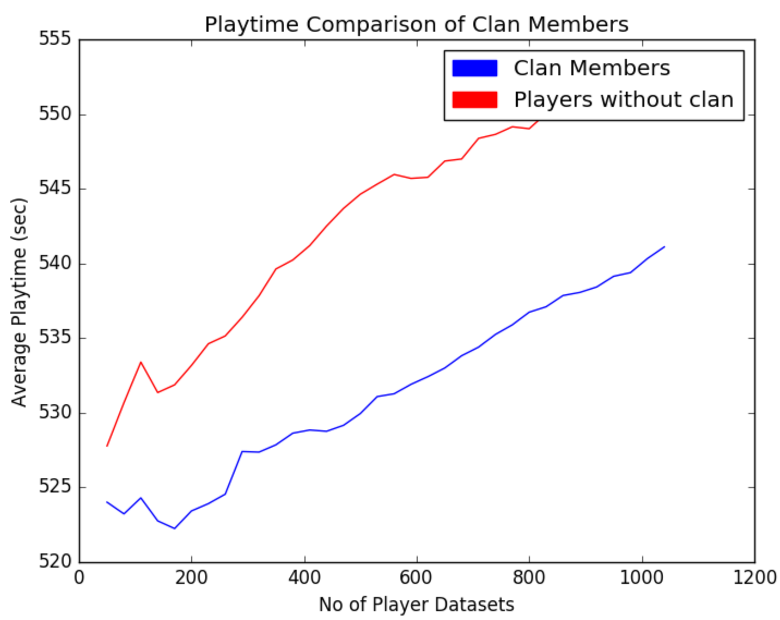


Figure 5.11: Average Time spend in Matches (Pirker et al., 2017)

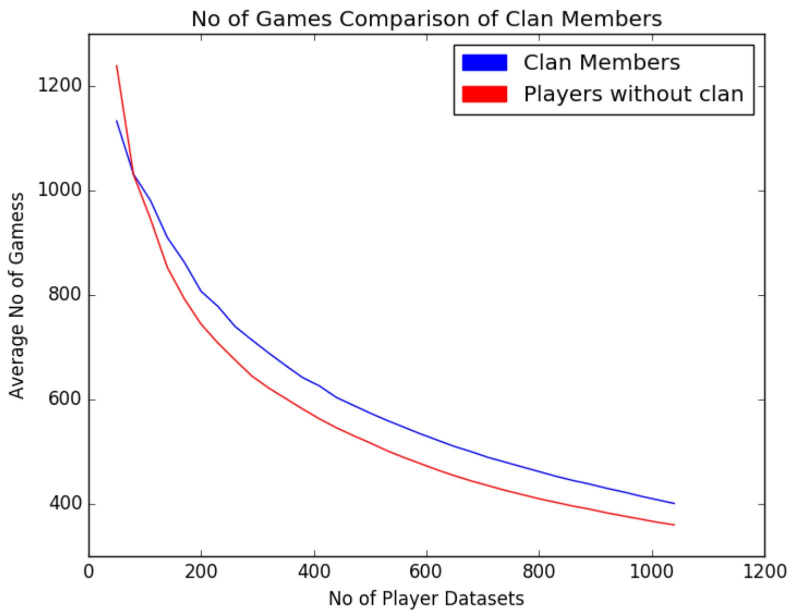


Figure 5.12: Average Number of Matches played by Clan Members (Pirker et al., 2017)

Table 5.2: Overview of the different archetypes (Rattinger et al., 2016)

Archetype	**Description**
Ranged Elites (AT1)	pretty good scores, auto-rifle focus, higher killing spree, unique precision kills, kill/death ratio, and win rate than AT3
Melee (AT2)	melee focused, medium performance, win rate similar to AT4
Mixed Weapon Elites (AT3)	high scores everywhere, more medals than AT1, better weapon scores except for auto-rifle, slightly lower win rate
Short Range (AT4)	medium performance, heavy use of shotgun, some melee
Newbies (AT5)	low performance everywhere

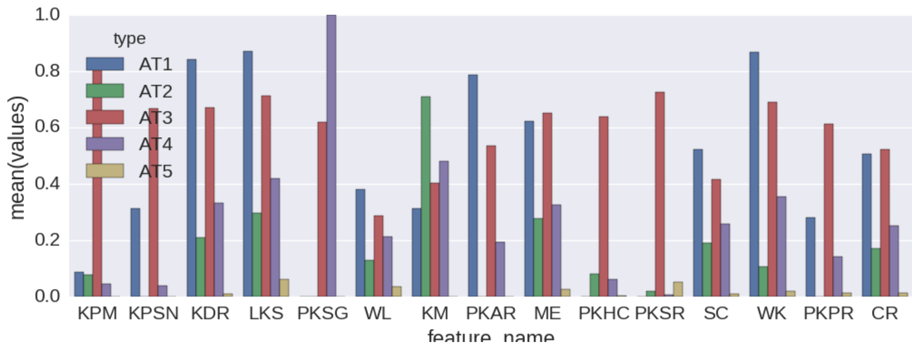


Figure 5.13: Features Importance of the different Archetypes (Rattinger et al., 2016)

KPM = Kills Machinegun, KPSN = Kills Sniper, KDR = Kill/Death Ratio, LKS = Longest Kill Spree, PKSG = Kills Shotgun, WL = Win/Loss Ratio, KM = Kills Melee, PKAR = Kills AutoRifle, ME = Medals earned, PKHC = Kills Handcannon, PKSR = Kills Scoutrifle, SC = Average Score, WK = Unique Weapon Kills, PKPR = Kills Pulserifle, CR = Combat Rating

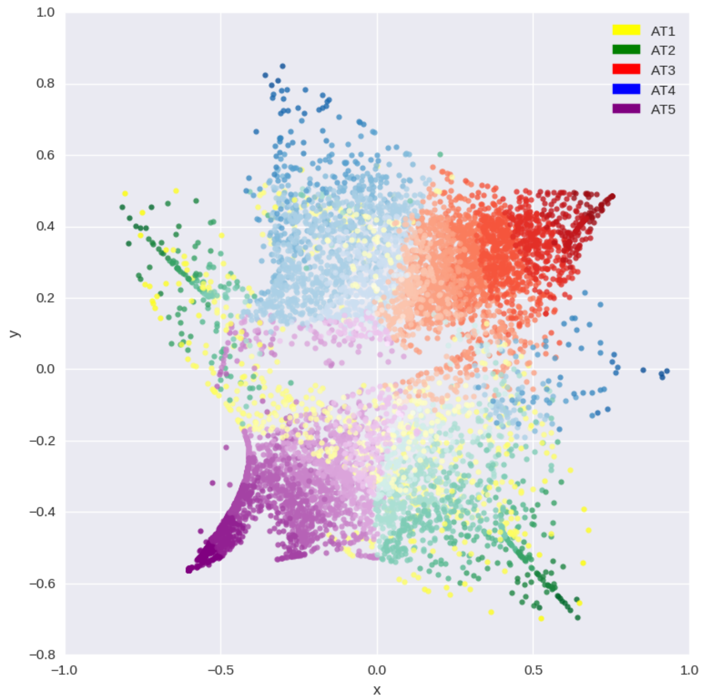
compared to other clustering methods, assigns each player multiple probabilities to determine to which archetype they belong. A player could for example belong to Archetype 1 for 60% and to Archetype 2 for 40%.

Visualization

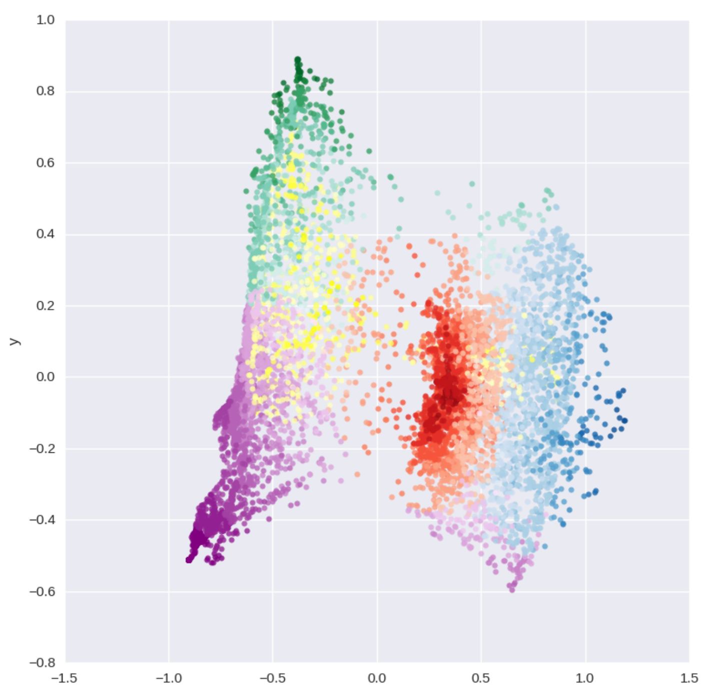
Although the results of the archetypal analysis step give a clearer picture of the overall dataset and the player classes, some aspects of the behavioral pattern can be visualized in a different way. The dataset consists out of features mapping to over a multitude of dimensions, making it hard to get a picture of how the clusters lie in the multidimensional space. Rattinger et al. (2016) showed that the high-dimensional spaces of the *Destiny* results can be visualized in a two dimensional space using multidimensional scaling (Kruskal & Wish, 1978). This is also shown in Fig. 5.14a. Compared to the approach in Rattinger et al. (2016), archetypes are only visualized with the color of their most prominent class, with more saturated colors corresponding to a higher percentage of affiliation. The coloring methods for the data points is also used in all further examples.

Fig. 5.14b shows another popular dimensional reduction method, which provides similar results to multidimensional-scaling, only moving the successful archetype 3 into the middle.

Due to the nature of AA, that creates cluster associations for every point in the data, some information might get lost when using the most prominent class.



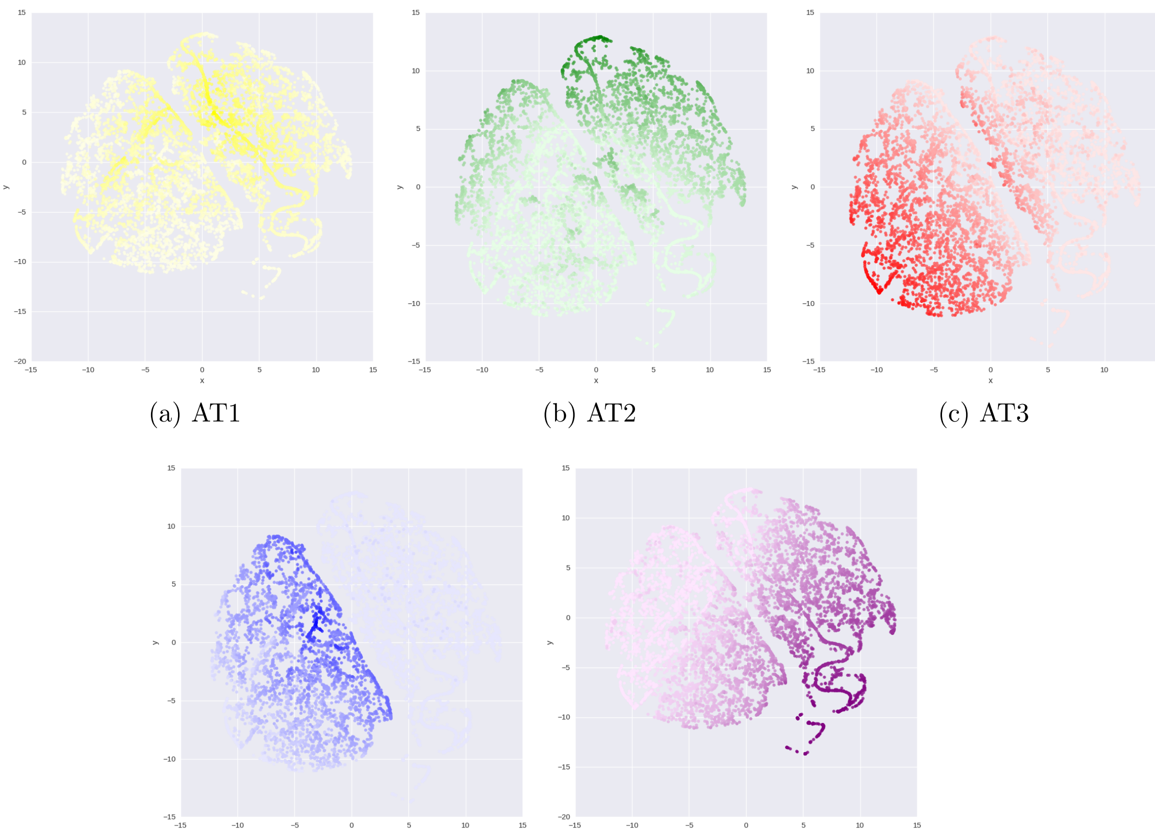
(a) Multidimensional scaling



(b) Isomap

Figure 5.14: Visualization of Archetype Clustering with Multidimensional Scaling and the Isomap Algorithm

Fig. 5.15 visualizes the clusters for every single archetype, showing the strength of the affiliation through color saturation as in the previous examples. This and the following parts of the archetype is done via a technique called t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008). The structure revealed here shows that Archetypes 3 and 4, which are the archetypes with better performance than the others, strongly overlap, but the better performing Archetype 3 further outside. The overlap is consistent with the feature importance shown in Fig. 5.13.

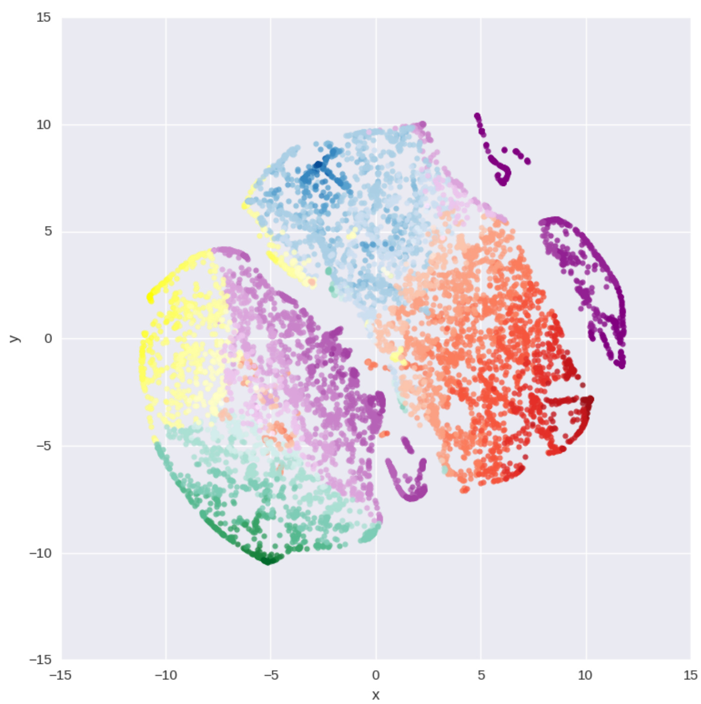


(a) AT1 (b) AT2 (c) AT3

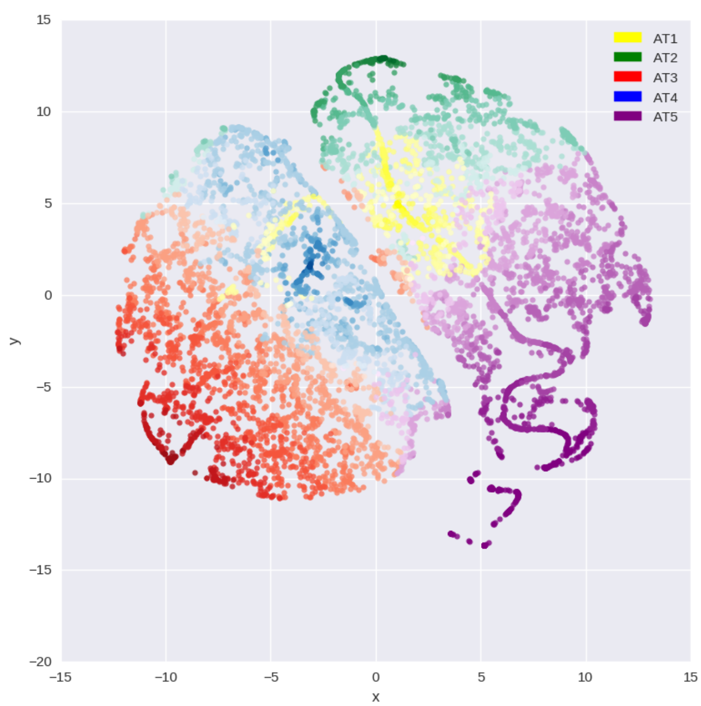
(d) AT4 (e) AT5

Figure 5.15: Single Class Archetype Clusters visualized using t-distributed stochastic neighbor embedding (t-SNE)

Fig. 5.16 shows two variation on t-SNE with different initialization methods. Both versions do not use Barnes-Hut approximation (Barnes & Hut, 1986) run time was not a concern for the relatively dataset, and uses the exact gradient calculation algorithm instead, leading to better results. The main difference between the two Figures is the initialization method. Fig. 5.16a was initialized using the random initialization method, where Fig. 5.16b was initialized using Principal components (PCA). The multidimensional scaling method and t-SNE create similar results, the most successful archetype 3 is the furthest from the worst performer archetype 5. The two methods also demonstrate that archetype 1 usually takes the middle-ground between all other archetypes. This can also be seen when looking at Fig. 5.13.



(a) t-SNE, exact, random initialization



(b) t-SNE, exact, PCA initialization

Figure 5.16: Visualization of Archetype Clustering with t-SNE

5.2 Galileo

The main measures that relate to success in the *Galileo* dataset are the grades the users received upon course completion. A positive completion in the dataset are all total grades that reach more than 50% of the possible score of the course. There are variation on the degree of success with this metric though. A person who finished with a score of 100, did much better than a person who scored the necessary minimum to pass the course. Another important factor in the dataset is how and how often the users access the content, and what kind of content they access. Table 5.3 shows the tool usage of *Galileo* users. Not every tool is equally represented in the dataset, and might not be as important as the other tools for our purposes. Other than the analysis of single network data the analysis also looked at the overall grade distribution and the success of all combined users.

Table 5.3: Overview of the Galileo access methods in the dataset

Tool	**Description**
Assessment	Tests the knowledge or satisfaction of student, consists of online quizzes and surveys
Assignment	Link to the assignments and their descriptions, and all other resources used in the course description (images, videos)
File Storage	Links to all hosted coursework, contains all files used in the course
Forums	Forum entries by students and by the instructors
Learning Content	Content uploaded by the instructors, can include video, audio, mind maps, images and external resources
Evaluation	Used to upload and download tasks in the MOOC
Peer Evaluation	Peer review tool for students, students evaluate each other

5.2.1 Performance & Engagement

The factors that define performance and engagement in *Galileo* are similar to the ones used in the *Destiny* Analysis, and we perform similar experiments on the *Galileo* dataset to examine if parallels in the performance, engagement and user behavior can be drawn, although some information differs. One of the main differences between the datasets is related to engagement: The *Galileo* dataset does not provide session length that are similar to the time of a single game in *Destiny*. Instead we focus the engagement experiments on the tool usage, specifically the amount of tool usage users exhibit. The following experiments are conducted for the *Galileo* dataset and networks:

1. Has the strength of the implicit relationships any impact on course performance?
2. Has the amount of implicit relationships any impact on course performance?
3. Does engagement have an impact on the performance?

Has the strength of the implicit relationships any impact on course performance?

Fig 5.17 shows the overall distribution of points the user reached in the courses. The high amount of people who had overall scores below 20% of the possible maximum course points can be explained by the high churn rate online courses usually display. Notable is also the second peak in the distribution which shows that the overall users of *Galileo* mostly reach the top grades when they complete the course successfully. It also demonstrates that users barely score the full amount of points available in the courses, as the greater than 95% mark sees the fewest users.

Fig. 5.18 shows the final grades of course participants compared to the weight they have with their connections in their respective networks. The weight used here is the sum of all weights a user has in the network. Users who participated in multiple courses would have a higher overall combined weight. There are only 139 out of the total 3157 users in the dataset that participates in more than one course. Due to the small amount of users this applies to, they are excluded from further analysis. Notable is the peak around the 50% mark. This might be explained by people doing what is necessary to barely pass the course, but do not to anything to improve their grades further. Further investigation would need to be conducted to see if this is indeed the case. When compared to the overall grade distribution from Fig. 5.17, it can be observed that they deviate. This implies that users that are more socially active users achieve different and mostly better scores than the average user.

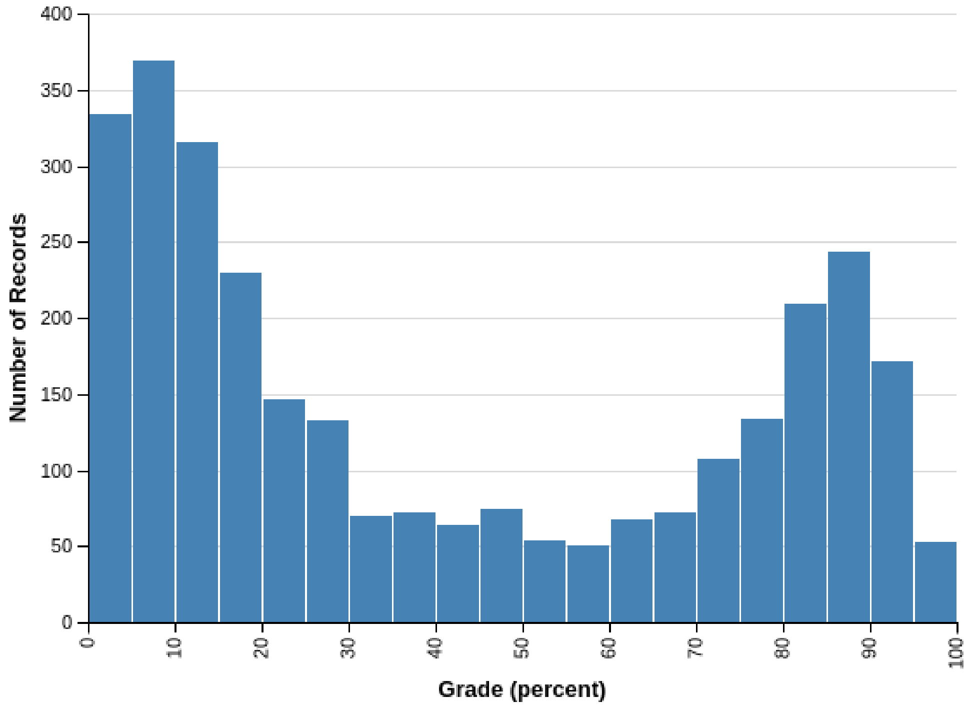


Figure 5.17: MOOC Course Score Distribution

Has the amount of implicit relationships any impact on course performance?

Fig. 5.19 visualizes the network degree compared to the grades users receive. Similar to the grade-weight distribution in Fig. 5.18, this also indicates that users that have a higher degree of social interaction achieve higher scores in the courses. The degrees also have a peak around the 50% mark, but it is not nearly as pronounced as in the weight-distribution example. This might be caused by the fact that this figure treats all relationships similarly and does not consider the strengths of the relationships. The reason behind this is in how we build the networks in the previous chapters. Two users who interact with the system once might have similar overall weights, because weights are calculated with a distance, but they would exhibit different degrees. One possibility to achieve a more expressive result in this figure is applying a threshold of a minimal amount of social interaction similar to the threshold applied to the *Destiny* dataset.

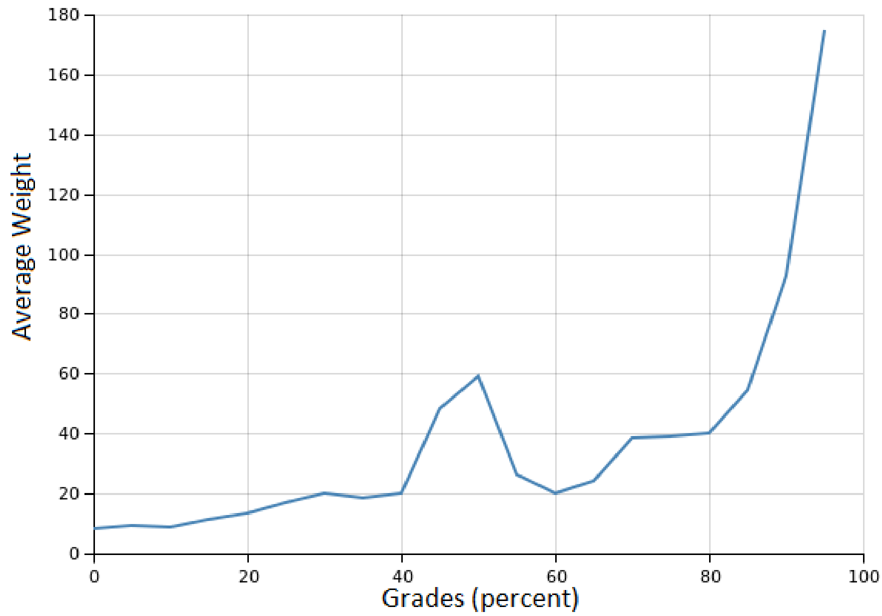


Figure 5.18: Average weights of course participants in different final score groups

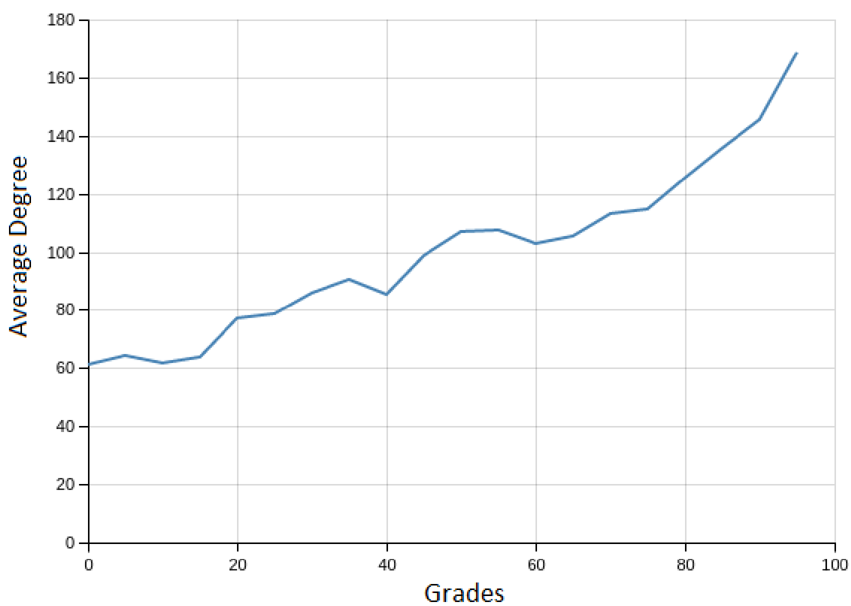


Figure 5.19: Average degrees of course participants in different final score groups

Does engagement have an impact on the performance?

The main way interaction data in the *Galileo* dataset is mapped, is which tools the participants use when accessing the site. The frequency of the tool usage provides an insight on how frequent the users connect to the site, and preform the actions. Fig. 5.21 shows which tools are accessed the most by the users. The two types that are accessed the most are file storage and learning content. All other types are only accessed marginally. As can be seen in Fig. 5.21a, the access of different tools varies quite a bit between single courses, but the two most prominent tools are always learning content and file storage. Due to the more consistent nature of learning content, it is used as the basis for further analysis. Fig. 5.20 visualizes the grades users receive compared to their learning content tool usage. When looking at how users with different grades behave, it can be seen that learning content tool usage is only relevant for the bottom 20% of the grades. This might be related to the high dropout rate shown in 5.17, which also exhibits a small peak around 90%. The same conclusion can be drawn from Fig. 5.22. Both the file storage tool and the learning content tool fulfill similar tasks in *Galileo*. Another tool that is related to the rest of the experiments and might give more insight is the forum tool. The relation of the forum tool to the grades is visualized in Fig. 5.23. The forum tool exhibits similar properties to previous experiments, because the social network properties like weights and degree are dependent on it.

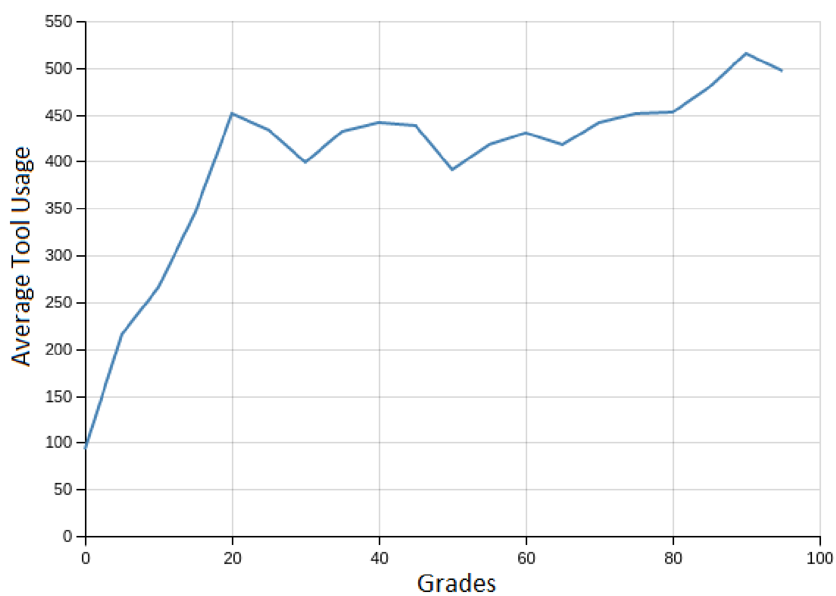
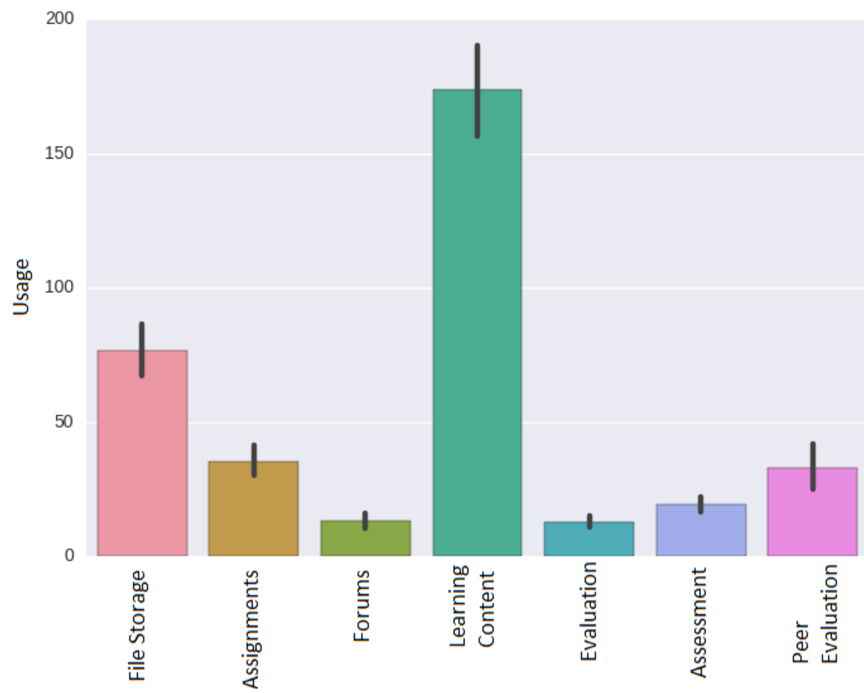
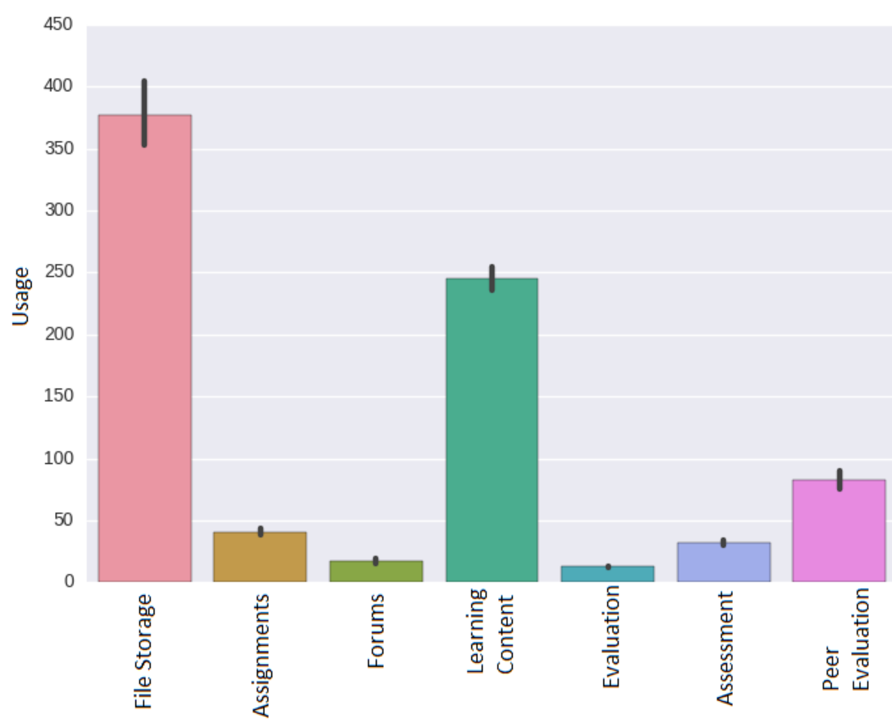


Figure 5.20: Tool Grade Distribution: Learning Content



(a) Android Tool Usage



(b) Tool Usage of all MOOCs

Figure 5.21: Visualization of Tool Usage

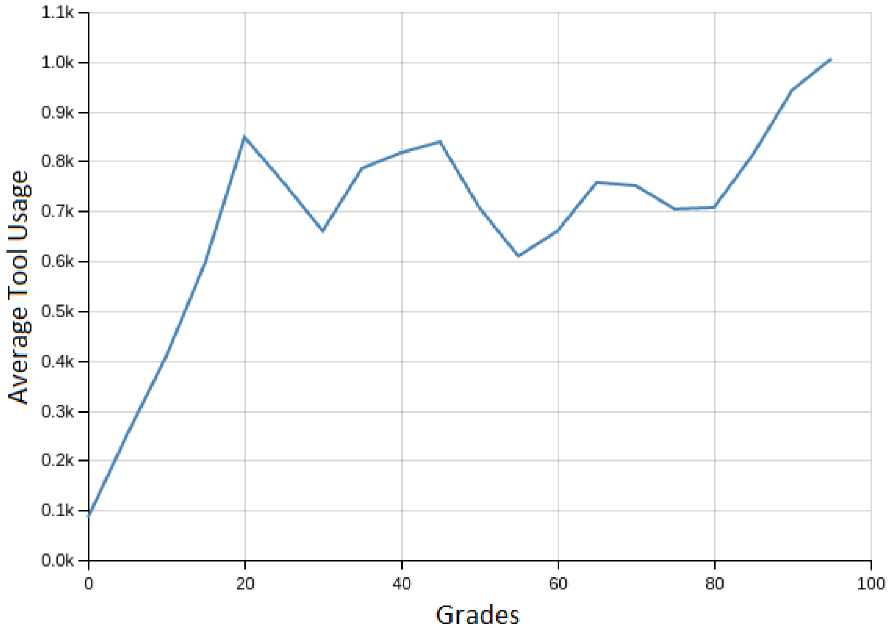


Figure 5.22: Tool Grade Distribution: File Storage

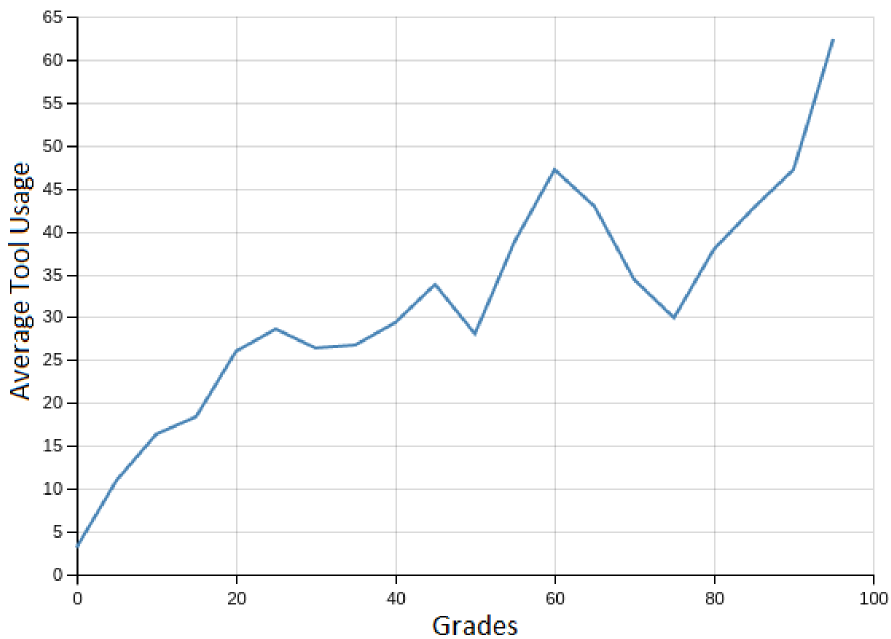


Figure 5.23: Tool Grade Distribution: Forum

5.3 Summary

To get insight about the performance, engagement and behavioral aspects of the datasets, several experiments were performed on each of them. Both datasets include metrics that can give a detailed look at those factors. Visualizing the impact of relationships in *Destiny* is done by splitting the results in different groups. The most important groups in terms of the analysis are players who play a lot with the same players and players who use the match-making to play with random people. The first group is used to give a insight on how players performance changes when they have more distinct relationships. Important performance metrics like win/loss ratio or kill/death ratio are significantly better for players who belong to the first group. Another group we take a look at are successful player independently of their relationship. Successful Players are determined by their win/loss ratio and generally have shorter games than other players. The first player group exhibits many traits that are similar or close to the successful players. The last experiment performed for the *Destiny* dataset is related to clan membership. Clan members display improved performance and engagement compared to other player groups. To gain deeper insights into the dataset, another behavioral analysis is conducted. With the unsupervised soft-clustering method archetypal analysis, 15 weapon and performance related features are chosen and 5 player archetypes are observed. The archetypes vary between their success and weapon choices. For further visualization multiple dimensionality reduction techniques and algorithms are performed, mapping the archetypes to 2 dimensions easier interpretation.

Experiments for the *Galileo* dataset are mostly related to the user performance, where the grades users receive is chosen as the main metric. The *Galileo* dataset exhibits two peaks in the grade distribution. The first one is below the 20% mark, and the second one is above 80%. When observing the social interaction, which is defined as the sum of all weights in the graph, the peaks shift. The first peak can be seen just after the 50% mark, which indicates a drop of social interaction after a user reached the minimal assignment goal. The degree of the nodes is also used in the same capacity, but it is shown that the degree does not provide better insights than the sum of weights as metric. The most used tools in the dataset are the file storage and the learning content tools. The amount of use of the specific tools only increases slightly after a certain grade threshold was reached. This might be caused by the high early dropout rate the MOOC experiences.

Chapter 6

Summary and Outlook

The work presented here discusses an approach to create social network structures out of inferred implicit relationships between members of online systems. Implicit relationships are different from traditional explicit relationships as they reveal more relevant social structures, but it is harder to find a metric that maps relationships in the right way. Compared to their counterpart, they usually correspond to underlying network in a more natural way, caused by the direct actions users have to take to be associated instead of an action only taken once. Furthermore it is also discussed how information about user performance and engagement is extracted, as well as how behavioral features are inferred from network structure. As an important first step in this work, both datasets and their particularities are explored at the beginning. This reveals important aspects for the analysis steps following the initial assessment.

6.1 Conclusion

Destinys data exploration step revealed that most of the players participated in only a small amount of games, that a great part of the players reached the maximum level in the game and that some in-game characters are preferred over others. To create the *Destiny* networks information from 930,720 games is used containing information from over 3.5 million users. These users are the source of the nodes in the graph, and the matches they play together in is used to define their relationship. The three networks that were build: 1) Team: Players who play together 2) Adversarial: Players who play against each other 3) Match: Players who participate in the same game. Through the analysis of network measures it is observed that the adversarial network is not a very significant network for the analysis due to the mostly random nature that is caused by the random match-making. The data analysis conducted a number of experiments to show if social factors influence game performance and engagement. It was found that players interacting with the same group of people compared to players who play with random players scored higher in all performance and engagement related metrics, and are generally more efficient when playing. Those results also were true for members of clans.

Galileos data exploration step revealed the high drop-out rate of the MOOC Dataset, and that the amount of social interaction is largely dependent on the course. The course with the highest drop-out rate is "Medical Emergencies" with 89.71% of participants not finishing the course with success. A network was build for each of the courses, resulting in 11 networks. The participants are the nodes of the network, and vary between 169 and 918 users, which is really small compared to *Destiny*. All of the resulting networks are viable for further analysis, and they are in fact oftentimes combined to create results that are valid throughout the courses. Participants who are more active in their online relations are found to be more successful in the coursed. Success in the *Galileo* dataset is defined by the grades the users reach. Users seem also to be more active around the 50% mark of grades, and drop their social activity once they pass that threshold. The reasoning and implication of those actions in context of forum interaction would need further study. After the 80% mark a sharp rise in social activity was observed, hinting that the top performers are much more likely to use the tools provided.

To explore the behavioral features of *Destiny* further, a technique called "Archetypal Analysis" was employed. With this technique 5 profiles were found that correspond to the users behavior and success in the game. Compared to other clustering methods, each of the users is assigned a probability for belonging to a specific archetype. The 15 features chosen for the analysis were all either based on performance or the users weapon choices. The analysis showed that the users weapon choices have influence on the success related parameters. To complement the archetypal analysis results, a few different visualization techniques were used. The main problem in visualizing the archetypes is that the user data is multidimensional, and needs to be transformed to a 2 or 3 dimensional space for visualization. The main way this was done is by using a machine learning algorithm for probability-based dimensionality reduction called t-SNE. The visualization of the archetypes with the algorithm create a good overview of their relation to each other, and it was found that the positioning corresponds to their performance in the game. Overall it was found that users that engage in more social interaction are more successful in the two systems analyzed. Users who display more interactive behavior or have stronger relationships generally outperform users that do not engage socially. One of the problems that arises with certain implicit connections is that not all social behavior might be captured: In the *Galileo* dataset, forum discussions are the basis for the edges in the network. Socially active users that used other means like e-mail communication are not captured by this approach. Engagement metrics depict a similar picture within the analysis. Users who have stronger ties are much more likely to use the system frequently and for longer periods. This also applies in the other direction: Users who display higher and recurrent usage of the tools provided to them are more likely to have stronger ties. The same behavior can be observed for *Destiny* clan members. Group membership improves the performance and engagement of those users. The presented framework builds a basis for further analysis of other datasets, and how their metrics change in the context of their relationships.

6.2 Suggestions for Future Work

As mentioned in the previous chapters, there are some areas in this work that could be explored in more detail. When building upon the work presented in this thesis, one valuable approach would be to create a new network and analyze a dataset from a different domain. This would improve comparability and would show the viability of the approach outside of games and MOOCs, which both have naturally build-in social mechanisms. Another approach would be to use community detection algorithms to identify more of the underlying structure. Behavioral and success factors could then be analyzed as well, which possibly would make them more comparable to the smaller networks created in this work. In addition some communities possibly exhibit different properties than others, and community detection algorithms might identify communities that are close to *Destiny* clans or find learning groups in *Galileo*.

Some of the social structure this work maps could be improved upon. The *Galileo* analysis uses forum interaction as its base, but it is possible that more of the underlying social structure could be uncovered. One important step in retrieving more of the inherit implicit relationships would be to gather enhanced data sources, which contain more data that is easier to identify as a social connection. Especially in *Galileo* an approach to gather this information might be to take a more detailed look at information on how and when users access the online tools provided to them. Similarly *Destiny* might have more useful social structure one can infer, with usage of certain communication tools being one of them. Clan membership could also be used as a guide to help weighing the strength of those connections. The analysis features a few of the most important network measures, but could be improved by using more complex network measures beyond that. This also might add to comparability between the created networks. Another approach to the data analysis would be to look at how the networks change over time. This could give an insight on how and why the connections are formed in the system. It is likely that some of the connections in *Galileo* are formed right before an assignment is due, or players get to know each other if a special event takes place in *Destiny*. An area that could be explored further is network visualization. Some work was done to visualize subsets of the networks, but visualizing networks considering aspects as user performance or behavior might lead to new insights. The following sections discuss a few suggestions for future work that applies to the analyzed dataset in particular.

6.2.1 Destiny

In addition to the implicit relationships, *Destiny* also features traditional friendships. The friendship relationships are available in online tracking systems, and would need to be extracted using web crawling techniques. This data could than be compared to the edges created through the implicit relationships to analyze how the playing behavior differs from traditional friendships. The *Destiny* analysis could also be improved through gathering more data from the games. As mentioned in Chapter 3, the gathered data is a subset of the actual played games. More structure could be revealed if the dataset was more complete. This would also improve community detection efforts.

6.2.2 Galileo

The *Galileo* analysis features figures that are related to the grade of its participants, and the distribution contains two peaks. It was reasoned in Chapter 5 that users might drop their interaction after they reached the passing grade, but the data could need further inspection to confirm this. A part of the analysis that was not performed for the *Galileo* dataset yet is archetypal analysis. The main reason for this stems from the size of the dataset. When a bigger dataset is gathered or more data for other courses is available, it might be worthwhile to create learning archetypes.

Appendix A

List of Abbreviations

AA	Archetypal Analysis
ACC	Average Clustering Coefficient
BA	Behavioral Analysis
KDR	Kill/Death Ratio
LCC	Largest Connected Component
MDS	Multidimensional Scaling
MMORPG	Massively Multiplayer Online Role-Playing Game
MOG	Multiplayer Online Game
MOOC	Massive Open Online Course
PCA	Principal Component Analysis
RPG	Role-Playing Game
SNA	Social Network Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
WR	Winrate or Win/Loss Ratio

Bibliography

Anderson, A., Huttenlocher, D., Kleinberg, J. & Leskovec, J. (2014). Engaging with massive online courses. In *Proceedings of the 23rd international conference on world wide web* (S. 687–698).

Barnes, J. & Hut, P. (1986). A hierarchical o (n log n) force-calculation algorithm. *nature*, *324* (6096), 446–449.

Bauckhage, C., Drachen, A. & Sifa, R. (2015). Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games*, *7* (3), 266–278. doi: 10.1109/TCIAIG.2014.2376982

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D. & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edx's first mooc. *Research & Practice in Assessment*, *8*.

Brown, J., J., B. A. & N., L. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing*, *21* (3), 2–20.

Clow, D. (2013). Moocs and the funnel of participation. In *Proceedings of the third international conference on learning analytics and knowledge* (S. 185–189).

Cutler, A. & Breiman, L. (1994). Archetypal analysis. *Technometrics*, *36* (4), 338–347.

Destiny overview. (2016). `https://www.destinythegame.com/game`. (Accessed: 2017-01-22)

Deterding, S., Sicart, M., Nacke, L., O'Hara, K. & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In *Chi'11 extended abstracts on human factors in computing systems* (S. 2425–2428).

Di Battista, G., Eades, P., Tamassia, R. & Tollis, I. G. (1994). Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry*, *4* (5), 235–282.

Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development*, *53* (2), 67–83.

Drachen, A., Sifa, R., Bauckhage, C. & Thurau, C. (2012). Guns, swords and data: Clustering of player behavior in computer games in the wild. In *2012 ieee conference on computational intelligence and games (cig)* (S. 163–170).

Drachen, A., Thurau, C., Togelius, J., Yannakakis, G. N. & Bauckhage, C. (2013). Game data mining. In *Game analytics* (S. 205–253). Springer.

Ducheneaut, N., Yee, N., Nickell, E. & Moore, R. J. (2006). Alone together?: exploring the social dynamics of massively multiplayer online games. In *Proceedings of the sigchi conference on human factors in computing systems* (S. 407–416).

El-Nasr, M. S., Drachen, A. & Canossa, A. (2013). *Game analytics: Maximizing the value of player data.* Springer Science & Business Media.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486* (3), 75–174.

Goggins, S. P., Galyen, K. & Laffey, J. (2010). Network analysis of trace data for the support of group work: activity patterns in a completely online course. In *Proceedings of the 16th acm international conference on supporting group work* (S. 107–116).

Haythornthwaite, C. (2005). Social network methods and measures for examining e-learning. *Social networks*, *2005*, 1–22.

Huberman, B. A., Romero, D. M. & Wu, F. (2008). Social networks that matter: Twitter under the microscope.

Jia, A. L., Shen, S., Bovenkamp, R. V. D., Iosup, A., Kuipers, F. & Epema, D. H. J. (2015). Socializing by gaming: Revealing social relationships in multiplayer online games. *ACM Trans. Knowl. Discov. Data*, *10* (2), 11:1–11:29. doi: 10.1145/2736698

Kizilcec, R. F., Piech, C. & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (S. 170–179).

Kop, R. & Hill, A. (2008). Connectivism: Learning theory of the future or vestige of the past? *The International Review of Research in Open and Distributed Learning*, *9* (3).

Kruskal, J. B. & Wish, M. (1978). *Multidimensional scaling* (Bd. 11). Sage.

Lim, C.-U. & Harrell, D. F. (2015). Revealing social identity phenomena in videogames with archetypal analysis. In *Proc. aisb symp. on ai and games.*

Lin, N. (2002). *Social capital: A theory of social structure and action* (Bd. 19). Cambridge university press.

Maaten, L. v. d. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9* (Nov), 2579–2605.

Nawrot, I. & Doucet, A. (2014). Building engagement for mooc students: introducing support for time management on online learning platforms. In *Proceedings of the 23rd international conference on world wide web* (S. 1077–1082).

Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, *28* (6), 441–453.

Phan, T., McNeil, S. G. & Robin, B. R. (2016). Students patterns of engagement and course performance in a massive open online course. *Computers & Education*, *95*, 36–44.

Pirker, J., Rattinger, A., Drachen, A. & Sifa, R. (2017). Analyzing player networks in destiny (under review). *Entertainment Computing*.

Rattinger, A., Wallner, G., Drachen, A., Pirker, J. & Sifa, R. (2016). Integrating and inspecting combined behavioral profiling and social network models in destiny. In *International conference on entertainment computing* (S. 77–89).

Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P. & Sherer, J. (2014). Social factors that contribute to attrition in moocs. In *Proceedings of the first acm conference on learning@ scale conference* (S. 197–198).

Scott, J. (2012). *Social network analysis.* Sage.

Shin, D.-H. & Shin, Y.-J. (2011). Why do people play social network games? *Computers in Human Behavior*, *27* (2), 852–861.

Sifa, R. & Bauckhage, C. (2013). Archetypical motion: Supervised game behavior learning with archetypal analysis. In *Ieee conference on computational intelligence in games* (S. 1–8). doi: 10.1109/CIG.2013.6633609

Sifa, R., Bauckhage, C. & Drachen, A. (2014). Archetypal game recommender systems. In *Proceedings of the 16th LWA workshops: Kdml, IR and fgwm* (S. 45–56).

Sifa, R., Drachen, A. & Bauckhage, C. (2015). Large-scale cross-game player behavior analysis on steam. In *Proc. aiide* (S. 198–204).

Sifa, R., Hadiji, F., Runge, J., Drachen, A., Kersting, K. & Bauckhage, C. (2015). Predicting purchase decisions in mobile free-to-play games. In *Eleventh artificial intelligence and interactive digital entertainment conference.*

Simon, S. & Apt, K. R. (2015). Social network games. *Journal of Logic and Computation*, *25* (1), 207–242.

Sinha, T. (2014). Supporting mooc instruction with social network analysis. *arXiv preprint arXiv:1401.5175*.

Stafford, G., Luong, H., Gauch, J., Gauch, S. & Eno, J. (2012). Social network analysis of virtual worlds. In *International conference on active media technology* (S. 411–422).

Thurau, C. & Bauckhage, C. (2010). Analyzing the evolution of social groups in world of warcraft®. In *Proceedings of the 2010 ieee conference on computational intelligence and games* (S. 170–177).

van de Bovenkamp, R., Shen, S., Jia, A. L., Kuipers, F. et al. (2014). Analyzing implicit social networks in multiplayer online games. *IEEE Internet Computing*, *18* (3), 36–44.

Vitiello, M., WALK, S., HERNÁNDEZ, R., HELIC, D. & GÜTL, C. (2016). Classifying students to improve mooc dropout rates. *Research Track*, 501.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. & Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the 4th acm european conference on computer systems* (S. 205–218).

Wohn, D. Y., Lee, Y.-H., J., S. & Bjornrud, T. (2010). Building common ground and reciprocity through social network games. In *Proc. chi extended abstract on human factors in computing systems* (S. 4423–4428). ACM publishers.

Xiang, R., Neville, J. & Rogati, M. (2010). Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on world wide web* (S. 981–990).

Yee, N. (2007). Motivations for play in online games. *CyberPsychology & Behavior*, *9* (6), 772–775. doi: 10.1089/cpb.2006.9.772

Zafarani, R., Abbasi, M. A. & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

List of Figures

List of Tables