DESIGN THE CHURN ANALYSIS ON GAMES

A REVIEW ON TECHNIQUES FOR CHURN ANALYSIS

Thesis Presented

by

Qi Wang

to

The Department of Game Design

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Game Science and Design

Northeastern University

Boston, Massachusetts

May 2018
DESIGN THE CHURN ANALYSIS ON GAMES

A REVIEW ON TECHNIQUES FOR CHURN ANALYSIS

by

Qi Wang

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Game Science and Design
in the Graduate School of the College of Arts, Media and Design of
Northeastern University

<May>, <2018>
TABLE OF CONTENTS

Introduction 7

Background 9

Meta-Analysis Review 12

A. Studies in Game Industry 12
B. Studies in Telecommunication Industry 20

Further Analysis of Essential Components of Churn 27

C. Churn definition 27
D. Feature selection and churn extraction 31
E. Statistical Methodologies 35

Churn Analysis Design 39

F. Preliminary Design 40
G. Revised design 43

Analysis Implementation 48

H. Brief Introduction of Wuzzit Trouble 48
I. Churn definition 49
J. Data formatting 50
K. Collinearity Test 51
L. Discriminant Analysis 51
Abstract

This is a study of churn analysis design based on the meta-analysis method used in previous related research in game and telecommunication industries. The aim is to facilitate the general use of churn analysis in game development.

This study focuses on the statistical modeling and machine learning techniques used to analyze player churn by researchers in game and telecom studies over the past ten years. This study uses the meta-analysis methodology for the literature review. It will provide recommendations as to what techniques are suitable to perform churn analysis. The final section applies Network Analysis and Belief Bayesian Network to data from the game, Wuzzit Trouble, in order to prove that these methods, adopted from the telecom industry, can be used to perform churn analysis in the game industry.

The paper consists of five significant parts. 1) Reviewing the chosen studies to summarizing the most popular techniques in churn analysis to deal with the specific types of datasets or research problems. 2) Analyzing the similarities and differences between data manipulation techniques in churn analysis used in game industry and the other industries. 3) Picking the proper data manipulation techniques for churn analysis in educational games by meta-analyzing the selected studies. 4) Providing a full research plan including research questions, research approaches and research processes for Churn analysis of game such as the Wuzzit Trouble. 5) An implementation on the sample data from the game Wuzzit Trouble.

The investigation in this paper addresses the churn definition, feature selection, data processing statistical models and the machine learning techniques, especially statistical graphics model, to present the analysis on churn problem as well as the possible relationship between each variable. The contribution of the paper provides an analytical research design of the churn
analysis graphically, which has never been utilized in churn behavior analysis in the game industry previously. The implementation of the techniques mentioned in the research design, the result of the paper, proved its applicability by taking the Wuzzit Trouble data as the sample. The methodologies include Discriminant Analysis, Collinearity Test, Network Analysis on features and Bayesian Belief Network visualization on churn probability.

*Keywords:* churn, Wuzzit Trouble, Education game, data processing, Network Analysis, BBN, Statistical Graphic Model
Introduction

This study is a meta-analysis of churn analysis in the game research. The paper will focus on the statistical models and the machine learning techniques that utilized to analyze players’ churn in the chosen articles within the recent ten years on the game studies. The analytical results of the meta-analysis are expected to help to design the user experience analysis of an educational game named Wuzzit Trouble and to recommend the techniques that suitable to do a churn analysis based on this game.

According to previous studies, churning action is the process where a customer ends his or her subscription from the existing provider and subscribes to another one (Kisioglu, 2011). Wikipedia defines “Churn” as the number of customers or subscribers who cut ties with the service or company during a given period (Wikipedia, 2017). Concluded from varies descriptions and applied in our case, the “Churn” action is the progress of a player who begins to play the game until the moment that he or she quits within a specific period. “Churn” Analysis is an idea that provides the numerical result to predict the players churning, which could be a resource to recommend the resolution to prevent it from happening.

Predicting and improving player retention has become a crucial way to measure how much and how successful a game is. Previous studies have targeted game analysis or player behavior on a larger scale by implementing varieties of machine learning-based models (El-Nasr et al. 2013). Most of the churn analysis studies of games focused on a week worth of data derived from the players, however, churn action is possible to begin as soon as the player starts the game (Anders Drachen et al. 2016). It is crucial to perform a meta-analysis on the techniques that are utilized in previous studies from diverse fields to nail down the most suitable methodology for an educational
game churn analysis without being distracted by the noise of individual error within individual studies.

To hypothetically design the churn analysis on the Wuzzit Trouble game, this study will recommend machine learning techniques addressing the game features based on a small sample of data from the Wuzzit Trouble. The initial variables are Points earned, level reset times, the number of bonus item accumulated, stars earned, moves count, and so forth. This theoretical research aims to provide a methodology that can help identify which features stating above will be the most critical variables that affect the player's behaviors. This design is expected to contribute to the future user experience studies on how to retain or prevent players from leaving an educational game such as Wuzzit Trouble.

The research questions are as follows: 1) What are the most popular techniques to address churn problem? 2) Which are the most suitable techniques to perform churn analysis on the Wuzzit Trouble game? Moreover, the study will try to provide a research design of churn analysis on Wuzzit Trouble.
Background

Educational games have become more popular for both instructional methodology as well as entertainment. With numerous games with pure entertainment purpose releasing every year, educational games are standing out to learning age students due to the advantage of providing an opportunity for the user to gain knowledge while entertaining. Educational game plays a significant part in overall learning process, it increases both inside and outside of institutional educational settings, which attribute to the growth in consumption of educational game. Player retention is essential for extending the life cycle of a service or a game since retaining existing users is a more cost effective in comparison to marketing new players. In essence, the study can help us to unveil the regularity and causality behind the numbers are critical in business development. However, the difficulties appear on educational games vary from the regular entertainment games due to educational games are less appealing to the age group that is outside the learning age and the purpose of learning. The churn analysis on the educational game based on the studies of the ordinary games and the other virtual economies but more specifically on the time range, target population and so on.

Previous studies on churn prediction in the game and telecommunication industries mainly applied data mining techniques such as neural networks, decision trees, and cluster analysis to predict churn rate. However, to our knowledge, there has been no research on an educational game using Bayesian Belief Network (BNN) to identify the factors that have effects on users to churn.

1) In a churn analysis of the Turkish Telecommunication (Pınar Kısıoglu, Y. İlker Topcu, 2011), the author used Bayesian Belief Network to generate the graphics model to dig the pattern how a customer terminates his or her subscription from the existing provider and go for another service provider. This study provides a fundamental academic reference for BNN churn analysis. 2) In
another telecommunication field study (Saravanan M., Vijay Raajaa G.S., 2012), churn analysis was taken as a social phenomenon instead of a purely mathematical problem by constructing a graph parameter analysis to the churner behavior. 3) A study from within the game industry that focused on Free-to-Play mobile games (Anders Drachen, Eric Thurston Lundquist, Yungjen Kung, Pranav Simha Rao, Diego Klabjan, Rafet Sifa, Julian Runge, 2016) used heuristic modeling approaches to build simple rules for predicting short-term retention model. Their heuristic-based approach achieves reasonable and comparable performance using information from the first session, day, and week of player activity. There is no study that has implemented BNN causality analysis to educational games to perform player’s churn analysis thus far.

Compared to the research of the game for entertainment that is designed for all age groups, research in educational game has more difficulties in data collecting and modeling due to the limit that the most players are pre-high school. Holding the educational purpose, the churn analysis in this study differs from case studies which targeted players are mostly adults. Additionally, the collected data is not assigned explicitly to individuals due to the state Law of Massachusetts. Thus, the pre-manipulation of the data that assuming logins with same IP addresses and similar game version/device type belongs to one player. These conditions that pre-addressed will reflect the time ranges that this study is going to set will explain the unmanaged noise from the analysis.

To discover the best machine learning techniques for the churn analysis of the educational game, I plan to address the research question by reviewing previous articles for churn analysis. The paper consists of five significant parts. 1) Reviewing the chosen studies to summarizing the most popular techniques in churn analysis to deal with the specific types of datasets or research problems. 2) The similarities and differences between data manipulation techniques in churn analysis used in game industry and the other industries. 3) Picking the proper data manipulation
techniques for churn analysis in educational games by meta-analyzing the selected studies. 4) Providing a full research plan including research questions, research approaches and research processes for Churn analysis of game such as the Wuzzit Trouble. 5) An implementation on the sample data from the game Wuzzit Trouble.

Meta-analysis is a popular and frequently used statistical technique used to combine data from several studies and reexamine the effectiveness of treatment interventions (Israel H, Richter RR, 2018). Wikipedia defines meta-analysis as a research method that combines the results of multiple scientific studies (Wikipedia, Meta-analysis, 2018). The key benefit of this approach is the aggregation of information leading to higher statistical power and more robust point estimation than the measurement derived from an individual study. In performing a meta-analysis, an investigator must make choices which can affect the results, including how to search for reviews, selecting the studies based on a set of objective criteria, dealing with incomplete data, analyzing the data, and accounting for, or choosing not to account for and publication bias. The selected studies for analyzing and comparing came from varies industries, which their techniques for analyzing vary between industries since they weigh factors differently. Thus, meta-analysis meets the needs of this study. Despite the different errors derived from the various research conditions, however, the similarity of the algorithms that used a tool to manipulate the data is the standard truth that this paper takes as the research aim as well.

In the latter part of this study, an analysis of a small sample of the data from Wuzzit Trouble will be used to illustrate the potential factors and questions. The sample data that this paper used for the experiment is information from 83 logins that represents 83 users. This 83-user information is sampled from a larger data set that contains inaccurate and duplicated entries within one day. To present research approaches and avoid the privacy issue that the paper declared previously, I
merged those massive login entries in one day to 83 users’ information by logins with similar IP address, game version and so on.

**Meta-Analysis Review**

**A. Studies in Game Industry**

The game industry grew from focused markets to the mainstream began from the 1940s to the present days, meanwhile its revenue multiplied. In 2007, the market cap has increased to US$9.5 billion in the US, it increased multiple times in just two years according to the ESA annual report. Along with significant revenue, the competition among different games was rising at the same time. In the game industry, competition is essential to take place in a free market environment. As a result, being able to answer such questions as how players behave, engage and leave can help the game industry to improve game design or user-oriented testing. According to *Leading the edges of Chaos* (Mark Murphy, Emmett C. Murphy, 2002), reducing customer churn by 5% can increase profits 25-125%. Thus, implementing churn analysis may reduce production and marketing costs on a large scale. In comparison to Customer Acquisition Cost, the efforts of keeping and retaining current players are gradually taking more significant part in the marketing scheme. The endorsement is considering the research conducted by Frederick Reichheld of Bain & Company that increasing customer retention rates by 5% increases profits by 25% to 95% (Reichheld, 2001).

Churn analysis in games brings significant advantages to the enhanced overall effectiveness of the business. It allows the game developers to uncover user’s behavior (El-Nasr et al, 2013), transactions (T. Fields, B. Cotton, 2011), demographics data (E. G. Castro, M. S. G.
Tsuzuki, 2015), and usage pattern from the modeling data. The meaningful insights that converted data that is not only valuable for predicting players who are likely to churn, but also identify the causes for churn and discovering the potential solution accordingly. In conclusion, it is important to engage players to foster relationships to implement effective programs for player retention.

As the “churners” defined in previous game analysis paper, it refers to the subscriber, user or the player in the game that who left a game or service. The ratio of churners over non-churners as a function of the time determines the churn rate (Fabian Hadiji, Rafet Sifat, Anders Drachen, Christian Thurau, Kristian Kersting, Christian Bauckhaget, 2014). “ (Gallo, 2014). Typically, the churn rate is measured by month, quarter, or year, depending on the industry and the product being sold. Therefore, churn is a time-sensitive action that depends on how to determine at what point the players leaving should be considered as churning. To better understanding why a particular player churned, features that are derived from the historical telemetry data that might affect the action must be clean, non-volatile, and independent. In this paper, I’ve selected from eight game-industry churn analysis studies and ten telecommunication-industry churn analysis studies to compare the result on how the researchers use data mining to predict the customer churn in such a massive-audience base. The following reviews will mainly show the detailed analysis of two papers in the game industry on game genres, platform types, research questions, datasets characteristics, processing models, and results. Two churn analysis papers from telecommunication industry will also be analyzed as the endorsement but only focus on the data mining and modeling methodology to process the churn analysis. Table 1 presents the details of the targeted churn analysis papers in the game industry.
Table 1: Churn analysis in the game industry

<table>
<thead>
<tr>
<th>No.</th>
<th>Topic</th>
<th>Game Type</th>
<th>Contribution</th>
<th>Data Modelling</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing Times</td>
<td>action-adventure, shooter, P2P</td>
<td>Random process theory Lifetime analysis.</td>
<td>Weibull distribution on playing time</td>
<td>A formalization of the temporal evolution of people’s interest in playing a game.</td>
</tr>
<tr>
<td>2</td>
<td>Predicting Player Churn In the Wild</td>
<td>F2P games, social, mobile</td>
<td>Testing using data from five commercial games across mobile- and web based social-online platforms.</td>
<td>Neural Networks, Logistic Regression, Naive Bayes and Decision Trees.</td>
<td>Defining churn into ways and providing two predicting systems.</td>
</tr>
<tr>
<td>3</td>
<td>Rapid Prediction of Player Retention in Free-to-Play Mobile Games</td>
<td>Mobile, F2P</td>
<td>Predicting retention in F2P mobile games based on very short-term user behavior.</td>
<td>Heuristic-based decision tree</td>
<td>Retention prediction models are developed using a number of machine-learning models covering different windows of observation.</td>
</tr>
<tr>
<td>4</td>
<td>When Players Quit (Playing Scrabble)</td>
<td>Desktop, casual</td>
<td>Find the sequences of churn action in five transformed features.</td>
<td>Naive Bayes</td>
<td>Using a naive Bayes model and used it calculate the probability of quitting the game early.</td>
</tr>
</tbody>
</table>

Note: This table presents the selected churn studies in the game industry, along with the featured information.


Based on Distributions of Total Playing Times
The analyzed in-game data in this selected study are from five action adventure shooting games that all based on the desktop computer. These included two single-player games (Just Cause 2, Tomb Raider: Underworld) as well as three multi-player games (Battlefield Bad Company 2, Medal of Honor, Crisis 2). Table 2 showed the details of data when each player was actively playing the corresponding game.

**Table 2: Description of game telemetry data**

<table>
<thead>
<tr>
<th>Game Name</th>
<th>Game Genre</th>
<th>Modes</th>
<th>Online</th>
<th>Platform (Desktop/Mobile)</th>
<th>Monetization type</th>
<th>Players Observed</th>
<th>Period (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tomb Raider: Underworld</td>
<td>action-adventure</td>
<td>Single-player</td>
<td>No</td>
<td>Desktop</td>
<td>Paid to Play</td>
<td>146,233</td>
<td>2</td>
</tr>
<tr>
<td>Just Cause 2</td>
<td>action-adventure</td>
<td>Single-player</td>
<td>No</td>
<td>Desktop</td>
<td>Paid to Play</td>
<td>5,331</td>
<td>7</td>
</tr>
<tr>
<td>Battlefield Bad Company 2</td>
<td>First-person shooter</td>
<td>Multi-player/Single-player</td>
<td>Yes</td>
<td>Desktop/Mobile</td>
<td>Free to Play/Paid to Play</td>
<td>87,126</td>
<td>21</td>
</tr>
<tr>
<td>Crysis 2</td>
<td>First-person shooter</td>
<td>Multi-player/Single-player</td>
<td>Yes</td>
<td>Desktop</td>
<td>Paid to Play</td>
<td>4,364</td>
<td>6</td>
</tr>
<tr>
<td>Medal of Honor</td>
<td>First-person shooter</td>
<td>Multi-player/Single-player</td>
<td>Yes</td>
<td>Desktop</td>
<td>Paid to Play</td>
<td>12,328</td>
<td>14</td>
</tr>
</tbody>
</table>

*Note: This table explains the data from each game that involved in the research How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing*
The mode column focuses on whether the game is single-player or multi-player. The Platform column focuses on distinguishing mobile game from desktop games.

In this paper, the fundamental difficulty was how to model the dynamics of player engagement. Christian, the author, addressed the challenge in two aspects. He morphed the player’s interests into a numeric quantity by fitting playtime to the Weibull distribution and generate specific numbers. The time featured in this study described as the first passage times to represent the player’s interest towards the game. They labeled the gaming interest of players to different degrees at particular time-stamps, and then fitted to the first passage times with the player’s interest from a certain level to zero interest (Christian Bauckhage, Kristian Kersting, Rafet Sifa, Christian Thurau, Anders Drachen, Alessandro Canossa, 2012).

Since the time featured in the game engagement involved significant randomness, while the interest of players is an uncountable element. As a result, the model that is suitable for distributing time data in the game has to perform well in the random process in nature. To infer the details about these unobservable process, the paper that labeled as 1 in Table 1 considered that the lifetime distribution, or per say the first passage time distributions, better perform by Weibull distribution, Gamma distribution, Log-normal distribution and Inverse Gaussian distribution. The main reason, according to the author, was that these distribution methods are known to be in one-to-one correspondence to the distinct random processes that occur in nature frequently and well-understood. With further analyzing, the paper indicated that the underlying latent process is a power law process, which is a particular inhomogeneous Poisson process through Weibull distributed first passage time that helped on discussing how the latent variable, the player interest, in this case, influenced the lifetime analysis. In other words, the paper inferred that the average player’s interests in playing action games drops following the probability density function of the Weibull random variable. The most identifying characteristic is that the “failure rate”, or the rate
that players lost their interests in the games, is affected by the shape parameter, which refers to the first passage time. Since the beginning passage time is larger than 1, the rate that players lost interests is increasing with time (Wikipedia, 2018), (Christian Bauckhage, Kristian Kersting, Rafet Sifa, Christian Thurau, Anders Drachen, Alessandro Canossa, 2012).

In this paper, one of the suggestions is that the player mode cause the potential differences on churn prediction. One histogram in this study visualized the player activities of each game as normalized the observed data by day of the week. It revealed that only the player activity in *Just Cause 2* noticeably spikes on Saturday, for the rest of the games player activities distributed evenly in terms of the average daily playing time. The assumption addressed in the play mode that *Just Cause 2* is a single-player game. Meanwhile, the tendency towards higher activity on the weekend was captured in *Tomb Raider: Underworld* as well, minuscule but noticeable. Combined the information from table 2, which shows the characteristics of the observed data, the single-player mode not only exists in these two games. The other three games are multiplayer games because the majority player would maximize the use experience in multiplayer mode, which aligns better with the initial design purpose as oppose to playing in single mode. In conclusion, the playing mode cannot explain the churn activity of the players, so it is not vital in this game analysis.

A feature that influences the research sharply and appeared as a problem during the data modeling is the data size. With the purpose of quantifying the distribution quality, the study ran the Kolmogorov-Smirnov tests to resort the goodness of fit. The interpreted KS scores shows how good the Weibull, Gamma, Log-normal and Gaussian fitted each game. However, the p-values of *Tomb Raider: Underworld* showed 0 in every distribution method as mentioned. It was a consequence of the vast amount of the data, which the observed players were over 100,000, rather than a failure of modeling. Some statistical tests are useful under particular conditions, for instance,
Kolmogorov Smirnov test is especially useful for processing small data samples. Details about choosing a suitable test method to test the fitness of result will be further discussed in the next section. What should learn here is that to pre-determine the data size in game analysis that involving multiple games are necessary in the initiative on data retrieval to forbid the other difficulties it may happen in downsizing the sample data.

It is necessary to consider the effect that the online or offline play mode would cause, especially when targeting multiple games. The Christian’s study tried to eliminate the inconsistency caused by the differences between the online game and the standalone game. Removed the *Tomb Raider: Underworld*, which is an independent game and showed an abnormal p-value in distribution performance, the remaining three games were all fit well in Weibull distribution despite online or offline.

**b. Review of Predicting Player Churn In the Wild**

Free-to-Play(F2P) is the revenue model that is the most widespread nowadays in the game industry. A significant number of the games shifted their business to a F2P model to extend the life cycle of the game. The number of times that players accesses the game does not directly drive revenue, but the in-game purchase does. As a result, the ability to predict the timing of when a player will churn is more important, especially in finding out a player’s in-game purchase behavior before they churned, it is crucial for F2P games to build a sustainable revenue model. Hadiji’s paper focused on the churn prediction on F2P games and provided many insights of the churn prediction for the future analysis. With two definitions of game churn prediction that proposed from the real-world perspective, this study came up with the adapted statistical model and methods to analyze correspondingly. As declared in the paper, this is the first churn prediction analysis that
is across multiple games, and it proposed various statistical models of the retention likelihood and non-linear behavior function, the Hadiji’s study offers good examples to discover the player’s engagement behavior in games.

There are five unidentified games analyzed in Hadiji’s paper and they were in the mobile F2P and social game categories. All data were collected by a third-party service named GameAnalytics (www.gameanalytics.com). The author did not specify either the titles or the genres of these five games due to confidentiality reasons for this service. The paper only indicated that all the data from these five games were gathered over a five-month timespan in the same year and reached twenty million play sessions entirely. The primary features centered around sessions and were normalized by discarding the game content dependent features as well as the invalid data.

The two definitions proposed were the tight-churn and the relaxed-window-churn. Rather than only considering the last engagement of a player as churn, which refers the tight-churn as the most churn analysis did, this study also labeled players as churners who have a low number of play sessions after the cutoff date. Under the comparison of hard churn (tight-churn) and the soft churn (relaxed-window-churn), the churn action was discussed aligning with the game industry standards as a formal problem.

This study compared various classifiers with two types of features proposed, including Decision Trees, Logistic Regression, Naïve Bayes and Neural Networks. By weighting the averaged F1-score of each classifier corresponding to two definitions of churn, the differentials appeared on the feature importance between the two approaches that derived from the two churn definitions. The final results intended to show better performance when practically considering the churn as the relaxed-window-churn, which consider a player churned even if they play a little.
B. Studies in Telecommunication Industry

Churn analysis did not apply to game analysis at first. Plenty of studies on customer churn analysis of the telecommunication companies lead the current techniques for churn research, not only on data mining techniques but also the process of conducting research. The wireless industry has been through the same scenario as the game industry, in particular, the period after the revenue model for game industry become more dependable on continuous subscriptions, where multiple provider has made it competitive. By retrieving the previous studies, I noticed that there is not much churn analysis in game research, but a large number of churn analysis in the telecommunication industry, see for example (Owczarczuk, 2010), (Saravanan M., Vijay Raajaa G.S., 2012). In order to predict the number of users who would leave their service, those wireless companies put lots of resources to reveal the customer churn pattern. While they were defining the churn action from various perspectives and situations, they also proposed different data mining algorithms to fit the corresponding types of research data. Thus, the churn analysis for games that share a similar life cycle to the telecommunications industry, there are many insights to learn from their studies of churn.

The needs of the churn model, regardless of the economic sector, are documented in the previous study (Owczarczuk, 2010) which states the customer's retention is one of the significant activities of customer relationship management. The dramatic increase of competition in the telecommunication industry also makes customer churn a great concern for the providers (Richeldi & Perrucci, 2002). There is a study that even declared that churn is a significant problem for companies with many customers, e.g., credit card providers or insurance companies (Saravanan M., Vijay Raajaa G.S., 2012). Because of its massive audiences, the games industry is facing the
same circumstances as those industries that listed above. The similarities between the telecommunication and game industries indicate that previous work on churn analysis from within the telecommunications industry could provide a valuable resource for analyzing churn problems in games.

Another similarity lays between two industries is the monetization approach. The telecommunication companies usually offer their customers two payment plans, prepay and postpay. The plans resulting in a two to three years contracts are called postpay, versus the prepay plans requires a month-to-month payment basis. Adapting the two monetizing methods to games, they are the free-to-play with in-app purchase and the pay-to-play game version. Almost all free-to-play (FTP) games attempt to monetize players through in-app purchases, even though the games present themselves as free. Once the player spent any amount of money on a game item within the FTP game, it becomes a prepay plan that is charging on an on-going basis. On the other hand, the PTP (pay-to-play) game is postpay where a customer must pay in advance. A playing period of a PTP game before churned follows the similar logic of the postpay wireless plan that user must sign a contract except for PTP game that has no contract binding. Considering how incredibly important churn research has been to wireless industry, it is undoubtedly wise for the game companies to figure out the factors that might affect the stay duration of its customers. Furthermore, it is crucial for game companies to discover the algorithm that enable them to predict the timing of churn.

The selected studies in churn analysis that focus on telecommunication demonstrate how to predict the customer churn rates under certain kinds of data, addressing different research questions from different perspectives. Each study provides valuable insight by using different data mining models to predict customer churn based on the characteristics of the collected data. By presenting each of the step during the research, the articles demonstrate how problems usually
show up in churn analysis. More importantly, by observing the frequency of occurrence of those problems from the literature, it reveals the popular data mining techniques used. Table 3 shows the overview of the selected papers in telecom industry.

**Table 3: Churn analysis in the telecommunication industry**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Focus</th>
<th>Methodology</th>
<th>Data Modelling</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey</td>
<td>Statistic Graphic model of churn problem</td>
<td>A Bayesian Belief Network to identify the behaviors of customers with a propensity to churn.</td>
<td>Bayesian Belief Networks, CHAID</td>
<td>Causal map</td>
</tr>
<tr>
<td>Churn models for prepaid customers in the cellular telecommunication industry using large data marts</td>
<td>Most influential Features discovery</td>
<td>Dealing with prepaid clients and their life curve for all the percentiles, even after six months.</td>
<td>Logistic Regression, linear regression, Fisher linear discriminant analysis and decision trees.</td>
<td>Linear models as logistic regression perform better than decision tree in high percentiles of the lift curve.</td>
</tr>
</tbody>
</table>
A Graph-Based Churn Prediction Model for Mobile Telecom Networks

Network visualization of Churn

Apply graph parameters into churn prediction to compute a network analysis

Network analysis, Multivariate Discriminant, Linear Regression

Provided a novel idea of using the graph-based network to visualize churn in Telecom industry

Note: This table presents the selected churn studies in the telecommunication industry, along with the featured details.

a. Review of Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey

Research into behavior across a large customer base is mature in the telecommunications industry. Not only the amount of the churn related papers I found within the telecom industry revealed this fact, but also the quality of every paper regarding the completeness of the approach design as well as the practicality of the given results. In Pinar’s study, the churn analysis of the customers from a wireless company in Turkey was performed in a casual network map based on the Bayesian Belief Network, via combining the data modeling results of correlation analysis, multicollinearity test with the experts’ suggestions.

This is aligned with the reviewed churn prediction papers in the game industry, where the focuses are also on the definition of churn, the features selected and the statistical models to process the datasets in different characteristics in this study.

The definition of churn in this study followed the traditional logic that the moment when a customer terminates a subscription. Thus, the churn rate can be calculated as the result of numbers of churn divided by the total number of users.
The data collected is from 2000 subscribers, including 534 known churners, within six months during 2008, nine features modified from the initial dataset became the critical variables for analysis. These variables are **Place of residence, Age, Tariff type, Average billing amount, Average minutes of usage, the Average frequency of usage** and **Trend in billing amount**. In particular, Trend in billing amount, which is a latent variable created in addition to discover if the billing amounts are going up, down or stay constant. Compared to the conventional features in churn analysis research of games, the variables generated here are neither temporarily nor consumer-related, they are spatial and customer-based. The divergence of the critical feature selecting between the churn analysis on games and the churn analysis on wireless service presented the importance and relativity in people’s life.

To reveal the relationship between each variable, researchers depicted three scenarios to assume the cause-effect casual map among the nine variables, according to the suggestions from the experts. Each of the three scenarios showed a network plot, with directions between each variable. By assuming the relation flows that exist potentially, and by adjusting the weighting factor based on experience, the study is constructed by the Bayesian Belief Network in each of the three scenarios. The paper later provided the details of the significance of influence that a specific feature has on the other features under different situations, and further determined the most critical factors that have the effects on customer churn in the telecom industry. Therefore, the methodology of utilizing industry experts’ knowledge to propose a hypothesis on correlations of factor had illustrated an excellent example of network research between features in churn analysis.

*b. Review of Churn models for prepaid customers in the cellular telecommunication industry using large data marts*
Owczarczuk’s study is another churn analysis within the telecommunication industry mainly focus on the behaviors of the prepaid customers. Since prepaid customers does not sign any contract and their identities are kept anonymous, the study does not provide personal information. Apparently, it is easier to compare the users in the prepaid program in wireless service to the players in games because most of the churn analysis in games has tried to avoid undetectable noise that is easily caused by the personal preferences. Consequently, the Owczarczuk’s paper is an excellent resource to learn experiment design, feature selecting and modeling strategies for constructing a general churn analysis in games.

The data has been collected in two different methods in this study. The training sample was collecting simultaneously with the validation sample, which was randomly split from the dataset obtained in the first term. The test sample was received in the second term that was six months after the train and calibration dataset. The total amount of the observations was over 167,000.

The variable selection was accomplished in the preliminary test on training dataset in a computational method rather than determined by the experience of the researchers. From the 1381 primary variables of the raw data, each variable has been applied the Student’s t-test to distinguish the 50 variables with the highest absolute value in the t-statistics. Initially, the study proposed two null hypotheses. One stated that the means of a particular variable among churners and non-churners do not differ from each other. Another study suggested the existence of the significant deferens between these two means. The researcher then inferred upon the hypothesis that the variables interested in churn modeling have a substantial difference between their means. The study applied the Student’s t-test on every variable and mined out 50 variables in which they have the highest t-statistical numbers. Then the additional variable selection methods such as stepwise-
backward-, forward-regression were applied on these 50 variables to further prove that they are the most influential variables in this model.

The study performed four models towards the selected features: logistic regression, linear regression, Fisher linear discriminant analysis and decision trees. After measuring the quality of each fitting used the life curves, it showed all linear models performed similarly, and the logistic regression outperforms the decision trees. Since churn is a time-sensitive behavior that change with time continuously, it is not surprising that the study suggested the linear models as the most stable method in retention problem.
Further Analysis of Essential Components of Churn

Based on the views in previous articles, multiple crucial parts for churn analysis can be concluded from both game and telecom industries. These essentials include the definition of the churn action, the feature selection, models for different characteristics and the classification methodologies. Defining the churn accurately in various aspects is crucial for executing the churn prediction, including the actual action, timing and conditions. In sum, defining churn is the prerequisite for the research, it leads the analysis directions in every aspect.

The goal is to design a generic churn analysis plan that can apply to games to predict the players’ churn behaviors, details of the essential aspects mentioned above are extracted from more than ten studies and being discussed separately. Further comparison of the algorithms and techniques that implemented in the previous studies is the focus, including but not limited to the ones in data preparation, feature determination or churn labeling.

C. Churn definition

As I defined the churn at the beginning, it is the action of users’ leaving in general. However, defining the churn by the actual meaning of words is not as important as defining the duration of user leaving and counted as churning. Some of the studies took churn as the behavior that take place suddenly. They named the churn that occurred “within a second” as the sharply churn. Considering churn action as an immediate move can simplify the logic by transferring a complicated prediction problem that might be influenced by varies factors into a logical explanation. The sharply churn is ideal for simple data modeling, whereas it is hard to say whether the user who he or she left at one point will never return. In conclusion, the propose of considering
churn as a continuous action that happens over a period or happens over terms is more practical. Nevertheless, the difficulties of dividing the time periods on the exact time stamps and scaling the churn level in equal terms raised complexity of data modeling. The two directions of definition bring churn analysis with the advantages and the disadvantages, which is a tradeoff that requires balancing.

The Hadiji’s paper that analyzed a sample in the previous section named the churn behavior over a period as relaxed-window-churn and named the sharply churn as tight-churn. The most significant difference between these two types of churn is on whether the player who returned after the cutoff date counted as churner. The tight churn considers all the players who returned after the cutoff date as a returner, however, the relaxed-window-churn intends to take the player who retained only once after the data being cutoff. The following definition towards churner is more useful in the real-world applications since getting bored with a game is a gradual process that take place over time.

The crucial problem in churn is where to set the cutoff time. The Hadiji's study provides a methodology to determine the exact cutoff point in a logical way. The negative or positive value will be marked on each part of the dataset after randomly choosing a game session as the breakdown point and subset the data into the churn users or the returned users. The process of dissecting the useful part from the useless part should be kept running until one session with the proper cutoff time is found. In which a session that, stops most of the players from playing directly. However, this approach in separating churn and non-churn data is ambiguous in some respects. For example, the initial dataset contains the players, who bounced after playing once. It is evident that the one-time playing is not a typical circumstance in gaming behaviors, either the valid data for churn analysis. It is more genuine to discard the players who only have played one session
before the cutoff time stamp. Secondly, the size of data may reduce the effectiveness of subsetting. The times of data dividing is unknown until the final determination of the best split point between churn and return is established. Similar to cutting papers, the number of datasets after splitting grows exponentially as the size of data expanding. Therefore, this method is not the best choice for processing the massive dataset.

Another churn study in the game industry, A. Drachen’s study, also considered splitting the churned data and retained data by the binary classification task. The data was scaled weekly, and the first-week data after the installation of has been eliminated from the raw dataset before labeling. Retained class refers to the players who register at least one game round in 7-14 days following installation. Taking the data without the first week minimized the confounding of considering immediate-bounced players. Scaling the examination period from 7 days to 14 days also avoided the seasonal patterns in playing, such as weekdays versus the weekend.

The rest players who are not in accord with the retained criteria are the dataset that needs to be defined as churners. The Drachen’s paper suggested three feature windows, that are the usual cutoff points during the game playing: first, the end of the first play session; second, the end of the first day of game playing; and third, the end of the first week of playing. Accordingly, different classification strategies that are dealing with specific churn feature windows should be proposed, so applying heuristic-based decision rules rather than the traditional classifier can achieve more accuracy and more actionability.

A customer churn study in the telecommunication industry, the Saravanan’s study visualized churn behavior by network analysis, defined churn as the action with multiple stages (Saravanan M, Vijay Raajaa G.S, 2012). As Saravanan’s paper presents, the time unit to predict churn is a week, and a window frame was constructed to show the process of churning. The Saravanan's
study did not determine the churn time precisely; instead, it named the churn window frame by stages. The first and the second stages are before and during the cutoff time. The third stage is churn window which is the stage that the customers are most likely to churn. While the scene is moving as time goes by, the graph parameters that utilized to analyze the churn network is changing symmetrically with variations marked as probable churners. Defining the churn into multiple stages affect the result of churner splitting in the later analysis in Saravanan’s study. The customers who churn out in the window slot, which is the third stage are considered as real churners, in which the variation in selected graph parameters will be the mainly tested, rather than the first and the second. In other words, how to define the churn behavior and the churning period determine the data size and the complexity of the prediction.

Summarizing the instances of the churn definition above, considered the churn behavior as a procedure is how most of the studies conducted the churn analysis in both game industry and telecom industry. To achieve the goal that constructs a generic churn analysis plan for games, define the churn in stages is complicated but accurate.

If we apply this understanding of churn analysis to the Wuzzit Trouble case, the methods introduced above for finding cutoff points by continually dividing the dataset would make it possible to extract the churn players from the data sample, although this would not be suitable for large datasets. It might be useful to locate churners by this method for the small games like the Wuzzit Trouble, which has a clear boundary of audiences that expected from the specific age groups, and which does not contain any social functionality within the game. However, for massive dataset games that have data running over the years, dividing data in this way is apparently redundant. Moreover, comparing to determine the initial session as cutoff point by randomly assuming, starting from the middle position is more legitimate and effective.
D. Feature selection and churn extraction

To dig out the features which are genuinely significant for research questions among the thousands of options preparing the analysis a good start. In this section, a summary of how the churn studies in both game industry and telecom industry select the research features based on the churn definition and data characters. There are two directions for selecting the features in churn prediction, which are underlying theory based and statistics based, or two steps of feature selection precisely since additional feature choosing based on the draft selection adopted in some papers as well. Through the analysis on the feature selection methods utilized in the review studies, the discussion on the suitable features selection method of constructing a general churn analysis in games will address on the Wuzzit Trouble as an example.

a. Underlying theory based feature selection

According to previous studies, most of the researchers decide the research features base on their research experience and common sense due to the difficulties of accessing data in targeted games. Furthermore, the gap of the features choosing between big games and small games performed how vital role that the game contents play in game churn prediction.

The conventions of feature selecting provide the instances. Study of C. Bauckhage, which targets on big desktop games with the research address on the scale the interest of game playing by temporal distributions, chose more features about time, for instance, the number of hours played, weekdays vs. weekends and the player's interest that the latent variables need to be valued. The Brent Harrison’s study that focuses on predicting when players quit Scrabblesque, a small desktop word game compares to the AAA games in C. Bauckhage study. Since the game does not
involve any human opponent, the features originated from the game such as the *mouse clicks times* and *player rack status*.

Another example that is worthwhile to mention is a set of generic behavioral features introduced in Hadiji’s paper (Fabian Hadiji, Rafet Sifat, Anders Drachen, Christian Thurau, Kristian Kersting, Christian Bauckhaget, 2014) and defined in the study to build an analytic model for various games regardless of the game content. Besides the temporal features that have an apparent influence on churning, features related to in-app purchases were also included for fitting in characteristics of the F2P games. Table 4 summarizes the essential temporal features, and the Table 5 summarized features related to the virtual economy which are meaningful for the In-App purchase in the F2P games.

**Table 4: Overview of features in Hadiji’s paper.**

<table>
<thead>
<tr>
<th>Temporal features</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Sessions</td>
<td>The number of past sessions a player was actively involved.</td>
</tr>
<tr>
<td>Number of Days</td>
<td>The number of days since the player has signed up for the game.</td>
</tr>
<tr>
<td>Current Absence Time</td>
<td>The elapsed time since player's most recent activity.</td>
</tr>
<tr>
<td>Average Playtime per Session</td>
<td>The total playtime of the player divided by the number of sessions.</td>
</tr>
<tr>
<td>Average Time Between Sessions</td>
<td>The average time between the sessions, i.e., intersession time.</td>
</tr>
<tr>
<td>Playtime Model Parameters</td>
<td>i.e., parameters of the player-based power law to represent the player's playtime history until the day of prediction</td>
</tr>
<tr>
<td>Retention Value(latent)</td>
<td>Have the average player retention fitting in a model based on the day of play.</td>
</tr>
</tbody>
</table>

**Table 5: Overview of In-App Purchase features in Hadiji’s paper.**

<table>
<thead>
<tr>
<th>Virtual Economy related features</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premium User Flag</td>
<td>Denoted if a player made a purchase.</td>
</tr>
<tr>
<td>Predefined Spending Category</td>
<td>Categorized the purchase experience of the players.</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Number of Purchases</td>
<td>The total number of the purchases a player has made.</td>
</tr>
<tr>
<td>Average Spending per Session</td>
<td>The average amount spent per played sessions.</td>
</tr>
</tbody>
</table>

It is evident that temporal features are considered in multiple studies most frequently. Creating temporal features based on the given churn definitions and game content correspondingly is the trend as the examples show. Additionally, the difference in selecting the temporal features between F2P games and the P2P games is inconspicuous as the common that used the log-on and -off time, number play sessions, weekends or weekdays, etc. The latent variable created by indicating the inventiveness of churn are players’ interests and retention values in previous sample study in this section, however, the underlying meanings of these variables are still similar. Similarities can also be found on the features that created based on the game content, even the genre of the games are different.

b. Statistic based feature selection

Except for the experience-based feature selection, using machine learning approaches is another approach to analyze the most contributed features to the phenomenon, which is churn behavior in our case. Implementing machine learning algorithms into feature selection improves the accuracy of the research as well as the effectiveness of data retrieving. Due to the character of churn prediction, the ideal possible results are two classes, churner or non-churner. Extracting out churners from massive users becomes inevitable. The most common algorithms for discriminating the significant features on top of the classification background are Discriminant analysis according to the reviewed studies. There are examples of utilization in both churn study in the game and in the telecom industry.
Discriminant analysis is a statistical technique that is suitable for the research data the dependent variable is categorical and the independent variables is an interval in nature. The objective of the discriminant analysis is to develop the linear combinations of independent variables to discover the differences among the dependent variables. The difference among the predictor groups discriminates the category that each of the independent variables belongs to (Discriminant Analysis, 2018). The discriminant analysis procedure can be summarized with the following steps: 1) Determine and formulate the problem. 2) Compute the discriminant function coefficients. 3) Examine the significance of these discriminant functions. 4) Interpret the results obtained. 5) Validate the results. The Saravanan’s study is an example of utilizing discriminant analysis in telecom churn analysis. It applied linear discriminant analysis on the graph parameters that were developed for the churn network analysis to find the canonical weighted score stepwise and interpreting the results by Wald test. The value of canonical weighted score for each graph parameter determines the contribution it made to the churn behavior. The Wald test enable to number the true value of the parameter, which is another method that frequently appears in feature selecting. One point that should be noted is that the discriminant analysis is effective only in the processing of a small number of variables due to the possibility of collinearity and extremely long-time computation.

In addition to the multivariate model such as Discriminant analysis, the Student’s t-test approach, which defines the most valuable features, provides the ability to deal with a more extensive dataset with multiple variables. Student’s t-test is the statistical testing method used to verify how well a model fits. It has been widely adopted to cross verify the outputs by performing the mean and the standard deviation scores in different churn stages of periods.
E. **Statistical Methodologies**

c. Data recording

Proceeding the model based on the data types and adjusting the data into ideal format enable the analysis takes place. There are always two data types, discrete and continuous. Discrete data refers to the information that only in particular values such as an exact number or the non-numeric categories. Continuous data can be measured on a scale which can be meaningfully subdivided into more delicate increments, depending upon the precision of the requirements.

Chi-square automatic interaction detection (CHAID) is a solution to discretize the continuous variables. The statistical analysis begins with CHAID analysis for discretizing variables as step 3 in study workflow diagram. Chi-square automatic interaction detection (CHAID) is a decision tree technique, based on adjusted significance testing (Bonferroni testing). The CHAID algorithm has proven to be a practical approach for obtaining a quick but meaningful segmentation where segments defined regarding demographic or other variables that are predictive of a single categorical criterion (dependent) variable. One crucial advantage of CHAID over alternatives such as multiple regression is that it is non-parametric (Magidson, 2011).

d. Discriminant and collinearity

Discriminant analysis is a statistical technique that suitable for the research data where the dependent variable is categorical and the independent variables is an interval in nature. The objective of the discriminant analysis is to develop the linear combinations of independent variables to discover the differences among the dependent variables. The difference among the predictor groups discriminates the category that each of the independent variables belongs to (Discriminant Analysis, 2018).
The procedure of discriminant analysis can summarize as the following steps:

1) Determine and formulate the problem.
2) Compute the discriminant function coefficients.
3) Examine the significance of these discriminant functions.
4) Interpret the results obtained.
5) Validate the results.

The Saravanan’s study is an example of utilizing discriminant analysis in telecom churn analysis. It applies linear discriminant analysis on the graph parameters that developed for the churn network analysis to find the canonical weighted score stepwise and interpreting the results by a Wald test. The value of a canonical weighted score for each graph parameter determines contribution it made to the churn behavior. The Wald test enable to number the true value of the parameter, which is another method that frequently appears in feature selecting. A point that should be noted is that the discriminant analysis is effective only in the processing of a small number of variables due to the possibility of collinearity and extremely long-time computation. The stepwise regression that processes a new parameter each step based on the current parameter improves the accuracy, but it is prone to overfitting the data under specific circumstance which needs to do the folder-cross validation to reduce the risk.

It usually contains the binary classes for a churn problem, in which there are only two possible outcomes, such as true or false. Thus, the complexity of discriminant analysis is reduced. The Saravanan’s paper is one of the examples for implementation. It applied Logistic regression and Multivariate Discriminant analysis on selecting the best graph parameters on different stages of churning, which developed a graph-based model on mobile telecom network to visualize customers churn behaviors.
Collinearity is a phenomenon that the others can linearly predict one variable in a regression model with an influence on the accuracy of this multiple variables.

The coefficient estimation number of this regression model may have erratic changes due to the response of the other small transformation of the data which reduce the predicting power.

To test whether the variables this paper chose are collinear between each other, I calculate the variance inflation factor (VIF) for measuring the collinearity. Collinearity is a situation that usually happens during the modeling between the variables within the research, and it is an inevitable step in data preparation.

e. Statistical Graphics model
   i. Bayesian Belief Network

Bayesian Belief Network is a probabilistic directed acyclic graphic model that graphically presents a statistical model to illustrates a set of variables and their conditional dependencies with directions. This graphical model consists of nodes and edges. The nodes represent random variables and arcs represent direct probabilistic dependencies among them, in other words, the nodes represent the variables such as the observable quantities, the edges with row represent the connected directions.

The model is usually used to mimic the causal structure of a domain based on a joint probability distribution that updated from the Bayes theorem. In Kisioglu’s study, the researcher used a causality network map to illustrate the churn behavior for a telecom company in causal relationship diagram to visualize the statistical model. However, there are few churn-researches in game industry implementing this method to predict churn.

ii. Network Analysis (Link Analysis)
Network Analysis is a theory that utilized to depict the relationship of a network. It is a study of graphs as a representation of either symmetric or asymmetric relations between objects. Link analysis is one methodology under the network theory which widely used to analyze real-life problems such as the fraud detection, search engine optimization, marketing research and medical research. The application and construction procedure is similar to the regular network analysis and belongs to a knowledge discovery method.

Conventionally, the application of the network analysis is on social networks. This paper aims to attempt on implementing the network theory to describe the relationship between each feature, and trying to illustrate the affection the features caused to players’ churn behaviors in the game Wuzzit Trouble.
The research design determines the strategy to address the problem by integrating the knowledge from various sources into a systematic methodology coherently and logically (De Vaus, D. A, 2001), (Trochim, William M.K, 2006). A good plan, no matter in what kind of research, ensures the research problem to be efficiently revealed, which emphasize the priority among the other parts. However, it is a common mistake made by researchers to begins the investigations before thinking of a plan thoroughly (Organizing Your Social Sciences Research Paper: Types of Research Designs, 2018). Combined with churn analysis, the theme of the reviews, the study intends to address the focus to the research design in the churn analysis.

This section presents the design of a churn analysis research plan in an educational game, Wuzzit Trouble. It contains two versions of design that are the two components formed this section: the initial version and the revised version. Based on the same game, the final design is delivered through the comparison between the initial version and the revised version. Since the goal of this study is to bring a research design for churn analysis in the game industry, the research design as deliverables in this paper are composed of the aim, research procedure, variable selecting and the data modeling, the crucial parts in a churn analysis.

Wuzzit Trouble is a mobile educational game that helps children in elementary grades, second through eighth grades specifically, to learn math in an entertainment environment with the smartphones and digital tablets. It made by Brainquake Inc. and came out around 2001 with the original version played on the computer before becoming widely spread in application stores on mobile phones. The mechanics in Wuzzit Trouble based on the concept of integer partition, actualized by slipping the gear to express numbers in the game to solve the puzzles that are more
than 70 at different challenge levels. The contribution that the Wuzzit Trouble made, which is being a useful tool in the classroom-tech toolbox and developing the critical-thinking and problem-solving skills of students, broadly acknowledged by teachers and players (Klindt, 2015).

A Wuzzit Trouble dataset that consists of 83 login information is utilized as the sample dataset for presenting the concept of the churn analysis research design. It is important to notice that each login information is unidentified regarding the individual players, which cannot tell the players in the examined sessions or either categorize the login data into different players due to the regulation in children’s privacy in Massachusetts law.

F. Preliminary Design

The initial research design composed in the preliminary analysis. The sample data were collected in one week, from 07/20/2016 to 07/27/2016, stored in the format of JavaScript. For the convenience of data manipulation, I adjusted the sample dataset into a suitable shape that separates the tree-based data into columns by Studio 3T.

The following JavaScript shows the detail of one login data:

```javascript
{"r":"21c6da02-5720-350e-9610-99072c6c2a7e","s":1104016318447,"t":1469013187948,"m":"32.212.136.16 vn51o06m LOG level='Botany Lab Puzzle 1' stars_earned=2 event_name='Puzzle Complete' local_time='7/20/2016 7:13:06 AM' time_to_finish='120.43 seconds' time_to_turn_gear='112.26 seconds' points_earned=1000 moves=2"},
```

a. Research Question
The research question of in the initial churn analysis design is to discover the most influential features that contributed to the churn action.

Hypotheses are made that features in sample data are correlated to player's churn action. The features can be any in sample data and can be replaced by any feature combination in the sample data. For instance, taking the churn action as a binary feature, the feature *Number of Stars Earned* has no influences on churn action. By rejecting the hypotheses, the research may indicate the results and the relations between the features in graphics method.

b. Features

The initial analysis used the features in the original data sample with ignoring how the features have been determined beforehand.

Reading the login data from the left side to the right side, the features are:

*Series numbers; Device Number; IP Address; Game Level; Number of Stars Earned; Even Names; Login Time; Level Finished Time; Gear Turning Time; Points Earned; Level Reset times, Item Grabbed Number; Moves Count.*

In the initial research design, the data collection strategy is not included. Since the prior intention is to construct an overall plan through procedures, I took this sample data as the raw data for our research design. It is obtained from numbers of levels of a player finished and to show if the trend of level-accomplishment is in upward, downward or constant trend. Thus, the latent variable created to represent the churning process is:

*Trend in Level Finished*

The 83-login information from the sample data has been labeled into 83 users by taking the logins with the same device version, continued login time, same game level and game version. This 83-user information derived from a larger scale of data that contains accurate and duplicated entries within one day. To present research approaches and avoid the privacy issue that declared
previously, massive login entries in one day are mixed into 83 individual players by combining logins with similar IP address, game version and so on.

c. Research Workflow

The research procedures of the initial churn analysis consist of seven steps in general. It has two levels to accomplish the casualty exploration.
Before predicting which parameters will be used to predict the pattern of churn, the correlation analysis, and multicollinearity test probably are the excellent methods to choose the useful factors. After proved that the parameters are ideal for further study using correlation analysis, determining the most effective parameter that affects the churn action of players is the next step. The data manipulated techniques that derived from the meta-analysis result will be applied to the possible parameters to see the cause-effected relationship.

G. Revised design

The new design is the revision of the initial one after reviewed previous studies which are churn analysis in the game industry and the telecom industry. The revised design holds the same purpose of the initial ones that aims at discovering the relation between features and the churn behaviors. Meanwhile, it contains new features on statistical modeling method and the visualization of the relations that better legitimately performs the research.

a. Research goal

The goal of this research is to present the relation between the features related to players’ gaming behaviors and churning in the Wuzzit Trouble in a graphics model.

b. Churn Definition

To initialize the definitions of the churn behavior according to the churning period determines how the analysis processes.

Based on the data character of the sample data, the study uses the trace to scale churn behavior. Ideally, it should consider churn as a series of actions that contains three steps: Before churned, churning and churned. However, the data sample of Wuzzit trouble does not contain
any retention and mid-login, so the churned players under specific trace refer to the player who has no movement made.

c. Workflow

Based on the initial workflow, I made some changes from “prepare data set” to “construct BBN”.

d. Data Manipulating

Determine the features of data modeling is the next step in setting the research goal. To have detailed features that related to game content, I chose a period from 04/10/2017 to 04/15/2017
to record the column “m” from the raw data into a standalone dataset, with 20 traces of the 83 players valued in vary play behaviors. Following is the takeaway of the features that expected to collect from the raw JSON file:

- Series numbers; Device Number; IP Address; Game Level; Number of Stars Earned;
- Even Names; Login Time; Level Finished Time; Gear Turning Time; Points Earned;
- Level Reset times, Item Grabbed Number; Moves Count.

To reveal the relation of player’s behavior and to predict churning, determining the features that contribute to player’s churning is the direction to go. Thus, one feature will be created to represent where the players stop playing. It also stands for the feature representing churn actions in our case.

The following section, Analysis Implementation will discuss how to recode the data into the desired format and how to generate the features detailed.

a. Analysis

- Data Recoding

Figuring out the data types of the features is the first step. There are two types of data basically, discrete and continuous. Discrete data usually refers to the data that has a fixed value, either numeric or non-numeric. Categorical is a typical type of the discrete data. However, continuous data is an ambiguous value within a specific range, so its value is flexible to accommodate with the analysis standard that the data modeling requires.

Recoding the continuous variable into categorical is an unavertable process. Based on the effectiveness and applicability shows in the review part, Chi-square automatic interaction detection (CHAID) is an option to categorize the continuous data into discrete.

- Eliminate the Collinearity
Collinearity is a phenomenon that the others can linearly predict one variable in a regression model with an influence on the accuracy of this multiple variables. The coefficient estimation number of this regression model may have erratic changes due to the response of the other small modification of the data which reduce the predicting power.

To test whether the variables I chose are collinear between each other, I calculate the variance inflation factor (VIF) for measuring the collinearity. To calculate the VIF number, I follow these three steps:

1. Run the least square regression of the target variable to get the $R^2$
2. Calculate the VIF number by the r-square number by the equation:
   $$VIF = \frac{1}{1-R^2}$$
3. According to the rule of thumb that if $VIF > 10$ then the multicollinearity is high.

   ➢ Discriminant analysis

Since the number of the features is not enormous, the linear discriminant analysis is an excellent approach to discriminate the useful features. To conduct the discriminant analysis, I followed these steps:

1) Determine and formulate the problem.
2) Compute the discriminant function coefficients.
3) Examine the significance of these discriminant functions.
4) Interpret the results obtained.
5) Validate the results.

   ➢ Validation
Student’s t-test is the statistical testing that most commonly used to verify the goodness of model fitting. It is widely adopted to crossed verify the outputs by performing the mean and the standard deviation scores in different churn stages of periods.

➢ Network analysis

The Network Analysis is a theory that utilized to depict the relationship of a network. It is a study of graphs as a representation of either symmetric or asymmetric relations between objects. Conventionally, the application of the network analysis is on social networks. However, it also can be used to describe the relationship between each feature that I monitored for churn behaviors in the game. To develop the network analysis, I plan to use ggplot2, igraph, GGally and sna packages in R, which highly recommended in the network visualization tutorials (Ognyanova, 2016).

➢ BNN

BNN (Bayesian Belief Network) is a probabilistic graphical model represents the conditional dependencies via a directed acyclic graph (Wikipedia, Bayesian network, 2018). GeNIe, one of the free statistical software to process probabilistic causal map produced by Bayesfusion, LLC., is the tool to recommend for BNN analysis. This software enables the researchers to build the map by merely created the nodes and entered the probabilities manually without dealing with visualization part (BAYESFUSION, 2015-2018). Moreover, since probabilistic requires to be calculated preliminarily out of the GeNIe, R can complement the calculations.
Analysis Implementation

H. Brief Introduction of Wuzzit Trouble

Wuzzit Trouble is a math educational game for helping kids before high school to learn basic math and calculation concepts. Since this study introduced the game in the previous part, the game screenshots are presented below to show the game mechanics.

There are three levels totally, but only two levels are available initially, Botany Lab and Invention Room. There are 25 stages in each level for collecting stars. Accessing the third level requires the player to reach the specific number of stars in the first two levels, which means based on the accomplishment of the initial levels.

The mechanics of the game focus on a gear that required to switch. The gear represents different number in different level. With the increment of numbers, players will get the keys and special items. Keys and items affect the points that the player can earn. Fewer movements that she/he made, more stars she/he can get.

One level is showed in Figure 1. The further screenshots of the Wuzzit Trouble are listed in Appendix for supplement information.

Figure 1:
I. Churn definition

Adapting to the characteristics of the data sample, I scale the churn by the game trace. There are 21 traces entirely in the dataset, and each trace contains 83 players data recorded simultaneously. I distinguished the churned and non-churn player in a trace by checking if the data exits for the player under the particular trace. When the player churned, the corresponding trace is blank. For better understanding, table 6 shows part of the data in the churn binary matrix constructed according to the raw dataset.

Table 6: Binary matrix example

<table>
<thead>
<tr>
<th></th>
<th>Player0</th>
<th>Player1</th>
<th>Player2</th>
<th>Player3</th>
<th>Player4</th>
<th>…</th>
<th>Player 82</th>
<th>Churn Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trace0</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>…</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>Trace1</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>…</td>
<td>T</td>
<td>31.3253</td>
</tr>
<tr>
<td>Trace2</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>…</td>
<td>T</td>
<td>49.39759</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Trace20</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>…</td>
<td>F</td>
<td>98.79518</td>
</tr>
</tbody>
</table>

This paper adopted the theory of churn calculation in marketing and business, however, player retention and the time series are not considered in the Wuzzit trouble data, I adapted the theory into following equation:

\[
Churn \text{ Rate} = \frac{SUM(\text{Number of the active player in a trace})-\text{(Number of player in a trace)}}{\text{The number of players in this trace}} \times 100
\]

If the result is a negative value, it means the players number grow faster than lose them. Vice versa, if there is have a positive value, losing them is faster than keeping them playing.
J. Data formatting

I reformat the data into different shapes regarding the analysis requires. The initial data collected in JSON. The full data is attached in appendix. Data has been processed in NoSQL DB, Mongo DB, to load and process the JSON file. The idea of using a NoSQL DB rather than the traditional relational DB is that the NoSQL data technologies support scalable and schema-less structure that helps in analyzing and storing huge datasets (Saravanan M, Vijay Raajaa G.S, 2012). To manipulate the Wuzzit Trouble sample data in good shape, I used Studio 3T for MongoDB. The processing screenshots stores in Appendix.

Based on the original features in the sample data, the separated features are ID, Timestamp, Platform, Device, Event, Traces, Version, Session ID. To expend the column Moves, the details showing in Appendix. The moves illustrated the movements the players made in corresponding traces. However, the implementation will not separate the feature profoundly into the single steps of playing. That feature selection stops at the level of traces.

Now the features declared as follows:

Series numbers; Device Number; IP Address; Game Level; Number of Stars Earned; Even Names; Login Time; Level Finished Time; Gear Turning Time; Points Earned; Level Reset times, Item Grabbed Number; Moves Count.

The blank space represents no playing information, and the data does not contain reactive players.

To reveal the relation of player’s behavior and to predict churning, determining the features that contribute to player’s churning is the direction to go. The following section discussed the situations that usually occurs during the research.
K. Collinearity Test

The discriminant analysis requires the variables be dependent, thus, I conduct the collinearity test to eliminate the collinear variables.

```r
# Collinearity test
library(car)

## Loading required package: carData

lm <- lm(Churn ~ PointsEarned + BonusItemsGrabbed + MoveCount + StarsEarned + TimesLevelReset, data=AllTraceW)

vif<-vif(lm)
print(vif)
```

<table>
<thead>
<tr>
<th>PointsEarned</th>
<th>BonusItemsGrabbed</th>
<th>MoveCount</th>
<th>StarsEarned</th>
<th>TimesLevelReset</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.976917</td>
<td>5.768109</td>
<td>1.609131</td>
<td>4.426368</td>
<td>1.061502</td>
</tr>
</tbody>
</table>

Table 7: VIF scores

The VIF scores for all the variables showed in Table 7 display the result of the collinearity test. The feature, PointsEarned, has the greatest VIF value that over 10. Therefore, PointsEarned should be eliminated from the further analysis.

L. Discriminant Analysis

Means and variances of each class in Discriminant Analysis determine the linear boundary, or separation, between them. The coefficients delimit this boundary.

The analysis presents following:
Result of lda(Churn ~ PointsEarned + BonusItemsGrabbed + MoveCount + StarsEarned + TimesLevelReset, 
data=AllTraceW)

Call:
lda(Churn ~ BonusItemsGrabbed + MoveCount + StarsEarned + TimesLevelReset, 
    data = AllTraceW)

Prior probabilities of groups:
   FALSE    TRUE
0.8187034 0.1812966

Group means:
   BonusItemsGrabbed  MoveCount  StarsEarned  TimesLevelReset
FALSE    0.02242467  0.1723896  0.02592852    0.00000000
TRUE     1.68670886  4.6424051  1.57911392    0.07594937

Coefficients of linear discriminants:
   LD1
BonusItemsGrabbed  1.239433
MoveCount          0.140564
StarsEarned        1.918037
TimesLevelReset    -0.534105

Table 8: Linear Discriminant Coefficients

<table>
<thead>
<tr>
<th>BonusItemsGrabbed</th>
<th>MoveCount</th>
<th>StarsEarned</th>
<th>TimesLevelReset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.239433</td>
<td>0.140564</td>
<td>1.918037</td>
<td>-0.534105</td>
</tr>
</tbody>
</table>
There are four different models in this analysis; each model contributes to analysis the dependency between the target variables and the latent variable. The first one depends on the variable BonusItemsGrabbed, the next is on BonusItemsGrabbed and MoveCount. The third model depends on variables BonusItemsGrabbed + MoveCount + StarsEarned, and the forth is BonusItemsGrabbed + MoveCount + StarsEarned + TimesLevelReset. Results shows in Table 8.

The first thing should be checked is the Prior probabilities of groups. These probabilities are the ones that already exist in the training data. I.e., 81.87% of the training data corresponds to FALSE evaluated as churn. 18.12% of the training data corresponds to non-churn probability assessed as 1. These probabilities are the same among models.

The second thing should be checked is the Group means, which are the average of each predictor within each class. These values could suggest that the variable BonusItemsGrabbed might have a slightly higher influence on non-churn (TRUE) that is 1.686 than on churn (FALSE) that is 0.0224. This situation also happens with the variable MoveCount, StarsEarned, TimesLevelReset in the rest models.

The calculated coefficient for BonusItemsGrabbed in the first model is 1.239433. The following formula will specify the boundary between the two different classes:

\[ y = 1.239433 \times \text{BonusItemsGrabbed} \]

It can be represented as Figure 1 (x represents the variable BonusItemsGrabbed). Churn of FALSE or non-churn of TURE will be predicted depending on which side of the line they are.
The second model contains two dependent variables, BonusItemsGrabbed and MoveCount so that this formula will delimit the boundary between classes:

\[ y = 1.239433 * \text{BonusItemsGrabbed} + 0.140564 * \text{MoveCount} \]

This formula represents a plane as Figure 2 (x1 represents BonusItemsGrabbed, and x2 represents MoveCount). As I illustrated in the previous model, Figure 2 represents the difference between churn and a non-churn one.
In this second model, the BonusItemsGrabbed coefficient is much greater than the MoveCount coefficient, suggesting that the former variable has a greater influence on the churn probability than the latter variable.

Following the example of the first two models, the last model illustrates the influence generated by all four variable to churn probability.

Summarizing the Table 8, the StarsEarned has the most significant coefficient, BonusItemsGrabbed has the second largest coefficient, MoveCount has a minor number of coefficient and TimesLevelReset has the negative value on coefficient. The result suggests that the variable StarsEarned has a more significant influence on the churn behavior than any other ones.

M. Network Analysis

The Network Analysis is a theory that utilized to depict the relationship of a network. It is a study of graphs as a representation of either symmetric or asymmetric relations between objects. Conventionally, the application of the network analysis is on social networks. However, this study tries to implement the network theory to describe the relationship between each feature in the game, which is a new attempt. Figure 3 shows the result of Network Analysis by choosing a small set of data. The R analysis scripts stores in Appendix.

Figure 3:
Ten rows randomly from the dataset processed to plot an essential network relationship between each other. The number labeled on each node represents one single player’s behaviors including the bonus items grabbed, the star earned, points earned and the level finished. With no direction on the plotted model, the edge between each node represent the similarities, the connections in our case, between each player. Therefore, the random ten players at this moment have 4, 10, 19, 80, 16, 24 and 50 having the same status as player 0, churned. Players 1, 34, 2, 60 are still playing.

When applied all rows of data with direction according to the method used above, the plot is displaying as Table 9.

Table 9:
Taking the bottom right corner as an example, user 30, user 21 and user 1 shows the relationship underlying between them. The directed arrow shows that user 30 and user 21 has similarities in one of the features, and they are intending to grow like user 1. Therefore, if user 1 is churned in the current trace, the user 30 and user 21 intends to churn in next trace.

Below, the network analyzed in detailed.

```r
### basic functions
vcount(mis)  # list table of vertices (nodes)
## [1] 12

ecount(mis)  # list table of edges
## [1] 10

igraph::degree(mis)  # list table of degree values
## 60 80 19 10 4 24 34 59 16 2 0 1
## 1 1 1 1 1 1 2 1 1 2 7 1

#igraph::betweenness(mis)  # list table of betweenness values
igraph::closeness(mis)  # list table of closeness values

## Warning in igraph::closeness(mis): At centrality.c:2784 :closeness centrality is well-defined for disconnected graphs
## 60 80 19 10 4 24 34 59 16 2 0 1
## 0.009803922 0.016393443 0.016393443 0.016393443 0.016393443 0.016393443 0.016393443
```
##          34          59          16           2           0           1
## 0.010000000 0.016393443 0.016393443 0.010000000 0.018181818 0.009803922

V(mis)  # get vector with all vertices (nodes)
## + 12/12 vertices, named, from 249bcc1:
## [1] 60 80 19 10 4 24 34 59 16 2 0 1

E(mis)  # get vector with all edges
## + 10/10 edges from 249bcc1 (vertex names):
## [1] 60--2 80--0 19--0 10--0 4--0 24--0 34--2 34--1 59--0 16--0

###attach centralities as attributes to data the graph, normalized to 0-1
V(mis)$degree <- igraph::degree(mis, normalized = TRUE)
#V(mis)$betweenness <- igraph::betweenness(mis, normalized = TRUE)
#V(mis)$closeness <- igraph::closeness(mis, normalized = TRUE)

### set default options for layout
igraph.options(layout=layout.kamada.kawai,
vertex.label = AllTraceW$i..SessionID,
vertex.size = 25,
vertex.frame.color = NA,
vertex.color = "lightblue")

### set node size to a degree
V(mis)$size <- V(mis)$degree * 25  # assign size to closeness, adjust value for good size
#V(mis)$color <- V(mis)$betweenness * ncol  # assign color to betweenness, scaled to number of colors
#V(mis)$color <- V(mis)$closeness  * ncol  # assign color to closeness, scaled to number of colors

E(mis)$weight <- edge.betweenness(mis) * 0.000000003
E(mis)$width <- edge.connectivity(mis)
plot(mis)

#The following plots are the examples of manipulated aesthetics.

The Network map shows in different shapes every single time of execution. Further plots display as Figure 10 and Figure 11.

Figure 10: Figure 11:
N. BBN Map

Since to construct a Bayesian Belief Network needs to compute the probability before plotting, R Studio has been used to calculate the probabilities of each circumstance in each feature prior the map construction. The calculated results will be put in GeNIe.

```
# BBN
Prob <- function(x) {
  y <- unique(x)
  for (i in y) {
    result <- colSums(as.data.frame(x) == i, na.rm = T)/length(x)*100
    print(paste(i, result))
  }
}
Prob(AllTraceW$BonusItemsGrabbed)
## [1] "0 82.4440619621343"
## [1] "1 7.17154331612163"
## [1] "3 3.6144578313253"
## [1] "4 0.172117039586919"
## [1] "2 6.42570281124498"
## [1] "5 0.172117039586919"
Prob(AllTraceW$MoveCount)
## [1] "0 80.2639127940333"
## [1] "1 3.15547905909352"
## [1] "4 0.109007458405049"
## [1] "7 0.229489386115892"
## [1] "3 6.42570281124498"
## [1] "43 0.114744693057946"
## [1] "2 1.89328743545611"
```
After defining each node by the statistics computed from R, I construct the relation map in GeNle, a BBN constructor. Figure 12 shows the BBN causality network plot of Wuzzit Trouble.
generated by GeNIe. Every square shape node represents one “chance”, which refers to one variable/feature in the analysis. Particularly, the churn node is created as a latent variable representing the circumstances of “left” and “playing”. The causality map enables the readers to capture the underlay relationship of how the features effects the churn behavior as well as how the features effect each other in a simple visualization. Unconventionally, the arrow on every edge points to the cause not effect. For instance, the churn node is affected by PointsEarned, StarsEarned, BonusItemGrabbed and MoveCount. The direction of the edges and the nodes are constructed according to the relationship between features founded in Discriminant Analysis.

Figure 12:
To explain the map in detail, each square contains percentile grams illustrated the information of each feature. For instance, the BonusItemsGrabbed variable with 82% of players grabbed 0 items, has directly contributed to Churn for 82%, meanwhile, directly contributed to TimesLevelReset and StarsEarned feature as well. If any value in BonusItemsGrabbed changed to control the diagram, the corresponding changes in the other features who has connected to it will happen. This result also has been proved in the discriminant analysis.
**Conclusion and Discussion**

This paper has compared previous studies in churn analysis that focused on churn definition, feature selection and the classification method. The comparison of those previous study is used in the construction of the research design on churn analysis in games, taking Wuzzit Trouble as the research data sample. The research design contains two versions that display the progress of the unmatured initiative and the revision of the previous studies of churn prediction in game and telecommunication industries. The main difference between the preliminary design and the revision can be concluded as the workflow and the research method. The preliminary work initiates the workflow analysis, while the revision focused on how to implement the analysis detailed. The revision also offered a new graphical method, network analysis, to visualize the relationship between the features and to provide a theory to draw the relationship line in BNN causality map. More specifically, the revision replenished the first workflow on data set preparation, and graphical model construction.

As a proof of the revised research design, the Wuzzit Trouble dataset is implemented into discriminant analysis, collinearity test, network analysis and Bayesian Belief Network map. The results show the applicability of utilizing the statistical graphics model to depict the relationship between each player behavior in the game and to explain how these behaviors causing the player to churn, even the size of the sample data, 83, limit the accuracy of the implementation.

Lacking data is the biggest limitation of this study, which indirectly limit the setting of the research goal. The churn analysis on Wuzzit Trouble is hard to conduct even when the real-time data is available due to the data gathering permissions. The information gathered in Wuzzit trouble is limited by the scope due to the specific demographic that the game is targeting. The sample data exclude the identity feature of each user, with only the login time, IP address, Game version
remained. Since there is no precise association between the login session and the users, the sample data in preliminary research is using the login time, IP address, Device type and Game version to infer the boundary between different users. This will potentially cause errors on separating different users for analyzing if any one user churns within a period.

The future application of this research design can be used on enhancing data collection, detail churn percentage prediction and testing effort. Since the research on churn analysis targets human/players behavior, to further splitting the data into the types of features requires the data to be more extensive and specific, for instance, on temporal and spatial aspects. Furthermore, the accuracy of player churning percentage provided in this paper is limited by data size, and without testimony the result is relative rough given the initial purpose of this paper is to provide a research design in games. Thus, future research work in this area should focus on offering a model that is based on a larger dataset and under a streaming environment which provide visibility of churners with a more accurate dataset.
References


http://www.statisticssolutions.com/discriminant-analysis/


Israel H, Richter RR. (2018). *A guide to understanding meta-analysis*. Saint Louis University, Department of Orthopaedic Surgery, St Louis, MO 63104, USA.


http://libguides.usc.edu/writingguide/researchdesigns


https://en.wikipedia.org/wiki/Weibull_distribution
Appendix

- Game Overview

Screenshot 1: Two initial levels: Botany Lab and Invention Room

Screenshot 2: 25 stages in each level
The game works on a gear that needs the player to increment the number to get the keys and special items.

Screenshot 3: Game mechanics
Keys and items affect the points that the player can earn. Fewer movements that she/he made, more stars she/he can get.

Screenshot 4: Value of stars and points

- Raw data
Churn Binary Matrix
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>T0</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>T1</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>T2</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>5</td>
<td>T3</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>T4</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>T5</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>8</td>
<td>T6</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>9</td>
<td>T7</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>10</td>
<td>T8</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>11</td>
<td>T9</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>12</td>
<td>T10</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>13</td>
<td>T11</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>14</td>
<td>T12</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>15</td>
<td>T13</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>16</td>
<td>T14</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>17</td>
<td>T15</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>18</td>
<td>T16</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>T17</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>20</td>
<td>T18</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>21</td>
<td>T19</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>22</td>
<td>T20</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CF</th>
<th>CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHURNRATE</td>
<td>82</td>
</tr>
<tr>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>31.3253</td>
<td>T</td>
</tr>
<tr>
<td>49.39759</td>
<td>T</td>
</tr>
<tr>
<td>62.6506</td>
<td>T</td>
</tr>
<tr>
<td>67.46988</td>
<td>T</td>
</tr>
<tr>
<td>74.6988</td>
<td>F</td>
</tr>
<tr>
<td>81.92771</td>
<td>F</td>
</tr>
<tr>
<td>89.15663</td>
<td>F</td>
</tr>
<tr>
<td>92.77108</td>
<td>F</td>
</tr>
<tr>
<td>92.7108</td>
<td>F</td>
</tr>
<tr>
<td>96.38554</td>
<td>F</td>
</tr>
<tr>
<td>97.59036</td>
<td>F</td>
</tr>
<tr>
<td>97.59036</td>
<td>F</td>
</tr>
<tr>
<td>97.59036</td>
<td>F</td>
</tr>
<tr>
<td>97.59036</td>
<td>F</td>
</tr>
<tr>
<td>98.79518</td>
<td>F</td>
</tr>
<tr>
<td>98.79518</td>
<td>F</td>
</tr>
<tr>
<td>98.79518</td>
<td>F</td>
</tr>
</tbody>
</table>

- Studio 3T Processing
R markdown file for data processing

Wuzzit

Qi Wang

April 4, 2018

```r
#Load the data exported from MongoDB
wuzzit<- read.csv("W.csv", header = TRUE, sep = ",")

#Load data removed moves and extract the churn slot
wuzzitTrace<- read.csv("wuzzitTrace.csv")

#Transpose
ChurnBinary<- read.csv("binaryChurn.csv", row.names = 1)
Transposed<- t.data.frame(ChurnBinary)
#write.csv(Transposed, "WuzzitTrans.csv")
#RateList<-list()
#  for (i in 0:21){
#    rate<-rowSums(Transposed[i,]==0)/21*100
#    RateList<-list.append(Ratelist,rate)
#  }
#WuzzitTrace<-data.frame(cbind(wuzzitTrace, ChurnBinary$ChurnRate))
#Trace0<-data.frame((cbind(WuzzitTrace$SessionID, WuzzitTrace[9:15])))

# Linear Discriminant Analysis with Jacknifed Prediction
```
library(MASS)

AllTraceW <- read.csv("AllTraceW.csv")
AllTraceW[is.na(AllTraceW)] <- 0

lda<-lda(Churn ~ BonusItemsGrabbed + MoveCount + StarsEarned + TimesLevelReset,
    data=AllTraceW)

# Assess the accuracy of the prediction
# percent correct for each category of G
#ctW <- table(AllTraceW$Churn, fitAll$class)
#diag(prop.table(ctW, 1))
# total percent correct
#sum(diag(prop.table(ctW)))

# Normalization
NormPoints <- as.data.frame(scale(AllTraceW$PointsEarned))
NormBonus <- as.data.frame(scale(AllTraceW$BonusItemsGrabbed))
NormMoves <- as.data.frame(scale(AllTraceW$MoveCount))
NormStars <- as.data.frame(scale(AllTraceW$StarsEarned))
NormTraceW <- cbind.data.frame(points=NormPoints$V1, bonus=NormBonus$V1, move=
    NormMoves$V1, stars=NormStars$V1, churn=AllTraceW$Churn)

# Linear Discriminant Analysis with Jacknifed Prediction
library(MASS)
fitNor <- lda(churn ~ points + bonus + moves + stars,
    data=NormTraceW, CV=TRUE)

# Assess the accuracy of the prediction
# percent correct for each category of G
#ct1 <- table(NormTrace0$churn, fitNor$class)
#diag(prop.table(ct1, 1))
# total percent correct
#sum(diag(prop.table(ct1)))

# Collinearity test
library(car)
## Loading required package: carData

```r
lm <- lm(Churn ~ PointsEarned + BonusItemsGrabbed + MoveCount + StarsEarned +
         TimesLevelReset,
         data=AllTraceW)

vif<-vif(lm)
print(vif)
```

```
##      PointsEarned BonusItemsGrabbed         MoveCount       StarsEarned
##         10.976917          5.768109          1.609131          4.426368
##   TimesLevelReset
##          1.061502
```

## Network Analysis

```r
library(igraph)

## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##     decompose, spectrum

## The following object is masked from 'package:base':
##     union

library(GGally)
library(sna)

## Loading required package: statnet.common

## Attaching package: 'statnet.common'

## The following object is masked from 'package:base':
##     order

## Loading required package: network

## network: Classes for Relational Data
## Version 1.13.0 created on 2015-08-31.
## copyright (c) 2005, Carter T. Butts, University of California-Irvine
##         Mark S. Handcock, University of California -- Los Angeles
##         David R. Hunter, Penn State University
##         Martina Morris, University of Washington
##         Skye Bender-deMoll, University of Washington
```
lesmis <- AllTraceW[sample(nrow(AllTraceW), 10), ]

## parse the data frame as undirected graph
mis <- graph.data.frame(lesmis, directed = FALSE)
plot(mis, label=TRUE)  # render the igraph object
### basic functions

\texttt{vcount(mis)} \quad \# list table of vertices (nodes)

## [1] 12

\texttt{ecount(mis)} \quad \# list table of edges

## [1] 10

\texttt{igraph::degree(mis)} \quad \# list table of degree values

## 60 80 19 10 4 24 34 59 16 2 0 1
## 1 1 1 1 1 1 2 1 1 2 7 1

\texttt{igraph::betweenness(mis)} \quad \# list table of betweenness values

\texttt{igraph::closeness(mis)} \quad \# list table of closeness values

\texttt{Warning in igraph::closeness(mis): At centrality.c:2784 : closeness cannot be computed for disconnected graphs}

## 60 80 19 10 4 24 34 59 16 2 0 1
## 0.009803922 0.016393443 0.016393443 0.016393443 0.016393443 0.016393443
## 34 59 16 2 0 1
## 0.010000000 0.016393443 0.016393443 0.010000000 0.018181818 0.009803922

\texttt{V(mis)} \quad \# get vector with all vertices (nodes)

## + 12/12 vertices, named, from 249bcc1:
## [1] 60 80 19 10 4 24 34 59 16 2 0 1
E(mis)  # get vector with all edges
## + 10/10 edges from 249bcc1 (vertex names):
## [1] 60--2 80--0 19--0 10--0 4 --0 24--0 34--2 34--1 59--0 16--0
### attach centralities as attributes to data the graph, normalized to 0-1
V(mis)$degree <- igraph::degree(mis, normalized = TRUE)
#V(mis)$betweenness <- igraph::betweenness(mis, normalized = TRUE)
#V(mis)$closeness <- igraph::closeness(mis, normalized = TRUE)
### set default options for layout
igraph.options(layout=layout.kamada.kawai,
                vertex.label = AllTrace$SessionID,
                vertex.size = 25,
                vertex.frame.color = NA,
                vertex.color = "lightblue")
### set node size to degree
V(mis)$size <- V(mis)$degree * 25  # assign size to closeness, adjust value
#V(mis)$color <- V(mis)$betweenness * ncol  # assign color to betweenness, scaled to number of colors
#V(mis)$color <- V(mis)$closeness * ncol  # assign color to closeness, scaled to number of colors
E(mis)$weight <- edge.betweenness(mis) * 0.000000003
E(mis)$width <- edge.connectivity(mis)
plot(mis)
```r
# BBN

Prob <- function(x) {
  y <- unique(x)
  for (i in y) {
    result <- colSums(as.data.frame(x) == i, na.rm = T) / length(x) * 100
    print(paste(i, result))
  }
}

Prob(AllTraceW$BonusItemsGrabbed)
## [1] "0 82.4440619621343"
## [1] "1 7.17154331612163"
## [1] "3 3.6144578313253"
## [1] "4 0.172117039586919"
## [1] "2 6.42570281124498"
## [1] "5 0.172117039586919"

Prob(AllTraceW$MoveCount)
## [1] "0 80.2639127940333"
## [1] "1 3.15547905909352"
## [1] "4 1.09007458405049"
## [1] "7 0.229489386115892"
## [1] "3 6.42570281124498"
## [1] "43 0.114744693057946"
```