

**Computational Analysis of Churn in Multiplayer Online  
Games**

**A THESIS**

**SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL  
OF THE UNIVERSITY OF MINNESOTA**

**Zoheb Hassan Borbora**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
Doctor of Philosophy**

**Jaideep Srivastava**

**May, 2015**

© Zoheb Hassan Borbora 2015  
ALL RIGHTS RESERVED

# Acknowledgements

There are a number of people that I would like to thank without whose help this endeavor would not have been possible. First and foremost, I would like to thank my advisor and mentor Professor Jaideep Srivastava for his unwavering guidance, encouragement and support through this entire effort. I have learned many valuable life long lessons under his tutelage. I would like to thank Professor Ted Christ for believing in me and supporting me throughout this time. I would like to thank Professor Joseph Konstan and Rui Kuang for being part of my committee and for providing valuable feedback to better my work.

I would also like to thank all of my labmates for their collaborative efforts and their support. In particular, I would like to thank my senior colleague Muhammad Aurangzeb Ahmad for his guidance and collaborations during my early years in the PhD program.

I would like to thank my parents (Safiqur Rahman Borbora and Ashrafa Rasool Borbora), my brother (Zahid Hassan Borbora), my sister-in-law (Nahida Zahir) and my friends (Varun Mithal, Nipun Garg, Surabhi Mithal, Anurag Kumar and Jehwan Oh) for their love and support. Finally, I would like to thank Zainab Naz Hazarika for her love and support and for being patient with me all these years.

# Dedication

Dedicated to my parents..

To my father, Dr. Safiqur Rahman Borbora, his ideals of perseverance, hard work, kindness and service to others have been the guiding principles of my life.

To my mother, Ashrafa Rasool Borbora, for all of her sacrifices and for bestowing on me her unbounded love and blessings throughout my life.

## Abstract

Churn refers to loss of customers and understanding churn behavior and being able to accurately predict likely churners is important for any business as it directly affects the customer base and thus revenue. Analysis of churn behavior is also important in terms of understanding factors of user engagement. As such, churn behavior has been studied across a wide range of industries such as telecom, banking and online social networks. However, most of existing churn research have focused on modeling individual churn behavior and the type of questions has also been limited by the types of datasets which are available to researchers. In this thesis, different aspects of churn in a Massively Multiplayer Online Role Playing Games (MMORPGs) are studied in depth.

MMORPGs are persistent virtual environments that mimic complex physical spaces and many of the behaviors which are observed in the real world are also observed in MMORPGs. Millions of players interact in an online manner in these environments and the game logs capture player activities in great detail. We first use a behavior modeling approach to analyze the player's behavior leading up to the point of churn and discover key indicators or behavioral trends which can help identify players who are going to churn. We do an extensive evaluation and comparison of two types of churn - *Cancellation of Subscription* and *Dormancy*, using this approach. MMORPG environments are characterized by collaboration among players to achieve common goals in activities such as raids and group quests. We identify player communities which evolve over time in such game environments and extend the lifecycle -based approach to build models for predicting churn of these dynamically evolving communities.

Models of player motivation seek to identify factors that motivate player behavior and can be helpful in analyzing and predicting churn behavior. We study the impact of different *achievement* and *socialization*-based player motivational factors on player churn. Specifically, we are interested in studying how socialization serves to increase player engagement and decrease churn.

Contagion processes arise broadly in the social and biological sciences and can be seen in, for example, the spread of infectious diseases, the diffusion of innovations, dissemination of religious doctrine and information diffusion in online social networks.

As per theories of social contagion, behavior and emotions can be transmitted between individuals in a population. We study the relationship between player churn and social contagion i.e when a player leaves a network, what is the impact on its immediate neighborhood. All of the existing churn research have focused on factors which lead to churn. We study the interpersonal effects which can cause spread of churn behavior in a network as well as the factors which keep a player in the network after his neighbor has churned.

# Contents

<b>Acknowledgements</b>	<b>i</b>
<b>Dedication</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Literature Review . . . . .	3
1.2.1 User Churn . . . . .	3
1.2.2 Social Science Background . . . . .	6
1.2.3 Network Diffusion and Dynamic Networks . . . . .	11
1.2.4 Game Design and Flow Experiences . . . . .	16
1.3 MMOPRGs as Testbeds for Study of Churn . . . . .	17
1.3.1 User Churn in MMOPRGs . . . . .	17
<b>2 User Behavior Modelling Approach for Churn Prediction</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 Background and Related Work . . . . .	21
2.3 Methodology . . . . .	22
2.3.1 Player Lifecycle Analysis . . . . .	23

2.3.2	Time-series clustering . . . . .	24
2.3.3	Semantic Dimensions and Supervised Approach . . . . .	25
2.3.4	Hybrid Approach ( <i>wClusterDist</i> ) - Classification Based on Weighted Distance from Labeled Clusters . . . . .	26
2.4	Experimental results and analysis . . . . .	29
2.4.1	Data Description . . . . .	29
2.4.2	Experimental Setup - Cancellation . . . . .	30
2.4.3	Analysis of Player Profiles - Cancellation . . . . .	31
2.4.4	Classification using Supervised and Hybrid Approaches - Cancel- lation . . . . .	35
2.4.5	Evaluation of <i>wClusterDist</i> - Cancellation . . . . .	38
2.4.6	Experimental Setup - Dormancy . . . . .	40
2.4.7	Analysis of Player Profiles - Dormancy . . . . .	40
2.4.8	Classification using Supervised and Hybrid Approaches - Dormancy	43
2.4.9	Evaluation of <i>wClusterDist</i> - Dormancy . . . . .	45
2.5	Application Impact . . . . .	46
2.5.1	Lift Analysis - Cancellation . . . . .	47
2.5.2	Lift Analysis - Dormancy . . . . .	49
2.6	Discussion . . . . .	50
2.6.1	Comparison of Cancellation and Dormancy . . . . .	50
2.6.2	Other Applications . . . . .	54
2.7	Conclusion and Future Work . . . . .	54
<b>3</b>	<b>Community Churn Prediction in Online Games Using Supervised Learn- ing</b>	<b>56</b>
3.1	Introduction . . . . .	56
3.2	Background and Related Work . . . . .	57
3.3	Methodology . . . . .	59
3.3.1	Choice of Algorithms and Parameters . . . . .	64
3.4	Experimental Results and Analysis . . . . .	65
3.4.1	Data Description . . . . .	65
3.4.2	Identification of persistent groups . . . . .	66



3.4.3	Network Construction . . . . .	67
3.4.4	Community detection . . . . .	69
3.4.5	Community Evolution . . . . .	70
3.4.6	Community Churn Prediction . . . . .	71
3.5	Conclusion and Future Work . . . . .	73
<b>4</b>	<b>Impact of Achievement and Socialization Factors on Individual Churn</b>	<b>76</b>
4.1	Data Description and Experimental Setup . . . . .	78
4.2	Feature Space . . . . .	81
4.2.1	Weekly Activity Metrics . . . . .	81
4.2.2	Trend Metrics or Semantic Dimensions . . . . .	83
4.3	Experimental Results and Analysis . . . . .	85
4.3.1	Analysis of Behavioral Profiles . . . . .	86
4.3.2	Impact of Achievement Factors on Churn . . . . .	89
4.3.3	Impact of Socialization Factors on Churn . . . . .	89
4.4	Conclusion and Future Work . . . . .	93
<b>5</b>	<b>On Player Churn and Social Contagion</b>	<b>94</b>
5.1	Introduction . . . . .	94
5.2	Background and Related Work . . . . .	96
5.3	Methodology . . . . .	98
5.3.1	Classification Labels . . . . .	98
5.3.2	Dyadic Influence of <i>ego</i> on <i>alter's</i> activity behavior . . . . .	99
5.3.3	Neighborhood Influence on <i>alter's</i> activity behavior . . . . .	103
5.4	Experimental Results and Analysis . . . . .	104
5.4.1	Data Description . . . . .	104
5.4.2	Experiment 1 - Dyadic Influence on <i>alter's</i> activity behavior . . . . .	105
5.4.3	Experiment 2 - Neighborhood Influence on <i>alter's</i> activity behavior . . . . .	107
5.5	Conclusion and Future Work . . . . .	108
<b>6</b>	<b>Conclusion</b>	<b>111</b>
	<b>References</b>	<b>113</b>

# List of Tables

2.1	Player Cancellation Dataset . . . . .	31
2.2	Player Cancellation Model: Info-gain ranking of features . . . . .	35
2.3	Player Dormancy Dataset . . . . .	40
2.4	Player Dormancy Model: Info-gain ranking of features . . . . .	44
3.1	Player Groups . . . . .	67
3.2	Weekly networks constructed from <i>persistent groups</i> . . . . .	69
3.3	Community Churn Datasets . . . . .	71
3.4	Community Churn Model: Info-gain ranking of features . . . . .	72
3.5	Community Churn Model: 10-fold cross validation results . . . . .	74
4.1	Model of Player Motivation . . . . .	77
4.2	Loners and Socializers: Weekly Datasets . . . . .	79
4.3	Weekly Achievement and Socialization Metrics . . . . .	82
4.4	Loners and Socializers: Week 12 Dataset . . . . .	85
4.5	Loners and Socializers: Performance of Achievement Model . . . . .	89
4.6	Dataset for testing hypothesis $H'_0$ . . . . .	91
4.7	Performance of Achievement and Socialization Factors . . . . .	92
5.1	Weekly Activity Networks - breakdown by dormancy labels . . . . .	104
5.2	EQII Weekly Group Networks - average degree by dormancy labels . . . . .	104
5.3	Dyadic Influence Dataset . . . . .	105
5.4	Dyadic Influence Model - Info-gain ranking of features . . . . .	106
5.5	Dyadic influence Model - 10-fold cross validation results . . . . .	107
5.6	Neighborhood Influence Dataset . . . . .	107
5.7	Neighborhood Influence Model - Info-gain ranking of features . . . . .	108
5.8	Neighborhood Influence - mean, median and SD of features . . . . .	108

5.9	Neighborhood Influence Model - 10-fold cross validation results . . . . .	109
-----	---	-----

# List of Figures

2.1	Cancellation: Player Lifecycle Scenarios . . . . .	23
2.2	Illustration of wClusterDist: Classification Based on Weighted Distance from Labeled Clusters . . . . .	28
2.3	Within cluster SSE for canceled and active non-cancelled samples . . . . .	32
2.4	Cluster centroids for cancelled training samples . . . . .	33
2.5	Cluster centroids for active non-cancelled training samples . . . . .	33
2.6	Player Cancellation Model: Performance on 1-week test set . . . . .	37
2.7	Player Cancellation: wClusterDist Classifier Performance on 1-week test set . . . . .	39
2.8	Within cluster SSE for dormant and non-dormant samples . . . . .	41
2.9	Cluster centroids for dormant training samples . . . . .	41
2.10	Cluster centroids for non-dormant training samples . . . . .	42
2.11	Player Dormancy Model: Performance on 1-week test set . . . . .	46
2.12	Player Dormancy: wClusterDist classifier performance on 1-week test set	47
2.13	Lift analysis for canceled accounts using J48 decision-tree model . . . . .	48
2.14	Lift analysis for dormant accounts using J48 decision-tree model . . . . .	49
2.15	Player Disengagement Model . . . . .	53
3.1	Supervised Learning Approach to Community Churn Prediction . . . . .	59
3.2	Algorithm for tracking communities over time . . . . .	60
3.3	Group with members A, B, C, D as identified through weeks 1 through n	66
3.4	Network constructed from members of persistent groups - weeks 8 and 9.	68
3.5	Communities for weeks 8 and 9. . . . .	70
3.6	Samples of dynamic communities identified using community tracking algorithm . . . . .	70

4.1	Loners and Socializers: Experimental Setup . . . . .	78
4.2	Loners and Socializers: Data description . . . . .	80
4.3	Loners and Socializers: Proportion of Churners . . . . .	81
4.4	Loner segment - comparison of session activity for churners and non-churners . . . . .	85
4.5	Loner segment - comparison of leveling activity for churners and non-churners . . . . .	86
4.6	Socializer segment - comparison of session activity for churners and non-churners . . . . .	87
4.7	Socializer segment - comparison of leveling activity for churners and non-churners . . . . .	88
4.8	Socializer segment - comparison of grouping activity for churners and non-churners . . . . .	90
4.9	Socializer segment - comparison of interaction with previous week churners	91
5.1	Dormancy - Classification Labels . . . . .	98
5.2	Dyadic Influence of <i>ego</i> on <i>alter</i> - illustrative example . . . . .	99
5.3	Neighborhood Influence on <i>alter</i> - illustrative example . . . . .	103

# Chapter 1

## Introduction

### 1.1 Introduction

Churn refers to loss of customers and churn rate is defined as the total gross number of subscribers who leave the service in the period divided by the average total customers in the period. Understanding churn behavior and being able to accurately predict likely churners is important for any business as it directly affects the customer base and thus revenue. It has also been estimated that it is much cheaper to retain an existing customer than to acquire a new one. For these reasons, churn behavior has been explored across a wide range of industries, particularly in the telecom sector [1, 2, 3, 4, 5, 6], but also in other domains such as banking [7, 8], Internet Service Provider [9], P2P Networks [10], insurance [11], credit card [12] and MMORPGs [13].

Churn may be defined differently across domains. In the telecom sector, churn is defined as customers switching from one carrier to another, the domestic monthly churn rates being 2-3% of the customer base [2]. In the banking sector, a customer could be said to have churned based on whether the customer has a current account in the time period [8]. Churn in digital social networks differ in many ways from traditional subscription-based services. For example, a telecom subscriber is generally bound by a service contract and there is a switching cost associated in moving to another service provider. In online communities, on the other hand, there is a low-entry and exit barrier and activity-based definitions of churn would be more meaningful [14].

This dissertation delves into different aspects of churn behavior in online games.

Chapter 2 proposes a novel player lifecycle based approach for churn prediction. A regular player has a natural lifespan within the game environment where he interacts with the game environment in a sequence of sessions and eventually leaves the game. At that point, the player is declared to have churned from the game and has thus reached the terminal state of the player lifecycle. The motivating idea behind using a lifecycle-based approach for churn prediction is to analyze the player's behavior leading up to the terminal state and discover key indicators or behavioral trends which can help identify players who are moving towards that state. The vast majority of the churn prediction studies in the literature use features which are only cross-sectional in nature i.e only look at a snapshot of the population. As part of the lifecycle-based approach, we use trend metrics that capture the level and rate of user activity before the point of churn to provide a better discriminate churn and non-churn behavior.

We look at two definitions of churn - *Cancellation* and *Dormancy*. *Cancellation* relates to the player stopping to pay - by either pro-actively canceling the subscription or by not renewing the subscription. *Dormancy*, on the other hand, has to do with the player stopping to play the game. Both of these types of churn are indicators of a player getting disengaged (losing interest or getting bored) from the game experience.

Chapter 3 extends the idea of a lifecycle -based approach to build models for predicting the likely churn of a dynamically evolving community. The main contribution of this chapter is a framework for predicting churn of dynamically evolving communities using a supervised learning approach. To the best of our knowledge, no prior work has addressed the specific problem of community churn prediction in online games or proposed a general framework for predicting churn of dynamically evolving communities using a supervised learning approach.

Chapter 4 studies the impact of different *achievement* and *socialization*-based player motivational factors on player churn. We posit that players with greater motivation and involvement are less likely to leave a game and hence player motivation is one of the most important factors that can help in analyzing and predicting churn behavior. Specifically, we are interested in studying how socialization serves to increase player engagement and decrease churn.

Chapter 5 looks at the relationship between player churn and social contagion i.e when a player leaves a network, what is the impact on its immediate neighborhood

or group. To the best of our knowledge, this question has not been studied in the churn literature and we believe this is a novel contribution to the area. Specifically, we address two research questions in this chapter. First, we address the question - When an active node, *ego*, becomes dormant, what is the impact on the activity of *ego*'s immediate neighbor, *alter*, based on *ego*'s characteristics and *ego*'s relationship with *alter*? Second, we address the question - When an active node, *ego*, becomes dormant, what is the impact on the activity of *ego*'s immediate neighbor, *alter*, based on the activity behavior of *alter*'s remaining neighbors? Existing contagion models deal with phenomena such as information diffusion, word of mouth in viral marketing and emotional and behavioral contagion - in these cases, contagion does not change the underlying network. However, we address the problem of player churn - an event which changes the underlying network and propose a supervised learning approach to study the problem.

## 1.2 Literature Review

We first do a literature review of the various aspects of the problem that we propose to study.

### 1.2.1 User Churn

In this section, we do a brief survey of churn research across different domains.

#### Telecom

Churn has been studied extensively in the telecom domain and a wide-array of data mining and statistical techniques have been used for this purpose. Among the earliest studies. Masand et al. described CHAMP (CHurn Analysis, Modeling, and Prediction), an automated system for modeling cellular customer behavior on a large scale [15]. The system was tested on the telecom company's 20 largest customers totaling over 30 million customers and demonstrated the possibility of automatically modeling churn on a large scale and that customized monthly models and scores can be provided on a timely and cost effective basis. Decision Tree models were used in the system which could identify churn factors for several geographic regions.



Mozer et al used logit regression, decision trees, neural networks and boosting for churn prediction and showed that under a wide variety of assumptions concerning the cost of intervention and the retention rate resulting from intervention, using predictive techniques to identify potential churners and offer incentives can yield significant savings to a carrier [2]. Similarly, Hung et al compares neural networks and decision trees and their empirical evaluation shows that data mining techniques can effectively assist telecom service providers to make more accurate churner prediction [5].

Ferreira et al compared the predictive and explanatory power of four families of models: neural networks, decision trees, genetic algorithms and neuro-fuzzy systems [4]. They found that variables related to the pattern of air time consumption by subscribers (such as roaming time, night usage and long distance air time) are decisive in defining churn. The study also demonstrated the importance of careful data representation by showing how enhanced models significantly outperform their simple counterparts.

Advanced data mining techniques have also been explored to study the churn prediction problem. Tsai et al used a hybrid model to combine two different neural network techniques for churn prediction - the techniques being back-propagation artificial neural networks (ANN) and self-organizing maps (SOM) [16]. Experimental results showed that the two hybrid models outperform a single neural network baseline model. Zhao et al have proposed an Improved One-class Support Vector Machine technique for the problem and have shown in their study that the proposed method performs very well compared with other traditional methods, ANN, Decision Tree, and Naive Bayes [17].

Survival analysis techniques have also been used to help telecommunications companies understand which customer will churn and when they will churn [18]. Customer survival and hazard functions were used in this study to estimate the risk of customer churn over time.

Few studies have also incorporated social network analysis techniques to study customer churn in the telecom domain. Dasgupta et al have examined the communication patterns of millions of mobile phone users to analyze the propensity of a subscriber to churn out of a service provider's network depending on the number of friends that have already churned [1]. They propose a spreading activation-based technique that predicts potential churners by examining the current set of churners and their underlying social

network. The results indicate that social relationships play an influential role in affecting churn in the operator’s network. Richter et al have proposed a group-first social networks approach for predicting customer churn in mobile networks [3]. They identify tightly-knit social groups, assign a churn risk score for each group and finally assign an individual churn score to each subscriber based on the churn score of her social group as well as her personal characteristics.

### **Other traditional domains**

In the banking domain, Mutanen et al applied logistic regression techniques to predict customer churn using data from a personal retail banking company [8]. Xie et al proposed an improved balanced random forest model and found the method to improve prediction accuracy compared with other algorithms, such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM) [7]

Huang et al presented a new set of features for *broadband internet* customer churn prediction, based on Henley segments, broadband usage, dial types, the spend of dial-up, line-information, bill and payment information, account information. They used Logistic Regressions, Decision Trees, Multilayer Perceptron, Neural Networks and Support Vector Machines to build churn prediction models. Experimental results showed that the high true churn of 77% with the low false churn rate of 2% can be achieved using the proposed features [9]. In the *insurance* domain, Morik et al used a TF/IDF representation from information retrieval for compiling time-related features of the data set and used Decision Tree, SVM and Naive Bayes models. Experimental results show that these new features lead to superior results in terms of accuracy, precision and recall [11]. Nie et al used logistic regression to build churn prediction models for *credit card* holders [12]

### **Online Games and Social Networks**

Recent work has looked at user churn in online social networks such as discussion boards [14], Q&A forums [19] and mobile social networks [20]. Churn in digital social networks differs in many ways from traditional domains. Karnstedt et al have proposed different activity-based definitions of churn to capture the nuances of such online environments and performed an empirical analysis of different user activity profiles

in a discussion board to explore churn behavior under different conditions [14].

Oentaryo et al proposed a churn prediction approach based on collective classification (CC), which accounts for both the intrinsic and extrinsic factors by utilizing the local features of, and dependencies among, individuals [20]. The method was evaluated in a mobile social networking site to predict chat activity churn and compared to a traditional Support Vector Machine model. Kawale et al have proposed a modified diffusion model that propagates the social influence (which has both a positive and a negative component) of a player and combines it with player engagement factors to predict churn in MMORPGs [13].

### 1.2.2 Social Science Background

We study the impact of socialization on churn of a node, *ego* in a network and conversely, the impact of *ego's* churn on its immediate neighborhood. Furthermore, we study these questions in the context of small groups, which is a key aspect of game play in online games. In this section, we give an overview of different aspects of social science which relate to *social influence* and *group dynamics*.

#### Social Influence and Contagion

Dorwin Cartwright argues that there is no single body of literature on influence but rather a collection of discrete and more or less independent literatures concerned with various aspects of influence such as leadership, attitude change, conformity, persuasion, communication, social learning and socialization [21]. Cartwright identifies three major aspects of the influence process a) the *agent exerting influence*, denoted by O, b) the method of *exerting influence*, and c) the *the agent subjected to influence*, denoted by P. When an agent, O, performs an act resulting in some change in another agent, P, O is said to *influence* P.

Marsden and Friedkin argue that "influence does not even require face-to-face interaction; the only precondition of social influence is information (which allows social comparison) about the attitudes and behavior of other actors [22]. They identify relations of authority, identification, expertise and competition as some of the diverse substantive processes that underlie *influence*. As per the social comparison framework [23], people obtain normative guidance by comparing their attitudes with those

of a reference group and attitudes are reinforced or altered based on this comparison. This is consistent with the argument that the resolution of intra- and interpersonal conflict is the driving force in social influence processes [24]. Social power [25] and control over information [26, 27] have also been considered as bases of social influence. French and Raven identified five sources of social power: the capacity to coerce, the ability to reward, incumbency in a position of legitimacy or authority, recognized expertise and referent power [25]. Festinger proposed that influence between two actors declines as a function of the discrepancy of their opinions or behaviors [28].

Contagion is a type of social influence [29] and in its most general sense, is the spreading of an entity or influence between individuals in a population, via direct or indirect contact. Contagion processes arise broadly in the social and biological sciences [30], seen in, for example the spread of infectious diseases [31, 32, 33] and computer viruses, the diffusion of innovations [34, 35, 36], political upheavals [37] and the dissemination of religious doctrine [38, 39].

Social contagion research can be broken down into two major areas - emotional contagion and behavioral contagion. Emotional contagion can be thought of as "a process in which a person or group influences the emotions or behavior of another person or group through the conscious or unconscious induction of emotion states and behavioral attitudes" [40] and is a process that can occur at both subconscious and conscious levels [41]. A laboratory study of emotional contagion and its influence on work group dynamics found the predicted effect of emotional contagion among group members [42], but the study did not find any hypothesized differences in contagion effects due to the degree of pleasantness of the mood expressed and the energy level with which it was conveyed.

Polansky et al. operationally defined *behavioral contagion* as an event in which a recipient's behavior has changed to become "more like" that of the actor or initiator and the change has occurred in a social interaction in which the actor has not communicated intent to evoke such a change [43]. In order to distinguish contagion from other types of social influence (such as conformity), Wheeler [44] defined behavioral contagion in terms of reduction statements, labeled as  $S(1)$  -

*S(1). If the set of test conditions  $T_1$  exists, then contagion has occurred if and only if Person X (the observer) performs behaviour N ( $B_N$ ) where  $T_1$  is specified as follows:*

a) A set of operations has been performed on Person X which is known to produce instigation toward  $B_N$  in members of the class to which X belongs: b)  $B_N$  exists in the response repertoire of X, and there are no physical restraints or barriers to prevent the performance of  $B_N$ ; c) X is not performing  $B_N$ ; d) X observes the performance of  $B_N$  by Person Y (the model)." (p. 180)

Wheeler reviews empirical research dealing with contagion [43, 45, 46, 47, 48, 49, 50] and derives theoretical statements from the review. Wheeler's central theoretical statement [44] regarding behavioral statement is that -

*Behavioral contagion as defined in S (1) is mediated by the lowering of the observer's avoidance gradient in an approach-avoidance conflict*

Wheeler thus defines behavioral contagion in terms of Lewin's classical socio-psychological theory of approach-avoidance conflict. Behavioral contagion research can itself be broken down into six broad areas, based on the nature of the behaviour that is spread [51]; hysterical contagions [52, 53], deliberate self-harm contagions [54, 55], contagions of aggression [46, 56], rule violation contagions such as teenage smoking [57], consumer behaviour contagions [58], and financial contagions [59].

Burt [60] examined two factors which can bring about social contagion from a network theory perspective - *cohesion* and *structural equivalence*. The cohesion model focuses on socialization between ego and alter and makes the argument that the more frequent and empathic the communication is between ego and alter, the more likely that alter's adoption will trigger ego's. The structural equivalence mode, on the other hand, highlights competition between ego and alter and argues that the more intense that ego's feelings of competition with alter are the more likely it is that ego will quickly adopt any innovation perceived to make alter more attractive as the object or source of relation. Results and analysis on the *Medical Innovation* data showed that a) product adoption was strongly determined by a physician's personal preferences and b) when contagion was involved in the diffusion of the product, its effect was through structural equivalence rather than cohesion.

### **Group Dynamics and Group Cohesion**

Dorwin Cartwright and Alvin Zander defined group dynamics as a "field of inquiry dedicated to advancing knowledge about the nature of groups, the laws of their development,

and their interrelations with individuals, other groups, and larger institutions” [61]. Group cohesiveness has always played a central role in theories of group dynamics [62]. In addition, cohesion has been cited as a contributing factor in various group processes, including conformity [63, 64, 65], productivity [66] and behavior change [67]. The extensive literature on the topic has led to a proliferation of definitions of cohesion which have proved difficult to combine or reconcile [68, 69].

Festinger et al had defined *cohesiveness* as ”the total field of forces which act on members to remain in the group” [70] and then had redefined *cohesiveness* as ”the resultant of all forces acting on the members of a group to remain in the group,” [64]. This represents a subtle but important shift from the causal mechanisms to outcome variable. In a causal system, group cohesion might be defined according to membership duration [70, 71] or it might be defined according to one of the antecedent conditions that affect membership duration, such as a persons intention to remain in the group [61], identification with the group [72], or interpersonal ties [73]. Contemporary analyses of social cohesion treat it either as a multidimensional phenomenon or as a latent construct with multiple indicators [74, 62, 68].

Positive interpersonal ties among persons has long been proposed as a basis of cohesion [61, 73] with the density of interpersonal relations in a group being treated as a group-level measure of cohesion [70]. Gross and Martin defined cohesiveness as ”the resistance of a group to disruptive forces” and proposed that such cohesiveness is associated with the strength of the relational bonds among group members [75]. Lott and Lott [73] noted that ”if we define the cohesiveness of small groups in terms of the positive judgments which members make of one another, then a great deal is known about the conditions under which cohesiveness is likely to develop.”

In general, it can be argued that there is an intricate relationship between group-level cohesion and the attitude and behavior of individual members. Friedkin et al have described the individual membership attitudes and behaviors and group-level conditions, based on which investigators have defined groups as more or less cohesive [69]. They then develop an argument that these group level conditions of cohesion are either derivative properties of the distribution of individual-level indicators of cohesion or causal antecedents of these indicators.

A cohesive group is one in which there is a high level of uniformity in individual

attitudes and behaviors and an important mechanism that brings about attitudinal consensus and behavioral uniformity are endogenous mechanisms of interpersonal influence [76]. The susceptibility of a group member to interpersonal influence and the structural features of a group's social network both contribute to the spread of influence. In terms of network structure, while small networks, high density networks or networks based on strong interpersonal ties can bring about cohesion, other structural characteristics in larger networks can facilitate cohesion as well. Granovetter found that *weak ties* were indispensable to individual's opportunities and their integration to communities while *strong ties* increase local cohesion but can lead to overall fragmentation [77].

Moody and White extended the concept of *structural cohesion* and provide an algorithm for its use in empirical analysis [78]. They define structural cohesion as a measure of the *relational* component of social solidarity and give its operational definition as the minimum number of actors who, if removed from a group, would disconnect the group. They do an empirical study on two different types of networks. In the first study they identify cohesive groups within the friendship network of a high-school and in the second study they apply their method on the interlocking directorate networks of 57 large US firms and show how structural cohesion relates to similar political activity behaviors.

Hogg has discussed the limitations and criticisms of group cohesiveness [79] - the primary criticisms being a) the formulation of group cohesiveness as an attraction-to-group or interpersonal attraction and b) measuring group level phenomena in terms of aggregation. He describes how self-categorization theory, social identity theory and a general intergroup perspective, go some way towards overcoming limitations of the original group cohesiveness concept, and resolving shortcomings of recent reformulations. *Social Identity* is the portion of an individual's self-concept derived from perceived membership in a relevant social group and is a theory that seeks to explain intergroup relations and group processes [80]. *Self-categorization* theory takes a social cognition viewpoint on social cohesion and basically involves an individual identifying oneself as members of a particular in-group and *not* as members of other groups [81]. During self-categorization, people cognitively represent social groups in terms of prototypes, which are a subjective representation of the defining attributes (beliefs, attitudes, behaviors etc) of the group, and associate their personal attitudes and behaviors to these prototypical norms.

### 1.2.3 Network Diffusion and Dynamic Networks

This section gives an overview of computational models and recent data-analytic work in the fields of network diffusion and and dynamic networks.

#### Network Diffusion Models

Models of diffusion by which information, ideas and influence spread through a network has been studied extensively in a number of domains such as diffusion and adoption of innovations [34, 36, 35], spread of infectious diseases in epidemiology [82, 33], the effects of "word of mouth" and "viral marketing" in the promotion of new products [83, 58, 84], trust propagation through networks [85, 86] and more recently, information diffusion in online social networks [87, 88, 89, 90].

The goal of viral marketing is to rely on existing networks of influence among customers to promote a product. Such "word-of-mouth" advertising can be quite economical as compared to traditional direct marketing or mass marketing campaigns. Also, one would expect them to be more effective since the promoters of the product tend to be friends and acquaintances of the target customers in most cases. A classic example of this is the Hotmail free email service, which grew from zero to 12 million users in 18 months on a mere USD 50,000 on traditional marketing, thanks to the inclusion of a promotional message with the service's URL in every email sent using it [91].

The basic premise of viral marketing is that by initially targeting a few influential members of the network, a cascade of influence can be triggered which will cause promotion of the product through the network. Towards that end, Domingos et al first proposed an approach to model a customer's *network value* using Markov random field where the *network value* becomes an estimate of the customer's influence in the network [84]. Since the problem of finding the optimal set of customers to market on is combinatorial in nature, they evaluated greedy search and hill-climbing search approaches to find an approximate solution. In a follow-up work, they extended their technique to handle continuously variable marketing actions and partial network knowledge. The approach is evaluated on data mined from a real-world knowledge sharing site and showed that it scales efficiently to networks of hundreds of millions of customers [92].

Two basic diffusion models have been proposed in the literature



- *Linear Threshold Model* [77]: In this model, a node  $v$  is influenced by each neighbor  $w$  according to a weight  $b_{v,w}$  and each node in the network has a threshold  $\theta_v$  typically drawn from some probability distribution. Under such a setting, the diffusion process then unfolds such that at time step  $t$ , node  $v$  adopts the behavior if the total weight of its neighbors that have already adopted the behavior is greater than the threshold  $\theta_v$

$$\sum_{w \text{ neighbor of } v} b_{v,w} \geq \theta_v \tag{1.1}$$

- *Independent Cascade Model* [93]: In this model, when a node  $v$  adopts a behavior, it has a chance to infect each of its neighbor  $u$  with success probability of  $p_{u,v}$

Kempe et al have studied the problem of choosing influential sets of individuals as a problem in discrete optimization [94]. They show that the *influence maximization* problem is NP-hard for both the linear threshold and independent cascade models. Using an analysis framework based on submodular function they are then able to show that a hill-climbing strategy obtains a solution that is provably within 63% of optimal for both these classes of models. In a follow up work, they also show that the influence maximization problem can be approximated in a very general model they call the *decreasing cascade* model [95]. The basic intuition behind this model is that a contagious node's probability to activate its neighbor  $v$  decreases if more nodes have already tried to *infect*  $v$ .

Leskovec et al did a study where they were able to directly measure and model the effectiveness of recommendations by studying an online retailer's incentivised viral marketing program [96]. It was observed that the recommendation chains are usually short, often terminating with the initial purchase of a product. However, occasionally a product will propagate through a very active recommendation network. They proposed a simple stochastic model that explains the propagation of recommendations and were able to gain several interesting insights into the underlying process. They found that the probability of purchasing a product increases with the number of recommendations received, but quickly saturates to a constant and relatively low probability. For high degree sender nodes, they found that the success rate of recommendations declines as the number of recommendations made increase. Finally, they found that the category and price of product plays a role, with recommendations of expensive products of interest

to small, well connected communities resulting in a purchase more often.

Information diffusion in online social networks has been another area of active research where network diffusion models have been studied a lot. Diffusion of ideas and news through online social networks such as Facebook, Twitter, Youtube and online blogs have played a significant role in major global events such as during the 2010 Arab spring [97] or during the 2008 U.S. presidential elections [98]. Key elements of research in this field have been topic detection [87, 88], modeling information pathways [99, 100, 89] and finding influential members [94] in the network, as discussed by Guille et al in their recent survey [90].

Gruhl et al have studied the dynamics of information diffusion through blogspace by characterizing topic patterns and how topics propagate across individuals [101]. They characterize the patterns of topic postings into *chatter* (internally driven, sustained discussion) and *spikes* (externally induced short-term, high-intensity discussion). Then they characterize four categories of individuals based on their typical posting behavior within the life cycle of a topic and propose and evaluate a model for information diffusion based on the theory of the spread of infectious diseases. Leskovec et al proposed a framework for identifying and clustering time-evolving quotes or phrases that travel through on-line text and show how such an approach can provide a coherent representation of the news cycle [88]. The analysis revealed that typically a phrase is first picked up and quoted with great intensity for a short duration in the mainstream news media and is then handed off to blogs where the item can persist for a longer duration and there is a typical lag of 2.5 hours between the peaks of attention to a phrase in the news media and in blogs.

Kossinets et al have studied the temporal dynamics of communication using online data and proposed the idea of a *network backbone*, which is defined as the subgraph consisting of edges on which information has the potential to flow the quickest [102]. Using this approach, direct connections with low rates of communication can be viewed as being longer than multi-step paths along which information flows more rapidly. In their recent work, Gomez-Rodriguez et al propose an algorithm to provide a time-varying estimate of edges as well as the dynamic edge transmission rates and thus detect evolution of information pathways in a network [89]. Evaluation of the proposed approach on real-world mainstream media and blog sites reveal that information pathways for

general recurrent topics are more stable across time than for on-going news events.

Network diffusion models have also been used in trust propagation models [85, 86, 103]. Ziegler et al proposed a spreading energy-activation model for trust propagation in which energy is assigned to a set of seed nodes and then propagates through active nodes according to a global spreading factor  $d$ . At each trust propagation step, an active node  $X$  transfers a portion of its energy  $d.E(X, i)$  to its neighbors, while retaining  $(1 - d).E(X, i)$  [85]. Gray et al propose a trust-based security architecture for networks that exhibits small world properties such as mobile ad-hoc networks [86]. Guha et al propose a trust propagation scheme wherein given a small number of expressed trusts/distrust in the network, the algorithm can infer trust between any two nodes in the network by using a set of basic trust propagation rules - *direct propagation*, *co-citation*, *transpose trust* and *trust coupling* [103].

The problem of identifying influential nodes in a network is closely related to that of network diffusion since one can maximize the spread of the desired behavior or minimize the spread of undesirable by identifying such nodes. As such, this has been an active area of research in recent years - particularly in the field of blogs and social media [104, 105, 106, 107, 108]. Romero et al proposed an algorithm, IP (Influence Passivity), that assigns an influence score to an actor based on the size and passivity of the influenced audience [104]. The proposed algorithm was used on a 2.5 million user dataset - a key finding of the study was that having a large number of followers does not necessarily mean that the actor is influential if the followers are passive consumers rather than active propagators of information. Weng et al first observed the presence of homophily [109] in the *Twitter* context. Based on that, they propose the *Twitter-Rank* algorithm (a topic-sensitive version of the Page Rank algorithm) to measure the topic-sensitive influence of twitter users, which takes into account both the topical similarity between twitterers and their link structure [106]. Along similar lines, Pal et al proposed a topic-sensitive, feature-centric approach to identify the most authoritative and influential authors for a topic [105]. The approach involves characterizing authors based on nodal and topical metrics, clustering over the feature space and then doing a within-cluster ranking procedure.

## Evolution of Online Networks, Communities and Groups

All real world networks evolve over time. Barabasi and Albert [110] first proposed the *Preferential Attachment* model to explain the scale-free power-law degree distribution observed across diverse networks like genetic networks or the World Wide Web. There are two key ingredients of this model (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected. Similarly, Leskovec et al studied a range of networks from several domains and made two key observations (i) networks become denser over time with the average degree increasing (the number of edges growing super-linearly in the number of nodes) and (ii) the network diameter decreases as the network grows [111]. They propose two probabilistic generative models to capture these properties - the *Community Guided Attachment* model and the *Forest Fire* model.

As has been shown by the above models, most real-world networks follow non-uniform degree distributions with a tail that often follows a power law. In fact, vertices in such networks can often be organized into clusters such that there are many edges joining vertices within the clusters and comparatively fewer edges joining vertices of different clusters - such clusters are called *communities*. Community detection in graphs has been researched extensively across diverse fields such as sociology, biology and computer science [112]. While most of the existing work have focused on detecting communities in static networks [113, 114, 115, 116], the study of dynamic communities is still in a nascent stage.

Palla et al. performed an analysis of dynamic communities in a cell phone network and a collaboration network [117] using their *Clique Percolation Method* [118]. The study revealed a significant difference between the dynamics of small groups and large institutions. Small groups remain stable and persist longer if the core membership/composition of the group remain unchanged. Large groups, on the other hand, persist for longer if there is continuous change in the group size and composition. Berger-Wolf et al have proposed a framework for analysis of dynamic social networks [119, 120] that explicitly makes use of information about when social interactions occur. The key element of their framework is a *metagroup* which is defined with three parameters *persistence*(group duration), *turnover*(member churn rate) and *membership*(group size)

Backstrom et al investigated the membership, growth and evolution of online communities using data from LiveJournal (a friendship network) and DBLP (a co-authorship network) [121]. While the underlying premise of diffusion models is that an individual's probability of adopting a new behavior depends primarily on the number of friends already engaging in the behavior, a key result from Backstrom et al's study is that the adoption probability also depends, to a significant extent, on the internal connectedness of the individual's friends in the target community. Specifically, the results show that that individuals whose friends in a community are linked to one another are significantly more likely to join the community.

#### 1.2.4 Game Design and Flow Experiences

Game development has a strong focus on the individual player experience as evident from the considerable amount of time and money invested in market research and game testing by game publishers and developers. Pagaluyan et al have reviewed the principles and challenges in game design and evaluation and discussed user-centered techniques that address those challenges [122]. Key game design principles they have outlined include identifying the right kind of challenges, addressing different skill levels, appropriate rewards, collecting all items and completing the game, storyline and technological innovations. Some of the subjective attributes they have identified for game evaluation include overall quality (the *fun* factor), ease of use, challenge and pace. Traditional approaches to player-centred game design and evaluation include user studies [123] and heuristic testing [124, 125].

Charles et al described an approach to player-centred game design through adaptive game technologies [126] to address some of the challenges outlined by Pagaluyan et al. A game may be adapted through changes to: i) a player's character, ii) non-player characters in the game and iii) the game environment or game state. Charles et al propose a potential framework for an adaptive game system in which a player's in-game performance is monitored and the game is adapted based on player type and preferences [126].

Ultimately, the goal of player-centric game design is to facilitate pleasurable and/or rewarding experiences for the player through his/her interactions with the game system. The optimal level of experience during work or play is defined as flow [127]. When a

player enters the flow state while playing an online game, this means that he/she is focused on playing the game with no other distraction, has full control over the game and finds the experience to be intrinsically rewarding. A balance between a player's ability level and the challenge of the task at hand can lead to a flow experience. So, one of the key challenges of game design is to maintain the right level of challenge as compared to the player's skill level so that the player is engaged in the game experience.

Cowley et al studied the relationship between player and game, characterized by learning and enjoyment, and tried to come up with a practical mapping of flow onto game-play [128]. Choi et al proposed a theoretical model using the concepts of customer loyalty, flow, personal interaction, and social interaction to explain why people continue to play online games; with the basic premise that if somebody experiences the flow state more often while playing an online game, he/she will have higher customer loyalty to the game. [129].

### 1.3 MMOPRGs as Testbeds for Study of Churn

Massively Multiplayer Online Role-playing Games (MMORPGs) are a popular genre of computer-based game which is characterized by a persistent virtual world maintained by the game developer where millions of players can interact with one another in real time. A player in an MMORPG can control one or more characters through which they interact with the game environment. MMORPGs offer rich social environments where players can engage in a multitude of in-game activities e.g., trading, raiding, exploring, questing etc and also multitude of social interactions e.g., grouping, mentoring, chatting etc. The online activities are captured in great detail in game logs and thus provide us the opportunity to study online behavior in these rich environments.

#### 1.3.1 User Churn in MMOPRGs

MMORPGs constitute a multi-billion dollar industry and revenue sources include player subscription or in the case of free-to-play games, through virtual item sales, subscription tiers and advertisements displayed within the games. For subscription based games, churn can mean the player has *stopped paying* while for free-to-play games churn can mean the player has either *stopped playing* or their activity level has gone abnormally

low. The rich social environments and detailed game logs of MMORPGs allow us to study different aspects of user churn we are interested and for that reason, we use MMORPG game data as the test bed for our experiments and analysis.

## Chapter 2

# User Behavior Modelling Approach for Churn Prediction

### 2.1 Introduction

A massively multiplayer online role-playing game (MMORPG) is a popular genre of computer-based game which is characterized by a persistent virtual world maintained by the game developer. A player in an MMORPG can control one or more characters through which they interact with the game environment. Activities in this environment are mostly driven by individual or group *quests*. *World of Warcraft* is arguably the most popular game in this genre while other game titles from this genre include *Star Wars: The Old Republic* and *Eve Online*. MMORPGs constitute a multi-billion dollar industry and revenue sources include player subscription or in the case of free-to-play games, through virtual item sales, subscription tiers and advertisements displayed within the games. Given the market size and increased competition, customer acquisition and retention is of major concern to gaming companies.

Studies have shown that MMORPG players spend an average of 25.86 hours per week gaming [130] and such online activities are captured in great detail in the game logs. Also, these games have subscriber bases which run into millions of players. Under such a scenario, techniques such as web usage mining [131] can be used to analyze the huge amount of usage logs and build accurate models of user behavior. Such models can capture different aspects of user behavior. In this chapter, we are interested in



modelling user churn behavior and we use a lifecycle-based approach for this purpose.

A regular player has a natural lifespan within the game environment where he interacts with the game environment in a sequence of sessions and eventually leaves the game. At that point, the player is declared to have churned from the game and has thus reached the terminal state of the player lifecycle. The motivating idea behind using a lifecycle-based approach for churn prediction is to analyze the player’s behavior leading up to the terminal state and discover key indicators or behavioral trends which can help identify players who are moving towards that state.

The goal of this chapter is to analyze player churn behavior and build predictive models for churn in an MMORPG. Towards that effect, we look at two definitions of churn - *Cancellation* and *Dormancy*. *Cancellation* relates to the player *stopping to pay* - by either pro-actively canceling the subscription or by not renewing the subscription. *Dormancy*, on the other hand, has to do with the player *stopping to play* the game. Both of these types of churn are indicators of a player getting disengaged (losing interest or getting bored) from the game experience. We address the problem of predicting players who are likely to churn in the week following the date of analysis. *Key contributions* of the chapter are listed below.

First, we propose a player lifecycle-based approach for churn prediction in MMORPGs. The key idea behind this approach is to analyze the activity traits of churners in the weeks leading up to their point of leaving the game and compare it with the activity traits of a regular player. We also propose three semantic dimensions of *engagement*, *enthusiasm* and *persistence* along which the weekly player history of an observed variable is recomputed to give derived features. Experimental results show that models built using this approach have good predictive power and outperform previous models on the same dataset.

Second, we identify several intuitive behavioral profiles for churners and non-churners. The profiles of the two populations were observed to be quite distinct from each other and can thus discriminate between the two classes.

Third, we propose a distance-based classification scheme, which we call *wClusterDist*, which classifies a new point by comparing the sum of its weighted normalized Euclidean distance from two labeled clusterings. The motivating idea behind this scheme is to

benefit from the distinct behavioral profiles of churning and active players. Experimental results showed that the proposed classification scheme is well-suited for this problem formulation and its performance is better than or comparable to other traditional classification schemes.

Fourth, we do an extensive evaluation and comparison of the two types of churn. Based on our analysis, we propose a model of player disengagement and claim that *Dormancy* is a better indicator of player disengagement over *Cancellation* - both in terms of its immediacy to the underlying construct and its universal applicability to any online game, especially to games which do not have a subscription-based model.

The chapter is organized as follows: Section II talks about churn analysis in different domains and other background work. Section III describes in detail the four main components of the lifecycle-based approach that we have used. Section IV describes the dataset, experimental setup, results and analysis for the *Cancellation* and *Dormancy* problems. Section V highlights the application impact of this approach in terms of lift analysis. Section VI has a detailed comparison of the two types of churn with emphasis on their relation to player engagement. Section VII is conclusion and future work.

## 2.2 Background and Related Work

Understanding churn behavior and being able to accurately predict likely churners is important for any business as it directly affects the customer base and thus revenue. For that reason, churn behavior has been explored across a wide range of industries. Traditionally, churn has been analyzed in great detail in the telecom sector [4, 5, 2, 132, 3]. There has also been churn related work in other domains such as retail business [7], banking [133, 8], Internet service providers [9], P2P networks [10], insurance [11], credit card [12] and MMORPGs [134, 13].

Recent work has looked at user churn in social networks such as Q&A forums [19] and discussion boards [135]. Churn in digital social networks differs in many ways from a traditional domain such as telecom [14] and as such, activity-based definitions of churn have been proposed to capture the nuances of such online environments [14]. In addition, the effect of different behavioral and structural features on the users churn likelihood in an online social network has been analyzed [136]. Among other studies,

a modified diffusion model has been used that propagates the social influence (which has both a positive and a negative component) of a player and combines it with player engagement factors to predict churn in MMORPGs [13].

Various machine learning models have also been used for churn analysis including logistic regression models [2, 11, 12, 8], decision tree models [4, 5, 2], neural networks [5, 2] and support vector machines [132, 7, 133]. The churn prediction models may use features which are only cross-sectional in nature i.e only look at a snapshot of the population, or they may include time-related variables [12, 8]. While the cross-sectional variables have sufficient predictive power in most cases, including temporal information can lead to better prediction results as was shown for insurance data [11].

User behavior modeling has been used extensively for web personalization [137, 138] and techniques such as content-based filtering, collaborative filtering and rule-based filtering are commonly used for web personalization [137]. In this chapter, we propose a lifecycle-based approach for modeling and predicting player churn behavior. Also, we propose a distance-based classification scheme which labels a new point by comparing the sum of its weighted normalized Euclidean distance from two labeled clusterings. Topics related to this approach are time-series clustering [139], Gaussian Mixture Models [140] and temporal data mining [141].

In an earlier work, we investigated the problem of churn prediction in MMORPGs from a social science perspective and developed models incorporating theories of player motivation [134]. The goal of this research is to identify and analyze the temporal behavioral profiles of churners and use such insights for building prediction models.

## 2.3 Methodology

The motivating idea behind a lifecycle-based approach for churn prediction is to study activity traits of churners in the weeks leading up to their point of leaving the game and compare it with the activity traits of a regular player. We come up with behavioral profiles from these activity traits and use the insights to propose semantically derived features and a distance-based classification scheme *wClusterDist*.

The following sub-sections describe in detail the main components of the approach.

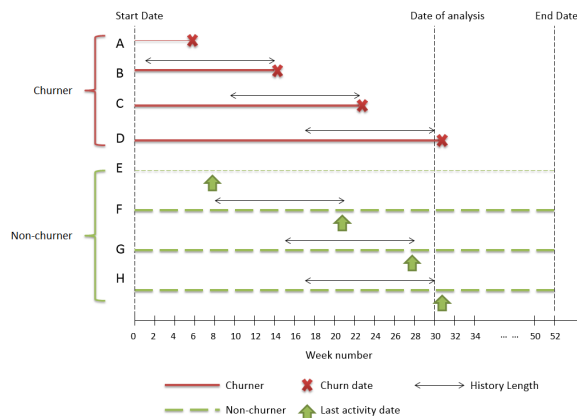


Figure 2.1: Cancellation: Player Lifecycle Scenarios

### 2.3.1 Player Lifecycle Analysis

Figure 2.1 illustrates the different player scenarios that are possible with respect to a date of analysis. This example is for the *Cancellation* type of churn. Let us assume that a year's worth of player logs are available and we are running the analysis, say at the beginning of week number 30. The solid lines denote churners (cases A through E) with the crosses indicating the point in time when they canceled their subscriptions. The dotted lines denote non-churners (cases F through J) with the bolded up-arrow indicating the point of last activity of the non-churners. As illustrated in the figure, there can be different scenarios with respect to the date of analysis for both the player populations.

For the churners, cases A-C represent the scenarios where the player canceled their subscription before the date of analysis, case D is the one where the player unsubscribed in the week following the date of analysis and finally, case E is the one where the player left later on during the game. Depending on their level of activity, non-churners are divided into two categories - active and inactive. We use an *activity threshold* to segment the non-churner population into these categories.

- *Active non-churners* are those whose last activity was after the activity threshold, represented by Cases H-J in Figure 2.1
- *Inactive non-churners* are those whose last activity was before the activity threshold, represented by Cases F and G in Figure 2.1

In player lifecycle analysis, we look at the aligned weekly history of the players for the observed variable(s) of interest. A finer (hourly or daily) or coarser (fortnightly or monthly) level of aggregation could have been chosen but one would expect players to have more of a weekly cycle (e.g they might play more during the weekends). For that reason, we chose our basic time unit of analyses to be a week. The double-headed arrows in Figure 2.1 indicate the length of player history we look at for each of the cases. The players in cases A and F are ignored since they do not have sufficient history. We also ignore the inactive non-churners (case G) in our analysis. For the churners with sufficient history who left before the date of analysis (cases B and C) or in the week following the analysis date (case D), we look at the weekly history leading up to week when they unsubscribed. For the active non-churners with sufficient history, we look at their weekly history leading up to their week of last activity (case H) or till the date of analysis (cases I and J).

For the *Cancellation* churn problem, the objective is to predict the players who are going to cancel their subscriptions in the week following the date of analysis, week 30 for this example. Also, we are primarily interested in the active non-churners and the differences in behavioral profiles of this population from the churners. So, for our classification tasks, positive class would refer to churners and negative class would be active non-churners.

With some minor modifications to the above setup, we can define the *Dormancy* churn problem. The objective here is to predict the players who are going to become dormant (or not play) in the week following the date of analysis, say week 30 for the above example. Since *Dormancy* is based on a behavioral definition which can be easily observed, we can label accounts appropriately for any point of analysis. Thus, we just look at the weekly histories of the dormant and non-dormant accounts leading up to the point of analysis.

### 2.3.2 Time-series clustering

After we have the aligned lifecycle history of the players, we perform time-series clustering of the weekly points using simple K-means. Since the histories are aligned, we can use pair-wise distance measures between two weekly time-series [142]. We use Euclidean distance for our analysis. A key point to note here is that we perform separate

clusterings on the two sets of training samples so that we have one clustering for the positive class (churners) and another clustering for the negative class (non-churners), as illustrated in Figure 2.2.

Analysis of the time-series clusters of churners and non-churners reveal several distinct behavioral profiles of the two population and provide us with key insights into churner behavior.

### 2.3.3 Semantic Dimensions and Supervised Approach

We propose three semantically meaningful dimensions along which the weekly player history of an observed variable, say *number of sessions per week*, should be recomputed to get derived semantic features. These dimensions are -

- *Engagement*: This dimension is intended to capture the engagement level of the player for the observed variable and is computed as the simple average over all weeks.

$$x_{engage} = \frac{1}{N} \sum_{i=1}^N x_i \quad (2.1)$$

where,  $x_i$  is observed value for variable  $x$  in the  $i^{th}$  week  
and,  $N$  is the number of weeks

- *Enthusiasm*: This dimension is intended to capture increase or decrease in the enthusiasm/immersion level of the player and is captured using the magnitude of increase or decrease in the observed behavior over successive weeks. The measure is computed using the sum of the slopes over successive weeks. Since the recency of the changes matter, we use a linear weighting function such that the recent changes get more weightage.

$$x_{enthu} = \sum_{i=1}^N w_i * (x_i - x_{i-1}) \quad (2.2)$$

where,  $x_i$  is observed value for variable  $x$  in the  $i^{th}$  week and  $x_0 = 0$ ;  
 $w_i$  is the week number such that higher-numbered weeks are closer to the date of analysis;  
and,  $N$  is the number of weeks

- *Persistence*: This dimension is intended to capture the general *mood* of the player in the observed past and is captured using the direction (up, down or level) of the change in the observed behavior over successive weeks. The measure is computed using an indicator function  $Ind(x_i - x_{i-1})$  with a linear weighting function so that the recent changes get more weightage.

$$x_{persist} = \sum_{i=1}^N w_i * Ind(x_i - x_{i-1}) \quad (2.3)$$

where,  $x_i$  is the observed value for variable  $x$  in the  $i^{th}$  week and  $x_0 = 0$ ;

$w_i$  is the week number ;

$N$  is the number of weeks;

and,

$$Ind(x_i - x_{i-1}) = \begin{cases} 1, & \text{for } x_i - x_{i-1} > 0 \\ 0, & \text{for } x_i - x_{i-1} = 0 \\ -1, & \text{for } x_i - x_{i-1} < 0 \end{cases} \quad (2.4)$$

$$(2.5)$$

$$(2.6)$$

After we compute the derived features for each of the observed variables, we apply binary classification schemes to distinguish churners from non-churners.

### 2.3.4 Hybrid Approach (*wClusterDist*) - Classification Based on Weighted Distance from Labeled Clusters

We propose a distance-based classification scheme, which we call *wClusterDist*, that classifies a new point by comparing the sum of its weighted normalized Euclidean distance from two labeled clusterings. The motivating idea behind this scheme is to benefit from the distinct behavioral profiles of the two classes (churner and non-churner). Since k-Means is used, each clustering can be thought of as a Gaussian mixture model [140] in which the covariance matrices of the mixture components are identical. Furthermore, each cluster is a D-dimensional multivariate Gaussian where D equals the number of weeks of player history i.e the *History Length* from Figure 2.1.

In the following paragraphs, we describe how we compute the distance of a point  $y$  from the clusters constructed from the samples with label  $l$

We use Mahalanobis distance to compute the the distance of  $y$  from a cluster - this normalizes for different variances along the D dimensions of a cluster [143]. Since

K-means is used, we can assume the covariance matrix of a cluster to be diagonal and thus, the Mahalanobis distance reduces to the normalized Euclidean distance. The normalized Euclidean distance of a new point  $y$  from a cluster  $c$  is given by -

$$euc_{norm}(y, c) = \sqrt{\sum_{i=1}^D \frac{(y_i - \bar{c}_i)^2}{s_i^2}} \quad (2.7)$$

where  $D$  is the number of dimensions

$\bar{c}_i$  is the cluster mean along the  $i^{th}$  dimension

and  $s_i$  is the cluster standard deviation along the  $i^{th}$  dimension

To be noted that a NaiveBayes classifier makes a similar assumption as above of a diagonal covariance matrix when estimating the discriminant function of a class [143].

If we consider the clustering for a particular class label  $l$ , then the weight of the  $k^{th}$  cluster is given by the fraction of all the training samples of that class belonging to the  $k^{th}$  cluster. The cluster weights are nothing but the mixing coefficients of the underlying Gaussian mixture model.

$$\pi_k^l = \frac{N_k^l}{N^l} \quad (2.8)$$

where  $N^l$  is the total number of training samples of class  $l$

and  $N_k^l$  is the number of training samples of class  $l$  in the  $k^{th}$  cluster

Since we want our distance measure to be sensitive to the weight of the cluster, we use the weighted normalized Euclidean distance of a point  $y$  to the  $k^{th}$  cluster of class label  $l$  as given by -

$$wd_{norm}(y, c_k^l) = (1 - \pi_k^l) * euc_{norm}(y, c_k^l) \quad (2.9)$$

where,  $euc_{norm}(y, c_k^l)$  is the normalized Euclidean distance

and  $\pi_k^l$  is the weight of the  $k^{th}$  cluster with class label  $l$

The intuition behind equation 2.9 is that if the Euclidean distance of a point  $y$  to two clusters  $c_1$  and  $c_2$  are the same but  $c_1$  has more number of samples than  $c_2$  (i.e  $\pi_1 > \pi_2$ ), then there is a higher likelihood of the point belonging to cluster  $c_1$  and hence, we treat the point as being closer to  $c_1$  (i.e  $wd_{norm}(y, c_1) < wd_{norm}(y, c_2)$ ).

For a given point  $y$  and class label  $l$  we can thus compute the sum of its weighted normalized Euclidean distance from each of the clusters for that label.



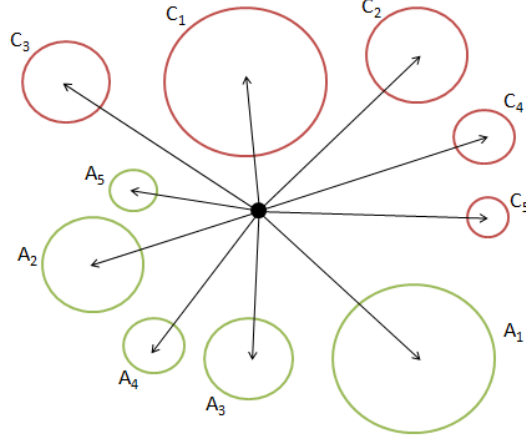


Figure 2.2: Illustration of  $wClusterDist$ : Classification Based on Weighted Distance from Labeled Clusters

$$swd_{norm}(y, c^l) = \sum_{i=1}^K wd_{norm}(y, c_k^l) \quad (2.10)$$

where  $K$  is the number clusters belonging to class label  $l$

Thus, given the clusterings of churners ( $C$ ) and non-churners ( $\bar{C}$ ), we use the above constructs to compute  $swd_{norm}$  of any test point  $y$  from both of these clusterings and label  $y$  based on whichever clustering its closer to, as shown below

Figure 2.2 illustrates the approach with 5 sets of clusters each from the churner ( $C$ ) and non-churner ( $\bar{C}$ ) training samples and the point  $y$  we would like to classify into one of these classes. The arrows denote normalized euclidean distances  $euc_{norm}(y, c_k)$  of  $y$  from each of the clusters and let the cluster weights for a label be such that  $\pi_i^l > \pi_j^l$ ; for  $i > j$ . In this example,  $y$  is closer to the largest cluster of churners  $C_1$  than to the largest cluster of non-churners  $\bar{C}_1$ . So, even though  $y$  is nearer to the smaller non-churner clusters ( $\bar{C}_2$  through  $\bar{C}_5$ ) than to the 4 smaller churner clusters ( $C_2$  through  $C_5$ ), one would expect  $swd_{norm}(y, c^C)$  to be lesser than  $swd_{norm}(y, c^{\bar{C}})$  and  $y$  would thus be labeled as a churner.

Two key advantages of  $wClusterDist$  are -

- Even though this is a distance-based classification scheme, it only needs to compute distances to the labeled cluster centroids when making a classification decision. So it is much faster than instance-based learning schemes such as kNearest-Neighbor which needs to compute distances to every other point in order to make a classification decision.
- The cluster centroids for each class, which represent behavioral profiles, are highly interpretable. Furthermore, when a new test instance is labelled as a churner, the distance functions can be used to make judgements regarding the exact profile of the churner. Both of these pieces of information can be very useful for Customer Relationship Management (CRM) analysts to understand the different populations behind each label.

A *limitation* of the *wClusterDist* is that since it has to first construct the clustering for each label from the training samples, it needs as input the number of clusters to use for each class label.

## 2.4 Experimental results and analysis

### 2.4.1 Data Description

Data from the MMORPG - Sony Everquest II was used for the experiments and analysis. The game had four servers and we used data from all four servers to make sure we were capturing all the activities of an account across the servers. The game activity logs spanned 9 months from January to September, 2006.

In the setup described in Figure 2.1, there are a couple of parameters that need to be chosen for the lifecycle-based analysis: a) *Activity Threshold* and b) *History Length* - the length of weekly player history to look at. We choose the history length to be 13 weeks since that corresponded to a quarterly cycle. The activity threshold was chosen to be 1 month since billing cycles are usually monthly and a 1-month activity threshold would give us all the subscribers who have been active at least once in the last billing cycle. The impact of these parameters on the analysis is discussed in the sub-section *Impact of Experimental Parameters* later on.

We used the session-related weekly history of the accounts. Since the activity logs only record player actions, we had to define player-sessions using a simple heuristic, as used in our previous work [134]. A session consists of sets of activities which are separated by no more than 30 minutes. Using this definition, we collect the following session-related variables for the training and test accounts for their appropriate lifecycle weeks.

- Number of sessions per week
- Sum of session length per week
- Sum of inter-session length per week

#### 2.4.2 Experimental Setup - Cancellation

During the time period of January to September, 2006, 25190 accounts canceled their subscription and never returned. On the other hand, there were 74062 accounts which had some activity during this time period but never canceled their subscription. We picked the date of analysis as Sunday, Aug 27th, 2006. So, our training period was Jan 1st, 2006 to Aug 26th, 2006 and our hold-out test period was the week of Aug 27th-Sep 2nd.

A one month *Activity Threshold* divided the 74062 subscribers into 28057 active and 46005 inactive subscribers, which means 62.12% of the subscribers were inactive at that point (i.e they had no activity after 7/27/2006). This is an interesting finding in itself as it means that a vast number of the subscribers were not actively engaged in the game. At this point, we are not sure of the reason for this seemingly counterintuitive trend but such an investigation could be helpful for game developers both in terms of tracking the engagement level of their player base and identifying any game-specific factors which may lead to player dissatisfaction and premature disengagement.

Table 2.1 gives a breakdown of the samples in the dataset. If we were to relate these numbers to the different player lifecycle scenarios described in Figure 2.1, then the 16119 canceled accounts in the training period are the player samples belonging to cases B and C and the 1008 canceled accounts in the test period are those belonging to case D. There were a total of 28057 active subscribers available (i.e players with last

Table 2.1: Player Cancellation Dataset

Period	Duration	Canceled	Non-canceled	Total samples
Training	Weeks 1-34	16119	25033	41152
Test	Week 35	1008	3024	4032

activity date on or after 7/27/2006) - these are the samples belonging to cases H, I and J. Of these, 16032 of them had a last activity date after the date of analysis. We randomly chose 3024 from these 16032 samples as the active subscriber test samples to get a canceled:non-canceled class distribution of 1:3 in the test set. Finally, we chose the remaining 25033 samples as the active non-canceled training samples. Our goal is to predict players who are going to cancel their subscription on week 35. As observed

### 2.4.3 Analysis of Player Profiles - Cancellation

In this section, we visualize and discuss the results of clustering the canceled and active non-canceled training samples using their weekly history of *number of sessions*. Figure 2.3 shows the decrease in the *Within Cluster Sum of Squared Errors* as the number of clusters is decreased and we observe that the knee of the curve for both the canceled and active non-canceled samples is observed around  $k = 5$ . Therefore, we perform k-means clustering on the canceled and active non-canceled training samples using 5 clusters each as input.

Figures 2.4 and 2.5 show the cluster centroids for the two population training samples, with the x-axis having the aligned 13-week lifecycle history and the y-axis having the number of sessions per week. The difference in behavioral trends between the two samples immediately come to light and in both the cases, we can identify few distinct behavioral profiles. We use *activity* to mean the observed number of sessions in a week

Four behavioral profiles of canceled accounts are observed from Figure 2.4, which are described next. To be noted here that the 13 weekly points for an account represent its weekly history leading up to week when the account canceled its subscription.

- *Canceled Profile A*: These are the canceled accounts who show little to no activity before they cancel their subscription and they constitute the large majority

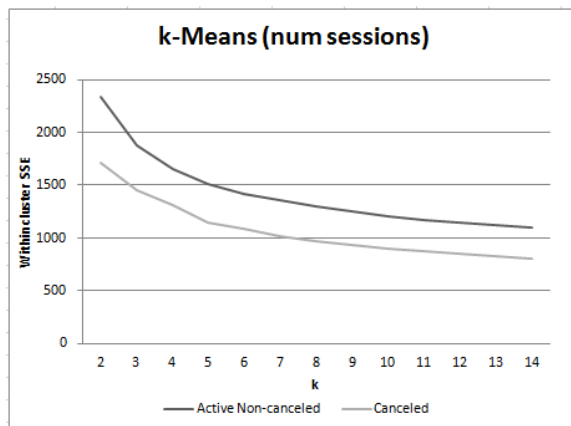


Figure 2.3: Within cluster SSE for canceled and active non-cancelled samples

(75.98%) of the canceled population.

- *Canceled Profile B*: The players fitting this profile show an increase in their activity and then a sharp decrease leading up to the week of cancellation - they constitute around 13.22% of the canceled account samples. Two clusters are observed within this profile -  $B_1$  (6.51%) and  $B_2$  (6.71%). The canceled accounts belonging to the  $B_2$  profile show an increase in their activity but then the activity has a downward slope after around the mid point whereas  $B_1$  profile has canceled accounts who are initially inactive followed by an increasing activity level in the latter half and then a drop in the last week or so.
- *Canceled Profile C*: Around 7.42% of the churners belong to this profile which is characterized by a gradual decrease in activity throughout the 13 week history.
- *Canceled Profile D*: Only a small fraction (3.37%) of the churners fit this profile. The churners belonging to this group have a very high activity level in the first half but as in the case of profile B, there is sharp decrease in their activity leading up to the week of churn.

Four behavioral profiles of active non-cancelled accounts are observed from Figure 2.5. In this case, the 13 weekly points represent the weeks leading up to point of last activity (refer case H in Fig 2.1) or the weeks prior to the week of analysis (refer case I and J in Fig 2.1).

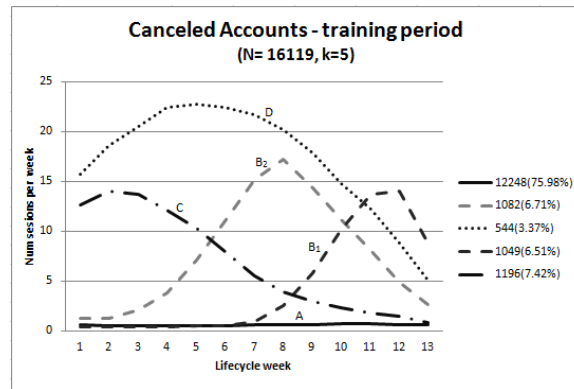


Figure 2.4: Cluster centroids for cancelled training samples

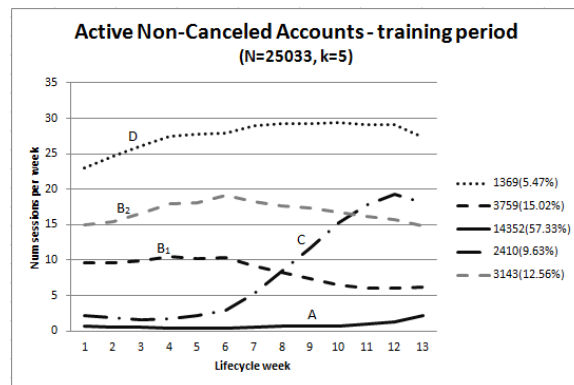


Figure 2.5: Cluster centroids for active non-cancelled training samples

- *Active Non-Canceled Profile A*: The active non-cancelled accounts fitting this profile comprise 57.33% of the active players. This is an interesting behavior profile since it looks very similar to the Canceled profile A in that the players show little to no activity for most of their weekly history. However, they show some minimal activity beyond the activity threshold (refer case H in Fig 2.1) because of which they are picked up as active non-churner samples. The choice of the activity threshold would impact the proportion of samples in this profile since the number of such samples would reduce as the activity threshold is brought closer to the point of analysis (there would be fewer samples whose last activity was within the last week as compared to the number of samples whose last activity was within the last month).

- *Active Non-Canceled Profile B*: The players fitting this profile maintain an almost constant level of activity in the first half followed by a slight and gradual decrease in the activity level in the second half. They comprise 27.58% of the population. Depending on the average level of activity, two sub-profiles are observed in this group -  $B_1$  (15.02%) and  $B_2$  (12.56%). The players in the  $B_1$  sub-profile have a lower mean activity level of around 9 sessions per week throughout the history whereas the  $B_2$  sub-profile has a higher mean activity level of around 16 sessions per week.
- *Active Non-Canceled Profile C*: Around 9.63% of the active non-canceled belong to this profile which is characterized by a gradual increase in activity throughout the 13 week history. To be noted that this active behavioral profile is almost a mirror opposite of the canceled profile C.
- *Active Non-Canceled Profile D*: The players fitting this profile are highly engaged in the game with both a high activity level and a general trend of increasing activity level throughout the 13-week history. They comprise around 5.47% of the active non-churner population.

Overall, if we analyze the trends for the two population samples, we observe that a vast majority of the canceled accounts are characterized by little to no activity while the rest display a decrease in their activity levels in the weeks prior to canceling their subscription. Both of these are expected behavior for disengaged players leading upto cancellation of subscription. Similarly, it is observed that a large portion (42.67%) of the active non-canceled accounts either display relatively consistent activity levels or show an increase in their activity levels. Again, both of these are expected behavior for engaged players. These two sets of behavioral profiles are quite distinct from each other and can thus discriminate between the samples.

The behavioral profile which is somewhat *counterintuitive* is *Active Non-Canceled Profile A* for the non-canceled accounts which in fact, comprise a majority (57.33%) of the active non-canceled accounts. These are the players who have not canceled their subscription and yet they show little to no activity for the past billing quarter. This profile more fits that of a canceled account or a inactive non-canceled account. The presence of such samples are a major source of prediction errors in the models.

Table 2.2: Player Cancellation Model: Info-gain ranking of features

Feature	Info-gain
num_sessions_enthu	0.1088
num_sessions_engage	0.1065
sl_mins_enthu	0.0912
isl_mins_engage	0.0859
isl_mins_enthu	0.0847
num_sessions_persist	0.0833
sl_mins_persist	0.0799
sl_mins_engage	0.0673
isl_mins_persist	0.0628

#### 2.4.4 Classification using Supervised and Hybrid Approaches - Cancellation

##### Feature Construction and Evaluation

We have the weekly history of session-related variables for each player - *Number of sessions per week* (num\_sessions), *Sum of session length per week* (sl\_mins) and *Sum of inter-session length per week* (isl\_mins). Using the semantic dimensions proposed in Section III C, we get a total of nine derived features. Table 2.2 lists the nine features in decreasing order of information gain. Information gain measures how well a given attribute separates the training examples according to the target classification and is given by the expected reduction in entropy caused by partitioning the examples according to the attribute [144].

We also used the Correlation based Feature Selection (CFS) technique [145] to identify the subset of features which are highly correlated with the class while having low intercorrelation. The features identified by this measure were - *num\_sessions\_engage*, *num\_sessions\_enthu*, *num\_sessions\_persist*, *sl\_mins\_enthu*, *isl\_mins\_engage* and *isl\_mins\_enthu*.

The feature evaluation thus tells us that the features derived from the *number of sessions per week* variable are highly discriminating and also uncorrelated with one another.



## Classification Results

We used the nine features from Table 2.2 and ran traditional classifiers using the Weka [146] machine learning tool to classify the 4032 test instances (refer Table 2.1). The goal was to correctly predict the test instances that canceled their subscription on week 35. The same experiment was also performed using *wClusterDist* - the proposed classification scheme which is based on the weighted distance from labeled clusters, using the same set of nine features. Figure 2.6 shows the precision, recall and f-measure values on the 1-week test-set for the churner class using JRip, J48, NaiveBayes, BayesNet, k-NearestNeighbor (k=3), Logistic Regression, MultiLayerPerceptron, SVM (using RBK kernel) and *wClusterDist*.

Key observations from this experiment are

- Among the traditional classifiers SVM gives the best precision (84.4); NaiveBayes gives the best recall (83.3) and Logistic Regression gives the best F-measure (64.5).
- Among all the classifiers, *wClusterDist* gives the best F-measure of 65.2, the best recall (77.6) while maintaining a good precision (56.2) value.
- *wClusterDist* performed better than the distance-based k-NearestNeighbor classifier, which was using 3-nearest neighbor and Euclidean distance.

Thus, the basic intuition behind *wClusterDist* of using a classification scheme based on weighted normalized Euclidean distance from labeled clusters, does seem to hold for this problem formulation with performance which is better than or comparable to traditional classification schemes.

Since the test set contains 25% canceled accounts, a naive classifier for the minority class would have an accuracy (precision) of 25. As observed in Figure 2.6, all the classifiers have a better accuracy than this naive baseline for the minority (canceled) class. Similarly, a naive classifier for the majority class would have an accuracy (precision) of 75 on the test set. The accuracy of all the classifiers for the majority (non-churner) class is better than this naive baseline as well - the highest being 92.6 for NaiveBayes and lowest being 77.2 for SVM.

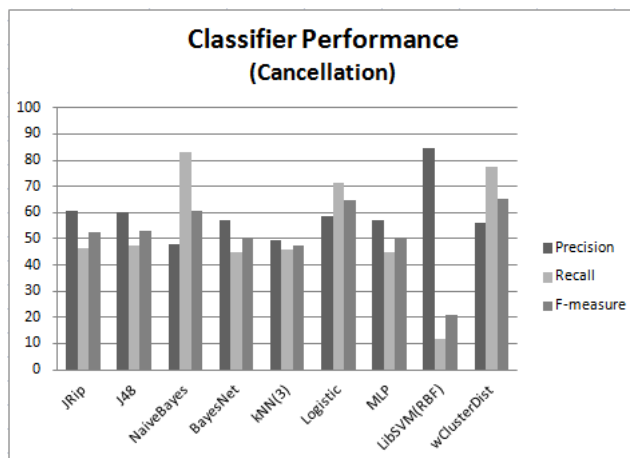


Figure 2.6: Player Cancellation Model: Performance on 1-week test set

### Comparison with previous results

Our best prediction model, *wClusterDist*, with an F-measure of 65.2 outperforms a modified diffusion model [13] which gave a best f-measure of 43.7 using an ADTree classifier on the same game dataset. In our previous work [134], we got an F-measure of 76 using the same game dataset. However, in that paper we had used a modified definition of churners to include accounts with no recorded activity in the last two months of the time period used for model building. Also, the results reported were for 10-fold cross-validation. In this case, we are addressing the more specific and harder problem of predicting players who are likely to cancel their subscription in the week following the date of analysis. We believe that our new model is better in terms of deployment within a real-time game analytic environment. In fact, lift analysis in the next section shows that a J48 classifier built using this lifecycle-based approach gives better lift numbers than our previous model.

### Impact of Experimental Parameters

The performance of a classifier built using the lifecycle-based model would be impacted by a couple of choices made during the experimental setup, as described below-

- *Activity Threshold*: The choice of the activity threshold impacts the number of active non-canceled accounts in the training set - the number of such samples

would reduce as the activity threshold is brought closer to the point of analysis (refer Figure 2.1). Specifically, this would reduce the number of samples belonging to *Active Non-Canceled Profile A*, which is characterized by little to no activity for most of the observed weekly history. Since this behavioral profile more resembles that of a canceled account, i.e. *Canceled Profile A*, fewer number of such samples would help the classifier to discriminate better. To summarize, *the closer the activity threshold is to the point of analysis, the better should be the precision of the classifier.*

- *History Length:* We choose the history length to be 13 weeks since that corresponded to a quarterly cycle. However, the behavioral profiles from Figures 2.4 and 2.5 seem to indicate that even the trends in the last 2-3 weeks of history may be sufficient to discriminate between the two population samples. So, *the classifier performance should not degrade significantly even if we were to use a shorter history length.*

Of course, both of the above claims need to be empirically verified and we propose to pursue that as part of future work.

#### 2.4.5 Evaluation of *wClusterDist* - Cancellation

Experimental results in the previous section showed that the *wClusterDist* classification scheme is well-suited for this problem formulation and its performance is better than or comparable to other traditional classifiers (JRip, J48, NaiveBayes, BayesNet, kNN, LogisticRegression, MultilayerPerceptron and SVM). In this section, we describe additional experiments performed to evaluate *wClusterDist*. As mentioned earlier, the underlying motivation of this technique is to benefit from the distinct canceled and active non-canceled behavioral profiles, as indicated by the cluster centroids of Figures 2.4 and 2.5.

We constructed the canceled and active non-canceled training clusterings using the three options listed next. The number of input clusters for each option was determined by doing a *knee-of-the-curve* analysis as shown in Figure 2.3. The label *DimN* indicates that the training clusters had dimension equal to N.

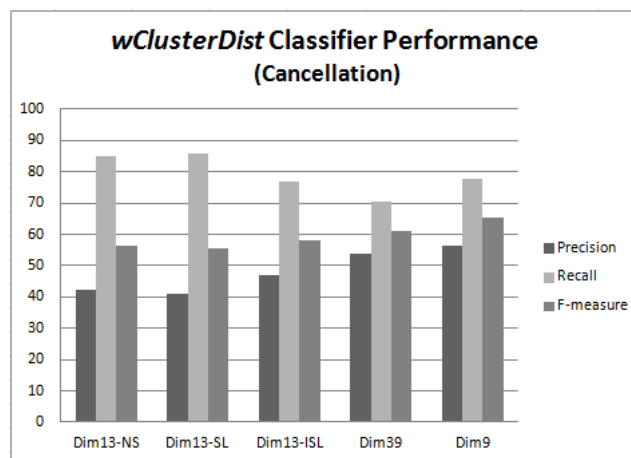


Figure 2.7: Player Cancellation: wClusterDist Classifier Performance on 1-week test set

- *Dim13*: The 13-week history of a single observable was used in this case. Three sets of clusters are possible corresponding to the three observed variables - these are labeled Dim13-NS, Dim13-SL and Dim13-ISL. Figures 2.4 and 2.5 are the clusterings constructed using the observable *Number of sessions per week*.
- *Dim9*: The 9 derived semantic features from Table 2.2 are used in this case.
- *Dim39*: The concatenated 13-week history of the three variables are used in this case.

Figure 2.7 shows the performance of the *wClusterDist* classification scheme using the different cluster configurations listed above. We observe that if we just use clusters based on 13-week history of a single variable (refer Dim13-NS, Dim13-SL, Dim13-ISL), we get higher recall numbers but the precision is low. One reason for the lower precision could be attributed to the similarity of *Canceled Profile A* and *Active Non-Canceled Profile A*, as observed in Figures 2.4 and 2.5. On the other hand, using combined weekly history (Dim39) or the nine derived semantic features (Dim9) reduces recall but improves precision and gives slightly better F-measures.

Table 2.3: Player Dormancy Dataset

Dataset	Label Week	Dormant	Non-dormant	Total samples
Training	Week 34	17680	16370	34050
Test	Week 35	17159	15247	32406

#### 2.4.6 Experimental Setup - Dormancy

In order to build a prediction model for the *Dormancy* problem, we modify the experimental setup used for the *Cancellation* problem. Our goal here is to predict players who are going to be dormant in week 35. In order to get training samples, we use the labels from week 34. The accounts which did not have any activity in week 34 are used as positive training samples (dormant accounts) whereas the accounts which had some activity in week 34 are used as negative training samples (non-dormant accounts). The hold-out test samples are obtained in a similar manner by using labels from week 35.

Table 2.3 gives a breakdown of the samples in the dataset. We observe that dormancy prediction is no longer an imbalanced class problem as opposed to predicting cancellations (refer table 2.1). In fact, the dormant samples form the majority class with a class distribution of 51.92:48.08 in the training dataset and 52.95:47.05 in the test dataset.

#### 2.4.7 Analysis of Player Profiles - Dormancy

In this section, we visualize and discuss the results of clustering the dormant and non-dormant training samples using their weekly history of *number of sessions*. Figure 2.8 shows the decrease in the *Within Cluster Sum of Squared Errors* as the number of clusters is decreased and we observe that the knee of the curve for both the canceled and active non-canceled samples is observed around  $k = 5$ . Also, we would like to compare the behavioral profiles of the dormancy model with those of the canceled subscription model (refer Section IVC). Therefore, we perform k-means clustering on the dormant and non-dormant training samples using 5 clusters each as input. With  $k = 5$ , we

Figures 2.9 and 2.10 show the cluster centroids for the two population training samples, with the x-axis having the aligned 13-week lifecycle history and the y-axis

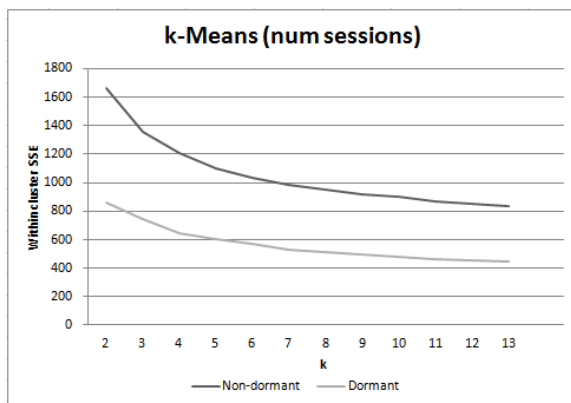


Figure 2.8: Within cluster SSE for dormant and non-dormant samples

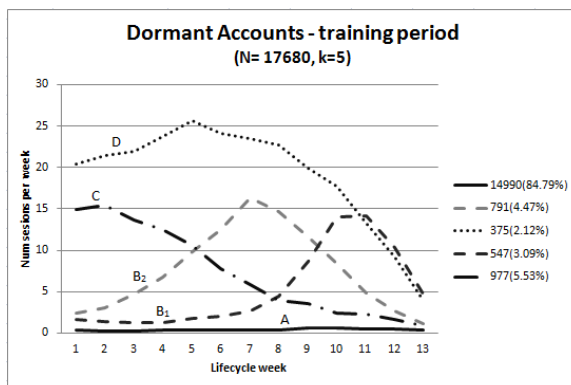


Figure 2.9: Cluster centroids for dormant training samples

having the number of sessions per week. As earlier, the behavioral trends between the two samples are quite different from each other and in both the cases, we can identify few distinct behavioral profiles. We use *activity* to mean the observed number of sessions in a week

Four behavioral profiles of dormant accounts are observed from Figure 2.9, which are described next. To be noted here that the 13 weekly points in the X-axis represent the account's weekly history leading up to week 34 when it was labeled dormant.

- *Dormant Profile A*: These are the dormant accounts who show little to no activity and they constitute the large majority (84.79%) of the canceled population.
- *Dormant Profile B*: The players fitting this profile show an increase in their activity

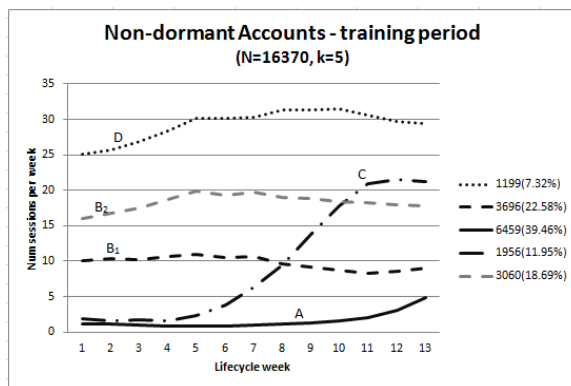


Figure 2.10: Cluster centroids for non-dormant training samples

and then a sharp decrease before dormancy - they constitute around 7.56% of the dormant account samples. Two clusters are observed within this profile -  $B_1$  (3.09%) and  $B_2$  (4.47%). The dormant accounts belonging to the  $B_2$  profile show an increase in their activity but then the activity has a downward slope after around the mid point whereas  $B_1$  profile has canceled accounts who are initially inactive followed by an increasing activity level in the latter half and then a drop in the last week or so.

- *Dormant Profile C*: Around 5.53% of the dormant accounts belong to this profile which is characterized by a gradual decrease in activity throughout the 13 week history.
- *Dormant Profile D*: 2.12% of the dormant accounts fit this profile. The dormant accounts belonging to this group have a very high activity level in the first half but as in the case of profile B, there is sharp decrease in their activity leading up to the week of dormancy.

Four behavioral profiles of non-dormant accounts are observed from Figure 2.10. In this case, the 13 weekly points represent the weeks leading up to week 34.

- *Non-Dormant Profile A*: The non-dormant accounts fitting this profile comprise 39.46% of the non-dormant players. The players fitting this profile have little to no activity for most of their history but show a significant increase in activity in the weeks immediately before the week of analysis.

- *Non-Dormant Profile B*: The players fitting this profile maintain an almost constant level of activity in the first half followed by a slight and gradual decrease in the activity level in the second half. They comprise 41.27% of the non-dormant population. Depending on the average level of activity, two sub-profiles are observed in this group -  $B_1$  (22.58%) and  $B_2$  (18.69%). The players in the  $B_1$  sub-profile have a lower mean activity level of around 9 sessions per week throughout the history whereas the  $B_2$  sub-profile has a higher mean activity level of around 17 sessions per week.
- *Non-Dormant Profile C*: Around 11.95% of the non-dormant accounts belong to this profile which is characterized by a gradual increase in activity throughout the 13 week history.
- *Non-Dormant Profile D*: The players fitting this profile are highly engaged in the game with both a high activity level and a general trend of increasing activity level throughout the 13-week history. They comprise around 7.32% of the non-dormant population.

Overall, if we analyze the trends for the two population samples, we observe that a vast majority (nearly 85%) of the dormant accounts are characterized by little to no activity while the rest display a decrease in their activity levels in the weeks prior to the date of analysis. Both of these are expected behavior for disengaged players before they become dormant. Similarly, the behavioral profiles of the non-dormant players indicate that they are consistently active in the game or show an increase in level of activity in the weeks prior to the date of analysis. Again, these are expected behavior for players who are engaged in the game experience.

#### 2.4.8 Classification using Supervised and Hybrid Approaches - Dormancy

##### Feature Construction and Evaluation

We have the weekly history of session-related variables for each player - *Number of sessions per week* (num\_sessions), *Sum of session length per week* (sl\_mins) and *Sum of inter-session length per week* (isl\_mins). Using the semantic dimensions proposed in



Table 2.4: Player Dormancy Model: Info-gain ranking of features

Feature	Info-gain
isl_mins_engage	0.289
num_sessions_engage	0.287
num_sessions_enthu	0.287
isl_mins_enthu	0.285
sl_mins_engage	0.281
sl_mins_enthu	0.269
isl_mins_persist	0.244
sl_mins_persist	0.202
num_sessions_persist	0.201

Section III C, we get a total of nine derived features. Table 2.4 lists the nine features in decreasing order of information gain. We observe that all the features have an info-gain value in the range of 0.2-0.29 and in fact, the top five features are quite close to each other in their info-gain value. Thus, each feature by itself has good discriminating power.

We also used the Correlation based Feature Selection (CFS) technique [145] to identify the subset of features which are highly correlated with the class while having low intercorrelation. The features identified by this measure were - *sl\_mins\_engage*, *sl\_mins\_enthu*, *isl\_mins\_engage* and *isl\_mins\_enthu*.

### Classification Results

We used the nine features from Table 2.4 and ran traditional classifiers using the Weka [146] machine learning tool to classify the 32406 test instances (refer Table 2.1). The goal was to correctly predict the test instances that became dormant on week 35. The same experiment was also performed using *wClusterDist* - the proposed classification scheme which is based on the weighted distance from labeled clusters, using the same set of nine features. Figure 2.11 shows the precision, recall and f-measure values on the 1-week test-set for the dormant class using JRip, J48, NaiveBayes, BayesNet, k-NearestNeighbor (k=3), Logistic Regression, MultiLayerPerceptron, SVM (using RBK

kernel) and *wClusterDist*.

Key observations from this experiment are

- Among all the classifiers SVM gives the best precision (85.8); JRip gives the best recall (88.7) and J48 decision tree gives the best F-measure (85.5).
- *wClusterDist* performs quite well as compared to the best traditional classifiers with precision, recall and F-measure values of 82.6, 84.7 and 83.7.

As opposed to its performance in the *Cancellation* classification problem, *wClusterDist* does not give the best performance among the same set of traditional classifiers. But its performance is within 4% of the best performance in each of the evaluation metric - Precision, Recall and F-measure and within 2% of the best F-measure. A major difference between the two experimental setups is that *Cancellation* prediction was an imbalanced class problem where the target class (Cancellation) was the minority class comprising 25% of the test samples. On the other hand, *Dormancy* prediction is no longer an imbalanced class problem with the target class (Dormant) being a majority class comprising of 52.95% of the test samples.

Since the test set contains 52.95% dormant accounts, a naive classifier for the majority class would have an accuracy (precision) of 52.95. As observed in Figure 2.11, all the classifiers have a better accuracy than this naive baseline for the majority (dormant) class. Similarly, a naive classifier for the minority class would have an accuracy (precision) of 47.05 on the test set. The accuracy of all the classifiers for the minority (non-dormant) class is better than this naive baseline as well - the highest being 88.3 for Logistic Regression and lowest being 66.3 for LibSVM.

#### 2.4.9 Evaluation of *wClusterDist* - Dormancy

In this section, we describe additional experiments performed to evaluate *wClusterDist* for the Dormancy problem under different feature spaces. We constructed the dormant and non-dormant training clusterings using the three options listed next. The number of input clusters for each option was determined by doing a *knee-of-the-curve* analysis as shown in Figure 2.8. The label *DimN* indicates that the training clusters had dimension equal to N.

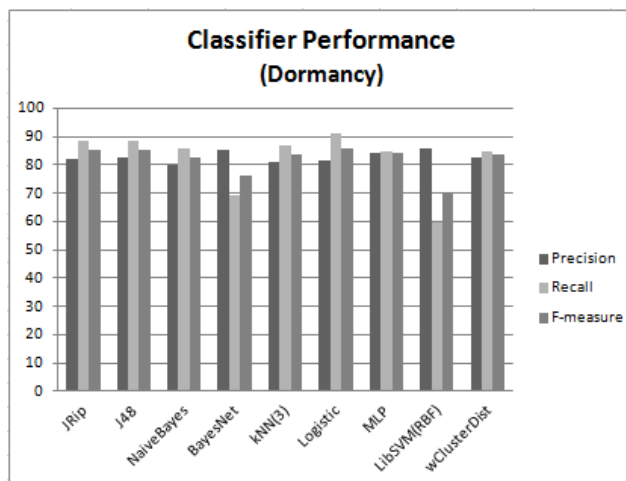


Figure 2.11: Player Dormancy Model: Performance on 1-week test set

- *Dim13*: The 13-week history of a single observed variable was used in this case. Three sets of clusters are possible corresponding to the three observed variables - these are labeled Dim13-NS, Dim13-SL and Dim13-ISL. Figures 2.9 and 2.10 are the clusterings constructed using the observable *Number of sessions per week*.
- *Dim9*: The 9 derived semantic features from Table 2.4 are used in this case.
- *Dim39*: The concatenated 13-week history of the three variables are used in this case.

Figure 2.12 shows the performance of the *wClusterDist* classification scheme using the different cluster configurations listed above. We observe that if we just use clusters based on 13-week history of a single variable (refer Dim13-NS, Dim13-SL, Dim13-ISL), we get higher recall numbers but the precision is low. On the other hand, using combined weekly history (Dim39) or the nine derived semantic features (Dim9) reduces recall but improves precision and gives slightly better F-measures.

## 2.5 Application Impact

In this section, we evaluate the lifecycle-cycle based approach by doing lift analysis of a classifier built using this approach. A lift curve is an important tool for direct

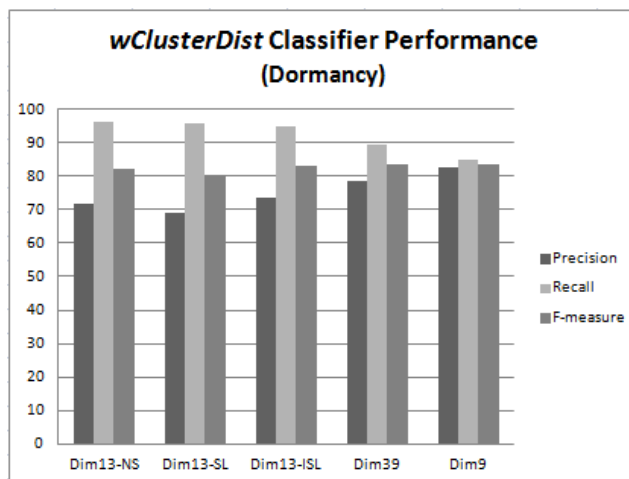


Figure 2.12: Player Dormancy: wClusterDist classifier performance on 1-week test set

marketing when a subset of customers are to be contacted [147]. The lift is a measure of a predictive model calculated as the ratio between the results obtained with and without the predictive model [8]. In order to generate a lift curve, we need to sort the test instances in order of their decreasing churn probabilities. So, for this analysis, we use the J48 decision tree model since it easily gives us the churn probability of an account.

### 2.5.1 Lift Analysis - Cancellation

For the *Cancellation* problem, the J48 model was trained on the 41152 training samples and evaluated on the 4032 test samples (refer Table 2.1). As shown in Figure 2.6, the model had precision:recall:f-measure of 60.8:46.4:52.6 on the test set.

The top chart of Figure 2.13 shows the cumulative lift curve generated using the J48 decision tree model on the test set. We also wanted to analyze the model accuracy for each part of the lift curve - for that reason we plotted the predicted churn probabilities of each test instance in the bottom chart of the figure. To be noted that the J48 decision tree model uses a threshold of 0.5 to label unseen instances as churner or non-churner. *Model accuracy* refers to whether the predicted label matched the actual label (for both churner and active non-churner test instances).

We observe three regions in the churn probability curve of Figure 2.13-

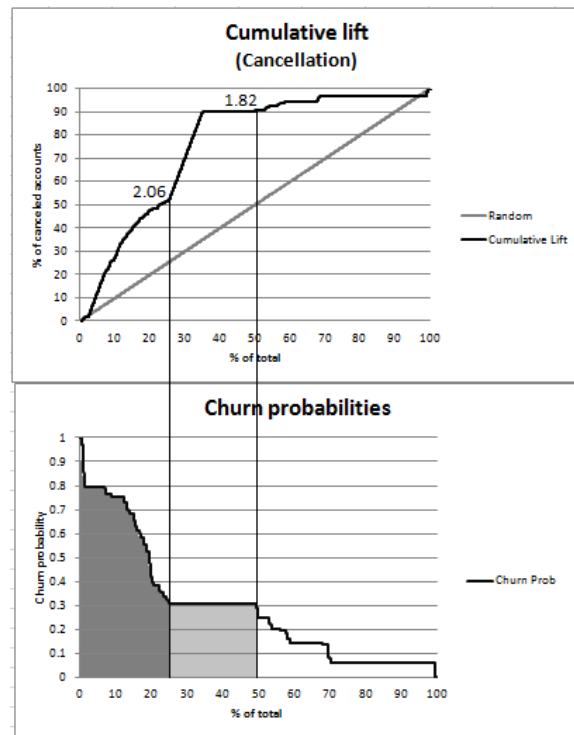


Figure 2.13: Lift analysis for canceled accounts using J48 decision-tree model

- In the dark-shaded region to the left, the churn probabilities drops from 1 to 0.3 and the model accuracy is 64.79%. At the end of this region, one can reach 51.79% of the canceled accounts with 25.12% of the total test instances i.e the lift at the end of the region is 2.06. To be noted that since the churn probability has dropped below 0.5 by this point, all test instances in the next two regions are predicted to be active non-churners.
- In the light-shaded region in the middle, the churn probabilities remain more or less constant around 0.3 and the model accuracy is 60.79%. At the end of this region, one can reach 90.18% of the canceled accounts with 50% of the total test instances i.e a lift of 1.82.
- In the unshaded region to the right, the churn probabilities drops from 0.3 to 0 and the overall model accuracy is 95.13%.

To summarize, using a player lifecycle-based model, one can contact 51.79% of the

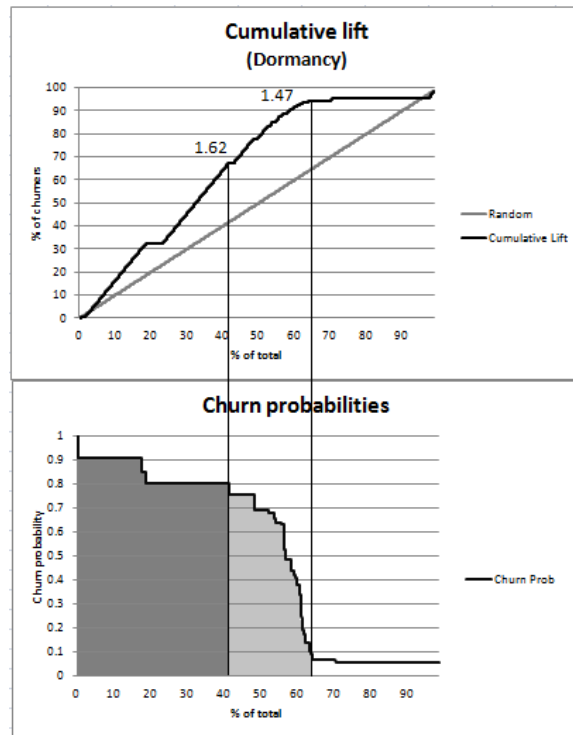


Figure 2.14: Lift analysis for dormant accounts using J48 decision-tree model

players who are about to cancel their accounts next week by reaching out to 25% of the total population and 90.18% of the churners with 50% of the total population. This is better than our previous model's lift numbers [134] where one could reach 90% of the actual churners with 67% of the population.

### 2.5.2 Lift Analysis - Dormancy

For the *Dormancy* problem, the J48 model was trained on the 34050 training samples and evaluated on the 32406 test samples (refer Table 2.3). As shown in Figure 2.11, the model had precision:recall:f-measure of 82.7:88.5:85.5 on the test set. As earlier, the top chart of Figure 2.14 shows the cumulative lift curve generated using the J48 decision tree model on the test set and the bottom chart shows the predicted churn probabilities of each test instance.

We observe three regions in the churn probability curve of Figure 2.14-

- In the dark-shaded region to the left, the churn probabilities are in the range of 0.8 to 1 and the model accuracy is 85.65%. At the end of this region, one can reach 67.07% of the canceled accounts with 41.46% of the total test instances i.e the lift at the end of the region is 1.62. To be noted that since the churn probability is above 0.8, all test instances in this region are predicted to be dormant.
- In the light-shaded region in the middle, the churn probabilities drop sharply from 0.8 to 0.07 and the model accuracy is 70.42%. At the end of this region, one can reach 94.02% of the dormant accounts with 64% of the total test instances i.e a lift of 1.47.
- In the unshaded region to the right, the churn probabilities drops from 0.07 to 0 and the overall model accuracy is 91.21%.

To summarize, using a player lifecycle-based model, one can contact 67.07% of the players who are about to become dormant next week by reaching out to 41.46% of the total population and 94.02% of the dormant accounts with 64% of the total population.

For the *Cancellation* model, we found that we could reach 51.79% of the target class with 25% of the total instances (lift = 2.06) and reach 90.18% of the target class with 50% of the total instances (lift = 1.82). On the other hand, for the *Dormancy* model, we found that we could reach 35.98% of the target class with 25% of the total instances (lift = 1.44) and reach 78.64% of the target class with 50% of the total instances (lift = 1.57). This indicates that the *Cancellation* prediction model does better than the *Dormancy* prediction model in terms of its lift performance. This is an interesting result because the *Dormancy* prediction model does much better than the *Cancellation* prediction model in terms of the other evaluation metrics of Precision, Recall and F-measure.

## 2.6 Discussion

### 2.6.1 Comparison of Cancellation and Dormancy

In this chapter, we did an extensive evaluation of the two types of player churn - *Cancellation* and *Dormancy*. Cancellation relates to the player *stopping payment* - by either pro-actively canceling the subscription or by not renewing the subscription. Dormancy,

on the other hand, has to do with the player *stopping to play* the game. Both of these types of churn are indicators of a player getting disengaged from the game. Some of the key similarities and differences observed between these two phenomenon, based on our experiments, are listed below

### **Class Imbalance**

In case of *Cancellation*, only a small fraction of the base population unsubscribe from the game in any given week (refer Table 2.1) so that the prediction problem is an imbalanced class one with the target class being the minority one and therefore that much more difficult. On the other hand, we find that nearly half of the base population become dormant in a given week (refer Table 2.3) so that we are no longer dealing with an imbalanced class problem.

### **Behavioral Profiles**

Based on time-series clustering, we could identify distinct behavioral profiles that can help identify players who are about to churn from the game. A striking fact is that the behavioral profiles for the *Cancellation* problem (refer figures 2.4 and 2.5) and the *Dormancy* problem (refer figures 2.9 and 2.10) are near identical.

In case of the *Cancellation* problem, we observe that nearly 76% of the canceled accounts become inactive (stopped playing the game) long before they decide to cancel their subscription or their subscription expires. Interestingly, we found that about 57% of the non-canceled accounts also display similar behavior as above i.e they stop playing the game. Such a strongly disengaged behavioral profile for non-canceled accounts seem to be a major source of prediction errors for the model. More importantly, this indicates that dormancy is a highly prevalent phenomenon among the player population. This fact is also borne out by our earlier observation that nearly half of the player base becomes dormant in a week.

In case of the *Dormancy* problem, we observe that nearly 85% of the dormant accounts have become inactive (stopped playing the game) long before the date of analysis. Though we observe that around 39% of the non-dormant accounts are marked by a long period of low level of activity, they shows an upward slope in the weeks immediately prior to the date of analysis.



## Inertia Effect

The behavioral profiles described above indicate that there is an inertia effect driving both the churn phenomena - *Cancellation* and *Dormancy* and this inertia effect seems to be stronger in the case of the *Dormancy* phenomenon. A vast majority of the churners are characterized by little to no activity while the rest display a decrease in their activity levels in the weeks leading up to churn. Similarly, players who are unlikely to churn in a given week display consistent level of activity in the game or show an increase in level of activity in the weeks prior to the date of analysis.

Our proposed semantic dimensions do a good job of capturing this inertia effect and thus the derived features do a good job of discriminating between the target classes.

## Model Performance

In terms of model performance, the prediction models for both the types of churn do much better than a random classifier or previous models on the same dataset [134, 13]. The *Dormancy* prediction model (refer figure 2.11) does much better than the *Cancellation* model (refer figure 2.6) though. This could be attributed to the fact that the class compositions for the two problems are quite different (refer *Class Imbalance* section above) with the *Dormancy* classification problem being one of balanced class. Also, there is a much better distinction in the behavioral profiles for the *Dormancy* problem as compared to the *Cancellation* problem (refer *Behavioral Profiles* section above).

Interestingly, we find that the *Cancellation* prediction model does better in terms of its lift performance. Also, our proposed distance-based classification scheme, *wCluster-Dist*, gives the best overall performance for the *Cancellation* problem and is within 2% of the best overall performance for the *Dormancy* problem. Since the proposed scheme is based on the player's behavioral profiles in the underlying feature space, the results from this approach can be helpful to analysts in terms of being able to identify players as belonging to nicely interpretable behavioral profiles.

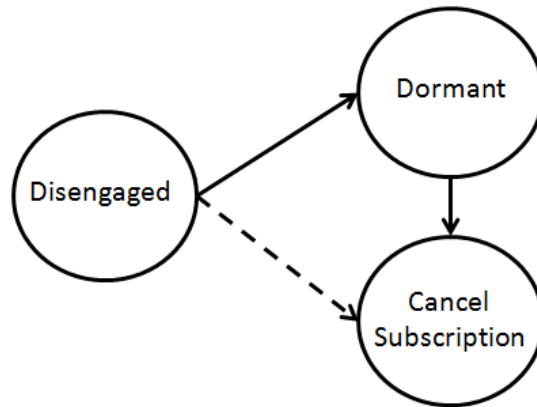


Figure 2.15: Player Disengagement Model

### Relationship with Player Engagement

From a behavioral or even cognitive point of view, churn is an indicator of the player getting disengaged (losing interest or bored) in the game experience. Such disengagement can be manifested in either the player *stopping to pay* or *stopping to play*. In this chapter, we have analyzed both of these problems as *Cancellation* or *Dormancy* respectively.

Based on the experimental results, we find that *Dormancy* is a much more prevalent phenomenon among players than *Cancellation*. Figure 2.15 illustrates the resulting player disengagement model with the states and transitions. As per the model, when a player becomes disengaged in the game, they are most likely to become dormant (solid line in figure) while few of them can directly cancel their subscription (dotted line in figure). The dormant accounts will eventually cancel their subscriptions. Since a majority of the disengaged players *stop to play* long before they *stop to pay*, we claim that ***Dormancy* is a better indicator of player disengagement in terms of its immediacy rather than *Cancellation*.**

Furthermore, since we can obtain the true labels for the *Dormancy* model using purely behavioral information, it can be useful for all kinds of games. However, not all games have a subscription-based business model and thus the *Cancellation* model would not be applicable to such games.

### 2.6.2 Other Applications

In this chapter, we have used a player lifecycle-based approach for churn prediction in MMORPGs. If we consider a player’s natural lifecycle, then churn is a terminal state in which the player leaves the game. There can be other events where the game administrators can cancel an account for reasons such as the account being compromised, use of illegal software by the account or gold farming. Gold farming refers to the illicit practice of gathering and selling virtual goods in online games for real money [148]. All of these again are terminal events in the player lifecycle and hence, a lifecycle-based approach proposed in this chapter should suit the problem of predicting such events quite well.

In fact, the above claim was verified recently when a lifecycle-based change point detection method was used for detecting compromised accounts in the same Sony Everquest II MMORPG dataset [149]. Among the observable variables evaluated in that paper, *Session Length* and *Experience Points Earned* over the weeks leading upto the account being labeled as compromised were found to be highly discriminative in nature in terms of predicting compromised accounts with F-measure of around 67.1 and 67.6 respectively.

## 2.7 Conclusion and Future Work

We investigated the problem of churn prediction in MMORPGs using a player lifecycle-based approach for modeling user behavior. In particular, we addressed the problem of predicting players who are likely to churn in the week following the date of analysis. We looked at two definitions of churn - *Cancellation* and *Dormancy*.

We adopted a player lifecycle-based approach where we analyzed the activity traits of churners in the weeks leading up to their point of churn and compared it with the activity traits of a regular player. The analysis indicated that such an approach has good predictive power but more importantly it gave us key insights into churn behavior in an MMORPG. We could identify distinct behavioral profiles associated with churners and non-churners which can discriminate between the two populations. We proposed three semantic dimensions of *engagement*, *enthusiasm* and *persistence* along

which weekly player history of an observed variable is recomputed to give derived features. Using three session-related variables alone and the features derived from them, we achieve good classification performance with our prediction models. We also proposed a distance-based classification scheme, which we call *wClusterDist*, which benefited from the distinct behavioral profiles of the two populations. Experimental results show that the proposed classification scheme is well-suited for this problem formulation and its performance is better than or comparable to other traditional classifiers. Finally, we did an extensive evaluation and comparison of the two types of churn and based on our analysis, we propose a model of player disengagement and claim that *Dormancy* is a better indicator of player disengagement over *Cancellation* - both in terms of its immediacy to the underlying construct and its universal applicability to any online game, especially to games which do not have a subscription-based model.

We have only used session-related variables in this chapter. As part of our future work, we would also like to explore if other features based on player motivation theories [150] can improve model performance. We will empirically test the impact of the two experimental parameters - *Activity Threshold* and *History Length* on the performance of classifiers build using lifecycle-based approach. Finally, we will investigate *wClusterDist*, the proposed distance-based classification scheme, in terms of its generalizability to other problems.

## Chapter 3

# Community Churn Prediction in Online Games Using Supervised Learning

### 3.1 Introduction

A massively multiplayer online role-playing game (MMORPG) is a popular genre of computer-based game which is characterized by a persistent virtual world maintained by the game developer. A player in an MMORPG can control one or more characters through which they interact with the game environment. Activities in this environment are mostly driven by individual or group quests. While group quests may not be very important in early levels within the game, they become a necessity in later levels for players to progress through the game. In our previous chapter, we addressed the problem of modeling individual churn behavior using a lifecycle-based approach. The motivating idea behind using a lifecycle-based approach for churn prediction is to analyze the player's behavior leading up to the terminal state and discover key indicators or behavioral trends which can help identify players who are moving towards that state.

The focus of this chapter is to build models to predict the likely churn of a dynamically evolving in-game community. We use an activity-based definition wherein a community is said to have churned when the community size falls below a user-defined

*churn threshold*. A churn threshold of  $x\%$  means that a community is considered to have churned when more than (or equal to)  $x\%$  of its current members are no longer part of the community in the next week. The problem addressed in this chapter is one of accurately predicting if a player community is likely to churn in the week following the date of analysis.

The main contribution of this chapter is a framework for predicting churn of dynamically evolving communities using a supervised learning approach. First, we identify *persistent groups* that are indicative of long-term social relationships among the group members. Using such relationships, we can measure the strength of connections among players based on the amount of time they spend together in these groups. We then construct weighted networks for each time period and run community detection algorithm to identify tightly knit communities in the weighted networks. Next, we track the evolution of the communities to identify churned and non-churned communities. We calculate several intuitive community health metrics for each time period and use a lifecycle-based approach to build the final prediction model. Using such an approach we can provide key insights into behavioral traits that can distinguish between churned and non-churned communities.

To the best of our knowledge, no prior work has addressed the specific problem of community churn prediction in online games or proposed a general framework for predicting churn of dynamically evolving communities using a supervised learning approach.

The chapter is organized as follows: Section 2 talks about background and related work. Section 3 describes the overall framework for predicting churn of dynamically evolving communities using a supervised learning approach. Section 4 describes the dataset, experimental setup, results and analysis and Section 5 is conclusion and future work.

## 3.2 Background and Related Work

Churn refers to loss of customers and thus loss of revenue. Therefore, understanding churn behavior and being able to accurately predict likely churners is important for businesses. While most of the research has focused on modeling individual churn behavior,

some of the work take into account the structural properties of the underlying network [151] and the influence of social ties [152], [153] in individual churn analysis. Recently, Oentaryo et al proposed a churn prediction approach based on collective classification that exploits intrinsic (user profile) and extrinsic (social ties) factors underlying churn behavior [154]. Churn in digital social networks differs in many ways from a traditional domain such as telecom and as such, activity-based definitions of churn have been proposed to capture the nuances of such online environments [155]. User churn in social networks have been explored in different settings such as Q&A forums [156] and discussion boards [155].

Community detection in graphs is of great importance in several fields such as sociology, biology and computer science and as such has become a fast-growing research area. In a recent survey, Fortunato et al have discussed various aspects of the problem, from the definition of the main elements of the problem, to the presentation of most methods developed [112]. The study of dynamic communities, however, is still in its infancy. Palla et al. performed an analysis of dynamic communities in a cell phone network and a collaboration network [117] using their *Clique Percolation Method* [118]. Berger et al have proposed a framework for analysis of dynamic social networks [119], [120] based on social cost model.

Related work to this chapter would be that of Richter et al [157] who proposed a group-first social networks approach for predicting customer churn in mobile networks. They identify tightly-knit social groups, assign a churn risk score for each group and finally assign an individual churn score to each subscriber based on the churn score of her social group as well as her personal characteristics. However, they use heuristics for identifying communities and their approach do not take into account the evolution of communities over time. Our proposed framework takes into account the dynamic nature of communities and the supervised learning approach allows us to gain insights into key discriminators of churn at a community level. Also, we use an activity-based definition of churn which we believe to be a much better indicator of disengagement than cancellation of subscription.

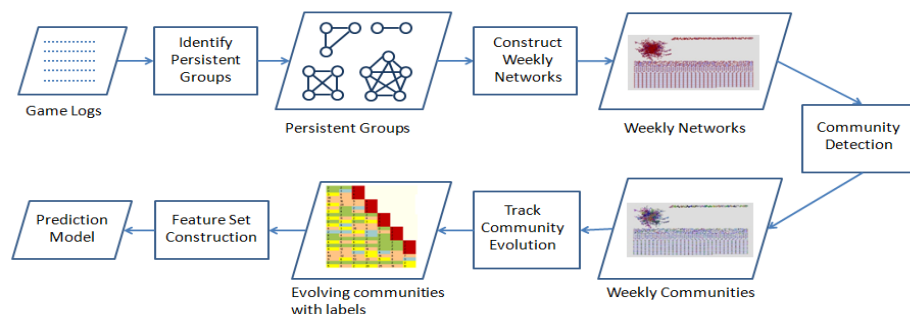


Figure 3.1: Supervised Learning Approach to Community Churn Prediction

### 3.3 Methodology

Figure 3.1 outlines the proposed framework for predicting churn of dynamically evolving communities using a supervised learning approach. The parallelograms represent input and output to processes (square boxes).

First, *persistent groups* are identified from game logs. Persistent groups are those which represent non-transient relationship among its members. Using such relationships, we can measure the strength of connections among players based on the amount of time they spend together in such groups. Second, networks are constructed from the persistent groups for every time period based on the weighted connections among the group members. An edge in this network means the two players were part of a group in that time period and the edge weight is based on the amount of playtime the two players spent together. We have used week as the basic unit of analysis in this chapter. One can choose a finer (hourly or daily) or coarser (fortnightly or monthly) level of analysis depending on application requirements.

Next, community detection algorithm is then applied to each network to get communities (partitions) for the time period. The quality of the partitions resulting from these methods is measured by the modularity of the partition. The modularity of a partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities [158] and in the case of weighted networks, it is defined as [159] -



---

**Algorithm 1** Community Tracking Algorithm

---

```

for all  $Community(week_i) \in$  list of communities in week  $i$  do
   $highest\_ji = 0$ 
   $Community\_Next(week_i) = null$ 
  for all  $Community(week_{i+1}) \in$  list of communities in week  $i+1$  do
     $actual\_ji = JaccardIndex(i, i + 1)$ 
    if  $actual\_ji > highest\_ji$  then
       $highest\_ji = actual\_ji$ 
       $Community\_Next(week_i) = Community(week_{i+1})$ 
    end if
  end for
   $current\_size = |Community(week_i)|$ 
   $next\_size = |Community\_Next(week_i)|$ 
   $Community\_Label(week_i) = f(current\_size, next\_size, highest\_ji)$ 
end for

```

---

Figure 3.2: Algorithm for tracking communities over time

$$Modularity, Q = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \quad (3.1)$$

where  $A_{ij}$  is edge weight between nodes  $i$  and  $j$ ,

$k_i = \sum_j A_{ij}$  is the sum of the weights of the edges attached to vertex  $i$ ,

$c_i$  is the community to which vertex  $i$  is assigned,

the  $\delta$  function  $\delta(u, v)$  is 1 if  $u = v$  and 0 otherwise

and  $m = \frac{1}{2} \sum_{i,j} A_{ij}$

Once the communities have been identified for each time period, we track the evolution of communities across time periods. The basic metric used to link communities across two time periods can be any set similarity measure. For the purposes of our analysis, we have used *JaccardIndex*, as defined in 3.2

$$JaccardIndex(i, i + 1) = \frac{|Community(week_i) \cap Community(week_{i+1})|}{|Community(week_i) \cup Community(week_{i+1})|} \quad (3.2)$$

where  $Community(week_i)$  is the set of community members in week  $i$

Algorithm 3.2 outlines how the evolution of a community is tracked across time periods. We consider communities for two consecutive weeks  $i$  and  $i + 1$ , compute similarity metric for every pair of communities  $(i, i + 1)$  and assign the highest match

to the pair. Based on the value of the similarity metric and the user-specified *churn threshold*, the algorithm outputs a label for every community in week  $i$ . The labels are described below -

- EXPIRED: A community is labeled as EXPIRED if the similarity match between communities in consecutive weeks is below the specified *churn threshold*.
- CONTINUE: A community is labeled as CONTINUE if we find a perfect match for the community in the next week.
- CHANGE\_MEMBER - There was a partial match of the communities in consecutive weeks but the community size remains unchanged.
- MERGE\_EXPAND - There was a partial match of the communities in consecutive weeks and the size of the matched community increased from week  $i$  to week  $i + 1$
- SHRINK - There was a partial match of the communities in consecutive weeks and the size of the matched community decreased from week  $i$  to week  $i + 1$

The *Community Tracking Algorithm* allows us to assign labels to a community for each time period and thus track the evolution of the community over time (refer figure 3.6) and its eventual churn. This allows us to construct a labeled dataset for a binary classification problem in which for any point in time, the communities which are labeled as EXPIRED becomes a churning sample (target class) and the samples with other labels are treated as non-churning samples.

Finally, we compute several intuitive community-based metrics (below) across the lifetime of a community and analyze the efficacy of these metrics in predicting community churn. In this chapter, amount of playtime is represented by SL Mins (session length in minutes) where a session consists of sets of activities which are separated by no more than 30 minutes.

- *Size*: The size of the community in the week

$$Size(week_i) = |Community(week_i)| \quad (3.3)$$

- *Connectedness*: Measure of how connected the community is

$$Connectedness(week_i) = \frac{Observed\ intracommunity\ edges\ in\ week\ t}{Possible\ intracommunity\ edges\ in\ week\ t} \quad (3.4)$$

- *Intra Community SL Mins*: A measure of the strength of connections among the community members and is given by the total amount of playtime the community members spend within the community.

$$Community\_intra(week_i) = \sum_{e \in community(week_i)} weight(e) \quad (3.5)$$

where e is an intra-community edge

- *Inter Community SL Mins*: A measure of the strength of connections to other communities and is given by the total amount of playtime the community members spend with members of other communities.

$$Community\_inter(week_i) = \sum_{e \in community(week_i)} weight(e) \quad (3.6)$$

where e is an inter-community edge

- *Community Social SL Mins*: A measure of social engagement of the community members given by -

$$Community\_social(week_i) = \sum_{member \in community(week_i)} Social\ SL\ Mins(member) \quad (3.7)$$

where *Social SL Mins(member)* is the amount of time the community member has spent in groups during week *i*

- *Community Overall SL Mins*: A measure of overall engagement of the community members given by -

$$Community\_overall(week_i) = \sum_{member \in community(week_i)} Overall\ SL\ Mins(member) \quad (3.8)$$

where  $Overall\ SL\ Mins(member)$  is the amount of time the community member has spent overall during week  $i$

- *Member Loyalty*: A measure of how loyal members are to the community.

$$Member\_loyalty(week_i) = \frac{Community\_intra(week_i)}{Community\_social(week_i)} \quad (3.9)$$

- *Member Socialization*: A normalized measure of social engagement of the community members.

$$Member\_social(week_i) = \frac{Community\_social(week_i)}{Community\_overall(week_i)} \quad (3.10)$$

We then apply three semantically meaningful dimensions, as proposed in our previous chapter, along which the weekly history of an observed metric is recomputed to get derived semantic features. These dimensions are -

- *Engagement*: This dimension is intended to capture the engagement level of the community for the observed metric and is computed as the simple average over all weeks.

$$x_{engage} = \frac{1}{N} \sum_{i=1}^N x_i \quad (3.11)$$

where,  $x_i$  is observed value for metric  $x$  in the  $i^{th}$  week  
and,  $N$  is the number of weeks

- *Enthusiasm*: This dimension is intended to capture increase or decrease in the enthusiasm/immersion level of the community and is captured using the magnitude of increase or decrease in the observed metric over successive weeks. The measure is computed using the sum of the slopes over successive weeks. Since the recency of the changes matter, we use a linear weighting function such that the recent changes get more weight.

$$x_{enthu} = \sum_{i=1}^N w_i * (x_i - x_{i-1}) \quad (3.12)$$

where,  $x_i$  is observed value for metric  $x$  in the  $i^{th}$  week and  $x_0 = 0$ ;  
and,  $w_i$  is the week number such that higher-numbered weeks are closer to the

date of analysis;

- *Persistence*: This dimension is intended to capture the general *mood* in the observed past and is captured using the direction (up, down or level) of the change in the observed metric over successive weeks. The measure is computed using an indicator function  $Ind(x_i - x_{i-1})$  with a linear weighting function so that the recent changes get more weight.

$$x_{persist} = \sum_{i=1}^N w_i * Ind(x_i - x_{i-1}) \quad (3.13)$$

where,

$$Ind(x_i - x_{i-1}) = \begin{cases} 1, & \text{for } x_i - x_{i-1} > 0 \\ 0, & \text{for } x_i - x_{i-1} = 0 \\ -1, & \text{for } x_i - x_{i-1} < 0 \end{cases} \quad (3.14)$$

$$(3.15)$$

$$(3.16)$$

### 3.3.1 Choice of Algorithms and Parameters

One needs to make certain choices of algorithms and parameters at different stages of the proposed approach (refer figure 3.1). The key decision points are described below -

- *Unit of Analysis* We use week as the basic unit of analysis in this study and the prediction task is one of identifying if a community is going to churn in the week following the date of analysis. One can choose a finer (hourly or daily) or coarser (fortnightly or monthly) level of analysis depending on application requirements. But one would expect players to have more of a weekly cycle (e.g they might play more during the weekends).
- *Churn Threshold* The parameter defines when a community is said to have churned. In this study, we have tested our approach at three different thresholds - 100%, 75% and 50% i.e a community is said to have churned when all, 3/4th and 1/2 of its members are no longer part of the community in the next time interval.
- *Identification of Persistent groups* The key idea behind this step is to identify

long-term recurring relationships among players based on which we can build communities for subsequent steps. In this study, we have used recurring group quests that players embark upon to identify long-term relationships between players.

- *Community Detection algorithm* Several community detection algorithms have been proposed in the literature [112] and these approaches are based on traditional methods such as hierarchical, partitional and spectral clustering, modularity-based methods or divisive algorithms (e.g, Newman-Girven). In this study, we use the Louvain community detection algorithm which is a modularity-based method that has been shown to outperform other algorithms in terms of computation time and also gives high-quality partitions [160]. An implementation of the algorithm is also available in Pajek [161] which allows the communities and underlying networks to be visually analyzed.
- *Tracking dynamic communities* Existing approaches that track the evolution of communities over time use set similarity measures to compare communities in successive time periods [118, 119] and they use some form of the Jaccard similarity measure, as we have done in this chapter.

The choices to be made for these decision points would depend on the requirements of the application. For e.g, a Customer Relationship Management (CRM) analyst would typically choose the churn threshold depending on their domain and knowledge of customer base.

## 3.4 Experimental Results and Analysis

### 3.4.1 Data Description

Data from the MMORPG - Sony Everquest II <sup>1</sup> was used for the experiments and analysis. The game data had four servers and in-game relationships among players include housing (player grants access to his/her house to another player), mentoring (experienced player mentoring another player to help the latter acquire skills and level up within the game), trading (buying and selling of in-game items), chat and grouping.

---

<sup>1</sup> <https://www.everquest2.com/>

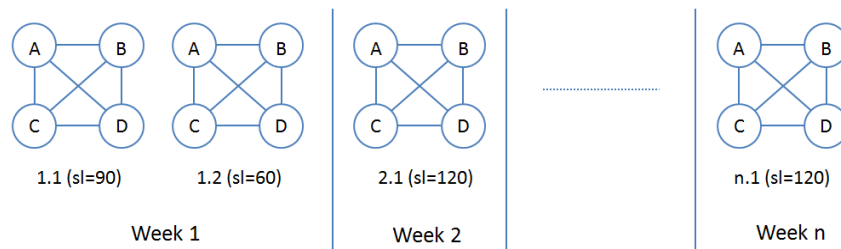


Figure 3.3: Group with members A, B, C, D as identified through weeks 1 through n

As a player levels up within the game, the quests become progressively more difficult and it becomes essential for players to play in groups to accomplish tasks and progress through the game<sup>2</sup>. For the purposes of our analysis, we used data from the *guk* server and the grouping relationships to build communities.

### 3.4.2 Identification of persistent groups

The game logs contain information of individual player actions but they do not explicitly identify group activities. However, the logs contain *group size* - so, if the player activity was logged while being part of a group, the value of *group size* would be greater than 1. In order to identify the groups, we parse the logs sequentially and look for entries which has the exact same values of *server name*, *log time*, *location id*, *group level* and *group size*. A tuple which meets these parameters is defined as a *group instance*. Identification of such group instances allow us to define groups which is a sequence of such group instances.

Figure 3.3 illustrates a group identified from game logs with members *A*, *B*, *C*, *D*. In week 1, the members participated in two group sessions of length 90 and 60 mins respectively. In week 2, they participated in a single group session of length 120 mins and so on. Using such an approach, we identify groups from Feb-May, 2006. We could identify 125758 such groups. Table 3.1 gives a breakdown of the groups by size and weekspan. Weekspan refers to the number of weeks across which the instances of a group were identified. We also identified 2260 groups which had sizes greater than 6 but none of these large-sized groups had lifespan greater than 0 and are not shown in Table 3.1, hence the total is listed as 123498. A key observation from Table 3.1 is that,

<sup>2</sup> <http://eq2.wikia.com/wiki/Groups>

Table 3.1: Player Groups

Group Size	2	3	4	5	6	Total
Weekspan						
0	27878	21305	19003	22327	22570	113083
1	1656	696	338	218	197	3105
2	815	340	140	68	34	1397
3	609	240	105	49	18	1021
4	534	210	85	31	14	874
5	434	170	51	26	7	688
6	389	115	47	10	6	567
7	279	90	47	13	3	432
8	269	79	26	10	5	389
9	248	73	22	8	5	356
10	205	61	20	8	7	301
11	194	53	19	4	2	272
12	197	46	8	3	1	255
13	167	45	11	5	1	229
14	179	17	6	1	2	205
Persistent Groups (1-14) Total	6175	2235	925	454	302	10415
Total	34327	23580	19937	22782	22872	123498

for each group size, as the lifespan increases the number of groups decrease i.e there are fewer groups which live longer.

We are interested in churn of long-lasting dynamically evolving communities and since our basic unit of analysis weekly, we only consider groups which have a weekspan of at least 1 - we call such groups as *persistent groups*. Such groups are indicative of a stronger sense of relationship as they represent recurring interactions among the same group of players over extended period of time. There are 10415 such *persistent groups* in our dataset, as highlighted in Table 3.1.

### 3.4.3 Network Construction

We next construct weekly networks from the *persistent groups*. An edge in this weekly network means the two players were part of at least one group session in that week and the edge weight represents the total amount of time the two players spent together. The



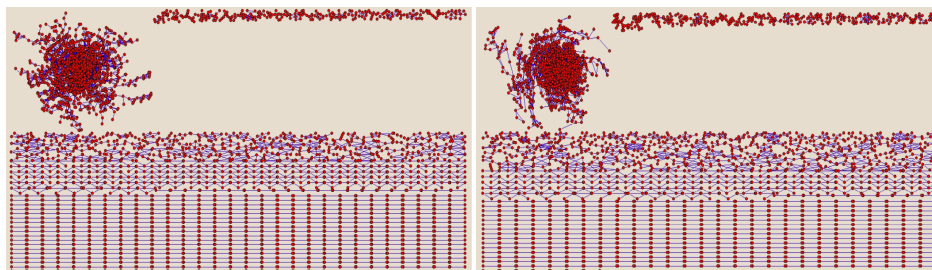


Figure 3.4: Network constructed from members of persistent groups - weeks 8 and 9.

process of edge weight calculation is described below.

In our illustrative example (refer figure 3.3), the first group session among members A, B, C and D last 90mins. We divide the total time with the assumption that each group member spends an equal amount of time with the other members in the group. In this case, we say that A spends 30 minutes each with B, C and D during the overall session of 90 minutes. Similarly each of the group members B, C and D spend 30 mins with each of the other members so that the weight of every edge for the first group session is 30. We repeat this process for every group session in a week. If there is more than one edge between two players in a week (i.e they were part of multiple sessions within the same group or part of multiple groups), we add up all the edges to get the final edge weight between the two players.

Figure 3.4 shows the networks constructed from persistent group for weeks 8 and 9 respectively. Pajek [161] program for analysis and visualization of large networks was used for this purpose. Table 3.2 shows some basic statistics for the networks constructed across weeks 8 through 15. Key observations from this table are -

- The largest component for each week comprises of only a small fraction of the nodes, an average of 22.43% across different networks. We saw from Table 3.1 that all the persistent groups ranged in size from 2 to 6 players. A network component of size greater than 6 thus means that two or more groups shared players in common during the week. The small size of the largest component indicates that most of the persistent groups in a week do not share common players. This is also evident from the network components for weeks weeks 8 and 9 (refer figure 3.4) where most of the components are of size six or smaller (bottom half of both

Table 3.2: Weekly networks constructed from *persistent groups*

Week	8	9	10	11	12	13	14	15	Avg
Nodes	3845	3571	3479	3317	3241	3154	2973	2893	3309
Edges	7915	7530	5759	5111	4445	4479	4006	3688	5366.6
Avg. degree	4.12	4.22	3.31	3.08	2.74	2.84	2.69	2.55	0.001
Node churn rate (%)	-	33	31.7	34.46	35.18	34.68	37.48	37.94	34.92
Edge churn rate (%)	-	62.36	64.2	60.93	60.4	55.59	55.3	60.83	59.94
Components	832	709	762	766	792	782	764	769	772
LC nodes	1520	1328	1046	877	444	489	254	244	775.25
LC node(%)	39.53	37.19	30.07	26.44	13.70	15.50	8.54	8.43	22.43
Communities	898	767	813	807	827	805	788	795	812.5
Modularity	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98

figures).

- The average node churn rate across consecutive weeks is 34.92% (std dev= 2.24) and the average edge churn rate is 59.94% (std dev= 3.33). We observe from Table 3.1 that as the lifespan increases the number of groups decrease i.e there are fewer groups which live longer. The short lifespan of most groups explain the relatively high node and edge churn rate from week to week and indicate that the networks are quite dynamic in nature.

### 3.4.4 Community detection

We run community detection algorithm on each weekly network to get communities for each week. The communities are identified using the Louvain community detection algorithm [160] in Pajek. The bottom two rows of Table 3.2 show the number of identified communities alongwith the modularity values of each weekly partition (set of communities). We observe that the identified communities have very high modularity value (avg=0.98, std dev=0.005) which implies that the identified clusters are very pure. Figure 3.5 shows the communities identified for weeks 8 and 9. As expected, the largest component is partitioned further into communities whereas the smaller components (dyads, triads etc) are treated as single communities.

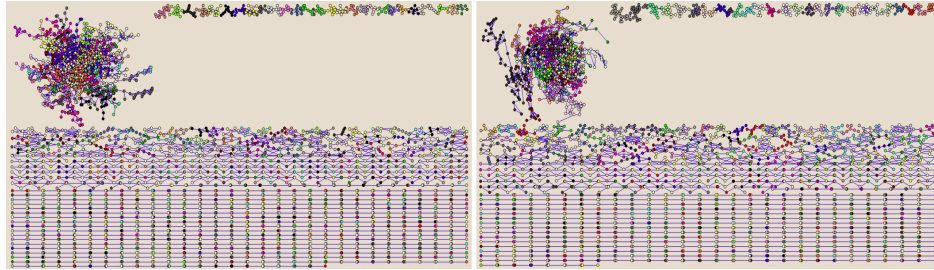


Figure 3.5: Communities for weeks 8 and 9.

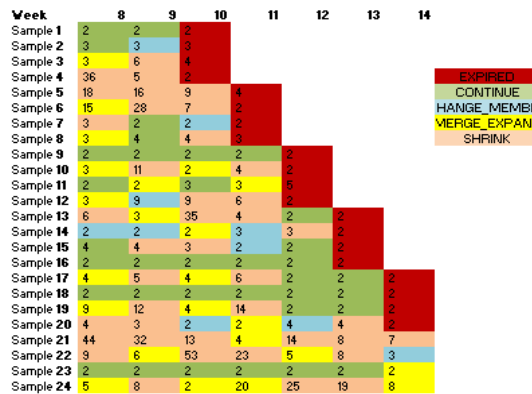


Figure 3.6: Samples of dynamic communities identified using community tracking algorithm

### 3.4.5 Community Evolution

Once the communities have been identified for each week, we track the evolution of communities across weeks using Algorithm 3.2. Figure 3.6 illustrates the output of the Community Tracking Algorithm for a churn threshold of 100 (i.e a community is labeled as EXPIRED when all of its members leave). We start with the communities identified in week 8 and then track the evolution of a community from weeks 8 through 14. Color indicates the output label assigned to the community based on its transition from current week to next. The samples in figure 3.6 explained below -

- Sample 1: The community starts in week 8 with size 2, remains unchanged in weeks 9 (green) and 10 (green) and expires in week 11 (red).
- Sample 2: The community starts in week 8 with a size of 3, remains unchanged

Table 3.3: Community Churn Datasets

Churn Threshold	Churner(Y)	Non-churner(N)	Total samples	Ratio(Y:N)
100%	380	2396	2776	13.69:86.31
75%	494	425	919	53.75:46.25
50%	365	171	536	68.09:31.93

in week 9 (green), changes membership but not size in week 10 (blue) and expires in week 11 (red).

- Sample 3 - The community starts in week 8 with a size of 3, merges/expands in week 9 (yellow), shrinks in week 10 (pink) and expires in week 11 (third red).
- Sample 4 - The community starts in week 8 with size 36, shrinks in week 9 (pink) and week 10 (pink) and expires in week 11 (third red).

Samples 5 through 20 are examples of communities which expire in weeks 11, 12, 13 and 14 and samples 21 through 24 are examples of communities which do not expire during the whole period.

### 3.4.6 Community Churn Prediction

The output labels of the *Community Tracking Algorithm* allows us to construct a labeled classification dataset. For each of the weeks 10, 11, 12, 13, 14 (refer figure 3.6), positive samples (churners) are communities which are labeled EXPIRED for the following week whereas negative samples (non-churners) are communities which have non-EXPIRED labels. We use samples from all the weeks to get the overall classification dataset.

We test our approach at three different levels of churn threshold - 100%, 75% and 50%. Table 3.3 gives a breakdown of the samples in the datasets built using the three thresholds. We observe that at 100% threshold, this is an imbalanced class problem with the target class (churner) comprising only 13.69% of the samples. At 75% threshold, the classes are almost equally balanced with a churner:non-churner ratio of 53.75:46.25 whereas at the 50% threshold, the target class becomes the majority class with 68% of the total samples.

As part of the lifecycle-based approach we consider the activity traits of each community leading upto the week of analysis and in this chapter, we consider three weeks of

Table 3.4: Community Churn Model: Info-gain ranking of features

Info-gain Ranking	Feature	Info-gain Value
1	Community_overall_engage	0.21847
2	Connectedness_enthu	0.20904
3	Community_social_engage	0.20779
4	Community_intra_engage	0.20273
5	Community_overall_enthu	0.18617
6	Size_engage	0.18055
7	Connectedness_engage	0.17697
8	Size_enthu	0.14952
9	Community_social_enthu	0.14704
10	Member_loyalty_engage	0.1293
11	Community_intra_enthu	0.12554
12	Member_loyalty_enthu	0.10485
13	Member_loyalty_persist	0.10485
14	Connectedness_persist	0.10353
15	Size_persist	0.0929
16	Community_inter_enthu	0.09289
17	Community_inter_engage	0.08297
18	Community_inter_persist	0.06347
19	Member_social_engage	0.04627
20	Member_social_enthu	0.04019
21	Community_intra_persist	0.00721
22	Community_social_persist	0.00378
23	Member_social_persist	0
24	Community_overall_persist	0

history. This choice of history length is based on previous work where we had found that three weeks of history is sufficient for getting high quality prediction results. So, for each sample in the dataset and for the three weeks prior to the week of analysis, we compute each of the eight community metrics (refer *Methodology* section). We then apply the three semantic dimensions (refer *Methodology* section) to get 24 derived features for every sample. Table 3.4 lists the features in decreasing order of information gain for the 100% churn threshold dataset. Information gain measures how well a given attribute separates the training examples according to the target classification and is given by the expected reduction in entropy caused by partitioning the examples according to the attribute [144].

We also used the Correlation based Feature Selection (CFS) technique [145] to identify the subset of features which are highly correlated with the class while having low intercorrelation. As per this measure, the following features were found to be highly discriminating and uncorrelated to one another

- *Community\_intra\_engage*; three-week average of member’s within community playtime.
- *Community\_overall\_engage*; three-week average of member’s overall playtime.
- *Connectedness\_engage*; three-week average of member’s connectedness.
- *Size\_enthu*: the change in community size leading upto the week of analysis. We found that communities that churned progressively decreased in size before churning.

These results indicate that communities display an overall decrease in engagement leading upto the point of churn, which is reflected in lower average playtime and connectedness of its members and also the gradual decrease in community size. We performed info-gain experiments on the 75% and 50% churn threshold datasets (not listed here) as well and found the feature ranking to be similar to the above.

Finally, we perform 10-fold cross-validation on the three datasets. Table 3.5 shows the precision, recall and F-measure values for the churning class with J48, JRip, Logistic Regression, NaiveBayes and k-NearestNeighbor (k=3) using the Weka [146] machine learning tool. At the 100% churn threshold, we observe that J48 and JRip give near equal F-measure performance of around 75. The JRip rule-based classifier gives precision, recall and F-measure of 78.8, 72.37 and 75.45 respectively. This is quite significant since the target class is only 13.69% - so a naive minority classifier would have a precision of 13.69 only. For the other two churn thresholds, we find that JRip and J48 perform well with F-measure values above 75.

### 3.5 Conclusion and Future Work

In this chapter, we propose a framework for predicting churn of dynamically evolving communities using a supervised learning approach. We first identify persistent in-game

Table 3.5: Community Churn Model: 10-fold cross validation results

Churn Threshold	Classifier	Precision	Recall	F-measure
100%	J48	83.39	68.68	75.32
	JRip	78.8	72.37	75.45
	Logistic	62.93	38.42	47.71
	Naive Bayes	30.56	92.89	45.99
	kNN	81.44	56.58	66.77
75%	J48	83.15	77.94	80.46
	JRip	71.71	80.57	75.88
	Logistic	65.48	78.34	71.34
	Naive Bayes	59.8	86.44	70.7
	kNN	81.39	66.4	73.13
50%	J48	86.88	81.64	84.18
	JRip	77.54	87.95	82.41
	Logistic	72.23	91.23	80.63
	Naive Bayes	75	18.08	29.14
	kNN	85.08	73.42	78.82

player groups which represent strong long-term social relationship among players and using such relationships, we measure the strength of connections among players based on the amount of time they spend together in such groups. We then construct weighted networks for each time period, identify player communities for each week and then track the evolution of a community over weeks. We calculate several intuitive community-based metrics for each time period and use a lifecycle-based approach to build the final prediction model. We test our approach at three different levels of community churn thresholds. In all three cases, we find that J48 and JRip classifiers perform quite well with F-measures above 75. Furthermore, analysis of the community metrics indicate that communities display an overall decrease in engagement leading upto the point of churn which is evident from the lower average playtime and connectedness of its members and also the gradual decrease in community size.

To the best of our knowledge, no prior work has addressed the specific problem of community churn prediction in online games or proposed a general framework for predicting churn of dynamically evolving communities using a supervised learning approach. There are several choices that one can make at each step of the approach. As

future work, we would like to investigate the impact of other parameter and algorithmic choices on the overall results. We would also like to use the proposed approach to study the impact of individual community members on community churn and identify members who are influential for community survival.



## Chapter 4

# Impact of Achievement and Socialization Factors on Individual Churn

The goal of this chapter is to study the impact of different motivational factors on player churn. We posit that players with greater motivation and involvement are less likely to leave a game and hence player motivation is one of the most important factors that can help in analyzing and predicting churn behavior. Bartle had proposed player "types" based on a taxonomy of player motivation [162]. While the model had served as the industry standard for a long time, it did not have an empirical basis. A more recent and empirically based taxonomy by Yee uses validated scales to detect player motivations, but does not preclude multiple "types" from existing within a single player. Yee used a factor analytic approach to create an empirical model of player motivations [150]. The analysis revealed 10 motivation subcomponents that grouped into three overarching components (achievement, social, and immersion) with underlying relationships between motivations and demographic variables (age, gender, and usage patterns), as illustrated in Table 4.1.

In order to study the impact of different motivational factors on player churn, we first segment the player population into loners and socializers, identify features which represent achievement and socialization orientation of players and then apply these two

Table 4.1: Model of Player Motivation

<b>Achievement</b>	<b>Socialization</b>	<b>Immersion</b>
<i>Advancement</i> Progress, Power, Accumulation, Status	<i>Socializing</i> Casual Chat, Helping Others, Making Friends	<i>Discovery</i> Exploration, Lore, Finding Hidden Things
<i>Mechanics</i> Numbers, Optimization, Templating, analysis	<i>Relationship</i> Personal, Self-disclosure, Find and Give, Support	<i>Role-playing</i> Story Line, Character History, Roles, Fantasy
<i>Competition</i> Challenging Others, Provocation, Domination	<i>Teamwork</i> Collaboration, Groups, Group Achievements	<i>Customization</i> Appearances, Accessories, Style, Color Schemes
		<i>Escapism</i> Relax, Escape from Real Life, Avoid Real-Life Problems

models to the two players segments. We also study and compare the behavioral profiles of the different player segments.

We find that loners are much more likely to churn than socializers. A large fraction of loners churn from the game and they tend to be players who are in the initial stages of the game and are not very engaged or progressing through the game levels. Among socializers, we observe that there is a near equal likelihood for a player to churn or not. In the socializer segment, churners show little to no activity leading up to the point of analysis whereas non-churners show a steady activity level. Also, non-churners were observed to be in higher levels within the game as compared to churners. Furthermore, we observe that churners interact more with other churners in the weeks leading up to the point of churn and consequently, the fraction of activity in groups decrease for churners as compared to non-churners. Finally, we build models using Achievement and Socialization-oriented features and find the models to have good predictive performance in identifying churners. Furthermore, the socialization features were found to improve model performance by a significant amount.

The experimental setup and analysis of the study are described next.

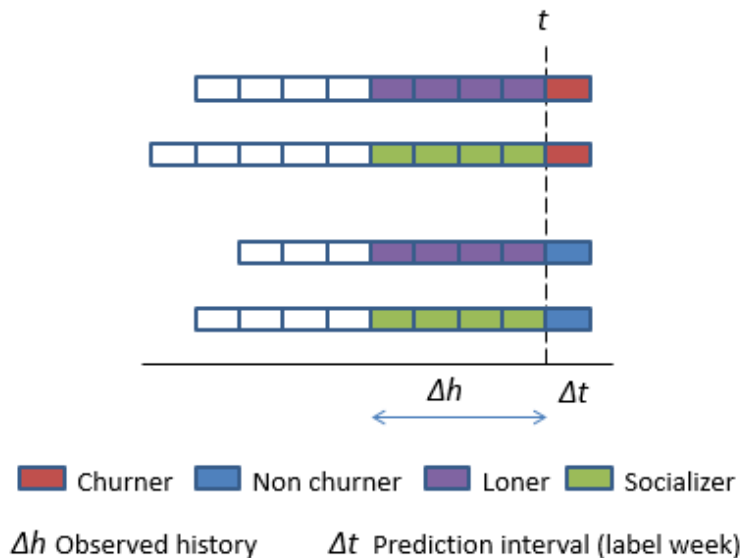


Figure 4.1: Loners and Socializers: Experimental Setup

## 4.1 Data Description and Experimental Setup

Data from the MMORPG - Sony Everquest II was used for the experiments and analysis. The game data had four servers and in-game relationships among players include housing (player grants access to his/her house to another player), mentoring (experienced player mentoring another player to help the latter acquire skills and level up within the game), trading (buying and selling of in-game items), chat and grouping. As a player levels up within the game, the quests become progressively more difficult and it becomes essential for players to play in groups to accomplish tasks and progress through the game. For the purposes of our analysis, we used data from the guk server and the grouping relationships to build communities.

Figure 4.1 shows the experimental setup. Lets say  $t$  is the point of analysis. In keeping with our lifecycle-based approach proposed in the previous chapter, we observe various activity traits of the player for history period  $(t - \Delta h)$  to  $t$  and predict whether the player is going to churn in the prediction period  $t + \Delta t$ . The basic time unit of analysis in this setup is  $\Delta t$ . If we assume that the history length  $\Delta h$  can be split into  $h$  intervals, each of length  $\Delta t$ , then we observe and measure various player features for

Table 4.2: Loners and Socializers: Weekly Datasets

<b>Week Num</b>	11	12	13	14	15
<b>Loner</b>	12263	11020	10094	9687	9327
<b>Socializer</b>	13358	12698	11738	11369	11040
<b>Churner</b>	17222	15479	13774	13293	12634
<b>Non-churner</b>	8399	8239	8058	7763	7733
<b>Total</b>	25621	23718	21832	21056	20367

each of the basic time intervals. Key labels used in this setup are defined below -

- *Churner*: defined as a player with *no* activity in the prediction interval  $\Delta t$ .
- *Non-churner*: defined as a player with activity in the prediction interval  $\Delta t$ .
- *Loner*: defined as a player who only engaged in solo-play during the observed history period  $\Delta h$ .
- *Socializer*: defined as a player who engaged in any group activity with other players during the observed history period  $\Delta h$ .

We segment the player population into *loners* and *socializers* to clearly measure the efficacy of socialization on churn. We identify features which represent *achievement* and *socialization* orientation of players and apply these two models to the two player segments in order to understand the impact of these motivational factors on player churn. The operational definitions of the various elements used in the experimental setup are listed below -

- *Unit of Analysis,  $\Delta t$* : The basic time unit of analysis used in the experiments is *weekly*. A finer (hourly or daily) or coarser (fortnightly or monthly) level of aggregation could have been chosen but one would expect players to have more of a weekly cycle (e.g they might play more during the weekends).
- *Observed History  $\Delta h$* : We observe four weeks of player history.
- *Prediction Interval  $\Delta h$* : We predict if the player is going to churn in the week following the point of analysis  $t$ .

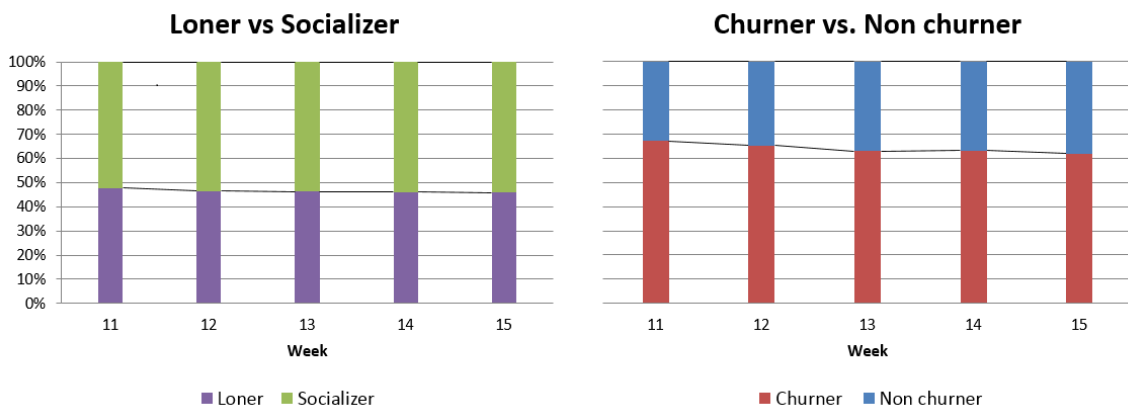


Figure 4.2: Loners and Socializers: Data description

With this setup, we thus observe roughly a month of player history to predict *next week* churn. The game dataset has complete player logs from January to September, 2006 - this is about 39 weeks of player activity. We consider weeks 7 through 15 for our analysis (2/12/06 - 4/15/06). Table 4.2 gives a breakdown of churners/non-churners and loners/socializers for weeks 11 through 15. If we consider week 11, we observe that there were a total of 25621 players who had some activity in the previous four weeks (7 through 10) and is our sample population for the week. Out of these, 17222 are churners (had no activity in week 11) and 8399 are non-churners. Similarly, 12263 of them were loners (played solo-only in weeks 7 through 10) and 13358 are socializers (engaged in some group activity in weeks 7 through 11).

Figure 4.2 shows the percentage breakdown of player samples over the weeks by loner/socializer and churner/non-churner. We observe that there is roughly an equal distribution of loners and socializers every week while 60-70% of the players are not active (churner) in any given week. Figure 4.3 shows the percentage breakdown of churners and non-churners for the *loner* and *socializer* segments. We clearly observe that among loners, the percentage of churners far outweigh the non-churners across all weeks with a ratio of about 81:19. On the other hand, for socializers, there is roughly an equal distribution of churners and non-churners every week. Thus, there is strong evidence to suggest that *loners are more likely to churn than socializers*. In order to

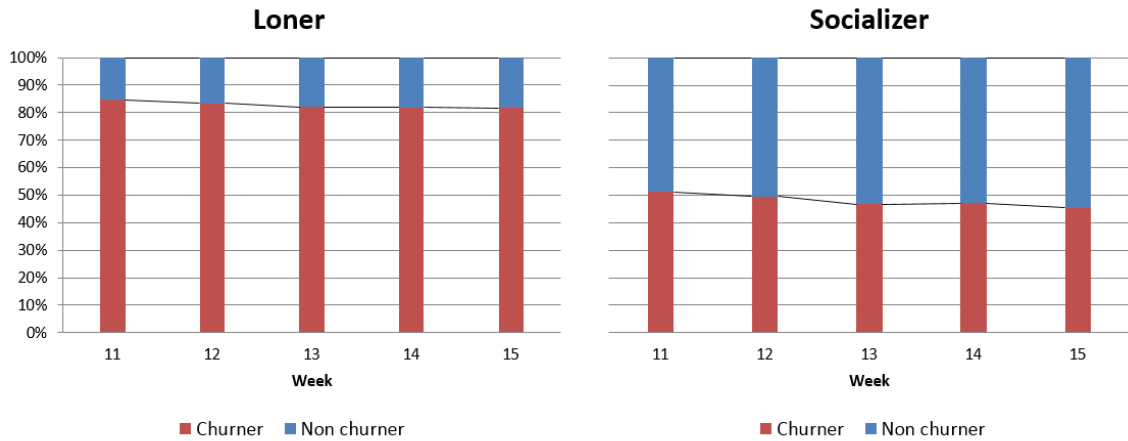


Figure 4.3: Loners and Socializers: Proportion of Churners

validate this, we pose the following null hypothesis

$H_0$ : There is no difference in the proportion of churners for loners and socializers.

For our dataset, we consider the 52931 loners and 60203 socializer for weeks 11 through 15 (refer table!4.2). The proportion of churners in these populations were - 0.83 for loners and 0.48 for socializers. A two-sample significance test for difference in proportions rejects the null hypothesis  $H_0$  ( $p < 0.001$ ).

## 4.2 Feature Space

### 4.2.1 Weekly Activity Metrics

As described in the experimental setup section, we calculate various activity metrics of a player every week in the history period  $\Delta h$  leading upto the point of analysis  $t$ . Table 4.3 lists the weekly player metrics, categorized by whether the metric is achievement or socialization-oriented - the categorization was done based on the Model of Player Motivation 4.1.

The *Achievement-oriented* metrics capture the psychological factors which motivate

Table 4.3: Weekly Achievement and Socialization Metrics

<b>Achievement-oriented</b>	<b>Socialization-oriented</b>
Number of Sessions	Number of Group Sessions
Session Length (mins)	Group Session Length (mins)
Inter-session length (mins)	Degree Centrality
Character Level	Closeness Centrality
Total Experience Points	Betweenness Centrality
Number of Kills	Eigenvector Centrality
Number of Deaths	Clustering coefficient
	Number of Churner Neighbors
	Fraction Churner Neighbors
	Churner Session Length (mins)
	Fraction Churner Session Length (mins)

a player based on the his/her degree of in-game engagement and success. The session-related metrics - *Number of Sessions*, *Session Length (mins)* and *Inter-session length (mins)*, capture the player’s overall engagement within the game. The activity logs available to us only record player actions and did not explicitly define sessions. So, we used a simple heuristic to define player-sessions - a session consists of sets of activities which are separated by no more than 30 minutes. The other achievement-oriented metrics - *Character Level*, *Total Experience Points*, *Number of Kills*, *Number of Deaths*, capture how successful the player is in the various in-game activities. For example, lets say Player *A* is making constant progress in the game by accumulating experience points, killing in-game monsters and leveling quickly while Player *B* is frequently dying and unable to progress within the game. In such a scenario, one would expect player *A* to have a greater sense of success than player *B* and hence is more likely to continue playing.

The *Socialization-oriented* metrics capture the social interactions of players - we use the in-game group interactions for this purpose. For each week in the history period  $\Delta h$ , we first construct the group interaction network; where a node represents a player, an edge represents a group interaction between two players and the edge weight is calculated based on the session length (in minutes) the two players spent together as part of groups in that week. So, the node degree and sum of edge weights in the weekly network gives us the metrics - *Number of Group Sessions* and *Group Session Length*

(mins). We then compute various node centrality metrics - *Degree Centrality* [163], *Closeness Centrality* [164], *Betweenness Centrality* [165] and *Eigenvector Centrality*, which measure the relative importance of a node within a graph. We also measure *Clustering Coefficient* [166]; which measures the fraction of a node's neighbors who are also neighbors with one another. We think these network topological measures can be a good indicator of a player's degree of social connectedness and their relative importance within their game community and can thus have an impact on his/her churn likelihood.

Studies have found that interactions with other churners in the network have an impact on an actor's churn likelihood [1, 13] - this is an important aspect that we investigate in this paper. For this purpose, we first identify players who have churned in the week before the point of analysis  $t$ . Thereafter, for each weekly group network constructed as described earlier, we compute the metrics - *Number of Churner Neighbors*, *Fraction of Churner Neighbors*, *Churner Session Length (mins)* and *Fraction Churner Session Length (mins)*.

#### 4.2.2 Trend Metrics or Semantic Dimensions

We apply three semantically meaningful dimensions or trend metrics, as proposed in our previous paper [167], along which the weekly history of an observed metric is recomputed to get derived semantic features. The semantic dimensions are -

- *Engagement*: This dimension is intended to capture the engagement level of the player for the observed variable and is computed as the *weighted moving average* over the history period  $\Delta h$

$$x_{engage} = \frac{1}{N} \sum_{i=1}^N x_i \quad (4.1)$$

where,  $x_i$  is observed value for variable  $x$  in the  $i^{th}$  week  
and,  $N$  is the number of weeks in the history period  $\Delta h$

- *Enthusiasm*: This dimension is intended to capture increase or decrease in the engagement level of the player and is captured using the magnitude of increase or decrease in the observed behavior over successive weeks. The measure is computed using the sum of the slopes over successive weeks. Since the recency of the changes



matter, we use a linear weighting function such that the recent changes get more weightage.

$$x_{enthu} = \sum_{i=1}^N w_i * (x_i - x_{i-1}) \quad (4.2)$$

where,  $x_i$  is observed value for variable  $x$  in the  $i^{th}$  week and  $x_0 = 0$ ;  
 $w_i$  is the week number such that higher-numbered weeks are closer to the date of analysis;  
and,  $N$  is the number of weeks in the history period  $\Delta h$

- *Persistence*: This dimension is intended to capture the general *mood* of the player in the observed past and is captured using the direction (up, down or level) of the change in the observed behavior over successive weeks. The measure is computed using an indicator function  $Ind(x_i - x_{i-1})$  with a linear weighting function so that the recent changes get more weightage.

$$x_{persist} = \sum_{i=1}^N w_i * Ind(x_i - x_{i-1}) \quad (4.3)$$

where,  $x_i$  is the observed value for variable  $x$  in the  $i^{th}$  week and  $x_0 = 0$ ;  
 $w_i$  is the week number ;  
 $N$  is the number of weeks;  
and,

$$Ind(x_i - x_{i-1}) = \begin{cases} 1, & \text{for } x_i - x_{i-1} > 0 \\ 0, & \text{for } x_i - x_{i-1} = 0 \\ -1, & \text{for } x_i - x_{i-1} < 0 \end{cases} \quad (4.4)$$

$$0, \quad \text{for } x_i - x_{i-1} = 0 \quad (4.5)$$

$$-1, \quad \text{for } x_i - x_{i-1} < 0 \quad (4.6)$$

It should be noted that in the final feature space, we get three features from each weekly observed metric (one for each trend metric). With reference to table 4.3, there are 7 achievement-oriented weekly metrics and 11 socialization-oriented weekly metrics. So, in terms of the final prediction models, the *Achievement* model has 21 features and the *Socialization* model has 33 features.

Table 4.4: Loners and Socializers: Week 12 Dataset

	<b>Loner</b>	<b>Socializer</b>	<b>Total</b>
<b>Churner</b>	9197	6282	15479
<b>Non-churner</b>	1823	6412	8239
<b>Total</b>	11020	12968	23718

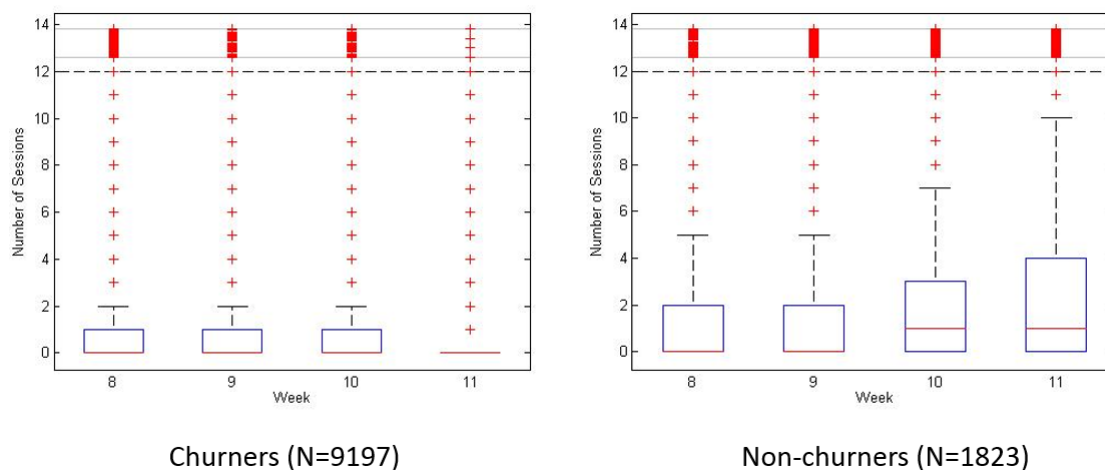


Figure 4.4: Loner segment - comparison of session activity for churners and non-churners

### 4.3 Experimental Results and Analysis

We have outlined the experimental setup where we segmented the player population into *loners* and *socializers*. Next, we described the *Achievement* and *Socialization* models in terms of their features. We now apply the two models to the two player segments in order to understand the impact of these motivational factors in player churn.

For the following experiments, we construct the models using four weeks of data (weeks 8 through 11) and predict whether a player is going to churn in week 12 in order to check the efficacy of the models. The breakdown across categories is given in Table 4.4

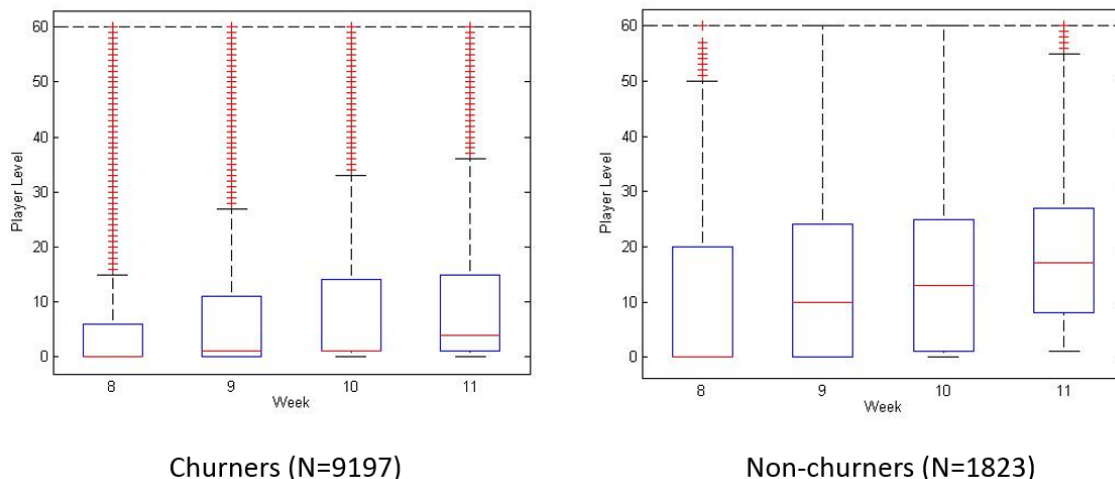


Figure 4.5: Loner segment - comparison of leveling activity for churners and non-churners

### 4.3.1 Analysis of Behavioral Profiles

We first visualize and compare the behavioral profiles of churners and non-churners. In order to do this, we take the weekly histories of the activity metric we would like to compare (refer table 4.3) and then for both churners and non-churners, we construct boxplots for each of the weeks in the history period (weeks 8 through 11).

Figure 4.4 shows the behavioral profiles of churners and non-churners based on the the metric - *Number of Sessions* for the *loner* segment. We observe that for loners in general, the number of sessions every week is quite low (the median being less than 2) and there are differences in the trend between the two populations. For churners, the median number of sessions is 0 throughout the four weeks - there is a higher spread in weeks 8 through 10 which drops to 0 in week 11 (the week before prediction). For non-churners, on the other hand, the median number of sessions shows an increase leading upto the week of analysis - the median is 0 for weeks 8 and 9 and increases to 1 for weeks 10 and 11.

Figure 4.5 shows the behavioral profiles of churners and non-churners based on the the metric - *Character Level* for the *loner* segment. For churners, the character level is quite low throughout the four weeks (median being less than 5) and we do not observe a

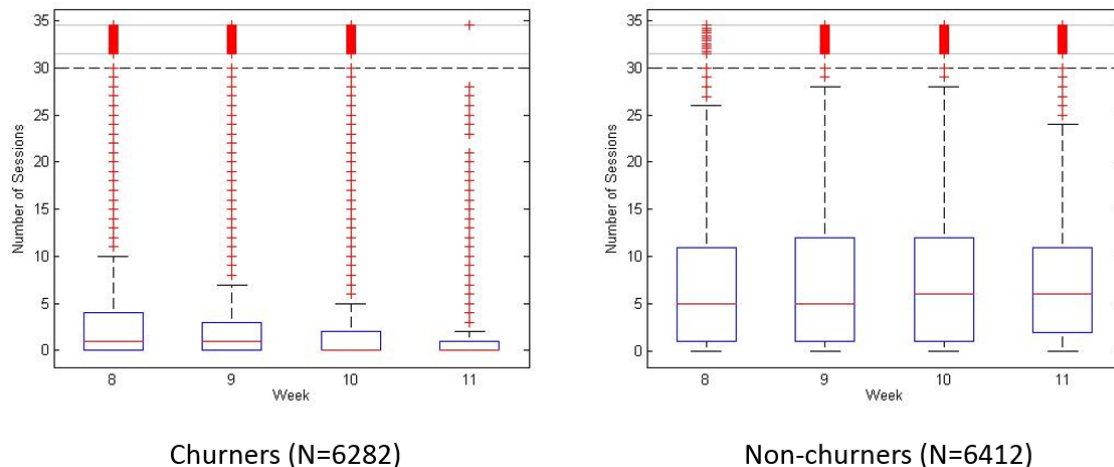


Figure 4.6: Socializer segment - comparison of session activity for churners and non-churners

substantial increase in character levels before the point of analysis. This indicates that in general, loners who churn tend to be players who are in the initial stages of the game and are not very engaged or progressing through the game levels. For non-churners, on the other hand, we observe the median character level is more than 10 in weeks 9 through 11, which is substantially higher than that for churners.

Figure 4.6 shows the behavioral profiles of churners and non-churners based on the the metric - *Number of Sessions* for the *socializer* segment. We observe that for churners, the median number of sessions is quite low throughout the four weeks with the median being less than 2 in weeks 8 and 9 and dropping to 0 in the week before prediction. For non-churners, on the other hand, the median number of sessions remain more or less constant around 5 throughout the four weeks.

If we compare the behavioral profile of loners and socializers based on the the metric - *Number of Sessions* (refer figures 4.4 and 4.6), we observe differences between the two segments; particularly for non-churners. Among non-churners, loners show little to no activity whereas socializers show a constant level of activity with a median of around 5 throughout the history period.

Figure 4.7 shows the behavioral profiles of churners and non-churners based on the

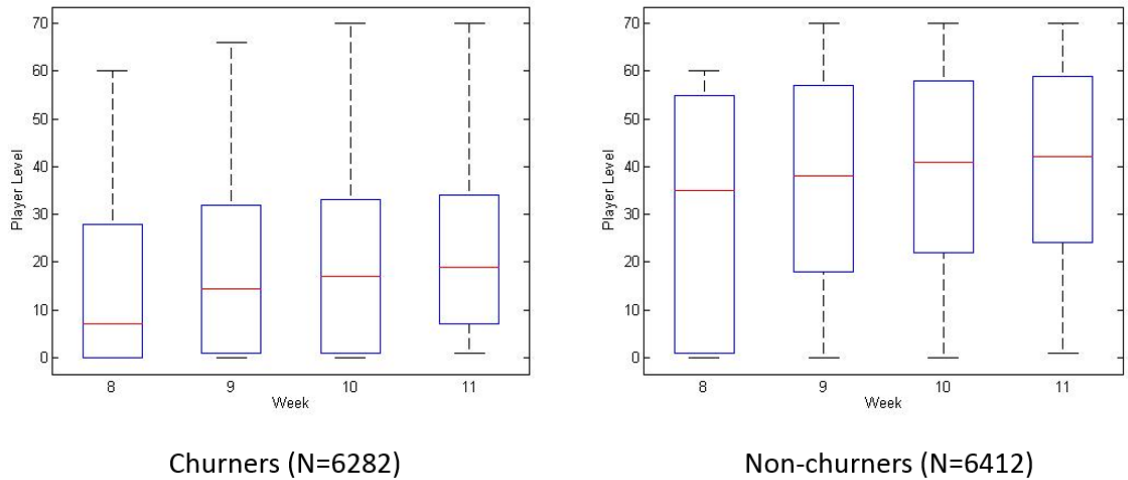


Figure 4.7: Socializer segment - comparison of leveling activity for churners and non-churners

the metric - *Character Level* for the *socializer* segment. For churners, the median character level is less than 20 throughout the four weeks while for non-churners, the character level is close to 40 in the weeks leading up to the point of analysis. If we compare the profiles to that of loners (Figure 4.5), we find that in general socializers are in higher levels of the game.

If we compare the behavioral profile of loners and socializers based on the the metric - *Character Level* (refer figures 4.5 and 4.7), the differences between the two segments are even more pronounced. Among churners, loners tend to be players who are in the initial stages of the game and are not very engaged or progressing through the game levels while socializers are at higher levels in the game, with a median close to 20 in the two weeks before the point of analysis. We observe a similar trend of differences among non-churners - loners are at lower levels in the game as compared to socializers. The median level for loners increase steadily from 10 to 20 in the three weeks leading up to the point of analysis whereas the median level for socializers remains close to 40 throughout the history period.

Table 4.5: Loners and Socializers: Performance of Achievement Model

Model	Loner Segment			Socializer Segment		
	Precision	Recall	F-measure	Precision	Recall	F-measure
J48	88.1	95.04	91.44	76.34	79.83	78.05
JRip	88.59	95.24	91.79	77.66	79.62	78.63
Logistic	87.21	97.63	92.13	74.3	83.7	78.72
Naive Bayes	88.65	92.67	90.62	66.69	88.11	75.92
kNN	87.85	93.2	90.45	74.26	77.87	76.02
AdaBoost	88.4	92.93	90.27	76.49	76.92	76.7
Bagging	88.5	95.51	91.87	78.34	80.63	79.47

### 4.3.2 Impact of Achievement Factors on Churn

We used the *Achievement* model (refer Table 4.3) to predict churners for loners and socializers. Table 4.5 shows the precision, recall and f-measure results for the two segments using JRip, J48, NaiveBayes, Logistic Regression, NaiveBayes, k-NearestNeighbor (k=3), AdaBoost and Bagging. JRip gives the best F-measure in both scenarios with (precision, recall, f-measure) values of (88.59, 95.24, 91.79) and (77.66, 79.62, 78.63) respectively for loners and socializers respectively. All the models perform better than random for both scenarios, as the ratio of churners to non-churners is 83:17 for loners and 49:51 for socializers. An *info-gain* ranking of the 21 features of the Achievement model listed showed that the engagement metric related to *Number of Sessions*, *Session Length*, *Experience Points* and *Character Level* are highly discriminating in predicting churners for both loners and socializers.

The *Achievement* model capture the psychological factors which motivate a player based on the his/her degree of in-game engagement and success. We found that, in general, players who are more active and making progress within the game are less likely to churn.

### 4.3.3 Impact of Socialization Factors on Churn

In this section, we to study the impact of socialization factors on player churn. For each week in the history period  $h$ , we first construct the group interaction network from the game dataset; where a node represents a player, an edge represents a group interaction between two players and the edge weight is calculated based on the session length (in

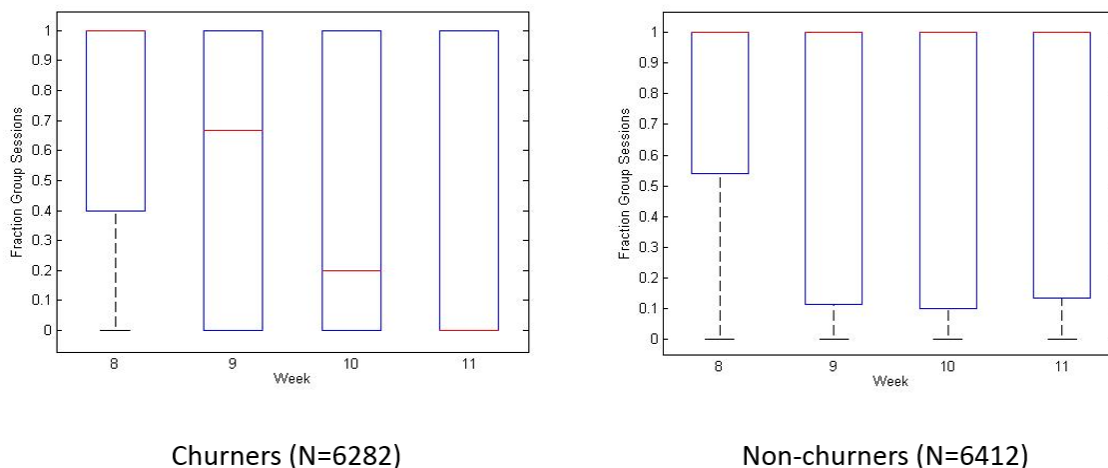


Figure 4.8: Socializer segment - comparison of grouping activity for churners and non-churners

minutes) the two players spent together as part of groups in that week.

Figure 4.8 shows the behavioral profiles of churners and non-churners based on the metric - *Fraction of Group Sessions*; which captures the fraction of overall number of sessions that the player spends in groups. We observe a sharp contrast between the churners and non-churner profiles. For churners, the median fraction of group sessions decrease steadily from 1 to 0 in the weeks leading upto the point of analysis whereas for non-churners, the median fraction of group sessions remains consistently at 1 throughout the history period.

Studies have found that interactions with other churners in the network have an impact on an actor's churn likelihood - this is an important aspect that we investigate in this chapter. We observed from figure 4.8 that the fraction of group activities for churners decrease before the point of churn. We wanted to investigate if there was an interaction effect with previous week churners. Specifically, we would like to answer the question - *Do churners interact more with other churners in the weeks leading upto the point of churn?*

In order to answer the above question, we consider the node labels in the weekly group interaction networks. As earlier we consider the churners and non-churners for

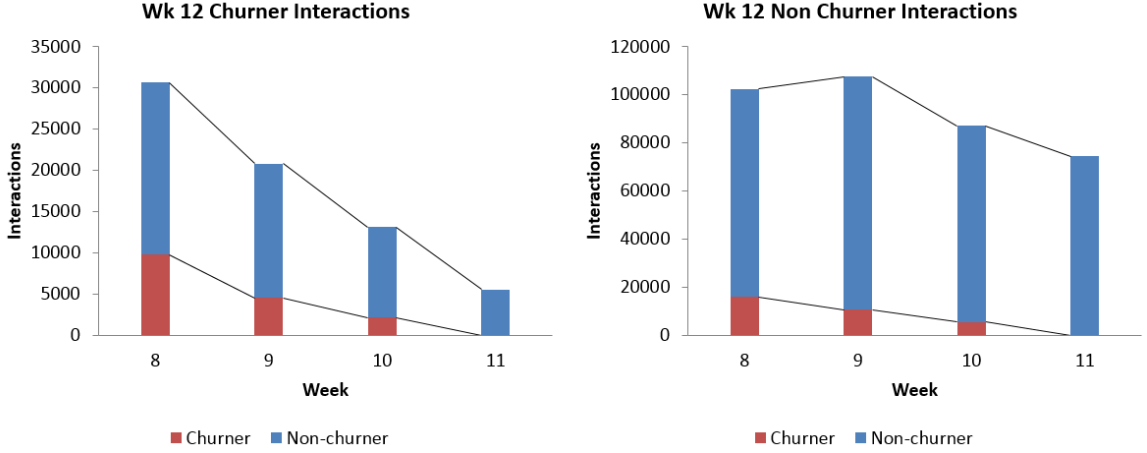


Figure 4.9: Socializer segment - comparison of interaction with previous week churners

Table 4.6: Dataset for testing hypothesis  $H'_0$

Population	n	X	p=X/n
Churner	70203	16492	0.24
Non-churner	371490	32506	0.09

week 12; but we also identify the player labels in the previous week i.e week 11. Figure 4.9 shows the comparison of interactions with previous week churners. The left graph shows the interactions of week 12 churners with week 11 churners and non-churners while the right graph shows the interactions of week 12 non-churners with week 11 churners and non-churners. The overall number of interactions for week 12 churners drops steadily over time from 30000 to around 5000 whereas for non-churners, the overall number of interactions is quite high as compared to churners and does not drop as steadily. From the breakdown of the bar graphs in both the figures, we also observe that fraction of overall interactions with previous week churners is higher for churners than for non-churners. Thus, there is strong evidence to suggest that churners interact more with other churners than non-churners. In order to validate this, we pose the following null hypothesis

$H'_0$ : There is no difference in the proportion of interactions with prior week churners between current churners and non-churners.



Table 4.7: Performance of Achievement and Socialization Factors

Model	Achievement			Achievement + Socialization		
	Precision	Recall	F-measure	Precision	Recall	F-measure
J48	76.34	79.83	78.05	85.52	83.78	84.64
JRip	77.66	79.62	78.63	81.72	82.05	81.88
Logistic	74.3	83.7	78.72	80.47	82.23	81.34
Naive Bayes	66.69	88.11	75.92	69.83	87.91	77.83
kNN	74.26	77.87	76.02	76.92	83.42	81.48
AdaBoost	76.49	76.92	76.7	78.39	79.25	78.82
Bagging	78.34	80.63	79.47	80.14	81.89	81.01

Table 4.6 shows the dataset used for testing hypothesis  $H'_0$ . The two populations are the week 12 churners and non-churners;  $n$  is the number of overall interactions for the history period (weeks 8 through 11);  $X$  is the number of interactions with week 11 churners during the history period and  $p$  is the proportion being tested in the hypothesis. A two-sample significance test for difference in proportions rejects the null hypothesis  $H'_0(p < 0.001)$ .

Finally, we look at the impact of socialization factors on player churn. Table 4.7 shows the precision, recall and f-measure results using just Achievement factors alone and using both the Achievement and Socialization metrics. The metrics used in the models are listed in Table 4.3. As earlier, JRip, J48, NaiveBayes, Logistic Regression, NaiveBayes, k-NearestNeighbor (k=3), AdaBoost and Bagging models were used to evaluate performance. For each of the models, we observe that including the Socialization variables does improve performance by a significant amount. An *info-gain* ranking of the features showed that the Socialization factors which were most discriminating and improved overall performance (in combination with the Achievement factors) were the *Fraction of Churner Neighbors* and *Closeness Centrality*. We have already seen that churners interact more with other churners (refer hypothesis  $H'_0$ ) - the info-gain analysis further indicates that the fraction of neighbors who are churners is an important feature in improving the performance of a churn prediction model.

## 4.4 Conclusion and Future Work

We studied the impact of different motivational factors in player churn in this chapter. Towards this end, we first segmented the player population into loners and socializers, identified features which represent achievement and socialization orientation of players and then applied these two models to the two player segments. We also studied and compared the behavioral profiles of the different player segments. We were able to gain several insights based on the study.

We found that loners were much more likely to churn than socializers. A large fraction of loners churn from the game and they tend to be players who are in the initial stages of the game and are not very engaged or progressing through the game levels. Among socializers, we observed that there was a near equal likelihood for a player to churn or not. In the socializer segment, churners show little to no activity leading up to the point of analysis whereas non-churners show a steady activity level. Non-churners were observed to be in higher levels within the game as compared to churners. We also observed that churners interact more with other churners in the weeks leading up to the point of churn and consequently, the fraction of activity in groups decrease for churners as compared to non-churners. Finally, we built models using Achievement and Socialization-oriented features and found the models to have good predictive performance in identifying churners. Furthermore, the socialization features were found to improve model performance by a significant amount.

The impact of motivation on learning and behavior is an important aspect of educational psychology [168]; where motivation is categorized as *extrinsic* or *intrinsic*. In behavioral psychology, reinforcement is a consequence that increases the frequency of a particular behavior and there can be different schedules of reinforcement. As future work, one can also examine if there are patterns of in-game behavior which causes a player to be motivated and continue playing.

## Chapter 5

# On Player Churn and Social Contagion

### 5.1 Introduction

Most of existing churn research have focused on modeling individual churn behavior without considering network effects - for a literature review, we refer the reader to Section 1.2.1. Recent work have started to investigate the influence of social ties in individual churn analysis [1, 13]. However, all the studies so far have focussed on the impact on individual churn based on neighbors who have already churned. In this chapter, we would like to study what is the impact on an actor's neighborhood if the actor churns from the network. To the best of our knowledge, this question has not been studied in the churn literature and we believe this would be our novel contribution to the area.

An important aspect of any churn prediction model is how early the model can identify that an actor is at risk of churn. Typical churn prediction models would predict whether an actor is likely to churn in the next time window; which could be a week or month. However, at this point it may already be too late for the CRM folks to take effective counter-measures to prevent the churn from happening. In our previous chapter, we have seen that socialization has an impact on individual churn and this has also been observed in previous studies [1, 13]. We can therefore make the assumption that when an actor churns, it has an impact on the actor's neighborhood. Specifically, we

would expect the churn likelihood of the actor's neighbor to change and the magnitude of the change would depend on the degree and nature of influence that the actor has on the neighbor. This can provide early warning regarding an actor's propensity to churn rather than just doing next-interval prediction. *Key contributions* of the chapter are listed below.

First, we address the question - When an active node, *ego*, becomes dormant, what is the impact on the activity behavior of *ego*'s immediate neighbor, *alter*, based on *ego*'s characteristics and *ego*'s relationship with *alter*? We consider individual game-based and node-based features for *ego* and also features based on the existing relationship between *alter* and *ego*. Results indicate that *ego*'s centrality/prestige in the network is a key determinant of *alter*'s activity behavior after *ego* becomes dormant. *Ego*'s character level, which is indicative of expertise level, is also a key factor in *alter*'s change in behavior. Among the features based on the existing relationship between *alter* and *ego*, we find that the number of common neighbors and the adar-adamic index are key determinants in the contagion process. Finally, results indicate that homophily-based features between *alter* and *ego* are not very discriminating in predicting dyadic influence.

Second, we address the question - When an active node, *ego*, becomes dormant, what is the impact on the activity behavior of *ego*'s immediate neighbor, *alter*, based on the activity behavior of *alter*'s remaining neighbors? Results indicate that *alter*'s behavior is strongly impacted by the number of remaining active neighbors and the strength of the relationship with those neighbors. Thus, we find that there is a strong social influence in effect wherein *alter*'s activity behavior is impacted by the activity behavior of the players around *alter*. This is in keeping with existing models of diffusion in the literature [77].

Third, we use a supervised learning framework to study the impact of player churn and social contagion. Experimental results show that the classification models perform substantially better than random for both the research problems.

To the best of our knowledge, the problem of player churn and social contagion has not been studied in the literature and in general, the problem of social contagion has not been studied in a supervised learning framework.

## 5.2 Background and Related Work

In this chapter, we focus on the impact of a node's churn behavior on its immediate neighborhood or group - the underlying sociological processes of this study relate to the fields of social influence, social contagion and group dynamics. For a brief overview and social science background of these fields, we refer the reader to Section 1.2.2. Specifically, the research questions in this chapter relate to behavioral contagion [44] with the behavior in question being a player's churn in an online game setting.

Recent work has started looking at social contagion and behavior cascades in human social networks [169, 170] - this encompasses a diverse spectrum of behavior and affective states such as happiness [171], loneliness [172], depression [173], smoking [174], alcohol consumption [175] and even obesity [176]. Centola investigated the effects of network structure on diffusion by studying the spread of health behavior through artificially structured online communities [169]. Results indicated that individual adoption was much more likely when participants received social reinforcement from multiple neighbors in the social network. Furthermore, the behavior spread farther and faster across clustered-lattice networks than across corresponding random networks.

Bond et al performed a randomized controlled trial of political mobilization messages delivered to 61 million Facebook users during the 2010 US congressional elections [177]. The results show that the messages directly influenced political self-expression, information seeking and realworld voting behaviour of people and in general, suggest that strong ties are instrumental for spreading both online and real-world behaviour in human social networks.

Aral and Walker used in vivo randomized experimentation to identify influence and susceptibility in networks using a sample of 1.3 million Facebook users [178]. Key findings from the study were a) Younger users are more susceptible to influence than older users, men are more influential than women, women influence men more than they influence other women, and married individuals are the least susceptible to influence in the decision to adopt the product offered and b) Influential individuals are less susceptible to influence than noninfluential individuals and that they cluster in the network while susceptible individuals do not. In a separate study, Aral et al [179] distinguished between contagion and homophily [109] effects in a dataset documenting product adoption in a

large network. Results show that previous methods overestimated peer influence in this network by 300-700% and that homophily explains more than 50% of the perceived behavioral contagion in mobile service adoption. These findings demonstrate that homophily can account for a great deal of what appears at first to be a contagious process.

Traditionally, models of diffusion by which information, ideas and influence spread through a network has been studied extensively in a number of domains such as diffusion and adoption of innovations [34, 36, 35], spread of infectious diseases in epidemiology [82, 33], the effects of "word of mouth" and "viral marketing" in the promotion of new products [83, 58, 84], trust propagation through networks [85, 86] and more recently, information diffusion in online social networks [87, 88, 89, 90]. For a brief literature review of the field, we refer the reader to Section 1.2.3.

We study behavioral contagion in small groups within online games in this chapter. To the best of our knowledge, no other work has studied this problem. Previous studies have only looked at in-game player organizations in MMORPGs called *guilds*. Williams et al did an extensive study of the social life of guilds in *World of Warcraft* by interviewing a representative sample of players in the game [180]. They created a basic typology of guilds based on goals, size and membership; much like Mintzberg's [181] typology of organizations. The basic guild types by goal were identified as social, PvP, raid and role-play - with overlap between the types. Guild sizes were found to be interwoven with the goals and were classified as being small, medium and large. In general, smaller guilds tend to be more focused on social bonds, whereas larger guilds focused more on game goals. They observed a high level of guild churn in the game [182] - approximately 21% of the guilds present on a WoW server at any given time were found to disappear after a month. Lack of alignment between the player's individual objectives and the guild's objectives was often cited as an important reason for leaving a guild. Other common reasons were found to be - social distance, poor leadership, a lack of players at their level to play with, and the wrong level of seriousness (both too high and too low).

Ducheneaut et al analyzed the structural properties of guilds in *World of Warcraft*, and examined some of the factors that could explain the success or failure of these groups [183]. Their findings reinforced earlier research showing that there might be a hard limit on the size of a viable organic group online [184], possibly set at around 35 group members or less. They looked at guild organization and its impact on guild

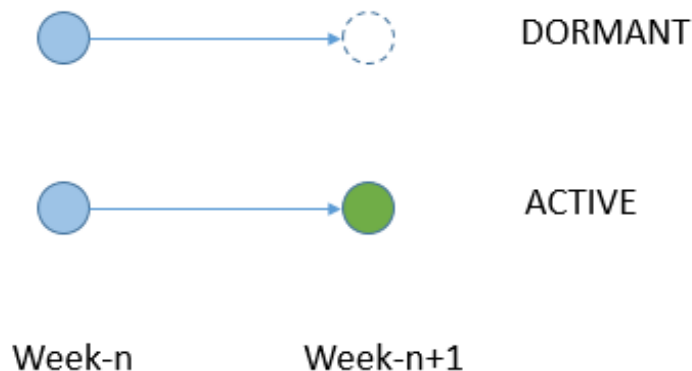


Figure 5.1: Dormancy - Classification Labels

survival and player achievement. They found that some of the factors which contribute to higher guild survival were - balanced classes of players, larger guild size, a wider spread of player levels and smaller subgroups within the guild.

Kang et al [185] studied the social dynamics of a guild in its early stages and found that social activities that lead to an increase in an individual's profit are the main factors in strengthening and expanding the lifetime of the guilds in these stages, more so than pure friendship and loyalty, especially in the very early stages. Patil et al [186] used a classification-based approach to predict guild stability and constructed a range of features that describe group composition, group structure and group activities. Chung et al [187]

### 5.3 Methodology

We use a supervised learning approach to study the impact of *ego's* churn on its immediate neighborhood or group.

#### 5.3.1 Classification Labels

A node is labeled based on whether or not it becomes dormant going from one week to the next. Figure 5.1 illustrates the labelling scheme.

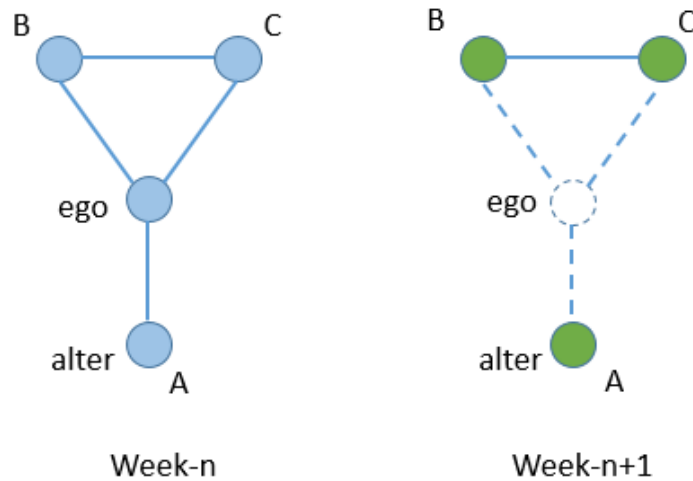


Figure 5.2: Dyadic Influence of *ego* on *alter* - illustrative example

- **DORMANT:** A node is labeled as DORMANT in week  $n$  if its session length drops to zero in week  $n + 1$ .
- **ACTIVE:** A node is labeled as ACTIVE in week  $n$  if its session length has a non-zero value in week  $n + 1$ .

### 5.3.2 Dyadic Influence of *ego* on *alter*'s activity behavior

The first research question we address is

**RQ1:** When an active node, *ego*, becomes DORMANT, what is the impact on the activity behavior of *ego*'s immediate neighbor, *alter*, based on *ego*'s characteristics and *ego*'s relationship with *alter*?

The purpose of this study is to gain insight into the factors that influence *alter*'s behavior when *ego* is no longer in the network, as illustrated in Figure 5.2 where *ego* becomes DORMANT in week  $n + 1$ . In this particular scenario, *ego* has three neighbors in week  $n$ . We consider individual game-based and node-based features for *ego* in week  $n$  and also features of the edges *ego*-A, *ego*-B and *ego*-C in week  $n$  in order to predict the label for A, B and C in week  $n+1$ .



### Game-based features for *ego*

These features are intended to capture *ego*'s engagement and experience in the game. A key indicator of engagement is the amount of playtime *ego* puts into the game and this is captured using session-related features. The activity logs available to us only record player actions and did not explicitly define sessions. So, we used a simple heuristic to define player-sessions - a session consists of sets of activities which are separated by no more than 30 minutes. We consider *ego*'s overall sessions and sessions spent in groups. While the *Overall Session* features give a measure of *ego*'s overall engagement, the *Overall Session* features are a measure of *ego*'s degree of socialization/group-oriented activities during the week. The features in this category are listed below -

- *Overall Number of Sessions* for *ego* during the week.
- *Overall Session Length (mins)* for *ego* during the week.
- *Overall Inter-Session Length (mins)*: Time between sessions for *ego* during the week.
- *Number of Sessions in Groups* spent by *ego* during the week.
- *Session Length (mins) in Groups* spent by *ego* during the week.
- *Inter-Session Length (mins) in Groups* spent by *ego* during the week.
- *Player Character Level* - max character level for *ego* during the week. This feature is intended to capture *ego*'s level of expertise in the game. A player with greater expertise in the game can be expected to have greater influence on its immediate neighbors.
- *Experience Points* collected by *ego* during game play in the week. This feature is intended to capture *ego*'s degree of success in game-play activities during the week. Successful players can be expected to have greater influence on its immediate neighbors in terms of motivating them towards game-play.

### Network-based features for *ego*

These features are intended to capture the importance/prestige of *ego* in the network. The features were computed using the Pajek [161] program for analysis and visualization of large networks.

- *Degree Centrality*: Degree centrality is based on the number of edges incident upon the node (i.e., the number of ties that a node has) [163]

$$C_D(ego) = Degree(ego) \quad (5.1)$$

- *Closeness Centrality*: The *farness* of a node  $s$  is defined as the sum of its distances to all other nodes, and its *closeness* is defined as the reciprocal of the farness. Thus, the more central a node is the lower its total distance to all other nodes.
- *Betweenness Centrality*: Betweenness centrality of a node is a measure based on the number of shortest paths that a vertex lies in [165]

$$C_B(ego) = \sum_{s \neq i \neq t \in V} \frac{\sigma_{st}(ego)}{\sigma_{st}} \quad (5.2)$$

where  $\sigma_{st}$  is the number of shortest paths from  $s$  to  $t$  and  $\sigma_{st}(i)$  is the number of shortest paths from  $s$  to  $t$  that pass through vertex *ego*.

- *Clustering Index*: Clustering Index is the fraction of pairs of a person's collaborators who have also collaborated with one another [166].

$$C(ego) = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} \quad (5.3)$$

### Edge-based features for *ego* and *alter*

These features are intended to capture various facets of the existing relationship between *ego* and *alter* such as a) shared sessions, b) common neighbors in the network and c) homophily. Shared sessions between *alter* and *ego* are a good indicator of the strength of interpersonal connection between the nodes. Features based on common neighbors and similarity metrics are intended to capture . Homophily-based features are intended to

capture similar tastes, preferences and demographics between *alter* and *ego*. The features in this category are listed below -

- *Shared Number of Sessions* between *alter* and *ego* during the week.
- *Shared Session Length (mins)* between *alter* and *ego* during the week.
- *Common Neighbors*: as given by

$$n(\textit{ego}, \textit{alter}) = |\mathbb{G}(\textit{ego}) \cap \mathbb{G}(\textit{alter})| \quad (5.4)$$

- *Jaccard Index* [188]: a statistic used for comparing the similarity and diversity of sets.

$$\gamma(\textit{ego}, \textit{alter}) = \frac{|\mathbb{G}(\textit{ego}) \cap \mathbb{G}(\textit{alter})|}{|\mathbb{G}(\textit{ego}) \cup \mathbb{G}(\textit{alter})|} \quad (5.5)$$

- *Adar-Adamic Index*: Neighbors with few connections have more weight in capturing the similarity of nodes *i* and *j* [189].

$$\alpha(\textit{ego}, \textit{alter}) = \sum_{k \in \mathbb{G}(\textit{ego}) \cap \mathbb{G}(\textit{alter})} \frac{1}{\log(|\mathbb{G}(k)|)} \quad (5.6)$$

- *Character Race Homophily Indicator*: identifies whether *alter* and *ego* belong to the same race within the game.
- *Character Class Homophily Indicator*: identifies whether *alter* and *ego* belong to the same class within the game. In a generalized setting, this could be interpreted as whether the two actors have similar skills and interests.
- *Character Guild Homophily Indicator*: identifies *alter* and *ego* to the same guild in the game. In a generalized setting, this could be interpreted as whether the two actors belong to the same organization or group and hence share a common goal.
- *Character Gender Homophily Indicator*: identifies *alter* and *ego* have the same character gender in the game.
- *Real Country Homophily Indicator*: identifies whether *alter* and *ego* belong to the same real-world country.

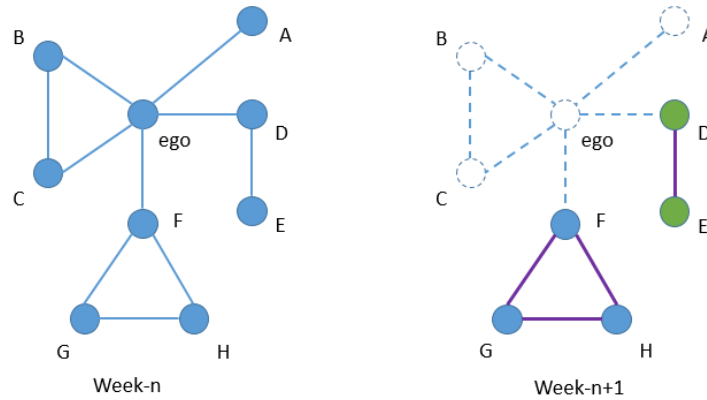


Figure 5.3: Neighborhood Influence on *alter* - illustrative example

### 5.3.3 Neighborhood Influence on *alter's* activity behavior

The second research question we address is

**RQ2:** When an active node, *ego*, becomes DORMANT, what is the impact on the activity behavior of *ego's* immediate neighbor, *alter*, based on the activity behavior of *alter's* remaining neighbors?

The purpose of this study is to gain insight into the factors that cause *alter* to leave or stay after *ego* has left the network, as illustrated in Figure 5.3 where *ego* becomes DORMANT in week  $n + 1$ . In this particular scenario, *ego* has five neighbors in week  $n$  and depending on the remaining neighbors, each of these *alter* may undergo a different change in their activity behavior. For example, B and C may become DORMANT as well since they were part of a triad with *ego*. D might continue to remain ACTIVE because of E's influence. F, G and H maybe relatively unaffected because they are part of a stable triad.

In order to study this question, we look at the neighbors for *alter* for the week  $n$ . The general idea is that because of social influence [77], a node's behavior will primarily be impacted by the behavior of the players around *alter*. The specific features are listed below -

- *Number of DORMANT neighbors:* for *alter* in the current week.
- *Number of ACTIVE neighbors:* for *alter* in the current week.

Table 5.1: Weekly Activity Networks - breakdown by dormancy labels

Ego Label	Week			
	14	15	16	17
DORMANT	1266 (21.69%)	1308 (22.16%)	1488 (24.23%)	1480 (24.69%)
ACTIVE	4572 (78.31%)	4594 (77.84%)	4652 (75.77%)	4515 (75.31%)
Total Nodes	5838	5902	6140	5995
Total Edges	35713	35592	38114	37928

Table 5.2: EQII Weekly Group Networks - average degree by dormancy labels

Ego Label	Week			
	14	15	16	17
DORMANT	4.92	5.21	5.08	5.03
ACTIVE	14.22	13.99	14.75	15.09

- *Edge-weighted DORMANT neighbors*: for *alter* in the current week.
- *Edge-weighted ACTIVE neighbors*: for *alter* in the current week.

## 5.4 Experimental Results and Analysis

### 5.4.1 Data Description

Data from the MMORPG - Sony Everquest II <sup>1</sup> was used for the experiments and analysis. The game data had four servers and in-game relationships among players include housing (player grants access to his/her house to another player), mentoring (experienced player mentoring another player to help the latter acquire skills and level up within the game), trading (buying and selling of in-game items), chat and grouping. As a player levels up within the game, the quests become progressively more difficult and it becomes essential for players to play in groups to accomplish tasks and progress through the game <sup>2</sup>. For the purposes of our analysis, we used data from the *guk* server and the grouping relationships to build networks across the weeks.

<sup>1</sup> <https://www.everquest2.com/>

<sup>2</sup> <http://eq2.wikia.com/wiki/Groups>

Table 5.3: Dyadic Influence Dataset

<b>Alter Label (Week 14)</b>	<b>Sample Size</b>	<b>Percentage</b>
DORMANT	1228	19.7%
ACTIVE	5005	80.3%
<b>Total</b>	<b>6233</b>	<b>100%</b>

Table 5.1 gives a breakdown of the nodes and edges in the group networks across 4 weeks for the two labels - DORMANT and ACTIVE. For example, there are a total of 5838 nodes and 35713 undirected edges in the week 14 group network. Of the 5838 nodes in week 14, 1266 become DORMANT (session length goes to zero) and 4572 remain ACTIVE (non-zero session length). We observe that, on average, around 23.19% of the nodes become DORMANT and 76.81% of the nodes remain ACTIVE going into the next week.

Table 5.2 shows the average degree of nodes for the two labels. We observe that, on average, DORMANT nodes have 5.06 neighbors and ACTIVE nodes have 14.52 neighbors. This indicates that nodes which become dormant have a disproportionately smaller number of neighbors. This is in keeping with our observation in the previous chapter.

#### 5.4.2 Experiment 1 - Dyadic Influence on *alter*'s activity behavior

In this experiment, we use a binary classification approach to study the factors that impact activity of *ego*'s immediate neighbor, *alter*, when *ego* becomes DORMANT - based on *ego*'s characteristics and *ego*'s relationship with *alter*. The week 14 group network is used for these purposes. There are 1266 nodes in the week 14 network which become DORMANT and there are 6233 edges connected to these nodes. Thus, we get 6233 samples for the experiment (refer Table 5.3). Of these samples, *alter* itself becomes DORMANT in 1288 (19.7%) of the samples and *alter* remains ACTIVE in 5005 (80.3%) of the samples.

Table 5.4 lists the 22 features (refer Section 5.3.2) in decreasing order of their info-gain value. Based on this table, we observe that the top five features in predicting change in *alter*'s behavior after *ego* has become DORMANT are -

Table 5.4: Dyadic Influence Model - Info-gain ranking of features

Info-gain Ranking	Feature	Info-gain Value
1	ego_curr_week_closeness_cent	0.0856
2	alter_ego_curr_week_common_neighbors	0.07826
3	ego_curr_week_pc_level	0.06883
4	alter_ego_curr_week_adar_adamic_index	0.0638
5	ego_curr_week_degree_cent	0.05962
6	alter_ego_curr_week_shared_num_sessions	0.04531
7	ego_curr_week_clustering_coef	0.03559
8	ego_curr_week_total_xp	0.02996
9	alter_ego_guild_indicator	0.02622
10	alter_ego_country_indicator	0.02337
11	alter_ego_race_indicator	0.02331
12	alter_ego_gender_indicator	0.02202
13	alter_ego_class_indicator	0.02127
14	ego_curr_week_sl_mins	0.0143
15	ego_curr_week_betweenness_cent	0.01413
16	ego_curr_week_group_sl_mins	0.01385
17	ego_curr_week_num_group_sessions	0.0137
18	ego_curr_week_num_sessions	0.01357
19	alter_ego_curr_week_shared_sl_mins	0.00948
20	ego_curr_week_isl_mins	0.00557
21	ego_curr_week_group_isl_mins	0.00257
22	alter_ego_curr_week_jaccard_index	0

- *Ego's closeness centrality*: Closeness centrality of a node is a centrality measure based on how close the node is to all other nodes. The importance of this feature indicates that ego's location within the network is indicative of its social influence on its neighbors and is highly discriminating in terms of determining *alter's* behavior when *ego* leaves.
- *Number of common neighbors between alter and ego*: Common neighbors between *ego* and *alter* are an indicator of the strength of the connection between .
- *Ego's character level*: This feature indicates that *ego's* level of expertise is important in determining *alter's* behavior.
- *Adar-adamic index for alter and ego*: Adar-adamic index between two nodes is greater when common neighbors of the two nodes have fewer connections. The

Table 5.5: Dyadic influence Model - 10-fold cross validation results

Model	DORMANT			ACTIVE		
	Precision	Recall	F-measure	Precision	Recall	F-measure
J48	54.95	36.64	43.97	85.63	92.63	88.99
Logistic	62.63	23.62	34.3	83.74	96.54	89.69
kNN	46.6	32.41	38.23	84.57	90.89	87.62
Naive Bayes	36.6	55.7	44.17	87.53	76.32	81.55
Perceptron	60.54	23.86	34.23	83.74	96.18	89.53
AdaBoost	53.02	40.8	46.11	86.25	91.13	88.62
Bagging	63	38.27	47.62	86.19	94.49	90.14

Table 5.6: Neighborhood Influence Dataset

Alter Label (Week 14)	Sample Size	Percentage
DORMANT	770	24.39%
ACTIVE_DECREASE	1499	47.48%
ACTIVE_INCREASE	888	28.13%
<b>Total</b>	<b>3157</b>	<b>100%</b>

importance of this feature indicates that when the common neighbors between alter and ego are primarily connected to them, the influence is stronger.

- *Ego's degree centrality*: This feature indicates that *ego's* connectivity, in general, is indicative of its influence on *alter*.

Furthermore, we observe that homophily-based features between *alter* and *ego* are not very discriminating in predicting dyadic influence.

Table 5.5 shows the precision, recall and f-measure results for the dyadic influence model on the DORMANT and ACTIVE labels. We observe that all the models do much better than random for both the labels (refer 5.3).

### 5.4.3 Experiment 2 - Neighborhood Influence on *alter's* activity behavior

In this experiment, we use a binary classification approach to study the factors that impact activity of *ego's* immediate neighbor, *alter*, when *ego* becomes DORMANT - based on the activity behavior of *alter's* remaining neighbors. The week 14 group



Table 5.7: Neighborhood Influence Model - Info-gain ranking of features

Info-gain Ranking	Feature	Info-gain Value
1	num_curr_week_neighbors_ACTIVE	0.1881
2	weighted_curr_week_neighbors_DORMANT	0.1437
3	weighted_curr_week_neighbors_ACTIVE	0.1265
4	num_curr_week_neighbors_DORMANT	0.0167

Table 5.8: Neighborhood Influence - mean, median and SD of features

Feature	DORMANT	ACTIVE
num_curr_week_neighbors_ACTIVE	<1.59, 1, 1.13>	<2.03, 1, 1.58>
weighted_curr_week_neighbors_DORMANT	<4.23, 2, 7.65>	<18.41, 13, 16.48>
weighted_curr_week_neighbors_ACTIVE	<0.27, 0.24, 0.22>	<0.1, 0.05, 0.12>
num_curr_week_neighbors_DORMANT	<0.26, 0.21, 0.27>	<0.47, 0.49, 0.21>

network is used for these purposes. There are 1266 nodes in the week 14 network which become DORMANT and there are 3157 unique *alter*'s connected to these nodes. Thus, we get 3157 samples for the experiment (refer Table 5.6). Of these samples, *alter* itself becomes DORMANT in 770 (24.39%) of the samples and *alter* remains ACTIVE in 2387 (28.13%) of the samples.

Table 5.7 lists the 4 neighbor based features (refer Section 5.3.3) in decreasing order of their info-gain value. Based on this table, we observe that *alter*'s behavior is strongly impacted by the number of remaining active neighbors and the strength of the relationship with those neighbors. This is indicative of a social influence effect going on.

Table 5.9 shows the precision, recall and f-measure results for the neighborhood influence model on both the labels. We observe that the models do much better than random for the both the labels (refer 5.6).

## 5.5 Conclusion and Future Work

We address two research questions related to player churn and social contagion in this chapter. We used a supervised learning framework for the study and found that the classification models perform substantially better than random for both the research problems.

Table 5.9: Neighborhood Influence Model - 10-fold cross validation results

Model	DORMANT			ACTIVE		
	Precision	Recall	F-measure	Precision	Recall	F-measure
J48	65.83	44.29	52.95	83.74	92.58	87.94
Logistic	65.32	42.08	51.18	83.24	92.79	87.76
kNN	54.22	48.44	51.17	83.93	86.8	85.34
Naive Bayes	54.63	60.52	57.42	86.81	83.79	85.27
Perceptron	63.93	43.51	51.78	83.48	92.08	87.57
AdaBoost	60.74	47.01	53	84.07	90.2	87.03
Bagging	66.54	45.19	53.83	83.98	92.67	88.11

First, we address the question - When an active node, *ego*, becomes dormant, what is the impact on the activity behavior of *ego*'s immediate neighbor, *alter*, based on *ego*'s characteristics and *ego*'s relationship with *alter*? We consider individual game-based and node-based features for *ego* and also features based on the existing relationship between *alter* and *ego*. Results indicate that *ego*'s centrality/prestige in the network is a key determinant of *alter*'s activity behavior after *ego* becomes dormant. *Ego*'s character level, which is indicative of expertise level, is also a key factor in *alter*'s change in behavior. Among the features based on the existing relationship between *alter* and *ego*, we find that the number of common neighbors and the adar-adamic index are key determinants in the contagion process. Finally, results indicate that homophily-based features between *alter* and *ego* are not very discriminating in predicting dyadic influence.

Second, we address the question - When an active node, *ego*, becomes dormant, what is the impact on the activity behavior of *ego*'s immediate neighbor, *alter*, based on the activity behavior of *alter*'s remaining neighbors? Results indicate that *alter*'s behavior is strongly impacted by the number of remaining active neighbors and the strength of the relationship with those neighbors. Thus, we find that there is a strong social influence in effect wherein *alter*'s activity behavior is impacted by the activity behavior of the players around *alter*. This is in keeping with existing models of diffusion in the literature [77].

The two research questions addressed in this chapter deal with gaining insights into the churn contagion process and discovering the network substructures which facilitate or inhibit the contagion process. We believe understanding these factors can be quite

helpful in early identification of at-risk players and give the CRM folks a window of opportunity to address and mitigate the risk. One possible action that CRM folks could take would be to recommend friends and groups to players who are at risk of churn.

## Chapter 6

# Conclusion

Churn behavior has been studied across a wide range of industries such as telecom, banking and online social networks. However, most of existing churn research have focused on modeling individual churn behavior and the type of questions has also been limited by the types of datasets which are available to researchers. In this thesis, different aspects of churn in a Massively Multiplayer Online Role Playing Games (MMORPGs) are studied in depth.

From a behavioral or even cognitive point of view, churn is an indicator of the player getting disengaged (losing interest or bored) in the game experience. Such disengagement can be manifested in either the player *stopping to pay* or *stopping to play*. We analyzed both of these problems as *Cancellation* or *Dormancy* respectively and found that *Dormancy* is a better indicator of player disengagement in terms of its immediacy rather than *Cancellation*.

We investigated the problem of churn prediction in MMORPGs using a player lifecycle-based approach for modeling user behavior. We analyzed the activity traits of churners in the weeks leading up to their point of churn and compared it with the activity traits of a non-churner. The analysis revealed distinct behavioral profiles associated with churners and non-churners which can discriminate between the two populations. We proposed three trend metrics along which weekly player history of an observed variable is recomputed to give derived features and found that these metrics have good discriminating power. We also proposed a distance-based classification scheme, which we call *wClusterDist*, which benefited from the distinct behavioral profiles of the two

populations. Experimental results show that the proposed classification scheme is well-suited for this problem formulation and its performance is better than or comparable to other traditional classifiers. We then extended the idea of a lifecycle -based approach to build models for predicting the likely churn of a dynamically evolving community.

We studied the impact of different *achievement* and *socialization*-based player motivational factors on player churn. Specifically, we looked at how socialization serves to increase player engagement and decrease churn. We found that loners are much more likely to churn than socializers. A large fraction of loners churn from the game and they tend to be players who are in the initial stages of the game and are not very engaged or progressing through the game levels. Among socializers, we observed that churners interact more with other churners in the weeks leading up to the point of churn and consequently, the fraction of activity in groups decrease for churners as compared to non-churners.

Finally, we looked at the relationship between player churn and social contagion i.e when a player leaves a network, what is the impact on its immediate neighborhood or group? Specifically, we addressed two research questions. First, we addressed the question - When an active node, *ego*, becomes dormant, what is the impact on the activity of *ego's* immediate neighbor, *alter*, based on *ego's* characteristics and *ego's* relationship with *alter*? Results indicate that *ego's* centrality/prestige in the network is a key determinant of *alter's* activity behavior after *ego* becomes dormant. *Ego's* character level, which is indicative of expertise level, is also a key factor in *alter's* change in behavior. Among the features based on the existing relationship between *alter* and *ego*, we find that the number of common neighbors and the adar-adamic index are key determinants in the contagion process. Finally, results indicate that homophily-based features between *alter* and *ego* are not very discriminating in predicting dyadic influence.

Second, we addressed the question - When an active node, *ego*, becomes dormant, what is the impact on the activity of *ego's* immediate neighbor, *alter*, based on the activity behavior of *alter's* remaining neighbors? Results indicate that *alter's* behavior is strongly impacted by the number of remaining active neighbors and the strength of the relationship with those neighbors. Thus, we find that there is a strong social influence in effect wherein a player's activity is impacted by the activity of the its neighbors.

# References

- [1] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A. Nanavati, and Anupam Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '08, pages 668–677, New York, NY, USA, 2008. ACM.
- [2] M.C. Mozer, R. Wolniewicz, D.B. Grimes, E. Johnson, and H. Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *Neural Networks, IEEE Transactions on*, 11(3):690–696, may 2000.
- [3] Yossi Richter, Elad Yom-Tov, and Noam Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, pages 732–741, 2010.
- [4] Jorge B. Ferreira, Marley Vellasco, Marco Aurlio Pacheco, and Carlos Hall Barbosa. Data mining techniques on the evaluation of wireless churn. In *In ESANN*, pages 483–488, 2004.
- [5] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006.
- [6] Bingquan Huang, B. Buckley, and T. M. Kechadi. Multi-objective feature selection by using nsga-ii for customer churn prediction in telecommunications. *Expert Syst. Appl.*, 37(5):3638–3646, May 2010.

- [7] Yaya Xie, Xiu Li, E.W.T. Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3, Part 1):5445 – 5449, 2009.
- [8] Teemu Mutanen. Customer churn analysis a case study. *Research Report*, 2006.
- [9] B. Q. Huang, M-T. Kechadi, and B. Buckley. Customer churn prediction for broadband internet services. In *Proceedings of the 11th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK '09, pages 229–243, Berlin, Heidelberg, 2009. Springer-Verlag.
- [10] Octavio Herrera and Taieb Znati. Modeling churn in p2p networks. In *Proceedings of the 40th Annual Simulation Symposium*, ANSS '07, pages 33–40, Washington, DC, USA, 2007. IEEE Computer Society.
- [11] Katharina Morik and Hanna Köpcke. Analysing customer churn in insurance data: a case study. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD '04, pages 325–336, New York, NY, USA, 2004. Springer-Verlag New York, Inc.
- [12] Guangli Nie, Guoxun Wang, Peng Zhang, Yingjie Tian, and Yong Shi. Finding the hidden pattern of credit card holder's churn: A case of china. In *Proceedings of the 9th International Conference on Computational Science*, ICCS 2009, pages 561–569, Berlin, Heidelberg, 2009. Springer-Verlag.
- [13] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, pages 423–428, Washington, DC, USA, 2009. IEEE Computer Society.
- [14] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, Partha Basuchowdhuri, Conor Hayes, and Thorsten Strufe. Churn in social networks. In Borko Furht, editor, *Handbook of Social Network Technologies and Applications*, pages 185–220. Springer US, 2010.

- [15] Brij Masand, Piew Datta, D.R. Mani, and Bin Li. Champ: A prototype for automated cellular churn prediction. *Data Mining and Knowledge Discovery*, 3:219–225, 1999. 10.1023/A:1009873905876.
- [16] Chih-Fong Tsai and Yu-Hsin Lu. Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10):12547 – 12553, 2009.
- [17] Yu Zhao, Bing Li, Xiu Li, Wenhuan Liu, and Shouju Ren. Customer churn prediction using improved one-class support vector machine. In *Proceedings of the First International Conference on Advanced Data Mining and Applications*, ADMA’05, pages 300–306, Berlin, Heidelberg, 2005. Springer-Verlag.
- [18] Junxiang Lu and D Ph. Predicting customer churn in the telecommunications industry an application of survival analysis modeling using sas. *Techniques*, 114-27:11427, 2002.
- [19] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW ’12 Companion, pages 829–834, New York, NY, USA, 2012. ACM.
- [20] R.J. Oentaryo, Ee-Peng Lim, D. Lo, F. Zhu, and P.K. Prasetyo. Collective churn prediction in social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 210–214, Aug 2012.
- [21] Dorwin Cartwright. Influence, leadership, control. *Handbook of Organizations*, pages 1–47, 1965.
- [22] Peter V Marsden and Noah E Friedkin. Network studies of social influence. *Sociological Methods & Research*, 22(1):127–151, 1993.
- [23] Bonnie H Erickson. The relational basis of attitudes. *Social structures: A network approach*, 99:121, 1988.



- [24] Serge Moscovici and Claude Faucheux. Social influence, conformity bias, and the study of active minorities. volume 6 of *Advances in Experimental Social Psychology*, pages 149 – 202. Academic Press, 1972.
- [25] RP John Jr, Bertram H Raven, et al. The bases of social power. 1959.
- [26] Bertram H Raven. Social influence and power. Technical report, DTIC Document, 1964.
- [27] Andrew M Pettigrew. Information control as a power resource. *Sociology*, 6(2):187–204, 1972.
- [28] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.
- [29] David A Levy and Paul R Nail. Contagion: A theoretical and empirical review and reconceptualization. *Genetic, Social, and General Psychology Monographs*, 1993.
- [30] Peter Sheridan Dodds and Duncan J Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604, 2005.
- [31] James D Murray. Mathematical biology i: An introduction, vol. 17 of interdisciplinary applied mathematics, 2002.
- [32] Daryl J Daley, Joe Gani, and Joseph Mark Gani. *Epidemic modelling: an introduction*, volume 15. Cambridge University Press, 2001.
- [33] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford science publications. OUP Oxford, 1992.
- [34] J.S. Coleman, E. Katz, H. Menzel, and Columbia University. Bureau of Applied Social Research. *Medical innovation: a diffusion study*. Advanced Study in Sociology. Bobbs-Merrill Co., 1966.
- [35]
- [36] E.M. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 2003.

- [37] Susanne Lohmann. The dynamics of informational cascades: the monday demonstrations in leipzig, east germany, 1989–91. *World politics*, 47(01):42–101, 1994.
- [38] Rodney Stark. Why religious movements succeed or fail: a revised general model. *Journal of contemporary religion*, 11(2):133–146, 1996.
- [39] Robert L Montgomery. *The diffusion of religions: A sociological perspective*. University Press of America, 1996.
- [40] Gerald Schoenewolf. Emotional contagion: Behavioral induction in individuals and groups. *Modern Psychoanalysis*, 1990.
- [41] Peter Totterdell. Catching moods and hitting runs: mood linkage and subjective performance in professional sport teams. *Journal of Applied Psychology*, 85(6):848, 2000.
- [42] Sigal G Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47(4):644–675, 2002.
- [43] Norman Polansky, Ronald Lippitt, and Fritz Redl. An investigation of behavioral contagion in groups. *Human Relations*, 1950.
- [44] Ladd Wheeler. Toward a theory of behavioral contagion. *Psychological review*, 73(2):179–192, 1966. ID: 1966-05303-001.
- [45] Daniel Grosser, Norman Polansky, and Ronald Lippitt. A laboratory study of behavioral contagion. *Human Relations*, 4:115–142, 1951. ID: 1952-01418-001.
- [46] Albert Bandura, Dorothea Ross, and Sheila A Ross. Imitation of film-mediated aggressive models. *The Journal of Abnormal and Social Psychology*, 66(1):3, 1963.
- [47] Albert Bandura, Dorothea Ross, and Sheila A Ross. Vicarious reinforcement and imitative learning. *The Journal of Abnormal and Social Psychology*, 67(6):601, 1963.
- [48] Don L Kimbrell and Robert R Blake. Motivational factors in the violation of a prohibition. *The Journal of Abnormal and Social Psychology*, 56(1):132, 1958.

- [49] RH Walters, Marion Leat, and L Mezei. Response inhibition and disinhibition through empathetic learning. *Canad. J. Psychol.*, 17:235–243, 1963.
- [50] Richard H Walters and Ross D Parke. Influence of response consequences to a social model on resistance to deviation. *Journal of Experimental Child Psychology*, 1(3):269–280, 1964.
- [51] Paul Marsden. Memetics and social contagion: Two sides of the same coin? *Journal of Memetics - Evolutionary Models of Information Transmission*, 2(2):68, 1998.
- [52] Alan C Kerckhoff and Kurt W Back. *The June bug: A study of hysterical contagion*. Appleton-Century-Crofts, 1968.
- [53] Michael J Colligan and Lawrence R Murphy. A review of mass psychogenic illness in work settings. *Mass psychogenic illness: A social psychological analysis*, pages 33–52, 1982.
- [54] David P Phillips. The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American Sociological Review*, pages 340–354, 1974.
- [55] Madelyn Gould, Patrick Jamieson, and Daniel Romer. Media contagion and suicide among the young. *American Behavioral Scientist*, 46(9):1269–1284, 2003.
- [56] Ladd Wheeler and Anthony R Caggiula. The contagion of aggression. *Journal of Experimental Social Psychology*, 2(1):1–10, 1966.
- [57] Susan T Ennett, Robert L Flewelling, Richard C Lindrooth, and Edward C Norton. School and neighborhood characteristics associated with school rates of alcohol, cigarette, and marijuana use. *Journal of Health and Social Behavior*, pages 55–71, 1997.
- [58] Vijay Mahajan, Eitan Muller, and Frank M. Bass. New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, 54(1):pp. 1–26, 1990.

- [59] Thomas Lux. Herd behaviour, bubbles and crashes. *The economic journal*, pages 881–896, 1995.
- [60] Ronald S. Burt. Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92(6):pp. 1287–1335, 1987.
- [61] Dorwin Philip Cartwright and Alvin Frederick Zander. Group dynamics: Research and theory. chapter The Nature of Group Cohesiveness. Tavistock Publications, 1968.
- [62] Nancy J. Evans and Paul A. Jarvis. Group cohesion: A review and reevaluation. *Small Group Research*, 11(4):359–370, 1980, <http://sgr.sagepub.com/content/11/4/359.full.pdf+html>.
- [63] Kurt W. Back. Influence through social communication. *The Journal of Abnormal and Social Psychology*, 46(1):9–23, 1951. ID: 1951-07362-001.
- [64] Leon Festinger, Harold B. Gerard, Bernard Hymovitch, Harold H. Kelley, and Bert Raven. The influence process in the presence of extreme deviates. *Human Relations*, 5(4):327–346, 1952, <http://hum.sagepub.com/content/5/4/327.full.pdf+html>.
- [65] Harold B. Gerard. The anchorage of opinions in face-to-face groups. *Human Relations*, 7(3):313–325, 1954, <http://hum.sagepub.com/content/7/3/313.full.pdf+html>.
- [66] Edwin J. Thomas. Effects of facilitative role interdependence on group functioning. *Human Relations*, 10(4):347–366, 1957, <http://hum.sagepub.com/content/10/4/347.full.pdf+html>.
- [67] Irvin D Yalom. *The theory and practice of group psychotherapy*. Basic Books, 1995.
- [68] Peter E. Mudrack. Defining group cohesiveness: A legacy of confusion? *Small Group Research*, 20(1):37–49, 1989, <http://sgr.sagepub.com/content/20/1/37.full.pdf+html>.

- [69] Noah E. Friedkin. Social cohesion. *Annual Review of Sociology*, 30(1):409–425, 2004, <http://dx.doi.org/10.1146/annurev.soc.30.012703.110625>.
- [70] Leon Festinger. Informal social communication. *Psychological review*, 57(5):271–282, 1950. ID: 1951-04528-001.
- [71] Miller McPherson and Lynn Smith-Lovin. *Cohesion and membership duration: Linking groups, relations and individuals in an ecology of affiliation*, chapter 2, pages 1–36. <http://www.emeraldinsight.com/doi/pdf/10.1016/S0882-6145>
- [72] Michael A Hogg. The social psychology of group cohesiveness: From attraction to social identity author: Michael a. hogg, publisher: Prent. 1992.
- [73] Albert J. Lott and Bernice E. Lott. Group cohesiveness as interpersonal attraction: A review of relationships with antecedent and consequent variables. *Psychological bulletin*, 64(4):259–309, 1965. ID: 1965-15044-001.
- [74] Kenneth A. Bollen and Rick H. Hoyle. Perceived cohesion: A conceptual and empirical examination. *Social Forces*, 69(2):479–504, 1990, <http://sf.oxfordjournals.org/content/69/2/479.full.pdf+html>.
- [75] Neal Gross and William E Martin. On group cohesiveness. *American Journal of Sociology*, pages 546–564, 1952.
- [76] Noah E Friedkin and Eugene C Johnsen. Social influence networks and opinion change. *Advances in Group Processes*, 16(1):1–29, 1999.
- [77] Mark Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
- [78] James Moody and Douglas R White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, pages 103–127, 2003.
- [79] Michael A. Hogg. Group cohesiveness: A critical review and some new directions. *European Review of Social Psychology*, 4(1):85–111, 1993, <http://dx.doi.org/10.1080/14792779343000031>.

- [80] Henri Tajfel and John C Turner. An integrative theory of intergroup conflict. *The social psychology of intergroup relations*, 33(47):74, 1979.
- [81] John C. Turner, Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. *Rediscovering the social group: A self-categorization theory*. Basil Blackwell, Cambridge, MA, US, 1987. ID: 1987-98657-000.
- [82] N.T.J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Mathematics in Medicine Series. Griffin, 1975.
- [83] Jacqueline J. Brown and Peter H. Reingen. Social Ties and Word-of-Mouth Referral Behavior. *The Journal of Consumer Research*, 14(3):350–362, 1987.
- [84] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 57–66, New York, NY, USA, 2001. ACM.
- [85] C.-N. Ziegler and G. Lausen. Spreading activation models for trust propagation. In *e-Technology, e-Commerce and e-Service, 2004. EEE '04. 2004 IEEE International Conference on*, pages 83 – 97, march 2004.
- [86] Elizabeth Gray, Jean-Marc Seigneur, Yong Chen, and Christian Jensen. Trust propagation in small worlds. In Paddy Nixon and Sotirios Terzis, editors, *Trust Management*, volume 2692 of *Lecture Notes in Computer Science*, pages 1072–1072. Springer Berlin / Heidelberg, 2003.
- [87] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 91–101, New York, NY, USA, 2002. ACM.
- [88] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

- [89] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, pages 23–32, New York, NY, USA, 2013. ACM.
- [90] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, July 2013.
- [91] Steve Jurvetson. What exactly is Viral Marketing?
- [92] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02*, pages 61–70, New York, NY, USA, 2002. ACM.
- [93] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.
- [94] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03*, pages 137–146, New York, NY, USA, 2003. ACM.
- [95] David Kempe, Jon Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *Proceedings of the 32Nd International Conference on Automata, Languages and Programming, ICALP'05*, pages 1127–1138, Berlin, Heidelberg, 2005. Springer-Verlag.
- [96] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.
- [97] Philip N. Howard, Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid. Opening closed regimes: What was the role of social media during the arab spring? 2011.

- [98] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.
- [99] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1019–1028, New York, NY, USA, 2010. ACM.
- [100] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schlkopf. Uncovering the temporal dynamics of diffusion networks. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 561–568. Omnipress, 2011.
- [101] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.
- [102] Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 435–443, New York, NY, USA, 2008. ACM.
- [103] R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA, 2004. ACM.
- [104] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 113–114, New York, NY, USA, 2011. ACM.
- [105] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 45–54, New York, NY, USA, 2011. ACM.



- [106] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twiterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.
- [107] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 241–250, New York, NY, USA, 2010. ACM.
- [108] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 207–218, New York, NY, USA, 2008. ACM.
- [109] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001, <http://www.annualreviews.org/doi/pdf/10.1146/annurev.soc.27.1.415>.
- [110] Albert-Lszl Barabasi and Rka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999, <http://www.sciencemag.org/content/286/5439/509.full.pdf>.
- [111] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Den-sification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM.
- [112] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
- [113] M.E.J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330, 2004.
- [114] B. W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell system technical journal*, 49(1):291–307, 1970.

- [115] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.
- [116] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.
- [117] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.
- [118] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
- [119] Tanya Y Berger-Wolf and Jared Saia. A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528. ACM, 2006.
- [120] Chayant Tantipathananandh and Tanya Y Berger-Wolf. Finding communities in dynamic social networks. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 1236–1241. IEEE, 2011.
- [121] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, pages 44–54, New York, NY, USA, 2006. ACM.
- [122] Randy J. Pagulayan, Kevin Keeker, Dennis Wixon, Ramon L. Romero, and Thomas Fuller. The human-computer interaction handbook. chapter User-centered Design in Games, pages 883–906. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003.
- [123] Heather W. Desurvire. Usability inspection methods. chapter Faster, Cheaper!! Are Usability Inspection Methods As Effective As Empirical Testing?, pages 173–202. John Wiley & Sons, Inc., New York, NY, USA, 1994.

- [124] Melissa A. Federoff and Melissa A. Federoff. Heuristics and usability guidelines for the creation and evaluation of fun in video games. Technical report, Indiana University, Bloomington, 2002.
- [125] Heather Desurvire. Using heuristics to evaluate the playability of games. In *In: the Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI 2004)*, pages 1509–1512. ACM Press, 2004.
- [126] Darryl Charles, Michael Mcneill, Moira Mcalister, Michaela Black, Adrian Moore, Karl Stringer, Julian Kücklich, and Aphra Kerr. Player-Centred Game Design: Player Modelling and Adaptive Digital Games. In *Digital Games Research Association 2005 Conference: Changing Views - Worlds in Play*, 2005.
- [127] Isabella Selega Csikszentmihalyi. *Optimal experience: Psychological studies of flow in consciousness*. Cambridge University Press, 1992.
- [128] Ben Cowley, Darryl Charles, Michaela Black, and Ray Hickey. Toward an understanding of flow in video games. *Comput. Entertain.*, 6(2):20:1–20:27, July 2008.
- [129] Dongseong Choi and Jinwoo Kim. Why people continue to play online games: in search of critical design factors to increase customer loyalty to online contents. *Cyberpsychology Behavior - the impact of the Internet, multimedia and virtual reality on behavior and society*, 7(1), 2004.
- [130] Dmitri Williams, Nick Yee, and Scott E. Caplan. Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, 13(4):993–1018, July 2008.
- [131] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, January 2000.
- [132] Y Zhao, B Li, X Li, W Liu, and S Ren. *Customer churn prediction using improved one-class support vector machine*, volume 3584, pages 300–306. 2005.

- [133] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Syst. Appl.*, 34(1):313–327, January 2008.
- [134] Z. Borbora, J. Srivastava, Kuo-Wei Hsu, and D. Williams. Churn prediction in mmorpqs using player motivation theories and an ensemble approach. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 157–164, oct. 2011.
- [135] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, and Conor Hayes. Churn in social networks: A discussion boards case study. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10*, pages 233–240, Washington, DC, USA, 2010. IEEE Computer Society.
- [136] Marcel Karnstedt, Matthew Rowe, Jeff Chan, Harith Alani, and Conor Hayes. The effect of user features on churn in social networks. *Human Factors*, pages 1–8, 2011.
- [137] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Trans. Internet Technol.*, 3(1):1–27, February 2003.
- [138] Bamshad Mobasher. The adaptive web. chapter Data mining for web personalization, pages 90–135. Springer-Verlag, Berlin, Heidelberg, 2007.
- [139] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recogn.*, 38(11):1857–1874, November 2005.
- [140] Christopher M. Bishop. Mixture models and em. In *Pattern Recognition and Machine Learning (Information Science and Statistics)*, pages 423–455. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [141] John F. Roddick and Myra Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. *SIGKDD Explor. Newsl.*, 1(1):34–38, June 1999.

- [142] Michail Vlachos, Marios Hadjieleftheriou, Dimitrios Gunopulos, and Eamonn Keogh. Indexing multi-dimensional time-series with support for multiple distance measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 216–225, New York, NY, USA, 2003. ACM.
- [143] Ethem alpaydin. Multivariate classification. In *Introduction to Machine Learning*, pages 94–99. The MIT Press; second edition edition (December 4, 2009), 2009.
- [144] Tom Mitchell. Decision tree learning. In *Machine Learning*, pages 52–78. The Mc-Graw-Hill Companies, Inc., 1997.
- [145] Mark A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. pages 359–366. Morgan Kaufmann, 2000.
- [146] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.
- [147] Charles Ling, , Charles X. Ling, and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 73–79. AAAI Press, 1998.
- [148] M.A. Ahmad, B. Keegan, J. Srivastava, D. Williams, and N. Contractor. Mining for gold farmers: Automatic detection of deviant players in mmogs. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, pages 340 –345, aug. 2009.
- [149] J. Oh, Z. Borbora, and J. Srivastava. Automatic detection of compromised accounts in mmorpgs. In *Proceedings of the ASE International Conference on Social Informatics*, Washington D.C, USA, December 2012.
- [150] Nick Yee. Motivations for play in online games. In *CyberPsychology & Behavior*, pages 772–775, 2006.

- [151] Amit A Nanavati, Siva Gurumurthy, Gautam Das, Dipanjan Chakraborty, Koustuv Dasgupta, Sougata Mukherjea, and Anupam Joshi. On the structural properties of massive telecom call graphs: findings and implications. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 435–444. ACM, 2006.
- [152] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A Nanavati, and Anupam Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 668–677. ACM, 2008.
- [153] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 423–428. IEEE, 2009.
- [154] Richard J Oentaryo, Ee-Peng Lim, David Lo, Feida Zhu, and Philips K Praseetyo. Collective churn prediction in social network. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 210–214. IEEE Computer Society, 2012.
- [155] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, and Conor Hayes. Churn in social networks: A discussion boards case study. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 233–240. IEEE, 2010.
- [156] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 829–834. ACM, 2012.
- [157] Yossi Richter, Elad Yom-Tov, and Noam Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, pages 732–741, 2010.
- [158] Mark EJ Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.
- [159] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

- [160] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [161] Vladimir Batagelj and Andrej Mrvar. Pajek-program for large network analysis. *Connections*, 21(2):47–57, 1998.
- [162] Richard Bartle. Hearts, clubs, diamonds, spades: Players who suit muds. <http://mud.co.uk/richard/hcdfs.htm>.
- [163] J. Nieminen. On the centrality in a graph. *Scandinavian Journal of Psychology*, 15:332–336, 1974.
- [164] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239, 19781979.
- [165] L. Freeman. A set of measures of centrality based upon betweenness. *Sociometry*, pages 35–41, 1977.
- [166] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.
- [167] Z. Borbora and J. Srivastava. User behavior modelling approach for churn prediction in online games. In *Proceedings of the ASE/IEEE International Conference on Social Computing (SocialCom 2012)*, Amsterdam, The Netherlands, September 2012.
- [168] Jeanne Ellis Ormrod. *Educational psychology: Developing learners*. Merrill Upper Saddle River, NJ, 2000.
- [169] Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.
- [170] James H Fowler and Nicholas A Christakis. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12):5334–5338, 2010.

- [171] James H Fowler, Nicholas A Christakis, et al. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338, 2008.
- [172] John T Cacioppo, James H Fowler, and Nicholas A Christakis. Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6):977, 2009.
- [173] J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281, 2011.
- [174] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008.
- [175] J Niels Rosenquist, Joanne Murabito, James H Fowler, and Nicholas A Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, 2010.
- [176] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.
- [177] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.
- [178] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.
- [179] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.
- [180] Dmitri Williams, Nicolas Ducheneaut, Li Xiong, Yuanyuan Zhang, Nick Yee, and Eric Nickell. From tree house to barracks: The social life of guilds in world of warcraft. *Games and Culture*, 1:338–361, 2006.



- [181] Henry Mintzberg. The structuring of organizations: A synthesis of the research. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1979.
- [182] Nicolas Ducheneaut, Nicholas Yee, Eric Nickell, and Robert J. Moore. "alone together?": Exploring the social dynamics of massively multiplayer online games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 407–416, New York, NY, USA, 2006. ACM.
- [183] Nicolas Ducheneaut, Nicholas Yee, Eric Nickell, and Robert J. Moore. The life and death of online gaming communities: A look at guilds in world of warcraft. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 839–848, New York, NY, USA, 2007. ACM.
- [184] Robin IM Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(04):681–694, 1993.
- [185] Ah Reum Kang, Juyong Park, and Huy Kang Kim. Loyalty or profit? early evolutionary dynamics of online game groups. In *Proceedings of Annual Workshop on Network and Systems Support for Games*, NetGames '13, pages 2:1–2:6, Piscataway, NJ, USA, 2013. IEEE Press.
- [186] Akshay Patil, Juan Liu, and Jie Gao. Predicting group stability in online social networks. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1021–1030, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [187] Taejoong Chung, Jinyoung Han, Daejin Choi, Taekyoung Ted Kwon, Huy Kang Kim, and Yanghee Choi. Unveiling group characteristics in online social games: A socio-economic analysis. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 889–900, New York, NY, USA, 2014. ACM.
- [188] Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272, 1901.

- [189] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, July 2003.