(HOW) DOES DATA-BASED MUSIC DISCOVERY WORK?

Complete Research

Akarapat Charoenpanich, LSE, UK, a.charoenpanich@lse.ac.uk

Aleksi Aaltonen, Warwick Business School, UK, aake@iki.fi

Abstract

This paper analyses a new type of business operations that mediate the production and consumption of music. Online environment has largely abolished constraints on the variety of music that can be economically distributed, but, at the same time, it reveals another problem. How do people learn what music items do they want to listen to? In the music industry, the product space consists of thousands of artists, songs and albums, and is expanding rapidly. More effective forms of music discovery could therefore create considerable new value by allowing people to listen to music that better matches their taste. We analyse data from Last.fm music discovery service that deploys a collaborative filtering recommender system and social media features to aid music discovery. The analysis finds evidence that the new form of music discovery is valuable to consumers, yet it is relatively less important than an opportunity to listen to music for free. The findings lead us to discuss how the nature of analytical problem and product space, consumer taste, and social media features shape the potential value of created by big data.

Keywords: Big data, Collaborative filtering, Last.fm, Music industry, Path analysis

1 Introduction

In this paper, we study data-based business operations that mediate the production and consumption of music in the digital ecosystem. We define music discovery as a process by which people identify new music items that are subsequently incorporated into individual music consumption. Music discovery can happen in many different ways. For instance, consumers may actively browse racks of CDs, search online catalogues, or be guided by social cues and recommendations from their environment. Importantly, people often do not know what they want to listen until they have actually started listening to it, which gives music discovery often an exploratory nature. It differs considerably from known-item type seeking, that is, locating items that they already know (Morville and Rosenfeld, 2006). The limitations of physical distribution channels have traditionally pushed music consumption toward the most popular artists, and there is a hope that new digital platforms could unleash the potential of niche items in the long tail of consumer demand (Anderson 2006; Celma 2008).

Assuming that people's 'true' music taste is more diverse than what traditionally narrow distribution channels have been able to serve, more effective forms of music discovery can create considerable new value by allowing people to listen to music that better matches their taste. Our empirical analysis focuses on an online service that musters social consumption data to provide personalized music recommendations. We analyse a dataset retrieved from Last.fm music discovery service that deploys a collaborative filtering recommender system and social media features to aid music discovery. The company was founded in 2002 and is one of the most popular services of its kind today. Last.fm users submit listening data from over 600 playback applications, services and devices to receive recommendations for further music items. Whether these recommendations are valuable to consumers is, however, an empirical question that goes to the heart of a business built on data and data analytics (Aaltonen and Tempini, 2014). We ask the following research question: *Does the new form of data-based music discovery provide value to consumers?*

To answer such a question, one needs to be able to separate music discovery and its value from the value of sheer music acquisition, that is, the surplus between price and the utility of consuming music. Consumers can undoubtedly find value in dirt-cheap music streaming, but do they find the new form of music discovery *per se* valuable? The answer has important managerial implications and can help us better understand data-based innovations and business models. We develop a theoretical model of music discovery and consumption, and harness changes in Last.fm consumer offering to separate music acquisition from music discovery.

The analysis uses the amount music consumption as an indicator that a consumer finds the service worth using (and hence valuable) in a competitive market environment (Oestreicher-Singer and Zalmanson, 2013). The main dependent variable is thus not a direct measure of commercial success but an important prerequisite for generating revenues and increasing the valuation of online businesses (Brynjolfsson, Kim and Oh, 2013). We find evidence that the new form of music discovery is valuable to consumers, yet it is relatively less important than an opportunity to listen to music for free. Also, whether the value of music recommendations can be captured to support a viable business is a different matter. We will return to these matters in the discussion of findings, which call for more attention to the underlying mechanisms of value creation and capture for data-based business.

2 Music discovery through Last.fm

Last.fm is one of the oldest and most popular online music discovery services. The service collects music listening and social data from over 600 playback applications, services and devices to create personalized music recommendations. Since the inception of Last.fm in 2002 to early 2009, users could stream free music directly from the service. This undoubtedly contributed to the rapid growth seen in Figure 1 below. In April 2009, the company limited free streaming to the US, UK and Germany, citing inability to recover music licensing fees from advertising. Users in other countries were then required to pay a subscription fee for streaming and. Over the last five years, Last.fm has gradually wound down all streaming operations, focusing its business exclusively on music discovery. The service is based on continuously amassing user-generated music consumption data from the digital ecosystem, and carefully distilling them into personalized music recommendations that are supported by various social media features. The users are encouraged to stream music provided by partners such as Spotify and YouTube.

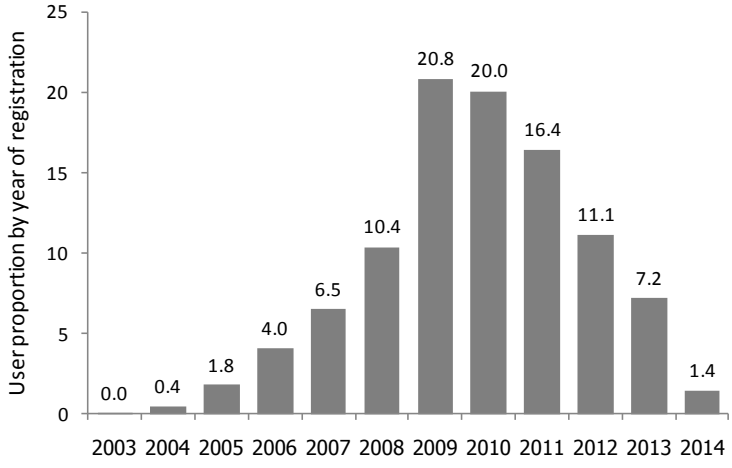

Figure 1. Last.fm User Growth (user proportion by the year of registration)

In the following brief literature review, we describe three factors that shape data-based music discovery and consumption through Last.fm. At the heart of the Last.fm there is a collaborative filtering rec-

ommender system (CFRS) that is a popular approach to providing product recommendations. Second, raw data that feeds the Last.fm CFRS is retrieved from hundreds of different sources and must be cleansed and amended with appropriate metadata so that the data becomes a reliable resource for computational processing. Finally, music discovery and consumption are highly social activities. The service allows users to interact around music items, artists and events, which can both enhance music discovery and make users more committed to the service.

2.1 Collaborative filtering recommender system

People reveal their music preferences by listening more often and repeatedly to music they like. Playback counts can therefore be used as ratings data for a collaborative filtering recommender system (Ekstrand et al. 2010). CFRS uses the data to construct a collective similarity network between music items, maps an individual user's preferences to the network, and, finally, produces recommendations regarding nearby products in the network. Anderson (2006) believes the approach could increase demand for niche products and there are some highly successful applications such as Amazon's "Customers Who Bought This Item Also Bought" feature that drive sales by automatically created recommendations. However, CFRS alone is not a panacea for finding relevant products. This may be due to the lack of suitable, high quality data but also relate to the specific nature of items to be recommended. Goldenberg, Oestreicher-Singer and Reichman (2012) point out that the fact that the underlying product network is constructed simply on similarities between items can be particularly problematic for music.

Music recommendations should be both novel and relevant (Celma and Lamere 2011). Recommending *Nine Inch Nails* to a *Prince* fan may be novel but probably not very relevant whereas *Michael Jackson* would be perhaps relevant but most likely redundant. There are different opinions on what kind of dynamics recommender systems stimulate in general and specifically in the context of music business. Oestreicher-Singer and Sundararajan (2012a) study of Amazon.com shows that the structure of a product network does not only reflect people's past purchases but it can also affect subsequent demand for products. Computational recommendations may not thus be a straightforward reflection of similarities between items in the product space but a part of complex, dynamic process that, at least to a degree, shapes what it is supposed to reflect. Celma and Cano (2008) find that the Last.fm similarity network suffers from a popularity bias and go on to argue that this may be an inherent problem associated with the use of social data to organize content (see also Hosanagar, Fleder, Lee and Buja 2013). In Last.fm, the play count of artists is strongly correlated with the play count of other similar artists. Popular artists are more likely to act as hubs within the similarity network, while less popular artist are less likely to be recommended even if discovering those long tail items could often be most valuable. On the other hand, Levy and Bosteels (2010) who are Last.fm employees defend the service against popularity bias criticism using an internal dataset.

2.2 Metadata infrastructure

The CFRS can only work if it can reliably identify two pieces of music content as the same or different items, and associate them with a correct description. It is important to bear in mind that the system does not operate directly on music content but analyses its metadata that is an important infrastructure for many operations in the industry (Jannach et al. 2011; Brookes, 2014a; 2014b). Metadata is a description of a resource. It informs about the structure and content of a bundle of data that may represent a song, photograph, database or any other digital artefact (Kallinikos, Aaltonen and Marton, 2013). Critical metadata used to be permanently printed on top of vinyls and CDs, whereas digital music files are much more loosely coupled with their metadata and may easily lose it. Without appropriate metadata, it is impossible to manage music content and its copyrights on a commercial scale, or even find anything from tens of millions songs available through online music services. At the moment, there is no authoritative, institutionally controlled source of music metadata and a lot of music circulates in the digital ecosystem with partial, inconsistent and simply incorrect metadata.

Humans can deal with small inconsistencies and errors in music metadata, but they pose considerable problems for computational processes that underpin CFRS and music business in general. Last.fm needs to be able to analyse music consumptions across a broad range of services and devices that source metadata from at least five different vendors all imposing their own metadata standards. Furthermore, peer-to-peer file sharing services allow user-generated ID3 metadata tags propagate throughout the digital ecosystem as the tags are not controlled by anybody (Morris, 2012; Brookes, 2014a; 2014b). Even more importantly, no metadata is useful unless it can be associated with the right content. A music identifier is a unique token that ties metadata to a music item, and allows identifying two pieces of music content as the same or different items. The lack of reliable identifiers makes it difficult to calculate play counts, compute recommendations and, in general, to ensure items are presented with the correct description of the music content. Last.fm relies mainly on the combination of artist name and song name as the identifier, but given the poor quality of existing metadata the approach is far from perfect. For instance, the company has found that there are over 100 ways to spell artist name – song name combination *Guns N' Roses – Knockin' on Haven's Door*.

2.3 Social media features

Music discovery and consumption are typically highly social activities. Last.fm users often begin as mere consumers of recommendations but may eventually start to participate more intensively, for example, by creating and organizing content, participating in discussions, and even become informal leaders in the community (Oestreicher-Singer and Zalmanson, 2013). This is important because socially engaged users have been found to be more likely to pay for Last.fm as well as other services (Fullerton, 2003; 2005; Oestreicher-Singer and Zalmanson, 2013), and once a user subscribes to a premium service, the likelihood that his or her friends subscribes increases (Bapna and Umyarov, 2012).

Social media engagement can mitigate the shortcomings of CFRS in providing valuable music recommendations. Recommendations that are based on a similarity network constructed from user data can be too successful in connecting similar products together and, arguably, biased toward popular items. This can easily render the output from CFRS less useful for the users. The integration of a similarity (product) network with a social network into a dual network approach can alleviate the problem. Users create idiosyncratic links to the similarity network as they participate in social media activities by posting comments, writing reviews, tagging content, etc. These links can be seen as their personal recommendations and ways of grouping products, which can complement similarity network and help users to discover more relevant items (Chen, Boring and Butz, 2010; Goldenberg, Oestriecher-Singer and Reichman, 2012).

3 Data collection and the dataset

We collect social consumption data from Last.fm to evaluate a theoretical model of music discovery and consumption. The data are mainly retrieved via the Last.fm Application Programming Interface (API) without personally identifying information. The only exception to this is the username that may sometimes represent the real identity of a sample user. Usernames are not included anywhere in the reported findings. The construction of a dataset for statistical analysis involves three steps: 1) identifying a representative sample of Last.fm users, 2) retrieving data for each user in the sample, and 3) assembling the dataset with variables that operationalize music discovery and consumption.

We apply a rejection sampling method proposed by Gjoka et al. (2010) to retrieve a representative sample of users. Each Last.fm user has a unique positive integer as his or her identifier. The identifiers are generally assigned so that a user who registers later will receive a larger number, and the entire user population should comfortably fall within a range of 1 and 100,000,000. We draw a random integer from the range and query Last.fm for data by using the number as the user identifier. We repeat the procedure storing raw data until we end up with a random sample of 12,839 users.

3.1 Dataset construction

We retrieve five types of data for each user in the sample and assemble them to a panel dataset that traces users through time along several variables. We divide the temporal dimension of the panel dataset into yearly increments, which allows us to separate the impact of changes to Last.fm consumer offering without breaking the dataset into too small subsamples. The dataset describes individual users with five main variables that allow us to unpack the impact of data analytics and social media features on music discovery and on music consumption.

Playcount measures the amount of music consumption. The variable is based on listening event data that represent the playback or streaming of individual songs. The data include the title, artist name and time for each song a user has listened. We simply count the number of annual listening events per user, which is the sole input to the variable.

Listening concentration measures the relative success of music discovery. Chen, Boring and Butz (2010) found that after a successful discovery there is often a burst of listening as the user keeps listening to music from the same artist for a period of time in Last.fm. Consequently, we assume that a more concentrated listening profile at the artist level signals more successful music discovery. We use the listening event data to compute a Herfindahl Index (HI) as an operational measure of concentration (Benkler, 2006; Kwoka 1985; Rhoades, 1993). Note that we also normalize our HI to ensure we can better use it to compare listening concentration of different users across time. HI is described in more detail in Statistical Appendix.

Friends measure social media engagement in Last.fm. We retrieve a friend list for each user, which represents social relationships that the user has actively acknowledged at the end of the observation period. We also retrieve the time each of public communication between the user and his or her friends. Using these two types of data, we construct a proxy variable that traces the number of friends at different points in time by assuming that the time at which each connection of friendship is established coincides with the time at which the users communicated for the first time in Last.fm. This should give a reasonably accurate, lower bound estimate of the number of friends at different times, since our panel dataset observes the temporal dimension only at the annual level. Although users can communicate without adding each other to the list of friends, by combining the two types of data we intend to increase the reliability of the measure and ensure that we capture positive emotional relationships within the user community (Chmiel et al., 2011).

Auto-corrections measure the quality of metadata that makes personalized music recommendations possible. Last.fm introduced in January 2009 a system that can automatically correct artist and song names, and therefore counter problems stemming from incorrect music identifiers circulating in the digital ecosystem. We retrieve all auto-correction mappings applied to the listening events of our sample users. The mappings consist of artist names submitted by users that are deemed incorrect, and the correct names to which they are mapped to. We construct a proxy variable to trace the number of corrections made to the listening data of each user over time. This is done by estimating the number of artist names that have been corrected for each user by comparing auto-correction mappings with the listening events of each user. We assume for certain names on auto-correction mapping to appear for the first time when those names appear on listening data of users for the first time. Since we can only retrieve listening events whose metadata has been already corrected, we do not know the original metadata that the user submitted to Last.fm.

Past listening similarity measures the utilization of data analytics. Ideally, we would like to observe actual personal recommendations produced by the CFRS through time. However, since no such data is easily available, our next best option is to analyze the similarity of current listening to past listening. This is because the logic of CFRS is to use product similarity network to recommend products similar to those that the user has ranked highly in the past. Therefore, we expect the listening events to be relatively more similar to the music which the user has listened previously if the user relies on the rec-

ommender system to discover new music. The approach has been previously implemented by Celma and Cano (2008), and Oestreicher-Singer and Sundararajan (2012b).

We compute proxy variable for past listening similarity. Since a similarity network between products is known to be usually relatively stable over time (Konstan and Riedl, 2012), we simply use a static network at the time of data retrieval retrospectively for all calculations. First, we pull a list of top 50 similar artists for each listening event in the sample. We then identify different artists that user has listened to previously in relation to each listening event. Finally, we count number of those different artists with the list of top 50 similar artists, whereby each listening event fall into, and average these values annually. This is done by computing the value at each listening event for a year and taking their arithmetic mean. Note that we calculate past listening similarity for users with more than 10,000 listening events by randomly selecting only 10,000 events for calculation. This improves computational speed, while the accuracy of the calculation is also ensured since it is based upon random selection.

Variable	Data	Concept
PLAYCOUNT	Listening events*	Playcount is the main dependent variable that measures the amount of music consumption.
LISTENING CONCENTRATION	Listening events*	Artist-level concentration of music consumption measures the success of music discovery.
FRIENDS	Friend list Time of communication between users*	The number of friends is a proxy for the use of social media features.
AUTO-CORRECTIONS	Auto-correction mappings Listening events*	The number of auto-corrections is a proxy for the quality of underlying metadata
PAST LISTENING SIMILARITY	Lists of 50 most similar artists Listening events*	Past listening similarity is a proxy for the use of data analytics.

*Table 1. Dataset Construction (time series data marked with *)*

Table 1 summarizes the data and concepts that are used to construct the five main variables for path analysis. Out of the five types of retrieved data, listening events and the time of communication between users are time series while auto-correction mappings, friend lists and the lists of 50 most similar artist represent the situation at the end of the observation period 30 June 2014. FRIENDS, AUTO-CORRECTIONS and PAST LISTENING SIMILARITY variables are therefore constructed as retrospective estimates in our panel dataset by computing proxies for them. Finally, users can opt to hide their listening event data and to turn the auto-correction feature off, but this in practice rare.

3.2 Descriptive statistics

In our sample of 12,839 Last.fm users, 57 per cent have submitted at least one listening event, 22 per cent have had their music metadata corrected by the auto-correction system, 9 per cent have at least one person in their friend list, and 5 per cent have communicated publicly with their friends through Last.fm. Listening data for the sample users amounts to 18,804,414 events that, by construction, is the sample users' total aggregate playcount. These listening events include 221,614 different artist names. For each of the artists, we retrieve the list of 50 most similar artists. Since some of the variables cannot be calculated for a user that has zero playcount, we have dropped such users during such time periods from the table and any further analyses. We also drop users who listen to only one artist, since normalized HI cannot be computed for them.

Variable	Mean	Median	Std. deviation
PLAYCOUNT	1659.6	101.0	6199.2
LISTENING CONCENTRATION	317.9	93.4	752.5
FRIENDS	1.2	0.0	6.8
AUTO-CORRECTIONS	13.6	1.0	40.9
PAST LISTENING SIMILARITY	7.3	5.1	7.2

Table 2. Descriptive Statistics for the Five Variables

Table 2 reports descriptive statistics for the five main variables in our dataset. The means for each variable are considerably higher than medians that suggest highly skewed distributions, which is common in social data (Shirky 2008). Hence, we transform the variables logarithmically and then compute a correlation matrix presented in Statistical Appendix. We find that the variables are significantly correlated and include a variance inflation factor (VIF) check for potential multicollinearity problems. Note that although mean and median of friends is relatively low (since only 5% of our users have publicly communicated with their friends), almost 20% of our data points register positive number of friends.

Year	PLAYCOUNT	LISTENING CONCENTRATION	AUTO-CORRECTIONS	PAST LISTENING SIMILARITY	FRIENDS
2005	3508.9	434.2	6.9	4.2	0.0
2006	2308.5	415.7	8.4	4.7	0.4
2007	2166.2	458.3	9.4	5.6	0.8
2008	1932.8	431.9	11.0	6.5	1.2
2009	1310.0	250.5	9.1	5.8	0.8
2010	1248.1	243.8	9.7	6.4	0.9
2011	1281.7	247.2	10.5	7.4	1.0
2012	2278.6	312.5	18.4	8.7	1.8
2013	2332.9	427.5	25.6	10.2	2.0
2014	1747.7	549.4	41.8	13.3	3.3

Table 3. Annual Means for the Five Variables in the Panel Dataset

Figure 1 shows that the number of sample users peaks in 2009 and declines thereafter, which matches the overall pattern of new users in Last.fm. The peak coincides with the changes to the Last.fm consumer offering indicating that these changes may have had a significant impact on Last.fm usage. Table 3 presents annual means for the five key variables. It is worth pointing out that the auto-correction variable gets positive values even before the system was activated in 2009. This is because we rely on a proxy variable and do not know the time when a particular correction was first applied to user-submitted metadata. We compensate for this problem in our main estimations by running our estimation twice, before and after 2009. Differences between the two estimations allow us to make inferences about the impact of auto-correction system.

4 Theoretical model

The empirical analysis consists of estimating two equations that capture together eight hypotheses on how Last.fm works. Five of these are captured in Figure 2 that shows a path diagram for a theoretical model of music discovery and consumption. The first equation (Model 1, note that we read the diagram from right to left) estimates factors that influence music consumption, while the second equation (Model 2) opens up music discovery. As the model itself is relatively straightforward mapping of causal relationships found in the previous literature, our main interest is on the changes before and after the major changes to the consumer offering in 2009.

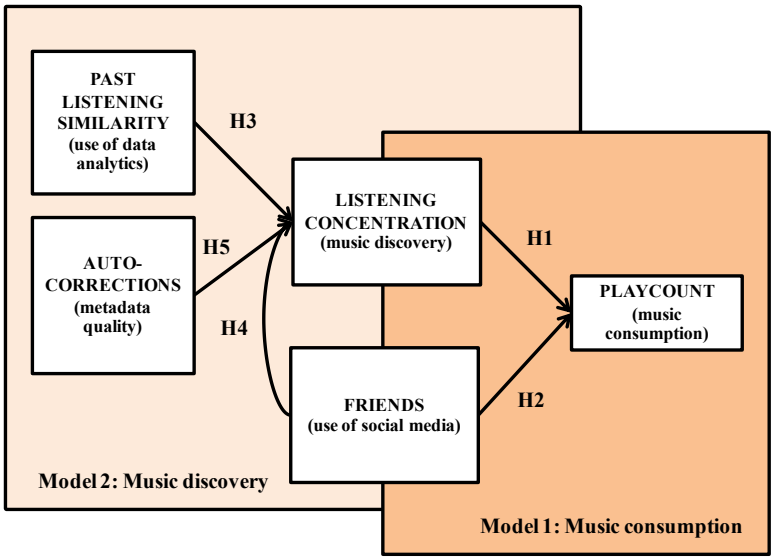

Figure 2. Path Diagram of Music Discovery and Consumption

We expect *ceteris paribus* that users consume more music if they are able to discover interesting artists and engage social media around music. This is captured in Model 1. Bateman et al. (2011) show that online participation is directly linked to commitment, as defined by organizational commitment theory. There are three types of commitment: continuance, affective and normative commitment.

H1: Better music discovery leads to more music consumption (continuance commitment)

H2: More intensive social media engagement leads to more music consumption (affective/normative commitment)

Model 2 opens up data-based operations underpinning music discovery in more detail. We expect users to be more successful in discovering music if they use the recommendations (Chen, Boring and Butz, 2010), engage in social media activities (Goldberger, Oestriecher-Singer and Reichamn, 2012), and their music items have correct metadata (Brookes, 2014a; 2014b). We assume the following hypothetical relationships in Model 2.

H3: More utilization of data analytics leads to better music discovery (dual network)

H4: More intensive social media engagement leads to better music discovery (dual network)

H5: Better metadata quality leads to better music discovery (metadata problem and information infrastructure)

Closing down music streaming from 2009 onward has had a major impact on Last.fm users, which may cast a direct negative impact not only on both music consumption but also, interestingly, on music discovery. Slowing growth and decline in users number as a result of closing down streaming operations 2009 onward may affect music discovery since Last.fm gets less timely data on new artists and songs. Therefore, we estimate the models separately for periods before and after 2009 to formulate two additional hypotheses that isolate the effect of business model change.

H6: The intercept term for Model 1 (music consumption) is lower after 2009 as compared to estimation before 2009 (changes of consumer offerings and continuance commitments)

H7: The intercept term for Model 2 (music discovery) is lower after 2009 as compared to estimation before 2009 (changes of consumer offerings and slower growth of social data)

Since the auto-correction variable gets positive values even before the system was activated in 2009, running the estimation separately before and after 2009 also allows us to better assess the impact of the auto-correction system. One of the main effects of the auto-correction is to lump data that was previously associated with different artists together, increasing, naturally, listening concentration. Because we only observe listening events that have already been mapped by the auto-correction system, effect of auto-correction variable should be positive and significant for music discovery even before 2009. This purely illustrates the lumping effect. We expect the positive effect associated with the estimation after 2009 to be even stronger. And, here, the incremental positive effect can be interpreted as the positive impact of metadata quality upon music discovery.

H8: Positive association between auto-correction variable and listening concentration is stronger for estimation after 2009 as compared to estimation before 2009

5 Findings from a path analysis

We conduct a path analysis by estimating the two models (music consumption and discovery) for two time periods (before 2009 vs. 2009 and onward) using simple ordinary least square estimation. For this purpose, we need to make a few additional adjustments to our panel dataset. First, we pool our panel dataset across time since, here, we run our estimation based on pooled panel data. Second, we include only data points with at least one listening event, since some of the variables can only be computed for users with a positive PLAYCOUNT value. Also, we drop data points, whereby users listen to only one artist since we cannot compute normalized HI for those data points. Third, we apply a logarithmic transformation to adjust the highly skewed distributions of the five main variables. After that, we apply ordinary least square estimation to estimate the following two equations for the two time periods.

(1) $\text{Log (PLAYCOUNT or music consumption)} = \alpha_0 + \alpha_1 \text{Log (LISTENING CONCENTRATION or music discovery)} + \alpha_2 \text{Log (FRIENDS or use of social media)}$

(2) $\text{Log(LISTENING CONCENTRATION or music discovery)} = \alpha_4 + \alpha_5 (\text{PAST LISTENING SIMILARITY or use of data analytics}) + \alpha_6 \text{Log (FRIENDS or use of social media)} + \alpha_7 \text{Log(AUTOCORRECTION or metadata quality)}$

Table 4 and Table 5 report the estimation result for the two models during the two periods. All coefficients show hypothesized signs and most of them are statistically significant at one per cent level (Sig ≤ 0.01). The tables compare the relative importance of different factors for music discovery and music consumption in Last.fm for the two time periods. Further, policy changes from 2009 onward have a significant negative effect on music consumption and, more interestingly, on music discovery, as reflected in changes to the intercept terms for estimations before and after 2009.

5.1 Music discovery (Model 2)

Coefficients[a]

Dummy_2009onward		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	T	Sig.	Tolerance	VIF
0	(Constant)	3.656	.080		45.938	0.000		
	log_analytics	.397	.051	.175	7.854	.000	.504	1.986
	log_dataquality	.095	.029	.077	3.330	.001	.467	2.142
	log_friends	.076	.041	.034	1.865	.062	.729	1.372
1	(Constant)	2.245	.050		45.329	0.000		
	log_analytics	.805	.030	.312	26.538	.000	.774	1.292
	log_dataquality	.328	.023	.178	14.247	.000	.685	1.459
	log_friends	.133	.054	.028	2.475	.013	.825	1.212

a. Dependent Variable: Log_HI_normalized

Table 4. The Impact of Data Analytics, Data Quality and Social Media Engagement on Music Discovery

Table 4 reports the impact of data analytics, data quality and social media engagement on music discovery. The first four rows (Dummy_2009onward = 0) present the estimation for 2002–2008, while the last four rows present the estimation for 2009–2014 (Dummy_2009onward = 1). The variance inflation factor (VIF) values in the two last columns show that the findings are not subject to multicollinearity problems. We also test if the observed differences between the time periods are statistically significant.

We find that the use of music recommendations and data quality have considerable effects on music discovery, while social media engagement has only a weak effect. Most importantly, successful music discovery is expected to increase by 0.397 per cent (before 2009) and 0.805 per cent (2009 and after) against 1 per cent increase in the use of music recommendations. This shows that consumers find data-based recommendations useful. Also, we find that the difference between the time periods is statistically significant, which means that the company is able to make its recommender system progressively more effective. Furthermore, music discovery is expected to increase by 0.095 per cent (before 2009) and 0.328 (2009 and after) against 1 per cent increase in the data quality. The difference between the time periods is statistically significant and results most probably from the activation of auto-correction system in 2009. At the same time, social media engagement has a very weak effect on music discovery as the latter is expected to increase only by 0.076 per cent against 1 per cent increase in social media engagement (the difference between time periods was found statistically insignificant). On the other hand, changes to consumer offerings also cast significant influences upon music discovery as the intercept term of the estimation for the 2009 and after time period is much lower than that for the before 2009 time period. The difference is statistically significant and indicates that music discovery is expected to decrease by as much as 75.6% because of the policy changes. The dependent variable of Model 2, music discovery, enters into Model 1 as an explanatory variable. The variables in the equation may therefore cast an indirect effect on music consumption.

5.2 Music consumption (Model 1)

Coefficients[a]

Dummy_2009onward		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	T	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
0	(Constant)	3.639	.087		41.992	0.000		
	Log_HI_normalized	.377	.017	.308	22.044	.000	.977	1.023
	Log_friends	1.052	.038	.388	27.730	.000	.977	1.023
1	(Constant)	2.375	.040		59.422	0.000		
	Log_HI_normalized	.465	.009	.477	51.338	0.000	.967	1.034
	Log_friends	1.360	.043	.297	31.957	.000	.967	1.034

a. Dependent Variable: log_playcount_annual

Table 5. The Impact of Music Discovery and Social Media Engagement on Music Consumption

Table 5 reports the impact of music discovery and social media engagement on music consumption. The specification of the estimation is nearly identical to Model 2 above, but since social media engagement is an independent variable for our Model 2, it also has an additional indirect effect on music consumption through music discovery. The indirect effect is, however, relatively small. Again, the first three rows (Dummy_2009onward = 0) present the estimation for 2002–2008, while the last three rows present the estimation for 2009–2014 (Dummy_2009onward = 1). The variance inflation factor (VIF) show that the findings are not subject to multicollinearity problems, and we also test if the observed differences between the time periods are statistically significant.

In contrast to music discovery in Model 2, we find that social media engagement has a considerable impact on music consumption. Music consumption is expected to increase by 1.052 per cent (before 2009) and 1.360 per cent (2009 and after) against 1 per cent increase in social media engagement. Also, music discovery has a strong impact on music consumption that is expected to increase by 0.377 per cent (before 2009) and 0.465 per cent (2009 and after) against 1 per cent improvement in music discovery. The differences between time periods are statistically significant for both independent variables, which allows further interpretation. This indicates that Last.fm use has become increasingly focused on data-based music discovery that provides clear value to consumers, albeit with small indirect support from social media engagement. At the same time, social media features remain very important for user retention.

On the other hand, policy changes regarding changes to consumer offerings also cast significant influences upon music consumption as intercept term of the estimation for the 2009 and after time period is much lower than that for the before 2009 time period. The difference is statistically significant and it translates to the direct negative impact of as much as 71.7%. Further, policy changes cast indirect negative impact upon music consumption through music discovery to reduce music consumption by further 35.2% (75.6% x 0.465). Henceforth the total effect of policy changes is to reduce music consumption by a whopping 81.3% $[1 - (1 - 71.7\%) \times (1 - 35.2\%)]$

Finally, Table 6 summarizes direct and indirect impact of data analytics, data quality, social media engagement and policy changes related to consumer offerings upon music consumption. It demonstrates that indirect effect of data analytics, metadata quality and social media via music discovery upon music consumption is relatively small. Although the direct impact of use of social media upon music consumption is relatively large, such impact is still relatively small as compared to that related to

policy changes. Henceforth, arguably, although new form of music discovery is valuable to consumers, the value is relatively modest compared to music acquisition, that is, music streaming.

	Before 2009			2009 and after		
	Direct	**Indirect**	**Total**	**Direct**	**Indirect**	**Total**
Increase in use of data analytics by 1%	n/a	+0.15%	+0.15%	n/a	+0.30%	+0.30%
Increase in metadata quality by 1%	n/a	+0.04%	+0.04%	n/a	+0.12%	+0.12%
Increase in use of social media by 1%	+1.05%	+0.03%	+1.08%	+1.36%	+0.05%	+1.41%
Changes to consumer offerings	n/a	n/a	n/a	-71.7%	-35.2%	-81.3%

Table 6. The Direct and Indirect Impact of Data Analytics, Data Quality, Social Media Engagement, and Policy Changes upon Music Consumption

6 Discussion and conclusions

We find evidence that the new form of music discovery and social media features are valuable to Last.fm users. However, value created by such operations need to be understood in context. The declining number of active users since 2009 suggests that the overall consumer value created by such operations is relatively modest compared to an opportunity to listen to music for free. Also, the value of data-based music discovery may not be perceived equally by all consumers but is likely more relevant to a specific type of music listeners. For instance, the Phoenix 2 UK project found that the proportion of music listeners who are enthusiasts is relatively small (Jennings 2007).

The findings raise questions whether big data supporting the venture can alone generate enough competitive advantage to sustain the business. In 2013, Last.fm made 2.1 million GBP loss, its revenues plummeted by 20 per cent, and the number of employees was halved (Sweney, 2014). Together with our finding that Last.fm depends heavily on social media features to retain its users beyond 2009, these observations call attention to key assumptions underpinning data-based music discovery business and, as we will elaborate below, big data innovations in general.

The new form of music discovery may well serve the needs of particular music enthusiasts whose music consumption is indeed limited by difficulties in findings interesting new music. Yet, for the majority of consumers this is probably not a major issue. Many people prefer to listen to *popular* music, that is, the very opposite of the long tail items. The concentration of music consumption on a relatively few popular items can look like a problem to some but it is also a testament to the social nature of music consumption. Popular music functions as a platform for socializing and makes it possible to share common experiences. Against this background, it is not surprising that significant amount of consumer value in Last.fm would seem to emerge from the use of its social features. This makes declining user numbers particularly problematic.

To an extent that the consumer value of Last.fm is created by social network externalities, loss of users numbers can perpetuate itself unless the service is able counter the loss of network externalities with increasingly successful music discovery. We find that music discovery has improved significantly over the years as Last.fm has enhanced its recommender engine and released new features such as the auto-correction system. At the same time, however, our findings show that the declining user base can also have a direct negative effect on digital music discovery. There are two reasons for this. First, it is less likely that the dual network is able to mitigate the problems of collaborative recommender filtering system if there are less people on the platform. Second, the product space expands continuously with new music items that are new to all users. The less users there are submitting data, the longer it takes to capture enough ratings to incorporate new items. More generally, the importance of network externalities in social media is a well-known topic, and our study shows that the relative size of user base can also matter for the value of big data as a resource for product/service innovation.

Our analysis does not allow pinning down a causal model of data-based music discovery business, but it certainly opens up complexities involved in creating a big data business in a specific domain. These involve consumption patterns and the product space of a particular industry, the nature of analytical problem and its applicability to computational processing, and the role of social media and social data as a part of big data operations. In the case of music industry, the new form of data-based music discovery is valuable to some consumers and hence potentially a source of competitive advantage. At the same time, it may require sourcing data from a broader population to generate good recommendations. This leaves open a question, what is the benefit for those consumers?

Discussion about data-based businesses can become highly technical (e.g. Celma, 2008). Technical analyses are important and often insightful, but at the same time they may overlook other factors that are crucial for the successful operation of these businesses. Social media features (Oestriecher-Singer and Zalmanson, 2013; Goldenberg, Oestreicher-Singer and Reichman, 2012), industrial metadata (Brookes, 2014a; 2014b) and the nature of recommender systems are all important (Celma, 2008), but it is their interplay in a specific field of consumption that a company needs to understand if it is to reap sustained competitive advantage from products/services based on big data.

Statistical Appendix

Herfindahl Index

Herfindahl Index (HI) is commonly used by competition economists to measure market concentration in mass media and in other types of markets (Benkler, 2006; Kwoka 1985; Rhoades, 1993). Here, we compute listening concentration of each user in accordance to HI and the formula is as followed.

$$HI = \sum_{i=1}^{n} \left(\frac{y_i}{\sum_{i=1}^{n} y_i} \right)^2 \times 10000, \text{ where}$$

y_i Total number of track of music associated with a particular artist within a particular year for a user

n Total number of different artist whom the user listen to within a particular year

$i \in \{1,2,3,\ldots,n\}$

Unfortunately, the problem with HI is that its lower bound depends on the number of different artist whom the user listen to. Since this varies considerably between the users in our sample, we use a normalized version of HI that ranges between 0 (extremely diverse) and 10,000 (extremely concentrated) regardless of the number of different artists the user has listened to. More precisely, it can be shown that HI would range between $1/n \times 10000$ and 10000 by its construct. Henceforth, sometimes the index is normalized with the following formula:

$$Normalized\ HI = \frac{HI - (1/n \times 10000)}{10000 - (1/n \times 10000)}$$

Correlation Matrix

	PLAY COUNT	LISTENING CONCENTRATION	AUTO-CORRECTIONS	PAST LISTENING SIMILARITY	FRIENDS
PLAYCOUNT	1	.500**	.728**	.721**	.449**
LISTENING CONCENTRATION	.500**	1	.331**	.377**	.198**
AUTO-CORRECTIONS	.728**	.331**	1	.576**	.514**
PAST LISTENING SIMILARITY	.721**	.377**	.576**	1	.362**
FRIENDS	.449**	.198**	.514**	.362**	1

** *Correlation is significant at the 0.01 level*

References

Aaltonen, A., and Tempini, N. (2014). *"Everything Counts in Large Amounts: A Critical Realist Case Study in Data-based Production" Journal of Information technology* 29 (1), 97-110.

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More.* 1st Edition. New York: Hyperion.

Bapna, R., and A. Umyarov (2012). *Are Paid Subscriptions on Music Social Networks Contagious? A Randomized Field Experiment.* SOBACO Working Paper. Carlson School of Management, University of Minnesota.

Bateman, P. J., Gray, P. H. and B. S. Butler (2011). "The Impact of Community Commitment on Participation in Online Communities." *Information Systems Research* 22 (4), 841-854.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* 1st Edition. New Haven: Yale University Press.

Brookes, T. (2014a). "Descriptive Metadata in the Music Industry: Why It Is Broken and How to Fix it – Part One." *Journal of Digital Media Management* 2 (3), 263-282.

Brookes, T. (2014b). "Descriptive Metadata in the Music Industry: Why It Is Broken and How to Fix it – Part Two." *Journal of Digital Media Management* 2 (4), 359-374.

Brynjolfsson, E., Kim, S. T. and J. Oh (2013). "User Investment and Firm Value: Case of Internet Firms." In: *Workshop for Information Systems and Economics (WISE) 2013*

Celma, O. (2008). "Music Recommendation and Discovery in the Long Tail." PhD thesis. Universitat Pompeu Fabra.

Celma, O. and P. Cano (2008). "From Hits to Niches? Or How Popular Artists Can Bias Music Recommendation and Discovery." In: *2nd on Large-Scale Recommender Systems and the Netflix Prize Competition (ACM KDD).*

Celma, O. and P. Lamere (2011). "If You Like Radiohead, You Might Like This Article." *AI Magazine* 32 (3), 57-66

Chen, Y., Boring, S. and A. Butz (2010). "How Last.fm Illustrates the Musical World: User Behaviour and Relevant User-Generated Content." In: *International Workshop on Visual Interfaces to the Social and Semantic Web (VISSW) 2010.*

Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., and J. A. Holyst (2011). "Negative Emotions Boost Users Activity at BBC Forum." *Physica A: Statistical Mechanics and its Application* 390 (16), 2936-2944.

Ekstrand, M. D., Riedl, J. T. And Konstan, J. A. (2010). "Collaborative Filtering Recommender Systems." *Foundations and Trends® in Human-Computer Interaction* 4 (2), 81-173.

Fullerton, G. (2003). "When Does Commitment Lead to Loyalty?" *Journal of Service Research* 5 (4), 333-344.

Fullerton, G. (2005). "The Service Quality – Loyalty Relationship in Retail Services: Does Commitment Matter?" *Journal of Retailing and Consumer Services* 12 (2), 99-111.

Gjoka, M., Butts, C. T., Kurant, M. and A. Markopoulou (2010). "Multigraph Sampling of Online Social Networks." *IEEE Journal of Selected Areas in Communications* 29 (9), 1893-1905.

Goldenberg, J., Oestreicher-Singer, G. and S. Reichman (2012). "The Quest of Content: How User Generated Links Can Facilitate Online Exploration." *Journal of Marketing Research* 49 (4), 452-468

Hosanagar, K., Fleder, D., Lee D. and A. Buja (2013). "Will the Global Village Fracture into Tribes? Recommender Systems and their Effects on Consumer Fragmentation." *Management Science* 60 (4), 805-823.

Jannach, D., Zanker, M., Felfernig, A. and G. Friedrich (2011). *Recommender Systems: An Introduction.* 1st Edition, Cambridge: Cambridge University Press.

Jennings, D. (2007) *Net, Blogs and Rock 'n' Roll: How Digital Discovery and What it Means for Consumers.* 1st Edition, London: Nicholas Brealey Publishing.

Kallinikos, J., Aaltonen, A. and A. Marton (2013). "The Ambivalent Ontology of Digital Artifacts." *MIS Quarterly* 37 (2), 357-370.

Konstan, J. A., and J. Riedl (2012). *Deconstructing Recommender Systems.* URL: http://spectrum.ieee.org/computing/software/deconstructing-recommender-systems (visited on 11/21/2014).

Kwoka, J. E. (1985). "The Herfindahl Index in Theory and Practice." *The Antitrust Bulletin* 30, 915-947

Levy, M. and K. Bosteels (2010). "Music Recommendation and the Long Tail." In: *Workshop on Music Recommendation and Discovery (Womrad)* 2010.

Morris, J. W. (2012). "Making Music Behave: Metadata and the Digital Music Commodity." *New Media & Society* 14 (5), 850-866.

Morville, P. and L. Rosenfeld (2006). *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites.* 3rd Edition, California: O'Reilly Media.

Oestreicher-Singer, G. and L. Zalmanson (2013). "Content or Community? A Digital Business Strategy for Content Providers in the Social Age." *MIS Quarterly* 37 (2), 591-616.

Oestreicher-Singer, G. and A. Sundararajan (2012a). "The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets." *Management Science* 58 (11), 1963-1981.

Oestreicher-Singer, G. and A. Sundararajan (2012b). "Recommendation Networks and the Long Tail of Electronic Commerce." *MIS Quarterly* 36 (1), 65-83.

Rhoades, S. A. (1993). "The Herfindahl-Hirschman Index." *Federal Reserve Bulletin* 79 (3), 188-189.

Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations.* 1st Edition. New York: Penguin Books.

Sweney, M. (2014). *Last.fm made loss of £2.1m last year.* URL: http://www.theguardian.com/media/2014/oct/08/last-fm-made-loss (visited on 11/21/2014).