博士論文

Empirical Studies on Freemium Business Model:
Case of Online Music Streaming Provider
（オンライン音楽プロバイダーの事例を用いた
フリーミアムビジネスモデルに関する実証研究）

ポンヌムクン スチット
Suchit Pongnumkul

UNIVERSITY OF TOKYO

Empirical Studies on Freemium Business Model : Case of Online Music Streaming Provider

A dissertation submitted for the degree
Doctor of Philosophy
in
Engineering
By

Suchit Pongnumkul

2017

University of Tokyo

Abstract

Empirical Studies on Freemium Business Model : Case of Online Music Streaming Provider

Suchit Pongnumkul

This dissertation studies three aspects of freemium-based music streaming services that have potential to increase profits for the providers, which are income, cost and user's satisfaction. Empirical data from user's social network and historical usage data are used in this study. Freemium business model, in which some parts of the services are offered to users for free and advanced services are charged some premium, has gained popularity, especially for digital products and services. The motivation of this dissertation comes from the fact that freemium business providers still struggle to make profits. Therefore, in order for such business to continue to provide service sustain-ably, this dissertation investigates approaches to alleviate the problem. The struggle to make profits is especially true for subscription-based music streaming services, because of three main problems; first, the conversion rates for freemium is normally low; second, the licensing costs for music are high; and lastly with the ever increasing number of music, finding the music that fits a user's need is a challenging task. Therefore, this dissertation is organized into three parts: income loss prevention, cost reduction and user's satisfaction improvement.

The first part of this dissertation aims to increase income by analyzing premium subscriptions and un-subscriptions. Social network of users, usage information and subscription history of both user and friends are used in the analysis. The model used empirical data from last.fm, an online music streaming service. The second part of this dissertation aims to improve user's satisfaction of the service by presenting the right information to the user with an improved recommendation system. Bayesian probability was adopted to use with the recommendation system based on random walk with restart and yield better accuracy. The last part investigates cost reduction through a theoretical model that explains music consumption behavior. Long-tail distribution of content consumption of last.fm data is analyzed and a new evolution mechanism with bipartite network of users and songs is proposed to improve the match between the model and real-world phenomenon in music consumption. Parameters in the proposed model are then used to understand the changes in the shapes of the longtail and the changes in music licensing costs for the providers. In summary, freemium-based music streaming services could benefit from the study in this dissertation to increase profit and make the business more sustainable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Freemium is a popular business model for online services that has become increasingly prevalent for not only business providers but also for customers and investors. Freemium services naturally have low barrier of entry for users because they allow customers to use the services for free and only users who require advanced features need to pay. Therefore, service providers can quickly gain new users, which is attractive for investors because investors normally look for services with large user bases. While freemium has been used in practice for a long time, it has only gained interests from a wide audience and are widely studied after the best seller book, "Free: The future of a radical price" by Anderson [12] was published in 2009.

While freemium services are easy to deploy and gain users, freemium providers often struggle to make profit. Increasing the total number of users is not difficult, because users normally do not have to pay for freemium services. However, these users do not incur any income to the business. Music streaming services also struggle with this problem. Spotify, one of the most popular music streaming services, reported a net loss in 2015 [66].

Nowadays, the communication over the Internet is widely adopted and essentially incurs no additional costs for users due to the design of Internet subscription packages. These make communication over the Internet, especially on social network become popular and social network is becoming important information source. The social network information is important; therefore, a lot of researches use social network as source information and show the relationships between social network position of the users and similarity between friends and trend to use premium service. However, while unsubscriptions of premium users decrease income of freemium service, the relationships between these parameters and unsubscriptions of premium service has not been studied.

Profit making is problem for freemium content providers, so decreasing unsubscriptions is one of the important things that freemium providers should explore. If soon-to-unsubscribe members can be predicted, the provider can make a focused marketing campaign for the users to continue being premium members by using less marketing cost than make general campaign to all users.

The user behaviors show some longtail phenomenon such as distribution of listening frequency

per songs or distribution of listening frequency per user, and there are many researches that study the factors that generate the longtail. The longtail is getting more interests as the online distribution has less marginal cost to add more items. These create unlimited varies of items and services that have both merits and demerits. The merits are there are more capacity and variety of products and services to serve varies needs of customers. But the demerit is additional cost and time to find matched products/services due to increasing of number of items.

For the problem of information overwhelming due to the longtail effect, there are 2 type of solutions, search engine and recommendation system. The search engine will provide the most related information that relate to search keyword and user's information. But for the recommendation system that have no search keyword, user's information is the most important input to the system to find the information that user may want. User's information contains user's profile, user's behaviors and user's relation to the other users.

Cold start is the main problem of the recommendation system that will occur when the information is not enough to find the relation. For example, the new user that have no user's behavior information, the system will cannot make the recommendation to such that user. But there are some studies show that friends have similar behaviors, the social network information can be used to find new user's friends behaviors and use as input into the recommendation system. Random walk with restart is one of recommendation system's algorithm that can use both users' behaviors and social network information to make the efficient recommendation. But the process can be improved by change to random parameters.

For the providers, such as e-commerce provider that have income related to number of transactions or service used, the marginal cost will not exceed the marginal income. But for subscription based freemium provider, increasing of usages not guarantee to increasing of income, but certainly increase the cost. Longtail distribution of usage will change the cost as the shape changed. As the copyright is one of the main cost for content provider and increased follow to the number of usage and popularity of creator, changing the shape of longtail usage can reduce the copyright cost that have to pay to the right holders. But there is no study about the controllable parameters that effect the shape of longtail.

The goal of this dissertation is to explore approaches to improve profitability of freemium music streaming services as shown in Figure 1.1. Three main factors were explored, which are conversion rate, recommendation system and royalty fee, with the aim to increase profit for service providers by increase the income and reduce the cost and to allow the services to continue to be offered.

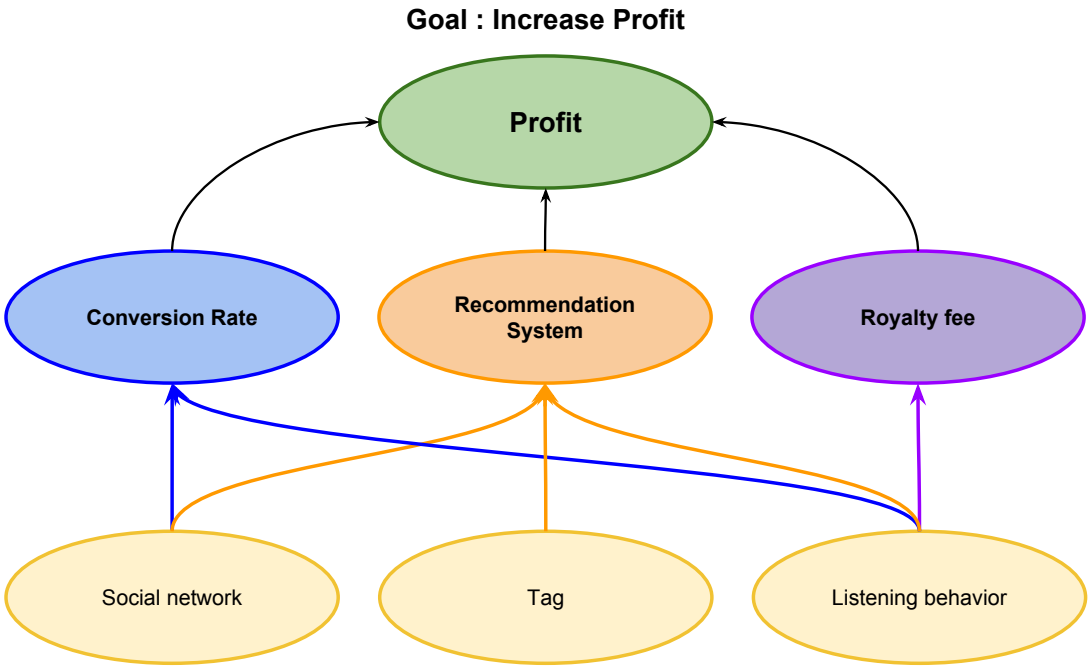

Figure 1.1: Goal and scope of this research

To increase the income, I analyzed the information about when premium users subscribe to the service and unsubscribe from the service by using the user's listening behavior and subscription history, friends' subscription history, and how important the user is in the social network. The strategic to increase conversion rate was proposed from the result of the proposed model. This is described in chapter 3.

Together, as the number of songs exponentially increases, finding the right songs for each user is challenging. I developed a recommendation system based on historical data of listening behavior and tags and social network of users to allow users to find the songs that suit their taste better. This can improve the user experience for the service and increase number of users.This is described in chapter 4.

To decrease the cost, I proposed a evolution model to explain the longtail phenomena in the usage distribution of music streaming service. An analysis of how each parameter affect the change in the shape of the longtail was performed. Then the model is used to propose licensing cost reduction by a simplified cost model in which popular songs costs more for the streaming service providers. This is described in chapter 5.

Combining the three factors discussed above, this dissertation provides information and managerial recommendation for freemium music streaming service providers to utilize in their business with the aim to increase profits.

Chapter 2

OBJECTIVE OF THESIS AND LITERATURE SURVEY

This chapter discusses the objective of this thesis and reviews main related knowledges and literatures, which are freemium business model and social network analysis.

2.1 The Objective of Thesis

Freemium business model becomes popular, because of clear marketing and monetization model. However, making profit from the service is the most difficult problem of the providers. To make sustainable services, the strategy to increase profits are needed which will make the provider can use the increased profit to improve the services. We can increase profits for service providers by increasing **income** or reducing **cost** of the service. The income of all freemium services is related to two factors which are **the total number of the users** and the **conversion rate** of the service. One of the strategy to increase both of them is to increase the **user satisfaction** which will lead to word of mouth of the service and the continuous using of premium service. For the service which have too many choices to the users, the recommendation system is need to help solving information overwhelming problem and increase the customer satisfaction. Together, the effect from social network have to be concern for online services. The social network can be the media for word of mouth and spread the technology or service usage. The strategy to increase premium service usage related to their users' social network can help provider to increase income too. On another hand, one of the biggest cost of the music streaming service is **royalty fee**, which is related to the frequency of the listening of the music. The frequency of listening is said to be **longtail distribution**. If the shape of longtail can be changed by some parameters, the royalty fee cost is also changed too. The tree factors social network, tag and listening behaviors, which shown as yellow items in figure 1.1, can affect the conversion rate, recommendation results and royalty fee, which lead to change in the profit.

The main objective of this thesis is to explore strategies to increase profits for freemium service providers, which can be done by increasing conversion rate, increasing performance of recommendation system that affects user satisfaction, and reducing the royalty costs. For the conversion rate, dynamic use-diffusion model, which can be used to explain the continuous usage of the product and

service, does not contain the social network and historical premium service usage factors. One of the sub-objectives of this thesis is to modify the dynamic use-diffusion model to be able to explain the factors that effect the premium service usage and use the new model to explain the behavior of the users and find the strategy to increase conversion rate. This sub-objective is illustrated using blue borders and arrows in Figure 1.1. Next, as recommendation systems can increase the user satisfaction, improving the recommendation system can further increase the satisfaction. Therefore, one of the sub-objectives is to improve the performance of recommendation system, which is illustrated using orange borders and arrows in Figure 1.1. Another sub-objective is to study the longtail phenomenon of music listening behavior to find factors which can change the shape of longtail distribution that affect the royalty cost of the freemium music streaming service. This sub-objective is illustrated using purple borders and arrows in Figure 1.1.

2.2 Freemium Business Model

Before making a purchase, customers often want to assess the products, both in terms of the product's quality and whether the product fits their needs. One of the known effective method that companies assist customers in the process is to allow customers to test the products for themselves. Free samples of products and test driving of cars are marketing approaches that companies frequently use to acquaint consumers with their product before they decide to make purchases. For physical products, such as cosmetics, food, cars, this type of direct marketing is costly, but it is known to be effective [1]. For digital products, such as software, mobile application, online services, this type of direct marketing has created a business model, Freemium, where users can test the products for themselves before making the decision to pay for the product.

The term "Freemium" is coined by Jared Lukin in 2006 [143] and has been published in the book "Free: The future of a radical price" of Chris Anderson [12] and the book "Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue" by Eric Seufert [119]. Freemium refers to the strategy to set the prices for services that are partly free but charges for services or features that are premium. This model includes, shareware, open-source software, many internet services and in-app purchases.

There has been many examples of successful business that use freemium business model. Popular examples are Skype, Dropbox and Microsoft Office. Skype is a cross platform software that allows users to chat and make calls over VoIP for free. Their premium service allows calls to home phones and mobile phones and also allows others to call from home phone and mobile phones by

giving each premium user a phone number to receive calls. Dropbox is a service that backs up user's data on the cloud and synchronizes these data across user's multiple machines automatically and easily. Users are allowed up to 2 GB of data for free. Users who wish to use more data storage can pay premium prices for increased amount of storage. Microsoft Office provides Microsoft office online for free, which allows viewing and limited editing capability of documents. Users can purchase offline software for a fee for advanced viewing and editing features.

One of the reasons that freemium pricing strategy has gained attention nowadays is because products and services that have been provided for free has incurred only small cost for providing services to more users due to the IT development that follows Moore's law [116] and Neilsen's law of Internet Bandwidth [95] which say that the processing capability, storage and speed of communication increase exponentially without much increase in cost. Therefore, service providers can bare the cost of providing services and products to large amount of users. On the users side, the free services and products provided is also an incentive to allow users to try out their products, where users only have to learn new technology, but no monetary cost. It also helps increase the number of total users of their services/products.

2.2.1 Monetizing Approaches for Freemium

Monetization is the main concern in most business. Freemium business makes money by innovative business model. LinkedIn allows 80 million users to post their personal profile to the internet for free. These free users incur cost to the service, but does not contribute directly to pay for these costs. However, they contribute indirectly by being a part of a large number of users, in which advertisers, recruiters and lead-seeking professionals who want to reach them are willing to pay. LinkedIn obliges this demand with ad packages, premium subscriptions and hiring solutions that together have resulted in positive cash flow. The three most common approaches to monetize a freemium business are advertisement, subscription, and in-app purchase. Additionally, link the LinkedIn example, combination of the three approaches are commonly employed.

2.2.2 Conversion Rates

The main factor that are often considered to affect the success in freemium business model is the "conversion rate". Conversion rate is the ratio of the number of premium users (users who pay for premium services) to the total number of users. The conversion rate shows the ability to convert

free users to premium users and the income of the product/service can be calculated by

$$\text{Income} = (\#\text{ of total users}) \times (\text{conversion rate}) \times (\text{average income per premium user})$$

The business sustains when this income exceeds the cost of providing services to both free and premium users and the fixed cost of running the business.

From the formula to calculate income above, there are three ways to increase income in freemium business model. The first approach is to increase the number of all users. The second approach is to increase the conversion rate. The third approach is to increase the average income per premium user. The third approach is difficult to execute without new features to offer to premium users. Increasing the total number of users is an attractive approach if the cost per extra user is small. However, the second approach is the most interesting because it increases the total income from the same customers.

There is a body of research that studies the factors that affect conversion rates. Factors that have been found to affect conversion rates include the amount of interactions with other users and the amount of service usage [92, 46], the difference between privileges of free and premium users [118, 80], ease of use of the product/service [40], type of freemium service (time-lock, feature-lock, hybrid, uniform seeding) [31, 91], prices , duration and quality of service [42].

2.2.3 Types of Freemium

There are four main strategies for freemium business model. This list is adapted from the ones described in [105, 2, 3].

Quantity Limited

Limiting quantity is the traditional approach to allowing customers to test their products before buying. Product samples are a small quantity of the full product, which are given out to their target customers for free. For consumable goods (i.e. items which get used up), quantity can be limited by volume, i.e. **volume limitation**. For durable goods (e.g. vacuum cleaners, cars), free sampling can come in a form of product demonstration or test driving, which allows customers to experience the product for a limited time.

In terms of digital products, quantity limitation is often done by allowing users to test the full product for a limited time, or **time limitation**. After the specified time has passed, the users will not be able to use the product.

A different version for quantity limitation in digital product is **capacity limitation** or **storage limitation**. For example, Gmail, a free email service, allows anyone to use their service for free with storage up to 17 GB (in 2016). Users who require more than the storage limit can pay a monthly fee to increase the storage limit. Dropbox also allows users to use their service for free up to 2 GB, and charges a monthly subscription fee to users who wish to increase this storage capacity.

As recent platforms offer users pay-per-user services, in order for them to provide a trial, they offer free credits, or seed money, to new users. This type of freemium, **free credit limitation**, allows users to explore new service without up-front payment.

Quantity limitation is easy to execute for digital products. Therefore, we see many variance of this type of freemium.

Feature Limited

The idea of feature limitation is to gives users real utility initially for free, and provides an option to pay if they want better or more advanced services by the same service provider. Feature limitation separates functionalities into basic functionality which all users can use, and advanced functionality where only paid users can use. Limiting the ability to accomplish certain tasks is a **capability limitation**. Skype, a VOIP service, allows anyone to make VOIP call between computers for free, but charges for calls to phones, sms, and multi-party conference calls. Google Voice is a US-based VOIP service that allows free calls and sms to US phone numbers, but charges for international calls and sms.

Mobile applications often offer freemium in the form of **in-app purchases**, in which the applications can be downloaded and used for free, but advanced features are offered for a price. This also include **effort limitation** such as Temple Run, a game in which all or most features are available for free, but require extended unlocking or slowly obtained in-game currency which can be accelerated or purchased for a fee.

In some segment of service, advanced service can be differentiated by quality, or **quality limitation**. Specific examples can be said about music streaming services. Free users can listen to music at lower quality (i.e. bit rates), while paid users will get the same song at higher quality. Other services that speed of services depends on the network speed also employ **bandwidth limitation** to differentiate their service quality.

The challenges of feature limited freemium model is to create a good balance between the basic and advanced functionalities. The basic functions need to be able to convince users of the quality

of the product, but does not provide too much that the users feel that they do not need to pay for extra functions provided for the premium users.

Distribution Limited

Distribution limitation distinguishes the free and paid version by the way that the product or service is distributed. This scenario is mostly found today in packaged software that includes a user license. The license, open source or commercial, dictates the use cases where the product may be used at no charge and those that require a commercial right. Depending how the end user obtains the product and what she does with it, it is free or it costs money. For example, TeamViewer is a software that allows their users to access and control a remote computer via GUI, which employs distribution limitation by supporting personal use for free, but has different premium packages for business uses. **Seat limitation** freemium services allow software to be only usable on one computer rather than across a network.

A specific approach to distribution limitation is **customer class limitation**, in which some segments of users are offered the product for free. The most common segment is educational users. Most Autodesk software with full features, Matlab, Tableau Desktop are free for students. A few reasons to offer products for students for free are to allow users with limited purchase power to use their software, and once students who are used to a software graduate, they are likely to continue to use the product, which will results in sales.

Support Limited

Support for software can come in many forms. Users of a ”lite” version do not receive telephone and/or email support. A different type of support is customization of software. Open source software are one of the support limitation, as everyone can access the source code. Many companies are founded to provide support to implement open source software. Some of the well-known software include Pentaho.

While this list separates the dimensions that freemium services differentiate their free and premium services, many business often combine the dimensions to create hybrid freemium service. For example, Tableau, a visualization software and service, provides free Tableau Public by limiting distribution of the final visualization to public access (no privacy or security to the data), but provides Tableau Desktop, the same software, which users can save files offline for a licensing fee. At the same time, students can get Tableau Desktop for free.

These can tell that monetize of freemium service is related on willing to pay of the users which has conversion rate as the main measurement and the provider can increase willing to pay of the users by make the limitation of quantity, feature, distribution or support.

2.3 Social Network Analysis and Network Matrices

Online social network (OSN) has become very popular and attained a large user base. Therefore, many business became interested in using OSN in advertising their services and products. Some business has created their own OSN for their customers for their specific products to provide extra features for their customers. However, due to the large number of users and active users in OSN, there has been many studies that aimed to find effective targeted advertising strategy. The main goal is to discover the users who has high probability of being premium users. For example, there were studies that found peer pressure to link to the probability of being premium users. That is, users that are friends with premium users have high probability of being premium users. Moreover social interactions, such as when premium users favorite or comment on photos in flickr, also increase probability of being premium users [139, 54, 61]. Therefore, if we can identify the users who have potential in influencing other users to change member status, we can create an effective strategy for target marketing with such user group. The network position in addition to some private information can be used to predict the ability to influence their neighbors [69].

An analysis of OSN which has gained traction nowadays is Social Network Analysis (SNA) [44], which utilizes theories from Network (which is part of Graph theory) to analyze relationships within OSN. Users or objects in OSN are represented with nodes and relationships are represented with edges between corresponding nodes, resulting in a big graph representing the entire OSN. Metrics of the resulted graphs can be calculated, where each node and each edge can has their own metrics. For example, the most effective person to spread information in the network can be found [15]. Structure holes are a metric that looks for a node that if removed from the network, the number of clusters in the network will increase [145]. Data visualization can also be used to further utilize this data.

Chapter 3

SOCIAL DYNAMIC USE-DIFFUSSION OF PREMIUM SERVICE IN SUBSCRIPTION BASED ONLINE MUSIC STREAMING FREEMIUM SERVICE

This chapter proposes a social dynamic use-diffusion model, which is improved from dynamic use diffusion model[125], by investigates technology adoption and usage behavior in online-social-network based freemium services, which is described in section and 2.2 and 2.3. As the use-diffusion model focused on technology devices, such as computer[123], tablet[77] and smart products[99], the determinants do not fit to analyze the online-social-network based freemium services. This paper modified determinants and index in dynamic use diffusion model[125], by creating hypotheses about relationships between social network metrics, such as degree and clustering coefficient, and used patterns of premium service. Empirical data from last.fm, a global company that provide the online music streaming service since 2003, was used to assess the regression models. Data was collected at three time points, Aug 2013, Jun 2014 and Dec 2014, which allow us to construct history of subscription status of users and their friends. The network of users were also constructed to calculate network metrics of users. Activities of users were measured using the number of songs that the user listened to, which is main usage of the music streaming service. The subscription status in each time point of the users are used as both usage behavior and the proxy of the outcome. One linear regression model and two logit regression models were proposed to analyze the subscription status of at the last time point of users. The result of the regressions support the hypotheses, which means the modification of the model is supported.

3.1 Introduction

Freemium has been a popular business model for digital products and services in the past decade, especially for the services that are provided through the Internet. Freemium services offer their core features to customers for free and charge money for premium features. For example, some providers provide only browser based online contents for free, but also provide the mobile device based online contents and offline contents for premium features. One key success factor of freemium services is the *conversion rate*, which is the ratio of the number of premium users to the

total number of users. Conversion rate shows the ability to convert free users to premium users which generate revenue through the usage of premium product/service. This conversion rate directly affects the income of the business due to the income of freemium business can be calculated from the product of the total number of users times the conversion rate times the average income per premium user. In this research, users can be divided by subscription status into free users and premium users. Both type of users are service adopters and the usage was determined by number of music that user listening to. A large body of research have studied the factors that affect user's conversion from free user to premium user, which include difference between privileges of free and premium users [118, 80], user's usage rates [92, 46], peer influence [139], etc. The typical conversion rate in freemium business is quite small, which is 5-10%. While researchers have studied factors that would encourage free users to convert to premium users, little is known about factors that make premium users unsubscribe in freemium services. This chapter also explores such factors.

While premium unsubscribes in subscription-based freemium services have not been widely researched, studies related to users dropping out of a service are not new. In traditional business models, churn rates are refereed to the rates at which users stops subscribing from a service. Churn management and prediction have been extensively explored for mobile network services [109, 70] and internet service providers [84]. Decisions to stop subscription in traditional business is different from freemium business, because the absent of subscription in traditional business means no access to the service, but users in freemium services could access some features of the service. For example, unsubscribed last.fm users are able to listen to free music but cannot listening to premium radios which can be listened while subscribed. To the best of our knowledge, there is no studies of churn on premium service of freemium business. On the other hand, there are multiple approaches to monetizing freemium services, including subscription, and in-app purchases. Detection and prediction of churners in freemium which employed in-app purchase to monetize has been explored for online games [113, 53, 47]. However, these studies also predict and detect cases when users stop using the services instead of stop paying for the services, which is different from when premium users in subscription-based freemium services stops paying for subscription. Therefore, premium unsubscribes in freemium services need to be further explored.

The dynamic use-diffusion model [125], which can explain about satisfaction of technology usage and used to analyzed the difference of usage of personal computer in 3 countries, was modified by adding the social network metrics and subscription history. The hypotheses about relation be-

tween social network metrics and pattern of premium service usage are constructed and empirical data from last.fm, which is an online music streaming service, is used to verify the hypotheses and test the models. The results show that the proposed parameters can be used as the determinants and index in the dynamic use diffusion model [125], which can improve the model.

3.2 Related Works

3.2.1 Technology adoption and use diffusion

There are many studies about decision of product and technology adoption and diffusion. Rogers's Diffusion of Innovation (DoI), which is one of the most referred study, described innovation, adopters, communication channels, time and social network as 5 elements of the theorem. The innovation spread to the adopters through the communication channels of the social network by time. DoI separated the decision into 5 steps as knowledge, persuasion, decision, implementation and confirmation. Each adopter have own decision process, which start from get the information in knowledge process and compare to other choices in persuasion process. If the adopter decide to adopt the innovation in decision process, adoption will be done in the implementation process and confirmed the results in confirmation process. These processes will take time deference in each adopter which made the theory divide the adopters into 5 groups as innovators, early adopters, early majority, late majority and laggards.

Social cognitive theory [16, 17, 18] explain the information received process from social network by divide the cognitive processes into 4 steps as attention, retention, reproduction, and motivation. The information is received in attention step, and repeated receiving in retention step. After try in reproduction step and understand the results, the learner will be motivated in motivation step. These steps is close to decision processes in DoI, without decision step.

To understand an individual decision making, Ajzen and Fishbein proposed that attitude and subjective norm affected behavioral intention in theory of reasoned action(TRA) [9] and perceived behavioral control that control intention of behavior was added by Ajzen in theory of planed behavior(TPB) [7, 8]. The attitudes is a person's opinion about behavior and subjective norm is perceived social pressure while perceived behavioral control is a perceived difficulty of performing the behavior. These affect the intention to perform behavior and the intention can be used as the predictor of behavior performing.

Technology Acceptance Model(TAM) [37, 136] apply the TRA to explain the innovation adoption behavior that affected by attitude and perceived of usefulness. Attitude are affected by per-

ceived of usefulness and perceived ease-of-use, where perceived of usefulness is affected by perceived ease-of-use and external factors. And perceived ease-of-use is affected by external factors too.

To explain external factors, there are Social identity theory [130] which tells that the social behavior will vary along a continuum between interpersonal behavior and intergroup behavior and purely interpersonal or purely intergroup behavior is unlikely to be found. Together, individuals are intrinsically motivated to achieve positive distinctiveness. Self-categorization theory [134], which is related to social identity theory, explained the grouping categorization and leveling the group by the difference between the individual and the others in the group as the more difference, the longer distance between the individual and the group. These 2 theories can tell the power from the social network separated by group of source information and the intent behavior of the individual.

In the confirmation process, there is the expectancy theory [137] which explain that the behavior is motivated by expectation of rewards from performing the behavior. With 3 elements as expectancy, which is believe of effort will result performance, instrumentality, which is believe of performance will result reward on outcome, and valence, which is valuation of the reward. Expectation confirmation theory, or expectation disconfirmation theory [93, 94], explain post-purchase or post-adoption satisfaction as a difference of expectations and perceived performance which make a disconfirmation of beliefs and concluded into satisfaction. These two theory explain the confirmation process as the comparison of expected value created from the information that received from social network through communication channel before adoption and perceived value after adoption.

Shih and Venkatesh proposed use-diffusion model that shows relation between determinants and patterns and outcomes of innovative products and services usage [124]. Users was divided into 4 groups as intense users, specialized users, nonspecialized users, and limited users by rate of use and variety of use as in Table 3.1. Use patterns are effected by determinants and the outcome is results of UD patterns. Where UD determinants can be divided into 4 dimension as: household social context, technological dimension, personal dimension, and external dimension. This model was confirmed and used to examine varies usage of products and services such as mobile 3G service [68], broadband Internet [120], IPTV [126, 117, 115, 89], hypermarket [62], and digital library [50].

Use-diffusion model [124] use information only from one time point, which cannot include time effect into model, so Shih et al. improved the model to take such effect by replaced UD patterns with

Table 3.1: Users group in use-diffusion model

	low rate of use	high rate of use
high variety of use	Nonspecialized Users	Intense Users
low variety of use	Limited users	Specialized Users

dynamic use diffusion index (UDI) and added satisfaction as final outcome and used the modified model to analyze and compare the usage of personal computer in United States, Sweden, and India. The results show that in India which has shortest used period have positive relation between usage and social status.

3.2.2 Increasing Freemium Conversion Rates

There is a body of research that studies the factors that affect conversion rates. Factors that have been found to affect conversion rates include the amount of interactions with other users and the amount of service usage [92, 46], the difference between privileges of free and premium users [118, 80], ease of use of the product/service [40], type of freemium service (time-lock, feature-lock, hybrid, uniform seeding) [31, 91], prices , duration and quality of service [42].

Enders et al. [46] proposed strategies to earn income for social networking sites in terms of a long tail model (a model where a lot of users has low service usage and a small number of users has high service usage, resulting in a power law graph of the amount of service usage to the number of users). Social networking sites were studied and their income models were categorized into three categories : advertising, subscription and transaction. The revenue drivers for all three models were the number of users, willingness to pay, and trust in service. For subscription model, willingness to pay is the most important driver. The strategy proposed for increasing income for subscription model is to fatten the tail by encouraging user-generated contents, increasing site activities and creating different subscriptions based on the demands and amount of usage of the users. In freemium business model that uses subscription for their premium services, the proposed strategy to increase income is to fattening the tail, which is a way to increase conversion rate. Therefore, Enders et al. [46] proposed a way to increase conversion rate, but the study did not include or use the structure of social network or network positioning in a meaningful way.

Lopes and Galletta [80] studied perception and willingness to pay for online contents using questionnaires that address factors affecting decision to pay for online contents. Answers from 392 students were used to construct a model that explains the relationship between perception of reputation, perception of technical quality, expected benefit and willingness to pay. The results show that perception of reputation is a predictor of both perception of technical quality and expected benefit. Perception of technical quality is a predictor of expected benefit, and expected benefit is a predictor of willingness to pay. While Lopes and Galletta [80] did not directly study conversion rate, they propose a model that are predictors of willingness to pay. Therefore, they studied the decision to pay for online contents which can be applied to services in freemium model which are mostly services that offer online contents. However, social network was not included in [80], so social network for this context can be further explored.

Doerr et al. [42] studied factors influencing willingness to pay for music as a service (MaaS) by means of questionnaires with response from 132 users of MaaS. From the questionnaire responses factors influencing willingness to pay were evaluated. The results show that factors that have negative influence on willingness to pay are prices and duration of subscription. Factors that have positive effects on willingness to pay are sound quality, offline feature for listening to music, in-browser player, feature for editing music, mobile application availability and community features. The work of Doerr et al. [42] has confirmed the positive influence of community or social network on the decision to pay for premium music streaming service in freemium model. However, no model for predicting conversion rate was proposed. Also the factor about the positions of users in the network was not included in the study of factors affecting the willingness to pay.

3.2.3 Relation with Social Network Analysis and Network Metrics

Katona et al. [69] created prediction models of diffusion process of technology adoption from social network data and compared the prediction accuracy between a model that using social network metrics to a model that uses just demographic information. The social network data was obtained from a popular social network site. Two periods of user and friendship data were captured in this study. In the first period, daily data were obtained for 3.5 years (1247 days), resulting in data from 138,964 users. Then, 3 years later, the second period of data were captured as a snapshot of friendship data of the same users, which contains 111,036 additional users. The second period of data were used as a ground truth for how users adopt this technology. Katona et al. [69] created a prediction model using network position to predict how users adopt the technology. The results

show that using social network properties, which include degrees, clustering coefficients and local betweenness, together with demographic information of users (gender, age, and density of population of the users' cities) increases the accuracy of prediction 50% to 100% from the model that uses just demographic information. Even though Katona et al. [69] did not conduct their study on freemium model, but converting from free users to premium users is a form of technology adoption. Therefore, Katona et al. [69]'s work was applied to predict how users would change services level by using social network metrics. Specifically, in freemium services that social network data of users are available, the pay users in this study was compared to adopted users in Katona et al. [69] and free users was compared to non-adopted users.

Wang and Chin [139] studied the relationships peer pressure to the probability of being a pay user. Peer pressure in this study includes social connections (the percentage of friends that are pay users) and social interactions (the percentage of people that interact with this user that are pay users). Data used in this study are from Last.FM and Flickr that are freemium services that are also a social network. The result shows that probability of being pay users increases as the number of friends that are pay users increases. The probability of being pay users also increases as the number of interactions from pay users increases. However, there seem to be a limit as the additional benefit drops. In Wang and Chin [139]'s study, the probability of being pay users are calculated as a ratio between the number of pay users to the number of total users, which is the conversion rate mentioned above. Therefore, Wang and Chin [139] has concluded that if the number of friends who are pay users increases, the conversion rate increases. However, Wang and Chin [139] only studied this from one point of time, there is no temporal confirmation that there is a conversion from free users to pay users. Therefore, additional study is need to further understand the relationship between number of friends who are pay users to the conversion of user types.

Wagner et al. [138] conducted a survey with Music-as-a-Service users and found that the increase in the similarity between the free and the premium function leads to the increase in user's conversion. Therefore, companies should consider providing time limitation freemium instead of feature limitation freemium. Sylvester and Rand [129] found in-game social network characteristics, namely the local clustering coefficient among only those of their friends who have subscribed, to link to users' conversion. Peer influences have also been found to affect user's conversion in other studies including [19].

Oestreicher-Singer and Zalmanson [92]'s study found that consumer's willingness to pay increases as the level of their community participation rises. The study was conducted using Last.FM

data to find factors that influence the willingness to pay. The factors that were studied include content consumption, content organization and community participation, and usual factors that were included in previous research, such as user's demographics, amount of service usage, number of friends who are premium users and number of all friends. Result from logistic regression shows that the factors that have strongest positive influence on the consumer's willingness to pay are the number of friends that are pay users and the number of songs that the user listens to. Other factors that have positive influence on user's willingness to pay are the number of playlist created, the number of songs that the user likes, the number of groups that the user belongs to, the number of groups that the user is a leader, the number of blogs posted, and age. The factors that has negative influence on the willingness to pay are the number of all friends, duration of membership. Therefore, Oestreicher-Singer and Zalmanson [92] proposed a model that predicts the willingness to pay for premium service, which in other words is conversion rate, from various factors including demographic information, content consumption rate, user's relationship to other users in the social network. However, other social network metrics have not been explored, which could be factors that help increase the precision of conversion rates. The goal of this research is to discover the relationship between the position of the network and the probability that a user is a premium user in order to provide effective targeted advertisement. The position of the network can be measured by the selected metrics above, which are centrality and clustering coefficient.

Metrics that are selected for this paper are degree centrality and local clustering coefficient. Centrality is a measure of the connectedness to other nodes, or how close to the center of the network the node is. There are many ways to compute centrality. Degree centrality is the connectedness measured by the number of edges, so the node with high number of edges has high degree centrality. Local clustering coefficient is an indicator of the degree to which nodes in a graph tend to cluster together to other neighboring nodes. In the social network, this indicator show the connection strength between one user's friends.

3.2.4 Churns Detection and Prediction

Churn rates, or churns, are the rates at which users stops subscribing from a service. Traditional business, such as mobile network operators, found that it is more expensive to earn new customers than to convince existing users not to leave the services. Therefore, churn management and prediction have been extensively explored both in research and in practice for mobile network services [109, 70] and internet service providers [84]. Kim and Yoon [70] surveyed 973 mobile users in

Korea and, using a binomial logit model, found that the probability that a subscriber will switch carrier depends on the level of satisfaction of service attributes including call quality, tariff level, handsets, brand image, as well as income, and subscription duration.

Churn in freemium services have been studied for games [113, 53, 47], which employ in-app purchase to monetize instead of subscription. However, these studies also predict and detect cases when users stop using the services instead of stop paying for the services, which is different from when premium users in subscription-based freemium services stops paying for subscription. Runge et al. [113] predicts when high-value players (top 10% of paying players) will completely quit online social games using in game activity tracking data, revenue related data, and user's profile data. Four classifiers were compared and single hidden layer neural network with fine-tuned learning rate and momentum out performs other algorithms.

3.3 Hypotheses

In UD models by Shih and Venkatesh [124], Shih et al. [125], family was used as one of UD determinants because the technology in the past were physical devices. The closer in physical distance, the more frequency that usage can be seen. So the family who stay in the same house was the most effective people to influence the usage and give the knowledge of the technology. However, online technology and service cannot be seen in physically, but can be seen in the social network. For example, users can see the music listening list of their friends in last.fm; users can see score rankings of friends in social-network based online games; users can see the uploaded photos of friends in social network site such as instagram or facebook. Therefore, the physical distances become less effective than distances on online social network. The distance on online social network can be described in many ways. One is transitivity or clustering coefficient which shows how strongly connected their friends are. If friends are connected together, or can be said that the node is in the close group of friend, should have shorter distance.

As users are influenced by their friends on social network, the more friends they have, the more information they get. Therefore, the number of songs that are listened increases in relation to the number of both friends who are premium users and the number of friends who are free users. However, the stronger connection their friends are, the more information they share. This makes the variety of information that the user received from friends deceased compare to user who have same number of friends but have weak connection strength between his/her friends. Therefore, the information from their friends decreases as the strength of their friend's connection increases. The

decreasing information will lead to the decreasing of usage in the music streaming service, which means that the number of songs that are listened to will decrease too. This makes the number of songs that are listened to decreases in relation to the strength of friends' connection.

Hypothesis 1. *Higher number of premium friends, higher number of free friends and weaker connection between friends lead to higher usage.*

Social identity theory [130] describes that people will behave in the way that their society perceives as distinctively positive. Premium users perceive paying for the service as a positive behavior. Therefore, the more premium friends a user has, the higher the chance that a user is a premium user.

Hypothesis 2. *Higher number of premium friends leads to higher chance to be a premium user.*

Self-categorization theory [134] states that a person will perceive collections of people as a group by the similarities of the people. Therefore, free users will be closer to free user's friends than premium user's friends. On the other hand, premium user will be closer to premium user's friend than free user's friends. And social identity theory [130], which says that individual will have behavior that is positive distinctive, made us can create the hypotheses about relationship between behavior and social network as follows. Because using only free service is normal behavior of free users and using of premium service is normal behavior of premium users, using only free service should be positive to free users and using premium service should be positive to premium users.

Hypothesis 3. *Stronger connections between free friends of a free user leads to lower chance to become a premium user.*

Hypothesis 4. *Stronger connections between premium friends of a premium user leads to higher chance to continue to be premium user.*

Social identity theory describes that people have intentions to perform positive distinctive behavior, and for the groups which can move in and out easily, one of the positive distinctiveness strategies that can be taken is individual mobility [57], which is to leave the group and increase social status in comparison to the others in the group. Premium users, who are paying to use the premium services that only paid users can use, can be viewed at as people who have higher social status in the community when compared to the free user who cannot use the premium services. And starting using of premium service of the free user can be implied as leaving the free user's

group and get higher social status. This makes the free users that change to premium users can be viewed by their free friends as a positive distinctive behavior, which is one type of individual mobility. Therefore, a free user who has a lot of free friends can be influenced to change to a premium user as it is perceived as a positive distinctive behavior.

Hypothesis 5. *Higher number of free friends of a free user leads to higher chance to become a premium user.*

Social identity theory describes that people have intentions to perform positive distinctive behavior [130]. Free users' positive distinctive behavior can be the use of the service for free. Therefore, a premium user who has a lot of free friends can be influenced to become a free user as it is perceived as a positive distinctive behavior. That means a premium user who has a lot of free friends will have less chance to continue to be premium user.

However, self-categorization theory [134] states that a person will perceive collections of people as a group by the similarities of the people and the distances between groups depend on the differences between the groups. This can lead to the assumption that the people who start using premium service is closer to the free user group than the people who continue using the premium service, because they just changed group that they were belonging. Therefore, the premium users who just started using premium services will be affected by free friends more than the people who have continued using premium service.

However, the previous hypothesis did not make a distinction between free users who have not used premium services and free users who just stopped using premium services. This is because the previous hypothesis focuses on the influence of free friends and the two subgroups are both free users. Therefore, there are no difference between these two subgroups of free users.

Hypothesis 6. *Higher number of free friends of a premium user leads to lower chance to continue to be a premium user, and the effect is higher for the premium users who were recently free users.*

From expectancy theory [137], which describes the expectation of behavior, and expectation confirmation theory, which describes the evaluation of adoption, the group of users which stopped using premium service got disconfirmation of beliefs. This means that there are differences between expectation and perceived performance.

The disconfirmation of beliefs occurs when the expectation is higher than the perceived performance. As perceived performance will not change after a user stopped using the premium service, the expectation have to be changed before reusing the premium service, and the new expectation

have to be lower than the previous perceived performance. However, if the usage is high, the perceived performance should be good too, which means that the expectation have to be much higher to have created disconfirmation of beliefs. This made the probability to lower expectation small. This means that, for the users who stopped using premium service, the more usage of service, the less chance that they will use the premium service again.

But for users who never use premium service, they have only the expectation, which is explained in expectancy theory as the valuation of reward related to performing behavior, which should be related to the usage of that user. The more usage that user has, the higher the expectation and the higher the chance to use the premium service.

Hypothesis 7. *Higher usage of a free user leads to higher chance to become a premium user, but leads to lower chance to become premium users again for the people who stop using premium service.*

Premium users are users who already have the perceived performance and still didn't get the disconfirmation of beliefs. Therefore, the expectation of premium users will be related to the usage. The more usage the user has, the higher performance they perceived, the less chance the expectation to be higher than perceived performance, and the less chance the disconfirmation of beliefs can occur. This should be the same for both people who start using the premium service and people who continue using the service.

Hypothesis 8. *Higher usage of a premium user leads to higher chance to continue to be a premium user.*

From dynamic use diffusion model [125], which shows the effects from determinants in 4 dimensions to the index and the effect from the index to the outcome such as user satisfaction, the hypothesis can be used to modify 2 parts of the model which are the determinants and the index. The number of free friends (#FreeFriend@t1 and #FreeFriend@t2) and premium friends (#PremiumFriend@t1 and #PremiumFriend@t2), which is degree of node in social network metrics, and connection strength between friends (Clustering@t1 and Clustering@t2), which is local clustering coefficient in social network metrics, can be used as social network dimension, which is new dimension in the model. The usage (Current Usage) and subscription history (#Payment@t2) can be used as the index of the model.

All of the proposed hypotheses can be summarized into one diagram as shown in Figure 3.1. The explanatory variables that related to UD determinants are colored in yellow and two variables

related to the dynamic use index are colored in blue and two dependent variables which are UD outcomes are colored in pink.

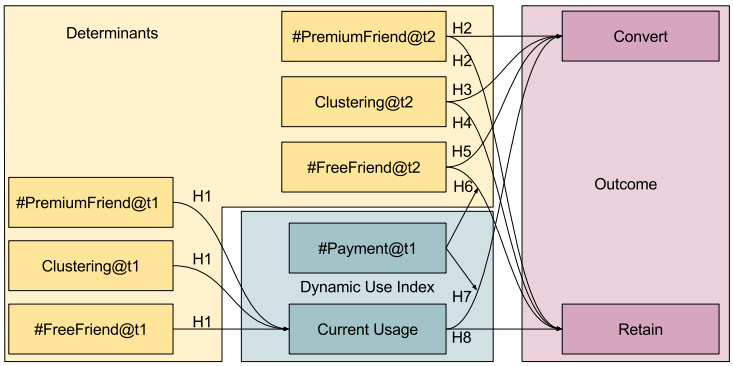

Figure 3.1: All hypotheses

3.4 Data

3.4.1 Data Source and Target User Group

Empirical data was collected from last.fm, an online music streaming service founded in UK in 2002 to provide music and radio. Last.fm collects listening data from users around the world, which contain users' information, friend relationship and subscription data. The data was collected at three points of time, which are August 2013 (t_1), June 2014 (t_2), December 2014 (t_3), to get the historical information of listening data, subscribe data and social network data. The aim of this chapter is to analyze the premium subscription and unsubscribes using user's usage data, user's network position and the subscription history of the user's friends.

3.4.2 Explanatory and Dependent Variables

There are five types of variables, some are explanatory variables and some are dependent variables in the regression models, that were calculated for each user. The first four types of variables are treated as input or explanatory variables to the models, which are an indication of previous subscription of the user, indications of how the user's friends change subscription in the previous time steps, the user's social network metrics, and the user's activity. User's activity is both explanatory and dependent variables up to the time point of the data and regression model. The fifth type of

variable is the user's subscription status at t_3, which is the dependent variable of the regression model.

The first group of variables is the user's own subscriptions status in time t_1 and t_2 for this group, which is sub1 and sub2 which have number of users as shown in 3.2. These variables are dummy variables. The subscription status at time t_2 was used to separate users into 2 groups as the free users and the premium users and use regression on each users group to confirm our hypotheses in sec 3.3. And sub1 is referred as #Payment@t1 in Figure 3.1

Table 3.2: User groups and statistics. Users are segmented into four groups according to their subscription statuses at t_1 (sub1) and t_2(sub2).

sub1	sub2	Changes in Subscription Status of Period 1 (p_1)	Number of Users
0	0	Users who remain as a free user (at both time t_1 and t_2)	117,840
1	0	Users who change from a premium user (at t_1) to a free user (at t_2)	2,179
0	1	Users who change from a free user (at t_1) to a premium user (at t_2)	417
1	1	Users who remain as a premium user (at both time t_1 and t_2)	2,477
Total			122,913

The second group of variables indicate number of friends separated by the subscription status of their friends at time t_1 and t_2. Due to the number of friends can increase exponentially, the natural logarithm of number of friends in each group are used. The #FreeFriend@t1 and #FreeFriend@t2 are indicated to the natural logarithm of number of friends who is free user at time t_1 and t_2, and the #PremiumFriend@t1 and #PremiumFriend@t2 are indicated to the natural logarithm of number of friends who is premium user at time t_1 and t_2.

The third group of variables are the network parameters of each user. From friends' information of same subscription status, a user network can be created, separated by subscription status at time t_1 and t_2, where each user is represented by a node, and an edge between two nodes represents that the users are friends. The clustering@t1 and clustering@t2 are indicated for local clustering coefficient which equal to the ratio of number of paired friends who connected together to the all possible number of connection. These indicate the strength of connection between each user's

friends, if all of the user's friends all know each others, the value equal to 1, if they are not know each others, the value is 0.

The fourth group of variable indicates the user's activity, for which the number of the unique songs that the user listened to between time t_1 and t_2 are calculated. The variable is called Playcount.ln, which indicated the natural logarithm of number of unique songs that user listened to.

The fifth group is the ground truth contains two variables, convert and retain, separated by subscription of the user at time t_2. The variable convert is for the free users and retain is for the premium users, which is 1 if the user is a premium user at t_3, and 0 otherwise. In summary, for each user, nine variables are calculated as listed in Table 3.3. Eight variables are input to the regression model and the output of the regression model is the premium status at time t_3, which is captured by variable convert or retain up the subscription status at time t_2 of each user.

The statistics of each parameter are shown in Table 3.4. Note that the transitivity cannot be computed if the number of connected node is less than 2, which results us to have to dropped some data. This made the number of data of Clustering@t1 is less than the others.

Table 3.3: List and description of user's calculated variables. For each user, the following variables are calculated.

Variable	Description
sub1	Dummy variable indicating that the user is a premium user at t_1
#FreeFriend@t1	Natural log of number of friends who are a free user at t_1
#FreeFriend@t2	Natural log of number of friends who are a free user at t_2
#PremiumFriend@t1	Natural log of number of friends who are a premium user at t_1
#PremiumFriend@t2	Natural log of number of friends who are a premium user at t_2
Clustering@t1	Local clustering coefficient for free and premium graph at t_1
Clustering@t2	Local clustering coefficient for free and premium graph at t_2
playcount.ln	Natural log of number of unique songs that user listened to between t_1 and t_2
convert	Dummy variable indicating that a free user at t_2 is a premium user at t_3
retain	Dummy variable indicating that a premium user at t_2 is a premium user at t_3

Table 3.4: Statistic of variables

Statistic	N	Mean	St. Dev.	Min	Max
sub1	121,430	0.031	0.173	0	1
#FreeFriend@t1	121,430	2.294	0.996	0.000	6.807
#FreeFriend@t2	121,430	2.675	1.068	0.000	7.422
#PremiumFriend@t1	121,430	0.439	0.632	0.000	4.997
#PremiumFriend@t2	121,430	0.434	0.689	0.000	5.505
Clustering@t1	116,432	0.071	0.140	0.000	1.000
Clustering@t2	121,430	0.016	0.028	0.000	0.237
playcount.ln	121,430	3.975	2.386	0.000	9.353

3.5 Experiment and results

The linear regression and logit regression were used to confirm the proposed hypotheses, and the results for coefficients was shown in Table 3.5 and the marginal in Table 3.6.

In the first model, the linear regression was used to confirm Hypothesis 1 by using the playcount.ln as dependent variable and sub1, #FreeFriend@t1, #PremiumFriend@t1 and Clustering@t1 as the explanatory variables. The F-Test's p-value of the regression is less than $2.2 * 10^{-16}$, which allows us to reject the equality of model with only intercept and model with explanatory variables, and accept that the dependent variable is related to the selected explanatory variables. The coefficient of #FreeFriend@t1 is equal to 0.356^{***}; the coefficient of #PremiumFriend@t1 is equal to 0.026^{*}; the coefficient of Clustering@t1 is equal to -0.301^{***}, which mean that the number of unique songs that are listened to between time t_1 and t_2 increased in relation to number of free friends and number of premium friends but decreased in relation to the strength of connection between their friends. Additionally, Hypothesis 1 is supported at significant level 90%

For the other hypotheses, the logit regression was done on 2 groups of users' data separated by the subscription status at time t_2. Both groups use the number of friends who are premium users, the number of friends who are free users, the strength of connection between friends of each user, the payment status at time t_1 and the number of unique songs that user listened to as

Table 3.5: Coefficient of regression results

	Dependent variable:		
	playcount.ln	convert	retain
	OLS	*logistic*	
	(1)	(2)	(3)
Constant	3.160^{***}(0.020)	-8.375^{***}(0.321)	0.519(0.507)
sub1	0.560^{***}(0.049)	3.825^{***}(0.361)	−0.294(0.530)
#PremiumFriend@t1	0.026^{*}(0.014)		
Clustering@t1	-0.301^{***}(0.050)		
#FreeFriend@t1	0.356^{***}(0.009)		
#PremiumFriend@t2		0.292^{**}(0.110)	0.348^{***}(0.077)
Clustering@t2		−6.947(4.737)	10.500^{***}(2.668)
#FreeFriend@t2		0.255^{**}(0.086)	-0.446^{**}(0.138)
playcount.ln		0.160^{***}(0.045)	0.059^{*}(0.025)
#FreeFriend@t2:sub1			0.331^{*}(0.144)
playcount.ln:sub1		-0.170^{*}(0.070)	
Observations	116,432	119,595	1,835
R^2	0.026		
Adjusted R^2	0.026		
Log Likelihood		−1,221.885	−1,091.008
Akaike Inf. Crit.		2,457.769	2,196.017
Residual Std. Error	2.349 (df = 116427)		
F Statistic	770.932^{***} (df = 4; 116427)		

Note: *p<0.1; **p<0.05; ***p<0.01

the explanatory variables and the status of subscription at time t_3 of each group as the dependent variable. The result of regression was shown as model (2) and model (3) in Table 3.5 and marginal effect are shown as model (2) and model (3) in Table 3.6.

From both model (2) and model (3), the coefficient, shown in Table 3.5, and the marginal effect, shown in Table 3.6, of #PremiumFriend@t2 are positive at significant level 95% and 99% (coefficient = 0.292^{**} and 0.348^{***}, marginal effect = 0.267^{**} and 74.705^{***}). This means that when the number of friends who are premium users increases, both the chance to start using premium service for free user and the chance to continue using premium service for premium user also increase. Therefore, Hypothesis 2 is supported.

The coefficient of Clustering@t2 of model (2) is negative; however the p-value is large, so Hypothesis 3 is not supported.

However, the coefficient and the marginal effect of Clustering@t2 of model (3) is positive with significant level 99% (coefficient = 10.500^{***} and marginal effect = 2251.850^{***}), which means that the chance to continue using premium service increases as the strength of connection between premium friends of the premium user increases. Therefore, Hypothesis 4 is supported.

As the coefficient and the marginal effect of #FreeFriend@t2 of model (2) is positive with significant level 95% (coefficient = 0.255^{**} and marginal effect = 0.234^{**}), the chance to become premium user increases as the number of friends who are free users increases. Therefore, Hypothesis 5 is supported.

The coefficient and the marginal effect of #FreeFriend@t2 of model (3) is negative with significant level 95% (coefficient = -0.446^{**} and marginal effect = -95.658^{**}). This means that the chance to continue to be premium user decreases as the number of friends who are free user increases for user who just start using premium service. Additionally, the coefficient and the marginal effect of interaction term #FreeFriend@t2:sub1 of model (3) is positive with significant level 90% (coefficient = 0.331^{*} and marginal effect = 70.939^{*}), which means the chance to continue to be premium user of the user that continue using premium service both time t_1 and t_2 is affected less from the number of the friends who are free user. However, as shown in Table 3.7, the coefficient and marginal effect of the group of user who continue using premium service are negative, the relationship is in the same direction with the users who start using premium service at t_2. Therefore, Hypothesis 6 is supported.

As the coefficient and the marginal effect of playcount.ln of model (2) is positive with significant level 99% (coefficient = 0.160^{***} and marginal effect = 0.146^{***}), the chance to become premium

Table 3.6: Marginal effect of logit models

	Marginal (x 1000):	
	convert	retain
sub1	$36.893^{**}(0.012)$	$-60.185(0.103)$
Clustering@t2	$-6.358(0.004)$	$2251.850^{***}(0.565)$
#PremiumFriend@t2	$0.267^{**}(0.000)$	$74.705^{***}(0.016)$
#FreeFriend@t2	$0.234^{**}(0.000)$	$-95.658^{**}(0.030)$
playcount.ln	$0.146^{***}(0.000)$	$12.565^{*}(0.005)$
#FreeFriend@t2:sub1		$70.939^{*}(0.031)$
playcount.ln:sub1	$-0.156^{*}(0.000)$	
Num. obs.	119595	1835
Log Likelihood	-1221.885	-1091.008
Deviance	2443.769	2182.017
AIC	2457.769	2196.017
BIC	2525.612	2234.621

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.7: Coefficient and marginal effect of retain

	COEF	COEFsub1	dYdX	dYdXsub1
(Intercept)	0.519	0.226	0	-0.060
friends.free.ln	-0.446	-0.115	-0.096	-0.025

user increases as the number of listened unique songs increases for user who don't use premium service. However, the coefficient and the marginal effect of interaction term playcount.ln:sub1 of model (2) is negative with significant level 90% (coefficient = -0.170^{*} and marginal effect = -0.156^{*}), and as shown in Table 3.8, the coefficient and marginal effect of the group of user who stop using premium service become negative. This means that the chance to become premium user of the user who stop using premium service between time t_1 and t_2 decreases in relation to the increase of number of the listened unique songs. Therefore, Hypothesis 7 is supported.

Table 3.8: Coefficient and marginal effect of convert

	COEF	COEFsub1	dYdX	dYdXsub1
(Intercept)	-8.375	-4.549	0	0.037
Playcount.ln	0.160	-0.011	0.0001	-0.00001

The coefficient and the marginal effect of playcount.ln of model (3) is positive with significant level 90% (coefficient = 0.059^{*} and marginal effect = 12.565^{*}), which means the chance to become premium user increases as the number of listened unique songs increases. This result supports Hypothesis 8.

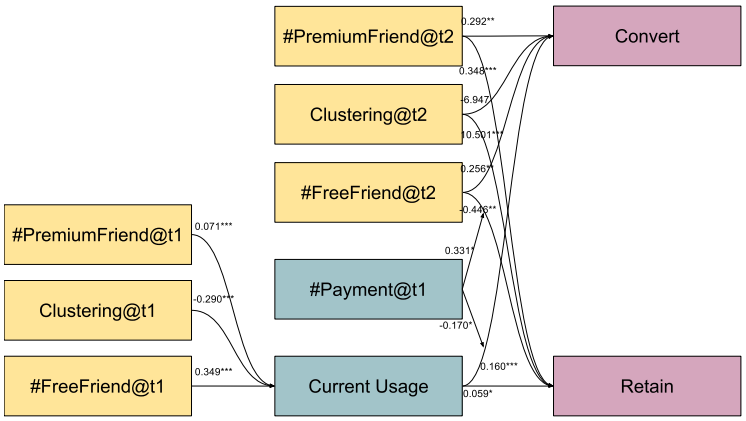

Figure 3.2: Results

3.6 Discussion and managerial recommendations

As hypothesis 1 was supported, the usage of a user can be said to be related to the amount of information that the user received from his/her friends. Additionally, the usage of a user and can be increased by increasing the number of free friends and premium friends but the connection strength between friends have to be reduced. To increase the number of friends without increase the connection strength between friends, the provider can recommend friends which have fewer connection to that user, which means that the criteria of friend's recommendation should use the behavior based instead of common friend based.

As hypothesis 2 was supported, increasing the connection from premium user to the other users will increase the chance of becoming premium user of that user, but due to the hypothesis 6 also be supported, connections to free users will increase chance of unsubscription of that premium user. This means that the provider should not recommend the connection between free and premium users, but should recommend only premium user to premium user. Together, as hypothesis 4 was supported, if connection strength between premium users increase, the chance to continue being premium user of their common premium friend will also increase.

As hypothesis 3 was not supported, the relationship between the strength of connection between friends of free users and chance of unsubscription can be ignored. This may be because of the free friends have both positive and negative effect to the free user due to the self-identity theory, and both effects are canceled in this case. However, due to the hypothesis 5 was supported, the provider should increase connection between free users to increase chance to become premium user of their common friend.

As hypothesis 7 was supported, the provider should increase the usage of free user who never subscribe to the premium services, and find the unsubscribed users who have low usage to proposed some marketing campaign such as discount promotion and community party because they have higher chance to become premium user again when compared to unsubscribed users who have high usage.

As hypothesis 8 was supported, the provider should increase the usage of the premium users to increase the chance to continue being premium user, which can be done by recommending new musics to that user.

For the understanding of the relationship between social network, usage and subscription behavior of the users in online-social-network based freemium music streaming service, the dynamic

use diffusion model can be modified by applying the **social network dimension** which contains social network metrics of the user into the determinants and historical subscription status into index of the model. The modified model can be used to analysis online social network based freemium service which is getting popular now.

From the results of hypothesis confirmations, the suggestions which can be proposed to the service provider are about increasing the number of premium users and preventing churns.

For free users, service providers want them to convert to premium users. From the results above, the number of friends is positive related to the chance to become premium user, for both premium users and free users. This means that if the service provider increases the connection to other users, the free user may have more chance to be premium user, which can be done by add friend recommendation system to the service.

For premium users, service providers want to prevent churns. From the results above, the number of friends who are premium users has positive relationship to the chance to continue using premium service but the number of friends who are free user has negative relationship. This means that the provider should recommend premium user to be friend with premium user than free user, because the connection to free user increase the churn rate. Together, the strength of connection between friends who are premium users also increase the chance to continue to be premium user. The more connection between premium users will increase strength of the connection and increase the chance to continue to be premium users as well. Furthermore, the usage also has positive relationship to the chance to continue to be premium user, which means that the recommendation of new songs that increases the usage are also able to prevent churns.

For people who stop using premium service, the provider can focus on user who has less usage and more number of friends because they have higher chance to re-use the premium service. The discount and special training for the premium service to increase perceived performance can be used to prevent the disconfirmation of belief again.

3.7 Conclusion and future works

Hypotheses in Section 3.3 and the results of hypothesis testing in Section 3.5 confirmed that the previous usage pattern and the social network is related to premium usage in next step. This can be used to improve the model by adding explanatory variables, which are related to social network, into the model. The variables that are used in this research are number of friends both premium users and free users, and clustering coefficient which shows how strong friends connected together.

The hypotheses and the testing results show that the payment and usage history of users also affect their change of premium service usage. There is one hypothesis which are not supported by the empirical hypothesis testing, which can be removed from the model improvement. However, the more dataset have to be used to significantly reject the unsupported hypothesis again.

As the premium service usage can be used as indicator of satisfaction to the service, or the final outcome of the dynamic use-diffusion model [125], the house-hold social context in the determinants can be changed to social network matrices and the payment pattern can be added to dynamic use-diffusion index for the freemium music streaming service.

From the hypothesis testing results, the managerial recommendation are proposed for the provider to increase premium users and prevent churns by making different friend recommendation to difference group of users, and providing the selection criteria of the users who stop using premium service to do targeted marketing to make them use the premium service again.

For future work, the proposed model can be used to predict the change of subscription by applying these parameters and may be able to increase conversion rate by both increasing the chance of subscription and decreasing the number of unsubscription by offering some special discount to users who have influence power. As the subscription of the user increase the chance to be premium user of their friends, this will decrease the chance of his/her friends to unsubscribe too. Together, the model can be applied to the other freemium online services such as video streaming services, or online data storage services.

Chapter 4

RECOMMENDATION SYSTEM BY RANDOM WALK WITH RESTART USING CONDITIONAL TRANSITION PROBABILITY ON SOCIAL INFORMATION

A recommendation system is an information retrieval system that employs user, product, and other related information to infer relationships among data to offer product recommendations. The basic assumption is that friends or users with similar behavior will have similar interests. The large number of products available today makes it impossible for any user to explore all of them and increases the importance of recommendation systems. However, a recommendation system normally requires comprehensive data relating users and products. Insufficiently comprehensive data creates difficulties for creating good recommendations. Recommendation systems for incomplete data have become an active research area. One approach to solve this problem is to use random walk with restart (RWR), which significantly reduces the quantity of data required and has been shown to outperform collaborative filtering, the currently popular approach. This study explores how to increase the efficiency of the RWR approach. We replace transition matrices that use information regarding relationships between user, usage, and tags with transition matrices that use conditional probability based on social information, and we compare the efficiency of the two approaches using mean average precision. An experiment was conducted using music information data from last.fm. The result shows that our approach provides better recommendations specially in limited data case.

4.1 Introduction

Exponential increases in available information emphasize the importance of information filtering systems. As stated by Chris Anderson, “The secret to creating a thriving Long Tail business can be summarized in two imperatives: (1) Make everything available. (2) Help me find it.” [11]. The success of any online service lies in helping users find what they are interested in, even before they realize that they are interested in it. Recommendation systems help online service providers offer personalized product suggestions by predicting user responses to items they have not yet considered.

Traditional recommendation systems achieve this goal by using either content-based filtering

(CBF), which analyzes product characteristics, or collaborative filtering (CF), which analyzes user behaviors. However, the popularity of social networking has prompted researchers to use the concept of friendship and social information to increase recommendation accuracy. Konstas et al. [74] showed that a generic framework of random walk with restart (RWR) that includes social annotation (tags) and friendship established among users outperforms the currently popular CF approach. While social annotations and friendship data capture some information regarding how a user is related to a product, it does not provide a straightforward probability. The authors hypothesize that this relationship is captured more effectively using conditional probability transition based on social network information of the user instead of the direction relation between users, items and tags. In this study, we extend the work of Konstas et al. [74] using conditional probability transition based on social network information of the user in RWR' s transition matrices to increase recommendation accuracy.

We evaluated our modified model against the method proposed by Konstas et al. [74] on a data set we collected from last.fm, an online music recommendation service. The dataset includes user, friendship, artist, and usage data. The modified model achieved better recommendation accuracy, particularly with limited data (80% of data removed). The contributions of this study include the following:

- We evaluated the use of social network based conditional probability in RWR transition matrices and found that it outperforms the original RWR model.
- We changed from track to artist recommendations.
- We found that changes in parameters (alpha) have little effect on recommendation accuracies.

The rest of this paper is organized as follows. Section 4.2 reviews previous related work. Section 4.3 describes the proposed method. The experiments, including how we collected data from last.fm, are explained in Section 4.4. We discuss the implications of our study and draw conclusions in Section 4.5 and Section 4.6, respectively.

4.2 Related Work

4.2.1 Information Retrieval

Information retrieval (IR) is the finding of information that is relevant and satisfies an information need from a normally unstructured and large collection of information resources. As collection of

information resources is big, people have difficulties in navigating within the collection manually. Such difficulties are called *information overload*, and automated information retrieval systems help reduce the difficulties [142].

The two most frequent and basic measures for information retrieval effectiveness are precision and recall [111]. Precision (P) is the percentage of retrieved information that are relevant and recall (R) is the percentage of relevant information that are retrieved, as calculated by formula (4.1) and (4.2).

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{4.1}$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \tag{4.2}$$

The standard measure in the Text Retrieval Conference (TREC http://trec.nist.gov/) community is the mean average precision (MAP), which provides a single-figure quality measure, across recall levels. In particular, among available evaluation measures, MAP has been shown to have effective discrimination and stability [85].

For each information query, the average precision (AP) is calculated by averaging the precision value of a set of documents after each relevant document is retrieved. MAP is the average AP for all queries for all the related documents. In other words, if the set of relevant documents for an information need $q_j \in Q$ is $d_1, \ldots, d_{m_j}$ and R_{jk} is the set of ranked retrieval results from the top result until you arrive at document d_k, then

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{4.3}$$

4.2.2 Recommendation systems

Recommendation systems, sometimes also called "recommender systems," are an extensive class of Web applications that involve predicting user responses to options [107].

There has been substantial research on this topic showing that a recommendation system can help users with information retrieval ([108, 67, 59, 135, 14, 56]).

Such systems typically use one of three approaches.

1. Collaborative filtering (CF) [128]: This method employs user behavior information, such as ratings or usage of an item, to find similar users and try to predict a missing usage or rating. There have been many studies using systems with this approach. As examples,

- Cho, et al. [32] use data mining and decision tree on web usage, while
- Kim, et al. [71] create groups of people who have similar activities with items.

2. Content-based filtering (CBF): This method uses item similarities and recommends the item closest to the items used by a target user. As examples,

   - Baraglia and Silvestri [22] cluster the contents and use the results for recommendation, while
   - Han, et al. [55] introduce an algorithm based on rules created from combinations of items selected together.

3. Hybrid recommendation systems [29]: This approach uses information regarding both users and items.

Recent rapid social network growth has created interest in using social information in recommendation systems. Studies have shown that people on the same network or having friendship relations share similarities ([38, 35, 6]). This shows that social information can be used to suggest directions for finding recommendations based on relationship and item usage information.

Tagging, in which user-generated keywords are attached to online contents, is also used in recommendation systems, for example, to identify items to be retrieved in the future ([102, 110, 63, 58, 133]). Studies have shown a relationship between friendship and tagging, or social bookmarks ([13, 24]).

A tagging system (folksonomy) model is often characterized by a tripartite graph with hyperedges. The three disjoint, finite sets of such a graph correspond to

1. a set of persons or users $u \in U$
2. a set of resources or objects $o \in O$ and
3. a set of annotations or tags $t \in T$

which are used by users U to annotate objects O. A very general model of folksonomies is defined by a set of annotations $F \subseteq U \times T \times O$ ([127, 76, 87, 60]).

Studies have shown that tags can be used as inputs to a recommendation system ([86, 43, 27, 52]).

However, since a recommendation system uses information as input, a system cannot provide suitable recommendations, when there is insufficient or sparse data.

Huang et al. [65] created social item-and-user graphs and used graph analysis on this problem.

4.2.3 Random walk with restart (RWR)

Random walk, a series of random variables [51], was first introduced by Karl Pearson in 1905 [101] and has been used in many fields. Google™'s well-known PageRank is based on random walk [97].

Random walk on graphs is a series of random variables X_i where X_i is a connected vertex selection for each node of each step ([10, 82, 81]).

RWR is the random walk that has probability α of jumping to the starting point, as shown in equation (4.4).

$$p^{t+1} = (1 - \alpha)Sp^t + \alpha q \tag{4.4}$$

where p^t and p^{t+1} are the probabilities of remaining at each node at steps t and $t + 1$, S is the transition matrix, α is the restart ratio, and q is the probability of remaining at each node at the starting step.

When the start stage probabilities are set equal, as in (4.5), the probabilities at the stable stage of node y show the relationship between node x and y ([144, 132])

$$q_i = \begin{cases} 1 & i = x \\ 0 & i \neq x \end{cases} \tag{4.5}$$

4.2.4 RWR-based Recommendation System

Studies on the use of RWR on related topics include the following.

- Clements et al. used RWR in information retrieval [34].
- Craswell, et al. used random walks to create rankings of documents for a given query [36].
- Barnd modeled consumer behavior as random walks on a weighted association graph [28].
- Fouss et al.present a new perspective on characterizing the similarities among elements of a database [48].

Furthermore, Konstas et al. [74] show results indicating that RWR outperforms the standard CF method using the four transition matrices shown in Fig. 4.1. The main target of using RWR is to deal with cold start problems, which will occur when there is insufficient information.

	User	Track
User	I	UTr
Track	TrU	$TrTr$

	User	Track
User	UU	UTr
Track	TrU	$TrTr$

	User	Track	Tag
User	I	UTr	UTg
Track	TrU	$TrTr$	$TrTg$
Tag	TgU	$TgTr$	$TgTg$

	User	Track	Tag
User	UU	UTr	UTg
Track	TrU	$TrTr$	$TrTg$
Tag	TgU	$TgTr$	$TgTg$

Figure 4.1: Four types of transition adjacency matrices for a recommendation system using RWR

Not only music ([74, 90]), the RWR based recommendation system was applied into many targets of recommendation such as, movies ([41, 48, 140, 122, 79, 141]), publications [131], point of interests (PoIs [83]), Dictionary/Internet/Citation/Social/Email [49].

However, there are no studies regarding the effect of data size, recommendation accuracy, or restart ratio.

4.3 Methodology

RWR is used in creating recommendations from user, item, and social information similar to the method used by Konstas et al. [74]. I modified the method by creating relationships for artists

instead of tracks and user-artist tags instead of user-track tags, as shown in Fig. 4.2 and equations (4.6)-(4.9). This modification reduces the number of calculations required.

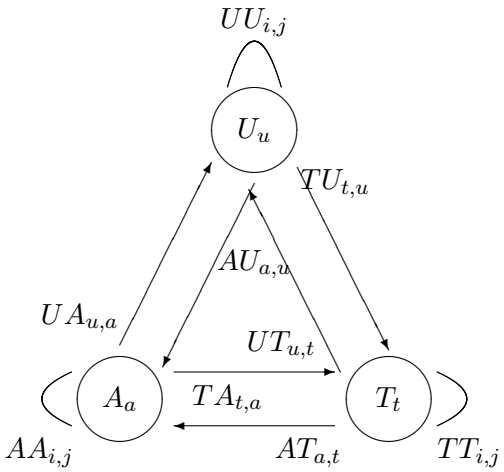

Figure 4.2: Graph for RWR

$$S_N = \begin{pmatrix} I & UA_{u,a} \\ AU_{a,u} & AA_{i,j} \end{pmatrix} \tag{4.6}$$

$$S_F = \begin{pmatrix} UU_{i,j} & UA_{u,a} \\ AU_{a,u} & AA_{i,j} \end{pmatrix} \tag{4.7}$$

$$S_T = \begin{pmatrix} I & UA_{u,a} & UT_{u,t} \\ AU_{a,u} & AA_{i,j} & AT_{a,t} \\ TU_{t,u} & TA_{t,a} & TT_{i,j} \end{pmatrix} \tag{4.8}$$

$$S_B = \begin{pmatrix} UU_{i,j} & UA_{u,a} & UT_{u,t} \\ AU_{a,u} & AA_{i,j} & AT_{a,t} \\ TU_{t,u} & TA_{t,a} & TT_{i,j} \end{pmatrix} \tag{4.9}$$

To improve recommendation accuracy, the traditional transition matrices are replaced with social-network-based conditional probability transition matrices. The main idea is to utilize both information from the user and information from the user's friends, instead of using only that user's information. The constructed social-network-based conditional probability transition matrices are used to perform RWR on the graph.

There are 9 transitions in the transition matrix as illustrated by the arrows in Fig. 4.2. The social information are added into the transition matrix as follows.

- For transition from user to artist ($AU_{a,u}$), the more listening counts from the friends of that users, the higher the transition probability. Therefore, the probability that the walker will move from user node i to artist node j was changed to add social information. In [74], this was the ratio of the number of times that user i listens to artist j to the number of times that user i listens to all songs. To add social information, this was changed to the ratio of the number of times that user i and his friends listen to artist j to the number of times that all users listen to artist j divided by the ratio of the number of friends of user i to the total number of users. This yields equations 4.10.

$$AU_{a,u} = \frac{\frac{\sum_{k=1}^{\#u} F_{k,u} P_{k,a}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} P_{i,j}}}{\frac{\sum_{k=1}^{\#u} F_{k,u}}{\#u}} \tag{4.10}$$

- For transition between users ($UU_{i,j}$), the more friends the two users have in common, the higher the transition probability.

$$UU_{i,j} = \frac{\sum_{k=1}^{\#u} F_{k,i} F_{k,j}}{\sum_{k=1}^{\#u} F_{k,j}} \tag{4.11}$$

- For transition from artist to user ($UA_{u,a}$), the more listening counts from the friends of that users, the higher the transition probability.

$$UA_{u,a} = \frac{\sum_{k=1}^{\#u} F_{k,u} P_{k,a}}{\sum_{k=1}^{\#u} P_{k,a}} \tag{4.12}$$

- For transition from tag to user ($UT_{u,t}$), the more tags from the friends of that users, the higher the transition probability.

$$UT_{u,t} = \frac{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} F_{i,u} T_{i,j,t}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} T_{i,j,t}} \tag{4.13}$$

- For transition from artist to artist ($AA_{i,j}$), the more common users that listen to their songs, the higher the transition probability.

$$AA_{i,j} = \frac{\sum_{k=1}^{\#u} L_{k,i} P_{k,j}}{\sum_{k=1}^{\#u} P_{k,j}} \tag{4.14}$$

- For transition from tag to artist ($AT_{a,t}$), the higher listening counts from user that tag to the artist, the higher the transition probability.

$$AT_{a,t} = \frac{\frac{\sum_{k=1}^{\#u} T_{k,a,t} P_{k,a}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} T_{i,j,t} P_{i,j}}}{\frac{\sum_{k=1}^{\#u} P_{k,a}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} P_{i,j}}} \tag{4.15}$$

- For transition from user to tag ($TU_{t,u}$), the more friends of the user that use the tag, the higher the transition probability.

$$TU_{t,u} = \frac{\frac{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} F_{i,u} T_{i,j,t}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} \sum_{k=1}^{\#t} T_{i,j,k}}}{\frac{\sum_{k=1}^{\#u} F_{k,u}}{\#u}} \tag{4.16}$$

- For transition from artist to tag ($TA_{t,a}$), the higher listening counts of the user that use the tag to that artist, the higher the transition probability.

$$TA_{t,a} = \frac{\frac{\sum_{k=1}^{\#u} T_{k,a,t} P_{k,a}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} T_{i,j,t} P_{i,j}}}{\frac{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} T_{i,j,t}}{\sum_{i=1}^{\#u} \sum_{j=1}^{\#a} \sum_{k=1}^{\#t} T_{i,j,k}}} \tag{4.17}$$

- For transition from tag to tag ($TT_{i,j}$), the more common users that tag both, the higher the transition probability.

$$TT_{i,j} = \frac{\sum_{x=1}^{\#u} \sum_{y=1}^{\#a} W_{x,y,i} T_{x,y,j}}{\sum_{x=1}^{\#u} \sum_{y=1}^{\#a} \sum T_{x,y,j}} \tag{4.18}$$

where,

$$\begin{aligned}
\#u &= \text{Number of users} \\
\#a &= \text{Number of artists} \\
\#t &= \text{Number of tags} \\
F_{i,j} &= \begin{cases} 1 & \text{User i and user j are friends} \\ 0 & \text{otherwise} \end{cases} \\
P_{u,a} &= \text{Number of playcount by user u on artist a} \\
T_{u,a,t} &= \text{Number of tag by user u on artist a with tag t} \\
L_{u,a} &= \begin{cases} 1 & P_{u,a} > 0 \\ 0 & P_{u,a} = 0 \end{cases} \\
W_{u,a,t} &= \begin{cases} 1 & T_{u,a,t} > 0 \\ 0 & T_{u,a,t} = 0 \end{cases}
\end{aligned}$$

These transition probabilities have to be normalized to 1 in all type of transition adjacency matrices shown in Fig. 4.1

4.4 Experiments

4.4.1 Data collection

Data from last.fm were collected through their free web service API which provided data in xml format. The essential data includes user information, which can be obtained from user.getInfo. The user id list was not public information, but the authors worked around this problem by requesting random user ids and obtained more information from the user's friend relations (user.getFriends). Repeating this process provided us with a quantity of user information. Another essential datum is the artist information, which was obtained through user.getTopArtists. The information associated with this includes play counts (the number of times this user plays songs from this artist) and social tags (user.getTopTags). The relationships between users, artists, and tags were also collected, through user.getPersonalTags. The last.fm license agreement forbids obtaining an exhaustive list of data; hence, the authors collected a subset of data including 11,239 users, 49,000 artists, and 11,726 tags.

4.4.2 Data set

The data from last.fm are too numerous to create an exhaustive transition matrix. To test our method, users and artists were randomly selected using the same method as that was used to collect the data. A user was first randomly selected and all the user's friends were included. Then, all the artists this user listened to and all the user's tags were included. This process was until the target data size was reached (Case 1: 400 users, 1,500 artists, and 600 tags; Case 2: 1,200 users, 3,000 artists, and 800 tags). For each case, some data regarding how users listened to their artists were randomly deleted by 20%, 50%, 80%, 90% and 95%. Then the remaining data were used to create a transition matrix, as described in the Methodology section. For performance comparison, the authors also created the transition matrices, using the method described in [74]. This resulted in 20 data sets. Each data set received nine restart rations (α alphas) 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

4.4.3 Evaluation

We evaluated our method by creating recommendations for each user using RWR for each data set and the 4 set of transition matrices. Each user has a sorted recommendation list based on artists' respective probabilities.

The RWR equation was solved $\hat{p} = \alpha S\hat{p} + (1-\alpha)q$ as

$$(I - \alpha S)\hat{p} = (1-\alpha)q \tag{4.19}$$

$$\hat{p} = (I - \alpha S)^{-1}(1-\alpha)q \tag{4.20}$$

and the probabilities p and q were obtained at the convergence and restart stages.

Then, the artists who already exist in the user's history were removed and the authors compared the remaining artists to calculate precisions and MAPs.

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \tag{4.21}$$

where $Precision(R_{jk})$ = #artists who user j listened to in recommendation results from rank 1 to k, and m_j is number of recommendations returned for each query (one query for each user).

4.5 Results and Discussion

4.5.1 Comparing Effects of Restart Ratio

Table 4.1 shows the averaged MAP from all case effected by restart ratio α. The $\alpha = 0.4$ yields the highest MAP, but the change is insignificant.

4.5.2 Comparing Effects of Data Size

Fig. 4.3 shows the MAP for each test case group, with 20%, 50%, 80%, 90% and 95% of the listening data deleted. This result shows that with the proposed method, the recommendation efficiency (MAP) is not affected by the quantity of data used. The previous method show the changed of performance effected by size of data. However, the proposed transition matrix show at the very small size of the data (where 90% and 95% of the data were deleted), the proposed method show the better performance.

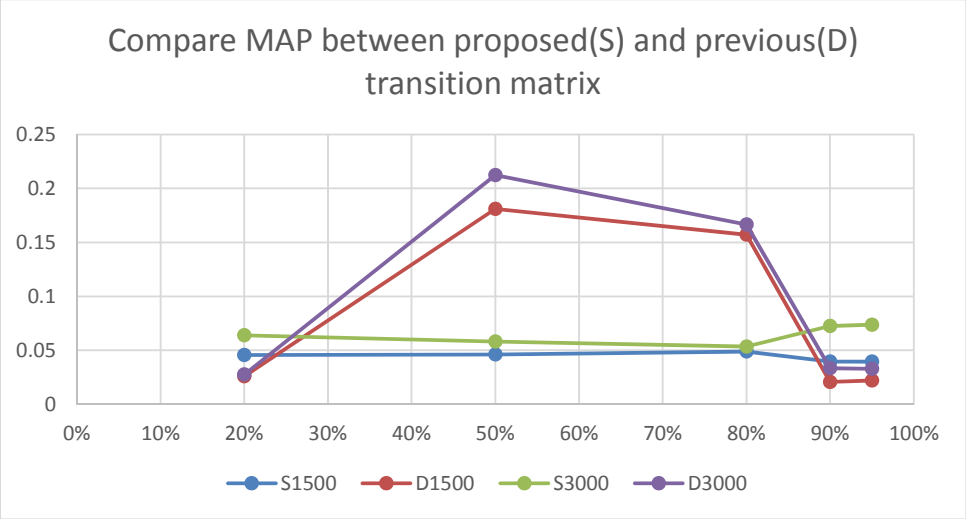

Figure 4.3: Comparing the MAP of each result with change in data size. The x-axis shows change in data size, when the listening data were deleted 20%, 50%, 80%, 90% and 95% and the y-axis shows the MAP. Four lines resulted from 2 types of transition matrix (proposed S for Social information based transition matrices and D for directed calculation) × two sets of data size (1,500 and 3,000 artists) which shown by type of transition matrix (S or D) and number of artists(1500 or 3000). The result shows that the proposed method performs better than previous methods for a small quantity of data.

Table 4.1: Long table caption.

Restart probability(α)	average MAP for all cases	SD for MAP
0.1	0.068917215	0.057843304
0.2	0.070890507	0.057721083
0.3	0.071567747	0.057584985
0.4	0.071676979	0.057623017
0.5	0.071568089	0.057660298
0.6	0.071474233	0.057663397
0.7	0.071395429	0.057678964
0.8	0.071233106	0.057557769
0.9	0.071123089	0.057525516

4.6 Conclusion

In the information overload era, recommendation systems are essential in helping users discover the information they are looking for. RWR has been used in recommendation systems to reduce the amount of information required to offer a recommendation. This study showed that RWR works well even in the case of limited data quantity. The effects of RWR' s restart ratio was examined and the restart ratio $\alpha = 0.4$ yielded the best recommendation accuracies. Social information based probabilities was used in creating transition matrices and tested with a data set collected from last.fm. The results showed that social information based probability transition matrices outperform the traditional transition matrices.

Our method relies on RWR, whose creation and inversion of transition matrices require substantial memory. Therefore, for an increasing number of users or items, straightforward RWR implementation could result in memory or speed limitations. There has been research on increasing the efficiency of RWR that can be applied here to work around this problem.

Major directions for future work include improving the efficiency of using RWR in recommendation systems, particularly with regard to calculation costs. One approach is to select suitable iterative RWR methods or to compute the inverse of the transition matrix. The matrix inversion approach requires computing the inverse, which can be expensive, but once computed, the inverse can be readily used to compute recommendations. The iterative approach allows more frequent up-

dates of the transition matrix. Another direction is to improve RWR scalability, because the sizes of social network graphs and items are constantly increasing. Yet another direction is to study how other social networking information (e.g., the number of messages sent or received between two friends, user viewing data, user following information, or friend suggestions) might improve recommendation accuracy.

Our proposed method is not limited to music domain. RWR-based recommendation systems have been shown to work in many domains ([41, 48, 140, 122, 79, 141, 131, 83, 49]), for which social information based conditional probability transition matrices can also be introduced.

Chapter 5

A BIPARTITE FITNESS MODEL FOR ONLINE MUSIC STREAMING SERVICES

This chapter proposes an evolution model and an analysis of the behavior of music consumers on online music streaming services. While previous studies have observed power-law degree distributions of usage in online music streaming services, the underlying behavior of users has not been well understood. Users and songs can be described using a bipartite network where an edge exists between a user node and a song node when the user has listened that song. The growth mechanism of bipartite networks has been used to understand the evolution of online bipartite networks [146]. Existing bipartite models are based on a preferential attachment mechanism [21] in which the probability that a user listens to a song is proportional to its current popularity. This mechanism does not allow for two types of real world phenomena. First, a newly released song with high quality sometimes quickly gains popularity. Second, the popularity of songs normally decreases as time goes by. Therefore, this paper proposes a new model that is more suitable for online music services by adding *fitness* and *aging* functions to the song nodes of the bipartite network proposed by Zhang et al. [146]. Theoretical analyses are performed for the degree distribution of songs. Empirical data from an online streaming service, *Last.fm*, are used to confirm the degree distribution of the object nodes. Simulation results show improvements from a previous model. Finally, to illustrate the application of the proposed model, a simplified royalty cost model for online music services is used to demonstrate how the changes in the proposed parameters can affect the costs for online music streaming providers. Managerial implications are also discussed.

5.1 Introduction

Online music streaming services have become an attractive means to consume music; however, despite the large and growing number of users, providers are still struggling to make a profit. Spotify, a music streaming service founded in Sweden in 2006, reported a net loss of €173.1M in 2015 [66]. To make effective business adjustments, providers need to understand how users use their services. The listening frequency of each song are longtail, and there are some studies which explain the longtail phenomenon. However the weighted bipartite graph which proposed by Zheng

and Chen, which is the closest representative model, cannot reproduce some properties of the real world.

Therefore, a large body of empirical research has explored the distribution of usage data [64, 121, 73] and found that the usage distribution follows a heavy-tailed distribution. The quantitative understanding of music listening behavior has been a subject of research interests. Hu and Han [64] studied the visiting log of a large Chinese online music service system for 105 days at the end of 2006. The results showed that distribution of inter-event time between two consecutive listening of music shows the fat tail feature. Koch and Soto [73] study the music listening behavior and confirmed that for each listener, the number of songs reproduced per artist follows a truncated power-law distribution. These shows that listening frequency per listener follows heavy-tailed distribution for both individual listeners and collectively. Our paper also found similar pattern for users of *Last.fm*. Some research paper view the relationship between a user and an object as a bipartite graph, or a two-mode network. If a user u uses or access object o, there is an edge between u and o Shang et al. [121]. While empirical research allows us to understand the current state of the usage of services, it does not allow us to understand the underlying evolution of the usage. Accurate models of human activities are crucial for better resource allocation and pricing plans for service companies, and to improve inventory and service allocation in both online and retail stores [20]. Research into the evolution of complex systems helps us understand the phenomena observed in real-world complex systems. However, the existing models are not suitable to model the network for online music streaming services. Therefore, this paper proposes a new model to solve the problem.

A seminal work on models of human activities by Barabási and Albert [21] proposed *preferential attachment* as an evolution mechanism for a unipartite graph, where nodes are created and attached to previous nodes with a probability proportional to the degree of the previous nodes. Such evolution mechanism results in networks with power-law distributions, also called long tailed distributions. Preferential attachment mechanism has been shown to describe real-world phenomena, such as the network of the World Wide Web [33, 75, 78] and the research citation network [33]. In a subsequent study, the Bianconi - Barabási model [26] added fitness to the model of Barabási and Albert [21]; when the fitness of a node is higher, the chances that it will be connected to other nodes increases, even if the node is created later. Lefortier et al. [78] and Ostroumova Prokhorenkova and Samosvat [96] added lifetime attractiveness to the Bianconi - Barabási model [26]. This makes the chances of being connected decrease as time passes. However, the aforementioned models

utilized unipartite graphs, which do not fit well to music listening.

A bipartite graph is a natural way to model online streaming services, where each user is represented by a user node and each song is represented by an object node. If a user u uses or accesses an object o, there is an edge between u and o. A weighted bipartite network based on the strength preferential attachment of Zheng and Chen [147] enables modeling of the usage of a scale-free network with a bipartite graph. However, the degree distribution for the user nodes is uniform. The Mandelbrot law distribution for the user nodes' degree in unweighted bipartite graph proposed by Zhang et al. [146] solves the problem of inappropriate degree distribution of the user nodes. Zhou et al. [148] add the aging parameter proposed in Lefortier et al. [78] to unweighted bipartite graph model and make popularity prediction model more accurate. However, they did not consider the fitness to market of each song.

This chapter proposes a new model, which combines the bipartite models [146, 147] and adds the fitness from the unipartite model [26] to the object nodes, as well as the recency property [96]. After the model is described, a theoretical analysis is performed on the degree distribution of the object nodes. Empirical data from *Last.fm* are collected, and the degree distribution of the objects is found to follow a power-law distribution. Using these empirical results, the fitness distribution of the objects is analyzed and new parameters are proposed. The new parameters are examined using a theoretical analysis to show how the changes in the parameters will affect the shapes of distributions of the objects.

Finally, we propose an example usage for the model in managerial decisions by constructing a simple cost model for online music streaming services and analyzing the effects of each parameter on the cost. While some studies on online music streaming services have proposed approaches for services and pricing strategies such as Adrian Maftei et al. [5] and Paul Thomes [100], none has analyzed costs, which directly affects profits. We focus on the royalty fee which is one of the various cost in the music streaming services.

The main contributions of this chapter are threefold. First, a weighted bipartite model for online music streaming services is proposed with a fitness function and a recency property. To the best of our knowledge, we are the first to incorporate a fitness function into a bipartite model. Second, the degree distributions are theoretically analyzed. Third, empirical data are used to demonstrate the usefulness of the model in making managerial decisions.

5.2 *Limitation of Previous Bipartite Models in Explaining Real-world Music Listening Behaviors*

The bipartite network model for modeling the content consumption of users proposed by Zhang et al. [146] has several behaviors that cannot reflect real world phenomenon. Because the model uses the preferential attachment model proposed by Barabási and Albert [21], the model will follow the rich-get-richer phenomena. The rich-get-richer phenomena makes older nodes more likely to get linked to than newer nodes. As in Barabási and Albert [21], the function of expected degree of node i, which was created at time t_i, at time t was shown as $k_i(t) = m(\frac{t}{t_i})^{\frac{1}{2}}$ where m is number of connected links for new node, which mean that the expected ratio of degree between the older node x, created at time t_x, and newer node y, created at time t_y is $(\frac{t_y}{t_x})^{1/2}$. As $t_y > t_x$, we have $(\frac{t_y}{t_x})^{1/2} > 1$, i.e. older nodes have higher degrees than newer ones.

However, real-world music listening behaviors establish a different phenomena. To illustrate this, we use the Echo Nest Taste profile subset[1] from the Million Song Dataset (MSD) [23]. The dataset contains the play counts of 384,546 unique MSD songs by 1,019,318 unique users, collected from 2005 to 2011. More specifically, the data is a set of tuples (u,s,n), in which each tuple indicates that a user u listened to a song s for n times between 2005 and 2011. To obtain the released year of each song, the song data were mapped with a list of the 515,576 songs with released year information [4] from the same dataset. We can match 60,715 songs, ranging from songs that were released in 1920's to 2011. Songs that were released from 2005 to 2010 are used here to illustrate the differences between usage of older songs and newer songs. Here, we note that newer songs have fewer number of years that are available to be listened to. Songs that were released in 2005 had six years to be listened to, while songs that were released in 2010 had one year to be listened to.

This dataset shows that newer songs are listened to more than older ones, as illustrated by the statistics of play counts shown in Table 5.1 and the distributions shown in Figure 5.1. In Table 5.1, to eliminate the effect of extreme values, the average, standard derivation, the maximum and minimum of the natural logarithm of the play counts are shown. Figure 5.1 shows the distributions of play counts of songs in each released year using a violin plot. Even though newer songs had fewer number of years that were available to be listened to, the play counts of newer songs were higher than those of older ones. Therefore, this set of data confirms the difference from rich-get-richer

[1]The Echo Nest Taste profile subset, the official user data collection for the Million Song Dataset, available at: http://labrosa.ee.columbia.edu/millionsong/tasteprofile.

phenomena, in which older songs should be listened more than newer ones. Therefore, we propose a new model to solve this problem by adding fitness and aging function to each object following the work of Bianconi and Barabási [25], Lefortier et al. [78] as described in Section 5.3.

Table 5.1: Natural logarithm of the play count for each song in each released year, shown the departure from rich-get-richer phenomena that the older songs have higher play counts.

Song's released year	Number of songs	Natural logarithm of play count per song			
		Mean	Min	Max	St.Dev.
2005	2,024	1.634	0	4.514	0.891
2006	6,336	1.648	0	5.285	0.873
2007	4,786	1.688	0	4.801	0.900
2008	3,711	1.665	0	4.864	0.884
2009	4,687	1.743	0	5.295	0.878
2010	1,865	1.784	0	5.094	0.898

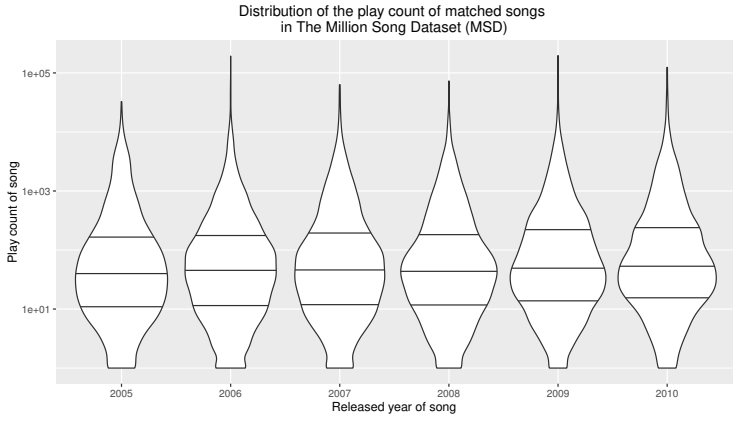

Figure 5.1: A violin plot shows the distribution of play counts of songs grouped by songs' released year from real data in the Million Song Dataset. The y-axis shows the play counts and the width of violin plot shows the density of songs that have the play count. Three lines in each violin plot show three quartiles of the distribution. The lines get higher in each year showing the newer songs have higher play counts than the older songs.

5.3 Model

To analyze the factors that affect the degree distribution of the content consumption, we constructed a model to explain the relationship between users and objects using a weighted undirected bipartite graph. The model proposed in this study is inspired by four related papers [146, 78, 26, 147]. Our model improved on the bipartite models of Zhang et al. [146] and Zheng and Chen [147], which did not model the intrinsic quality of the nodes or the order in which the nodes occur, by adding the *fitness* parameter that was proposed by Bianconi and Barabási [26] and the *lifetime attractiveness* parameter that was proposed by Lefortier et al. [78]. With the added parameters, our model can explain the intrinsic quality of songs and their reduced popularity after they have been released for some time. Users and songs can be modeled as a weighted bipartite graph where the set of users is the top nodes, the set of songs is the bottom nodes, and the weight of the edge is the number of times the user listened to that song.

The proposed network can be described as a weighted bipartite graph $G = \{U, O, E\}$ where U and O are two disjoint sets of nodes, where U is the set of users and O is the set of objects (e.g. songs) and $E \subset U \times O \times \mathbb{N}$ is the set of relationships between a user and an object with a natural number as a weight for each relationship. User u_i denotes a user node that is created at time step i. Object o_j denotes an object node created at time step j. User nodes and object nodes have properties and functions that are defined as follows:

- Strength, $s_u(t, i)$, of a user node u_i at time t is the sum of the weight of all the edges that connected to node u_i at time t. Therefore, we have $s_u(t, i) = \sum_{(u_i, o_j, w_{i,j}(t)) \in E} w_{i,j}(t)$.

- Strength, $s_o(t, j)$, of an object node o_j at time t is the sum of the weight of all the edges that connected to node o_j at time t. Therefore, we have $s_o(t, j) = \sum_{(u_i, o_j, w_{i,j}(t)) \in E} w_{i,j}(t)$.

- Fitness, $f_o(j)$, of object o_j is a random variable that indicates the object's quality.

- Age function, $a(t, j)$, at time t of an object node o_j is defined as $a(t, j) = e^{\frac{j-t}{\tau}}$, where τ is the mean lifetime of all the objects.

- Attractiveness function, attr(t, j), at time t of an object o_j is defined as attr$(t, j) = s_o(t, j) f_o(j) a(t, t_o)$.

The model starts from an initial state that $G = \{U = \{u_0\}, O = \{o_0\}, E = \{(u_0, o_0, 1)\}\}$. On each time step, t, do the following independently.

- Add a new user u_t and connect to m existing object nodes $o_j \in O$ according to the attractiveness probability $\frac{\text{attr}(t,j)}{\sum_{k<t}\text{attr}(t,k)}$ with weight = 1.

- Add a new object o_t with attractiveness defined above and connect to n existing user nodes $u_i \in U$ using the preferential probability $\frac{s_u(t,i)}{\sum_{k<t} s_u(t,k)}$ and weight = 1.

- Evolve edges with preferential attachment. If an edge between two nodes exists, increase the weight of the edge by one; otherwise add a new edge with weight = 1.

  - **Randomly**: Randomly select c users $u_i \in U$ by an equal probability of $\frac{1}{t}$, and connect to c objects $o_j \in O$ selected by an attractiveness probability of $\frac{\text{attr}(t,j)}{\sum_{k<t}\text{attr}(t,k)}$.
  - **Attractiveness**: Randomly select b users $u_i \in U$ by preferential probability $\frac{s_u(t,i)}{\sum_{k<t} s_u(t,k)}$ and connect to b objects $o_j \in O$ according to their attractiveness probability $\frac{\text{attr}(t,j)}{\sum_{k<t}\text{attr}(t,k)}$.

5.4 Empirical Analysis

This section studies the current usage distribution of an online streaming service and shows that play count distribution (object's strength distribution) follows a power-law distribution.

5.4.1 Description of the Last.fm Data

Last.fm is an online music streaming service, which allows developers to obtain music usage data from their service via a programmable API. We obtained usage data of 265,975 users for three one-month periods; August 2013, June 2014, and December 2014. The data specify which songs each user listened to. Out of all users, 125,822 users had listened to at least one song in the periods of data collection. The number of songs that were listened to was 13,040,032.

5.4.2 Object Strength Distribution

This section investigates Last.fm's object strength distribution. The strength of an object for Last.fm is the sum of play count of a song (i.e. object) by all users in the data set. Therefore, object strength indicates the popularity of the song. To investigate which distribution matches the data set, we use a visual comparison and a statistical method proposed by Clauset et al. [33].

For visual comparison, several distributions were tested, as shown in Figure 5.2, calculated by maximum likelihood estimation (mle) using R [106] package 'fitdistrplus' [39], where the distribution of the data seems to fit to power-law distribution. This is consistent with the results of Shang et al. [121], where the distribution of the degree is not uniform but follows a power-law distribution.

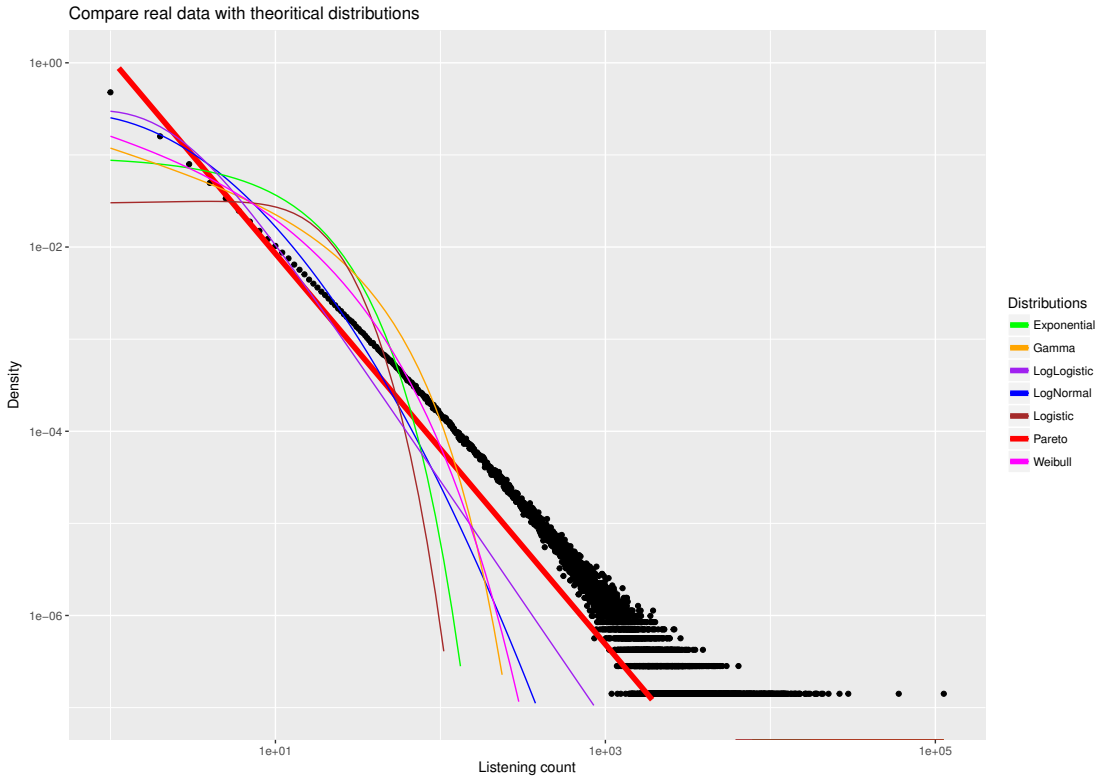

Figure 5.2: Comparison between the empirical data and theoretical data shown using density on a log-log scale. Empirical data is the number of times each song is listened to by users of Last.fm (play count per song), shown using the black dot, where x axis show the listening count and y axis show the density of songs that have such listening count. Theoretical data are different distributions fitted with the data shown using colored lines as listed in the legend.

The distributions' parameters estimations were compared by visualized method as shown in graph on Figure 5.2 and by using the information criterion both Akaike's Information Criterion(AIC) and Bayesian Information Criterion(BIC) as shown in table 5.2 ordered by information criterion value in ascending. The AIC and BIC of Pareto distribution is less than the other distributions which mean that the real data is closer to Pareto distribution when comparing with the other dis-

tributions.

Table 5.2: Compare the goodness of fit of the listening data from last.fm by the information criterion

Distributions	Akaike's Information Criterion	Bayesian Information Criterion
Pareto	**25,175,313**	**25,175,341**
LogLogistic	34,379,031	34,379,031
LogNormal	35,152,088	35,152,115
Weibull	39,604,048	39,604,076
Gamma	42,875,218	42,875,245
Logistic	61,411,392	61,411,419
Exponential	∞	∞

However, the graph in Figure 5.2 and Information Criterion in table 5.2 do not allow us to conclude that the distribution is a power law. Therefore, we utilize the method proposed by Clauset et al. [33] to more rigorously confirm that the distribution is a power law. The power law distribution has a shape parameter α and a scale parameter x_m and can be written as

$$f_{s_o(t,j)}(x; x_{\mathrm{m}}, \alpha) = \begin{cases} \dfrac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}. \end{cases} \tag{5.1}$$

where $s_o(t, j)$ is random variable of strength of object j at time t.

The method proposed by Clauset et al. [33] is a three-step process to determine whether a set of data is a power law. The first step is to estimate α and x_m to be used in the second and third steps. The second step calculates the goodness-of-fit between the data and the power law to test hypothesis and if the resulting p-value is greater than 0.1, the power law is a plausible hypothesis for the data. The third step compares the power law with an alternative distribution via a likelihood ratio test. If the calculated likelihood ratio is significantly different from zero, then its sign indicates whether the alternative is favored over the power-law model or not.

For the second step, Clauset et al. [33] suggested that the null hypothesis, *"the data is generated by a power-law distribution"*, can be tested using a goodness-of-fit (GoF), which generates a p-value that can be used to quantify the plausibility of the hypothesis. The null hypothesis will be rejected when the p-value is less than 0.1. To calculate goodness-of-fit test, Clauset et al. [33] estimated parameters and calculated the difference between the real data and the estimated model

with Kolmogorov-Smirnov (KS) statistic, then generated some datasets from the estimated model and calculated KS statistic in the same way as real data. The p-value is calculated from proportion of the dataset that have larger distance from the model than the real data. Clauset et al. [33] recommended to use 2,500 datasets to get reliability of p-value at 0.01, and recommended the default value of number datasets as 1,000 datasets; however, the process takes a long time for the large number of data in our dataset (using R [106] package 'poweRlaw' [114] on a personal computer with 2.5 GHz Intel Core i5 CPU and 8 GB 1600 MHz DDR3 RAM), we dropped the number of datasets to 100 datasets but this still have reliability of p-value at 0.05 which can be used to reject the null hypothesis. We get the GoF from KS statistic as 0.0134, and p-value = 0.69, which indicates that the data may be generated by a power-law distribution.

For the third step, Clauset et al. [33] suggested to compare the distance with other related distributions to confirm if the other distribution is closer to real data or not, by setting the null hypothesis as *"both distributions are equally far from the true distribution"* and testing with the same process as previous step. To confirm the longtail property of the data, the exponential distribution was compared by using the third step in the method. The log likelihood ratios = -10.470 and p-value = 1.181×10^{-25}, which means that Pareto distribution is closer than exponential distribution and this show the longtail property of the data. From comparing with log normal distribution, which sometime be confused with the power-law distribution, we get log likelihood ratios = -0.757 and p-value = 0.449. These made us cannot reject null hypothesis, i.e. the log normal distribution is not closer to the real data than power-law distribution.

Therefore, the object strength of our Last.fm data follow a power-law distribution. This result aligns with the previous studies by Lefortier et al. [78] and Ostroumova Prokhorenkova and Samosvat [96] that showed the power-law distribution of the song's play counts.

5.5 Theoretical Analysis

This section analyzes the degree distribution of the object nodes. The user nodes in our model are exactly as proposed and analyzed by Zhang et al. [146], who found the distribution of the user strength to follow a shifted power-law distribution. For the strength distribution of the object nodes, the method described in [25] and [78] is applied to analyze the effects of the fitness and aging functions. The proof follows the method in Lefortier et al. [78].

In the mean-field approximation, the dynamics of the attractiveness for an object o_j is

$$\frac{\partial s_o(t,j)}{\partial t} = (m+b+c)\frac{\text{attr}(t,j)}{\sum_{k<t}\text{attr}(t,k)} \tag{5.2}$$

At time t, there are $t+1$ objects and $t+1$ users with the sum of the weight of the users equal to the sum of the weight of the objects, which is $(m+n+b+c)t+1$.

Let $N=(m+b+c)$ and $W(t)$ be the expected value of sum of attractiveness at time t, then:

$$W(t) = \mathbf{E}[\sum_{k<t}\text{attr}(t,k)] = \mathbf{E}[\sum_{k<t} s_o(t,k)f_o(k)a(t,k)] \tag{5.3}$$

Then we have the following differential equation:

$$\frac{\partial s_o(t,j)}{\partial t} = N\frac{\text{attr}(t,j)}{W(t)} = N\frac{s_o(t,j)f_o(j)e^{\frac{j-t}{\tau}}}{W(t)} \tag{5.4}$$

$W(t)$ tends to some positive constant (see A.1 for a proof that $W : \lim_{t\to\infty} W(t) = W$). Therefore, with the initial condition $s_o(t,j) = n$ at $t=j$, we have the following solution of Eq. (5.2):

$$s_o(t,j) = ne^{\frac{\tau N}{W}f_o(j)(1-e^{\frac{j-t}{\tau}})} \tag{5.5}$$

This gives $\lim_{t\to\infty} s_o(t,j) = ne^{\frac{\tau N}{W}f_o(j)}$.

Because $ne^{\frac{\tau N}{W}f_o(j)}$ can be considered to be a one-to-one function to transform a random variable of fitness $f_o(j)$ to a random variable of object strength, the cumulative distribution of $ne^{\frac{\tau N}{W}f_o(j)}$ can be computed as

$$F_{s_o(t,j)}(y) = P\{s_o(t,j)\le y\} = P\{ne^{\frac{\tau N}{W}f_o(j)}\le y\} = P\{f_o(j)\le\frac{W\ln(\frac{y}{n})}{\tau N}\} = F_{f_o(j)}(\frac{W\ln(\frac{y}{n})}{\tau N}) \tag{5.6}$$

Next, the chain rule is used to compute the density of $s_o(t,j)$ such that

$$f_{s_o(t,j)}(y) = F'_{s_o(t,j)}(y) = \frac{W}{\tau N|y|}f_{f_o}(\frac{W\ln(\frac{y}{n})}{\tau N}) \tag{5.7}$$

Because $\lim_{t\to\infty} s_o(t,j) = ne^{\frac{\tau N f_o(j)}{W}}$, we obtain the relationship between the fitness distribution of objects to the strength distribution of objects shown in Table 5.3.

5.5.1 Analysis of the Fitness Distribution and Parameters

The results from fitting the play counts of the songs with a power law in subsection 5.4.2, allow us to use an exponential distribution for the fitness distribution of the objects. The probability density function(PDF) of the exponential distribution can be written as:

$$f_{f_o(j)}(x;\lambda) = \lambda e^{-\lambda x} \tag{5.8}$$

Table 5.3: Relationship between the fitness distribution of objects and the strength distributions of objects

Object's Fitness Distribution ($f_o(j)$)	Object's Strength Distribution ($s_o(t,j)$)
Exponential distribution	Power-law distribution
Normal distribution	Log normal distribution
Exponential gamma distribution	Gamma distribution

where $f_o(j)$ is random variable of fitness of object j.

The fitness of an object is the probability that the object is attractive. However, because it is a probability, some objects that are accessed more by users may have a lower fitness than some objects that are accessed less.

The strength distribution of object nodes can be rewritten as follows. The strength distribution of the object node is $f_{s_o(i)}(y) = \frac{W}{N\tau|y|} f_{f_o}(\frac{W\ln(\frac{y}{n})}{N\tau})$ and the fitness distribution of the objects follows exponential distribution $f_o(x) = f(x;\lambda) = \lambda e^{-\lambda x}$. Therefore, the distribution of the strength of the object nodes can be written as

$$f_{s_o(i)}(y) = \frac{W}{N\tau|y|} f_{f_o}(\frac{W\ln(\frac{y}{n})}{N\tau}) \tag{5.9}$$

$$= \frac{W}{N\tau|y|} \lambda e^{-\lambda(\frac{W\ln(\frac{y}{n})}{N\tau})} \tag{5.10}$$

$$= \frac{\lambda W}{nN\tau} (\frac{y}{n})^{-(\frac{\lambda W}{N\tau}+1)} \tag{5.11}$$

$$= \frac{\frac{\lambda W}{N\tau} n^{\frac{\lambda W}{N\tau}}}{y^{(\frac{\lambda W}{N\tau}+1)}} \tag{5.12}$$

This corresponds to a power-law distribution with a scale parameter n and a shape parameter $\frac{\lambda W}{N\tau}$, which is equal to $\frac{\lambda W}{(m+b+c)\tau}$. However, because the expected value of exponential distribution with a rate parameter λ is λ^{-1}, λW will not be affected by a change in λ.

From the empirical data, the song's play count distribution follows the power-law distribution; therefore, the fitness random variable should follows the exponential distribution $f(x;\lambda) = \lambda e^{-\lambda x}$. The resulting fitness distributions of objects and the other proposed model details create five parameters for the model. These parameters can be categorized into two groups; exogenous variables and endogenous variables. Two variables, the fitness distribution's parameter λ and the usage rate of new user m, are endogenous variables. There are three exogenous variables: the mean life time of objects τ, the object usage rate for each object created in each interval n, and the number of

interactions between the previously created users and the previously created objects in each interval $(b + c)$. These three variables can be adjusted using the content discovery functions of the system, such as the recommendation systems and the ranking systems. The values of n and $(b+c)$ can be influenced by the appropriate recommendation systems. If the recommendation system recommends recently created content, the usage rates of the new content, n, will increase while the usage rates of previously created content, $(b+c)$, will decrease. When n increases and $(b+c)$ decreases, the mean life-time of objects τ in the system will also decrease.

5.6 Comparison with a Previous Model

The proposed model was compared to the model proposed by Zhang et al. [146] to show the improvement in using the model to describe the real-world music listening behaviors. As mentioned in Section 5.2, we use the Million Song Dataset provided by The Echo Nest to examine the real-world usage distribution.

The Simulation was used to create a user-object network by setting parameters to $m = n = 100$ and $b = c = 50$ for both models and $\tau = \lambda = 1000$ in the proposed model run with a total of 10,000 time steps, Then, we compared the behaviors of the two models.

Both models show a power-law object strength distribution which is shown as a straight line when plotting the relationship between the rank and play count, as in Figure 5.3.

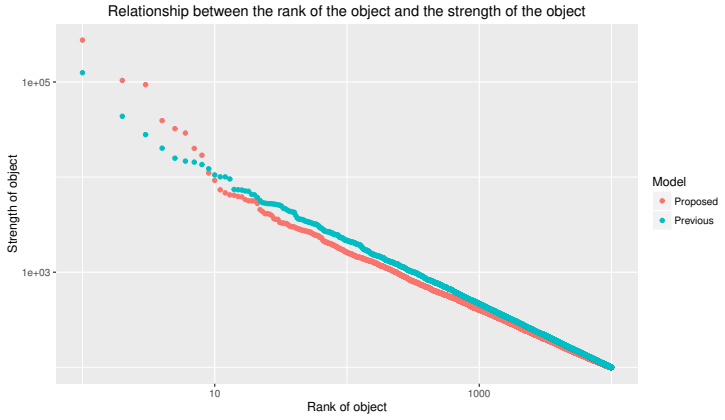

Figure 5.3: Relationship between the rank of the object and the strength of the object in the proposed and previous models. The parameter settings are $m = n = 100$ and $b = c = 50$ for both models and $\tau = \lambda = 1000$ in the proposed model and run with a total of 10,000 time steps. The x-axis shows the rank of the object node, and the y-axis shows the strength of the object node. The graph shows straight line that is one property of power-law distribution.

The density of object strength or playcount in real world show straight line in log-log scale, which have a similar shape to the real data as shown in Figure 5.4 for simulation from both models, which both models result in highly similar density, compared to Figure 5.2 and Figure 5.5, which plots the real data from last.fm and MSD.

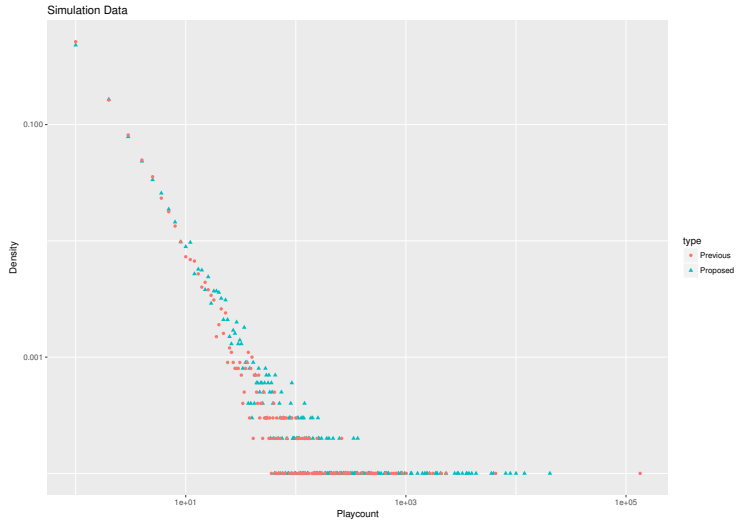

Figure 5.4: Density of the object nodes in the proposed and previous model in log-log scale. The parameter settings are $m = n = 100$ and $b = c = 50$ for both models and $\tau = \lambda = 1000$ in the proposed model and run with a total of 10,000 time steps. The x-axis shows the strength of object nodes, and the y-axis shows the density of the object nodes which have that strength.

To examine the rich-get-richer phenomena where nodes that are created before will have higher degrees than nodes that are created later, we made a scatter plot between the created time of each object and the strength of that node. The previous model shows that earlier created objects have higher strength than later created objects, as shown in the Figure 5.7; however, this did not occur in our proposed model as shown in Figure 5.6.

The violin plot was created to compare the simulation results with the real data. The objects were separated into five groups according to their created time as shown in Table 5.4, and the results are shown in Figure 5.8 for the previous model and Figure 5.9 for the proposed model.

The result of the old model simulation in Figure 5.8 shows the rich-get-richer phenomena as objects in the first group have relatively more strength than those in the second and third groups and so on. However, the proposed model's result in Figure 5.9 has a similar strength distribution in all groups, which closer to the real data shown in Figure 5.1.

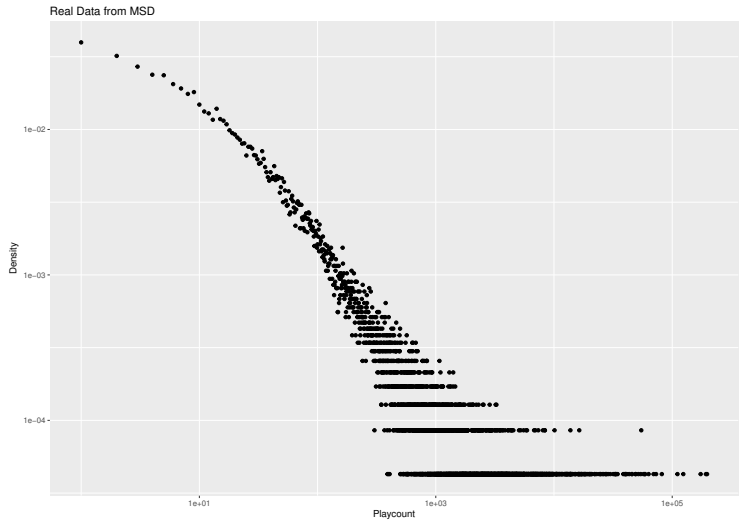

Figure 5.5: Density of playcount of the songs in the real data from MSD in log-log scale. The x-axis shows the play counts of songs, and the y-axis shows the density of the songs that got the playcount.

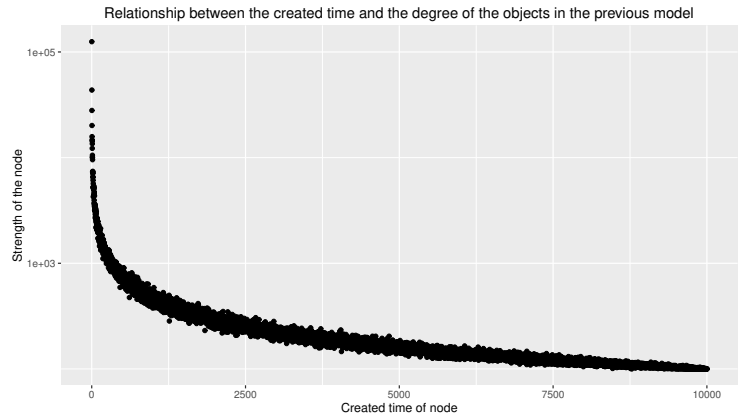

Figure 5.6: Relationship between the created time and the strength of the objects in the previous model. The x-axis shows the created time of each node and the y-axis shows the strength of the node. This shows the rich-get-richer phenomenon that the older nodes have more chance to get link than the newer nodes.

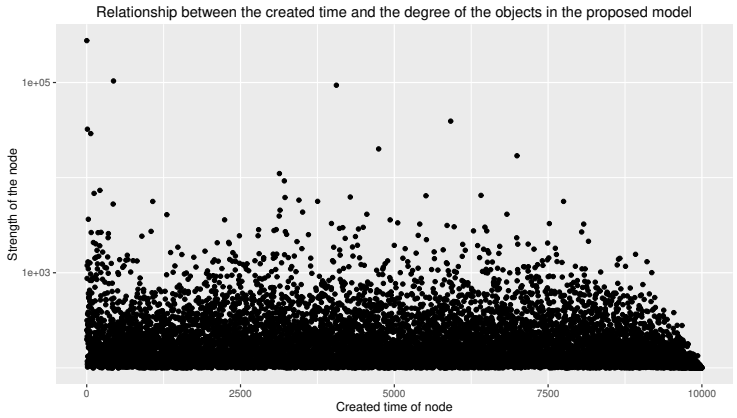

Figure 5.7: Relationship between the created time and the strength of the objects in the proposed model. The x-axis shows the created time of each node and the y-axis shows the strength of the node. The rich-get-richer phenomenon cannot be found in the result.

Table 5.4: Relationship between the created time and the group of objects.

Group	*Created time*	
	from	to
1	1	2000
2	2001	4000
3	4001	6000
4	6001	8000
5	8001	10000

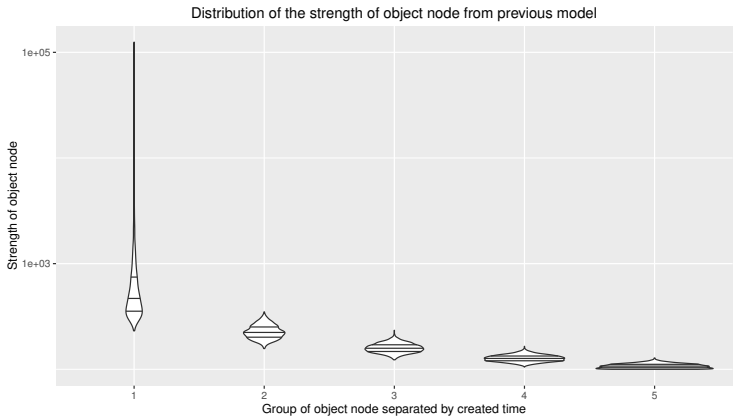

Figure 5.8: A violin plot shows the distribution of strength of the object nodes grouped by created time from previous model's simulation. The y-axis shows strength of the object node and the width of violin plot shows the density of object nodes that have the strength. Three lines in each violin plot show three quartiles of the distribution. The lines in the first group have relatively higher value than those in the second and third groups and so on.

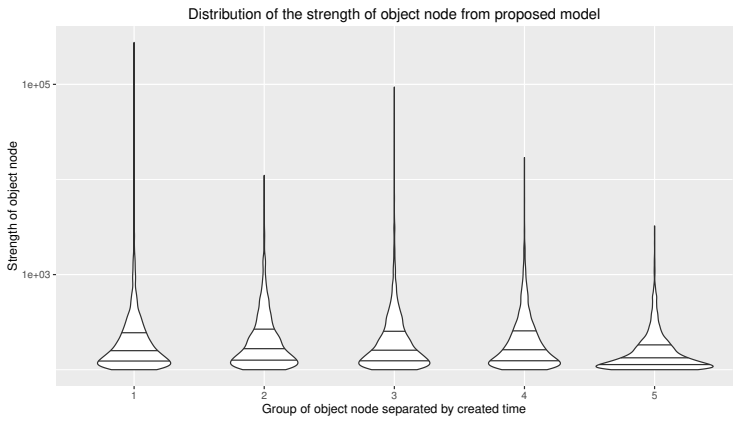

Figure 5.9: A violin plot shows the distribution of strength of the object nodes grouped by created time from proposed model's simulation. The y-axis shows strength of the object node and the width of violin plot shows the density of object nodes that have the strength. Three lines in each violin plot show three quartiles of the distribution. The lines in the first to the fourth groups has almost same value.

5.7 Application to the Cost Structure of Online Music Services

This section explores a sample application of the proposed weighted bipartite graph model with fitness and aging functions to be used for businesses. We constructed the simplified cost model for royalty fee and use the proposed model to show the effect of each parameter to the cost.

Some studies on online music streaming services have studied the factors that affect profits and suggests service and pricing strategies, but none has addressed the cost factor, which is a factor that directly affects profits. Adrian Maftei et al. [5] proposed that the critical success factors of online music streaming premium services are free music streaming, the ability to purchase music, the lack of advertisements and the satisfaction of supporting one's favorite artists. Paul Thomes [100] proposed an economic analysis on monopoly online music streaming premium services that showed the effect of piracy and advertising which causes a divergence in the profit equilibrium between users and the provider. These studies show the services and pricing strategies but still did not concern about cost which affect the profit of the service. We focus on the royalty fee which is one of the large expenses in the music streaming services. This royalty fee is calculated per play count per music and the price varies according to the popularity of that songs. Celma and Cano [30] showed that recommendations from collaborative filtering tend to steepen the slope of the distribution but content-based and human-based recommendations will drive usage to the tail part of the distribution, which indicates that there are methods that can change the distribution of the play count and also royalty fee of the service.

5.7.1 A Simplified Cost Model

To propose a way to reduce the cost of online streaming services, this section models how the changes in usage behavior affect the changes in the royalty cost, which is the biggest expense for this type of service [66]. The cost structure of online streaming services is simplified to create a mathematical model by making two assumptions that follow Elberse [45] findings about the royalty cost of songs in online streaming services. First, hit songs are more expensive than songs that are not famous. Second, a royalty fee is collected per play count per song. We simplified the cost model into two groups of songs: one group for hit songs and another group for non-hit songs. The number of hit songs is assumed to be $(\beta \times 100)\%$ (where $0 \leq \beta \leq 1$) of the total number of songs, and the number of non-hit songs is $((1-\beta) \times 100)\%$. The average royalty fee ratio of non-hit songs to hit songs is assumed to be γ (where $0 \leq \gamma \leq 1$)

If the play count distribution of each song follows a probability distribution function $f(x)$, where x is the play count for each song then we need to find a play count c that separates songs with play counts more than c to $\beta \times 100\%$ of all songs.

$$1 - \beta = \int_0^c f(x)dx \tag{5.13}$$

$$c = F^{-1}(1 - \beta) \tag{5.14}$$

Here, $F(x)$ is the CDF of $f(x)$. We can calculate the cost of the royalty fee (Cost$(f(x))$) to be

$$\text{Cost}(f(x)) = C(\gamma \int_0^{F^{-1}(1-\beta)} xf(x)dx + \int_{F^{-1}(1-\beta)}^{\infty} xf(x)dx) \tag{5.15}$$

$$= C(\mathbf{E}[f(x)] - (1 - \gamma) \int_0^{F^{-1}(1-\beta)} xf(x)dx) \tag{5.16}$$

where C is the cost per play count of each hit songs.

Because we know that the play count distribution follows a power-law distribution or $f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$, from the CDF of the power-law distribution, $F(x) = 1 - (\frac{x_m}{x})^\alpha$, $F^{-1}(x) = \frac{x_m}{(1-x)^{\frac{1}{\alpha}}}$ and $F^{-1}(1 - \beta) = \frac{x_m}{(1-(1-\beta))^{\frac{1}{\alpha}}} = \frac{x_m}{\beta^{\frac{1}{\alpha}}}$.

The cost can be calculated as

$$\text{Cost}(f(x)) = C(\mathbf{E}[f(x)] - (1 - \gamma) \int_0^{F^{-1}(1-\beta)} xf(x)dx) \tag{5.17}$$

$$= C\frac{\alpha x_m^\alpha}{1-\alpha}(\lim_{x\to\infty} x^{1-\alpha} - {x_m}^{1-\alpha} - (1-\gamma)((\frac{x_m}{\beta^{\frac{1}{\alpha}}})^{1-\alpha} - x_m^{1-\alpha})) \tag{5.18}$$

$$= C\frac{\alpha x_m^\alpha}{1-\alpha} \lim_{x\to\infty} x^{1-\alpha} + C\frac{\alpha x_m}{\alpha - 1}(1 + (1-\gamma)(\beta^{\frac{\alpha-1}{\alpha}} - 1)) \tag{5.19}$$

From Eq. 5.19, we get $\alpha > 1$ and

$$\text{Cost}(f(x)) = C\frac{\alpha x_m}{\alpha - 1}(1 + (1-\gamma)(\beta^{\frac{\alpha-1}{\alpha}} - 1)) \tag{5.20}$$

We can calculate the change in Cost$(f(x))$ related to the change in α by partially differentiating Cost$(f(x))$ by α such that

$$\frac{\partial}{\partial\alpha}\text{Cost}(f(x)) = C\frac{x_m}{(\alpha-1)^2}(\beta^{\frac{\alpha-1}{\alpha}}(1-\gamma)(\frac{\alpha-1}{\alpha}(\log\beta) - 1) - \gamma) \tag{5.21}$$

Because $1-\gamma \geq 0$ and $\frac{\alpha-1}{\alpha} > 0$ and $\log \beta \leq 0$, we get $\frac{\partial}{\partial \alpha}\text{Cost}(f(x)) \leq 0$ for all α, which means that α increases, the cost decreases.

5.7.2 Changes that Online Music Providers Can Adopt

Using the variables in Section 5.5.1, this section explores how changes in the exogenous variables will affect the shape parameter of the power-law distribution and the cost to online music streaming providers.

- *Recently created object usage rate (n)* : Because the new object usage rate n is a scale parameter, it will not affect the shape parameter of the power-law distribution. This means that the royalty fee will not be affected by the recently created object usage rate.

- *Previously created objects usage rate ($b+c$)* : Because the strength distribution follows the power-law distribution with a shape parameter $\frac{\lambda W}{(m+b+c)\tau}$, we can conclude that when the previously created object usage rate $b+c$ increases, the shape parameter of the power law will decrease. This will increase the cost of the royalty fee to the music online streaming service.

- *Mean life time parameter (τ)* : Because the strength distribution follows the power-law distribution with a shape parameter $\frac{\lambda W}{(m+b+c)\tau}$, we can conclude that when the mean life time τ increases, the shape parameter of the power law will decrease. This will also increase the royalty fee cost of the streaming service.

As discussed above, the changes in the two exogenous variables (the previously created object usage rate and the mean life-time parameter) will affect the shape of the power-law distribution. The results from Section 5.7.1 also suggest how the changes in the shape of the power law will affect the cost of online music streaming with respect to royalty fees. Therefore, except for n, which is a scale parameter that does not affect the shape of the power-law distribution, the following insights can be drawn:

- With respect to the previously created object usage rate ($b+c$), the recommendation of old songs should be avoided to reduce the usage of old songs.

- With respect to the mean life time parameter (τ), songs should be displayed in lists ordered from new songs to old songs. List ordering has been known to influence users' responses that the beginning of the list often received more responses [88, 72]. Therefore, ordering new songs at the beginning of the list will increase the access to new songs and decrease the access to old songs, which in turn will decrease the mean life-time.

5.8 Conclusion and Future Work

This chapter proposed a new bipartite evolution model for online music streaming usage to improve on the limitations of previous bipartite models that established rich-get-richer phenomena, which does not exist in music listening behaviors. The improvement was accomplished by adding fitness and aging functions to the object nodes. After theoretical analysis on object strength distribution, we also confirmed the distribution of the play counts of songs from Last.fm dataset is a power law.

Additionally, to find a way to reduce the cost of the royalty fee for content, which is based on the number of times the content has been accessed, and the price differences between popular and non-popular contents, we proposed a simplified model to explain the usage of the content. Based on the adjustments of model parameters, we suggested that royalty fee for the entire system can be reduced by recommendations of newer songs.

There are many directions for future work. First, due to the limitation of the datasets we have, we were not able to perform the statistical hypothesis testing and parameter estimation of recency property or aging phenomena in online music streaming service. To do such the testing, the dataset needs to contain the released time and the number of times each song is listened to for a few years following the released year. The MSD dataset contained only aggregated number of times the songs were listened to from 2005 to 2011. Last.fm did not contain the released year information. The dataset similar to the one used in Park and Kahng [98], but in larger scale, can be used to test the aging phenomena. The hypothesis testing and parameter estimation are our future work for this weighted bipartite graph model.

Moreover, the results of the cost model analysis suggest that the scale parameter is an indicator of the increased cost. We would like to further explore the results combining the scale parameter and the shape parameter of the power law in the cost model. This will further enhance our understanding of the cost structures. More factors can be added to the model to allow different factors to be explained by the model such as:

- The arrival rate of users and objects could follow the Poisson distribution and more param-

eters could be added.

- The content creators (artists) could be added to the model because content creators could influence the attractiveness of their content (songs).

- Personalized attractiveness could be added to the model and personal preferences (personalized attractiveness) should have a greater influence than the popularity (global attractiveness) of the songs

- The social network among users could be added by generating a network of users following the fitness-age scale-free network and adding parameters, which would also include the local attractiveness from the relationships between the users.

This chapter concerns only the cost of royalty fee, but in the real service, there are other costs such as server cost, marketing cost, and others. Together, to maximize profit, the provider should also be concerned about the income. Due to the variety of business model in the online music streaming services, we did not include these part in our paper but these should be investigated in the future.

Chapter 6
CONCLUSION AND FUTURE WORK

In this chapter, the implications from the results of the studies are summarized in section 6.1. Then all of the contributions are concluded in section 6.2 and the future works are listed in section 6.3.

6.1 Management Implications

From the objective of this study, which is to increase the profit of freemium music streaming business, the results in previous chapters can be applied to management implications both in income and cost part as shown below.

The result in chapter 3 shows that, chance to convert to premium user is related to the number of friends. This can imply that increase the connection between free users can lead to increase conversion rate too. This can be implemented by recommend other free users to be friend with that free user.

For a premium user, connection strength between premium friends have positive relation with chance to continue being premium user. This mean that when user change to premium user, the system should recommend other premium users to be friends and recommend some new songs to that user too.

While the recommendation system have to be used in the previous management implications, the performance of the recommendation system can be improved by using both data of social network of users and other information in the Bayesian statistic based random walk with restart recommendation system proposed in chapter 4. This recommendation system can be used in next implications too.

Due to the profit is related to both income and cost, the cost of providing the streaming service should be concerned. The result in chapter 5 shows that the cost of royalty fee for the music streaming service is related to number of usage of new and old songs. The more listening of the new songs and the less listening of old songs lead to the less royalty fee cost. This can imply that if the system can recommend new songs more that older songs, the cost of royalty fee should be reduced.

The managerial recommendations from these studies can be concluded as:

- Recommend other free users to be friend with free user.
- Recommend other premium users to be friend with premium user.
- Recommend newer songs to the users more than the older one.
- Use both individual behaviors and social network's data and Bayesian statistic to create better recommendation system.

6.2 Contribution

To address the profitability problem that freemium music streaming providers face, this dissertation studies three approaches to increase the ability to make profit for music streaming providers. The three approaches address the providers' income, the providers' cost and user satisfaction as shown in Fig. 6.1.

First, this dissertation examined the behavior of online consumers using historical usage log, social network information and user's history of premium subscription. The results show that for free user, number of both free and premium users' friends have positive relation to the chance to convert to a premium user, however, the usage rate and the chance to convert to a premium user shows a different relation between users who stop using premium service, as negative relation, and the other free users, which are positive relation. For premium users, the number of friends who are premium users and connection strength between premium friends and usage rates have positive relation to the chance to continue using premium service, but the number of friends who are free users has negative relation, which show stronger relation for the user who start using premium service than the other premium users. These results can suggest strategies to increasing conversion rates by recommending free users to be friends with other free users and recommend premium users to be friends with other premium users especially with the ones who just start using premium service. This is shown in blue in the diagram. Additionally, the recommendation of songs is also necessary.

To increase user satisfaction in music consumption, this dissertation improves a recommendation system, which was based on random walk with restart, by using social information based probability in the randomized process of the system. The accuracy of the system improves from the previous recommendation system specially when the data is limited, which would allow users

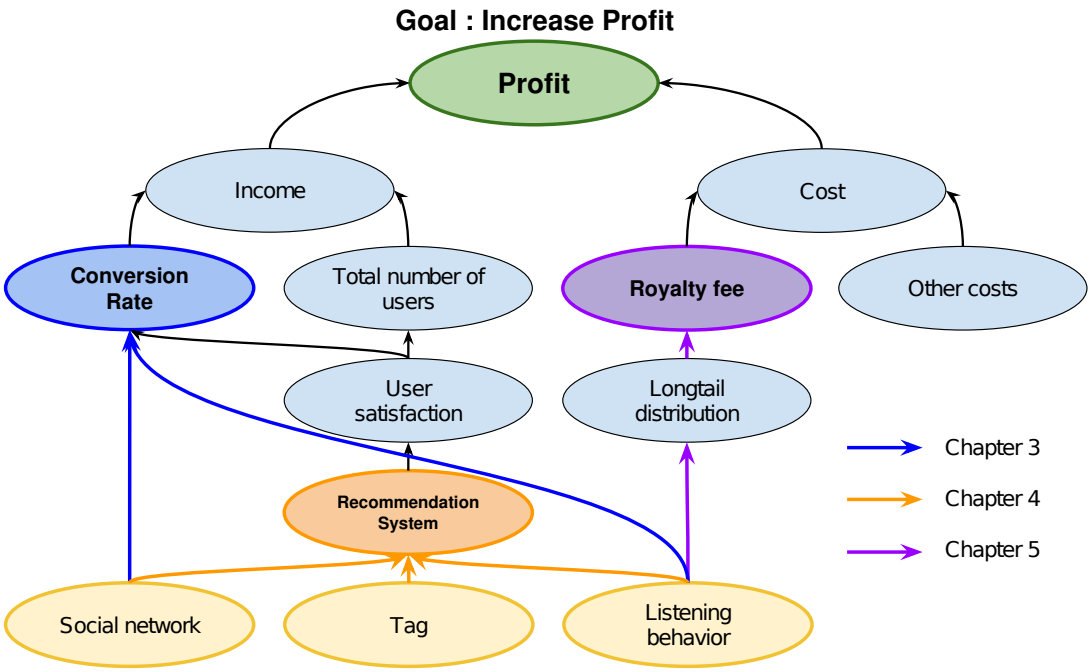

Figure 6.1: The contributions of this research

to find songs that they like, and improves user satisfaction to the system. This is shown in orange part in the diagram.

In the domain of longtail phenomenon in usage distribution of music consumption, this dissertation improves the generating model by Zhang et al. [146] to reduce the discrepancy between the model and the real world, in which the previous model did not allow new songs to gain popularity quicker than older songs. The new generating model can explain the real world data better. Moreover, the model parameters were analyzed in terms of how they affect the shapes of the longtail. This is used to make recommendations to providers about how to strategically adjust the user interfaces and recommendation system to reduce the usage of old contents in order to reduce the licensing costs that providers have to pay to music creators. This is shown in purple part in the diagram.

Overall, in order to increase profits for freemium music streaming providers, I studied the factors that affect both premium subscribes and unsubscribes in subscription-based freemium music streaming services and proposed a managerial recommendation to increase conversion rate. Moreover, I developed a recommendation system to improve the user experiences in discovering music to increase users satisfaction to the services. Finally, I examined longtail distribution in music consumption and developed a model that explains the longtail phenomena in music consumption and use the model to explore how providers can reduce royalty costs. Combined 3 parts which should be able to increase income and reduce cost that should increase profit to the provider.

6.3 Future Work

Many directions of future work can be explored. For increasing the conversion rate, both strategy for increasing the conversion and preventing unsubscribes, other parameters could be examined to improve the accuracy of the model. An example of the parameters is the similarity of usage between friends. For social-information-based conditional probability transition matrix RWR-based recommendation system, in order to effectively deploy, the computation needs to be optimized to increase the scalability of this approach. For the weighted bipartite graph that generates longtail distribution in music, users could access contents with Poisson process or new users and new objects could be added to the system with random process instead of the fixed rates of creating users and object in the current model.

This studies is based on information from last.fm, which is global music streaming service, therefore, the data refer to behavior of global user, which mean that users of localized service may

have difference behaviors and have to be studied separated in each countries.

Lastly, the author of this dissertation hopes that this research will be useful to providers of freemium music streaming services. While the empirical data in this dissertation is from a music streaming service, the methods can be applied beyond music domain. Because RWR recommendation system was applied to various target not only music ([74, 90]), but also movies ([41, 48, 140, 122, 79, 141]),and publications [131], this studies should be applicable to other social-based freemium services such as movie/video streaming services or publications or other contents providers too. However, the distribution of content consumption has to be confirmed to be power-law distribution and the cost structure of royalty fee for the contents has to be changed. The increased ability to make profits for content providers would allow the business to sustain-ably operate and provide quality services to a large number of users.

BIBLIOGRAPHY

[1] Product sampling: Costly but effective. http://www.fmcg.ie/product-sampling-on-the-increase/, . Accessed: 2016-08-27.

[2] Startup school: Wired editor chris anderson on freemium business models. https://techcrunch.com/2009/10/24/startup-school-wired-editor-chris-anderson-on-freemium-business-models/, . Accessed: 2016-08-27.

[3] Freemium. https://en.wikipedia.org/wiki/Freemium, . Accessed: 2016-08-27.

[4] List of the 515,576 tracks for which has the year information, ordered by year. http://labrosa.ee.columbia.edu/millionsong/sites/default/files/AdditionalFiles/tracks_per_year.txt, 2011. [Online; accessed 30-Aug-2016].

[5] Vlad Adrian Maftei, Vassilis C Gerogiannis, and Elpiniki I Papageorgiou. Critical success factors of online music streaming services-a case study of applying the fuzzy cognitive maps method. *International Journal of Technology Marketing*, 11(3):276–300, 2016.

[6] Charu C. Aggarwal. *Social Network Data Analytics*. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 1441984615, 9781441984616.

[7] Icek Ajzen. From intentions to actions: A theory of planned behavior. In *Action control*, pages 11–39. Springer, 1985.

[8] Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[9] Icek Ajzen and Martin Fishbein. The prediction of behavior from attitudinal and normative variables. *Journal of experimental social Psychology*, 6(4):466–487, 1970.

[10] David Aldous and Jim Fill. Reversible markov chains and random walks on graphs, 2002.

[11] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion Books. Hyperion, 2006. ISBN 1401302378.

[12] Chris Anderson. *Free: The future of a radical price*. Hyperion, 2009. ISBN 1401322905, 9781401322908.

[13] Ralitsa Angelova, Marek Lipczak, Evangelos Milios, and Pawel Pralat. Investigating the properties of a social bookmarking and tagging network. *International Journal of Data Warehousing and Mining*, 6(1):1–19, January 2010. ISSN 1548-3924. doi: 10.4018/jdwm.2010090801.

[14] Asim Ansari, Skander Essegaier, and Rajeev Kohli. Internet recommendation systems. *Journal of Marketing Research*, 37(3):363–375, August 2000. ISSN 0022-2437. doi: 10.1509/jmkr.37.3.363.18779.

[15] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM, ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935845. URL http://doi.acm.org/10.1145/1935826.1935845.

[16] Albert Bandura. Model of causality in social learning theory. In *Cognition and psychotherapy*, pages 81–99. Springer, 1985.

[17] Albert Bandura. Social cognitive theory of moral thought and action. *Handbook of moral behavior and development*, 1:45–103, 1991.

[18] Albert Bandura. Social cognitive theory: An agentic perspective. *Annual review of psychology*, 52(1):1–26, 2001.

[19] Ravi Bapna and Akhmed Umyarov. Do your online friends make you pay? a randomized field experiment on peer influence in online social networks. *Management Science*, 61(8): 1902–1920, 2015.

[20] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[21] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[22] Ranieri Baraglia and Fabrizio Silvestri. An online recommender system for large web sites. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, pages 199–205, 2004. doi: 10.1109/WI.2004.10158.

[23] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[24] Prantik Bhattacharyya, Ankush Garg, and ShyhtsunFelix Wu. Analysis of user keyword similarity in online social networks. *Social Network Analysis and Mining*, 1(3):143–158, 2011. ISSN 1869-5450. doi: 10.1007/s13278-010-0006-4.

[25] Ginestra Bianconi and A-L Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

[26] Ginestra Bianconi and Albert-László Barabási. Bose-einstein condensation in complex networks. *Physical review letters*, 86(24):5632, 2001.

[27] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 193–202, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458112.

[28] Matthew Brand. A random walks perspective on maximizing satisfaction and profit. In *SIAM International Conference on Data Mining*, pages 12–19, 2005.

[29] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002. ISSN 0924-1868. doi: 10.1023/A:1021240730564.

[30] Òscar Celma and Pedro Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, page 5. ACM, 2008.

[31] Hsing Kenneth Cheng, Shengli Li, and Yipeng Liu. Optimal software free trial strategy: Limited version, time-locked, or hybrid? *Production and Operations Management*, 2014. ISSN 1937-5956. doi: 10.1111/poms.12248. URL http://dx.doi.org/10.1111/poms.12248.

[32] Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim. A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications*, 23(3):329 – 342, 2002. ISSN 0957-4174. doi: 10.1016/S0957-4174(02)00052-0.

[33] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[34] Maarten Clements, Arjen P de Vries, and Marcel JT Reinders. Optimizing single term queries using a personalized markov random walk over the social graph. In *Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR)*, 2008.

[35] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 160–168, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401914.

[36] Nick Craswell and Martin Szummer. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277784.

[37] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. User acceptance of computer technology: a comparison of two theoretical models. *Management science*, 35(8): 982–1003, 1989.

[38] Maurits de Klepper, Ed Sleebos, Gerhard van de Bunt, and Filip Agneessens. Similarity in friendship networks: Selection or influence? the effect of constraining contexts and non-visible individual attributes. *Social Networks*, 32(1):82 – 90, 2010. ISSN 0378-8733. doi: 10.1016/j.socnet.2009.06.003. <ce:title>Dynamics of Social Networks</ce:title>.

[39] Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015. URL http://www.jstatsoft.org/v64/i04/.

[40] Debabrata Dey, Atanu Lahiri, and Dengpan Liu. Consumer learning and time-locked trials of software products. *Journal of Management Information Systems*, 30(2):239–268, 2013. ISSN 07421222. doi: 10.2753/MIS0742-1222300209. URL http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=93255289&lang=ja&site=ehost-live.

[41] Fernando Díez, J. Enrique Chavarriaga, Pedro G. Campos, and Alejandro Bellogín. Movie recommendations based in explicit and implicit features extracted from the filmtipset dataset. In *Proceedings of the Workshop on Context-Aware Movie Recommendation*, CAMRa '10, pages 45–52, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0258-6. doi: 10.1145/1869652.1869660.

[42] Jonathan Doerr, Alexander Benlian, Johannes Vetter, and Thomas Hess. Pricing of content services–an empirical investigation of music as a service. In *Sustainable e-business management*, pages 13–24. Springer, 2010.

[43] Frederico Durao and Peter Dolog. A personalized tag-based recommendation in social web systems. *Adaptation and Personalization for Web*, 2:40, 2009.

[44] Alessia D' Andrea, Fernando Ferri, and Patrizia Grifoni. *An overview of methods for virtual social networks analysis*. Springer, 2010.

[45] Anita Elberse. *A Taste for Obscurity: An Individual-Level Examination of'Long Tail'Consumption*. Harvard Business School, 2008.

[46] Albrecht Enders, Harald Hungenberg, Hans-Peter Denker, and Sebastian Mauch. The long tail of social networking.: Revenue models of social networking sites. *European Management Journal*, 26(3):199 – 211, 2008. ISSN 0263-2373. doi: 10.1016/j.emj.2008.02.002. URL http://www.sciencedirect.com/science/article/pii/S0263237308000200.

[47] PELIN ERCAN. *DETECTION OF CHURNERS IN INTERNET GAMES USING CRM APPROACH: A CASE STUDY ON PISHTI PLUS*. PhD thesis, MIDDLE EAST TECHNICAL UNIVERSITY, 2015.

[48] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative

recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, March 2007. ISSN 1041-4347. doi: 10.1109/TKDE.2007.46.

[49] Yasuhiro Fujiwara, Makoto Nakatsuji, Makoto Onizuka, and Masaru Kitsuregawa. Fast and exact top-k search for random walk with restart. *Proc. VLDB Endow.*, 5(5):442–453, January 2012. ISSN 2150-8097.

[50] R Gokhale and Ravi S Narayanaswamy. The role of experience in discontinuance of it innovations. In *Proceeding of the 2006 Southern Association for Information Systems Conference*, pages 1–7, 2006.

[51] Charles M Grinstead and J Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1998.

[52] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 194–201, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835484.

[53] Fabian Hadiji, Rafet Sifa, Anders Drachen, Christian Thurau, Kristian Kersting, and Christian Bauckhage. Predicting player churn in the wild. In *2014 IEEE Conference on Computational Intelligence and Games*, pages 1–8. IEEE, 2014.

[54] Michael Haenlein. A social network analysis of customer-level revenue distribution. *Marketing letters*, 22(1):15–29, 2011. ISSN 0923-0645. doi: 10.1007/s11002-009-9099-9. URL http://dx.doi.org/10.1007/s11002-009-9099-9.

[55] Eui-Hong (Sam) Han and George Karypis. Feature-based recommendation system. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 446–452, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6. doi: 10.1145/1099554.1099683.

[56] F. Maxwell Harper, Xin Li, Yan Chen, and Joseph A. Konstan. An economic model of user rating in an online recommender system. In *Proceedings of the 10th international conference on User Modeling*, UM'05, pages 307–316, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-27885-0, 978-3-540-27885-6. doi: 10.1007/11527886_40.

[57] S Alexander Haslam. *Psychology in organizations*. Sage, 2004.

[58] Sarah Hayman. Folksonomies and tagging: New developments in social bookmarking. In *Ark Group Conference: Developing and Improving Classification Schemes*. Citeseer, 2007.

[59] Thorsten Hennig-Thurau, André Marchand, and Paul Marx. Can automated group recommender systems help consumers make better choices? *Journal of Marketing*, 76(5): 89–109, 2012.

[60] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 195–206, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. doi: 10.1145/1341531.1341558.

[61] Shawndra Hill, Foster Provost, and Chris Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):pp. 256–276, 2006. ISSN 08834237. URL http://www.jstor.org/stable/27645754.

[62] Hayiel Hino. Use-adoption gaps in food retailing theoretical framework and application in an emerging economy context (jordan). *Journal of Macromarketing*, 35(3):368–386, 2015.

[63] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-34544-2. doi: 10.1007/11762256_31.

[64] Hai-Bo Hu and Ding-Yi Han. Empirical analysis of individual popularity and activity on an online music service system. *Physica A: Statistical Mechanics and its Applications*, 387 (23):5916–5921, 2008.

[65] Zan Huang, Daniel Zeng, and Hsinchun Chen. A link analysis approach to recommendation under sparse data. In *Proc. 2004 Americas Conf. Information Systems*, 2004.

[66] Tim Ingham. Spotify revenues topped $2bn last year as losses hit $194m, 2016. URL http://www.musicbusinessworldwide.com/spotify-revenues-topped-2bn-last-year-as-losses-hit-194m/.

[67] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. ISBN 0521493366, 9780521493369.

[68] Yong-Sheng Jin and Zhao-Hui Li. A use-diffusion model of 3g services in china. *African Journal of Business Management*, 5(27):11168, 2011.

[69] Zsolt Katona, Peter Pal Zubcsek, and Miklos Sarvary. Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research*, 48 (3):425–443, 2011. doi: 10.1509/jmkr.48.3.425. URL http://dx.doi.org/10.1509/jmkr.48.3.425.

[70] Hee-Su Kim and Choong-Han Yoon. Determinants of subscriber churn and customer loyalty in the korean mobile telephony market. *Telecommunications policy*, 28(9): 751–765, 2004.

[71] Jae Kyeong Kim, Hyea Kyeong Kim, Hee Young Oh, and Young U. Ryu. A group recommendation system for online communities. *International Journal of Information Management*, 30(3):212 – 219, 2010. ISSN 0268-4012. doi: 10.1016/j.ijinfomgt.2009.09.006.

[72] Nuri Kim, Jon Krosnick, and Daniel Casasanto. Moderators of candidate name-order effects in elections: An experiment. *Political Psychology*, 36(5):525–542, 2015.

[73] Nicolás Mongiardino Koch and Ignacio M Soto. Let the music be your master: Power laws and music listening habits. *Musicae Scientiae*, page 1029864915619000, 2016.

[74] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 195–202, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1571977.

[75] Jérôme Kunegis, Marcel Blattner, and Christine Moser. Preferential attachment in online networks: Measurement and explanations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 205–214. ACM, 2013.

[76] Renaud Lambiotte and Marcel Ausloos. Collaborative tagging as a tripartite network. In VassilN. Alexandrov, GeertDick Albada, PeterM.A. Sloot, and Jack Dongarra, editors, *Computational Science ¨ICCS 2006*, volume 3993 of *Lecture Notes in Computer Science*, pages 1114–1117. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-34383-7. doi: 10.1007/11758532_152.

[77] Michael Steven Lane and Adrian Stagg. University staff adoption of ipads: An empirical study using an extended tam model. *Australasian Journal of Information Systems*, 18(3), 2014.

[78] Damien Lefortier, Liudmila Ostroumova, and Egor Samosvat. Evolution of the media web. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 80–92. Springer, 2013.

[79] Bin Liu and Zheng Yuan. Incorporating social networks and user opinions for collaborative recommendation: local trust network based method. In *Proceedings of the Workshop on Context-Aware Movie Recommendation*, CAMRa '10, pages 53–56, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0258-6. doi: 10.1145/1869652.1869661.

[80] Alexandre B Lopes and Dennis F Galletta. Consumer perceptions and willingness to pay for intrinsically motivated online content. *Journal of Management Information Systems*, 23 (2):203–231, 2006.

[81] L. Lovász. Random walks on graphs: A survey. In D. Miklós, V. T. Sós, and T. Szőnyi, editors, *Combinatorics, Paul Erdős is Eighty*, volume 2, pages 353–398. János Bolyai Mathematical Society, Budapest, 1996.

[82] László Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2 (1):1–46, 1993.

[83] Claudio Lucchese, Raffaele Perego, Fabrizio Silvestri, Hossein Vahabi, and Rossano Venturini. How random walks can help tourism. In Ricardo Baeza-Yates, ArjenP. Vries, Hugo Zaragoza, B.Barla Cambazoglu, Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri, editors, *Advances in Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 195–206. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28996-5. doi: 10.1007/978-3-642-28997-2_17.

[84] Gary Madden, Scott J Savage, and Grant Coble-Neal. Subscriber churn in the australian isp market. *Information Economics and Policy*, 11(2):195–207, 1999.

[85] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

[86] Leandro Balby Marinho, Andreas Hotho, Robert Jschke, Alexandros Nanopoulos, Steffen Rendle, Lars Schmidt-Thieme, Gerd Stumme, and Panagiotis Symeonidis. *Recommender Systems for Social Tagging Systems*. Springer Publishing Company, Incorporated, 2012. ISBN 1461418933, 9781461418931.

[87] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In Yolanda Gil, Enrico Motta, V.Richard Benjamins, and MarkA. Musen, editors, *The Semantic Web - ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-29754-3. doi: 10.1007/11574620_38.

[88] Joanne M Miller and Jon A Krosnick. The impact of candidate name order on election outcomes. *Public Opinion Quarterly*, pages 291–330, 1998.

[89] Kazuyuki Motohashi, Deog-Ro Lee, Yeong-Wha Sawng, Seung-Ho Kim, et al. Innovative converged service and its adoption, use and diffusion: a holistic approach to diffusion of innovations, combining adoption-diffusion and use-diffusion paradigms. *Journal of Business Economics and Management*, 13(2):308–333, 2012.

[90] Makoto Nakatsuji, Yasuhiro Fujiwara, Akimichi Tanaka, Tadasu Uchiyama, and Toru Ishida. Recommendations over domain specific user graphs. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 607–612, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press. ISBN 978-1-60750-605-8.

[91] Marius F. Niculescu and D. J. Wu. Economics of free under perpetual licensing: Implications for the software industry. *Information Systems Research*, 25(1):173–199, 2014. doi: 10.1287/isre.2013.0508. URL http://dx.doi.org/10.1287/isre.2013.0508.

[92] Gal Oestreicher-Singer and Lior Zalmanson. Content or community? a digital business strategy for content providers in the social age. *MIS Quarterly*, 37(2):591–616, June 2013. ISSN 0276-7783. URL http://dl.acm.org/citation.cfm?id=2535658.2535672.

[93] Richard L Oliver. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology*, 62(4):480, 1977.

[94] Richard L Oliver. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, pages 460–469, 1980.

[95] Femi Olumofin and Ian Goldberg. Revisiting the computational practicality of private information retrieval. In George Danezis, editor, *Financial Cryptography and Data Security*, volume 7035 of *Lecture Notes in Computer Science*, pages 158–172. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-27575-3. doi: 10.1007/978-3-642-27576-0_13. URL http://dx.doi.org/10.1007/978-3-642-27576-0_13.

[96] Liudmila Ostroumova Prokhorenkova and Egor Samosvat. Recency-based preferential attachment models. *Journal of Complex Networks*, 4(4):475–499, 2016.

[97] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[98] Chan Ho Park and Minsuk Kahng. Temporal dynamics in music listening behavior: A case study of online music service. In *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, pages 573–578. IEEE, 2010.

[99] Hyun Jung Park and Hyung Seok Lee. Product smartness and use-diffusion of smart products: the mediating roles of consumption values. *Asian Social Science*, 10(3):54, 2014.

[100] Tim Paul Thomes. An economic analysis of online streaming music services. *Information Economics and Policy*, 25(2):81–91, 2013.

[101] Karl Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.

[102] Isabella Peters. *Folksonomies. Indexing and Retrieval in Web 2.0*. Walter de Gruyter & Co., Hawthorne, NJ, USA, 1st edition, 2009. ISBN 3598251793, 9783598251795.

[103] Suchit Pongnumkul and Kazuyuki Motohashi. Random walk-based recommendation with restart using social information and bayesian transition matrices. *International Journal of Computer Applications*, 114(9), 2015.

[104] Suchit Pongnumkul and Kazuyuki Motohashi. A bipartite fitness model for online music streaming services. *Physica A: Statistical Mechanics and its Applications*, 490:1125–1137, 2018.

[105] Nicolas Pujol. Freemium: attributes of an emerging business model. *Available at SSRN 1718663*, 2010.

[106] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

[107] Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011. ISBN 1107015359, 9781107015357.

[108] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010. ISBN 0387858199, 9780387858197.

[109] Yossi Richter, Elad Yom-Tov, and Noam Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, volume 2010, pages 732–741. SIAM, 2010.

[110] Marja riitta Koivunen. Semantic authoring by tagging with annotea social bookmarks and topics. In *In The 5th International Semantic Web Conference (ISWC2006) - 1st Semantic Authoring and Annotation Workshop (SAAW2006*, 2006.

[111] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294.

[112] E.M. Rogers. *Diffusion of innovations*. Free Press of Glencoe, 1962. URL https://books.google.co.th/books?id=zw0-AAAAIAAJ.

[113] Julian Runge, Peng Gao, Florent Garcin, and Boi Faltings. Churn prediction for high-value players in casual social games. In *2014 IEEE Conference on Computational Intelligence and Games*, pages 1–8. IEEE, 2014.

[114] Colin S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015. URL http://www.jstatsoft.org/v64/i02/.

[115] Yeong-Wha Sawng, Kazuyuki Motohashi, and Gang-Hoon Kim. Comparative analysis of innovative diffusion in the high-tech markets of japan and south korea: a use–diffusion model approach. *Service Business*, 7(1):143–166, 2013.

[116] Robert R Schaller. Moore's law: past, present and future. *Spectrum, IEEE*, 34(6):52–59, Jun 1997. ISSN 0018-9235. doi: 10.1109/6.591665. URL http://dx.doi.org/10.1109/6.591665.

[117] Dimitri Schuurman, Lieven De Marez, and Katrien Berte. Adoption versus use diffusion of idtv in flanders-personalized television content as a tool to cross the chasm? *Computers in Entertainment (CIE)*, 9(3):20, 2011.

[118] Davide Semenzin, Edwin Meulendijks, Wilbert Seele, Christoph Wagner, and Sjaak Brinkkemper. Differentiation in freemium: Where does the line lie? In MichaelA. Cusumano, Bala Iyer, and N. Venkatraman, editors, *Software Business*, volume 114 of *Lecture Notes in Business Information Processing*, pages 291–296. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-30745-4. doi: 10.1007/978-3-642-30746-1_27. URL http://dx.doi.org/10.1007/978-3-642-30746-1_27.

[119] Eric Benjamin Seufert. *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue*. Elsevier, 2013.

[120] Lisa F Seymour and Mogen Naidoo. The usage and impact of broadband: A south african household analysis. *Electronic Journal of Information Systems Evaluation*, 16(2):134–147, 2013.

[121] Ming-Sheng Shang, Linyuan Lü, Yi-Cheng Zhang, and Tao Zhou. Empirical analysis of web-based user-object bipartite networks. *EPL (Europhysics Letters)*, 90(4):48006, 2010.

[122] Yue Shi, Martha Larson, and Alan Hanjalic. Mining relational context-aware graph for rater identification. In *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, CAMRa '11, pages 53–59, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0825-0. doi: 10.1145/2096112.2096122.

[123] Chuan-Fong Shih and Alladi Venkatesh. Beyond adoption: Development and application of use diffusion (ud) model to study household use of computers. *J. Market*, 68:59–72, 2002.

[124] Chuan-Fong Shih and Alladi Venkatesh. Beyond adoption: Development and application of a use-diffusion model. *Journal of marketing*, 68(1):59–72, 2004.

[125] Eric Shih, Alladi Venkatesh, Steven Chen, and Erik Kruse. Dynamic use diffusion model in a cross-national context: A comparative study of the united states, sweden, and india. *Journal of Product Innovation Management*, 30(1):4–16, 2013.

[126] Minhee Son and Kyesook Han. Beyond the technology adoption: Technology readiness effects on post-adoption behavior. *Journal of Business Research*, 64(11):1178–1182, 2011.

[127] Markus Strohmaier. Purpose tagging: capturing user intent to assist goal-oriented social search. In *Proceedings of the 2008 ACM workshop on Search in social media*, SSM '08, pages 35–42, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-258-0. doi: 10.1145/1458583.1458603.

[128] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4:2–4:2, January 2009. ISSN 1687-7470. doi: 10.1155/2009/421425.

[129] Jared Sylvester and William Rand. Keeping up with the (pre-teen) joneses: The effect of friendship on freemium conversion. In *Winter Conference on Business Intelligence, Snowbird, Utah*, 2014.

[130] Henri Tajfel. Social identity and intergroup behaviour. *Information (International Social Science Council)*, 13(2):65–93, 1974.

[131] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1285–1293, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339730.

[132] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: fast solutions and applications. *Knowl. Inf. Syst.*, 14(3):327–346, March 2008. ISSN 0219-1377. doi: 10.1007/s10115-007-0094-2.

[133] Jennifer Trant. Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*, 10(1), 2009. ISSN 1368-7506.

[134] J.C. Turner. *Social Categorization and Self-Concept: A Social Cognitive Theory of Group Behavior*, pages 77–121. JAI Press, Greenwich, Connecticut, 1985.

[135] Rainer Typke, Frans Wiering, and Remco C. Veltkamp. A survey of music information retrieval systems. In *IN ISMIR*, pages 153–160, 2005.

[136] Viswanath Venkatesh and Fred D Davis. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2):186–204, 2000.

[137] Victor Harold Vroom. *Work and motivation*. San Francisco : Jossey-Bass Publishers, 1st ed edition, 1995. ISBN 0787900303 (pbk.). Originally published: New York : Wiley, 1964.

[138] Thomas M Wagner, Alexander Benlian, and Thomas Hess. Converting freemium customers from free to premium—the role of the perceived premium fit in the case of music as a service. *Electronic Markets*, 24(4):259–268, 2014.

[139] Hao Wang and Alvin Chin. Social influence on being a pay user in freemium-based social networks. In *Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on*, pages 526–533. IEEE, March 2011. doi: 10.1109/AINA.2011.35.

[140] Ziqi Wang, Yuwei Tan, and Ming Zhang. Graph-based recommendation on social networks. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*, pages 116–122, 2010. doi: 10.1109/APWeb.2010.60.

[141] Ziqi Wang, Ming Zhang, Yuwei Tan, Wenqing Wang, Yuexiang Zhang, and Ling Chen. Recommendation algorithm based on graph-model considering user background information. In *Creating, Connecting and Collaborating through Computing (C5), 2011 Ninth International Conference on*, pages 32–39, 2011. doi: 10.1109/C5.2011.11.

[142] Wikipedia. Information retrieval - wikipedia, the free encyclopedia, May 2013. URL http://en.wikipedia.org/wiki/Information_retrieval. [Online; accessed 27-May-2013].

[143] Fred Wilson. The freemium business model, jun 2006. URL http://avc.com/2006/03/the_freemium_bu/.

[144] Jing Xia, D. Caragea, and W. Hsu. Bi-relational network analysis using a fast random walk with restart. In *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pages 1052–1057, 2009. doi: 10.1109/ICDM.2009.134.

[145] Akbar Zaheer and Geoffrey G Bell. Benefiting from network position: firm capabilities, structural holes, and performance. *Strategic management journal*, 26(9):809–825, 2005. ISSN 1097-0266. doi: 10.1002/smj.482. URL http://dx.doi.org/10.1002/smj.482.

[146] Chu-Xu Zhang, Zi-Ke Zhang, and Chuang Liu. An evolving model of online bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 392(23):6100–6106, 2013.

[147] Xuguo Zheng and Qinghua Chen. A weighted bipartite network based on the strength preferential attachment. *Applied Mathematical Sciences*, 5(73):3619–3625, 2011.

[148] Yanbo Zhou, An Zeng, and Wei-Hong Wang. Temporal effects in trend prediction: identifying the most popular nodes in the future. *PloS one*, 10(3):e0120735, 2015.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my advisor Prof. Kazuyuki Motohashi for his continuous support of my Ph.D study. His guidance, patience, motivation, immense knowledge and pointers to related research have helped me in all the time of research and the writing of this thesis. With his help, we published 2 papers which formed the base of this dissertation [103, 104]. I could not imagine having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank my thesis committee: Prof. Kazumitsu Nawata, Prof. Kiyoshi Izumi, Prof. Yeong-Wha Sawng and Dr. Junichiro Mori, not only for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

I also would like to thank Ms. Miwa Abe, the secretary of Motohashi laboratory, who helped me a lot while I stayed in Japan and after I moved back to Thailand.

Last but not the least, I would like to thank my family: my parents and my sister for supporting me spiritually while writing this thesis and throughout my life.

Appendix A

A.1 Lemma and Proof

Lemma A.1.1. $\lim_{t\to\infty} W(t)$ *is a constant (for the Object Degree Distribution in Section 5.5).*

Proof. Let us now check that $\lim_{t\to\infty} W(t)$ is indeed a constant. Let $\rho(q)$ be the probability density function of $f_o(i)$. Therefore,

$$W(t) = \int_0^\infty (\int_0^t \rho(q) q s_o(x) e^{-\frac{t-x}{\tau}} dx) dq \tag{A.1}$$

$$= \int_0^\infty (\int_0^t \rho(q) q n e^{\frac{\tau N}{W} q} e^{-\frac{t-x}{\tau}} dx) dq \tag{A.2}$$

$$= \int_0^\infty (\rho(q) q \frac{nW}{Nq} (e^{\frac{N\tau q}{W}(1-e^{\frac{-t}{\tau}})} - 1)) dq \tag{A.3}$$

$$= \int_0^\infty \rho(q) \frac{nW}{N} e^{\frac{N\tau q}{W}(1-e^{\frac{-t}{\tau}})} dq - \int_0^\infty \rho(q) \frac{nW}{N} dq \tag{A.4}$$

$$= \frac{nW}{N} (\int_0^\infty \rho(q) e^{\frac{N\tau q}{W}(1-e^{\frac{-t}{\tau}})} dq - 1) \tag{A.5}$$

$$\lim_{t\to\infty} W(t) = \frac{nW}{N} (\int_0^\infty \rho(q) e^{\frac{N\tau q}{W}} dq - 1) \tag{A.6}$$

Let $F(W) = \frac{nW}{N} (\int_0^\infty \rho(q) e^{\frac{N\tau q}{W}} dq - 1)$, we confirm the monotone decreasing of the function $F(W)$ such that

$$F'(W) = \frac{nW}{N} \frac{d}{dW} (\int_0^\infty \rho(q) e^{\frac{N\tau q}{W}} dq - 1) + \frac{(\int_0^\infty \rho(q) e^{\frac{N\tau q}{W}} dq - 1)}{N} \tag{A.7}$$

$$= \frac{n}{N} (\int_0^\infty \rho(q) W \frac{d}{dW} e^{\frac{N\tau q}{W}} dq + \int_0^\infty \rho(q) e^{\frac{N\tau q}{W}} dq - 1) \tag{A.8}$$

$$= \frac{n}{N} (\int_0^\infty \rho(q) \frac{W - N\tau q}{W} e^{\frac{N\tau q}{W}} dq - 1) \tag{A.9}$$

Because $\frac{W - N\tau q}{W} e^{\frac{N\tau q}{W}} \leq 1$ and $\int_0^\infty \rho(q) dq = 1$, $F'(W) \leq 0$ which shows that $F(W)$ is a monotone decreasing function.

Because $\lim_{W\to\infty} \frac{nW}{N} e^{\frac{N\tau q}{W}} = \lim_{W\to\infty} n\tau q e^{\frac{N\tau q}{W}} = n\tau q$, $F(x) \to n\tau \mathbf{E}[q]$ as $x \to \infty$ and $F(x) \to \infty$ as $x \to 0$ tell us that $y = x$ and $y = F(x)$ have a unique intersection, which makes Eq. A.6 have a unique solution.

□