6-2018

Music popularity, diffusion and recommendation in social networks: A fusion analytics approach

Jing REN
Singapore Management University, jing.ren.2012@phdis.smu.edu.sg

MUSIC POPULARITY, DIFFUSION AND RECOMMENDATION IN SOCIAL NETWORKS: A FUSION ANALYTICS APPROACH

JING REN

SINGAPORE MANAGEMENT UNIVERSITY

2018

Music Popularity, Diffusion and Recommendation in Social Networks: A Fusion Analytics Approach

by

Jing Ren

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Information Systems

Dissertation Committee:

Robert J. Kauffman (Supervisor/Chair)
Professor of Information Systems
Singapore Management University

Zhiling Guo
Associate Professor of Information Systems
Singapore Management University

Qiuhong Wang
Assistant Professor of Information Systems
Singapore Management University

David R. King (External Reviewer)
Advisor, Board Member
Teuvonet Technologies, LLC, United States

Singapore Management University

2018

Music Popularity, Diffusion and Recommendation in Social Networks: A Fusion Analytics Approach

Jing Ren

Abstract

Streaming music and social networks offer an easy way for people to gain access to a massive amount of music, but there are also challenges for the music industry to design for promotion strategies via the new channels. My dissertation employs a fusion of machine-based methods and explanatory empiricism to explore music popularity, diffusion, and promotion in the social network context.

Essay 1 investigates the determinants of music track popularity and patterns, and their impacts. Music track popularity is the degree to which a track can satisfy an individual's listening tastes over time. By studying streaming music in the social network scenario, this essay assesses the effects of music semantic content, artist reputation and social context on music track popularity on the top-ranking chart for the streaming music service, Last.fm, from 2005 to 2015. This essay proposes two measurements for music track popularity, and constructs complete music track descriptions by combining machine learning and explanatory econometric methods, which leverages the power of these two methods to better explain and predict what kinds of music can be more popular. The results demonstrate the ability of this approach to music popularity estimation at an early stage after a track's release.

Essay 2 examines the impacts of external information on streaming music diffusion in a social network environment. Music social networks operate on a semi-closed platform, which makes music diffusion analysis a complex research process.

This essay uses propensity score matching to match the panel datasets from the listening records of over 557,000 users for 1,300 artists in a one-year period for analysis. Difference-in-differences and count data models are then developed to assess the effects of external information on streaming music diffusion at the macro- and micro-levels. This essay finds that external information has a significant impact on an artist's streaming music diffusion, and the impact and their persistence are related to artist information type and also listeners' geographic locations.

Essay 3 discusses the design of a two-sided value-based recommender system to help music industry professionals promote music and artists more effectively on streaming social networks. Traditional music recommendations usually only focus on improving recommendation accuracy for consumers, while ignoring the promotion requirements of a specific artist. This essay combines the analysis of the business value of music industry firms and the utility of consumers into an integrated model. Compared to commonly-used recommendation methods, the results show a clear increase in the conversion rate of listener recommendations for an artist by considering both sides' value and other factors, including geolocation, time, external information and listening behavior. This essay delivers new ways to develop online streaming music recommendations.

These essays involve fusion analytics and hybrid system design in a cycle that encompasses theoretical arguments, econometric analysis of big data, and construction of a software application. This dissertation contributes to understanding the new channels for music popularity and diffusion over time, and also paves the way for promoting music in social network scenarios in ways that go beyond traditional music recommendation.

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Professor Robert J. Kauffman, for his patient and continuous guidance, encouragement, and support throughout my Ph.D. study. He is always there whenever I need his help. I appreciate all the time he has spent on discussing my research ideas, editing my papers, helping me to practice my presentation skills, improving my posters and slides, and also sharing his personal experience in doing research and how to be a trustworthy researcher and colleague. I am extremely fortunate to have him as an advisor as he is cultivating my talents in many ways beyond just being a researcher and a teacher.

I would also like to thank Professors Zhiling Guo and Qiuhong Wang, and Dr. David R. King for being my committee members and giving valuable advice, guidance, and comments that were essential in shaping my dissertation well and for my future career. They inspired me in my research from various perspectives, especially on how to combine academic research with real world applications. I really appreciate their involvement.

I also would like to thank Professor Jialie Shen for his kindness in introducing me to the music research realm when I started my Ph.D. study. I am still a junior researcher, but this interesting research area he led me into will be a topic I would like to keep working on in my research career. In addition, I appreciate insightful discussions and general support from Professors Baihua Zheng, Feida Zhu, Ee-Peng Lim, David Lo, Mei Lin, Qian Tang, and my Ph.D. colleagues, Emmy Hoang Ai Phuong, Zhiyong Cheng, Kustini Lim-Wavde, Dan Geng, Zhiyuan Gao, Felicia Natali, and Deserina Sulaeman.

Last but not the least, I am indebted to my husband Li Deng and our dear parents

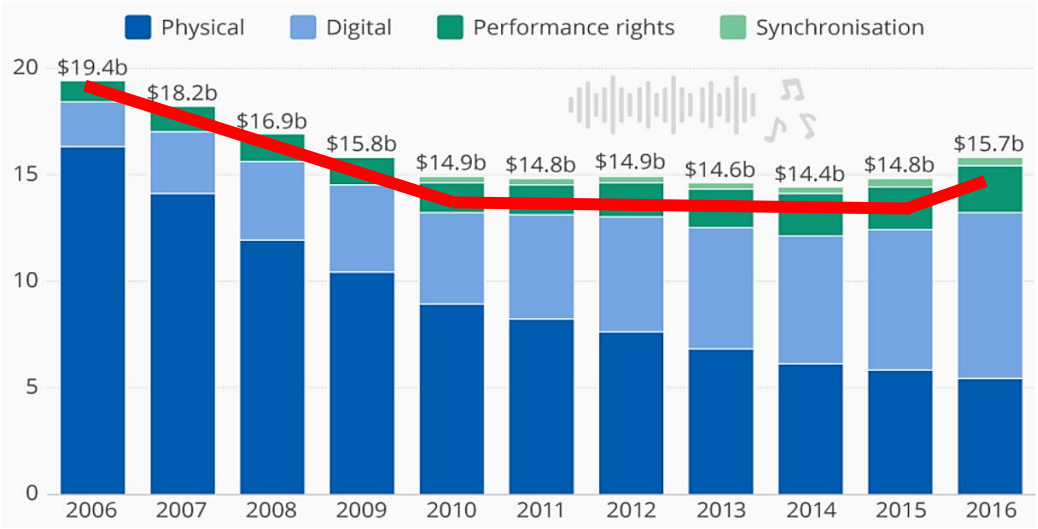
for always being on my side to lift my spirit. Their unconditional love, understanding, and emotional support have helped me get through every challenge along the winding road of Ph.D. study. I wish them to be forever happy and healthy.

To Li Deng and Our Dear Parents

Chapter 1. Introduction

The rapid evolution of contemporary digital entertainment and Internet technology has dramatically changed the way people produce and consume music. Today, a larger array of music is being enjoyed by more people in more ways than ever before, through traditional music recordings, mass media (radio, TV), digital streaming (Spotify, YouTube, Last.fm), live performances (music shows, concerts), and so on. Since 2004, digital music, such as digital downloads, paid subscriptions for music content access, and on-demand streaming music, has gradually chopped up the market share of physical music, including CDs, DVDs, LPs/EPs, cassettes and vinyl (RIAA 2016). The 2017 Global Music Report (IFPI 2017) suggested that, by the end of 2016, physical album sales sunk to a new historical low, while the gross annual revenue was increasing, turning around the decreasing trend that had occurred since the beginning of 21st century (see Figure 1.1.)

Figure 1.1. Global Recorded Music Industry Revenues (US$ bn) (IFPI 2017)



Although people are less likely to buy CDs now, they nevertheless are listening more than ever to downloadable digital music and streaming music. Digital music's sales volume today is over 10 times that of physical music in the U.S., according to

2016 data. In addition, Nielsen's (2017) U.S. music mid-year report for 2017 indicated a 62% increase in on-demand streaming audio listening compared to the same time in 2016. Digital music consumption via PCs and mobile phones has been gradually divvying up the music market and offsetting physical music sales.

One listening mechanism of digital music is through digital download tools such as Apple Music and Google Play Music. People now can choose between purchasing downloadable content and acquiring it for free (usually with ads), so they can use it on their local devices and enjoy it at any time. The other listening mechanism is streaming music services, such as Spotify, Pandora, Last.fm and YouTube, amongst others. Different from digital downloading customers, streaming music consumers have no need to download digital files to local devices, but only need to subscribe monthly or annually – or they can just enjoy the music for free whenever they have access to WiFi or a 3G/4G connection. Music collections are now so large that the limited storage of PCs and mobile devices, and the increasingly mature Internet environment, make it so that more and more people have switched to streaming music. The revenues for digital downloads in 2016, as a result, dropped to 60% of the 2012 peak in the U.S., while streaming revenues in 2016 doubled from those in 2015, and now exceed digital downloads (RIAA 2016). Increases in streaming music services have been continuing, with revenues accounting for 65% of the total music market in 2017, while with digital downloads only achieving a declining 19% share (RIAA 2017).

Streaming music services are making the depth and richness of every kind of music available to hundreds of millions of people. An important advantage is their close coordination with social networks, which are connecting artists with consum-

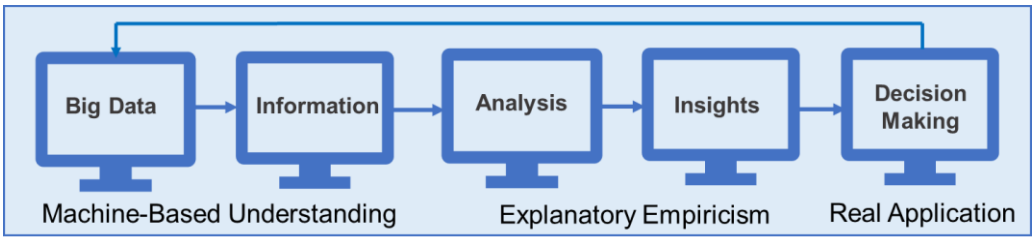
ers, and consumers with consumers in a more directly and faster ways. More interesting is that the "networks" are still functioning as open channels to the whole Internet and other mass media (Garg et al. 2011, Schedl 2011). Internal connections and external information go hand-in-hand today and affect music listening. In this context, rich music information, massive stores of music, diverse user listening behavior, and complex listening environments have brought challenges and opportunities on how to design and conduct market strategy for music information providers and platform service suppliers. These changes in streaming music consumption and promotion call for research on understanding the mechanisms of content consumption, the impact of social networks on consumer behavior and product sales, and strategy for product promotion and service design (Salo et al. 2013, IFPI 2012, 2013, 2015, 2017).

Multiple streaming music-related research topics have been attracting attention from researchers in various academic fields, including Information Systems (IS), Computer Science (CS), Economics, and Social Science, as well as industry professionals. The research has been trying to observe, explain, model, and predict the activities and changes in streaming music and consumer behavior from various perspectives. The research interests cover music popularity and value prediction (Karydis et al. 2016, Kim et al. 2014), music diffusion (Garg et al. 2011, Pálovics and Benczúr 2015), music retrieval (Skowron et al. 2017), music recommendation (Cheng and Shen 2016, Tan et al. 2011), music promotion (Scharff 2015), and the profitability and operation of streaming music services (Salo et al. 2013, Waldfogel 2015). Most of the existing research has studied specific topics in various areas, and has ignored the potential links among them. Very few works have tried to under-

stand streaming music services as covering the music industry, its artists, and consumers. For example, studies on music recommendation usually have focused on consumer satisfaction only, and have omitted addressing a key requirement from the music industry about artist promotion. The incomplete analyses may result in biases in understanding the issues, hide the potential business value, and fail to make the real nature of communication between music and its consumers more fully understood.

The availability of proprietary corporate and online public data and the development of innovative technology methods now allow us to explore about streaming music in more detail from a new and more complete perspective with the help of various data analytics methods. It links analysis and applications in the social media market by using a *fusion analytics framework* (Figure 1.2). leverage machine-based methods to transform *big data* (on social media and user behavior) into meaningful *information*. Then use *analysis* methods (econometrics, explanatory empiricism) to analyze and understand what happened in the process. Last, the learned *Insights* can be used to assist *decision-making* in real application setting.

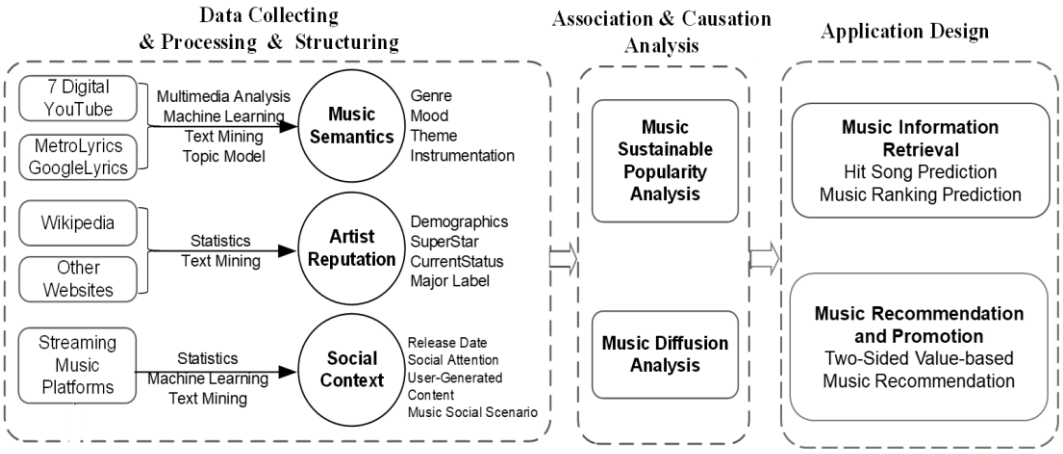
Figure 1.2. Fusion Analytics Framework (Kauffman et al. 2017)



My dissertation research applies the fusion analytics strategy, and examines a streaming music service from a new and effective perspective by considering the music industry, its artists, the music content and consumers as an ecosystem of interacting entities and stakeholders. It aims to construct an explanatory cycle of in-

sights, ranging from understanding the business value of music, to how music diffusion works in social networks, to how music information can be used to promote and recommend music products in streaming music settings. The three research essays lie in the interdisciplinary area of information, technology, and business sustainability and value in the music industry. These things involve data analytics, econometrics, data mining, social media analysis, recommender system design, and other methodologies for understanding and improving current and future industry capabilities. The detailed framework for the research is shown in Figure 1.3.

Figure 1.3. Content Framework for My Dissertation Research



Essay 1 studies streaming music popularity and offers new metrics for the study of popularity patterns and sustainability (Ren and Kauffman 2017, Ren et al. 2016). In the context of streaming music in music social networks, I have been exploring how music popularity develops and is sustained over time. This offers the potential to supply useful hints to music label firms, such as what type of music is more likely to become popular, and how long a new music track will match consumer listening tastes. These insights have the potential to assist the music industry to craft their promotions and make investment decisions more effectively.

Music is a *durable information good*, which may be popular for decades, or for

just several weeks. Based on observation of the actual popularity performance of music in the marketplace, I have been tracking the performance of music tracks and the related artists, from when the tracks' release occurs to the first time they enter the top-ranking chart, and then later when they drop off the top-ranking chart. I propose two popularity metrics to gauge the speed of music in achieving these kinds of top recognition, and how they sustain their popularity over time. To estimate the potential determinants and their influence, Essay 1 leverages machine-based methods to build a new *music semantics construct* and a variety of other track, artist and market descriptors for each music track. It also explores and predicts popularity by using econometrics and machine-learning models to assess how the track popularity duration and patterns are developing on the top-ranking charts.

This essay utilizes a large dataset that contains 78,000+ track-ranking observations and constructs over a 10-year period. I found that it is possible to explain the popularity duration and the weekly ranking of streaming music tracks. This research emphasizes the power of data analytics for knowledge discovery, and how explanations can be achieved with a combination of machine-based and econometrics-based approaches.

Essay 2 does a more detailed exploration into the development of music artists popularity and analyses the diffusion of streaming music over time (Ren and Kauffman 2018, Ren et al. 2014). Academic researchers and music industry marketers have attempted to figure out how music social networks operate, and how they can be used to effectively promote music. The insights from my research are critical for the design and implementation of marketing plans, to maximize the market value of the music and the artist. Also, understanding the diffusion of music can assist social networks to improve their services to retain a larger number of loyal and

satisfied users, through the provision of better music recommendations.

There are two different ways a piece of music reaches a listener on a streaming service. It may reach them through their connections in a social network (*social influence*), as well as through the influence of external sources (*information discovery*), such as mass media, newspapers, and other social networks. Most current work on music diffusion in social networks only considers the social influence, but the availability of massive online data now allows us to measure the effects of external influences on streaming music diffusion in more detail. Essay 2 uses listening data from a music social network for over 550,000 users to examine the effects of external information on streaming music diffusion at the macro-level and micro-level. Difference-in-differences and count data models are employed to assess the effects of new externally-released information on the diffusion of streaming music.

This study found that external information has a significant impact on an artist's music diffusion in a network, and the impact and persistence of popularity are related to artist information. This effect is confirmed at both the macro- and micro-levels. My study also found that a listener's geographic location limits the diffusion of an artist's music. Although people can access whatever music they like, their more limited access to external information may constrain their attention. This study offers managerial insights that ought to be useful for music promotion and personalized recommendations in online music platforms by combining the strength of multiple information channels.

Essay 3 proposes a streaming music recommendation algorithm based on the insights learned from music popularity and music diffusion analysis. The two main stakeholders of a music recommender system are the music industry – including social networks – and consumers. The music industry would like to promote its

music products to more consumers with lower investment costs, while consumers would like to get more accurate and novel music recommendations while controlling their search costs. Most of the existing recommender systems focus on the consumer level and have tried to improve recommendation accuracy, but they mostly have ignored the music industry's investment returns.

This essay develops a *two-sided value-based streaming music recommender algorithm*. The system combines the business value of the music artist's products and the utility of consumers into an integrated model. For the music industry, the system seeks to increase the conversion rate of potential listeners (Zhang et al. 2016). For consumers, the system aims to improve recommendation accuracy. The system design involves a new recommendation algorithm, leveraging the insights that were obtained in Essays 1 and 2 regarding the determinants of music popularity and diffusion.

Chapters 2, 3 and 4 discuss the related literature on interdisciplinary learning for streaming music and the three research essays. Chapter 5 delves into my experience as a Ph.D. student and the essential skills that I obtained during my doctoral research. Chapter 6 concludes with contributions, limitations, and future research.

Chapter 2. Understanding Music Popularity in Music Social Networks [1]

2.1. Introduction

With contemporary digital entertainment, people can easily access large music collections and stream content via social networks such as Last.fm, Spotify and YouTube. Streaming music and social networks have changed listener behavior dramatically. They can "listen," "like," and "comment" on music tracks, and communicate with and affect other listeners through social communication. In comparison to an album or a radio broadcast, listeners can make much richer selections also. They can listen to tracks repetitively or freely shift to other content.

Music is a *durable information good* that can bring utility to listeners and value to artists based on traditional music album sales (Bulow 1982, Poddar 2006). One on-going work by Hiller and Walter (2017) pointed out that streaming music services are changing this pattern. They found that the music industry encouraged musicians to release fewer and higher quality songs, leading to increased market demand and listeners, because people prefer to listen to individual songs but not buy an album.

In fact, when we explored the development of top-ranked musicians, we saw that even one strong and widely-appreciated song can lead to the rise of a new music superstar, such as "Rolling in the Deep" for Adele, or "Poker Face" for Lady Gaga. Moreover, a classic track can make people remember the singer, even many years after its release. Examples include "Hey Jude" by The Beatles, which Billboard named the tenth most popular song of all time in 2012, although it was first released

[1] An earlier published version of this work in a conference proceedings can be found in Ren and Kauffman (2017). Changes have been made to be responsive to my faculty committee's input.

in 1968 (Bronson 2012). Great business value for music and musicians is the natural outcome. *Forbes* (2017) has reported that Beyoncé's net worth was around $350 million in June 2017, and Adele's around $135 million. The music labels have paid attention to how music can be promoted by social networks to maximize its market value (IFPI 2012, 2013, 2015, 2017).

Researchers and industry pros have been exploring the ingredients for music to achieve sustainable popularity (Chon et al. 2006, Karydis et al. 2016, Nunes et al. 2015). It is possible to explain how a song became popular, and to predict future music superstars. Most have considered music and artist factors, or market and social factors (e.g., Bischoff et al. 2009, Koenigstein and Shavitt 2009). The present work aims to determine how effective music promotion investment activities will be, based on analyzing the popularity performance of a large set of tracks since their release in social media.

In this research, we focus on the analysis of music track popularity, and explore the following questions: (1) In a music social network, what factors produce a popular track? (2) Is the music content most important? (3) Can a song's popularity duration be predicted, based on hidden factors? (4) How much does the social context for a track affect the duration of its popularity? (5) And are there discernible popularity patterns for music tracks that are suggested by our research inquiry? To answer these research questions about music popularity, this research applies *computational social science* methods that combine machine-based methods for data analytics from CS and explanatory methods from IS research (Kauffman et al. 2017, Li et al. 2017, Chang et al. 2014, Chen et al. 2012). This permits a researcher to capture and analyze different kinds of data that would not be possible using non-machine methods, secondary datasets, or interviews.

We empirically examine the research questions using 78,000+ track ranking observations over 10 years collected from a streaming music service. We define two new measures, *Duration* and *Time2TopRank*, for music track popularity that consider the lifespan of a song from first release to top-chart popularity to chart drop-off. To better assess the factors' effect on music track popularity, for each music track, we assess a relatively complete construct covering musical and non-musical components that describe the social and market aspects of the track (music semantics, artist reputation, and social context). We further implement machine-based CS and Social Science methods to understand and predict track popularity and ranking based on the music track constructs that we have proposed.

We found that music track popularity is explainable and predictable in the early stage after its release. Music semantics, artist reputation and social context all have impact on the music popularity development, although the effect strengths are different from each other. Different genres of music have different popularity performance, including their speed to top-rank, and how long time they can remain at that level of popularity. The prediction work that we have done shows that artist reputation and social context information are important for whether a track can quickly gain enough attention and rise to the top-rank chart in a short period after its release. After it has reached the top-rank chart, music semantics can help to improve an analyst's prediction for how long time it will stay there. We also observe heterogeneous popularity patterns and the weekly rankings of music tracks, which can also be predicted through our use of the music constructs we propose.

This research emphasizes the power of data analytics for knowledge discovery and explanation that can be achieved with a combination of machine- and econometrics-based approaches. It contributes to the literature on music track popularity

in social networks scenario in several ways. We construct a relatively complete descriptive vector for each music track by leveraging machine-based methods. Our study supplements these factors (learned through non-machine methods, from secondary datasets, and via interviews) with more fine-grained music semantics to provide fuller information about the drivers of music track popularity. Also, we focus on the whole lifespan of a music track development, this can provide a more completed description to understand the popularity development. Finally, our findings on music popularity prediction provide useful insights for the music industry, which can assist the music labels to assess the potential popularity of a new track in its early stage, and further adjust their promotion strategy for music markets.

The remainder of this chapter is organized as follows. In Section 2, we review the literature related to music track popularity in social networks. Section 3 gives some new definitions of music track popularity and presents the empirical model. Section 4 describes the research context, data collection and variables in the empirical analysis. Our explanatory and predictive analysis are laid out in Section 5 and Section 6. Section 7 discusses the findings and draws conclusions of managerial interest. Section 8 concludes.

2.2. Theory and Literature Review

Music popularity analytics have attracted wide attention in multiple research fields, covering IS, CS, Society Science, and Psychology. We discuss and summarize the related literature from two perspectives: track popularity definitions and the related research methods.

2.2.1. What Is Music Track Popularity?

There are various ways to define the popularity of a music track. They include:

sales volume; the amount of audience listening that occurs via streaming music services; track performance on top-rank charts; and music industry awards received. No single standard to define popularity is recognized in the literature.

Most of the research on music popularity has been based on data from public sources (Chon et al. 2006, Herremans and Sörensen 2014). Some studied Billboard rankings (Karydis et al. 2016, Lee and Lee 2015, Nunes et al. 2015, Singhi and Brown 2015). Others chose rankings like UKTopChart, or streaming music services, such as Last.fm, Spotify, and Twitter (Dhanaraj and Logan, 2005, Kim et al. 2014, Pachet and Roy 2008). Some have observed music tracks' performance since they reached top-chart ranking (Frieler et al. 2015, Karydis et al. 2016, Ni et al. 2011). Many used a binary variable to define popular or non-popular music track, based on chart ranking at a point in time. For example, if a track reaches the Top-1 rank, it was labeled as popular; if it never climbed above Top 90, it was not popular (Nunes and Ordanini 2014). Lee and Lee (2015) explored various definitions of popularity, for chart performance based on the chart debut position, total weeks on the chart, and so on.

All of them focused on just one stage of a track's developing popularity: after it reached top-chart ranking. This reflects a bias for understanding how a track's popularity developed: it missed the stage of run-up to top-chart ranking and the long-term performance on the top-chart ranking. Some of the popular tracks reached the top or even Top 10 ranking immediately after their release, such as "Bad Romance" for Lady Gaga. Some others may spend a long time till first appearing in the top-chart ranking. They still may become very popular though, such as "Little Lion Man" from Mumford & Sons. So how to properly and completely define the popularity of a music track is still a challenge. In this research, we offer a relatively

complete measurement approach to music track popularity, from the release to ascension to the top-chart rankings till drop-off from the top-chart ranking.

2.2.2. Music Track Popularity Explanatory and Predictive Analysis

In Statistics, Econometrics, and IS, the link between explanatory analysis and prediction is common; they are useful methods expanding scientific knowledge and industry applications (Shmueli 2010). Simon (2001) distinguished between *basic science* and *applied science.* They are analogous to the difference between explanation and prediction. He pointed out that the former is aimed at knowing and understanding, to describe and explain the world. The latter is aimed at finding out unknown values of variables based on other known values, to make inferences or predictions about the world.

In general, explanatory analysis assesses causation between the independent variables and a dependent variable through the use of a model and a causation-focused research design. Prediction focuses on using the possible association between the two kinds of variables to improve predictive accuracy (Shmueli and Koppius 2011). Sometimes in real applications though, a good predictive model somehow is not explainable, like using a *convolutional neural network* to design a system for music recommendation. The output that the neural network produces is not so easily explained. Industry practice usually ignores whether the results can be explained; managers focus on good performance only. Although recently researches have started to pay attention on the explanatory side of an algorithm, such as explainable recommendation (Zhang 2017).

We next review the existing literature on explanatory analysis and predictive analysis of music track popularity, based on the differences that characterize them.

Explanatory analysis. Related research in IS has tried to figure out what are

the important factors for a music track or music album popularity. Some authors have assessed album popularity, including for those released during Christmas, and the impact of release timing on their success (Bhattacharjee et al. 2007). Other things that promote album popularity are highly correlated with artist reputation and superstardom (Chung and Cox 1994, Hamlen, 1991), label association, and the debut rank on Billboard (Strobl and Tucker 2000). Nunes and Ordanini (2014) tested the relationship between instrumentation combinations and the probability of high versus low-ranked tracks. Nunes et al. (2015) explored how a song's chorus lyrics affected how fast a Billboard Hot 100 song reached the Top-1 rank.

To assess music popularity, existing IS work has adopted two types of estimation models: logit regression and survival modeling (Bhattacharjee et al. 2007, Nunes and Ordanini 2014, Nunes et al. 2015, Strobl and Tucker 2000). Logit regression is used to test whether, or the probability of, a track or an album is high-ranked. For example, Nunes et al. (2014, 2015) utilized logit regression to estimate the relationship between lexical repetition, instrumentation, and the likelihood of being a top-ranked song. Survival models are used to gauge how long time a track or an album continues to be popular. Different survival functions are used based on different research perspectives. For example, Strobl and Tucker (2000) used a Kaplan-Meyer survival function to estimate album chart survival relative to the skewness of chart success. Bhattacharjee et al. (2007) estimated a Weibull survival function to assess album chart popularity under the impact of digital sharing technology. In general, models for music popularity are selected based on the music popularity measurement approach and research target.

Predictive analysis. Various CS authors have tried to find different combinations of musical and non-musical features to increase the accuracy of popular and

non-popular track prediction. They have used machine-based methods to extract feature sets for prediction, such as acoustic features (Borg and Hokkanen 2011, Herremans and Sörensen 2014, Frieler et al. 2015), social information (Koenigstein and Shavitt 2009, Schedl 2011, Kim et al. 2014), lyrics plus acoustic features (Dhanaraj and Logan, 2005, Singhi and Brown 2015), and acoustic features plus early stage popularity (Lee and Lee 2015). The prediction methods that have been used include *support vector machine* (SVM), *random forest* (RF), Bayesian network analysis, and so on. Most obtained no more than 67% in predictive accuracy. The best performance was achieved by Kim et al. (2014), with 92% accuracy for Top 10 song prediction, but with a limited dataset of 168 tracks over 10 weeks. No general conclusions were able to be drawn.

No matter whether it is for IS or CS, explanatory analysis or predictive analysis, a common challenge is that it is hard to compare the performance of different factor sets: there has been no standard dataset. Karydis et al. (2016) was the first work to construct a sharable musical track popularity dataset. This dataset covers 10 years of music ranking data from Last.fm, Spotify and Billboard. For each track in the dataset, its artist, album, acoustic features, ranking in the three charts, and similar tracks are included. And yet, other research has broadly shown that track popularity, especially in the social environment, cannot be explained or predicted by these attributes alone. Music, artist, and social context are the three key perspectives we cannot omit in music track popularity research.

In this research, we constructed a relatively complete music track popularity dataset, by integrating three key elements: music, artist and social context information. And implemented the explanatory and predictive analysis on the dataset to learn about music track popularity in the music social networks, we would like to

say whether the music construct vector we proposed can explain and predict the music track popularity.

2.3. A Model for Music Popularity Duration in a Social Network

2.3.1. Music Popularity Measurement

Music streaming services, such as Last.fm and Spotify, integrate music listening, social network activities, and social recommendation into a single platform. Listeners of music streaming services, no longer just listen to music tracks, they communicate with and affect other listeners through liking and commenting on specific tracks. In comparison to an album by one artist or a radio broadcast, listeners can make much richer and more colorful selections. They also can keep listening to a track repetitively or shift to other songs in different genres, by varied artists, and even in multiple languages more freely. Because of their different approaches as listeners in this setting, music in social networks has more staying power to achieve popularity over time and appeal to its audience's tastes. Therefore, music social networks are appropriate for a study that seeks to understand music popularity development in a "small society" context. Compared to some public music ranking charts such as Billboard and UKTopChart, streaming services record the listening logs for each track over time and rank their weekly listening time by streams. In addition, music streaming services have been shown to be good proxies for a music track's ranking based on their high correlation with Billboard.[2]

This research leverages the record of music track listening to investigate the development of music popularity in Last.fm. We focus on the full lifespan of a track

[2] Koenigstein and Shavitt (2009) and Kim et al. (2014) reported on the strong correlation between song popularity on Billboard's list and the extent of social media activity related to it in P2P networks and instant listening on Twitter. Schedl (2011) also offered evidence for high correlation between an artist's popularity on Twitter and the artist's ranking in Last.fm.

from its release, to when it reached a top-ranking on the chart, all the way until it dropped off and was no longer popular. Two related measures are:

- *Time2TopRank*: Total weeks from a song's release date to the first date it reached a top-chart ranking. It shows how long it took for a song to get enough attention to reach a top ranking.
- *Duration*: The total weeks a song appeared in the top-chart ranking for popularity. It suggests how long a song matched people's tastes and was highly rated on Last.fm and Billboard.

The measures describe the speed for achieving top recognition, and popularity sustainability over time.

2.3.2. Music Track Popularity Duration Model

A duration model is used to estimate when success in reaching a top-rank occurs (*Time2TopRank*), and when top-rank drop-off occurs (*Duration*). A *hazard function* specifies the duration until time *t* when this event happens. A *proportional hazard* (PH) *model* for this setting is:

$$\lambda\,(t \mid \boldsymbol{X_i}) = \lambda_0(t)\ exp\,(X_1\,\beta_1{}^{PH} + X_2\,\beta_2{}^{PH} + \ldots) = \lambda_0(t)\ exp\,(\boldsymbol{X_i} \cdot \boldsymbol{B}) \qquad (1)$$

Here, $\lambda(t)$ is *Time2TopRank* or *Duration.* $\lambda_0(t)$ is the *baseline hazard*, which represents the hazard value when all of the $\boldsymbol{X}_i$ are equal to zero. $\boldsymbol{X}_i$ are explanatory variables for a track i (i = 1, 2, …), and βi^{PH} are parameters to be estimated for all the data to gauge if there are modifications to the hazard rate of top-chart ascension or drop-off due to their influences (Kleinbaum and Klein, 2006). $\lambda(t)$ is the product of the baseline hazard $\lambda_0(t)$ and the exponential function of the linear combination of the explanatory variables $\boldsymbol{X}_i$. Thus, $\boldsymbol{X}_i$ have a multiplicative or proportional effect on the predicted $\lambda(t)$.

There are multiple functions for the different distributions of duration. A

Weibull hazard function $\lambda(t)$ for duration follows a monotonic curve, $\lambda(t) = \lambda z t^{z-1}$. In this model, λ is a scale parameter, and z is a shape parameter. The z value makes it so the hazard function can be constant, or steeply declining or increasing at an accelerating rate. This also fits situations in healthcare, finance, marketing and e-commerce.

Other distributions are non-monotonic, such as the log-logistic hazard, with $\lambda(t) = \lambda z t^{z-1}/(1 + \lambda t^{z})$: it decreases after peaking. It captures the dynamics of situations that involve an initially increasing and later decreasing hazard rate, as with the diagnosis and treatment of leukemia and cancer. In contrast, the log-normal distribution follows a normal distribution for the hazard function, which is positive and skewed with a lower mean and higher variance for the event timing. This distribution is often used in finance, so price observations that are less than the mean are not so extreme. It has also been applied in medicine to understand the occurrence of chest pain and the subsequent onset of heart disease (Hussain et al. 2014).

This research considers these three hazard function models for the analysis of track popularity *Time2TopRank* and *Duration*. A linear model estimated with *ordinary least squares* (OLS) (Bhattacharjee et al. 2007) is also considered. With a log-transformation, this approximates the more refined hazard models, and can act as a baseline for estimation. Time-invariant musical constructs, such as genre and mood, are used also. Non-musical constructs, such as the artist reputation and social context, are time-varying in contrast. By including fixed and time-varying covariates for each track, the general vector form of this model is:

$$\lambda(t) = f(\lambda, z, t; \boldsymbol{X}_{Music}\boldsymbol{B}_{Music}^{OLS}, \boldsymbol{X}_{Artist}\boldsymbol{B}_{Artist}^{OLS}, \boldsymbol{X}_{Social}\boldsymbol{B}_{Social}^{OLS}) \quad (2)$$

2.4. Research Setting, Dataset and Machine-Based Data Extraction

2.4.1. Research Setting

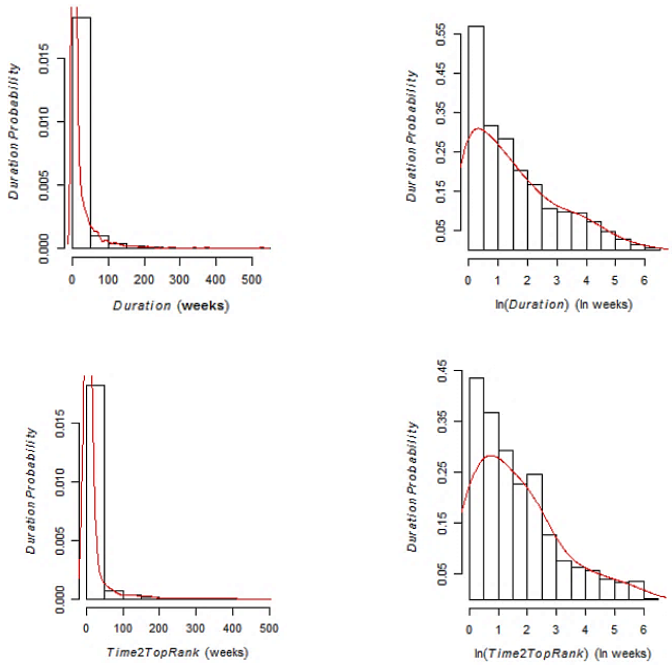
Spotify and Last.fm are two widely-adopted music streaming services. Both of them offer PC and mobile phone access. And both have a *scrobble function*, which connects a user's listening profile to other music-streaming services. This function links several music-streaming services to Last.fm and Spotify, such as Pandora Radio, iTunes, Windows Media Player, and Deezer, and supports the tracking of complete listening trends over time. Spotify has a listening limit for free-access use, while Last.fm has essentially unlimited listening. Flacy (2012) quoted Spotify's new terms of service in January 2012, when the free trials started to expire. It could be "accessed as an ad-supported free-to-the-user service having no monthly cap on listening hours or a cap on number of plays of a unique track during the first 6 months following creation of your Spotify account, but thereafter a cap of 10 listening hours per month and a cap of 5 plays per unique track." Last.fm users were less limited: to 1 million songs for listening in total, and around 3,000 songs a day free-to-the-users.[3] In addition, since our research studies a setting in which social sharing, comments and interaction are unconstrained, Last.fm is a better choice.

Last.fm puts out a Weekly Listening Chart based on its users' activities. It reports on the top-150 music tracks each week. For the 10 years of data, the track popularity *Duration* variable in Last.fm was 44.2% correlated with the *Song Popularity* duration variable for the Billboard Hot 100, as well as 34.3% correlated with Billboard's Streaming Songs, based on Spotify's data. This helps to verify Last.fm as a representative source of track popularity data, though some data were omitted

[3] Last.fm's and Spotify's Terms and Conditions of Use have been changing over time (Last.fm 2015, Spotify 2017). As the dataset used in this research is for 2005-2015, so the one with fewer limitations during this period is more suitable for this study.

due to imprecise song names that were hard to match across the services. We also collected a ranking dataset from February 2005 to May 2015 from Last.fm. This yielded 532 weeks and 12+ million streaming music tracks. Relatively few made it to the top-150 chart ranking though: only 4,410 tracks or 0.04% of the total.

Figure 2.1. Raw and Logarithmic Distributions of Popularity *Duration* and *Time2TopRank* (Weeks) for Music Tracks



Notes. The left column shows raw popularity *Duration* and *Time2TopRank* for tracks in weeks. The right column shows probability densities for *ln(Duration)* and *ln(Time2TopRank)* for the entire data set (obs. = 4,410)

Two popularity measures – *Time2TopRank* and *Duration* – were obtained for each track. Figure 2.1 shows the probability densities of *Duration* and *Time2Top-Rank* for the tracks in this study. The first column gives the distributions of raw values, which have positive skewness (6.07 for *Duration*, 7.00 for *Time2TopRank*) and large kurtosis (55.56, 68.59). We observe a long-tail distribution, with over 80% of the tracks appearing in the ranking for less than 18 weeks. We further show the distribution of ln(*Duration*) and ln(*Time2TopRank*) in the right column in Figure 2.1. We note that these are like a Weibull distribution, with positive skewness

of 0.89 for ln(*Duration*), 0.80 for ln(*Time2TopRank*), and kurtosis of 2.97 for ln(*Duration*), 3.28 for ln(*Time2TopRank*).

Some tracks have left-censored observations regarding their top-rank chart durations because they were released before February 2005. Bob Dylan's "The Times They Are a Changin" was released in 1964, but only reached the top-rank chart for Last.fm in March 2009, for instance. Right-censored observations of top-rank chart popularity include those that were popular across the 2005 to 2015 observation window and are still popular, such as Oasis' "Wonderwall" (released in 1995) and Coldplay's "The Scientist" (released in 2002), and they remained highly popular beyond 2015. For our empirical analysis of track popularity-related top-rank chart duration, we remove all censored music track data, including 421 left-censored and 108 right-censored tracks. Overall, we used 3,881 tracks by 477 music artists for our analysis. The bottom right of Figure 2.1 shows the logarithmic distribution for *Duration* and *Time2TopRank* (see Table 2.1.) We also give the distribution of our censored data on three different selected hazard functions in Appendix Figures A1 and A2 (Weibull, log-logistic, and log-normal distributions).

Table 2.1. *Duration, Time2TopRank* in Weeks for All Observations and Without Censored Observations

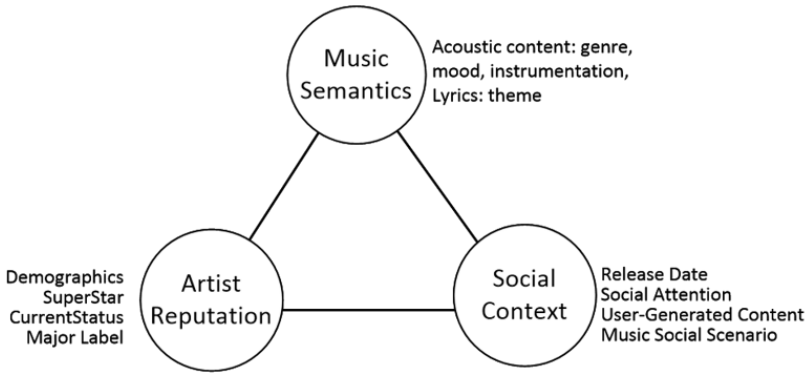
DATASET (ALL OBS.)	MIN	MAX	MEAN (SD)	1ST QUARTILE VALUE	MEDIAN	3RD QUARTILE VALUE
Duration	1	532	17.9 (47.2)	1	3	11
Time2TopRank	1	473	20.1 (53.7)	2	4	10
DATASET (WITHOUT CENSORED OBS.)						
Duration	1	504	13.1 (31.6)	1	3	9
Time2TopRank	1	395	11.4 (27.7)	2	3	9

Notes. Dataset with all obs.: 4,410 tracks, 550 music artists; dataset without censored obs.: 3,881 tracks, 477 music artists; values in weeks. Study period: February 2005 to May 2015, 532 weeks. In addition to the minimum, maximum and mean values of dependent variables' long-tail distributions, also included are the quartile values of the distributions.

Log-transformed *Duration* and *Time2TopRank* in weeks were used to measure the track's popularity in Last.fm. In the social network environment, three kinds of

constructs are relevant: (1) track semantics, acoustics and lyrics; (2) artist reputation and profile; and (3) social context. These were extracted from multiple sources. Through the measures associated with this musical construct vector (MCV), it is possible to assess how they affect music popularity (see Figure 2.2).

Figure 2.2. Drivers of Track Popularity in a Music Social Network Setting



2.4.2. Musical Construct Vector (MCV)

Music semantics. A music track has two components: *acoustic content* and *lyrics*. The content can be characterized as a *musical construct vector* (MCV), with the Theme, Mood, Instrumental, and Genre reflecting how acoustic content is perceived. High-level semantics can be extracted from lower-level musical features, such as timbre, rhythm, and tempo (Kim et al. 2010). Machine-based methods were used in this research to extract the music semantics.

*Acoustic content***.** For each track, a 30-second sample was collected from 7Digital or YouTube. A music track usually has an Introduction, Verse, Chorus, Bridge, and Conclusion. The Chorus is the key element of a track, and its music and lyrics are repeated. It is almost always of greater musical and emotional intensity than other structures in the track. 7Digital supplies 30-second samples for listeners to decide whether they would like to pay for an entire track. By 7Digital's design, most of these samples include the Chorus, while some offer Verse content in their 30-second clips. Our approach with 30-second samples of tracks is similar for

downloads from YouTube. We manually chose the Chorus tracks for analysis.

A four-step method was used to learn the constructs and implement filtering (Cheng and Shen, 2016):

- **Step 1.** *Segment music tracks* into clips of 1 to 5 seconds in length.
- **Step 2.** *Extract acoustic features* to identify a multi-dimensional low-level acoustic feature vector for all clips (Janani et al., 2012), via: *spectral features* (70 dim.); *timbral features* (23 dim.); *rhythmic features* (12 dim.); and *temporal feature* (62 dim.).
- **Step 3.** *Estimate musical construct probabilities*, based on track tags statistics for 18 genres on Last.fm, 12 types of instrumentation (Zhang et al., 2009), and 5 moods that were selected from the MIREX mood classification (Napiorkowski, 2015) for learning in the musical construct models (see Table 2.2). Although Last.fm offers well-defined categories for user tagging tracks, but the most are genre related only, and the tags are noisy. There are spelling errors and incorrectly applied labels. And, tagging in Last.fm tends to lack appropriate balance. So popular tracks tend to have more tags, less popular tracks less so, and niche tracks may have none. To make sure each track had proper musical semantic tags, machined-based methods were used to label them on genre, instrumentation and moods.

  100 labeled tracks were selected per subconstruct to train a multi-state vector model for each construct. An SVM with a Gaussian *radial basis function kernel* was trained on 80% of randomly-selected, labeled clips of tracks, and tested on the remaining 20% with 10 repetitions. SVM is a discrimination algorithm that classifies a subset of data introduced to it, creating a separating multidimensional hyperplane in the process, to categorize the other

remaining data. The theory behind this method is that it provides a means to construct a linear decision boundary between different classes of data, such that the margin or distance between them is maximized (Kecman et al. 2005).[4]

Five segmentation sets with lengths of 1 to 5 seconds were explored. Among then, 2-second clips were most effective for 53,296 clips, with prediction accuracies for: Genre – 70.5%; Instrumental – 85.6%; and Mood – 57.5%. The trained models were used to label each clip for tracks, using a 15 × 35 *acoustic* ***MCV*** *probability matrix*.

- **Step 4.** *Filtering of the learned constructs* resulted in only the useful ones being retained, while the noisy ones were cut. A 35-dimension acoustic **MCV** was produced for each track.

Table 2.2. Musical Constructs Used for the Machine-Based Content Analytics

CONSTRUCT	SUBCONSTRUCTS (VARIABLES)
Genre (18)	*Rock, Alternative, Indie, Pop, HipPop, Rap, Electronic, Metal, Folk, Soul, Blues, Country, R&B, Punk, Classic, Jazz, Experimental, Reggae*
Instrumental (12)	*Cello, Guitar, Drumkit, Violin, Piano, Tuba, Flute, Clarinet, Saxophone, Trombone, Trumpet, Snare*
Mood (5)	*Passionate, Lively, Brooding, Humorous, Intense*

Lyrics. They complement the acoustic content and give the artist's meaning behind the music (Hu et al., 2014). *Latent Dirichlet allocation* (LDA, Blei et al., 2003) was used to build a topic model to learn the semantic themes from the dataset of 4,410 tracks.[5] LDA exhibited effective performance for classifying topics in the text based on *document-word-topic* relationships it identified. The topic model was run by varying the number of topics from 3 to 15. The process identified 5 topics as

[4] The problem of *data sparseness* often arises with support vector machine-based learning. Training the learning algorithm is made more difficult due to the lack of a large enough number of instances for individual users (Cha et al. 2009, Li et al. 2015).

[5] Using machine learning methods to learn music semantics is unaffected by data censoring problems. Acoustic content and lyrics exhibit track-to-track variation, but the commonality is the time-invariant nature of any music track. Thus, we used the 4,410 tracks to learn the topic model.

providing the best summary, with LDA hyperparameters for the higher-level characteristics of $\alpha = 2.0$, and $\beta = 0.1$, which were established after 3,000 iterations. Table 2.3 shows the themes that emerged with representative words. About 65% of the tracks were about "love" and "life" (Themes 1, 2, and 4).

Table 2.3. Music Themes and the Representative Words for Each Theme Topic

TOPIC	MUSIC SEMANTICS THEMES	REPRESENTATIVE WORDS	# TRACKS (# IN SUBSET)
1	*Life, Dance, Passion*	We, like, dance, young, live, good, sweet, dream	589 (514)
2	*In Love, Relationships*	You, love, like, baby, wanna, need, girl, feel	967 (880)
3	*Soul*	Eyes, heart, soul, fall, cold, dark, blue, blood, left	1,041 (918)
4	*Sad Life, Love*	Back, alone, long, over, wrong, lost, leave, remember	1,290 (1,105)
5	*Anger, Hostility*	Like, fuck, shit, rock, bitch, fucking, hit, damn	523 (446)

Notes. The right-most column is the number of track with the labeled themes as their first-ranked theme. The numbers in parenthesis correspond to the number of tracks in the track popularity duration analysis dataset.

Artist reputation. The popularity of a track depends on who performs it to some extent, although other considerations may arise for some tracks and artists. Famous artists attract larger audiences. How to best measure reputation is open to debate though. The present research measures artist reputation, and leverages information on news on the Grammy, American and Billboard awards. Also relevant are their labels. Major labels have more resources to produce and promote high-quality tracks. This study covers 10 years, and 20 sub-labels associated with the 3 major labels that were considered.

Data were collected from Wikipedia and Billboard charts, and 8 dimensions were extracted and built:

- *Vocal*. Solo male, solo female, and group (3 dim.).
- *Major label*. Whether artist belonged to a major record label (1 dim.).
- *Pre-2005 reputation.* Times nominated or won award pre-2005 before a new

track (2 dim.).

- *Post-2004 reputation*. Times nominated or won award post-2005 before a new track (2 dim).

Social context. Last.fm had 59.2 million users in July 2015 when user growth plateaued. Its social environment is different from YouTube, Pinterest and Twitter. Users can "tag," "like," and "comment" on tracks and artists. Social comments offer a way to figure out what people are interested in and replace survey methods. Artists typically attract a group of followers as time passes, even when they are not famous. The social context subconstructs are as follows:

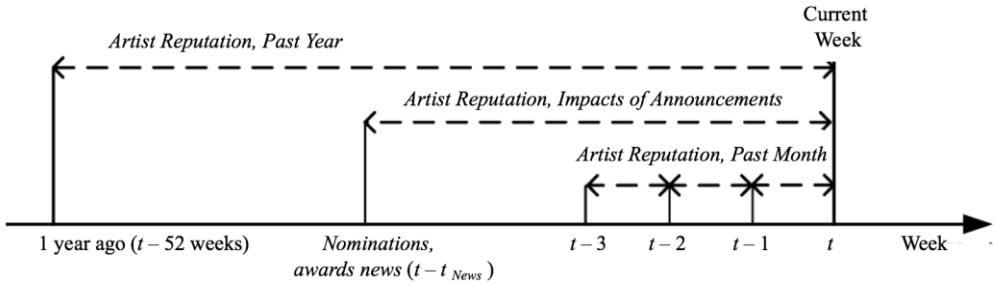
- *EarlyStageComments*. Cumulative comments since track release, time *t*. We observe the comments in first 5 weeks since the release. This is to assess whether early-stage attention has an impact on a track's future diffusion.
- *Top-rank before release*. If artist's track appeared in the top-rank chart in the top-50, top-100 or top-150 before a new track was released (for sensitivity analysis).
- *Holiday debut*. Binary variable for whether a music track released during the holidays, especially December in North America and Europe.
- *First top-chart rank.* The ranking when a track first reached a top-chart ranking on Last.fm.

For tracks that reached top-chart ranking, the median time *t* was 3 weeks, with a skewed distribution. Thus, it was appropriate to use several different periods to build an effective observation window. If the first few weeks of comments were sufficient to predict a track's popularity duration, then the number of weeks was set to the appropriate value of *t*. Overall, 54 **MCV** dimensions emerged for explanation.

2.4.3 Time-Wise Music Construct Vector (TMCV)

To analyze more about the development of a track's popularity over time, a *time-wise musical construct vector* (**TMCV**) was built and applied. Some dimensions of a track do not vary over time. They are non-musical constructs: an artist's voice; whether the artist had a major label; etc. Others vary: pre- and post-track release awards, top-rank in the past month or year, and social comments (see Figure 2.3):

Figure 2.3. Past Year, Post-Release Awards and Past Month Observation Windows



Notes. Our analysis involves multiple look-backs based on prior weeks of top-rank chart lists for: (a) the past year, (b) since the last music award news, and (c) during the past month from current week *t*.

- *Artist awards, past month.* Number of times artist nominated for, or won a major music award at *t, …, t* - 3.
- *Artist awards, past year*. Times artist was nominated for or won major music awards in before 2005.
- *Track comments, past month*. Number of track comments at *t, …, t* – 3.
- *Track comments, past year*. Number of comments on track, when a track reached top-rank during the past year based on the criteria of top-50, 100 or 150.
- *Track top-rank, past year*. Number of times artist's track appeared in the top-rank chart list, with top-rank varied based on the criteria of top-50, 100 or 150.
- *Rank change, past 2 weeks.* Change in rank during the previous week (*t* - 1)

compared to the week before (t - 2). A positive value indicates the ranking is ascending (getting worse), and a negative value indicates the rank is descending (getting better).

- *Holiday debut*. Whether current week is in holiday month of December, especially for North America and Europe.
- *Similar tracks, past month.* Number of similar tracks that reached top-rank in Weeks t, t - 1, t - 2, or t – 3.
- *Similar tracks, past year.* Number of similar tracks that reached top-rank during the past year.

The first 5 constructs of **TMCV** are similar to those in **MCV**, and they were calculated across different times in the 10-year dataset. The last two constructs describe the Last.fm effect. It offers a recommendation service for similar tracks to users in its network, so a user's listening choices may be affected. The similarity of two tracks is gauged via the *conceptual Euclidean distance* between 167-dimension low-level acoustic features of each. For a track in a week, a 19-dimension **TMCV** was produced.

Overall, 78,697 observations for **TMCV** were used to explain a track's ranking in a week, so right-censored data (when a track dropped off the chart) were not a problem. But the observations started in February 2005, 3 years after Last.fm was launched. For artists who obtained early social attention, no data were available. Tracks before February 2006, and for artists active before February 2005 were removed. This yielded: 67,508 observations on 2,989 tracks, and 450 music artists.

2.5. Explanatory and Predictive Analysis

2.5.1. Explanatory Analysis - Empirical Models

We present the explanatory results of popularity duration modeling in Table 2.4.

As expected, the subconstructs of **MCV**, music content, artist reputation and social context all have significant effects on music track duration popularity, but with different impact weights.

Table 2.4. Explanatory Results for Music Track *Duration*

CONSTRUCTS AND VARIABLES	LINEAR (SE)	WEIBULL HAZARD (SE)	LOG-LOGISTIC HAZARD (SE)	LOG-NORMAL HAZARD (SE)
Constant	1.23** (0.58)	-0.28 (0.29)	-0.21* (0.27)	-0.47 (0.30)
Genre				
Pop	0.75*** (0.07)	0.47*** (0.04)	0.45*** (0.04)	0.41*** (0.04)
Indie	0.42*** (0.05)	0.25*** (0.03)	0.24*** (0.03)	0.23*** (0.03)
Alternative	0.21** (0.07)	0.10** (0.04)	0.09** (0.03)	0.09** (0.03)
Soul	0.53*** (0.12)	0.25*** (0.06)	0.31*** (0.07)	0.27*** (0.06)
Folk	0.39*** (0.10)	0.25** (0.05)	0.25*** (0.06)	0.24*** (0.05)
Electronic	0.13* (0.06)	0.16*** (0.04)	0.09** (0.03)	0.09** (0.03)
Rap	0.13 (0.13)	0.16* (0.07)	0.09 (0.07)	0.07 (0.07)
Classic	0.79 (1.30)	0.50 (0.75)	0.42 (0.67)	0.47 (0.67)
Blues	0.21 (0.31)	0.20 (0.17)	0.10 (0.16)	0.11 (0.16)
Jazz	0.13 (0.28)	0.14 (0.15)	0.17 (0.16)	0.09 (0.14)
Reggae	0.04 (0.50)	0.05 (0.28)	0.07 (0.28)	0.04 (0.26)
Rock	-0.01 (0.06)	0.04 (0.03)	0.01 (0.03)	0.01 (0.03)
Hip-Hop	0.05 (0.13)	0.001 (0.07)	-0.07 (0.07)	-0.04 (0.07)
Experimental	-0.55*** (0.11)	-0.45*** (0.06)	-0.22*** (0.05)	-0.26*** (0.06)
Country	-0.40 (0.24)	-0.37** (0.12)	-0.38** (0.12)	-0.32* (0.13)
Punk	-0.26 (0.23)	-0.27* (0.12)	-0.14 (0.11)	-0.13 (0.12)
R&B	-0.25 (0.15)	-0.14* (0.08)	-0.08 (0.08)	-0.08 (0.08)
Metal	0.02 (0.13)	-0.05 (0.07)	0.02 (0.06)	-0.001 (0.07)
Instrumental				
Piano	-0.43** (0.07)	-0.21* (0.11)	-0.26* (0.11)	-0.23* (0.11)
Guitar	-0.05 (0.07)	-0.05 (0.03)	-0.04 (0.04)	-0.05 (0.03)
Trombone	1.55* (0.92)	0.82* (0.47)	0.83* (0.51)	0.85* (0.48)
Theme				
Life	0.35* (0.14)	0.13* (0.07)	0.21** (0.08)	0.20** (0.07)
LoveRelations	0.50*** (0.13)	0.27*** (0.07)	0.32*** (0.07)	0.30*** (0.07)
Soul	0.20 (0.14)	0.06 (0.07)	0.17* (0.07)	0.14 (0.07)
SadLifeLove	0.16 (0.13)	0.06 (0.07)	0.14 (0.07)	0.11 (0.07)
Hostility	0.30 (0.16)	0.14* (0.08)	0.22** (0.08)	0.18* (0.08)
Artist Reputation				
MajorLabel	0.02 (0.04)	0.02 (0.02)	0.01 (0.02)	0.02 (0.02)
Post-2004Awards	0.02 (0.02)	0.02 (0.01)	0.02* (0.01)	0.02 (0.01)
Post-2004Nominations	0.07*** (0.02)	0.04*** (0.01)	0.04*** (0.01)	0.04*** (0.01)
Pre-2005Awards	0.01 (0.01)	-0.004 (0.01)	0.003 (0.01)	0.002 (0.01)
Pre-2005Nominations	-0.04*** (0.01)	-0.03*** (0.01)	-0.02** (0.01)	-0.02** (0.01)
Social Context				
HolidayDebut	0.06 (0.07)	0.003 (0.04)	0.03 (0.04)	0.01 (0.04)
FirstTop-Rank#	-0.004*** (0.00)	-0.002*** (0.00)	-0.003*** (0.00)	-0.003*** (0.00)
EarlyStageComments	0.003*** (0.00)	0.002** (0.00)	0.001* (0.00)	0.001** (0.00)
Top-Rank 51-100	0.11*** (0.03)	0.045* (0.02)	0.06** (0.02)	0.06*** (0.02)
Top-Rank 100-150	-0.10*** (0.03)	-0.035* (0.02)	-0.04* (0.01)	-0.05** (0.01)
Model fit: Adj. R^2 or *LL*	Adj. R^2 = 0.269	*LL* = -4,592.4	*LL* = -4,327.8	*LL* = -4,268.8
Shape Parameter	—	2.016	3.326	(0.418, 0.524)

Notes. *LL* = log-likelihood. Estimated shape parameter of *log-normal* is: $\mu = 0.418, \sigma^2 = 0.524$. Omitted variables are shown in Appendix Table B1. Mood-related variables (*Passionate, Lively, Brooding, Humorous, Intense*) were not significant, and are also shown in Appendix Table B1. Signif: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Music semantics. In the 10-year period, 2005-2015, the *Pop* music Genre

seems to have most easily achieved longer popularity. This result follows since the dependent variable is in log form while the explanatory variable is not. Comparing *Pop* and non-*Pop* music, the difference is $e^{0.47}$ - 1 = 59.9% (Bhattacharjee et al. 2007). The *Indie*, *Soul*, *Alternative* and *Electronic Music* Genres had longer popularity too, while *Experimental* music had 36.0% ($p < 0.01$) shorter popularity. *Country* and *Punk* also had less sustainable popularity, and Instrumental music tracks with *Piano* or *Guitar* were less successful in maintaining high listener appeal. For Theme, tracks representing *Life, Love,* and *Relationships* were popular longer. For example, music related to *LoveRelations* had 31.0% ($p < 0.01$) more sustainable popularity duration.

Artist reputation. The *ArtistReputation* construct-related *Vocal* variable was not significant. Many tracks were vocal works, which may suggest regression to the mean for their popularity. *Major Label* was not significant, which suggests a different impact than for album sales, for which *Major Label* was positive and significant. But tracks in the same album are likely to be cointegrated, and exhibited correlation over time in their popularity, even if they are not identical.

People attend to recent tracks of famous artists more than older, less active ones. Nobel Prize winner, Bob Dylan, has over 60 music award nominations. His pre-2005 reputation was high, but not so post-2004. He had 2 tracks in our study that charted, were popular for 1 week, and dropped off. In contrast, Adele had no pre-2005 reputation, but her album "21" won music awards. She shot to stardom, and her tracks are top-ranked for a year now. *Primacy and recency effects* are at work it seems.

Also, *Post-2004Nominations* had a positive impact on the popularity for tracks released later, while *Post-2004Awards* did not. There were few awards and many

nominations, so the econometric estimation may not have been able to use information beyond what was present in the nominations. Still, if a musician was nominated, it had a reputation effect for the track's popularity.

Social context. *HolidayDebut* racks released at Christmas in North America and Europe had longer popularity on average, but not significant. When a track first rose to top-chart ranking is important, and the higher its first rank, the longer should be its popularity duration. The number of *EarlyStageComments* in each of the first 8 weeks were tested too. Those in the first 4 weeks had some explanatory power for popularity. We adopted $t = 4$ to maximize the likelihood of discovering an effect. Prior top-ranked tracks before a new track appeared demonstrate an artist's social network power. Locally, each artist has followers, and they will tend to adopt the artist's next album. The top-chart ranks from 51 to 100 had a positive impact, while those from 101 to 150 had a negative impact on popularity sustainability.

2.5.2. Predictive Analysis

The constructed **MCV** explained music *Duration*, so in this section we test whether it also can predict the popularity of a new music track, if we construct a prediction model based on the explanatory analysis. We used machine learning methods to train a prediction model and then test the prediction power of the **MCV** on new music tracks. The insights that are produced will help the music industry on how to assess specific tracks and artists.

Various combinations of constructs and subconstructs to predict popularity duration are used with multiple classification methods: *support vector regression* (SVR), *bagging*, and *random forest* (RF). These are the most used methods in previous research on popularity classification. SVR is a regression model for SVM. *Bagging* is short for bootstrap aggregating, a form of model averaging in machine

learning that combines the classifications of different training datasets to smooth discrimination and avoid unnecessary errors. Its roots are in decision tree-style learning algorithms, especially B-trees. The RF procedure builds on bagging by randomizing feature selection learned through iterative execution of the algorithm. See Breiman (1994) and James et al. (2013) for additional details.

The results of hierarchical prediction tests with *10-fold cross-validation* are shown in Table 2.5.

Table 2.5. *Duration* Prediction Results for Music Semantics, Artist Reputation, and Social Context

CONSTRUCTS (VARIABLES)	SUBCONSTRUCTS	SVR COEF. (SE)	BAGGING COEF. (SE)	RF COEF. (SE)
Music	See the notes below (Singhi and Brown, 2015)	0.26 (0.03)	0.41 (0.03)	0.38 (0.03)
Non-Music	*ArtistReputation*	0.22 (0.04)	0.42 (0.03)	0.43* (0.03)
	SocialContext (Schedl, 2011; Kim et al., 2014)	0.37*** (0.03)	0.62*** (0.02)	0.62*** (0.02)
Combined	*Music + ArtistReputation*	0.30*** (0.03)	0.56*** (0.03)	0.58*** (0.02)
	Music + SocialContext (Lee and Lee, 2015)	0.43*** (0.03)	0.69*** (0.02)	0.72*** (0.01)
	ArtistReputation + SocialContext	0.40*** (0.03)	0.68*** (0.02)	0.69*** (0.02)
	Music + ArtistReputation + SocialContext	0.45*** (0.03)	0.70*** (0.02)	0.73*** (0.01)

Note: Music includes *Genre, Instrumental, Mood,* and *Theme.* Related citations shown in table. Correlations between variables with top-rank chart popularity *Duration* are given by: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. For each prediction algorithm, significance is based on the prediction of using Music only.

Prediction performance overall is based on the correlations between the observed and estimated values of *Duration*. RF achieved the best prediction performance among the methods. This is consistent with the observation that it is the best algorithm to classify a large dataset (Fernández-Delgado et al. 2014). The future popularity of music is possible to be estimated in the early stage after release, and also to distinguish predictive power of various subconstructs.

The correlation between *Music* and top-chart rank *Duration* is positive but not significant ($\rho = 0.38$, $p > 0.10$). *ArtistReputation* was a more reliable indicator compared to *Music*, which captures previous top-chart ranking performance ($\rho = 0.43$,

$p = 0.10$). *SocialContext*'s correlation with duration was highly positive ($\rho = 0.62$, $p < 0.01$). This is consistent with the prior duration analysis because Last.fm's ranking is related to the listening behavior of its social network members. The *SocialContext* subconstructs cover the multiple social effects that were operating in Last.fm, and so the prediction should be the close to the true value for a track's popularity duration.

The assessment of combinations of constructs and their subconstructs – in pairs (*Music* + *ArtistReputation*, $\rho = 0.58$, $p < 0.02$; *ArtistReputation* + *SocialContext*, $\rho = 0.69$, $p < 0.02$; *Music* + *SocialContext*, $\rho = 0.72$, $p < 0.01$) – suggests that if listeners do not have knowledge of what a track is about, the artist and social context still will be useful to predict its future top-rank chart list popularity duration. If listeners do not know who an artist is, especially for new artists, the music content and social promotion are likely to be effective generating new interest in the market. When *Music*, *ArtistReputation* and *SocialContext* were all considered, the highest correlation between predicted and actual duration was achieved ($\rho = 0.73$, $p < 0.01$), but only marginal in terms of the new information it offered.

This information on the 70%+ correlation is helpful for a record label or independent producer to assess future performance, by improving the match of music tracks to market outcomes. This may clarify how much money marketers ought to be willing to spend to improve the likelihood that a track will have longer popularity, since popularity is correlated with future sales.[6]

[6] This perspective is not without controversy though. Many music artists object that record labels' interests in maximizing profit are at odds with an artist's interest in artistic creation and individual expression. An example is David Bowie's album, "Blackstar." Artist reputation was key in driving popularity for music produced just prior to Bowie's early passing in January 2016 – and even more so after that event occurred.

2.5.3. Robustness Check

Exploratory and predictive analysis indicated that the constructed **MCV** has good estimation performance on *Duration* at the very early stage after track release. In this section, we implemented a robustness check on another measurement of music track popularity *Time2TopRank*.

Time2TopRank measures how long time a track takes to attract enough listeners' attention before it first reaches the top-rank. Therefore, to eliminate possible future effects, we only looked at its performance in the first week after release (e.g., *EarlyStageComments*), and removed *FirstTop-Rank#* from the model and prediction algorithm. The explanatory results are shown in Table 2.6.

We found that **MCV** had similar performance for explaining *Time2TopRank* as *Duration* in general, although some other covariates have different effects on the value of *Time2TopRank*. For example, *Folk* tracks more easily gain high popularity *Duration*, but also take the most time to attract enough listeners. This offers useful input for the music industry on *Folk* music promotion: listeners to *Folk* music are loyal, and the music industry should pay more attention to early stage recommendations to increase the speed that they take to reach top-rank.

We also assessed the prediction performance of multiple combinations of constructs on *Time2TopRank*, and the results are shown in Table 2.7. Similar to *Duration*, RF yielded the best performance among the methods, by considering *Artist Reputation* and *Social Context* ($\rho = 0.81$, $p < 0.01$). In contrast to *Duration*, *Music* content did not help performance improvement, and even weakened the performance ($0.75 < 0.81$). This result indicates that, in the early stage after track release, people are more likely to be attracted because of the artists themselves, and not the music content. This attraction may change after people gain experience though,

which affects *Duration* later.

Table 2.6. Explanatory Results for Music Track *Time2TopRank*

CONSTRUCTS AND VARIABLES	LINEAR	WEIBULL HAZARD	LOG-LOGISTIC HAZARD	LOG-NORMAL HAZARD
Constant	0.936* (0.51)	-0.11 (0.25)	-0.36 (0.25)	-0.32 (0.27)
Genre				
Pop	0.31*** (0.06)	0.12*** (0.03)	0.24*** (0.03)	0.21*** (0.03)
Indie	0.12*** (0.04)	0.03 (0.02)	0.11*** (0.02)	0.09*** (0.02)
Alternative	-0.08 (0.06)	-0.02 (0.03)	-0.05 (0.03)	-0.05 (0.03)
Soul	0.21** (0.11)	0.14*** (0.05)	0.05 (0.06)	0.03 (0.06)
Folk	0.67*** (0.09)	0.24*** (0.05)	0.34*** (0.05)	0.03*** (0.05)
Electronic	-0.0001 (0.01)	-0.02 (0.03)	0.02 (0.03)	0.01 (0.03)
Rap	0.08 (0.12)	0.07 (0.06)	0.04 (0.06)	0.08 (0.06)
Classic	1.00 (1.16)	0.39 (0.61)	0.74 (0.52)	0.84 (0.60)
Blues	0.22 (0.28)	0.15 (0.14)	-0.02 (0.16)	0.06 (0.14)
Jazz	0.65*** (0.25)	0.07 (0.13)	0.38** (0.16)	0.28** (0.13)
Reggae	0.52 (0.45)	0.14 (0.22)	0.26 (0.22)	0.22 (0.23)
Rock	-0.22*** (0.06)	-0.16*** (0.03)	-0.05* (0.03)	-0.06** (0.03)
Hip-Hop	-0.23* (0.12)	-0.18*** (0.06)	-0.17*** (0.06)	-0.15* (0.06)
Experimental	0.01 (0.10)	0.004 (0.05)	0.005 (0.05)	0.01 (0.05)
Country	0.15 (0.22)	0.006 (0.10)	0.008 (0.11)	0.02 (0.11)
Punk	-0.43** (0.21)	-0.22* (0.11)	-0.14 (0.11)	-0.15 (0.11)
R&B	-0.12 (0.14)	-0.01 (0.07)	-0.16** (0.07)	-0.13* (0.07)
Metal	-0.22* (0.12)	-0.15*** (0.06)	0.01 (0.06)	-0.06 (0.06)
Instrumental				
Piano	0.30* (0.18)	0.14 (0.09)	0.13 (0.10)	0.13 (0.09)
Cello	-0.20*** (0.06)	-0.10*** (0.03)	-0.13*** (0.12)	-0.12*** (0.03)
Flute	0.12** (0.06)	0.08*** (0.03)	0.06* (0.03)	0.07* (0.03)
Violin	-0.48 (0.26)	-0.22* (0.13)	-0.29** (0.14)	-0.26* (0.14)
Theme				
Life	0.52*** (0.13)	0.25*** (0.06)	0.27*** (0.07)	0.26*** (0.07)
LoveRelations	0.41*** (0.12)	0.20*** (0.06)	0.24*** (0.06)	0.21*** (0.07)
Soul	0.43*** (0.13)	0.23*** (0.06)	0.24*** (0.07)	0.24*** (0.07)
SadLifeLove	0.18 (0.12)	0.09 (0.06)	0.13** (0.06)	0.12* (0.06)
Hostility	0.16 (0.14)	0.10 (0.07)	0.05 (0.07)	0.03 (0.07)
Artist Reputation				
MajorLabel	0.07** (0.04)	0.02 (0.02)	0.02 (0.02)	0.03 (0.02)
Post-2004Awards	-0.03* (0.02)	-0.01* (0.01)	-0.02** (0.01)	-0.02* (0.01)
Post-2004Nomin	-0.04*** (0.02)	-0.03*** (0.01)	-0.02** (0.01)	-0.02** (0.01)
Pre-2005Awards	-0.01 (0.01)	-0.008* (0.00)	-0.004 (0.00)	-0.003 (0.00)
Pre-2005Nomin	-0.02** (0.01)	-0.006 (0.01)	-0.013***(0.01)	-0.012** (0.01)
Social Context				
HolidayDebut	0.20*** (0.06)	0.09*** (0.03)	0.09*** (0.03)	0.08*** (0.03)
EarlyStageComm	-0.004*** (0.00)	-0.002*** (0.00)	-0.003*** (0.00)	-0.002***(0.00)
Top-Rank 51-100	-0.07** (0.03)	-0.008*** (0.00)	0.002 (0.02)	-0.01 (0.02)
Top-Rank 100-150	-0.02 (0.02)	0.009 (0.01)	-0.027** (0.01)	-0.02 (0.01)
Model fit	Adj. R^2 = 0.264	*LL* = -4,299.2	*LL* = -4,130.2	*LL* = -4,089.1
Shape Parameter	—	2.070	3.410	(0.464, 0.502)

Notes. *LL* = log-likelihood. Estimated shape parameter for the log-normal distribution is: $\mu = 0.464$, $\sigma^2 = 0.502$. Signif: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7. *Time2TopRank* Prediction Results for Music Semantics, Artist Reputation, and Social Context

CONSTRUCTS (VARIABLES)	SUBCONSTRUCTS	SVR COEF. (SE)	BAGGING COEF. (SE)	RF COEF. (SE)
Music	See the notes below	0.31 (0.04)	0.35 (0.04)	0.38 (0.04)
Non-Music	*ArtistReputation*	0.23 (0.05)	0.45* (0.04)	0.46* (0.04)
	SocialContext	0.18 (0.04)	0.70*** (0.04)	0.75*** (0.04)
Combined	*Music + ArtistReputation*	0.34 (0.05)	0.47** (0.04)	0.52** (0.04)
	Music + SocialContext	0.34 (0.05)	0.69*** (0.04)	0.73*** (0.02)
	ArtistReputation + SocialContext	0.26 (0.06)	**0.73*** (0.04)**	**0.81*** (0.02)**
	Music + ArtistReputation + SocialContext	0.36 (0.03)	0.70*** (0.04)	0.75*** (0.02)

Note: Music includes *Genre*, *Instrumental*, *Mood*, and *Theme*. Related citations shown in table. Correlations between variables with top-rank chart popularity *Time2TopRank* are given by: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. For each prediction algorithm, the result significance is based on the prediction result of Music only.

2.6 Extended Analysis on Music Popularity Patterns and Ranking

Track *Duration* is an *event-based performance measure* and *Time2TopRank* is a *speed-based performance measure*. Both supply a possible basis for predicting the popularity patterns of tracks, and how their ranks can be estimated.

2.6.1. Popularity Patterns

The dataset in this work demonstrates different patterns for track popularity performance with *Duration* and *Time2TopRank*. Some tracks may attract the attention of a large audience immediately – as soon as they are released – and keep satisfying their audience over time. In contrast, it may take a long time for an artist's track to attract attention – and then the track may lose its audience's interest fast. There seem to be different popularity patterns that are at work. Such patterns may create the impetus to forecast the potential value of a track, even if the artist has yet to achieve popularity. If one can predict a new track's future chart performance based on historical data, this will be a key advance toward the loftier goal of predicting emerging superstars.

Table 2.8. Music Popularity Patterns in Music Social Networks

PATTERN	POPULARITY PATTERNS FOR RISE TO TOP-RANK AND DURATION	*DURATION*	*TIME2 TOPRANK*	*#TRACKS*	*SMOTE*
1	Flash in the Pan, Short Popularity	[1, 3]	All	2,159	1,000
2	Overnight Sensation, Lengthy Popularity	>13	[1,2]	273	819
3	Slower Rise, Lengthy Popularity	>13	>12	232	928
4	Average Rise, Lengthy Popularity	>13	[3,12]	249	996
5	Faster Rise, Average Popularity	[4, 13]	[1, 12]	800	800
6	Slower Rise, Average Popularity	[4, 13]	>12	168	840

Notes. *Duration* and *Time2TopRank* are stated in weeks, and indicate the range of weeks for each pattern as [Lower, Upper] bounds. *#Tracks* measures the number of tracks observed for each popularity pattern. The numbers suggest there is an unbalanced distribution of the data, which is corrected to the number of observations indicated in the *SMOTE* column. Pattern 1, for this dataset, is the *majority pattern.* Its representation is reduced from 2,159 to 1,000 tracks with SMOTE and stays the majority. Over-sampling was done for the *minority patterns* (2, 3, 4, and 6), while Pattern 5 was sampled without change. This process yielded a balanced dataset with 5,383 tracks, that set up the appropriate conditions for prediction.

The insights learned from the top-rank chart list duration explanatory estimation that we conducted are helpful for analyzing the patterns of popularity, and for popularity prediction more generally. Based on the observed data for popularity duration (*Duration*) and the time it takes for an artist's track to become popular (*Time2TopRank*), we sought to categorize the top-rank chart list tracks as shown in Table 2.8, in the following way:

- **Flash in the Pan, Short Popularity (Pattern 1).** The artist's tracks stay in the top-rank chart for less than 3 weeks before dropping off. 56% of 3,881 tracks belong to this pattern, and 60% attract attention during the month since they were released. The effect does not persist.
- **Overnight Sensation, Lengthy Popularity (Pattern 2).** Tracks belonging to this pattern become truly popular. They attract attention since their release and stay on the top-rank chart for a long time. Popular and active artists, such as Beyoncé, Coldplay, Rihanna, and Linkin Park, contribute their long popularity. Other highly attractive songs appeal to general audience tastes, no matter whether the singer is famous already. Adele's "Rumor Has It" and Lady Gaga's "Poker Face" were like this. The pattern they followed may

suggest the rise of a new music superstar. Though "Rumor Has It" is on the same album "21" as "Rolling in the Deep," the former track was released later online, and its rapid rise to the top-rank chart list may have been a result of its association with the already highly-ranked "Rolling in the Deep."

- **Slower Rise, Lengthy Popularity (Pattern 3).** These tracks stay "under the radar" for quite a while, but eventually emerge on the top-rank chart list and attract more listeners. Adele's "Make You Feel My Love" (a past Bob Dylan song) seems to have followed this pattern of growing popularity over time. Adele covered it in her album "19," which was released in 2008. But this track did not reach the top-rank chart list for another 3 years (168 weeks) until after "Rolling in the Deep" became famous, and audiences recognized that they wanted more digital entertainment content from Adele.
- **Normal Rise, Lengthy Popularity (Pattern 4).** This pattern of popularity duration occurs when it takes an artist's track a normal amount of time to rise to the top-rank chart list, but a lengthy period of popularity is achieved. Adele's "Rolling in the Deep" followed this pattern. Before this song, Adele was not that well known around the world. This track took around 10 weeks to attract a large audience and reach top-rank.
- **Faster Rise, Average Popularity (Pattern 5).** This pattern is common among another large group of tracks. Those associated with this pattern seem to have average top-rank survival duration in our dataset, but they reached the top-rank chart list more quickly than other tracks typically did.
- **Slower Rise, Average Popularity (Pattern 6).** This is similar to Pattern 5, only the artists' tracks take a longer time to achieve a position in the top-rank chart list.

We sought to use a machine learning approach to classify and predict the pattern that a music track is associated with based on our calibrated 54-dimensional **MCV** to represent each track. However, imbalanced data related to the patterns or classes of observations that are identified must be considered; when this is not done, it is possible to achieve fewer incorrect predictions and improved accuracy, but nevertheless end up with a model that is not very useful. This is the *accuracy paradox*. For the track numbers, we can see that the dataset has an obvious imbalanced distribution related to the six different patterns. Previous research in healthcare accuracy is known to only reflect the underlying distribution of the classes (Cantrell and Conte 2009, Gonzalez et al. 2014). A similar situation exists in our dataset of music track popularity duration, so it makes sense to use a balanced set of evaluation metrics to assess prediction performance.

We implemented three prediction models to test prediction accuracy, including SVM, bagging, and RF. The prediction results with 10-fold cross validation are shown in Table 2.9.

The prediction accuracies based on the use of the bagging and RF methods were both acceptable for the original dataset (left columns in Table 2.9. 0.67, 0.69). When we examined the underlying classes, we found that accuracy was contributed by the most often observed patterns, in this case, the Flash in the Pan, Short Popularity Pattern (Pattern 1). Davis and Goadrich (2006) showed that the *area under the precision-recall curve* (AUC-PR) is more suitable compared to the *area under the relative operating characteristics curve* (AUC-ROC) for evaluating the classification performance of skewed or imbalanced data. The definitions of PR and ROC are shown in Equation 3.

PR: $Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$

$$\text{ROC: } TP\ Rate = \frac{TP}{TP + FN}, \quad FP\ Rate = \frac{FP}{FP + TN} \tag{3}$$

where *TP* = true positive, *TN* = true negative, *FP* = false positive, *FN* = false negative. AUC-PR does not account for true negatives, so it is suitable to calculate the accuracy for each pattern, because there are more negatives than positives.

Table 2.9. Performance of Three Algorithms for Popularity Pattern Prediction

	Imbalanced Dataset			Balanced Dataset		
	SVM Coef. (SE)	Bagging Coef. (SE)	RF Coef. (SE)	SVM Coef. SE)	Bagging Coef. (SE)	RF Coef. (SE)
Accuracy	0.56 (0.00)	0.67 (0.03)	0.69 (0.03)	0.57 (0.02)	0.73*** (0.01)	0.80*** (0.01)
K	0.00 (0.00)	0.40 (0.04)	0.43 (0.04)	0.48*** (0.02)	0.67*** (0.02)	0.76*** (0.02)
Patterns	**Precision**	**Recall**	**AUC**	**Precision**	**Recall**	**AUC**
1	0.71	0.94	**0.89**	0.69	0.80	**0.77**
2	0.70	0.47	**0.62**	0.90	0.81	**0.92**
3	0.57	0.31	0.47	0.88	0.89	**0.95**
4	0.52	0.18	0.34	0.87	0.84	**0.92**
5	0.62	0.48	0.59	0.64	0.62	**0.66**
6	0.47	0.09	0.16	0.90	0.86	**0.95**

Notes. Significance (*** $p < 0.01$) is for comparisons between the balanced and imbalanced datasets, made on the basis of random forest analysis, for precision, recall and AUC-PR. There is no significant difference for the imbalanced data-based estimation of patterns. This is because there is no comparison until we report the balanced dataset correlations. The bold numeric entries in this table indicate RF-produced AUC-PR metrics that suggest higher levels of predictive capability.

Thus, we use AUC-PR to study the performance of how each pattern performs in the classification of the various observations of music track top-rank chart list data relative to each of the patterns we have discussed. The higher AUC-PR is, the better the model is suited for pattern prediction. Among the methods that we selected for use, SVM is interesting because the 56% accuracy associated with the assignment of observations to Pattern 1 means that it has 100% recall, while the other 5 patterns have 0% recall, but this is not very useful. Even for RF, which has better performance, there still are 3 underlying classes have an AUC-PR of less than 0.50, so Pattern 1 still is represented as the *majority pattern*.

To address the imbalanced or skewed data issue, we applied the *synthetic mi-*

nority over-sampling technique (SMOTE) (Chawla et al. 2002, 2003). The algorithm calls for the under-sampling of the majority pattern – Pattern 1, in our case. Meanwhile, it applies over-sampling to the minority pattern – Patterns 2, 3, 4, and 6, while Pattern 5 is sampled without changed. This process enabled us to obtain a balanced dataset, including of total 5,383 tracks, as shown in the right-most column of Table 2.8. We then redid the top-rank chart list prediction procedure for the new dataset. Table 2.9 shows the results on the right side that were obtained from this process. Although SVM did not produce highly accurate predictions for the imbalanced dataset, the *K*-values increased and led to a significant increase for *K* for the balanced dataset. In contrast, the bagging and RF analyses yielded improved accuracy 2 times of out 3 for our methods and always for *K*. We can gauge these improvements based especially on the AUC-PR values that exceeded 0.60 (in bold) for all 6 patterns that were discovered by the RF analysis for the balanced dataset.

2.6.2. Ranking Prediction

Another research question is whether it is possible to predict a specific track's ranking in a music social network – up to real-time prediction. To do this, we used *ordinal regression* to estimate the ranking. The dependent variable is *polytomous ordinal* for the marginal effects of changes in **`TMCV`** on predicting an improving, stable, or declining weekly rank (see Table 2.10).

A rank correlation of 46.0% between the **`TMCV`** constructs and the one-period look-ahead weekly track ranks (*RankNextWeek). Awards*, *SocialComments*, and *Other* variables that had explanatory capability for popularity *Duration* are useful predictors here. *PriorRankChange* in the past 2 weeks was useful for forecasting *RankNextWeek* for a track. A positive coefficient for *PriorRankChange* indicates decreasing popularity. A negative coefficient suggests increasing popularity. The

estimates of the variables for *SimilarTracks* are in line with the outcomes of the recommender system that is at work among the social network members of Last.fm. When there are more similar tracks in the current week (*SimTracks-CurrWeek*: β = -0.012, $p < 0.01$), this seems to have helped the target track to move toward a more favorable top-chart rank. The opposite was true for more similar tracks in the past week though (*SimTracksPastWeek*: $\beta = 0.015$, $p < 0.01$): the positive coefficient points to a less favorable rank. Looking back at the data, there is evidence of oscillating signs in these estimates. So rather than suggest there is a final reading here, there is a need to investigate the effects closely, and not draw a quick conclusion.

Table 2.10. Ordinal Regression Results for Top-Chart Ranking Increases and Decreases

VARIABLES	COEF (S.E.)	VARIABLES	COEF (S.E.)
Awards		**Top-Rank, Past Year**	
AwardsPastMonth	-0.098** (0.047)	*Top 1-50*	-0.005*** (0.000)
NominationsPastMonth	-0.305*** (0.039)	*Top 51-100*	-0.001*** (0.000)
AwardsPast3Years	0.084*** (0.007)	*Top 101-150*	0.010*** (0.000)
NomininationsPast3Years	-0.018*** (0.007)		
SocialComments		**Similar Tracks**	
CommentsCurrWeek	-0.050*** (0.003)	*SimTracksCurrWeek*	-0.012*** (0.003)
CommentsPastWeek	0.010*** (0.002)	*SimTracksPastWeek*	0.015*** (0.004)
CommentsPast2Weeks	0.003** (0.002)	*SimTracksPast2Weeks*	-0.010** (0.004)
CommentsPastMonth	-0.010*** (0.001)	*SimTracksPastMonth*	0.008*** (0.003)
CommentsPastYear	-0.001*** (0.000)	*SimTracksPastYear*	-0.001*** (0.000)
Other		**Regression Metrics**	
PriorRankChange	0.014***(0.000)	Rank Correlation	46.0%
HolidayDebut	-0.033 (0.022)	Discrimination R^2	19.4%

Notes. Model: Ordinal regression; dep. var. = *RankNextWeek*; obs. = 67,508; Wald *Z* score used. Signif.: * *Pr*|*Z*| < 0.10, ** *Pr*|*Z*| < 0.05, *** *Pr*|*Z*| < 0.01.

The reader should further note that RF with 10-fold CV predicted *RankNextWeek* with an overall correlation of 81% ($\rho = 0.81$, $p < 0.05$) based only on `TMCV`. Better performance was achieved with the **`MCV`** with 91% ($\rho = 0.91$, $p < 0.01$). This shows the fundamental roles of the artist and social network context to compensate for limitations of the musical constructs for predicting track ranks.

2.7. Discussion

2.7.1. What Do the Duration Model Results Mean?

Empirical analysis based on a duration model yielded several key results on the impacts of multiple characteristics of musical and non-musical constructs in this research. From the model estimated with multiple baseline hazard function specifications, we found that music genre was the most impactful musical construct-related factor. In the last 10 years, the *Pop* Genre had the longest popularity, while the *Experimental* Genre had the shortest popularity. For music Theme, tracks representing *Life*, *Love,* and *Relationships* were popular longer. The second important musical construct in the estimated model for our dataset was the Theme. Specifically, music tracks that were related to *Love* and *Relationships* had even longer popularity duration, so this also suggests why the music of artists such as Beyoncé and Adele have enjoyed such high popularity in recent years.

Among the non-musical constructs, a track that was backed by a major label, rather surprisingly, was not more likely to achieve a longer duration for its top-rank chart list popularity. We thought this would have a rather large and significant impact on album popularity, but the backing of a major label apparently cannot guarantee the success of the tracks. People usually remember only one or two music tracks in an album, leading to their chart popularity. In contrast, whether an artist received music award nominations or important awards themselves, it turned out that the reputation of being an active and successful artist is an important driver of popularity and played an important role in a track's popularity duration, especially for newly-released tracks.

This finding also suggests why the record labels always try to identify the appropriate time to release a new digital single or an album to increase the likelihood

of its success for achieving top-rank and high sales. The importance of a music artist's reputation also suggests how the track is perceived in the larger environment of the Internet, while the variables for social context indicate the importance of the drivers of popularity in the local environment of the streaming platform on the Internet. We noted that the higher the numerical ranking when a track first reaches the top-rank chart list, the more likely will the track be able to achieve a longer duration of its popularity. We also observed that the cumulative effects that an artist benefits from who achieve more top-rank tracks. Such artists with top-rank chart listings above the level of the top-100 tracks are more likely to continue to achieve future success.

2.7.2. What Does the Out-of-Sample Prediction Tell Us?

In the analysis we did for the out-of-sample prediction capability for the musical constructs and variables that we analyzed in the duration empirics, we obtained evidence of high performance from a relatively simple prediction model of a track's top-rank chart list popularity at $\rho = 0.73\%$, and an even higher prediction accuracy for speed to top-rank chart at $\rho = 0.81\%$. These, we believe, are useful findings for the context and our data. Being able to predict the speed and duration of a track's future popularity will help a record label or an independent music production team to assess what kind of performance it will have. This kind of prediction capability can be improved to match actual music track outcomes in the marketplace also. This will benefit those who are involved with the production and marketing of music. They will be able to gain a clearer understanding of how much money they ought to be willing to spend to improve the likelihood that a track will have longer or sustainable popularity, since popularity is also likely to be correlated with sales revenue.

We offer several suggestions related to these findings. For example, since people only seem to remember 1 or 2 tracks at most in new albums, the music artists and the record labels should make strategic choices on which track to select as the title track. Why? Because they can create a means of promoting the album based on how their prediction for track-led popularity may create a halo effect and spillover benefits for revenues from the album as a whole. Our results suggest that with just 1 or 2 popular tracks, the entire album may also be successful. Also, so more of the tracks in an album will achieve an average duration for their top-rank chart list popularity, a music artist's record label may wish to release digital singles on a one-by-one basis to extract enough audience attention before they put the whole album out in the marketplace.

2.7.3. How Can the Track Popularity Patterns Be Understood?

We noted that some future superstars may stay "under the radar" for a long time before people realize they have tremendous star power. To address this interesting issue, we categorized music tracks into 6 different patterns of popularity based on their speed to become popular and the duration of their popularity. The patterns are named and noted, and the most interesting and valuable patterns are: Overnight Sensation, Lengthy Popularity (Pattern 2); Slower Rise, Lengthy Popularity (Pattern 3); and Faster Rise, Lengthy Popularity (Pattern 4). We wonder why music tracks and the artists that produced them all reached relatively high popularity but went through somewhat different ramp up processes. Going forward, it makes sense to invest additional time and effort with fusion analytics to probe more deeply into the speed of achieving popularity.

The popularity patterns of European music artists. To deepen the reader's

understanding of the nature of our findings, we selected 754 tracks from music artists that also were associated with Patterns 2, 3 and 4. Among these, 39.7% (299) of the tracks are from European music artists. They represent 17 female vocalists and 17 male vocalists, as well as 39 music groups. The artists include: Adele (UK), Amy Winehouse (UK), Avicii (Sweden), David Guetta (France), Coldplay (UK), Muse (UK), and Daft Punk (France), among others. An interesting observation on our dataset is that the tracks of artists such as these always seemed to contain all 3 patterns that we noted above before their top-rank chart list popularity occurs; of course, they needed to have 3 or more top-ranked tracks to get there but many managed to do that. Also, the artists' music seems to have followed Pattern 3 (for Slower Rise) if they only succeeded with 1 or 2 tracks at the top-rank level of popularity.

Superstars such as Adele and Coldplay, in contrast, always seem to have had all 3 patterns represented among their tracks. Coldplay, for example, has been active since 1996 and is still active as of 2018, with a large stadium tour in Asian cities in late 2017. Yet even for a band with this level of fame, there still were 27 tracks that reflect the various patterns we presented. For example, there are 5 instances of Pattern 3 (Slower Rise, Lengthy Popularity) and there are 7 instances of Pattern 4 (Normal Rise, Lengthy Popularity). In addition, 15 of their tracks reflect Pattern 2 (Overnight Sensation, Lengthy Popularity), an amazing achievement. And, also for Adele, it is the case that she has had 14 tracks that match these three patterns, with 3 in Pattern 3 (Slower Rise), 8 in Pattern 4 (Normal Rise), and 3 in Pattern 2 (Overnight Sensation), all with Lengthy Popularity. She was not well known until 2011, although she had been active since 2006. Her first famous track, "Rolling in the Deep," took 10 weeks to reach the top-rank chart list for the first time in her music career.

Another interesting finding is related to the 73 European music artists in our dataset who achieved top-rank music track popularity. Altogether, only 31 of them were active in producing highly popular music after 2005, while the other 42 artists have enjoyed an older and more long-standing popularity (e.g., The Beatles, The Rolling Stones, Elton John, etc.). For those with more recent music superstar credentials, their post-2005 top-rank chart listed-tracks all are covered by our data observation window. None of them had pre-2005 reputational assets (e.g., past music nominations and awards, past top-rank tracks, etc.). And just 4 of these 31 artists produced music tracks that were Overnight Sensations with Lengthy Popularity (Pattern 2) – all British: Adele, Alt-j, Florence and the Machine, Mumford & Sons. Prior to their great success with a track, all of them had an interesting comment: they all had at least one track that represented a Normal Rise (Pattern 4) to long-lived popularity. This indicates that many music superstars did not become overnight stars. Instead, their starpower required a long period of time to develop before they generated the spark that powered them to the heights of popularity in their careers (a year or longer). Further, they had to continue to be active to maintain their popularity and starpower.

Understanding the popularity patterns of music artists adds a different dimension to our understanding of how the successful ones among them grow over time, especially to the extraordinary level of the music superstar. Our research suggests that there is likely to be important hidden information that is present in an online music-based social network like Last.fm, which can play a useful role in identifying the popularity patterns associated with an artist's already-released tracks. They act as a basis for how far they can go in the extraordinarily competitive environment of the music industry.

2.8. Conclusion

We leveraged machine-based methods and constructed a relatively complete dataset for understanding how music track popularity develops in online music social networks. We considered three key elements: music content, artist reputation and social context effect. This dataset covers the top-chart ranking data in the past 10 years, from 2005 to 2015, and includes the popularity ranking of over 4,000 music tracks. Based on the constructed dataset, we defined two measures of music track popularity: *Duration* and *Time2TopRank.* They cover the lifespan of a music track from release till when it drops off of its audience's top-chart ranking.

We sought to understand the development of a music track's popularity based on the semantic aspects of the music, the reputation of the music artist, and the social aspects of the music content and its coverage in the social network environment. We used these higher-level constructs to create a musical construct vector. We also constructed and analyzed our data using a fusion analytics approach with methods from machine learning to discover how to construct a dataset that could be used for deeper explanatory analysis using duration modeling for popularity survival from econometrics. This permitted us to build a deeper and more complete understanding of how top-rank chart list popularity develops and is sustained for music tracks in our contemporary Internet environment. Although the dataset is not perfect, it can offer information on what we need to consider to understand the development of music popularity in streaming music platforms.

We further obtained insights that suggest the key constructs for a music track to survive longer on the top-rank chart list. In addition, our popularity pattern analysis suggests the presence of 6 different track popularity patterns. The prediction models

that we implemented yielded 73% accuracy for pattern matching with out-of-sample test data. The machine-generated estimates also can be used for predicting music artists who are more likely to become famous and for ranking prediction. In addition to our event-based analysis, we also constructed a time-wise musical construct vector to predict music track top-rank chart listing levels and achieved a relatively high correlation of 81%. More broadly, our results show the relevance of social context and music content for explaining the popularity duration and ranking of music tracks. These predictive analyses verified that the proposed **MCV** and **TMCV** are able to predict music track popularity duration, popularity patterns and weekly ranking, The offer useful insights for the music industry on how a music track can achieve a following by penetrating the market where it is targeted.

There are many sources of social context information that go beyond the boundaries of a given music social network. They include Twitter, a track's sales from Apple iTunes or other vendors, concert tour information, and music video releases related to top-ranked tracks. This is a limitation because the open Internet and other social networks limit our capability to achieve meaningful and reliable estimates. We did not look into the detailed development process of a music track becoming popular in this work based on these other sources of information and data. Next, we shift to focus on music artists, and observe the real-time diffusion of their music that occurs among listeners.

Chapter 3. Understanding Streaming Music Diffusion in a Semi-Closed Social Environment [7]

3.1 Introduction

Studies of information diffusion date back to the 1950s and focus on the diffusion of innovations, for example, new drug usage in physician's networks (Coleman et al. 1957), new technology adoption in the corporate community (Mansfield 1961), and new product growth among customers (Mahajan and Muller 1979). Researchers have tracked, explored, modeled and evaluated how information diffused under different scenarios over time. In more recent decades, with the rapid development of the Internet and the appearance of multimedia and multiple online channels, interest in information diffusion has expanded to include diverse information and digital products in various areas involving more complex environments.

The Internet and online digital channels provide new ways for diffusing information. People now can receive information based on their communication with others, via web searches, or over email, in social networks, and so on. The rich digital channels also create new challenges for estimating diffusion than ever before, because of the huge amount of information that is spread and exchanged on the Internet, and the uncertainty in information source identification and untraceable diffusion flows among people (Garg et al. 2011).

Measuring information diffusion in online social networks has become one of the hot topics in the past two decades, because of the significant growth in the use of social networks. Empirical research in various areas, including Social Science,

[7] The content of this chapter is based on the published research of Ren and Kauffman (2018), modified to include adjustments to be responsive to faculty reviewers of this thesis.

IS, Marketing, and CS, have tried to analyze and estimate how online social networks help users discover and diffuse new information, including news, music, books, movies, etc. (Dewan et al. 2017, Garg et al. 2011, Kumar et al. 2006, Myers et al. 2012). This research focuses on the analysis of streaming music diffusion in an online social network.

Music streaming via social networks involves a special platform for making the depth and richness of music available to millions of people. The platform creates social connections by introducing artists and music products more directly and faster to Internet audiences. This has brought challenges and opportunities to the music industry to figure out how social networks diffuse music, and how the platforms can strengthen the connection between audiences and artists, thereby effectively promoting music in real time (IFPI 2012, 2013, 2015). Such insights are critical for effective marketing plans, so as to maximize music and artist value. Also, understanding music diffusion can improve social network services to retain users.

Exploring streaming music diffusion is not just an important business problem for the music industry, but also a challenging research question from an academic perspective. The prior literature on music diffusion has studied the effect of users' status and activity in social networks, such as social connections, social recommendations, and user-generated content (Sharma and Cosley 2016, Dewan et al. 2017). However, the diffusion of streaming music occurs within platforms that represent a *semi-closed environment* (Garg et al. 2011). People's listening behavior for a song or an artist may be affected by other users on the platform through their social connections. It also may be guided by artist-related information from other out-of-network factors that we refer to as *external information*, which include music content news, and artists' social activities on social networks and mass media platforms,

such as newspaper, radio and TV.

Although social influence impacts artists' music diffusion (Bapna and Umyarov 2015, Pálovics and Benczúr 2015), without the external force associated with content promotion, the diffusion usually will be slower and decline over time to some relatively stable level, even for superstars like Adele. This suggests that firms may leverage external information for consumer engagement by bringing it into music social networks. An example is British vocalist Jessie J's appearance on the TV show, American Idol, in 2013. The exposure resulted in more than double the number of listeners in the month following her appearance. Another example is the burst of interest and listening to American vocalist, Ariana Grande, after the terror attack on her concert in Manchester, Great Britain. Although it was not a suitable time to promote Ariana's music, social listening still reacted positively to the negative event.

In this research, we add new knowledge by focusing on external information that may drive music diffusion. Using a large panel dataset on listening in a music social network, we ask: (1) What kinds of external information affect streaming music diffusion? (2) How large and persistent are the effects of external information? (3) What kinds of information can be used to improve personalized music recommendations?

We applied a *causal inference approach* to assess the effects of external information on streaming music diffusion at the macro- and micro-levels, using a large panel dataset on listening in a music social network. The raw data contain over 557,000 users' weekly listening records, and we tracked user listening for 1,300 artists over one year. To further analyze music diffusion, we collected two categories of external information, artist characteristics (e.g., artist popularity, music

genre, etc.), and user characteristics and listening behavior (e.g., listening taste, demographics, etc.). We constructed a large panel dataset to dissect the effects of external information on artist-related music adoption and user listening behavior.

This study found that external information has a significant impact on an artist's music diffusion in a network, and the impact and persistence are related to the details of artist information. This effect was confirmed at both the artist and user levels. This study also found that a listener's geographic location limits the diffusion of music. Although people can access whatever music they like, their more limited access to external information may constrain their attention.

This study contributes to the literature on music diffusion analysis in music social networks, by considering external information within a semi-closed music streaming environment. Prior research pointed out that there is media influence, however, what kind of information can be leveraged to promote artists and their music in social networks remains unclear. We also contribute to personalized music recommendation design by providing insights on user listening behavior.

The reminder of this chapter is organized as follows. In Section 2, we review the literature related to music diffusion in social networks. Section 3 describes the research context, data collection and variables in the empirical analysis. The econometric models are shown in Section 4, and their results are interpreted in Section 5. Section 6 discusses the findings and points out the managerial interest. Section 7 summarizes this chapter.

3.2. Literature Review

3.2.1. Information Diffusion Estimation in Social Networks

Research on information diffusion in social networks can be categorized into two big categories: *social influence* and *information discovery*. We give definitions

on these terms based on Garg et al.'s (2011) study and our research target, *external information*:

- **Social influence**. This is the effect of social relation or in-network recommendations related to a user's listening behavior, which involves a person's acceptance of what the system can supply.
- **Information discovery**. This is the effect of mass media information or out-of-network news related to a user's listening behavior, for a person to learn about what to buy or consume.

For *social influence,* research in CS, IS and Economics has estimated it impact for various kinds of information diffusion in social networks (Aral et al. 2009, Bakshy et al. 2012, Bapna and Umyarov 2015). From an econometric perspective, the problem in identifying social influence is to confirm that individuals' choices depend partly upon the choices of other individuals, which is referred to as the *reflection problem* (Manski 1993). Open information access via the Internet and other online channels is making this a more complex problem. Thus, most research on influence detection usually is undertaken for relatively closed social environments.

For example, Susarla et al. (2012) estimated the social influence on the diffusion of user-generated videos on YouTube. Xie et al. (2015) modeled the impact of followers on the diffusion of user-generated tweets on Twitter. One commonality in this research is that targeted information, such as user-generated videos and tweets, can only be accessed from their corresponding social networks, YouTube and Twitter. In this scenario, the biggest challenge for social influence detection is to distinguish it from the self-selection of users – *homophily* – in the local diffusion process. Previous research has arrived at diverse conclusions on the diffusion of different

kinds of information though. Even for a single type of information, such as streaming music, there are various findings on social influence.

For the other stream of information diffusion research, *information discovery*, studies are rare, because of the challenges in implementation and the unobservable process from discovery to consumption (Garg et al. 2011). People may receive a piece of new information through active search on the Internet, or they may come across it serendipitously, such as when they are reading a newspaper, watching TV, or attending an outdoor event (sports, music festival, etc.). But we can only observe the results of diffusion when people adopt new information.

It is truly hard to track what happens from the time of discovery all the way to adoption though. As a result, researchers have tried to consider the impacts of the interplay among multiple media channels on information diffusion for specific social network platforms. They usually focus on a specific period in the information diffusion process: the beginning stage after the information or related-content is released, when there is still not much social influence yet (Myers et al. 2012, Thies et al. 2014).

Information discovery involves unobservable external factors for information diffusion estimation. It points out a way to explore information diffusion in the complex media environment in the early stage, and the social influence that accrues over time. For the music industry, analyzing early developments related to streaming music diffusion can help them and their social media platforms to target the right users and improve services by leveraging external information. However, there still are few researchers who have paid attention to this topic.

3.2.2. Social Influence Effects on Music Diffusion

Streaming music services involve collaboration between music providers and social networks. People can listen, like, tag music and artists, and also make social friends, or join a listening group. Music is a special information product which is different from instant information, such as tweets and posts. People can keep listening to an artist or a song for a long time, or they may continue to listen to an artist after a long hiatus, due to a piece of news about the artist that gains the user's attention. This may be the release of new music content, or an artist's appearance in a big event.

Most research on streaming music diffusion in social networks has focused on the effects of influence, related to the users' status and activity. They include: the effects of social relations and capital (Ellison et al. 2011, Bapna and Umyarov 2015, Sharma and Cosley 2016); the role of weak and strong ties (Bakshy et al. 2012); the impacts of social recommen-dations (Garg et al. 2011); and what happens with user-generated content (Susarla et al. 2012). There have been varied conclusions on how such influence developed. For example, for example, Bapna and Umyarov (2015) confirmed the significance of social influence through friends related to music subscriptions. Sharma and Cosley (2016) indicated that the effect of social influence has been over-estimated. They also found that the majority of shared music listening be-tween friends is from homophily, not from influence. In their findings, less than 1% of users' actions can be explained by their friends' influence. Garg et al. (2011) and Dewan et al. (2017) also reported evidence for music diffusion and social influence, but social influence was more important for music with narrow or niche appeal, compared to more broadly appealing music. Prior research also has paid attention to the passive acceptance of a user's music listening selection. However,

it ignored the hidden reasons that drove a user's active decisions to change from one artist to another at a specific time.

3.2.3. Information Discovery Effects on Music Diffusion

Studies on the discovery effects on music diffusion have been rare. They are challenging to implement due to the unobservable process involving the discovery of the music all the way through to consumption (Garg et al. 2011). Music networks operate in a *semi-closed platform environment*, so listeners can discover information from the media or via content sampling. Internet technologies allow users to sample music via Twitter, TV, newspapers, websites, and email. As such, it is not easy to determine what is the relevant information source to prompt music diffusion without conducting a randomized experiment or a user study. Research in CS has sought to find the correlation between diffusion and external information via observational data and empirical designs, but has not done enough.

Schedl (2011), for example, explored music listening trends on Last.fm and Twitter, and reported that music popularity across platforms is correlated, so diffusion may involve platform interplay. However, few authors have studied the detailed impacts of external news on streaming music diffusion. To date, Myers et al.'s (2012) and Myers and Leskovec's (2014) work on Twitter are the most complete study of the effects of various types of external news on the diffusion of tweets in Twitter. The authors constructed a diffusion model of tweets over time, and found that 29% of Twitter information propagation is due to external information drivers. They also pointed out the different effects for different types of news information, including sports, business, entertainment, and travel news, among others. Myers et al. only studied the effects in general though, and did not analyse the effects of information content at a finer level of granularity, for example, by assessing music-

related news in the entertainment category.

Information that is available from different platforms is known to have different diffusion patterns. Myers et al. (2012) showed that news diffusion on Twitter has rapid information mobility, Thies et al. (2014) reported that the influence of social buzz from information diffusion only seems to persist for about 24 hours in Twitter and Facebook – not long. But music diffusion on a streaming music platform is a longer-term process, since music is a durable information good (Poddar 2006). More useful insights may be related to how long such information affects the choices users make of what music to listen to, and whether there are diverse effects for different types of music information sources. So in this work, we focus on the discovery effects on streaming music diffusion, to assess how external information drives music diffusion in social networks.

3.3 Research Setting and Data

In this research, we examined Last.fm, an online music community that integrates music listening, social activities, and social recommendations into a single platform. Besides music streaming, Last.fm also supplies a special "Events" column to broadcast important activity related to a specific artist, such as a coming concert or a live show. Last.fm users can access external information through its platform, and also have separate access to the Internet. To explain streaming music diffusion, we focus on both macro- and micro-level listening changes. For macro-level diffusion, we analyze the effects of external information on an artist's global listening, as well as in specific geographical locations. For micro-level diffusion, we analyze listening changes at the individual user level. To gauge the effects of external information, we use weekly listening log data to measure music diffusion in social

streaming, summed to the month level when more aggregated observations are necessary.

3.3.1 Data Collection

We acquired listening records from January 2013 to November 2013 in two stages. Totally 41 weeks data were collection.

User selection. Via the Last.fm API, we collected data on 110 top artists in the five-year period from 2008 to 2013. They were the most popular ones, with 100+ million play counts through 2017. For each artist, we also extracted the top fans of the artist's top songs. This yielded 18,933 seed users. They represent those who had already adopted at least one of the artists before the observation period and had the potential to keep listening to the artist in the future.

To obtain a more representative set of users, we also included users' social relations. Users on Last.fm are linked to each other through their social structure and relationships. Last.fm supplies data for: friends, possibly with an actual social relationship, or with similar listening tastes; and neighbors, who are recommended by Last.fm due to their listening similarities. These users are existing or potential listeners of an artist's music. We extracted 1-hop social relationships for the seed users and obtained information about 202,966 of their friends and 383,522 of their social network neighbors. Overall, the weekly listening logs of 557,554 users were downloaded for the observation period. The behavior of these users offers a representative snapshot of the basic music listening patterns on Last.fm. Based on statistical listening overlaps between seed users and their social relations, there was not very much overlap: only ~9.5% with friends, and ~20.5% with neighbors. This is consistent with Sharma and Cosley's (2016) findings on the over-estimation of social influence on music diffusion.

Artist selection. Based on preliminary statistics on the artists to whom users listened, we assessed 10,000+ artists for inclusion. The top 1,300 of them with at least 10 music tracks, including 110 who were popular through 2017, were finally selected as the study targets. We targeted popular artists because it is possible to observe visible diffusion within a short observation period for them.

3.3.2. Panel Dataset Construction

The raw data contain weekly listening records for 557,554 users. The first concern is whether user listening behavior changed in the one-year period around their observation. If user listening did not change over time for different artists, then listening tastes must have been stable in the observation period. This makes diffusion analysis unnecessary. Thus, we tested for artist listening changes for each user before doing diffusion analysis.

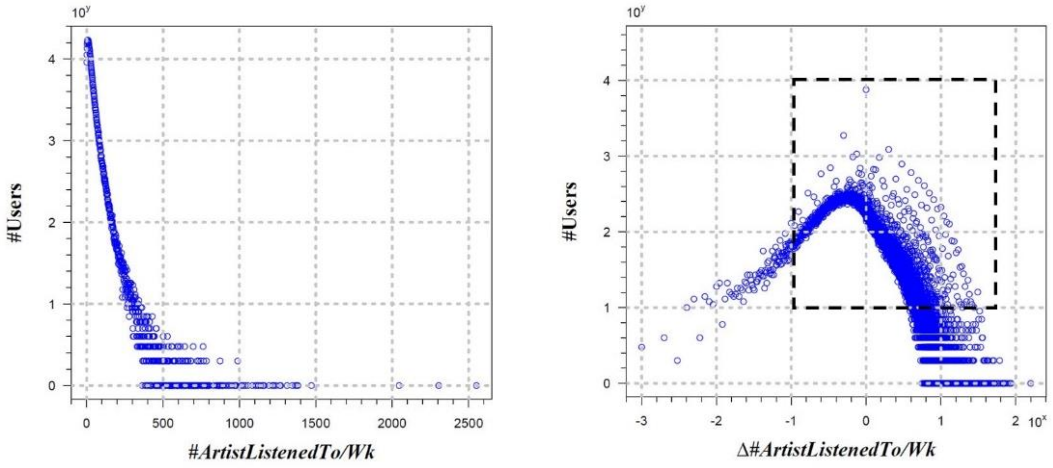
For each user, we captured data on the extent of their listening, *#ArtistListenedTo*, by *Week* and by new *Artist.* On average, each user listened to around 30 different artists weekly; statistics on weekly listening behavior are shown in Figure 3.1 (left). We also gauged each user's average changes in the different artists they listened to by *Week* and by *Artist* in the 41-week observation period, and calculated the proportion of new artists listened to, Δ, via this variable:

$$\Delta ArtistListenedTo_j = \frac{1}{40}\sum_{t=2}^{41}\frac{\#\ New\ Artists\ User\ j\ Listened\ to\ in\ Wk\ t}{\#\ Artists\ User\ j\ Listened\ to\ in\ Wk\ (t-1)} \qquad (1)$$

The data plot for Δ*#ArtistListenedTo* indicates that users had large listening weekly change rates (Figure 3.1, right). Most users listened to at least 10% (Δ*#ArtistListenedTo* > 10^{-1}) new artists compared to the week before (dashed box, Figure 3.1, right). This rate increased substantially (by 10^x with $x > 0$), which indi-

cates that user listening behavior is dynamic, and the change rate is salient and varied. Dynamic listening suggests the dataset is suitable for diffusion analysis to understand music network users better.

Figure 3.1. User Listening Behavior, by Artist and by Week



We created two panel datasets at the macro- and micro-level by merging user listening with artist characteristics, user characteristics, social network effects and country information data. Using *propensity score matching* (PSM) (Dehejia and Wahba 2002), we developed balanced panel data subsets for econometric analysis, to estimate music diffusion at the macro-level. Based on a more detailed scan of each user's listening behavior, we also constructed a more fine-grained panel dataset for assessing music diffusion at the micro-level. We next describe the dependent variables, the main effects, and controls for artists, users and countries (see Table 3.1 for the variables and notation).

Table 3.1. Notation and Definitions of the Study Variables

VARIABLES	DESCRIPTION	VALUE
DEPENDENT VARIABLES		
Artist#Plays/Wk	# times artist's music is played by all users each week	Numeric
Artist#Listeners/Wk	# unique users who listened to an artist's music each week (not Δ change rate)	Numeric
Artist#Plays/Mo	# times an artist's music played by all users each month	Numeric
Artist#Listeners/Mo	# unique users who listened to an artist's music each month (also not Δ)	Numeric
UserArtist#Plays/Mo	# times an artist's music played by a specific user monthly	Numeric
MAIN EFFECTS VARIABLES		
ArtistExtInfoRel	External info release occurred for artist, 1; 0 otherwise	Binary
AfterRelease	Period after artist's external info released, 1; 0 otherwise	Binary
ArtistExtInfoType	Type of external info released on an artist	Category
ExtInfoWeekAfter	Week # (-1, 1, 2, 3, 4) after external info released	Category
CtryExtInfoRel	Country where external info was released (multiple variables)	Binary
ARTIST CONTROL VARIABLES		
LongPopLast.fm	Top chart popularity on Last.fm, from 2005 to 2013	Numeric
LongPopBB	Top chart popularity on Billboard, from 2005 to 2013	Numeric
ShortPopLast.fm	Top chart popularity on Last.fm, 1 month before info release	Numeric
ShortPopBB	Top chart popularity on Billboard, 1 month before info release	Numeric
Artist	Two gender variables, Male (1, 0), female (0, 1) with band (0, 0) as base case	Binary
MajorLabel	Whether artist is connected with major music label	Binary
Genre	Artist's music genre (18-d numeric variable-based genre vector)	Vector
USER CONTROL VARIABLES		
ListeningScale	# of artists user listened to	Numeric
ListeningBreadth	User's diversity of music listening across artists	Numeric
ListeningTaste	User's listening taste (18-d numeric variable-based genre vector)	Vector
TasteSimilarity	Taste similarity for user with artist	Numeric
#Friends	# of friends of user who listened to artist	Numeric
#Neighbors	# of neighbors of user who listened to artist	Numeric
YrsSinceReg	# of years since registration	Numeric
Ctry	Country where user is from	Category
CtryExtInfo	External info released in user's country, 1; 0 otherwise	Binary
Artist#ExtInfoRelease	# of artists with external info in same period, listened to by a user	Numeric

3.3.3. Main Effects Variables: External Information

People access news and event information for artists through various Internet and other selected channels (e.g., Last.fm, Spotify). We captured such changes that affect music diffusion from multiple sources via the Internet. Considering just one kind of external source of information may create bias for geography, culture, information category, etc., we used Google Trends to support the identification of various sources of external information for an artist. Google Trends offers an unbiased sample of search data covering multiple categories, such as Entertainment, News, and other sources, like YouTube search.

Examining what people search for provides a perspective on their preferences and interests. External sources of information capture their immediate interest in a topic, compared to typical search volume. We used weekly change rates for the number of searches, and selected weeks in which rates of change were 50% larger than the prior week. For each external source, we filtered the information by checking its content based on what could be extracted from publicly-available data, for assessments from Wikipedia, Pitchfork, Setlist.fm, Google News, and Last.fm events. We then clustered them into two categories: *Music Content Information* and *Non-Music Content Information*, with eight types (see Table 3.2).

Table 3.2. External Information Source Type (*ArtistExtInfoType*)

TYPE	DESCRIPTION	# ARTISTS
Non-Music Content Information		**197**
1	*News, Artist Life*	48
2	*News, Music-Related Info, Music Awards*	47
3	*Tour, Concert*	40
4	*Live TV Show*	28
5	*Live Performance / Festival*	34
Music Content Information		**210**
6	*Single-Song Release*	66
7	*Album Release*	131
8	*Music-Video Release*	13

Music Content Information is directly connected with new music products, including *Single-Song*, *Album* and *Music-Video Releases*. *Non-Music Content Information* is more diverse and covers five types of artist social activity: *News, Artist Life* (e.g., birthday, marriage); *News, Music-Related*/*Music Awards* (e.g., Grammy Awards, news of a coming album); *Tour/Concert*; attending *Live TV Shows* (e.g., Saturday Night Live); and *Live Performances* at music festivals. Such information may not be directly connected with new music products, but still may attract people's attention when they are reading the news or watching TV.

Some artists had one instance of external information released in the study period, while others had multiple types: for example, they may have arranged a *Single-*

Song Release (Type 6), then an *Album Release* (Type 7), followed by a *Concert* (Type 3) week by week. To reduce the effects of multiple sources, we used only one external source of news and information for each artist during a week. For artists with multiple releases, we selected the one that had at least a two-month gap from others that were identified, to reduce possible over-estimation of the effect of a release. For artists who had multiple external information releases in the same week, we removed them from the candidate list to avoid the cumulative effect of multiple sources.

In total, for the 1,300 artists, 407 had external source-based information releases during the study period, 210 had new *Music Content Information*, and 197 released new *Non-Music Content Information* (see Table 3.2).

Time and geographical variables. Streaming music listening should not be bound by the time or any limits of geography, due to the ubiquitous nature of the Internet. However, the effect of an instance of external news may affect music listening and be limited both by the timing of its release and the country where it was released, especially for *Non-Music Content Information*, such as TV shows and concerts. A strong regional event may only have impact on local listeners, for example.

Understanding the type, time effect and geolocation of external information may offer deeper insights into the music labels for what kinds of news and information can be used for music promotion and effective implementation of related marketing efforts.

3.3.4. Dependent Variables: Streaming Music Diffusion at Macro- and Micro-Levels

Macro-level. Music social networks often use the number of times an artist's

music is played (*Artist#Plays*) and the number of users who listen to a unique artist's music at least once (*Artist#Listeners*) to rank their popularity. *Artist#Plays* reflects an artist's general popularity, while the latter indicates the artist's market penetration with users. YouTube uses *Artist#Plays* to measure track popularity; and Last.fm and Spotify track both *Artist#Plays* (scrobbles) and *Artist#Listeners*. Some users listen to a song once, while others listen more, so these dependent variables are complementary for assessing diffusion. We used both to measure each week's music diffusion at the macro-level, and the monthly diffusion for a robust check.

Diffusion statistics on *Artist#Plays* and *Artist#Listeners* are shown in Table 3.3. We note that *Artist#Plays* is five times as large as *Artist#Listeners*. On average, users listened to an artist five times.

Table 3.3. Streaming Music Diffusion Statistics, Averages for January to November 2013

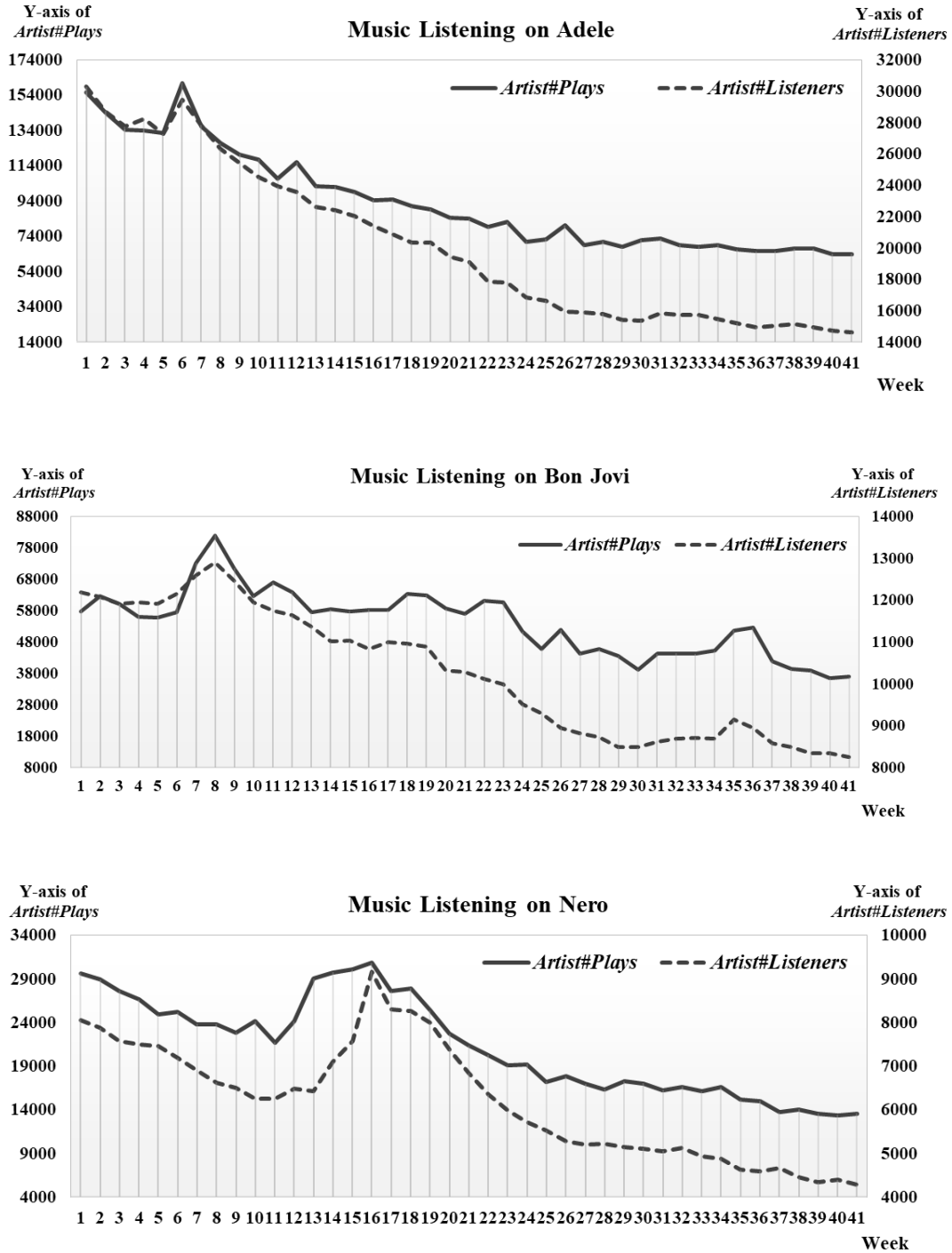
STREAMING MUSIC DIFFUSION	MIN	MAX	MEAN	SE
WEEKLY				
Artist#Plays/Wk	0.05	1,483.3	14.5	44.3
Artist#Listeners/Wk	0.03	85.7	2.8	5.0
MONTHLY				
Artist#Plays/Mo	0.45	4,649.8	59.2	172.3
Artist#Listeners/Mo	0.14	309.2	11.2	20.3

Notes. 53,300 obs., 41 obs. per artist for 1,300 artists; units: 000s of times.

Figure 3.2 gives three examples of weekly music listening of *Artist#Plays* and *Artist#Listeners* in the observation period. The artist in the first row, British vocalist Adele, had an immediate increase in music listening in Week 6. The increase was the listeners' reaction to her performance of "Skyfall" at the 85th Oscars, where it won the Best Original Song award. However, in the following weeks, there was no additional external information released. Thus, the rate of diffusion declined – even for a superstar like Adele. The middle row shows the listening of music by the American rock band, Bon Jovi. They had a new album "What About Now" released on March 8, 2013, in Week 7. The album release resulted in an increase and kept

the diffusion rate at a stable level over a long period before it dropped below the previous diffusion rate (Week 23).

Figure 3.2. Three Examples of Weekly Music Listening at the Artist-Level

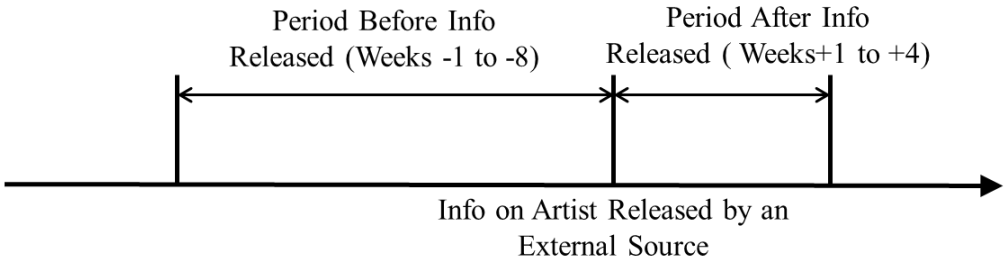


The third example in the bottom row comes from a British electronic music trio, Nero, who were not so popular as the previous two artists. A 30-second sample of

their song “Into the Past” was released on April 11, 2013 (Week 11) as a soundtrack for the movie “The Great Gatsby.” The listening to Nero’s music kept increasing during the movie promotion’s period, until the screening of the movie, May 16, 2013 in Week 17. After that, interest in Nero declined quickly. People may have shifted their attention to the movie or other artists.

Micro-level. Music diffusion at the micro-level is measured with *UserArtist#Plays*. This represents how many times a user listened to an artist after news information on the artist was released through an external source (see Figure 3.3). To eliminate the possible endogeneity effects, the users selected for analysis could not have listening records for a given artist prior to the release of external information, but they may have had listening records thereafter. We observed users' listening behavior for two months before the *ArtistExtInfoRelease* and one month afterwards, for the two different periods. We focused on the one-month period after an external release of information occurred. This is because it is not possible to guarantee that there are no influence effects mixed into the discovery effects that users experience, if the observation time is too long. We mitigated this kind of outcome by observing the effects of external information release in a limited 4-week time period after it occurred.

Figure 3.3. Observation Periods for Music Diffusion at the Micro-Level



3.3.5. Control Variables in Propensity Score Matching and Modeling

Music diffusion is affected by multiple factors, including artist characteristics, user diversity, the competitive environment of the music market, and so on. To obtain a balanced panel dataset that can be used in empirical testing for the effects of external information on music diffusion, we considered various control variables, and applied the PSM procedure to match observations on the basis of statistical matches. This enabled us to account for artist and user heterogeneity in music characteristics and listening behavior.

Artist. Artist music listening is influenced by a number of factors:

- **Popularity.** An artist's popularity contributes to the diffusion of their music (Bhattacharjee et al. 2007). We refer to this as *artist popularity*, and it is possible to distinguish the length of time that it lasts. Popularity exhibits an accumulating process and is dynamic over time, so long-term and short-term popularity are used to describe it. *Long-term popularity* (*LongPop*) is the accumulated level of popularity from 2005 until 2013, just before the study period. *Short-term popularity* (*ShortPop*) is defined as its observed level during the month prior to the time the observation week occurred. Both measures are based on the number of appearances on the Billboard Hot 100 and Last.fm weekly top-100 ranking charts at the relevant times.
- **Artist.** Two binary variables indicate male (1, 0), female (0, 1) and band (0, 0). Band (0, 0) is the base case.
- **Major Label.** One binary variable represents whether an artist had a major music label (1) or not (0). Three major labels (Universal, Sony/EMI, Warner) operated during the study period.
- **Genre.** A supervised machine learning algorithm, *support vector machine*

(SVM) was used to learn an 18-dimensional music genre vector from the genre of the artist's top songs (as in Essay 1). Each describes an artist's probability to belong to a genre.

User. A user's choice of what to listen to in a social network may be affected by multiple factors:

- **User listening behavior**. This includes how much and how diverse a user's music listening behavior was before information release occurred. *ListeningScale* is the number of unique artists a user listened to, and *ListeningBreadth* indicates how much a user listened to different artists (Garg et al. 2011). We used the *Gini coefficient* for the artist-level distribution of the number of songs to measure listening diversity. We also used *Euclidean distance* for the *TasteSimilarity* for a user and an artist with the user's current taste, and the artist's genre.
- **Social influence.** This is represented by a continuous variable for the number of a user's social friends or neighbors who listened to the artist, and they may have affected the user's listening choices.
- **External information released.** Two variables are included to indicate that news about an artist represents different types of information. One is *ArtistExtInfoType*, representing the base case (Type 1) and seven other types. In addition, for all the artists with external information releases in the same observation period, we measured the number of unique artists whose music each user listened to, *Artist#ExtInfoRelease*. This measure suggests a possible *crowding effect* that makes any individual external information release have less impact.
- **Geolocation.** The binary variable *CtryExtInfo* codes for whether an artist's

information release occurred in the country of the music-listening user's residence or if an artist-related external event (e.g., concert, TV show) occurred in the country.

3.4. Model and Methodology

A *difference-in-differences* (DiD) model was used to examine the causal effects of external information on music diffusion. It involves a hierarchical empirical method with steps from the general to a more fine-grained perspective. Prior to model estimation, we applied PSM to address artist and user endogeneity for their music listening choices and music diffusion. We constructed a macro-level control group, in which artists had no external information released in a common observation period, but where there were comparable artist characteristics and historical music diffusion in the treatment group. A fine-grained control group was also constructed by consider the user's listening behavior and a geolocation effect. By studying the individual user-level, we estimate the effect of external information on how the content of a user's music listening activities is changed.

3.4.1. PSM to Address Artist and User Endogeneity in Music Diffusion

Artist match. For each treated artist with external information released in the study period, we applied a time-dependent propensity score match with a control artist who had no external information released during the same week. For this, a logit model was estimated at the artist-level to obtain a balanced distribution of the observed covariates for the treatment and control group observations, including artists' characteristics, such as music genre, and long-term and short-term popularity.

We also balanced the artists' previous listening status one month before external information was released. This allowed us to observe the direct feedback on music diffusion for two similar artists, with and without external information in the same

observation period.

$$\Pr(ArtistExtInfoRel_{it} = 1 | \cdot) = f(Artist_i,\ MajorLabel_i,\ Genre_i, LongPopLast.fm_i, LongPopBB_i,\ ShortPopLast.fm_{it},\ ShortPopBB_{it}, Artist\#Plays/Mo_{i,t-1},\ Artist\#Listeners/Mo_{i,t-1}, t) \quad (2)$$

Finally, 407 control artist observations were selected from the total of 35,999 control observation for artists. The PSM results are summarized in Table 3.4, which shows that the treatment and control artists are properly matched. These 814 artists were further used to analyze music diffusion at the macro level.

Table 3.4. Propensity Score Matching Results for Artists

ARTIST CONTROL VARIABLES	TREATMENT	CONTROL	CONTROL MATCHED
Artist: Male	0.248	0.199	0.248
Artist: Female	0.155	0.102	0.194
MajorLabel	0.482	0.440	0.514
Music Genres			
Rock	0.618	0.684	0.601
Alternative	0.222	0.256	0.225
Indie	0.284	0.346	0.259
Pop	0.269	0.204	0.260
Hip-hop	0.067	0.073	0.059
Rap	0.033	0.033	0.027
R&B	0.048	0.020	0.065
Electronic	0.119	0.115	0.117
Metal	0.220	0.220	0.217
Folk	0.081	0.076	0.102
Soul	0.041	0.036	0.042
Experimental	0.087	0.001	0.089
Punk	0.041	0.093	0.045
Classic	0.014	0.047	0.006
Jazz	0.025	0.019	0.029
Blues	0.034	0.021	0.036
Country	0.004	0.014	0.010
Reggae	0.011	0.007	0.048
LongPopLast.fm	25.830	33.680	22.510
LongPopBB	20.290	11.430	23.640
ShortPopLast.fm	0.378	0.196	0.415
ShortPopBB	0.315	0.086	0.354
Artist#Plays/Mo	137,181	99,778	149,486
Artist#Listeners/Mo	23,209	16,689	23,290

Note. Numeric entries represent the statistical means of each control variable, used for artist matching. The *Control Candidate* column is the statistical result of 35,999 observation for artists who had no external information releases in the study period. *Treatment* and *Control Matched* are the statistics for 407 artists in each group.

User matches based on geolocation. Some of the external information has ob-

vious geographical bounds for its informational relevance, such as TV Shows, Music Festivals, and Concerts, which occur in a specific country or city. To estimate whether there are geographic restrictions on music diffusion related to external information releases, a fine-grained panel dataset was further constructed according to the country where the external information was released. Among the 407 artists in the treatment group with information released, 199 had external information releases in the U.S. and 76 in the U.K. We focused on these two top-ranked countries for comparison. There are 46,200 users in the dataset that are in the U.S., and 18,402 from the U.K., out of the total users. The listening change differences were tracked for U.S. and non-U.S. users, and for U.K. and non-U.K. users.

Similar to the artist treatment-and-control groups, the user's country, *Ctry*, was applied for matching, by balancing their registration time and listening behavior:

$$\Pr\left(Ctry_j = 1 \middle| \cdot\right) = f\left(RegSinceYear_j,\ ListeningTaste_j,\ ListeningScale_j,\ ListeningBreadth_j\right) \quad (3)$$

This resulted in 92,400 users tracked for U.S. external information releases, and 36,804 for the U.K. Tables 3.5 and 3.6 summarize the PSM results for the control and treatment groups of U.S. and U.K. users, which also show that the treatment and control group artists are properly matched.

3.4.2. Difference-in-Differences (DiD) Model at the Macro Level

The effects of external information were examined at the macro level with a DiD model (Imbens and Wooldridge 2007). This is ideal for the structure of the analysis. Diffusion at the artist level *i* is estimated via a pre-and-post DiD model:

$$\begin{aligned} DepVar_{it} = {} & Constant + \beta_1 ArtistExtInfoRel_{it} + \beta_2 AfterRelease_{it} \\ & + \beta_3 ArtistExtInfoRel_{it} \times AfterRelease_{it} \\ & + \beta_4 ArtistExtInfoType_{it} + \beta_5 ExtInfoWeekAfter_{it} + \epsilon_{it} \end{aligned} \quad (4)$$

Table 3.5. Propensity Score Matching Results for U.S. Users

Artist Control Variables	Treatment	Control	Control Matched
RegSinceYear	4.246	3.696	4.26
ListeningTaste			
Rock	0.263	0.328	0.260
Alternative	0.193	0.199	0.195
Indie	0.268	0.203	0.275
Pop	0.182	0.188	0.179
Hip-hop	0.111	0.063	0.112
Rap	0.049	0.028	0.051
R&B	0.051	0.031	0.047
Electronic	0.140	0.116	0.144
Metal	0.033	0.069	0.033
Folk	0.064	0.045	0.066
Soul	0.036	0.029	0.036
Experimental	0.000	0.000	0.000
Punk	0.031	0.030	0.030
Classic	0.007	0.007	0.006
Jazz	0.010	0.009	0.010
Blues	0.012	0.012	0.012
Country	0.015	0.010	0.014
Reggae	0.002	0.005	0.004
ListeningScale	60,042	56,838	60,442
ListeningBreadth	0.267	0.210	0.233

Note. Numeric entries represent the statistical means for each control variable used for user matching. The *Control Candidate* column is the statistical result of 173,365 observation for users, whose geolocation is not the U.S. *Treatment* and *Control Matched* are the statistics for the 46,200 users in each group.

Table 3.6. Propensity Score Matching Results for U.K. Users

Artist Control Variables	Treatment	Control	Control Matched
RegSinceYear	4.19	3.78	4.23
ListeningTaste			
Rock	0.298	0.316	0.295
Alternative	0.200	0.198	0.201
Indie	0.241	0.214	0.246
Pop	0.185	0.187	0.184
Hip-hop	0.071	0.073	0.072
Rap	0.032	0.033	0.032
R&B	0.038	0.035	0.039
Electronic	0.132	0.120	0.134
Metal	0.045	0.063	0.044
Folk	0.057	0.048	0.058
Soul	0.033	0.030	0.033
Experimental	0.000	0.000	0.000
Punk	0.032	0.030	0.032
Classic	0.007	0.007	0.006
Jazz	0.009	0.009	0.009
Blues	0.011	0.013	0.010
Country	0.011	0.011	0.011
Reggae	0.004	0.005	0.004
ListeningScale	57,987	57,468	58,215
ListeningBreadth	0.263	0.219	0.238

Note. Numeric entries represent the statistical means for each control variable and are used for user matching. *Control Candidate* column is the statistical result for 201,163 users, whose geolocation was not the U.K. *Treatment* and *Control Matched* are statistics for the 18,402 users in each group.

The $DepVar_{it}$ is for music diffusion, represented by weekly or monthly *Artist#Plays* and *Artist#Listeners* for artist *i* at observation week *t*. *ArtistExtInfoType* is the type of the external information, Type 1 (*News, Artist Life*) was used as the base case, because it has a distant relationship with the music product itself, and can be compared with other external information. This main effect variable indicates the validity of the effects of various types of information.

ExtInfoWeekAfter indicates the week after external information was released related to an artist. The listening records for users in the week prior to the external information release represent the basis for comparison. The changes in music diffusion after external information was release compared to before that. We next show the music diffusion changes at the macro-level for all users, and also at a more fine-grained level, by considering user geographical location information separately.

3.4.3. Counting Data Model for Micro-Level Analysis

At the micro-level, we model the listening change of a user after an artist's external information is released. We wish to know how much external information will affect an individual's listening behavior. As Figure 3.3 showed, we tracked users with no record of listening to an artist before the artist's information was released but may have listened to them after that. The effects of multiple variables were examined, including user listening taste, and social influence and external effects. Only seed users with observable social relations in the dataset were considered. The music diffusion of artist *i* to user *j* at observation time *t* is represented with a count data model:

$$UserArtist\#Plays_{ijt}$$
$$= Constant + \beta_1 ArtistExtInfoType_{it} + \beta_2 Artist_i + \beta_3 MajorLable_i$$
$$+ \beta_4 Genre_i + \beta_5 LongPopLast.fm_i + \beta_6 LongPopBB_i$$

$$+ \beta_7 ShortPopLast.fm_{it} + \beta_8 ShortPopBB_{it}$$

$$+\beta_9 Artist\#ExtInfoRelease_{jt} + \beta_{10} TasteSimilarity_{ijt}$$

$$+\beta_{11} ListeningScale_{jt} + \beta_{12} ListeningBreadth_{jt}$$

$$+ \beta_{13}\#Friends_{ijt} + \beta_{14}\#Neighbors_{ijt} + \beta_{15} CtryExtInfo_{ijt} + \epsilon_{ijt} \quad (5)$$

Here, the dependent variable is the cumulative number of streams of user *i* on artist *j,* one month before and after the release of external information, *t* is the week when the information was released. All covariates on the right side of Equation 5 are the corresponding descriptions for the possible factors.

For each music diffusion model, a negative binomial count data model was used. It is suitable for this study since the dependent variable has non-negative values of 0, 1, etc. Also, when the variance of the mean count is higher than the theorized count, it indicates that over-dispersion in the data is present, as was observed earlier.

3.5. Results and Interpretation

The estimation results for the DiD and count data models are presented next. The results include external information effects on music diffusion at the macro- and geographic levels, scale and persistence of various types of information effects, and micro-level listening changes.

3.5.1. Music Diffusion at the Macro-Level

I estimated the listening changes of 557,554 users for the 814 matched artists using Equation 4, to test whether an external information release had a positive impact on monthly and weekly music diffusion on total *Artist#Plays* and *Artist#Listeners.* Table 3.7 gives the DiD regression results.

Table 3.7. DiD Regression Results for Music Diffusion at the Macro-Level

Main Effect Variables	*Artist#Plays / Mo* (I) (SE)	*Artist#Listeners / Mo* (II) (SE)	*Artist#Plays / Wk* (III) (SE)	*Artist#Listeners / Wk* (IV) (SE)
Constant	11.39 *** (0.06)	9.59 *** (0.05)	9.87 *** (0.05)	8.00 *** (0.05)
ArtistExtInfoRel	0.05 (0.06)	0.03 (0.05)	-0.21 * (0.04)	-0.05 (0.03)
AfterRelease	-0.03 (0.06)	-0.03 (0.05)	-0.02 (0.03)	-0.01 (0.02)
ArtistExtInfoRel × *AfterRelease*	0.40 *** (0.08)	0.14 * (0.07)	0.34 *** (0.04)	0.13 *** (0.03)

Notes. Model: Neg. bin.; mo. = month, wk. = week ; 1,628 mo. obs. for I, II = (407 + 407) × 2; 6,512 wk. obs. for III, IV = (407 + 407) × 8. Pseudo-R^2: I – 37.6%, II – 36.4%, III – 44.9%, IV – 45.6%; shape parameter, α: I – .66, II – .52, III – .60, IV – a.46. Signif.: * $p < .10$; ** $p < .05$; *** $p < .01$.

In all models, *AfterRelease* was negative but not significant. A possible reason is that the control group had a larger listener base than the treatment group, as shown in Table 3.7. The increases were not enough to produce an average change for all 814 artists. Although PSM is used to match each treatment artist with the control artist, the scale of control candidates is limited (around 900 artists), therefore the matching may not that good enough. This coefficient may become positive, if we have much large control candidate group, and the treatment group's listener base were equal to or larger than the matched control group's though.

The coefficient of *ArtistExtInfoRel* was positive for the monthly dependent variables, but negative for the weekly ones. This occurred for the comparison between the treatment and control groups, both before and after external information was released. On a monthly basis at least, the treatment group achieved more music diffusion than the control group. When making a weekly comparison, the treatment group had no greater evidence of music diffusion. This indicates that the effect of external information may be limited as time passes.

Regardless of which dependent variable used, the targeted main effect *ArtistExtInfoRel* × *AfterRelease* was positive ($p < 0.01$). This suggests that external information release was not associated with a decline in music diffusion across the treatment and control groups, before and after the external information release. When monthly music diffusion was examined and compared to the control group, the

treatment group had a 49.2% increase in *Artist#Plays,* and a 15.0% increase in *Artist#Listeners* when an external information release occurred (*Artist#Plays*: ($e^{0.40}$ – 1) = 49.2%, *Artist#Listeners*: ($e^{0.14}$ - 1) = 15.0%). Weekly music diffusion had similar increase.

Although the DiD regression results in Table 3.7 indicate a significant impact of external information, the result was the average for all types of external information releases. However, the extent to which *Non-Music Content Information* affect diffusion is not clear. We further tested the effect of *Music Content* and *Non-Music Content Information* separately according to Equation 4, but only focused on the 407 artists who had external information. Table 3.8 shows the results for weekly music diffusion.

Table 3.8. Negative Binomial Regression Count Data Model Results for External Information

MAIN EFFECTS VARIABLES	MUSIC CONTENT INFO		NON-MUSIC CONTENT INFO	
	Artist#Plays / Wk (I) (SE)	*Artist#Listeners / Wk* (II) (SE)	*Artist#Plays / Wk* (III) (SE)	*Artist#Listeners / Wk* (IV) (SE)
Constant	9.23 *** (0.12)	7.35 *** (0.10)	9.63 *** (0.16)	7.94 *** (0.14)
ArtistExtInfoType				
News-Artist Life	Base case	Base case	Base case	Base case
News-Music-Related Info			0.10 * (0.13)	0.14 *** (0.05)
Tour, Concert			0.28 *** (0.06)	0.06 (0.05)
Live TV Show			0.31 *** (0.06)	0.35 *** (0.06)
Live Performance / Festival			0.09 (0.06)	-0.06 (0.06)
Single-Song Release	0.38 *** (0.07)	0.24 *** (0.06)		
Album Release	0.69 *** (0.06)	0.23 *** (0.05)		
Music-Video Release	0.96 *** (0.11)	0.84 *** (0.09)		
ExtInfoWeekAfter				
WeekAfter-1	Base case	Base case	Base case	Base case
WeekAfter1	0.47 *** (0.07)	0.15 *** (0.06)	0.12 * (0.06)	0.05 * (0.06)
WeekAfter2	0.51 *** (0.07)	0.19 *** (0.06)	0.11 * (0.06)	0.04 (0.06)
WeekAfter3	0.35 *** (0.07)	0.11 *** (0.06)	0.03 (0.06)	0.02 (0.06)
WeekAfter4	0.23 *** (0.07)	0.08 * (0.06)	0.001 (0.07)	0.01 (0.06)

Notes. Model: Neg. bin.; total obs. = 2,035; 985 *Non-Music Content Info* wk. obs. = 197 × 5; 1,050 *Music Content Info* wk. obs. = 210 × 5. Type 1, *News, Artist Life*: base case is *ArtistExtInfoType*. *WeekAfter*-1: base case is *ExtInfoWeekAfter*. We compared music diffusion for 1 wk. before and 4 wks. after info released. Shape parameters α: I – .57, II – .41, III – .42, IV – .11; pseudo-R^2: 1 – 48.7%, II – 49.6%, III – 55.6%, IV – 55.9%. Signif: * $p < .10$; ** $p < .05$; *** $p < .01$.

The results show the effects and persistence over time of *Music Content* and *Non-Music Content Information.* Type 1, *News, Artist Life*, is the base case for external information release. Not surprisingly, all types of *Music Content Information*

led to a significant increase in the *Artist#Plays* and *Artist#Listeners* dependent variable values. Among the three types, Type 8, *Music-Video Release*, resulted in the largest change in diffusion. This indicates that people are more attracted by 3D videos or stories compared to voice only, and also suggests why the music industry invests a lot in MTV videos. In addition, *Music Content Information* had a persistent effect on music diffusion in the month after information was released, and although it lessened over time, the trend was still increasing. This further verified why *ArtistExtInfoRel* for *Non-Music Content Information* was negative for weekly diffusion, when all 4 weeks after the external information released were considered (see Table 3.7).

Non-Music Content had some dissimilar effects compared to *Music Content Information*. Not all types had significant increases, though some led to music diffusion increases for both dependent variables. For example, Type 4, *Live TV Show*, had the highest effect, with 36.3% in *Artist#Plays* and 41.9% in *Artist#Listeners*. Type 2, *Music-Related Info* also has significant impact on music diffusion to new listeners. This kind of external information includes related news on music awards, a coming music album or tour arrangement. For example, listening to Adele increased sharply after her Oscar performance (see Figure 3.2). The other example is for a Swedish metal band, Sonic Syndicate. They released news about their new album in Week 15 and attracted 141 new listeners in the following week. Type 3, *Tour / Concert*, and Type 5, *Live Performance / Festival* seem to have increased total playing time, but this did not result in music diffusion to new listeners. A possible reason is that a tour and concerts are more likely to attract existing listeners, not new ones.

Another interesting finding is that the persistence of the effect over time of *Non-*

Music Content Information was only about 2 weeks after external information release occurred, although the coefficients for all 4 weeks after the event were positive. Compared to *Music Content Information,* the effect is small though (0.12 < 0.47, and 0.05 < 0.15, respectively). This further verified why *ArtistExtInfoRel* for *Non-Music Content Information* was negative for weekly diffusion, when we considered all 4 weeks after the external information released (see Table 3.7). Because of the sampling 1-hop subset of Last.fm users, only limited diffusion was observed, though it may have continued in the *N*-hop user groups.

3.5.2. Diffusion Diversity at the Geographic Level

Music diffusion at the macro-level also deserves comment. Some external information releases have obvious geographical bounds for their relevance, especially *Non-Music Content Information,* including TV shows, music festivals, and concerts, which occur in a specific geolocation. Diffusion diversity was tested for at the geographic-level with the fine-grained panel dataset, via this additional model, based on Equation 4.

$$\begin{aligned} DepVar_{it} = {} & Constant + \beta_1 CtryExtInfoRel_{it} + \beta_2 AfterRelease_{it} \\ & + \beta_3 CtryExtInfoRel_{it} \times AfterRelease_{it} \\ & + \beta_4 ArtistExtInfoType_{it} + \beta_5 ExtInfoWeekAfter_{it} \\ & + \beta_6 Artist_i + \beta_7 MajorLable_i + \beta_8 Genre_i \\ & + \beta_9 LongPopLast.fm_i + \beta_{10} LongPopBB_i + \epsilon_{it} \end{aligned} \quad (6)$$

The treatment group includes 46,200 U.S. users, and the control group has a matched set of 46,200 non-U.S. users. The treatment group variable was changed from *ArtistExtInfoRel* to *CtryExtInfoRel*. The dependent variables represent the cumulative listening counts for the treatment and control groups related to 199 artists with external information released in the U.S. The results are in Table 3.9.

Across the models that were used, the treatment variable *CtryExtInfoRel* was

positive and significant. This confirms the implied hypothesis that music diffusion has geographic bounds, not just for *Non-Music Content Information,* but also for *Music Content Information.*

Table 3.9. DiD Regression Results for External Information at the Geographic-Level for U.S. Users

Main Effect Variables	Monthly		Weekly	
	Music Content (I) (SE)	Non-Music Content (II) (SE)	MusicContent (III) (SE)	Non-Music Content (IV) (SE)
Constant	7.85 *** (0.27)	7.79 *** (0.27)	4.87 *** (0.15)	5.11 *** (0.14)
CtryExtInfoRel	0.26 *** (0.08)	0.16 * (0.08)	0.29 *** (0.07)	0.21 *** (0.07)
AfterRelease	0.48 *** (0.08)	0.04 (0.08)	0.10 * (0.06)	0.02 (0.06)
CtryExtInfoRel × AfterRelease	0.06 (0.11)	0.08 (0.39)	0.03 (0.08)	0.03 (0.08)
ArtistExtInfoType				
News-Artist Life	Base case	Base case	Base case	Base case
News-Music-Related Info		0.63 *** (0.15)		0.27 *** (0.08)
Tour, Concert		-0.01 (0.15)		-0.05 (0.08)
Live TV Show		0.45 *** (0.15)		0.43 *** (0.07)
Live Performance / Festival		0.08 (0.14)		-0.09 (0.07)
Single Song Release	0.24 * (0.13)		0.26 *** (0.07)	
Album Release	0.49 *** (0.12)		0.31 *** (0.06)	
Music Video Release	0.96 *** (0.19)		0.84 *** (0.10)	
ExtInfoWeekAfter				
WeekAfter-1			Base case	Base case
WeekAfter1			0.15 ** (0.05)	0.12 * (0.05)
WeekAfter2			0.28 * (0.05)	0.11 (0.05)
WeekAfter3			0.11 * (0.05)	0.04 (0.05)
WeekAfter4			0.005 (0.06)	0.001 (0.06)

Notes. Overall model obs.: I – 512, II – 376, III – 1,280, IV – 940. 105 artists had *Music Content Info*; 94 had *Non-Music Content Info*; I – 512 mo. obs. = (105 + 23 base case) × 4; III – 1,280 wk. obs. = (105 + 23 base case) × 10. Shape parameters α: I – .42, II – .46, III – .29, IV – .21; pseudo-R^2: 1 – 68.4%, II – 62.6%, III – 66.5%, IV – 74.5%. We only show regression results for U.S. *Artist#Plays* for treatment and control groups. *Artist#Listeners* had similar results. Results for other control variable are shown in Appendix Table B1. Signif: * $p < .10$; ** $p < .05$; *** $p < .01$.

The coefficient estimates for *AfterRelease* were all positive, but only significant in the case of *Music Content Information.* The reason is that *Music Content Information* seems to have had a positive impact on music diffusion, but the *Non-Music Content Information* variables were not as consistent in their estimated effects. The coefficients for *CtryExtInfoRel × AfterRelease* also were positive but not significant ($p = 0.43$ for the monthly *Music Content Information* variables; and $p = 0.48$ for the weekly *Music Content Information* variables – both in the DiD regression). A possible reason is that the existing user listening diversity across the selected 199 artists with external information was relatively high (*Artist#Plays/Mo,* mean = 17,730, SE

= 31,008; for *Artist#Plays/Wk*, the mean is 9,216, and the SE is 14,117). The average increase in diffusion was not big enough to produce a significant result. But the evidence for additional diffusion is further verification of the geographic effect.

The effects of *ArtistExtInfoType* and *ExtInfoWeekAfter* had similar results as music diffusion at the macro-level. But when they are compared to the estimation results at the macro-level, the coefficients of Type 2 *Music-Related Info* and Type 4 *Live TV Show* were larger. This indicates that TV show and music awards affected local users more compared to other countries.

For 76 external information releases in the U.K., there were 45 artists who had *Music Content Info*, and 31who had *Non-Music Content Info*. The data size was limited, so no converging regression results were observed. We further ran a DiD regression on the integrated U.S. and U.K. data. The results are shown in Appendix Table B2. The results interpretation is omitted here because it is similar to the U.S. results.

3.5.3. Listening Diversity at the Micro-Level

Next to report is the analysis of music diffusion at the user-level based on Equation 5. We focused on the new listeners of the 407 artists within one month after the artists' external information was released. Table 3.10 shows the results for the factors affecting listening diversity.

CtryExtInfo was positive and significant, so music diffusion was geographically bounded. For the social effects, even when all of the users' social relations for those who adopted an artist's music were considered, the effect on user listening choices was still smaller than the user's own listening behavior (*#Neighbors* = 0.04 < *ListeningBreadth* = 0.70 < *TasteSimilarity* = 1.10, p <0.01). External information releases had similar effects to what was observed at the macro-level. We assessed

what happened when many artists released information at the same time, and this negatively impacted an artist's music diffusion (-0.09; $p < 0.01$). Thus, the music industry must select a suitable time to release albums, to mitigate competitive effects from other artists.

Overall, the results indicate that external information, user geolocation and listening behavior, and social effects should be considered for the design of more effective, personalized music recommendations.

Table 3.10. Regression Results of Counting Data Model for Micro-Level Analysis

VARIABLES	COEEFFIECENTS
Constant	1.38 *** (0.05)
Geographic Effect	
CtryExtInfo	0.38 *** (0.04)
User's Listening Behavior	
ListeningScale	0.004***(0.00)
ListeningBreadth	0.70 *** (0.06)
TasteSimilarity	1.10 *** (0.07)
Artist Characteristics	
MajorLabel	-0.10 ** (0.02)
LongPopLast.fm	0.003** (0.00)
LongPopBB	0.004 ** (0.00)
ShortPopLast.fm	0.07 ** (0.02)
ShortPopBB	0.08 ** (0.02)
ArtistExtInfoType	
News-Artist Life	Base case
News-Music-Related Info	0.21 *** (0.03)
Tour, Concert	-0.04 (0.04)
Live TV Show	0.01 * (0.03)
Live Performance / Festival	-0.05 (0.03)
Single Song Release	0.21 *** (0.03)
Album Release	0.48 *** (0.03)
Music Video Release	0.03 * (0.05)
Social Effect and Crowding Effect	
#Friends	0.003*** (0.00)
#Neighbors	0.04 *** (0.00)
Artist#ExtInfoRelease	-0.09 *** (0.00)
Pseudo R^2	24.9%
Shape parameter, α	1.80
Note. Model: Neg. bin., 62,000 user-level listening obs. on 407 artists who had external information. Signif: * $p < .10$; ** $p < .05$; *** $p < .01$.	

3.6. Discussion

The music of an artist is a special information good, different from movies, news, and other instant information such as tweets and posts. Music attracts people's attention when it is released, the same as other types of information. But it also brings back their memories after a long period of low or no exposure to the artist. More interesting is that artists' social activities can also encourage music listening indirectly, even when there are no new songs or album that has been released. A distinct example as I cited in the beginning of this chapter, the terror attack that occurred at Ariana Grande's concert on May 22, 2017 brought a burst of listening interest to this artist on YouTube compared to regular listening. Although it was not suitable to promote her then, people's listening reacted to the terrorism events that occurred, especially in the open environment for information access.

In this research, we investigated how diffusion of an artist's music may be affected when new external information is released and enters a music social network. By analyzing Last.fm's data, this research was able to offer managerial insights that ought to be useful for music promotion and personalized recommendations in online music platforms.

First, external information from multiple channels had a positive impact on streaming music listening, and different kinds of information exhibited different impacts. New *Music Content Information*, for example, including new songs, new albums, and music video releases, was easier to use to attract new listeners in a short time. *Non-Music Content Information* was less effective in comparison though, while other media channels, including TV and newspapers, encouraged people to actively search for and listen to new artists. People may come across the artist via these media channels serendipitously. This may result in a listening cascade to the

artist via social networks.

Second, it is hard for an artist to always keep a high level of listening popularity, even for a superstar like Adele (see Figure 3.2). Existing listeners may lose interest if there are no new songs released. So how to attract new listeners and keep the current listeners' attention between new music releases is an important issue for music labels in the industry. *Non-Music Content Information* has shown its potential impact on streaming music diffusion. The study figured out that its effective time period is limited: only the first two weeks after the information release were significant for music diffusion to users in the dataset.

On one hand, this finding identifies the effective promotion period for the music industry and social network services. On the other hand, there may be still more diffusion that occurs due to social networks following the listeners in the sampled dataset. This kind of diffusion via peer influence may take a much longer time, such as half a year, especially for niche music (Garg et al. 2011). Only the listening behavior of seed users and their 1-hop friends were tracked though, so still more needs to be done to generate additional new knowledge about social network influences on music diffusion.

Third, streaming music diffusion has geolocation bounds, and external information is more likely to attract local listeners' attention. Although people can access whatever music they like online, their more limited access to external information may influence their choices, and this is true for both *Music* and *Non-Music Content Information*. The music industry may wish to pay greater attention to the release location, as they cultivate new consumers in different geographic regions for their artists. Although Last.fm has implemented an "Events" column to broadcast coming events for each artist, but there is no promotion, like local user recommendations

(www.last.fm/event/4393495+Fall+Out+Boy+M+A+N+I+A+Tour). For example, the U.S. rock band, Fall Out Boy, planned a tour stop for April 30, 2018 in Zepp@Bigbox Singapore. There were no further recommendations to listeners if they had no previous listening records for the band. Streaming services may consider to combine this kind of *Non-Music Content* release information, or collaborate with the industry, to increase the impact of external information on local listening.

Fourth, how to improve services to attract new users and keep loyal of existing users is one key task for streaming music services. From the analysis of music diffusion at the macro- and micro-levels, we found that external information, user geolocation, and listening behavior can be leveraged to a greater extent to improve personalization. For music consumers, this study offers a way to improve personalized music recommendations, by leveraging information from various channels for the music platform. For the music industry, this research also offers managerial insights on target consumer selection for more effective artist promotion.

3.7. Conclusion

Streaming music services have presented opportunities to promote artists and digital products online. Many music labels choose to release music for free listening before CDs are released. They aim to attract attention from existing and new consumers. Music social networks are especially interesting as semi-closed social environments: they encourage internal sharing of social information, and are open to external information that can influence their users too. With the complexity of this social environment, it is important for music social network providers to understand how the diffusion of music works in their networks. They may find that there is a hidden source for consumer engagement and higher profit.

This research drew on literature related to streaming music diffusion theory and methods and studied the impacts of external sources of information on the streaming music diffusion in online social networks. The empirical analysis dissected music diffusion from a different perspective, and the findings contribute useful insights for the music industry on music and artist promotion, especially in the open environment of the Internet.

There are some limitations to consider though. First, it is hard to capture all of the relevant external information for artists. We only considered a single channel for each artist. Some had more frequent releases, and we may not have been able to observe them all. Some artists also arranged tours just after a song release, enhancing their local diffusion. The effect of a new song on streaming music diffusion should reflect the cumulative impact of all external information releases. We think that this probably did not affect our estimation results for *Non-Music Content Information* very much, though it may have led to slight over-estimation of the impact of the effect of *Music Content Information*. Also, we cannot guarantee that there is absolutely no mixing of influence and discovery since we only observed the diffusion of music over a limited time range. To distinguish these effects, we need data over a longer period, and we will also consider effects of social relations and capital, and weak and strong social ties.

In addition, the estimation work was performed on a subset of Last.fm's data, so there may be selection bias. We plan to study music diffusion at the macro-level across the entire Last.fm platform, as a way to more comprehensively assess the effects of external information discovery. Last, we only considered if a user was from the country where an artist's external information was released. For our control group users, we did not do more fine-grained clustering, such as for country

traits, including language differences, and cultural and physical distance. U.S. artists may find it is easier to attract U.K. listeners compared to those in China, for example, due to language and cultural similarities. Based on data acquired from the Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) in France, the plan is to add these distance variables to more deeply analyze the geolocation effects.

By leveraging the insights obtained, additional research can be done on how to use them for the design and improvement of personalized recommendations for music social network settings.

Chapter 4. Two-Sided Value-based Music Promotion and Recommendation

4.1. Introduction

Music popularity development and music diffusion via streaming music services that were explored in the previous two chapters, are the understanding of streaming music listening, such as what kinds of music can easier become popular, what kinds of artist-related information can help the music diffuse on the platforms. All these insights are learned from analyzing on people's listening logs, how to use them to serve and improve the streaming music ecosystem is the ultimate goal. This chapter shows how these streaming platforms work in the center, and explores what can be done to shed light on making the streaming music ecosystem healthier. This is based on what was learned from the analysis of streaming music listening.

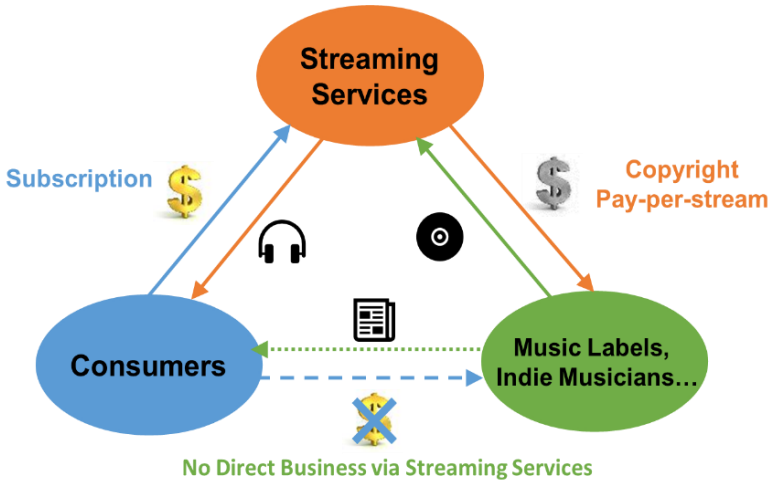
4.1.1. Streaming Music Ecosystem

Streaming music services, as the most popular third-party for music adoption in today, are worthwhile to learn about and understand more deeply so they can be improved. A simple description of the streaming music ecosystem is presented in Figure 4.1. It includes the three most important elements in the system: streaming services, consumers or listeners, and music labels, including independent musicians, songwriters, and producers.

Streaming service providers are middlemen that connect consumers with music. On one side, music labels and indie musicians use streaming services to upload their songs or albums. The service providers pay copyright fees based on *pay-per-stream* pricing, similar to paid software (Resnikoff 2016). On the other side, consumers subscribe to streaming services to listen to music. This enables them access music tracks and albums. They are not limited to any specific website or platform. They

can access to music based on their subscriptions. In general, the more consumers listen, the more the music labels benefit from their streaming services (Sanchez 2017, 2018).

Figure 4.1. Conceptual Framework for Streaming Music Ecosystem



How much the music labels can gain from the streaming music services per stream is changing too. Recent updates for January 2018 offer the following information: Apple Music pays $0.0078, Google Play $0.0061, Spotify $0.0040, Pandora $0.0013 and YouTube $0.0007. The payments have increased slightly compared to 2017 (Sanchez 2017, 2018). The market revenue is appreciable, and the three major labels (Universal Music Group, Sony Music Entertainment, and Warner Music Group) made an estimated $14.2 million per day. This amounted to $5.2 billion from streaming services in 2017, about 75% of the whole streaming music market, but with less than 10% of overall music artists included (Sanchez 2018).

Figure 4.1 shows that there is no direct business between the consumers and music labels in the current streaming music ecosystem. That is why we used different colors to represent money flows between the middleman and the two clients. Subscription fees do not yield the value that copyrights suggest though. The music labels' revenue is lower than expected (Parisi 2018). This is not critical for famous

artists with large listener bases, however, it is serious for indie musicians and independent songwriters, when they want to live on music-related income but do not have a stable listener group to support them.

If famous or budding artists want to promote their music via streaming services, they need the knowhow of streaming platforms to assist them. This is so they know about who their listeners are, and when and how many times they listen to their music. This will help the artists and labels to design promotions to attract new listeners. Music recommendation may be what they need to promote their products on streaming services. Recommender system capabilities, in this way, can serve as a tool to improve the profitability of streaming services too. Good recommendation services will make listeners more apt to stay with the service. This may also aid in converting free-listening consumers into subscribers, if their willingness-to-pay rises high enough. By the same token, good music recommendation services can also increase ads revenue from increased listening (Wlömert and Papies 2016).

4.1.2. Recommender Systems Design and Value

Recommendations assist consumers in finding interesting items or products by reducing their search costs (Brynjolfsson et al. 2003). In addition, recommender systems are widely used in e-commerce websites, such as Amazon, Netflix, Taobao, JD, and Hulu, because of the significant business value they create for online retailers (Culnan et al. 2010, Gomez-Uribe and Hunt 2016). The availability of massive historical user online behavior data also has brought the chance for vendors to explore the existing and potential connections between consumers and the products they are most likely to purchase. Nevertheless, recommender systems design in a specific environment for a specific requirement still has a long way to go.

The design of existing recommender systems has focused largely on computational approaches and how to improve recommendation accuracy (Adomavicius and Tuzhilin 2005). The target typically is to maximize user satisfaction by identifying the most relevant items for them. However, the selected items may not necessarily be the ones that maximize the value that the service providers can obtain, if business value, such as firm profit, the change in actual sales, conversion rates, and click-through rates, is considered (Zhang et al. 2016, Zhao et al. 2017).

Third-party platforms, such as Last.fm for music promotion and listening, act as a bridge between consumers and providers. Because of data and business confidentiality though, they rarely share knowledge about their customers or how their recommendations are designed to support the digital music product providers. The widely-used commercial recommender systems usually are designed for personalized recommendations but only for customers. Little attention has been paid to the value for the provider side, not to mention for the personalized promotion of a specific artist.

Although commercial recommender systems design needs to catch up, academic research in IS has demonstrated the importance of balancing the benefit of customers and providers in recommender systems design (Adomavicius et al. 2017, Panniello et al. 2016, Zhang 2017). Also academic attention to recommender systems design is beginning to shift to how to design recommendation algorithm by considering both kinds of interest. For example, Zhang et al. (2016) integrated the surplus derived from marginal utility and marginal cost in a collaborative filtering model. The model aimed to maximize both the provider's and users' surplus. Zhao

et al. (2017) also leveraged utility theory to improve complementary product recommendations for existing consumers. These are good starting points to improve existing recommender systems.

4.1.3. Summary of the Study

Similar to recommender systems design for other products and information, most existing work on music recommendations has been focusing on the consumer side. The new research has pointed out a different perspective to improve online music recommendation by considering the value that can be obtain on both sides. This is useful in the presence of the increasing streaming music market. In this research, a music recommendation algorithm is proposed by considering the value for service providers and consumers.

We developed a two-sided value-based streaming music recommender system. The system combines the value yielded for the music industry and consumers in an integrated model. For the music industry, the system aims to increase the conversion rate of potential listeners to adopters. At the same time, for consumers, the system aims to improve their satisfaction related to the recommendations they receive. The system design involves a new recommendation algorithm, by leveraging the insights gained about music popularity and diffusion in Chapters 2 and 3.

The basic idea of this research now follows. Promotions are modelled for an artist in a specific period based on identifying new listeners. These new listeners should have never listened to the artist. They can also be light and intermittent listeners, who have stopped listening to the artist for a long time, but may be called back by artist-related information. They are usually ignored by existing recommendation methods because of their listening records. The goal of this research is to help the music industry to extend their listener base. At the same time, another goal

is to retain loyal listeners by keeping their satisfaction up through social media services.

We used one year's listening records for over 15,000 Last.fm users to train and test a proposed two-sided value-based recommendation model. Compared to the most widely-used recommendation algorithm, collaborative filtering, the results of this study show a clear increase in the conversion rate of recommendations by considering both sides' value and other factors, including time, geolocation, external information and listening behavior.

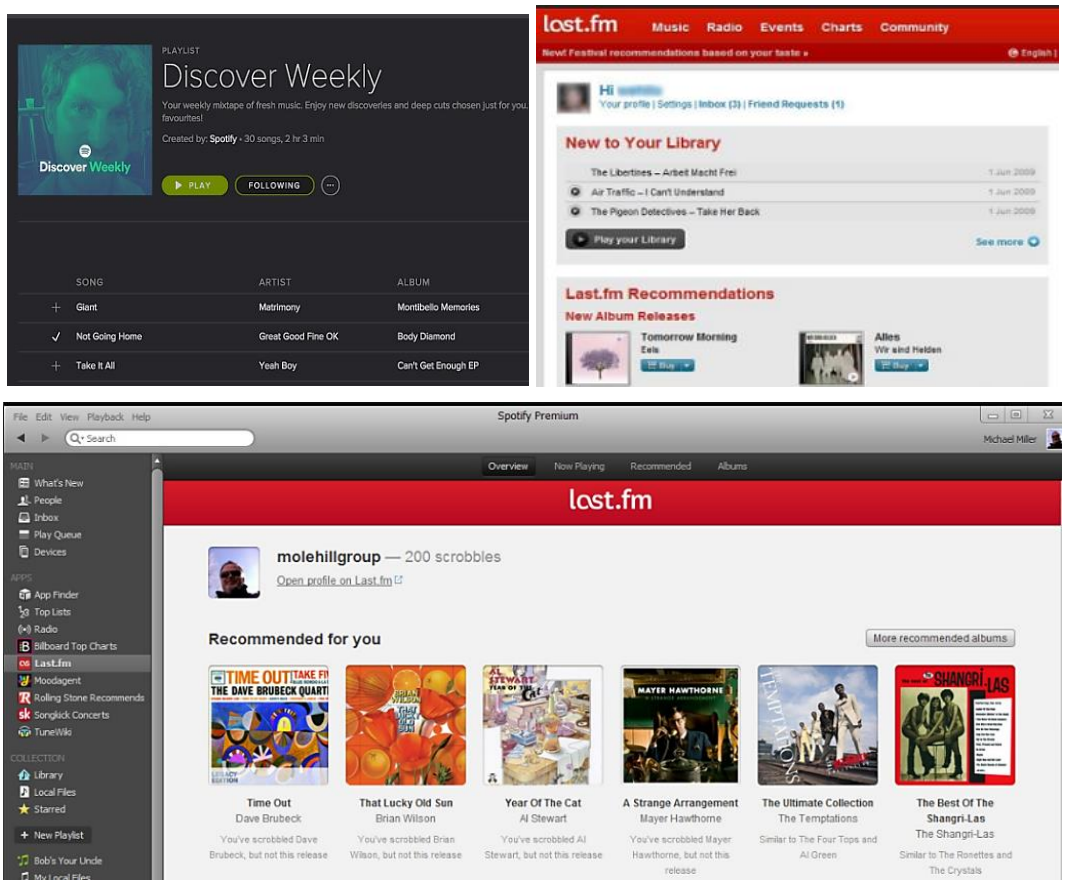
4.2. Literature Review

With the increases in online items and products, recommender systems have become commonplace in the online consumer environment, covering movies, music, books, retailing, on-demand video, and so on. Netflix, for example, reported that about 75% of the content watched by its subscribers was suggested by its recommender systems (Amatriain and Basilico 2012). Amazon also boosted its sales by 29% in its second fiscal quarter of 2012 after incorporating its recommendation mechanism into its website (Mangalindan, 2012).

Figure 4.2 gives some snapshots of the recommendation services supplied by Spotify and Last.fm. These streaming services have their own recommendation mechanisms, and some of them are trying to collaborate with each other. For example, Spotify has supplied trans-boundary recommendations related to Last.fm for their premium consumers since 2014. Spotify's users also can use the Last.fm application to view personalized recommendations based on their Spotify activity—so they can play new music on Spotify. Last.fm automatically creates a playlist of twenty tracks similar to the any song which is dragged from the Spotify library. Although Spotify is the hottest streaming music service in the market now and

claimed over 50% of the streaming market since 2017 (Richter 2017, MIDiA 2017), it is still in the process of opening up its service to third-party apps. Last.fm is an important collaborator which is worth to because of its recognized role as the "music recommendation experts" (Katz 2014).

Figure 4.2. Snapshots of Recommendations Supplied by Spotify, Last.fm and Their Collaborating Services



Note. Pictures were obtained from Google Images.

In this section, a summary of the literature on the techniques used in music recommendation and their limitations is presented. Then, related work on utility theory and the value that recommendations can create are also reviewed, and then we comment on their recent applications to recommender systems.

4.2.1. Music Recommendation

Music recommendation techniques can be categorized into four approaches: content-based, collaboration-based, context-aware, and hybrid-based.

Content-based methods, the assumption behind them is related to user listening preferences, which can be extracted from the acoustic content of the music the user has listened to. The key idea of content-based approaches is to extract measurable information directly from audio signals or lyrics, to represent a song using a vector or metrics of musical semantics (as in the research on music track popularity), and then recommend new songs to a user based on the similarity between the new songs and the user's listening preference (Cano et al. 2005).

Collaborative-based methods refer to the observation that similar users share similar listening preferences (Schafer et al. 2007). They estimate the similarity between users based on their listening history and recommend songs by referencing the preference similarities. The famous collaborative-based methods include *k-nearest neighbors* (Desrosiers and Karypis 2011) and *matrix factorization* (Koren et al. 2009).

Context-aware methods rely on the context besides the music content itself. Context refers to all music-relevant information but not music content itself. It is often used to refer to the user's context, the artist's context, or the usage context, including location, time, weather, artists' correlations, and the users' activities, including like, dislike, tag, and share (Adomavicius and Tuzhilin 2015, Knees and Schdel 2013). This category of music recommendation is attracting more attention as big data are available and web technology is changing the music listening environment.

Hybrid-based methods combine the techniques from the three basic approaches, such as content-based and context-aware by merging music semantics with venue information into a latent topic model (Cheng and Shen 2016) or combining content-based and collaborative-based approaches (Wang and Wang 2014).

As listening to streaming music and the resulting sharp increases in music market share, the music industry and independent (indie) musicians have paid more attention on how to collaborate with streaming music services and design strategy to maintain their effectiveness in the market (IFPI 2015, 2017). In this situation, they may consider several aspects when designing music promotion strategy. First, the use of *pay-per-stream* may force them to not just consider recommendation accuracy, but also consumer affinity to recommendations (e.g., how many times users listened to the music). Second, today we are in a free and open environment for accessing a massive amount of music-related information, Salo et al. (2013), Schdel (2011). Ren and Kauffman (2018) have identified the impact of the interplay of multiple media channels on music popularity and music diffusion for a streaming music service. The music industry, as a result, may want to leverage the strength of web technology to promote their artists.

We compare four categories of recommendation methods related to streaming music recommendations (see Table 4.1). Popularity bias reflects the long-tail phenomenon in music listening, and this issue exists in collaboration-based methods. Web technology, involving the interaction of multiple channels, emphasizes context-based methods, but may also involve hybrid-based methods. All of the four categories focus on increasing prediction accuracy, however, none of them has considered the business value of music recommendation. In fact, there is no guarantee that better prediction performance can translate into higher conversion rates for revenue per user (Belluf et al. 2012). Usually existing recommendation methods use a binary variable to indicate success. Some use a rating value (e.g., 1 to 5) to indicate the satisfaction level of the recommendation (Cheng and Shen 2014, Koenigstein et al. 2011). Rating alone, however, cannot assess satisfaction through the use of a

measure with a number for the music industry in the long term. The business and profit patterns may offer clues for music recommender systems designs though.

Table 4.1. A Comparison of Music Recommendation Methods

CHRATERISTICS	CONTENT-BASED	COLLABORATION-BASED	HYBRID-BASED	CONTEXT-BASED
Popularity bias	×	✓	Depends	Depends
Web technology	×	×	Depends	Depends
Accuracy	✓	✓	✓	✓
Business value	×	×	×	×

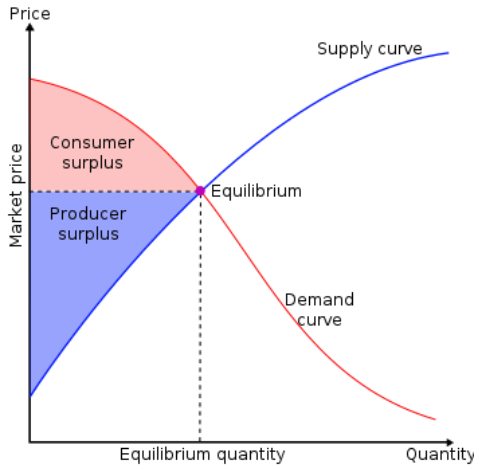
4.2.2. Utility Theory and the Value of Recommendation

Some IS related research has demonstrated the existence of important economic side-effects of recommender systems (Jannach and Adomavicius 2017, Adomavicius et al. 2017). For example, personalized recommendations can help increase consumer willingness-to-pay for digital music promotions. Some other work has further analyzed the relationships between provider profit and consumer surplus, as well as satisfaction and predictive accuracy. Panniello et al. (2016), for example, showed that there is a positive impact of the balance between accuracy and profit related to consumer online purchasing behavior, and it does not hurt the extent of their trust. The existing work has highlighted the necessity to balance the utility of the providers and the consumers in recommender system design.

In economics, *utility* is a quantitative proxy measure of one's preference over a set of goods or services. It represents the satisfaction experienced by the consumer for a good. It is an important concept that serves as the basis for rational choice theory (John 2000). Since one cannot directly measure the benefits that people can gain from the consumption experience, instead economists use satisfaction or happiness from a good or service consuming and have devised ways of representing utility in terms of measurable economic choices.

In recent several years, researchers have tried to use utility theory to estimate consumer surplus and provider surplus to assist search and recommendation for various products, for example, hotel search (Li et al. 2011), e-commerce (Zhao et al. 2017), P2P lending, and freelancing (Zhang et al. 2016). As shown in Figure 4.3, *surplus* is the net benefit associated with buying or selling a good or service. *Consumer surplus* is the amount of utility that the individual experiences beyond the amount that he pays (the price), while *provider surplus* is the amount that the provider earns beyond its expenses (the cost). How to measure the utility on both sides is the essential purpose of this research.

Figure 4.3. Consumer (Red) and Producer (Blue) Surpluses on a Supply and Demand Chart



A number of other research works considered the application of utility theory to music recommendation. Park et al. (2006) used utility theory in their context-aware music recommendation algorithm and estimated the probability of listening to music using a Bayesian network approach. Adomavicius et al. (2017) implemented controlled experiments to estimate the effects of recommender systems on consumer willingness-to-pay for digital songs and demonstrated the existence of important economic side-effects of personalized recommender systems.

The characteristics of streaming music consumption, and its profit patterns may

be an obstacle for how to calculate two-sided utility. On the consumer side, Varian (2010) pointed out the importance of the theory of the consumer, for which utility is a way of describing consumer preferences. Music is a kind of experience good, so consumers cannot gain any utility until they have listened to it. Therefore, according to the theory of the consumer, consumer utility for a given music track represents their satisfaction from a listening experience less the search cost they expend to find the recommended music. How to measure this kind of utility is a challenge for empirical research because different listeners may experience different satisfaction levels and different search costs for the same music product.

On the provider side, the theory of the firm and utility are not applicable. This is because the provider is involved in the production and marketing of music, not its consumption. Providers care mostly about the profit earned beyond the cost of selling music to the appropriate consumers and listeners. This makes the theory of the firm (Varian 2010), in which producers attempt to maximize sales revenue less the cost to produce them, much more relevant for analysis.

Creating business value is the target of product promotion and recommendation (Gomez-Uribe and Hunt 2016). Netflix, for example, has verified that personalization and recommendation can help to maintain subscribers' loyalty and reduce the number of members who decide to no longer use a service. Good recommendations have helped Netflix to create business value by saving the company more than $1 billion per year in its effort to acquire new consumers, simply because it has reduced the incidence of subscription cancellations (Gomez-Uribe and Hunt 2016).

Netflix has used the subscription fees it earns due to the recommendations it makes to consumer to measure the beneficial effects on business value. A possible

reason for this approach is that Netflix recognizes that it is hard to capture the benefits that are produced for each video that its users view. This is somewhat less true for other e-commerce products, such as clothes, cameras and food, where it is possible to count the number of units sold and the underlying cost to support such sales. For streaming music (same as Netflix), in contrast, it is not so easy to measure the revenues and the average costs of supplying the music. Because streaming music is not priced, nor is it obvious what the costs are since music that is acquired for streaming involves royalties for the music artists and fees for their music labels. Therefore, for the provider side at least, it makes sense to calculate the business value of recommendations that enhance consumption, while ignoring the associated costs. These are difficult or impossible to observe without direct access to the music labels' data sources also, which is a major roadblock for more in-depth empirical research.

4.3. Proposed Method

Next we describe the context and target for streaming music service promotion, and propose a model based on the theory of the consumer and utility and the theory of the firm for producers, along with the value of information.

4.3.1. Problem Description

The main goal is to design a personalized promotion method for an artist, to identify the most potential listeners for the artist when the music labels or independent musicians have a promotion plan via streaming music services. For example, imagine that a music label launches an artist promotion on Last.fm, so as to attract new listeners of the artist and increase the streaming track volume as much as possible in one month. The principle is to achieve effective listener selection to realize a high return on investment.

This is a different perspective from traditional music recommendation, which is to find the most suitable artists, songs, or albums for a listener. Here, the target moves from consumers to providers. Similar research is related to influential user identification (Ren et al. 2014, Trusov et al. 2010), however, there are very few research articles that have focused on music recommendation in a specific time period for a specific artist. This is a useful strategy for the music industry when there is a need to promote music products. This is especially true for niche music or independent musicians, whose music products are not easy to find or have been recommended by the widely-used collaborative filtering algorithm, because of popularity bias (see Table 4.1). Figuring out how to leverage the strength of the streaming platform to assist searching for potential listeners in a short time can increase their streaming over time. For example, Chapter 3 study found that the one month after the release of an artist-related external information is a suitable time period in which to do music product promotion.

The challenge here is to determine how the artists will know who the targeted listeners are. Here, the listener represents the consumer, the artist represents the provider (including music labels and streaming services). In other words, this allows the design of personalized recommendations for an artist to identify the most potential listeners by considering both sides' value.

Utility theory can be applied to measure the satisfaction of listeners for the music they consumer, but it is not possible to use the same theory for the producer side, including music artists and the labels that produce and sell their music content. To make this model workable for streaming music promotion, we had to simplify the approach to measurement for listeners on one side, and for artists and labels on the

other. For the listener side, we use utility theory but ignore the search costs that the listeners must pay for finding the music.

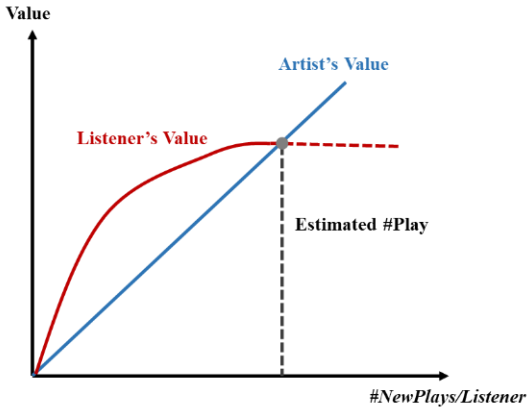
For the artist side, we borrowed the thinking that Netflix uses by comparing what happens to listening "with" (in the presence of) versus "without" (in the absence of) the availability of recommendation service (Gomez-Uribe and Hunt 2016). With recommendation services, the artist and music label side can predict the possible business value that targeted listeners may generate for them. This also can be regarded as a matter of the value of information, for which the common approach is to assess the value of consumer actions in the presence or the absence of advertising, for example, which is a choice made by a producer or seller to induce high purchase levels and revenue. As a result, the possible revenue that targeted listeners yield when they consume an artist's pay-per-stream music content is used as a way to measure the value of a recommendation service.

The proposed model is flexible in its approach to measures the value that recommendation services can create on both sides, and so it is a two-sided value model. Next, additional details about the proposed model are provided.

4.3.2. A Two-Sided Value Model

A description of two-sided value is shown in Figure 4.4. The red line represents the listener's value and the blue line is the artist's value. The *x*-axis is the time a listener has played an artist's music. The *y*-axis represents cumulative value. After some time passes, the marginal value for a listener to listen to additional music will monotonically decrease or even decline to 0, and the listener may no longer add to their listening repertoire. At the same time though, the artist's value is likely to continue to increase as new listeners discover their music.

Figure 4.4. Listener (Red) and Artist (Blue) Value for Music Listening



The goal is to find the listening level for each candidate listener *i*, to maximize the total value of the top-*N* recommended listeners for artist *j*, as Equation 1 shows:

$$TotVal_j = \text{Max}_{TopN} \sum_i \left(ListenerVal_{ij}(q_{ij}) + ArtistVal_{ij}(q_{ij}) \right) \quad (1)$$

TotVal(q), total value, is a function of the quantity *q* of streaming music consumed that produces value. Here, $ListenerVal_{ij}$ is the listener *i*'s value gained from listening to artist *j*, and $ArtistVal_{ij}$ is the artist's value gained from listener *i* 's streaming of her music. q_{ij} represents the amount of listening by the listener *i* to artist *j*. This function indicates that the selected top-*N* candidate listeners will yield the maximum total value by considering both sides' value.

Artist's value, $ArtistVal_{ij}(q_{ij})$. Morgan and Rego (2006) pointed out that customer satisfaction is a good predictor of firm business performance. Therefore, for the assessment of artist value $ArtistVal_{ij}(q_{ij})$, two aspects are considered. One is that artists hope to attract loyal listeners who will continue to stream the artist's music over a period time, and not just sample it and not return (Luarn and Lin 2003). This revenue source can be calculated with the pay-per-stream quantity as $a_0 q_{ij}$. The other source is from the potential listeners, their loyal listeners have the potential to affect other listeners who become aware of new artists via social influence

(Garg et al. 2011, Ren et al. 2014). It is labelled as p_{ij}, the number of the listener *i*'s friends who have not listened to artist *j*. Thus:

$$ArtistVal_{ij}(q_{ij}) = a_0(q_{ij} + Pr(q_{ij}) \times p_{ij}) \quad (2)$$

Here, $Pr(q_{ij})$ is the probability for how much user *i* listens to artist *j*. In this case, if two listeners have same listening quantity, then the one who has more social friends would be more attractive for an artist promotion, because of the potential social effect in long term.

Listener's value, $ListenerVal_{ij}(q_{ij})$. For the listener's value, $ListenerVal_{ij}$, there is no standard measure. Traditional music recommendation has used a binary or rating variable to represent the satisfaction level of the listener for the recommended music. This offers short-term feedback on the recommendation and ignores the diverse listening behavior. For example, the traditional recommendation estimates when two listeners give the same rating to an artist. One may listen 100 times, while the other one may only listen 5 times. Although they are considered identically in traditional recommendation, the strength of the value is different for each of them and also different for the artists.

Rios et al. (2013) indicated, utility is inherently governed by the *law of diminishing marginal utility*: when a person increases her consumption of a product, there will be a decline in the marginal utility that she derives from consuming each additional unit of the product. This is also true for music listening. If utility does not decline as listening quantity increases, this will be surprising, since most listeners stop listening over time as their utility declines. This approach does not calculate marginal utility directly. Instead the model estimates the possible listening quantity before the listener stops listening.

There are various functional forms for utility. We used King-Plosser-Rebelo (KPR) utility, based on Zhang et al.'s (2016) study on e-commerce product recommendation. The listener's value is shown in Equation 3, which yields 0 when the listening time is zero, $ListenerVal_{ij}(0) = 0$, based on:

$$ListenerVal_{ij}(q_{ij}) = a_{ij} \ln(1 + q_{ij}) \quad (3)$$

Here, a_{ij} is the weighted effect of user *i*'s utility for artist *j*'s music. It can be a binary rating or a probability value for listening utility.

Equation 1 can be transformed to yield:

$$TotVal_j = \text{Max}_{TopN} \sum_i (a_{ij} \ln(1 + q_{ij}) + a_0(q_{ij} + Pr(q_{ij}) \times p_{ij})) \quad (4)$$

Here, a_0 is the baseline value that an artist can gain through user listening, which can be represented by pay-per-stream revenues. We set $a_0 = 0.004$ based on the average pay-per-stream revenues that are generated by the major streaming music services. The detailed estimation of a_{ij}, q_{ij}, $Pr(q_{ij})$, and how the recommendation approach works is implemented, are presented next.

4.3.3. Model Specification

To calculate total value, Zhang et al. (2016) assumed that the consumption quantity value (q_{ij}) is a random variable. This is independent of weighted utility a_{ij} at first, and then they used the Poisson distribution to describe the consumption quantity of each consumer, as well as the collaborative filtering method to estimate utility. But the estimation results show that the quantity is associated with the utility level. Because of the nature of the Poisson distribution, $\widehat{a_{ij}} = \overline{Q_{ij}}$. Music listening has similar but also different characteristics. Listening quantity q_{ij} in this study is correlated with utility a_{ij}, and also may be affected by other factors, such as listening context. The difference is that the mean of the Poisson distribution is equal to

the variance, which is not suitable for music listening though. The dataset used in this study is over-dispersed: the variance is larger than the mean in the distribution. (See *#NewPlays/Listener* in Table 4.3.)

Based on these observations for the characteristics of streaming music listening, for each listener, the utility and revenue estimation results can be combined to create a unified value measure in this study. Moreover, collaboration-based and context-aware functions are estimated, by considering personal listening behavior and artist context information based on what we learned from music popularity and diffusion analysis, as shown in Equation 5. Combining a_{ij} with q_{ij} can reduce the missing potential listeners, for example, when $a_{ij} = 0$.

$$\begin{aligned} TotVal_{ij} &= \ln(1 + q_{ij}) + a_0(q_{ij} + Pr(q_{ij}) \times p_{ij}) \\ q_{ij} &= f(a_{ij},\ ListeningCharacteristics_i,\ ListeningContext_{ij}) \end{aligned} \tag{5}$$

Next, we explain the estimation and recommendation details.

Estimation of q_{ij}. The insights gained in this thesis demonstrate the variables that are useful in music promotion. They include music popularity, artist-related external information (*Music* and *Non-Music Content Info*), a listener's consumption behavior, and so on. These characteristics and context information drive service allocation in a specific listening context, which makes it possible to estimate the amount of listening potential listeners do.

To allow listening quantity to vary, q_{ij} is assumed to be a function of listener utility, characteristics and context (see Table 4.2). For example, all else equal, when artist-related *Non-Music Content* information is released in U.S. (e.g., a live TV show), listeners there may listen more times than those who are abroad. The empirical work in this thesis demonstrated that different kinds of artist-related information can affect a larger number of different new listeners on a more persistent

basis. In this study, the promotion for two types of external information were designed separately, based on the principle of *contextual pre-filtering* in context-aware recommender systems (Ricci et al. 2010).[8] For each type, *Music Content* and *Non-Music Content* information, we further characterized its new listeners by a set of features and listening context, and the listening quantity under either type is a function of them.

Table 4.2. Covariates Used for Listening Quantity Estimation

NOTATION	CONTEXT INFO	VALUE
Collaboration Estimation		
a_{ij}	User *i*'s weighted utility for artist *j*'s music	Numeric
Listening Characteristics		
$ListeningScale_i$	# of artists user listened to	Numeric
$ListeningBreadth_i$	User's diversity of music listening across artists	Numeric
$TasteSimilarity_{ij}$	Taste similarity of user for artist's music	Numeric
Context Information		
$MajorLabel_j$	Whether artist is connected with major music label	Binary
$LongPopLast.fm_j$	Top chart popularity on Last.fm, 2005 to 2013	Numeric
$LongPopBill_j$	Top chart popularity on Billboard Hot-100, 2005 to 2013	Numeric
$ArtistExtInfoType_j$	Type of external info released on an artist	Category
$Artist\#ExtInfoRelease_i$	# of artists with external info when user listened	Numeric
$CtryExtInfo_i$	(1, 0) if user country is the U.S.; (0,1) if user country is English-speaking; (0, 0) otherwise	Binary

Note. Listening characteristics variables are the same as in Chapters 2 and 3. $CtryExtInfo_i$ was adjusted to consider the effects of different languages.

The empirical study in Chapter 3 showed that the amount of listening that new listeners are observed to do can be estimated using a negative binomial distribution-based count data model. Negative binomial regression analysis is a method for predicting the value of a count variable for a set of predictor variables. It was used to predict the number of software bugs in previous work (Yu 2012). A negative binomial model is used with multiple covariates to perform an estimation of q_{ij}.

[8] *Contextual pre-filtering* means that contextual information drives data selection or data construction for a specific context. In this study, we constructed different datasets and estimated different models for *Music Content* context and *Non-Music Content* for the purpose of context recommendation.

For each artist-listener pair, the observation unit is (q_{ij}, X_{ij}), where $q_{ij} \geq 0$ is listening quantity and X_{ij} is a $k \times 1$ covariate vector to describe the listening characteristics and context information (as listed in Table 4.2). When either type of external information occurs, we focus on how the listening quantity varies with the covariates. The conditional mean μ and variance σ (Cameron and Trivedi 2013, Beaujean and Morgan 2016) are given by:

$$\mu = E\{q_{ij} | X_{ij}\} = \exp(X_{ij}^T \beta) \quad (6)$$

$$\sigma = \mu + \alpha\mu^2 \quad (7)$$

For each type of external context, β is a $1 \times k$ set of parameters to be estimated in Equation 6. Also α is the dispersion parameter of the negative binomial model. It involves a density function as in Equation 8 (Cameron and Trivedi 2013), which represents the probability that $q = q_{ij}$ when the listener's observation is X_{ij}.

$$Pr(q = q_{ij} | X_{ij}) = \frac{\Gamma(q + \alpha^{-1})}{\Gamma(q+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{q} \quad (8)$$

The function Γ(·) is the gamma function, and its definition and characteristics are given in Appendix C.

Maximum likelihood estimation (MLE) was implemented to estimate the parameters $\mu, \alpha,$ and β.[9] For each type of external information, 80% of the artist-listener pair data was used to estimate the model's β parameters. This allows estimation of the expectation $\widehat{q_{ij}}$ and $Pr(q_{ij})$ for a new listener. The expectation is used to calculate the total value.

Estimation of a_{ij}. The parameter a_{ij} describes the weighted utility of listener *i* for artist *j*, so we estimated its value for each artist. For artist *j*, we used a listener's

[9] $\text{Log } L = \sum_i \sum_j \{\log\left(\frac{\Gamma(q_{ij} + \alpha^{-1})}{\Gamma(q_{ij}+1)\Gamma(\alpha^{-1})}\right) - (q_{ij} + \alpha^{-1})\log(1 + \alpha \exp(X_{ij}^T \beta) + q_{ij} log\alpha + q_{ij} X_{ij}^T \beta)\}$

k-nearest neighbors (KNN) with a collaboration-based method (Desrosiers and Karypis 2011) to estimate the potential listening utility of a new listener. KNN is a memory-based recommendation method to estimate binary or rating feedback for a recommendation and supports listener- and artist-based estimation. The dataset has a relatively small number of artists though, so using artist similarity to estimate the number of listeners may result in somewhat more bias than listener similarity would. Thus, we used a listener-based KNN method to estimate general utility.

Pearson correlation was implemented to calculate the similarity, *sim*, between a potential listener *i* and another listener $v \in V$, V is set of existing listeners of artist $j \in J$, J is the set of artists in this study, with *r* representing the listening time to each artist:

$$sim(i,v) = \frac{\sum_{j \in J}(r_{i,j} - \overline{r_i})(r_{v,j} - \overline{r_v})}{\sqrt{(r_{i,j} - \overline{r_i})^2}\sqrt{(r_{v,j} - \overline{r_v})^2}} \quad (9)$$

Then, a_{ij} is estimated based on the weighted sum of the top-*k* neighbors' listening with similarity to listener *i*. This is based on collaborative filtering, which indicates that users who have similar taste will likely adopt similar products. We selected k = 15 based on the quality of the estimation performance. (See Appendix Figures C1 and C2).

$$a_{ij} = \frac{\sum_{v \in V} r_{v,j} \times sim(i,v)}{\sum_{v \in V} sim(i,v)} \quad (10)$$

Traditional music recommendation uses Equation 10 for recommendation, which is based on previous listening records for the artist but ignores the effects of the listening context. Examples are external and geolocation information that impacts music listening, as confirmed in Chapter 3.

Listener recommendations for an artist. Traditional music recommendation involves selecting a list of songs or artists for a listener, to enhance the listener's

value. This study selects a list of listeners for an artist in a context by considering the two sides' value. For each candidate artist-listener pair, we estimate the two sides' value for listening to an artist according to Equations 2 and 3. Next, for each artist promotion, Equation 1, the two-sided value is used to select the *N* candidate listeners that maximize the total value of the top-*N* recommended listeners. It uses a value-based ranking to realize the recommendation based on Tan et al. (2011). For each artist *j*, the goal is to find a set of listeners that will maximize the total value by considering both sides.

4.4. Research Setting and Data

This research used the dataset described earlier in this thesis. It is a subset of Last.fm's user data containing 18,933 seed users. Because the effects of external information and geographic characteristics are added to the proposed model, 143 artists who had external information released in the U.S. were selected for the recommendation test. Users with no observable geolocation information were removed, which led to 15,607 seed users being retained for further study. Related to these users, there were 1,796,932 appropriate listening records. The listening matrix of artist-listener observations was sparse, with an average density of only 3.22%. We obtained the listening records for the three months prior to the listening observations before the external information was released. The observations were used for estimating the effects of collaboration and listening characteristics for the covariates (see Table 4.2). We also used one month of listening records following the release of information to train and test the model.

The focus is on recommending an artist to new listeners, who had not listened to the artist in the previous observation period, but were more likely to listen to the artist after external information was released. Descriptive statistics for new listeners,

and the listening times for each listener and each artist are shown in Table 4.3. There is diversity in attracting new listeners by different artists when external information is released.

Table 4.3. Statistics for New Listeners and Plays during Jan. – Nov. 2013

	MIN	MAX	MEAN	S.D.
#NewListeners	6	681	163	152
#NewPlays/Listener	1	4,623	14	69
#NewPlays	67	22,943	2,284	3,631

Note. Obs.: 143 artists. *#NewPlays*: total listening time of all new users who listened to an artist within 1 month after external information was released. *#NewPlays/Listener*: number of times a new user listened to an artist within month after external information was released.

In this dataset, on average, an artist attracted 163 new listeners in the one-month period after new external information was released. The average number of times that a new listener listened to an artist was about 14. The diversity observed, based on a standard deviation greater than or equal to the mean of the distribution, indicates the effects of different types of external information. Thus, for music promotion for each artist, it is necessary to effectively identify the targeted 163 new listeners on average from the candidate pool of around 15,000 listeners.

4.5. Experiments and Results

We next investigated the use of the proposed approach for finding listeners to an artist when external information was released. For each type of external information, including *Music* and *Non-Music Content Info*, the corresponding artists were randomly segmented into 5 folds. A 5-fold cross validation (5-fold CV) was ran to obtain the recommendation results. In each training and testing dataset, the sizes of users' listening records on the corresponding artists are shown in Table 4.4.

In the training step, for each type of external information, we used users' listening count on the artist's music to learn a negative binomial model (by considering all the covariates listed in Table 4.2). In the testing step, the learned model was further used to estimate the listening quantity q_{ij} and probability $Pr(q_{ij})$ for the

test listener *i* of artist *j*. Finally, the recommended listener list for each artist in the test dataset was proposed based on the two-sided value: listeners' utility and provider's revenues.

Table 4.4. Training and Test Dataset Size for 5-Fold CV

	Music Content Info		Non-Music Content Info	
	Train	Test	Train	Test
Fold 1	1,038,541	281,233	649,423	164,228
Fold 2	1,042,915	276,859	647,204	166,447
Fold 3	1,043,929	275,845	653,466	160,185
Fold 4	1,036,439	280,555	648,928	164,723
Fold 5	1,035,860	282,134	655,583	158,068

Note. Data: 88 artists with *Music Content Info* were included in each round; 70 of them are for training, and the other 18 for testing. 55 artists were included with *Non-Music Content Info*, in each round, with 44 of them for training, and the other 11 for testing.

4.5.1. Evaluation Measures

#NewListeners and *#NewPlays* are used as dependent variables for this evaluation. Three evaluative measures are used to test the performance of the modeling perspectives. *Conversion%* (*C%*) measures the percentage of correct recommendations in the top-*N* listener list. *Recall%* (*R%*) measures the percentage of how many new listeners were found to be in top-*N* listener list. And *Value%* (*V%*) is the percentage of the total listening for the value-maximizing recommendations in top-*N* listener list. Different *N* values, 100, 1,000, 2,000, ……, 7,000, were selected to observe the performance. The evaluative measures are:

$$C\%\ @\ N = \frac{\#NewListeners\ in\ TopN}{N}$$

$$R\%\ @\ N = \frac{\#NewListeners\ in\ TopN}{Total\ \#NewListeners}$$

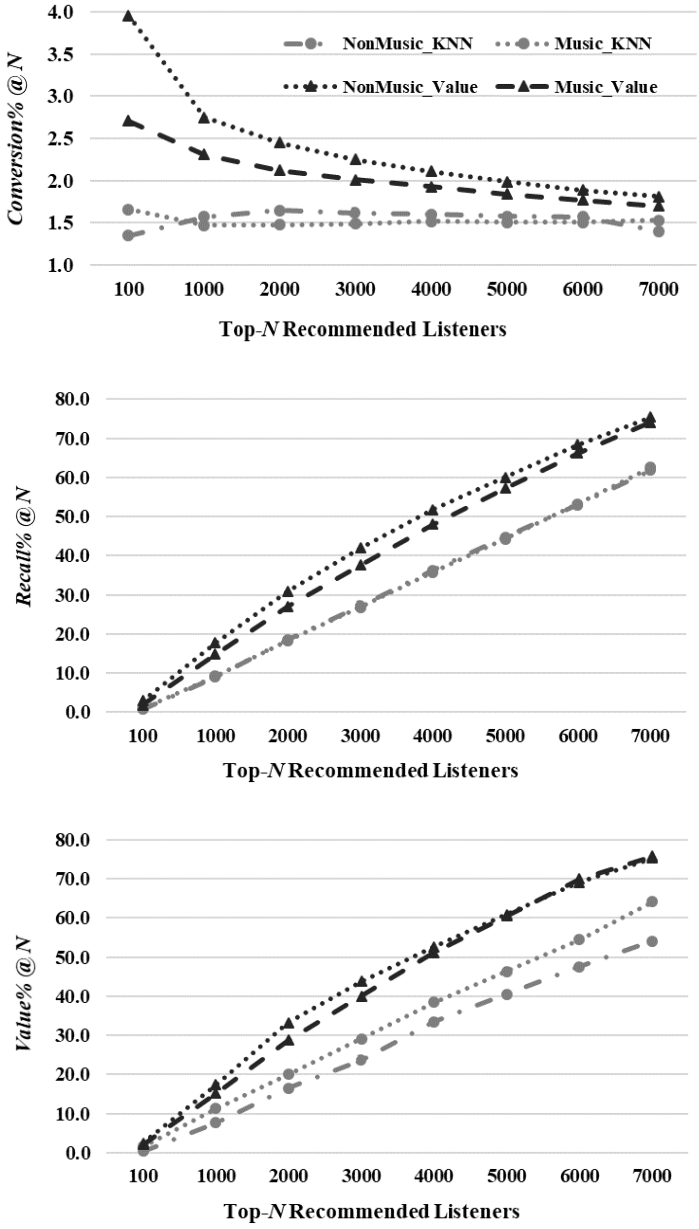
$$V\%\ @\ N = \frac{\#NewPlays\ in\ TopN}{Total\ \#NewPlays}$$

4.5.2. Performance Comparison

A traditional collaboration-based recommendation method was selected as the

baseline for this analysis work and comparisons. Listener-based *k-nearest neighbors* (KNN) was implemented, and potential listeners were ranked based on weighted utility a_{ij} value. We used k =15 as the number of neighbors parameter. This achieved stable performance, compared to k = 5, 10, and 20, as shown in Appendix Figure C2. The evaluation results for the measures are shown in Figure 4.5.

Figure 4.5. Evaluation Results of KNN and Value-based Methods for *Music* and *Non-Music Content Information* Contextual Recommendation



No matter which type of external information is considered, the proposed value-based method performs better than the baseline method. For *C%*, the value-based

method is around twice or even triple the value of KNN. Although both methods found a similar number of potential new listeners when N = 7,000, the value-based method did so faster, while KNN was still in process. For example, for the Top-1,000, the *C%* for the value-based method in the *Music Content* context was 2.34%, while for KNN it was 1.47%. This means that, on average, the value-based method was able to find around 23 new listeners in the Top-1,000 recommendation, while KNN was only able to find 15 at most.

For *R%* (the middle figure in Figure 4.5), the value-based method was better than the baseline method. It retrieved nearly 75% of all listeners in the Top-7,000 recommendations, while KNN only obtained around 60%. Similar conclusions pertain to *V%*.

These findings are useful for the personalized music promotion and recommendation design for a specific artist, especially for indie musicians with smaller listener bases, and less money to invest for music promotion. In this situation, this model can find the most potential listeners for them, satisfy the listeners' taste, and also maximize the possible pay-per-stream revenue value the artist can gain.

Figure 4.5 shows the recommendation performance of the value-based model for two categories of external information. As each category of external information has subtypes, so they may exhibit diverse performance levels. Next boxplots for each method show the maximum, minimum, median, and standard error (Figure 4.6 for *Music Content Info*, Figure C3 for *Non-Music Content Info*).

The figures on the top line are for KNN, and bottom line figures are for the value-based method. For *Music Content Info*, there is obvious diversity in *C%* and *V%*. *R%* has less diversity though. This means different subtypes or artists may have different levels of performance when they use the same recommendation method.

Non-Music Content Info is similar. It is worthwhile to check the recommendation performance of each subtype. A sample of the performance for each subtype is shown in Tables 4.5 and 4.6, and the full tables are in Appendix C.

Figure 4.6. Boxplot of Performances of KNN and Value-based Methods (*Music Content Information*)

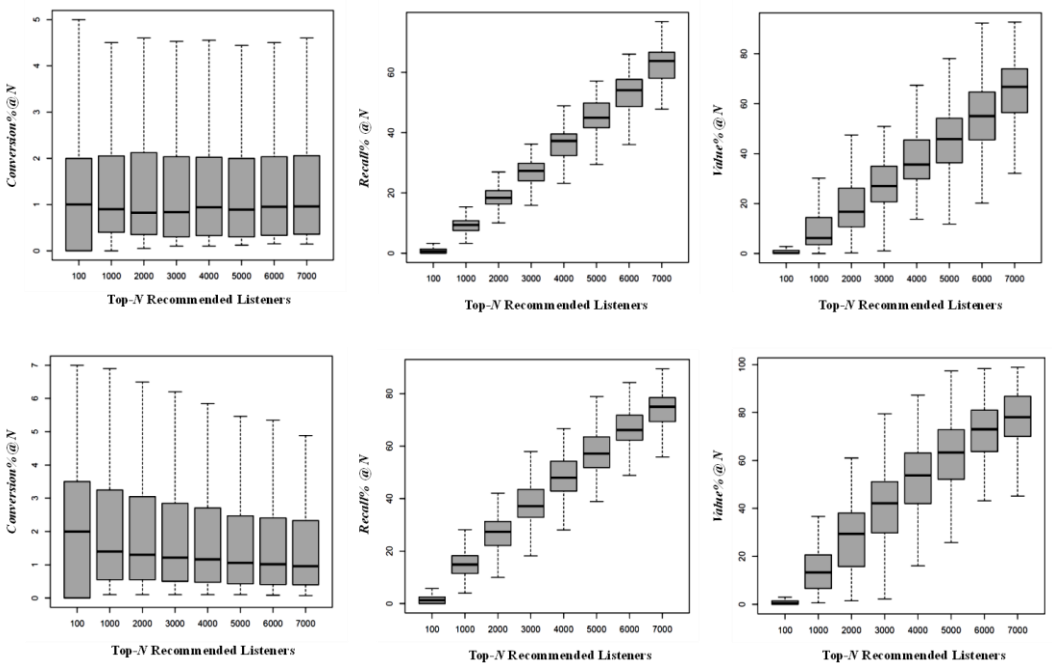


Table 4.5. Performance of Subtype of *Music Content Information*

KNN	*N* = 1,000 MEAN (S.D.)			N = 4,000 MEAN (S.D.)		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.47 (1.55)	9.30 (4.29)	11.27 (12.74)	1.52 (1.55)	35.75 (6.69)	38.41 (15.47)
Type6-Single-Song Release	1.47 (1.16)	9.52 (3.52)	10.51 (11.69)	1.55 (1.25)	38.04 (5.96)	38.13 (15.90)
Type7-Album Release	1.40 (1.65)	9.30 (4.77)	11.60 (13.70)	1.43 (1.60)	34.59 (6.86)	38.12 (15.88)
Type8-Music-Video Release	2.20 (2.55)	8.02 (3.29)	12.04 (8.93)	2.23 (2.55)	34.56 (6.75)	43.12 (8.50)
VALUE-BASED	***N* = 1,000 MEAN (S.D.)**			***N* = 4,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	2.34 (2.42)	14.88 (5.15)	15.20 (11.38)	1.93 (1.96)	47.99 (9.20)	51.21 (18.62)
Type6-Single-Song Release	2.41 (2.01)	15.61 (5.16)	15.72 (11.08)	2.04 (1.73)	49.45 (7.90)	50.77 (16.71)
Type7-Album Release	2.20 (2.50)	14.48 (5.00)	14.83 (11.62)	1.79 (1.99)	47.30 (9.65)	50.75 (20.34)
Type8-Music-Video Release	3.48 (3.79)	14.95 (7.19)	16.25 (12.79)	2.73 (2.99)	47.03 (12.19)	58.77 (5.27)

Note. Obs.: 88 artists with *Music-Content External Info. Type6:* 29; *Type7:* 54; *Type8:* 5. *C%: Conversion%, R%: Recall%; V%*: *Value%.*

For the subtype of *Music Content Information*, we can see that, regardless of either method, *Type8-Music-Video Release* gained the best performance but had large standard deviation. The *C%* was higher because this kind of external information typically has the largest impact on attracting new listeners (see Chapter 3). The possible reason for large standard deviation is the small number of observations – only 5 artists – in this subtype. *Type7-Album Release* had the worst performance but also the smallest standard deviation. There were more observations in this subtype, with 54 artists, and this type is not that effective as *Type-8* in attracting new listeners, therefore *C%* was smaller.

Table 4.6. Performance of Subtype of *Non-Music Content Information*

KNN	***N* = 1,000 MEAN (S.D.)**			***N* = 4,000 MEAN (S.D.)**		
Non-Music Content Info	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.57 (1.47)	8.97 (3.20)	7.67 (6.11)	1.60 (1.44)	36.24 (6.66)	33.48 (17.06)
Type1-News, Artist Life	1.41 (1.36)	9.18 (3.47)	8.18 (6.41)	1.56 (1.54)	36.00 (8.84)	38.01 (18.98)
Type2-News, Music-Related Info	1.64 (1.52)	9.08 (2.34)	8.15 (8.95)	1.73 (1.67)	38.22 (2.70)	33.92 (25.60)
Type3-Tour, Concert	0.42 (0.33)	7.55 (5.12)	8.29 (7.96)	0.55 (0.46)	33.99 (5.89)	29.69 (14.36)
Type4-Live TV Show	2.43 (1.50)	10.43 (1.66)	8.33 (3.26)	2.12 (1.33)	36.94 (5.81)	29.07 (6.92)
Type5-Live Performance / Festival	1.66 (1.73)	7.84 (2.81)	5.15 (4.00)	1.67 (1.46)	35.81 (5.15)	30.30 (13.49)
VALUE-BASED	***N* = 1,000 MEAN (S.D.)**			***N* = 4,000 MEAN (S.D.)**		
Non-Music Content Info	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	2.75 (2.21)	17.75 (5.78)	17.37 (14.96)	2.11 (1.78)	51.79 (8.02)	52.60 (20.08)
Type1-News, Artist Life	2.80 (2.47)	18.19 (6.48)	20.95 (17.06)	2.09 (1.92)	52.29 (6.97)	53.02 (21.11)
Type2-News, Music-Related Info	2.49 (2.20)	14.92 (2.98)	6.44 (3.87)	2.21 (1.98)	52.14 (7.61)	51.62 (22.09)
Type3-Tour, Concert	1.18 (0.73)	20.02 (3.43)	19.45 (11.57)	0.88 (0.62)	56.13 (6.34)	55.25 (20.31)
Type4-Live TV Show	3.84 (2.14)	18.20 (6.18)	20.27 (17.64)	2.90 (1.73)	51.97 (5.04)	55.92 (16.00)
Type5-Live Performance / Festival	2.72 (1.99)	17.29 (6.67)	14.43 (11.85)	2.03 (1.70)	7.69 (12.38)	47.61 (22.76)

Note. Obs.: 55 artists with *Non-Music Content External Info*. *Type1*: 21; *Type2*: 8; *Type3*: 6; *Type4*: 10; *Type5*: 10. *C%: Conversion%, R%: Recall%; V*%: *Value%.*

In this case, although the artists' characteristics (e.g., popularity, major label) were considered when training the model, there were not enough observations to

estimate a balanced model for every subtype and artist. This situation can be addressed if we model each subtype separately when a larger dataset is available. A similar conclusion can be reached for *Non-Music Content Info* too (see Table 4.6).

4.6. Discussion and Conclusion

Streaming services have reshaped people's music listening behavior and also the profit patterns of the music industry and the platforms that offer digital music. How to promote a song, album or artist via streaming music services also has become a hot topic for research and for the music industry. Although music recommendation methods have been explored since the 1990s, there are still many unsolved problems that need to be addressed.

Most of existing music recommender systems that have been adopted by streaming music services, such as Last.fm and Spotify, are focusing on the user side. Regardless of which recommender method is applied – content-based, collaboration-based or hybrid, their goals are to find the most value-enhancing music for a user via *personalized music recommendation*. Table 4.1 summarizes these methods from multiple perspectives, none of them have tried to design *personalized music promotions* for a specific artist. This is very useful, as streaming music has generated more and more market revenue, especially for independent musicians and song writers. Their music is usually not that easy to find among the available music choices, even when there are recommendation services available. So how to design a personalized music promotion for an artist has business and scientific value.

In this study, when an artist's external information is released, we showed that a related value-based music promotion method can be used to assist the artist to target potential listeners. It considers the listener's utility, with the artist's revenue from customer listening, with pay-per-stream as a basis for measuring value. The

proposed method demonstrated an improvement of the conversion rate for listeners which recommendations compared to the traditional method. The findings provide new design thinking for music recommendations and personalized artist promotions in online music platforms. Here is what was learned from this research.

First, considering external information can improve music recommendation accuracy, which validates the proposal in the empirical research of Chapter 3. The increase in *C%*, *R%* and *V%* confirm that the proposed value-based method can be used to assist artists to find potential listeners when they have promotion strategies, no matter whether they involve a new album release, or news about an upcoming concert. This method is different from existing approaches: it can identify new listeners who have never listened to the artist, and also can re-activate intermittent listeners who listened before but stopped for a while. Finer-grained modeling for different types of external information or a specific music genre can improve performance further. This paves the way for *on-demand music promotion*, especially for independent musicians.

Second, our method uses econometric analysis in real applications. Econometric modeling and estimation have uncovered many useful insights by considering many possible factors to analyze an event or a phenomenon. But industry applications usually focus on a single perspective, due to the complexity, data and modeling costs, and difficulty of achieving causal explanations, among other reasons. It also demonstrates the combination of statistics and econometrics with traditional music recommendation methods from CS. Although this research applies this kind of fusion analytics modeling, it can be further extended for other state-of-the-art recommendation methods, such as matrix factorization.

Third, another research stream in music information diffusion and recommendation is influential user identification. The goal is to figure out the "big" users who can influence others to adopt a product or information in the presence of some social structure. We also tried to detect the influential listeners in Last.fm by considering social relations (Ren et al. 2014). However, this is still an indirect way. If we can directly identify the "right" listener when there is a promotion strategy, the combination of the direct and indirect approach will be more effective. In this study, we worked toward a combination of direct and indirect promotion by considering the listener's utility and the number of her social friends. The latter is a proxy to measure the potential revenue the targeted listener may bring over time.

This research attempted to design music promotions for a specific artist at a specific time. There are several limitations though. First, in the proposed two-sided value-based model, we did not consider the costs of the two sides. Consumer utility was measured using the listening quantity, but we ignored the search cost for finding the music. It is hard to calculate the search cost for each listener because there are too many ways they can access music-related information before deciding to listen to it, so it is impossible to track the process.

For the provider's value, we borrowed the idea of "value in the presence of information less value in the absence of information" from the *value of information theory*. We think of this in terms of the availability of recommendation information, or the lack of it, to calculate the potential value that the providers can gain from the recommendation. This ignores the investment they need to make for music production and promotion. How to measure and combine the two-sided costs are worthwhile for future study.

Second, the model is in its infancy in developmental terms. The estimated expectation of listening quantity and probability was used to do the recommendation, but this is not that rigorous. In music listening, the expectation may be insufficient to describe how much a potential listener really listens. So the listening amount with the largest probability may be better to explain audience listening and estimate the its value, as in Equation 8. Thus, how to improve the estimation of the future listening quantity is on the list for future research.

Third, the dataset is limited, so memory-based music recommendation was selected as the base model on which to make improvements. This may not be workable for a very large dataset though, so how to make it into a scalable algorithm that can be implemented smoothly in industry environments still needs to be further explored. Last, historical data were used to train and test the proposed model but we did not do a randomized experiment or a user study to test the approach with unique new data.

Based on these limitations, our future work will emphasize several things. First, we will collect more data and work on the scalability of the proposed method. For example, transferring the base model from memory-based KNN to model-based matrix factorization (for estimation of a_{ij} value) can help to embed listener and artist value-aware attributes into the model-based recommendation algorithm. Second, we will seek to improve the estimation and representation of listening quantity by enhancing the econometric analysis, with finer-grained subtypes for external information and artists of various genres separately. Also, we will further explore the two-sided costs: search cost for the listeners and production and promotion costs for the providers. Last, we will expand the current research with a user study to test the proposed model using new data.

Chapter 5. Fusion Analytics Research Practice

Today, the availability of big data and fast developing technologies have resulted in new ways of doing data analytics. This has opened up the possibility for innovative thinking and novel contributions to scientific discovery in interdisciplinary contexts. The approach is *fusion analytics*, which brings different bodies of knowledge and research approaches to bear in order to reveal interesting insights and useful applications (Kauffman et al. 2017).

My research focuses on the interdisciplinary area of streaming music. My research journey started from studying streaming music data and provided a stepping stone toward the bigger picture of streaming music research. Reviewing prior music research in multiple areas and understanding streaming music services and the music industry further provided me with various theoretical perspectives and methodological approaches from different disciplines for addressing music-related issues. My increasing access to publicly-available big data, coupled with my training in data analytics, multimedia analysis, econometrics, machine learning and statistics, has opened up considerable opportunities for me to analyze, understand, design and improve streaming music services via fusion analytics research.

Research framework construction. My research began and developed by exploring streaming music in real world. Essay 1 was inspired by my curiosity about the sudden success of the music track "Rolling in the Deep," which paved the way for Adele to become a new superstar. My knowledge about music content analysis helped me to understand the nature of music, however, with the streaming music services and the openness of the Internet, there are many other potential factors that may affect music popularity.

My advisor, Prof. Robert J. Kauffman, introduced me to the Computational Social Science area, and I learned that combining econometric analysis, machine learning and prediction can assist in understanding music popularity development. This became Essay 1 and was the beginning of my fusion analytics journey. Essay 2 focused on a topic from the first study, related to the development of music popularity over time. The inclusion of external information and geographic effects was the result of discussions with Profs. Kauffman and Qiuhong Wang. Essay 3, on music promotion design with recommendations, was a natural development after I figured out the insights I was able to obtain about music popularity and diffusion. My overall research inquiry reflects my efforts with the full research cycle of observing, exploring, modifying and then repeating this procedure.

Data collection, extraction, cleansing and integration. Big data are defined in terms of four V's: *volume, velocity*, *variety*, and *veracity* (IBM 2017). The advantage of technological capabilities can be used to address these dimensions, to assist with research that uses public data. Before answering the research questions, I needed to collect large-scale data from different sites via a web crawling approach. The data were huge, raw, and noisy, in heterogeneous formats, and had diverse characteristics. I used machine learning, text mining, and statistics methods to extract information from them, for example, transferring music acoustic and lyric content into semantic descriptors in Essay 1.

The extracted data were further cleaned and integrated into panel data according to my specific research targets. For example, I used PSM to construct and analyze data to ensure that they were appropriately matched for inclusion in treatment and control groups for the diffusion analysis. At this point, the data became mature, and were usable in econometric analysis and estimation. Data collection, extraction,

cleansing and integration usually occupies over 80% effort of public data-based research. My Ph.D. research has helped me to become familiar with every step in the whole process. I am still learning how to leverage the advantages of multiple technologies to assist this key procedure.

Combination of econometrics and machine-based analytics in exploring business insights. I used econometric modeling and estimation to uncover the key relationships, influences, and marginal effects of relevant variables based on carefully structured and cleaned data. They generally have been undertaken separately with CS methods involving machine-based data analytics. In contrast, I mixed methods from these two fields in my fusion analytics. I used machine-based methods to help structure and clean the data, and then constructed meaningful econometrics models, such as a duration model, a geolocation-based DiD model, and a music listening-related count data model.

Through these analysis approaches, I gained useful insights and implications for business practice. I paid attention to interpreting the analysis results by considering the industry setting that the data came from. Traditional econometrics usually does not go as far as I did. My approach used the insights from the model estimation to improve business prediction and service capabilities, such as music popularity estimation and new music promotion algorithm design.

Effort to bridge academic research and industry relevance. Academic research always attempts to stay one step, or even multiple steps ahead of industry applications. How to quickly transfer academic findings into applicable methods and systems is still a complex problem though. Although I did not collaborate with an industry partner to create randomized experiments or to do recommender system

implementation, I nevertheless considered consumers, the music industry and platform providers as key stakeholders. For example, I proposed two-sided value-based recommendations to assist music labels, independent musicians, and song writers, and streaming music services to improve the music listener's experience and utility level based on consumer listening support.

This can be further expanded to on-demand recommendation for specific artists. This kind of targeted music recommendation is a new perspective compared to other more general recommendation implementations for streaming music services and other e-commerce websites. I still have a lot of work to do, including but not limited to making the algorithm scalable and workable in industry settings.

Scientific writing and presentation. Publishing research articles in conferences and journals in the future will allow me to communicate the research and findings in my work to my peers. While preparing my manuscripts for submission, I have learned so much from my advisors: how to write scientific papers, how to conceptualize and write effective peer review responses, how to modify my papers based on the review comments I receive, and so forth.

I was fortunate to have opportunities to present the preliminary results of my essays in three conferences and a Doctoral Consortium in both the CS and IS fields. Essay 1 was presented at the 2016 International Conference on the World Wide Web in Montreal, Canada, and the 2017 European Conference on Information Systems (ECIS) in Guimaraes, Portugal. Preliminary results of Essay 2 were presented at the 2014 International Conference on Internet Multimedia Computing and Services in Xiamen, China. And my overall thesis ideas were presented at the 2017 ECIS Doctoral Consortium. The feedback and comments I received were from different research perspectives, and invaluable to my research development. Attending

the other researchers' presentation sessions during the conferences also brought me new ideas and knowledge about the methods used in other related fields.

Next, I would like to share some personal thinking about the challenges in fusion analytics that Ph.D. students in the Data Mining and IS and Management (IS&M) arears may encounter.

Challenges for a Data Mining Ph.D. Student to Acquire Skills for IS&M Research. First, data mining Ph.D. students focus on accuracy for prediction, classification or other purposes, but their approaches only involve associational research designs. The research approach in IS&M is much different. It focuses on research designs for causation, to figure out the cause-and-effect relationships that are present for a data set and its research setting.

A data mining Ph.D. student needs to carefully extract, clean, select and construct a dataset for a specific research target, and the same is true for IS&M research. For example, in data mining, we usually do not consider whether a dataset is subject to censored data outside the period of observation, while it is an important issue in IS&M.

A second challenge is how to select an appropriate explanatory econometric model and how to interpret the regression results properly. It takes time to get familiar with the large variety of models that are available in econometrics. Data mining Ph.D. students learn about and obtain relevant methods knowledge in two ways.

One way is by reading authoritative books about econometrics, such as *Introductory Econometrics* (Wooldridge 2015). This way supports obtaining rigorous econometrics knowledge step by step, while analyzing data related to a research project. Another way is by reading related papers in top-tier journals and conferences, such as ISR, JMIS, MISQ, and conferences, including ICIS, PACIS, ECIS,

and others. The published works offer a quick way to learn about model usage and interpretation for different data and different research target. These two ways helped me in my Ph.D. study. Of course, I am still learning.

A third challenge is how to use insights about IS&M modelling to assist with algorithm design, which is usually the final target of data mining studies. The key point here is how to properly embed the significant variables into, or combine them with an existing machine-based model or algorithm to design an improved or a new one. For example, I tried to combine the negative binomial regression results with KNN recommendation to design a new two-sided value-based music recommendation algorithm. The research approach I have described can help me to do more in multiple research topics in the future, and is not limited to only music.

Challenges for an IS&M Ph.D. Student to Acquire Skills for Data Mining Research. The huge amount of public data and proprietary industry data in the complex environment of the Internet has led to data that are not well-structured, especially for multimedia analysis. Therefore, the challenges for an IS&M Ph.D. student to do research in this area are considerable. First, the collection of publicly-available data is not so easy. Some websites provide *application programming interfaces* (APIs) for sample datasets that can be used for academic studies. They include such sources as Last.fm and Spotify, as I used in my research, which are not difficult to access. Other websites do not supply APIs for data collection. Instead, Ph.D. students need to learn some simple website crawling methods using Java or Python coding. They may need to learn how to use HTTP methods like GET to search and retrieve data from HTML page sources.

The second challenge is how to transfer the data that are collected into a structured format. This challenge is similar to one that data mining students experience.

IS&M Ph.D. students need to be familiar with machine-based methods, at least insofar as knowing how to use basic methods to assist with data processing and storage. There are also existing tools for machine-based methods, such as Weka, LightSide and Stanford NLP, which are easy to learn for beginners.

If a study focuses on a specific kind of multimedia, such as music and movies, specific domain knowledge may need to be acquired. In addition, special tools for multimedia content analysis, such as MIRtoolbox for musical feature extraction, may be needed. Just as it takes time for a data mining student to acquire IS&M skills, it also takes time for an IS&M student to be able to learn to use machine-based methods well enough to be productive with them. Once they have acquired the skills, they will gain the power that those methods offer for processing huge and complex datasets.

Expanding my research network and framework. Business problems are complex and often require multi-disciplinary expertise. Through collaborating with my committee member, Dr. David R. King, and also attending conferences and workshops in multiple areas, I connected with experts and scientists from industry and the academic disciplines, including CS, IS, and Social Science. The research networks I developed will pave the way for future research collaborations.

The research experience during my Ph.D. program has given me a better understanding of fusion analytics research for social media-related issues, and it is not limited to music, but also includes cable TV programs, sports, and mobile apps. With the help of advances in data analytics, and novel methodologies in text mining, machine learning, and econometrics, I aspire to contribute to knowledge on social media, customer behavior and system design by undertaking research as a multidisciplinary scientist.

Chapter 6. Conclusion

This dissertation was motivated by the development of streaming music which is prevalent among audiences in the world today. Through *fusion analytics*, I looked into a streaming music service, Last.fm, from a new perspective by considering the music providers, content, and consumers as an ecosystem of interacting entities and stakeholders. I investigated three issues related to: the music business value, music diffusion, and how to promote products in streaming music settings. Through mining audience listening records over a 10-year period, I assessed the effects and predictive ability of musical and non-musical factors on music popularity development at the early stage after release. I also documented the interplay of external sources of information with streaming music diffusion. Furthermore, I designed a two-sided value-based music assessment and promotion approach that leverages my other research insights on what drives music popularity and diffusion.

Chapter 2 emphasized the power of data analytics for knowledge discovery and explanation that can be achieved with a combination of machine- and econometrics-based approaches. It contributes to the literature on music track popularity in the social network scenario in several ways. I constructed a relatively complete descriptive vector for music tracks. It supplements existing descriptors with more fine-grained music semantics to provide fuller information about the drivers of music track popularity. Also, my focus on the lifecycle of a music track's development provides a more complete description of what is happening during the process. Finally, the findings on music popularity prediction provide useful insights for the music labels, which can assist them to assess the potential popularity of a new track in its early stage after release and adjust their promotion strategies in the music market.

Chapter 3 demonstrated the impacts of the interactions among multiple media channels on streaming music diffusion in today's free-access platform environment. It complements studies on music diffusion by adding insights on how the external sources of information affect streaming music diffusion over time. The results provide strategic implications for the music labels on what kinds of artist-related news and information can be leveraged to promote streaming music diffusion by considering their impacts. The results also have implications for the music labels related to cultivating new consumers in different geographic regions according to the information release location.

Chapter 4 studied how to promote a music artist at a specific time by considering the utilities of both the music provider and consumer. I combined utility theory and value of information theory to create a music recommendation method, which allowed me to contribute knowledge about how to balance value for the two stakeholders. This is an interesting question for researchers as well as an important problem for the music labels, especially since the increases in streaming music services are continuing. Their effects have been to reduce revenue from half of the total music market, but they can also be used to promote music products via streaming services, especially for niche music and independent musicians. This study also has contributed a new research direction for online streaming music recommendation.

My dissertation lies in the interdisciplinary area of information, technology, business sustainability, and value in the music industry. These things involve data analytics, econometrics, data mining, social media analysis, and recommender system design. I studied how to use fusion analytics in an industry application. In Essay 1, I leveraged machine-based methods to collect and consolidate data from multiple sources with secondary and interview data to study the business value of music. I

also used PSM to resolve empirical estimation issues with respect to artists and user characteristics in Essay 2. The insights that were created by my data analysis work in the first two essays were further leveraged to design a new music recommendation approach in Essay 3. These three essays started from empirical analysis and ended with prediction applications, which support a positive feedback loop for understanding streaming music.

There are several limitations though. One limitation is that my dissertation is not a longitudinal study, which means that I cannot trace the music development process over time. For example, the Essay 1 described the social effects on music track potential popularity by only considering one social network. 24 x 7 free access to the Internet provides multiple sources of social context that go beyond the boundaries of a given music social network. But it is impossible to acquire data on all of the social developments in the past 10 years because we cannot trace back the data that describe the process.

For Essay 2, it was also hard to list all the different kinds of external information that is related to artists as time passes, although the Internet retains a massive amount of artist-related information in different locations. For Essay 3, I used historical data to train and test the proposed method, but did not run a randomized experiment or a user study to test the recommendation method used in Last.fm.

A second limitation is that I did not consider the culture effects that are related to streaming music listening. I considered the geolocation impacts on music diffusion though. A more fine-grained analysis of the country characteristics of consumers would help to address this problem, by considering language differences, and cultural and physical distance. For example, U.S.-based artists may find it easier to

attract U.K. listeners compared to those in China, due to language and cultural similarities.

Finally, in the three essays, I did not distinguish those who listened for free from other on-demand listeners (e.g., subscription-based and premium consumers) because of the lack of data. This limitation does not affect the results of Essay 1 because I focused on general listening for all of the users. It may affect Essay 2, however, because of user selection bias. I tried to trace the music diffusion for all users to control this bias. For Essay 3, free listeners and on-demand listeners may have different music access authorities that are likely to affect how they can benefit from music recommendations. Thus, the recommendation design in the marketplace may need adjustments to be made for the recommendations to be meaningful for consumers.

My future work will expand my research on music, so it has more longitudinal empirical scope. I will trace the development of music tracks in the marketplace, and also implement a randomized experiment to further test the proposed recommendation method in Essay 3 and explore whether music promotion can work in real time applications. To combine the power of econometrics and machine-based methods, I also plan to implement other methodologies including natural language processing and deep learning, to work with the massive and diverse business data available in the music domain so I can carry out theory-driven causal analysis.

My dissertation offers an example of fusion analytics, but I will explore the application of this blended methods approach to other digital products, such as TV programs and movies. Moreover, I plan to study other opportunities related to some of the high-level topics in my work: business intelligence, customer behavior anal-

ysis, and recommender systems. My hope is to investigate industry business problems for which insights can be obtained through extensive use of IT. I also seek to produce useful insights on customer behavior and managerial implications for e-commerce, social media, the financial services industry, and others.

References

Adomavicius, G., Bockstedt, J., Curley, S., Zhang, J., 2017. Effects of online recommendations on consumers' willingness to pay. Information Systems Research, in press.

Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734-749.

Adomavicius, G., Tuzhilin, A., 2015. Context-aware recommender systems. In Recommender Systems Handbook. Springer, Boston, MA. 191-226.

Amatriain, X., Basilico, J., 2012. Netflix recommendations: Beyond the 5 stars (part 1). Netflix Tech Blog, 6.

Aral, S., Muchnik, L., Sundararajan, A., 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. In Proceedings of the National Academy of Sciences 106(51), 21544-21549.

Bakshy, E., Rosenn, I., Marlow, C., Adamic, L., 2012. The role of social networks in information diffusion. In Proceedings of the 21st International Conference on World Wide Web, ACM Press, New York, 519–528.

Bapna, R., Umyarov, A., 2015. Do your online friends make you pay? A randomized field experiment on peer influence in online social networks. Management Science 61(8), 1902-1920.

Beaujean, A.A., Morgan, G.B., 2016. Tutorial on using regression models with count outcomes using R. Practical Assessment, Research & Evaluation 21(2), 1-18.

Belluf, T., Xavier, L., Giglio, R., 2012. Case study on the business value impact of personalized recommendations on a large online retailer. In Proceedings of the 6th ACM Conference on Recommender systems. ACM Press, New York, 277-280.

Bhattacharjee, S., Gopal, R.D., Lertwachara, K., Marsden, J.R., Telang, R., 2007. The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. Management Science 53(9), 1359-1374.

Bischoff, K., Firan, C., Georgescu, M., Nejdl, W., Paiu, R., 2009. Social knowledge-driven music hit prediction. In International Conference on Advanced Data Mining and Applications, Springer, Berlin / Heidelberg, Germany, 43-54.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3(1), 993-1022.

Borg, N., Hokkanen, G., 2011. What makes for a hit pop song? What makes for a pop song? Stanford University, Stanford, CA.

Breiman, L., 1994. Bagging predictors. Technical report 421, Department of Statistics, University of California, Berkeley, CA.

Bronson, F., 2012. Hot 100 55th anniversary: The all-time top 100 songs. Billboard, August 2.

Brynjolfsson, E., Hu, Y., Smith, M.D., 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. Management Science 49(11), 1580-1596.

Bulow, J.I., 1982. Durable-goods monopolists. The Journal of Political Economy 90 (2), 314-332.

Cameron, A.C., Trivedi, P.K., 2013. Regression Analysis of Count Data, 2nd edition. Econometric Society, Cambridge University Press, London, UK.

Cano, P., Koppenberger, M., Wack, N., 2005. Content-based music audio recommendation. In Proceedings of the 13th Annual ACM International Conference on Multimedia. ACM Press, New York, 211-212.

Cantrell, M.A., Conte, T.M., 2009. Between being cured and being healed: the paradox of childhood cancer survivorship. Qualitative Health Research 19(3), 312-322.

Cha, M., Mislove, A., Gummadi, K.P., 2009. A measurement-driven analysis of information propagation in the Flickr social network. In Proceedings of 18th International Conference on World Wide Web, ACM Press, New York, 721–730.

Chang, R.M., Kauffman, R.J., Kwon, Y.O., 2014. Understanding the paradigm shift to Computational Social Science in the presence of big data. Decision Support Systems 63, 67-80.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321-357.

Chen, H., Chiang, R.H., Storey, V.C., 2012. Business intelligence and analytics: from big data to big impact. MIS Quarterly 36(4), 1165-1188.

Cheng, Z., Shen, J., 2014. Just-for-me: An adaptive personalization system for location-aware social music recommendation. In Proceedings of International Conference on Multimedia Retrieval. ACM Press, New York, 185.

Cheng, Z., Shen, J., 2016. On effective location-aware music recommendation. ACM Transactions on Information Systems 34(2), 13.

Chon, S.H., Slaney, M., Berger, J., 2006. Predicting success from music sales data: A statistical and adaptive approach. In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, ACM Press, New York, 83-88.

Chung, K.H., Cox, R.A., 1994. A stochastic model of superstardom: An application of the Yule distribution. The Review of Economics and Statistics 76(4), 771-775.

Coleman, J., Katz, E., Menzel, H., 1957. The diffusion of an innovation among physicians. Sociometry 20(4), 253-270.

Culnan, M.J., McHugh, P.J., Zubillaga, J.I., 2010. How large US companies can use Twitter and other social media to gain business value. MIS Quarterly Executive 9(4), 243-259.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, ACM Press, New York, 233-240.

Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for non-experimental causal studies. Review of Economics and Statistics 84(1), 151-161.

Desrosiers, C., Karypis, G., 2011. A comprehensive survey of neighborhood-based recommendation methods. In Recommender Systems Handbook, Springer, Boston, MA, 107-144.

Dewan, S., Ho, Y., Ramaprasad, J., 2017. Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Music Community. Information Systems Research 28(1), 117–136.

Dhanaraj, R., Logan, B., 2005. Automatic prediction of hit songs. In Proceedings of the 6th International Society for Music Information Retrieval, ACM Press, New York, 488-491.

Ellison, N.B., Steinfield, C., Lampe, C., 2011. Connection strategies: Social capital implications of Facebook-enabled communication practices. New Media & Society 13(6), 873-892.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems. Journal of Machine Learning Research 15(1), 3133-3181.

Flacy, M., 2012. Unlimited listening on Spotify will vanish for U.S. early adopters next week. DigitalTrends.com, Portland, OR, January 6.

Forbes, 2017. Beyoncé: Musician, Adele: Musician.

Frieler, K., Jakubowski, K., Müllensiefen, D., 2015. Is it the song and not the singer? Hit song prediction using structural features of melodies. Yearbook German Society for Music Psychology, Hogrefe Verlag, Göttingen, Germany, 24, 41-54.

Garg, R., Smith, M.D., Telang, R., 2011. Measuring information diffusion in an online community. Journal of Management Information Systems 28(2), 11-38.

Gomez-Uribe, C.A., Hunt, N., 2016. The Netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems 6(4), 13.

Gonzalez, M. C., Pastore, C.A., Orlandi, S.P., Heymsfield, S.B., 2014. Obesity paradox in cancer: new insights provided by body composition. The American Journal of Clinical Nutrition 99(5), 999-1005.

Hamlen, W.A., 1991. Superstardom in popular music. The Review of Economics and Statistics 73(4), 729-733.

Herremans, D., Martens, D., Sörensen, K., 2014. Dance hit song prediction. Journal of New Music Research 43(3), 291-302.

Hiller, R.S., Walter, J., 2017. The rise of streaming music and implications for music production. Available at SSRN: https://ssrn.com/abstract=2670976 or http://dx.doi.org/10.2139/ssrn.2670976.

Hu, P., Liu, W., Jiang, W., Yang, Z., 2014. Latent topic model for audio retrieval. Pattern Recognition 47(3), 1138-1143.

Hussain, M., Khan, N., Uddin, M., 2014. Log-normal duration model is the best fitted model for duration from chest pain to coronary artery disease diagnosis:

and outcome of retrospective cross sectional study. Pakistan Journal of Statistics and Operation Research 10(4), 369-379.

IBM, 2017. The four V's of big data. Infographics and animations, Big Data & Analytics Hub, Armonk, NY.

IFPI (International Federation of the Phonographic Industry), 2012, 2013, 2015. Digital music report. Recording industry in numbers. London, UK.

IFPI (International Federation of the Phonographic Industry), 2017. Global music report. London, UK.

Imbens, G., Wooldridge, J., 2007. What's new in econometrics? Difference-in-differences estimation. Lecture Notes, NBER Summer Institute, Cambridge, MA.

James, G., Witten, D., Hastie, T., Tubshirani, R., 2013. An introduction to statistical learning: with applications to R. G. In Casella, S. Fienberg and I. Olkin (eds.) Springer, Berlin / Heidelberg, Germany.

Jannach, D., Adomavicius, G., 2017. Price and profit awareness in recommender systems. arXiv preprint arXiv:1707.08029.

Janani, S., Iyswarya, K., Priya, K., Visuwasam, L., 2012. Combining spectral features to identify the musical instruments and recognize the emotion from music. International Journal on Computer Science and Engineering 4, 1253-1259.

John, S., 2000. Rational choice theory. In G. Browning, A. Halckli, and F. Webster (eds.), Understanding Contemporary Society: Theories of the Present, Sage Publishing, Thousand Oaks, CA.

Karydis, I., Gkiokas, A., Katsouros, V., 2016. Musical track popularity mining dataset. In IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, Cham, Switzerland, 562-572.

Katz, C, 2014. Get on-demand music streaming on Last.fm with Spotify. Spotify News. https://news.spotify.com/us/2014/01/29/get-on-demand-music-streaming-on-last-fm-with-spotify/

Kauffman, R. J., Kim, K., Lee, S. Y. T., Hoang, A. P., Ren, J., 2017. Combining machine-based and econometrics methods for policy analytics insights. Electronic Commerce Research and Applications 25, 115-140.

Kim, Y., Schmidt, E., Migneco, R., Morton, P., Richardson, J., Speck, J., Turnbull, D., 2010. Music emotion recognition: a state of the art review. In Proceedings of the 11th International Society of Music Information Retrieval, Utrecht, Netherlands, 255-266.

Kim, Y., Suh, B., Lee, K., 2014. # now playing the future Billboard: Mining music listening behaviors of Twitter users for hit song prediction. In Proceedings of the 1st International Workshop on Social Media Retrieval and Analysis, ACM Press, New York, 51-56.

Kleinbaum, D.G., Klein, M., 2006. Survival Analysis: A Self-Learning Text. Springer Science & Business Media, New York.

Knees, P., Schedl, M., 2013. A survey of music similarity and recommendation from music context data. ACM Transactions on Multimedia Computing, Communications, and Applications 10(1), 2, 1-21.

Koenigstein, N., Shavitt, Y., 2009. Song ranking based on piracy in peer-to-peer networks. In Proceedings of the 10th International Society of Music Information Retrieval, Kobe, Japan, 633-638.

Koenigstein, N., Dror, G., Koren, Y., 2011. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. In Proceedings of the 5th ACM conference on Recommender systems. ACM Press, New York, 165-172.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. The Computer Journal, 42(8), 42-49.

Kumar, R., Novak, J., Tomkins, A., 2006. Structure and evolution of online social networks. In T. Eliassi-Rad (ed.), Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 611–617.

Last.fm, 2015. Terms of User. https://www.last.fm/legal/terms.

Lee, J., Lee, J. S., 2015. Predicting music popularity patterns based on musical complexity and early stage popularity. In Proceedings of the 3rd Edition Workshop on Speech, Language & Audio in Multimedia, ACM Press, New York, 3-6.

Lee, Y.J., Hosanagar, K., Tan, Y., 2015. Do I follow my friends or the crowd? Information cascades in online movie ratings. Management Science, 61(9), 2241-2258.

Li, B., Ghose, A., Ipeirotis, P.G., 2011. Towards a theory model for product search. In Proceedings of the 20th International Conference on World Wide Web, ACM Press, New York, 327-336.

Li, X., Wang, H., Gu, B. Ling, X., 2015. Data sparseness in linear SVM. In Proceedings of 24th International Conference of Artificial Intelligence, ACM Press, New York.

Li, Z., Kauffman, R.J., Dai, B., 2017. Can I see beyond what you can see? Blending machine learning and econometrics to discover household TV viewing preferences. In Proceedings of the 50th Hawaii International Conference on System Sciences, IEEE Computer Society Press, Washington, DC.

Luarn, P., Lin, H.H., 2003. A customer loyalty model for e-service context. Journal of Electronic Commerce Research 4(4), 156-167.

Mahajan, V., Muller, E., 1979. Innovation diffusion and new product growth models in marketing. Journal of Marketing 43(4), 55–68.

Mangalindan, J., 2012. Amazon's recommendation secret. CNN Money, July 30. http://tech.fortune.cnn.com/2012/07/30/amazon-5.

Mansfield, E., 1961. Technical change and the rate of imitation. Econometrica 29(4), 741-766.

Manski, C. F. 1993. Identification of endogenous social effects: The reflection problem. The Review of Economic Studies 60(3), 531–542.

MIDiA, 2017. Announcing MIDiA's streaming services market shares report. Music Industry Blog. https://musicindustryblog.wordpress.com/tag/streaming-market-share/

Morgan, N.A., Rego, L.L., 2006. The value of different customer satisfaction and loyalty metrics in predicting business performance. Marketing Science 25(5), 426-439.

Myers, S.A., Leskovec, J., 2014. The bursty dynamics of the twitter information network. In Proceedings of the 23rd International Conference on World Wide Web, ACM Press, New York, 913-924.

Myers, A., Zhu, C., Leskovec, J., 2012. Information diffusion and external influence in networks. In Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 33-41.

Napiorkowski, S., 2015. Music mood recognition: State of the art. RWTH Aachen University, Aachen, Germany.

Ni, Y., Santos-Rodriguez, R., McVicar, M., De Bie, T., 2011. Hit song science once again a science. In Proceedings of the 4th International Workshop on Machine Learning and Music, Sierra Nevada, Spain.

Nielsen, 2017. U.S. music mid-year report 2017. http://www.nielsen.com/us/en/insights/reports/2017/us-music-mid-year-report-2017.html.

Nunes, J. C., Ordanini, A., 2014. I like the way it sounds: The influence of instrumentation on a pop song's place in the charts. Musicae Scientiae 18(4), 392-409.

Nunes, J. C., Ordanini, A., Valsesia, F., 2015. The power of repetition: Repetitive lyrics in a song increase processing fluency and drive market success. Journal of Consumer Psychology 25(2), 187-199.

Pálovics, R., Benczúr, A., 2015. Temporal influence over the Last.fm social network. Social Network Analysis and Mining 5(1), 4.

Pachet, F., Roy, P., 2008. Hit song science is not yet a science. In Proceedings of the 9th International Society for Music Information Retrieval, Philadelphia, 355-360.

Panniello, U., Hill, S., Gorgoglione, M., 2016. The impact of profit incentives on the relevance of online recommendations. Electronic Commerce Research and Applications 20, 87-104.

Parisi, P, 2018. Copyright royalty board boosts songwriters' streaming pay nearly 50%. Variety. http://variety.com/2018/biz/news/copyright-royalty-board-boosts-songwriters-streaming-pay-nearly-50-1202679118/

Park, H.S., Yoo, J.O., Cho, S.B., 2006. A context-aware music recommendation system using fuzzy Bayesian networks with utility theory. In International Conference on Fuzzy Systems and Knowledge Discovery. Springer, Berlin / Heidelberg, 970-979.

Poddar, S., 2006. Music product as a durable good and online piracy. Review of Economic Research on Copyright Issues 3(2), 53-66.

Ren, J., Cheng, Z., Shen, J., Zhu, F., 2014. Influences of influential users: an empirical study of music social network. In Proceedings of International Conference on Internet Multimedia Computing and Service, ACM Press, New York, 411-415.

Ren, J., Kauffman, R. J., 2017. Understanding music track popularity in a social

network. In Proceedings of the 25th European Conference on Information Systems, Association for Information Systems, Atlanta, GA, 374-388.

Ren, J., Kauffman, R. J., 2018. Understanding streaming music diffusion in a semi-closed social environment. In Proceedings of Pacific Asia Conference on Information Systems, Association for Information Systems, Atlanta, GA, forthcoming.

Ren, J., Shen, J., Kauffman, R. J., 2016. What makes a music track popular in online social networks? In Proceedings of the 25th International Conference Companion on World Wide Web, ACM Press, New York, 95-96.

Resnikoff, P, 2016, How many streams does it take to earn $1? Take a look… Digital Music News. https://www.digitalmusicnews.com/2016/09/15/streaming-music-earn-1-dollar/

RIAA (Recording Industry Association of America), 2016. https://www.riaa.com/u-s-sales-database/.

RIAA (Recording Industry Association of America), 2017. http://www.riaa.com/wp-content/uploads/2017/09/RIAA-Mid-Year-2017-News-and-Notes2.pdf.

Ricci, F., Rokach, L., Shapira, B., 2010. Recommender Systems Handbook, 1st edition, Springer, New York.

Richter, F, 2017. The music streaming landscape. Statista. https://www.statista.com/chart/5152/music-streaming-subscribers/

Rios, M.C., McConnell, C.R., Brue, S.L., 2013. Economics: Principles, Problems, and Policies. McGraw-Hill, New York.

Salo, J., Lankinen, M., Mäntymäki, M., 2013. The use of social media for artist marketing: Music industry perspectives and consumer motivations. International Journal on Media Management 15(1), 23-41.

Sanchez, D, 2017. What streaming music services pay (updated for 2017). Digital Music News. https://www.digitalmusicnews.com/2017/07/24/what-streaming-music-services-pay-updated-for-2017/

Sanchez, D, 2018. What streaming music services pay (updated for 2018). Digital Music News. https://www.digitalmusicnews.com/2018/01/16/streaming-music-services-pay-2018/

Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S., 2007. Collaborative filtering recommender systems. In The Adaptive Web. Springer, Berlin / Heidelberg, Germany, 291-324.

Scharff, C., 2015. Blowing your own trumpet: Exploring the gendered dynamics of self-promotion in the classical music profession. The Sociological Review 63(1_suppl), 97-112.

Schedl, M., 2011. Analyzing the potential of microblogs for spatio-temporal popularity estimation of music artists. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 539-553.

Sharma, A., Cosley, D., 2016. Distinguishing between personal preferences and social influence in online activity feeds. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, ACM

Press, New York, 1091–1103.

Shmueli, G., 2010. To explain or to predict? Statistical Science, 25(3), 289-310.

Shmueli, G., Koppius, O.R., 2011. Predictive analytics in information systems research. MIS Quarterly 35(2), 553-572.

Simon, H.A., 2001. Science seeks parsimony, not simplicity: Searching for pattern in phenomena. Simplicity, inference and modelling: Keeping it sophisticatedly simple, Cambridge University Press, New York, 32-72.

Singhi, A., Brown, D.G., 2015. Can song lyrics predict hits? In Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR), Plymouth, UK.

Skowron, M., Lemmerich, F., Ferwerda, B., Schedl, M., 2017. Predicting genre preferences from cultural and socio-economic factors for music retrieval. In Proceedings of the European Conference on Information Retrieval. Springer, Cham, Switzerland, 561-567.

Spotify, 2017. Spotify terms and conditions of use. https://www.spotify.com/us/legal/end-user-agreement/#s10

Strobl, E.A., Tucker, C., 2000. The dynamics of chart success in the U.K. pre-recorded popular music industry. Journal of Cultural Economics 24(2), 113-134.

Susarla, A., Oh, J.H., Tan, Y., 2012. Social networks and the diffusion of user-generated content: Evidence from YouTube. Information Systems Research 23(1), 23-41.

Tan, S., Bu, J., Chen, C., Xu, B., Wang, C., He, X., 2011. Using rich social media information for music recommendation via hypergraph model. ACM Transactions on Multimedia Computing, Communications, and Applications, 78(1), 22.

Thies, F., Wessel, M., Benlian, A., 2014. Understanding the dynamic interplay of social buzz and contribution behavior within and between online platforms: Evidence from crowdfunding. In Proceedings of the 35th International Conference on Information Systems, Association for Information Systems, Atlanta, GA, 1-18.

Trusov, M., Bodapati, A.V., Bucklin, R.E., 2010. Determining influential users in internet social networks. Journal of Marketing Research 47(4), 643-658.

Varian, H. R., 2010. Intermediate Microeconomics: A Modern Approach, 8th ed. W.W. Norton, New York.

Waldfogel, J., 2015. Digitization and the quality of new media products: The case of music. In A. Goldfarb, S. Greenstein, and C. Tucker (eds.), Economic Analysis of the Digital Economy, University of Chicago Press, Chicago, 407-442.

Wang, X., Wang, Y., 2014. Improving content-based and hybrid music recommendation using deep learning. In Proceedings of the 22nd ACM International Conference on Multimedia. ACM Press, New York, 627-636.

Wlömert, N., Papies, D. 2016. On-demand streaming services and music industry revenues: Insights from Spotify's market entry. International Journal of Research in Marketing 33(2), 314-327.

Wooldridge, J.M., 2015. Introductory Econometrics: A Modern Approach, 5th ed.

Nelson Education, Toronto, Canada.

Xie, W., Zhu, F., Liu, S., Wang, K., 2015. Modelling cascades over time in microblogs. In Proceedings of 2015 IEEE International Conference on Big Data, IEEE Computer Society Press, Washington, DC, 677-686.

Yu, L., 2012. Using negative binomial regression analysis to predict software faults: A study of apache ant. International Journal of Information Technology and Computer Science 8, 63-70.

Yu, T., Benbasat, I., Cenfetelli, R.T., 2016. How to design interfaces for product recommendation agents to influence the purchase of environmentally-friendly products. In Proceedings of the 50th Hawaii International Conference on System Sciences, IEEE Computer Society Press, Washington, DC, 620-629.

Zhang, B., Shen, J., Xiang, Q., Wang, Y., 2009. CompositeMap: A novel framework for music similarity measure. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, 403-410.

Zhang, Y., 2017. Explainable recommendation: Theory and applications. Doctoral dissertation, arXiv preprint arXiv:1708.06409.

Zhang, Y., Zhao, Q., Zhang, Y., Friedman, D., Zhang, M., Liu, Y. Ma, S., 2016. Economic recommendation with surplus maximization. In Proceedings of the 25th International Conference Companion on World Wide Web, ACM Press, New York, 73-83.

Zhao, Q., Zhang, Y., Zhang, Y., Friedman, D., 2017. Multi-product utility maximization for economic recommendation. In Proceedings of the 10th ACM International Conference on Web Search and Data Mining, ACM Press, New York, 435-443.

Appendix

Appendix A. Understanding Music Popularity in Music Social Networks

Figure A1. Distributional Fit of Music Track Popularity *Duration*

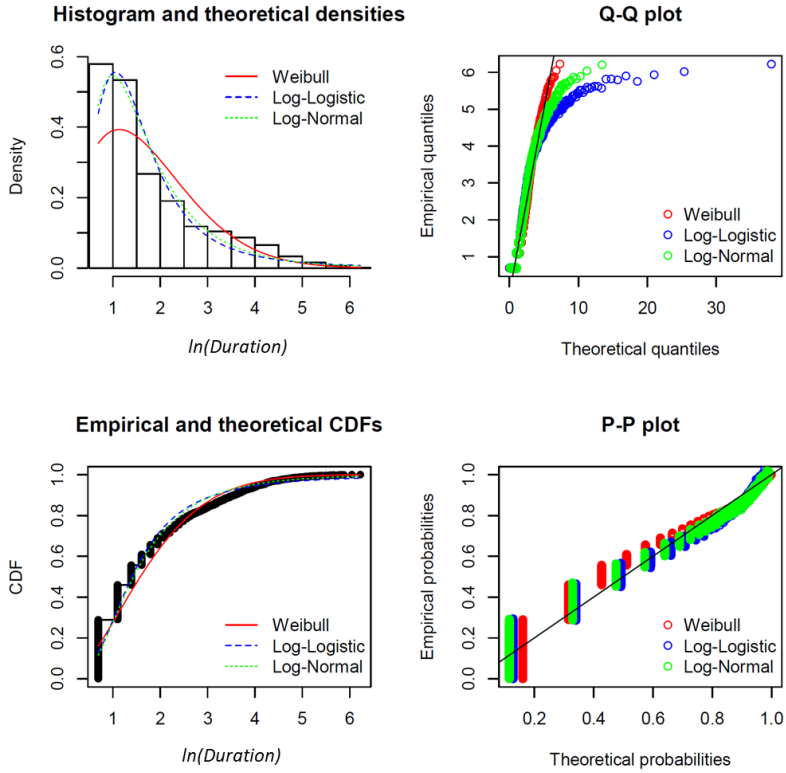


Figure A2. Distributional Fit of Music Track Popularity *Time2TopRank*

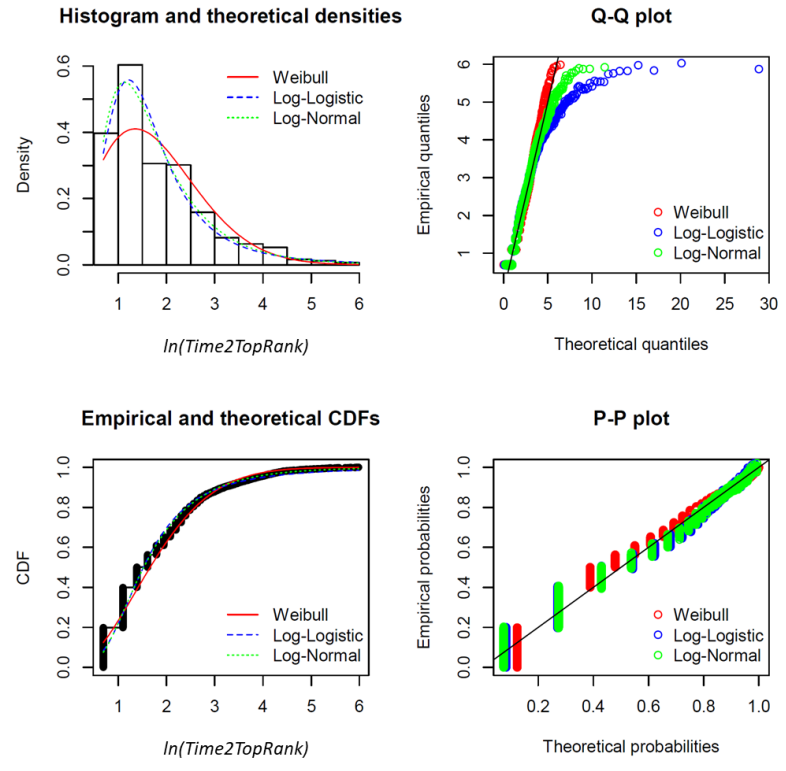


Table A1. Explanatory Results of Instrumental and Mood for Music Popularity *Duration* for Four Models

Constructs and Variables	Linear (SE)		Weibull Hazard (SE)		Log-Logistic Hazard (SE)		Log-Normal Hazard (SE)	
Instrumental								
Piano	-0.43*	(0.07)	-0.21*	(0.11)	-0.26*	(0.11)	-0.23*	(0.11)
Guitar	-0.05	(0.07)	-0.05	(0.03)	-0.04	(0.04)	-0.05	(0.03)
Trombone	1.55	(0.92)	0.82*	(0.47)	0.83*	(0.51)	0.85*	(0.48)
Cello	0.007	(0.064)	-0.02	(0.03)	0.006	(0.033)	0.004	(0.033)
Clarinet	0.06	(0.11)	-0.01	(0.06)	-0.002	(0.060)	0.005	(0.059)
Drumkit	-0.002	(0.059)	-0.005	(0.031)	0.01	(0.03)	0.002	(0.031)
Flute	-0.05	(0.07)	0.02	(0.04)	-0.03	(0.04)	-0.02	(0.04)
Saxophone	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Snare	0.11	(0.08)	0.04	(0.04)	0.06	(0.04)	0.06	(0.04)
Trumpet	0.25	(0.81)	-0.02	(0.04)	0.20	(0.41)	0.15	(0.42)
Tuba	-2.44	(4.89)	-3.03	(2.49)	-1.36	(2.07)	-1.71	(2.54)
Violin	-0.29	(0.29)	-0.14	(0.14)	-0.05	(0.15)	-0.10	(0.15)
Mood								
Rollicking	0.07	(0.07)	0.03	(0.04)	0.03	(0.04)	0.04	(0.04)
Literature	0.07	(0.07)	0.001	(0.036)	0.05	(0.04)	0.04	(0.04)
Aggressive	0.005	(0.070)	-0.02	(0.04)	-0.001	(0.037)	0.001	(0.036)
Humorous	0.03	(0.09)	-0.05	(0.05)	0.04	(0.05)	0.03	(0.05)
Passionate	-0.06	(0.10)	-0.08	(0.05)	0.02	(0.05)	-0.003	(0.052)

Notes. Signif: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A2. Explanatory Results of Instrumental and Mood for Music Popularity *Time2TopRank* for Four Models

Constructs and Variables	Linear (SE)		Weibull Hazard (SE)		Log-Logistic Hazard (SE)		Log-Normal Hazard (SE)	
Instrumental								
Piano	0.30*	(0.18)	0.14	(0.09)	0.13	(0.10)	0.13	(0.09)
Guitar	0.03	(0.06)	0.01	(0.03)	0.03	(0.03)	0.03	(0.03)
Trombone	0.28	(0.82)	0.09	(0.38)	0.26	(0.44)	0.27	(0.43)
Cello	-0.20***	(0.06)	-0.10***	(0.03)	-0.13***	(0.12)	-0.12***	(0.03)
Clarinet	0.06	(0.10)	0.03	(0.05)	0.03	(0.05)	0.03	(0.05)
Drumkit	-0.02	(0.05)	-0.01	(0.03)	-0.01	(0.02)	-0.01	(0.03)
Flute	0.12**	(0.06)	0.08***	(0.03)	0.06*	(0.03)	0.07*	(0.03)
Saxophone	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Snare	0.11	(0.07)	0.10*	(0.04)	0.05	(0.04)	0.06	(0.04)
Trumpet	-0.36	(0.72)	0.12	(0.31)	-0.87*	(0.45)	-0.52	(0.38)
Tuba	-3.43	(4.37)	-2.84	(2.15)	-1.57	(1.85)	-1.46	(2.28)
Violin	-0.48	(0.26)	-0.22*	(0.13)	-0.29**	(0.14)	-0.26*	(0.14)
Mood								
Rollicking	-0.02	(0.06)	-0.05	(0.03)	-0.02	(0.03)	-0.01	(0.03)
Literature	-0.07	(0.06)	-0.05	(0.03)	-0.05	(0.03)	-0.04	(0.03)
Aggressive	0.07	(0.06)	0.03	(0.03)	0.05	(0.03)	0.03	(0.03)
Humorous	-0.09	(0.08)	-0.05	(0.04)	-0.04	(0.04)	-0.04	(0.04)
Passionate	-0.10	(0.09)	-0.08*	(0.04)	-0.07	(0.05)	-0.09*	(0.05)

Notes. Signif: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Appendix B. Understanding Streaming Music Diffusion in a Semi-Closed Social Environment

Table B1. DiD Regression Results for Control Variables at the Geographic-Level for U.S. Users

	Monthly		Weekly	
Main Effect Variables	**Music Content (I) (SE)**	**Non-Music Content (II) (SE)**	**MusicContent (III) (SE)**	**Non-Music Content (IV) (SE)**
Artist Genre				
Rock	0.17 * (0.09)	-0.03 (0.10)	0.30 *** (0.05)	0.02 (0.05)
Alternative	0.50 *** (0.13)	0.57 *** (0.14)	0.36 *** (0.07)	0.22 *** (0.07)
Indie	0.41 *** (0.09)	0.59 *** (0.10)	0.39 *** (0.05)	0.54 *** (0.05)
Pop	0.23 * (0.12)	0.40 ** (0.17)	0.28 *** (0.06)	0.43 *** (0.09)
Hip-hop	0.42 (0.27)	0.15 (0.25)	0.65 *** (0.15)	0.08 (0.13)
Rap	1.48 *** (0.47)	0.08 (0.36)	0.88 *** (0.26)	0.25 (0.19)
R&B	0.59 ** (0.25)	0.14 (0.20)	0.57 *** (0.13)	0.19 * (0.11)
Electronic	0.61 ** (0.24)	0.98 *** (0.25)	0.86 *** (0.13)	1.05 *** (0.13)
Metal	0.01 (0.08)	0.11 (0.12)	0.03 (0.04)	-0.07 (0.06)
Folk	0.47 ** (0.23)	0.57 ** (0.26)	0.51 *** (0.12)	0.53 *** (0.13)
Soul	0.36 (0.23)	0.60 * (0.36)	0.86 *** (0.12)	1.33 *** (0.18)
Experimental	NA	NA	NA	NA
Punk	0.15 (0.11)	0.39 *** (0.09)	0.004 (0.059)	0.11 ** (0.05)
Classic	0.12 (0.47)	0.60 *** (0.23)	0.42 * (0.25)	0.37 *** (0.12)
Jazz	0.56 * (0.30)	0.60 ** (0.25)	0.79 ***(0.16)	0.83 *** (0.13)
Blues	0.86 (0.70)	0.46 ** (0.23)	0.29 (0.37)	0.59 *** (0.12)
Country	-0.11 (0.24)	-0.21 (0.28)	-0.11 (0.13)	-0.04 (0.14)
Reggae	6.26 * (2.57)	12.57 ***(3.67)	9.66 ***(1.85)	4.30 ** (1.89)
Artist Characteristics				
Artist: Male	-0.13 (0.12)	0.16 (0.14)	0.01 (0.06)	-0.09 (0.07)
Artist: Female	-0.27 (0.18)	-0.14 (0.16)	-0.10 (0.10)	-0.65 *** (0.08)
MajorLabel	0.003 (0.092)	0.22 ** (0.09)	0.17 *** (0.05)	0.41 *** (0.05)
LongPopLast.fm	0.013***(0.004)	0.007** (0.002)	0.011***(0.002)	0.005***(0.001)
LongPopBB	0.011** (0.005)	0.013** (0.006)	0.006***(0.003)	0.003 (0.003)

Notes. Overall model obs.: I – 512, II – 376, III – 1,280, IV – 940. 105 artists had *Music Content Info*; 94 had *Non-Music Content Info*; I – 512512 mo. obs. = (105 + 23 base case) × 4; III – 1,280wk obs. = (105 + 23 base case) × 10. Shape parameters α: I – .42, II – .46, III – .29, IV – .21; pseudo-R^2: 1 – 68.4%, II – 62.6%, III – 66.5%, IV – 74.5%. *Artist#Listeners* had similar results. Signif: * $p < .10$; ** $p < .05$; *** $p < .01$.

Table B2. DiD Regression Results for External Information at the Geographic-Level for both U.S. and U.K. Users

Main Effect Variables	Monthly		Weekly	
	Music Content (I) (SE)	Non-Music Content (II) (SE)	MusicContent (III) (SE)	Non-Music Content (IV) (SE)
Constant	7.57 *** (0.23)	7.44 *** (0.21)	4.92 *** (0.17)	4.95 *** (0.17)
CtryExtInfoRel	0.31 *** (0.07)	0.27 * (0.08)	0.17 * (0.09)	0.18 * (0.11)
AfterRelease	0.52 *** (0.07)	0.06 (0.08)	0.12 * (0.08)	0.02 (0.09)
CtryExtInfoRel × AfterRelease	0.03 (0.12)	0.04 (0.11)	0.02 (0.10)	0.03 (0.11)
ArtistExtInfoType				
News-Artist Life	Base case	Base case	Base case	Base case
News-Music-Related Info		0.40 *** (0.13)		0.12 * (0.08)
Tour, Concert		0.23 * (0.12)		0.06 (0.09)
Live TV Show		0.26 ** (0.11)		0.42 *** (0.09)
Live Performance / Festival		0.11 (0.12)		0.05 (0.09)
Single Song Release	0.33 *** (0.11)		0.23 *** (0.08)	
Album Release	0.48 *** (0.10)		0.26 *** (0.08)	
Music Video Release	0.94 *** (0.17)		0.79 *** (0.13)	
ExtInfoWeekAfter				
WeekAfter-1			Base case	Base case
WeekAfter1			0.12 * (0.06)	0.11 * (0.07)
WeekAfter2			0.24 * (0.06)	0.11 (0.07)
WeekAfter3			0.11 (0.06)	0.03 (0.07)
WeekAfter4			0.005 (0.05)	0.001 (0.06)
Artist Genre				
Rock	0.05 (0.08)	-0.16 * (0.09)	0.31 *** (0.06)	0.18 ** (0.08)
Alternative	0.34 *** (0.13)	0.85 *** (0.13)	0.27 *** (0.09)	0.38 *** (0.10)
Indie	0.23 *** (0.08)	0.46 *** (0.08)	0.29 *** (0.06)	0.46 *** (0.06)
Pop	0.004 (0.108)	0.66 *** (0.16)	0.27 *** (0.08)	0.93 *** (0.12)
Hip-hop	-0.10 (0.26)	0.24 (0.27)	0.57 *** (0.21)	0.56 *** (0.21)
Rap	1.82 *** (0.48)	0.20 (0.40)	1.08 *** (0.37)	0.22 (0.32)
R&B	0.37 (0.24)	0.12 (0.21)	0.77 *** (0.17)	0.60 *** (0.16)
Electronic	0.88 *** (0.16)	0.93 *** (0.21)	0.66 *** (0.11)	1.07 *** (0.17)
Metal	0.03 (0.07)	-0.05 (0.11)	0.18 *** (0.05)	0.14 * (0.08)
Folk	0.40 ** (0.17)	1.10 *** (0.23)	0.47 *** (0.12)	0.96 *** (0.18)
Soul	0.29 (0.22)	0.27 (0.23)	1.08 *** (0.16)	0.91 *** (0.19)
Experimental	NA	NA	NA	NA
Punk	0.06 (0.11)	0.10 (0.09)	0.22 *** (0.08)	0.22 *** (0.07)
Classic	0.21 (0.30)	0.85 *** (0.20)	-0.25 (0.22)	0.57 *** (0.16)
Jazz	0.53 ** (0.27)	0.42 * (0.23)	0.83 *** (0.20)	1.09 *** (0.18)
Blues	-0.13 (0.61)	1.13 *** (0.18)	0.89 ** (0.44)	1.06 *** (0.15)
Country	-0.40 * (0.23)	-0.67 ** (0.27)	0.11 (0.17)	0.08 (0.22)
Reggae	8.40 ** (3.52)	11.91 ***(3.52)	8.20 *** (2.59)	9.39 *** (2.82)
Artist Characteristics				
Artist: Male	-0.26*** (0.09)	-0.31*** (0.11)	-0.13** (0.06)	-0.08 (0.09)
Artist: Female	-0.10 (0.13)	-0.23* (0.13)	-0.32*** (0.09)	-0.74***(0.11)
MajorLabel	0.21 *** (0.07)	0.007 (0.008)	0.28 *** (0.05)	0.28 *** (0.06)
LongPopLast.fm	0.01 *** (0.003)	0.008** (0.001)	0.009***(0.002)	0.007***(0.001)
LongPopBB	0.008* (0.005)	0.004* (0.003)	0.008***(0.004)	0.007***(0.003)

Notes. Overall model obs.: I – 760, II – 500, III – 1,900, IV – 1,250. (105+45) = 150 artists had *Music Content Info*; (94+31) = 125 had *Non-Music Content Info*; I – 760 mo. obs. = (150 + 40 base case) × 4; II – 500 mo. obs. = 125× 4; III – 1,900 wk obs. = (150 + 40 base case) × 10; IV – 1,250 wk. obs. = 125 × 10. Shape parameters α: I – .49, II – .34, III – .66, IV – .56; pseudo-R^2: I – 63.7%, II – 67.8%, III – 43.9%, IV – 47.3%. *Artist#Listeners* had similar results. Signif: * $p < .10$; ** $p < .05$; *** $p < .01$.

Appendix C. Two-Sided Value-based Music Promotion and Recommendation

Definition C1. Gamma Function

The gamma function, denoted by $\Gamma(a)$, is:

$$\Gamma(a) = \int_0^{\infty} e^{-1} t^{a-1} dt, \qquad a > 0$$

The properties of the gamma function include:

1. $\Gamma(a) = (a-1)\Gamma(a-1)$
2. $\Gamma(a) = (a-1)!$ if a is a positive integer
3. $\Gamma(0) = \infty,\ \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
4. $\Gamma(na) = (2\pi)^{(1-n)/2}(n)^{na-1/2}\prod_{k=0}^{n-1}\Gamma(a+\frac{k}{n})$, where n is a positive integer.

Figure C1. Experiment Workflow of KNN Collaboration-based Recommendation

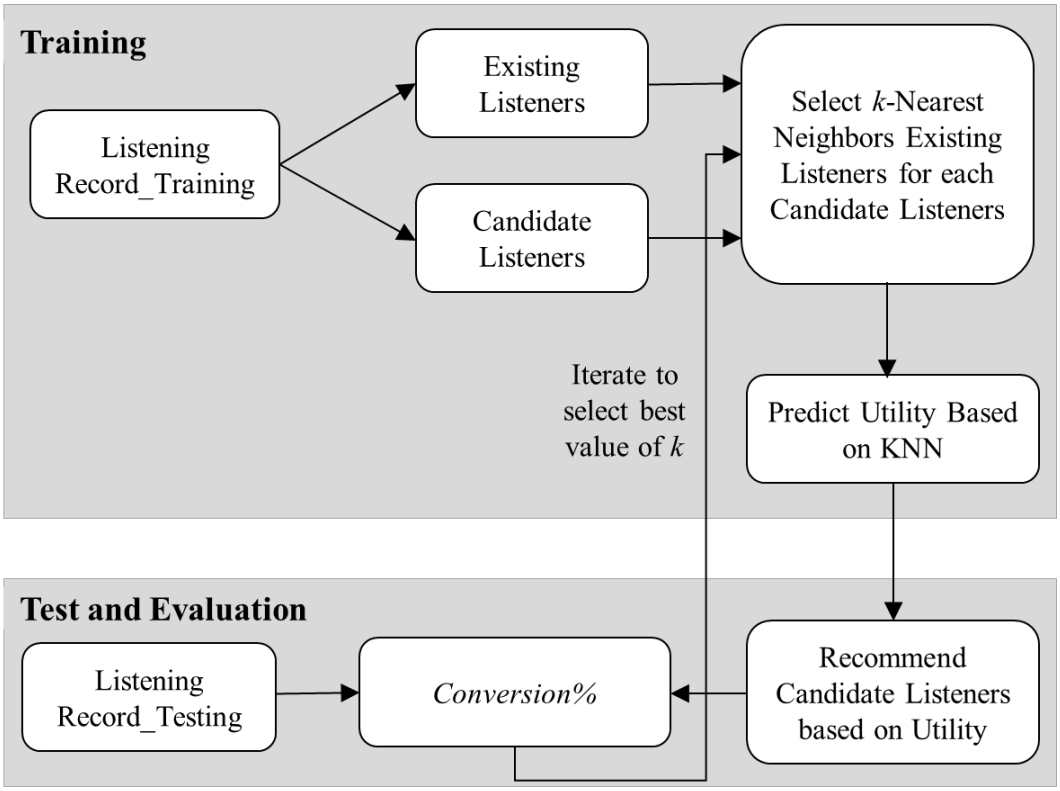


Figure C2. Conversion Rate of KNN Collaboration-based Recommendation

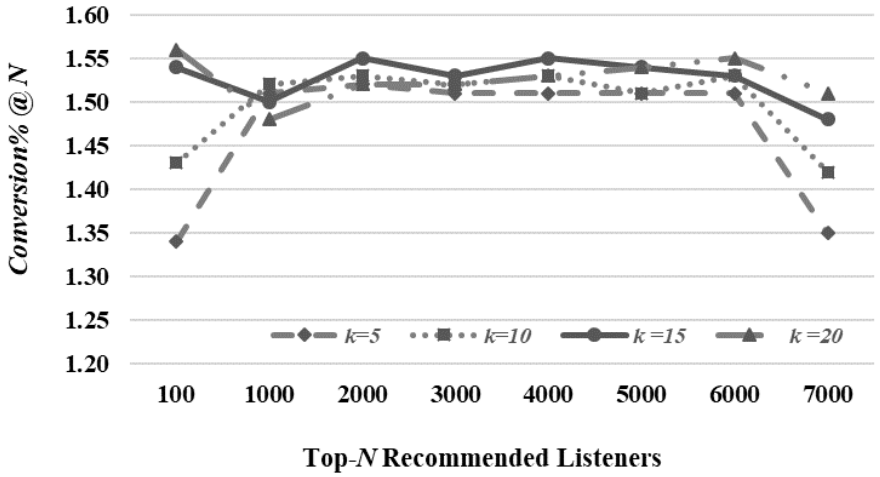


Figure C3. Boxplot of Performances of KNN and Value-based Methods (*Non-Music Content Information)*

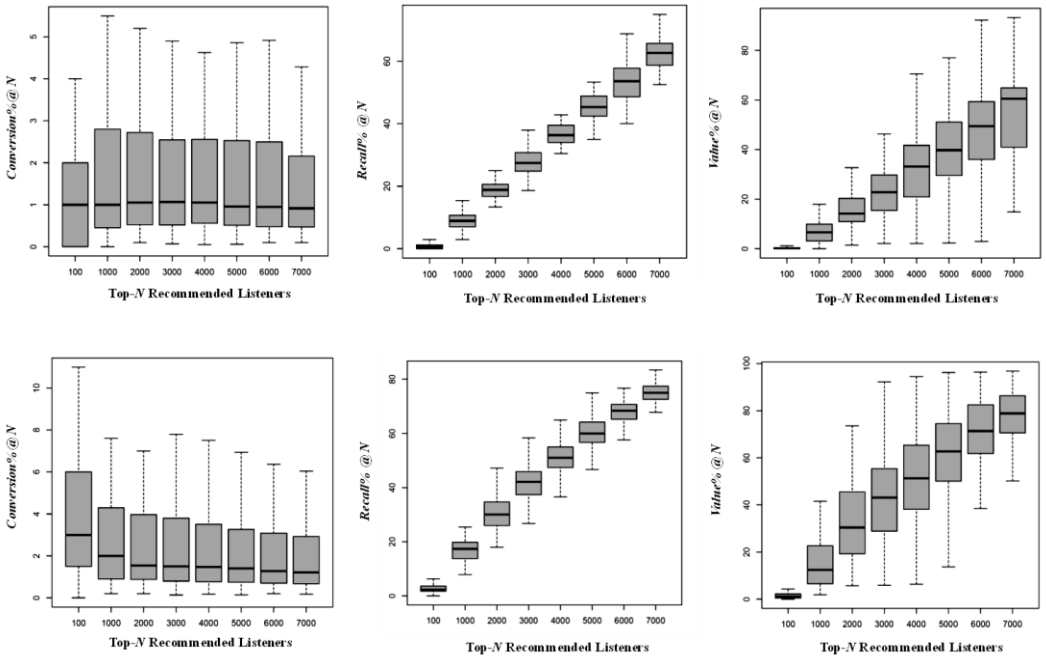


Table C1. Value-based Method Performance by Subtype of *Music Content Information* (Full)

TOP-*N*	*N* = 100 MEAN (S.D.)			*N* = 1,000 MEAN (S.D.)		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	2.71 (3.22)	1.84 (2.09)	2.11 (6.57)	2.34 (2.42)	14.88 (5.15)	15.20 (11.38)
Type6-Single-Song Release	3.14 (2.72)	2.60 (2.52)	2.59 (6.01)	2.41 (2.01)	15.61 (5.16)	15.72 (11.08)
Type7-Album Release	2.33 (3.23)	1.45 (1.76)	1.93 (7.16)	2.20 (2.50)	14.48 (5.00)	14.83 (11.62)
Type8-Music-Video Release	4.20 (5.36)	1.57 (1.86)	1.29 (1.71)	3.48 (3.79)	14.95 (7.19)	16.25 (12.79)
TOP-*N*	***N* = 2,000 MEAN (S.D.)**			***N* = 3,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	2.12 (2.17)	27.07 (7.94)	28.83 (15.32)	2.01 (2.04)	37.60 (8.92)	40.09 (16.68)
Type6-Single-Song Release	2.26 (1.83)	28.99 (7.17)	28.79 (15.36)	1.73 (1.77)	39.40 (8.55)	40.44 (15.05)
Type7-Album Release	1.97 (2.24)	26.13 (7.87)	28.88 (15.61)	1.87 (2.08)	36.58 (9.06)	38.93 (17.92)
Type8-Music-Video Release	3.07 (3.34)	26.09 (12.31)	28.51 (14.97)	2.87 (3.12)	38.09 (9.68)	50.54 (7.31)
TOP-*N*	***N* = 4,000 MEAN (S.D.)**			***N* = 5,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.93 (1.96)	47.99 (9.20)	51.21 (18.62)	1.84 (1.86)	57.22 (8.70)	60.57 (19.42)
Type6-Single-Song Release	2.04 (1.73)	49.45 (7.90)	50.77 (16.71)	1.93 (1.61)	58.70 (7.39)	59.99 (19.04)
Type7-Album Release	1.79 (1.99)	47.30 (9.65)	50.75 (20.34)	1.72 (1.91)	56.55 (9.56)	60.28 (20.46)
Type8-Music-Video Release	2.73 (2.99)	47.03 (12.19)	58.77 (5.27)	2.60 (2.87)	55.91 (5.34)	67.06 (8.11)
TOP-*N*	***N* = 6,000 MEAN (S.D.)**			***N* = 7,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.77 (1.80)	66.27 (7.64)	70.03 (16.99)	1.70 (1.72)	74.06 (7.04)	75.81 (16.75)
Type6-Single-Song Release	1.85 (1.53)	67.13 (6.22)	70.00 (15.77)	1.75 (1.44)	74.06 (6.06)	75.59 (15.97)
Type7-Album Release	1.66 (1.84)	65.79 (8.60)	69.49 (18.28)	1.59 (1.76)	73.84 (7.79)	75.46 (17.83)
Type8-Music-Video Release	2.56 (2.84)	66.55 (3.04)	76.08 (7.89)	2.50 (2.76)	76.45 (2.77)	80.85 (8.25)

Note. Data: 88 artists with *Music-Content* External Information. Type6: 29; Type7: 54; *Type9*: 5. *C%*: *Conversion%*, *R%*: *Recall%*; *V%*: *Value%.*

Table C2. Value-based Method Performance by Subtype of *Non-Music Content Information* (Full)

TOP-*N*	*N* = 100 MEAN (S.D.)			*N* = 1,000 MEAN (S.D.)		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	3.96 (3.16)	2.93 (2.22)	2.53 (5.71)	2.75 (2.21)	17.75 (5.78)	17.37 (14.96)
Type1-News, Artist Life	3.86 (3.37)	3.06 (2.44)	3.79 (8.54)	2.80 (2.47)	18.19 (6.48)	20.95 (17.06)
Type2-News, Music-Related Info	3.88 (3.98)	2.18 (1.20)	1.40 (2.80)	2.49 (2.20)	14.92 (2.98)	6.44 (3.87)
Type3-Tour, Concert	2.83 (2.04)	5.23 (3.69)	2.77 (3.15)	1.18 (0.73)	20.02 (3.43)	19.45 (11.57)
Type4-Live TV Show	4.90 (3.35)	2.29 (1.16)	0.93 (0.54)	3.84 (2.14)	18.20 (6.18)	20.27 (17.64)
Type5-Live Performance / Festival	4.00 (2.67)	2.50 (1.21)	2.22 (3.72)	2.72 (1.99)	17.29 (6.67)	14.43 (11.85)
TOP-*N*	***N* = 2,000 MEAN (S.D.)**			***N* = 3,000 MEAN (S.D.)**		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	2.45 (2.01)	30.99 (7.70)	33.28 (18.62)	2.25 (1.87)	41.96 (8.45)	43.80 (20.46)
Type1-News, Artist Life	2.46 (2.23)	31.84 (8.22)	34.44 (18.29)	2.25 (2.01)	42.92 (8.76)	45.71 (22.24)
Type2-News, Music-Related Info	2.45 (2.19)	29.62 (7.80)	34.58 (27.43)	2.32 (2.09)	41.20 (8.23)	39.95 (25.09)
Type3-Tour, Concert	1.06 (0.68)	35.20 (8.51)	31.80 (13.74)	0.96 (0.66)	46.31 (5.12)	47.81 (18.04)
Type4-Live TV Show	3.36 (1.98)	30.55 (6.56)	33.05 (15.01)	3.08 (1.81)	41.59 (4.68)	44.84 (14.77)
Type5-Live Performance / Festival	2.28 (1.75)	28.21 (7.13)	30.89 (20.23)	2.15 (1.79)	38.33 (11.78)	39.45 (21.67)
TOP-*N*	***N* = 4,000 MEAN (S.D.)**			***N* = 5,000 MEAN (S.D.)**		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	2.11 (1.78)	51.79 (8.02)	52.60 (20.08)	1.99 (1.68)	60.08 (6.88)	60.86 (19.18)
Type1-News, Artist Life	2.09 (1.92)	52.29 (6.97)	53.02 (21.11)	1.96 (1.80)	60.23 (5.87)	60.22 (22.35)
Type2-News, Music-Related Info	2.21 (1.98)	52.14 (7.61)	51.62 (22.09)	2.10 (1.89)	61.13 (6.89)	59.82 (21.45)
Type3-Tour, Concert	0.88 (0.62)	56.13 (6.34)	55.25 (20.31)	0.81 (0.59)	63.75 (5.36)	59.02 (22.14)
Type4-Live TV Show	2.90 (1.73)	51.97 (5.04)	55.92 (16.00)	2.73 (1.66)	60.76 (5.02)	66.18 (12.62)
Type5-Live Performance / Festival	2.03 (1.70)	7.69 (12.38)	47.61 (22.76)	1.93 (1.60)	56.05 (10.05)	58.82 (16.37)
TOP-*N*	***N* = 6,000 MEAN (S.D.)**			***N* = 7,000 MEAN (S.D.)**		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.89 (1.60)	68.34 (5.81)	69.11 (17.39)	1.81 (1.56)	75.38 (4.83)	75.30 (16.64)
Type1-News, Artist Life	1.84 (1.66)	68.94 (5.15)	68.74 (20.46)	1.75 (1.61)	75.69 (3.88)	74.46 (18.95)
Type2-News, Music-Related Info	2.00 (1.85)	67.56 (4.41)	65.60 (18.37)	1.92 (1.80)	74.78 (3.49)	71.93 (17.91)
Type3-Tour, Concert	0.78 (0.61)	72.93 (3.79)	71.15 (22.69)	0.72 (0.56)	78.57 (5.01)	73.62 (24.35)
Type4-Live TV Show	2.60 (1.63)	68.70 (3.84)	75.28 (8.76)	2.50 (1.61)	76.51 (3.92)	83.69 (6.02)
Type5-Live Performance / Festival	1.87 (1.54)	64.58 (8.62)	65.30 (13.69)	1.79 (1.46)	72.15 (6.89)	72.39 (12.12)

Note. Data: 55 artists with *Non-Music Content* External Information. *Type1*: 21; *Type2*: 8; *Type3*: 6; *Type4*: 10; *Type5*: 10. *C%*: *Conversion%*, *R%*: *Recall%*; *V%*: *Value%*.

Table C3. KNN Performance by Subtype of *Music Content Information* (Full)

TOP-*N*	*N* = 100 MEAN (S.D.)			N = 1,000 MEAN (S.D.)		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.66 (2.09)	0.97 (1.23)	1.60 (4.21)	1.47 (1.55)	9.30 (4.29)	11.27 (12.74)
Type6-Single-Song Release	1.76 (1.92)	0.95 (1.02)	1.57 (3.04)	1.47 (1.16)	9.52 (3.52)	10.51 (11.69)
Type7-Album Release	1.59 (2.18)	1.02 (1.37)	1.73 (4.91)	1.40 (1.65)	9.30 (4.77)	11.60 (13.70)
Type8-Music-Video Release	1.80 (2.49)	0.60 (0.67)	0.34 (0.41)	2.20 (2.55)	8.02 (3.29)	12.04 (8.93)
TOP-*N*	**N = 2,000 MEAN (S.D.)**			***N* = 3,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.48 (1.58)	18.30 (5.31)	20.11 (14.37)	1.49 (1.55)	26.77 (5.93)	29.10 (15.54)
Type6-Single-Song Release	1.50 (1.23)	18.50 (4.18)	18.12 (12.81)	1.52 (1.26)	27.47 (5.98)	28.02 (14.35)
Type7-Album Release	1.40 (1.64)	18.34 (5.91)	20.64 (15.50)	1.41 (1.62)	26.57 (6.09)	29.20 (16.71)
Type8-Music-Video Release	2.23 (2.58)	16.63 (4.92)	25.99 (9.10)	2.13 (2.43)	24.90 (4.01)	34.25 (8.74)
TOP-*N*	***N* = 4,000 MEAN (S.D.)**			***N* = 5,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.52 (1.55)	35.75 (6.69)	38.41 (15.47)	1.51 (1.55)	44.33 (7.15)	46.29 (15.09)
Type6-Single-Song Release	1.55 (1.25)	38.04 (5.96)	38.13 (15.90)	1.52 (1.25)	46.42 (6.41)	46.37 (14.80)
Type7-Album Release	1.43 (1.60)	34.59 (6.86)	38.12 (15.88)	1.43 (1.60)	43.31 (7.25)	45.16 (15.22)
Type8-Music-Video Release	2.23 (2.55)	34.56 (6.75)	43.12 (8.50)	2.22 (2.54)	42.83 (8.99)	57.58 (13.28)
TOP-*N*	***N* = 6,000 MEAN (S.D.)**			***N* = 7,000 MEAN (S.D.)**		
Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.51 (1.53)	53.23 (6.82)	54.51 (16.01)	1.53 (1.56)	62.58 (6.77)	64.21 (14.91)
Type6-Single-Song Release	1.49 (1.23)	53.97 (6.44)	53.43 (16.27)	1.51 (1.24)	63.53 (6.03)	64.89 (15.01)
Type7-Album Release	1.46 (1.60)	53.09 (6.69)	54.14 (16.28)	1.47 (1.62)	62.08 (7.20)	63.11 (15.39)
Type8-Music-Video Release	2.19 (2.52)	50.31 (10.51)	64.55 (8.88)	2.23 (2.55)	62.11 (7.13)	71.45 (6.67)

Note. Data: 88 artists with *Music-Content* External Information. T*ype6*: 29; *Type7*: 54; *Type9*: 5. *C%*: *Conversion%*, *R%*: *Recall%*; *V%*: *Value%.*

Table C4. KNN Performance by Subtype of *Non-Music Content Information* (Full)

TOP-N	N= 100 MEAN (S.D.)			N = 1,000 MEAN (S.D.)		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.35 (1.47)	0.82 (0.99)	0.45 (0.76)	1.57 (1.47)	8.97 (3.20)	7.67 (6.11)
Type1-News, Artist Life	1.05 (1.12)	0.80 (1.25)	0.52 (0.94)	1.41 (1.36)	9.18 (3.47)	8.18 (6.41)
Type2-News, Music-Related Info	1.50 (1.60)	0.73 (0.68)	0.50 (1.00)	1.64 (1.52)	9.08 (2.34)	8.15 (8.95)
Type3-Tour, Concert	0.50 (0.55)	0.90 (1.20)	0.32 (0.59)	0.42 (0.33)	7.55 (5.12)	8.29 (7.96)
Type4-Live TV Show	2.30 (1.70)	1.02 (0.85)	0.48 (0.51)	2.43 (1.50)	10.43 (1.66)	8.33 (3.26)
Type5-Live Performance / Festival	1.40 (1.84)	0.68 (0.68)	0.29 (0.52)	1.66 (1.73)	7.84 (2.81)	5.15 (4.00)
TOP-N	**N= 2,000 MEAN (S.D.)**			**N = 3,000 MEAN (S.D.)**		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.65 (1.54)	18.58 (3.91)	16.53 (10.43)	1.61 (1.48)	27.18 (5.46)	23.54 (12.43)
Type1-News, Artist Life	1.59 (1.60)	18.78 (4.26)	19.60 (14.23)	1.56 (1.57)	26.75 (6.48)	27.47 (16.32)
Type2-News, Music-Related Info	1.79 (1.71)	19.40 (1.36)	14.21 (9.40)	1.73 (1.64)	28.06 (3.06)	19.08 (11.16)
Type3-Tour, Concert	0.53 (0.46)	16.67 (5.59)	14.48 (9.34)	0.53 (0.42)	25.66 (6.98)	18.60 (10.61)
Type4-Live TV Show	2.28 (1.55)	18.70 (3.70)	15.85 (5.16)	2.18 (1.42)	28.01 (5.30)	22.68 (5.59)
Type5-Live Performance / Festival	1.71 (1.59)	18.55 (3.97)	13.86 (4.56)	1.68 (1.49)	27.48 (4.45)	22.67 (8.49)
TOP-N	**N= 4,000 MEAN (S.D.)**			**N = 5,000 MEAN (S.D.)**		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.60 (1.44)	36.24 (6.66)	33.48 (17.06)	1.58 (1.43)	44.71 (7.69)	40.57 (17.90)
Type1-News, Artist Life	1.56 (1.54)	36.00 (8.84)	38.01 (18.98)	1.55 (1.55)	44.28 (10.09)	42.93 (19.40)
Type2-News, Music-Related Info	1.73 (1.67)	38.22 (2.70)	33.92 (25.60)	1.69 (1.59)	47.09 (3.46)	40.88 (24.40)
Type3-Tour, Concert	0.55 (0.46)	33.99 (5.89)	29.69 (14.36)	0.56 (0.47)	42.26 (10.51)	36.25 (19.80)
Type4-Live TV Show	2.12 (1.33)	36.94 (5.81)	29.07 (6.92)	2.11 (1.40)	44.20 (5.49)	35.40 (9.62)
Type5-Live Performance / Festival	1.67 (1.46)	35.81 (5.15)	30.30 (13.49)	1.65 (1.38)	45.66 (4.19)	43.11 (15.86)
TOP-N	**N= 6,000 MEAN (S.D.)**			**N = 7,000 MEAN (S.D.)**		
Non-Music Content Info.	***C%***	***R%***	***V%***	***C%***	***R%***	***V%***
ALL	1.57 (1.44)	52.98 (7.44)	47.54 (18.78)	1.40 (1.20)	61.80 (6.28)	54.14 (18.49)
Type1-News, Artist Life	1.55 (1.55)	53.28 (9.26)	51.10 (21.40)	1.25 (1.10)	62.27 (7.45)	58.63 (17.75)
Type2-News, Music-Related Info	1.67 (1.62)	54.84 (6.16)	44.62 (25.94)	1.66 (1.62)	62.82 (5.88)	50.07 (27.08)
Type3-Tour, Concert	0.55 (0.44)	49.91 (9.17)	44.41 (17.97)	0.55 (0.43)	59.56 (7.44)	53.62 (20.24)
Type4-Live TV Show	2.10 (1.41)	52.95 (5.06)	42.44 (9.51)	1.79 (1.06)	61.99 (4.66)	48.20 (12.17)
Type5-Live Performance / Festival	1.62 (1.37)	52.75 (5.50)	49.41 (15.11)	1.60 (1.36)	61.33 (5.68)	54.95 (16.82)

Note. Data: 55 artists with *Non-Music Content* External Information. Type 1: 21; Type 2: 8; Type 3: 6; Type 4: 10; Type 5: 10. *C%: Conversion%, R%: Recall%; V%: Value%.*