Information Systems Research
Vol. 23, No. 3, Part 2 of 2, September 2012, pp. 1056–1067
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.1110.0405

Research Note

Music Blogging, Online Sampling, and the Long Tail

Sanjeev Dewan
Paul Merage School of Business, University of California, Irvine, Irvine, California 92697,
sdewan@uci.edu
Jui Ramaprasad
McGill University, Montreal, Québec H3A 1G5, Canada, jui.ramaprasad@mcgill.ca

Online social media such as blogs are transforming how consumers make consumption decisions, and the music industry is at the forefront of this revolution. Based on data from a leading music blog aggregator, we analyze the relationship between music blogging and full-track sampling, drawing on theories of online social interaction. Our results suggest that intensity of music sampling is positively associated with the popularity of a blog among previous consumers and that this association is stronger in the tail than in the body of music sales distribution. At the same time, the incremental effect of music popularity on sampling is also stronger in the tail relative to the body. In the last part of the paper, we discuss the implications of our results for music sales and potential long-tailing of music sampling and sales. Put together, our analysis sheds new light on how social media are reshaping music sharing and consumption.

Key words: blogs; social interactions; observational learning; word of mouth; long tail; music industry; social media
History: Vallabh Sambmurthy, Senior Editor; Siva Viswanathan, Associate Editor. This paper was received on March 27, 2008, and was with the authors 27 months for 4 revisions. Published online in *Articles in Advance* February 13, 2012.

> The debut Arcade Fire album, "Funeral," was released barely a month ago, on Sept. 14, by the indie label Merge, based in North Carolina. Enthusiastic reviews were written, even more enthusiastic blog entries were posted, MP3's circulated. It used to take months of touring and record-shop hype for an underground band to build a cult, but now it takes only a few weeks. "I'd like to thank the Internet," Mr. Butler said, and he wasn't serious, but he also wasn't wrong.
>
> (Sanneh 2004)

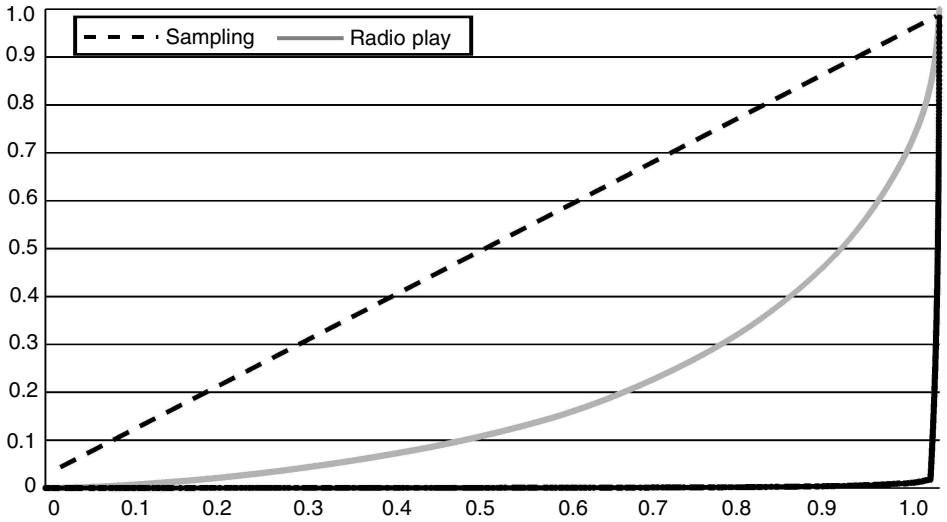
1. Introduction

Blogs and other social media are changing how consumers interact with each other, how they make decisions about consumption, and even how they actually consume products and services. These new media are particularly influential in the music world, as evidenced by the success of the Arcade Fire album *Funeral* (see the quote above). Traditionally, radio play has served as the primary mechanism for consumers to discover music before deciding whether to buy it or not. More recently, music blogs are emerging as an alternative new media competing with old media for consumer attention. The key difference is that whereas the old media have limited bandwidth and tend to focus on the most popular mainstream music, social media such as music blogs have a substantially higher bandwidth, bringing a far wider cross section of music to the attention of consumers. The purpose of this research is to examine how social interactions enabled by music blogs are shaping music sharing, sampling, and sales.

The stark contrast in bandwidth between new and old media is demonstrated in Figure 1, which displays the Lorenz curves for song radio play versus music blog sampling for the data set used in our empirical analysis.[1] It can be seen that the top 1% of songs on radio account for a full 50% of radio play, whereas the top 10% of songs consume more than 90% of radio time; the corresponding numbers for music blog

[1] Song radio play data are from Nielsen SoundScan, and music blog sampling data are from The Hype Machine. The Lorenz curves are constructed for the set of data that comprises the entire set of songs posted on The Hype Machine between July 1, 2006, and August 31, 2006 (for the sampling curve) and the set of songs played on the radio at least once during the first week of August (for the radio play curve).

Figure 1 Lorenz Curves for Radio Play vs. Music Blog Sampling



Note. The Gini coefficient, calculated as the proportion of the area between the 45° line and the Lorenz curve to the total area under the 45° line, is equal to 0.61 for music blog sampling and 0.93 for radio play.

sampling for the top 1% and top 10% of songs are only 8% and 40%, respectively.[2] As the use of social media expands alongside radio play, a key question is whether or not the exposure to a larger variety translates to a more diverse consumption of music by consumers.

Apart from their higher bandwidth, new media in the form of music blogs also enable consumers to immediately sample the music online—a type of consumption enabled by the "information good" nature of music. In this paper, we define sampling to be the streaming of full-track (as opposed to short clips) MP3 files posted on the blogs in a browser or music player. Despite the ease of music consumption through online sampling, the enormous variety of music available online[3] raises the question, how do consumers decide what music to sample? Any individual user is unlikely to be familiar with more than a tiny fraction of all those songs available. Given this, how does a user go about deciding which blogs to seek out for music recommendations and what songs to actually listen to? In this regard, with the advent of social media, consumers are increasingly relying on the opinions and actions of other consumers.

Music blogging and online sharing influences the choices of other consumers through two broad types of social interaction: word of mouth (WOM) and observational learning (OL). As discussed in Chen et al. (2010), WOM effects are meant to capture the impact of consumer *opinions* (such as product reviews) on other consumers' choices, whereas OL theories deal with the influence of consumer *actions* (such as frequency of purchase of different products). Word-of-mouth effects (see, e.g., Godes and Mayzlin 2004, 2009; Dellarocas 2003) are relatively less salient in our context because our data do not capture variation in consumer opinion—there are single blog posts per song, each of which typically signals a tacitly positive opinion about the music. Therefore, the thrust of our analysis is based on observational theories of learning as they apply to the role of blog/music popularity.

Understanding the role of popularity in consumer choice is the main goal of observational theories of learning. In the music blogging context, observational learning occurs when consumers draw inferences about music quality and likelihood of liking a piece of music from the past choices of other consumers as reflected in various popularity statistics (e.g., number of users who liked different songs or a list of most popular blogs). The initial focus of the OL literature was on developing analytical models (e.g., Banerjee 1992, Bikhchandani et al. 1998) to explain how consumers infer uncertain product quality from the choices of previous consumers. More recently, a distinctly empirical stream of the literature has emerged, and a common finding is that the release of popularity information results in popular products becoming even more popular, often leading to winner-take-all outcomes in product markets (e.g., Anderson and Holt 1997, Chen et al. 2010, Salganik et al. 2006). In contrast, Tucker and Zhang (2009) demonstrate that in certain settings the impact of popularity information might depend on the inherent market size of different products, so popularity information about a niche product might signal

[2] The Gini coefficients for *Sampling* and *RadioPlay* are 0.53 and 0.93, respectively, also indicating a substantially higher level of inequality in radio play compared with sampling.

[3] For example, The Hype Machine alone provides links to thousands of music blogs and through them to hundreds of thousands of songs.

high quality, further driving sales, whereas popularity information for a popular product is comparatively less informative in this regard—an insight that informs our analysis as well. Our theoretical discussions and empirical predictions draw on OL theories, emphasizing the differential impacts of blog and music popularity in mainstream versus niche music.

Our research also contributes to the emerging literature on the online long tail effect (Anderson 2004, 2006), where most of the prior work has focused on the changing distribution of book sales (e.g., Brynjolfsson et al. 2006). However, the idea of the long tail can also be applied to music consumption, as in the analysis of Bhattacharjee et al. (2007) and Chellappa et al. (2007). Given that music is a digital good that can be consumed immediately, we are able to explore how social media might contribute to the long tail in sales through first driving the long tail in sampling. Similar to the focus of these prior studies, we look at the potentially differential impact of blogging on the consumption of music in the body versus tail of music sales.

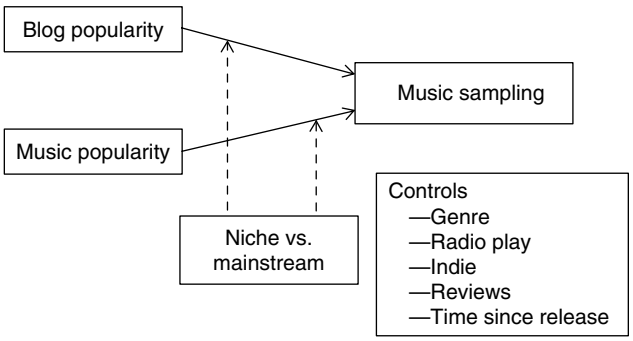
Highlighting our results, we find that blog popularity has a strong and positive effect on music sampling and that this effect is significantly stronger in the tail as compared to the body of music sales distribution. We also find evidence that music sampling is positively associated with the inherent popularity of the music item, and music popularity is also more important in the tail for driving sampling compared with the body of the song sales distribution. Finally, we find that sampling is positively associated with music sales. These results indicate that consumer blogging has an important influence on both music sampling and music sales, and the influence is different in the body versus tail of the sales distribution. These results have important implications for consumers and various participants of the music industry, as discussed later in the paper.

The rest of this paper proceeds as follows. Section 2 provides the theoretical underpinnings for our analysis and develops our hypotheses. Section 3 presents our data and empirical specifications. Section 4 presents our empirical results, and §5 provides some discussion and concluding remarks.

2. Theory and Hypotheses

In this section we develop our hypotheses, based on the conceptual framework shown in Figure 2. As shown in the figure, our primary interest is on the association between blog/music popularity and music sampling and on how this association is different for mainstream versus niche music (i.e., the long tail effect). We first look at the association between blog/music popularity and sampling (Hypotheses 1A and 1B) and then at how that association differs

Figure 2 Conceptual Framework



between the body and tail of music sales distribution (Hypotheses 2A and 2B). Our analysis is guided by the notion that music is inherently an *experience* good, in that its properties cannot be determined by inspection prior to consumption, as opposed to *search* goods, where this assessment can be done ex ante (see Nelson 1970 for the earliest distinction between search and experience goods). As such, consumers rely on the opinions and actions of their peers in choosing what music to consume. Now, one could argue that online sharing is turning all music into a search good because one can sample music prior to purchase. But the question relevant to this study is what music to sample in the first place, which is not trivial, given the enormous variety of both blogs and music.

As discussed in §1, sampling decisions might be guided by blog or music popularity, and the choice can be explained by theories of observational learning. The gist of the argument for blog popularity follows; the arguments for music popularity are analogous. In the presence of uncertainty about blog quality, or source credibility (Kelman 1961), consumers have to balance their own ex ante quality information with the inferences they draw from the prior consumption choices of their peers. Assuming that some fraction of the consumer population is able to discern blog quality, users would infer that a more popular blog is on average of higher quality than a less popular blog. In the extreme, when this correspondence between popularity and quality is made irrespective of the users' own private information, we have the so-called information cascade phenomenon (e.g., Banerjee 1992, Bikhchandani et al. 1992, Tucker and Zhang 2009). In general, OL theories predict that in the presence of quality uncertainty, consumers draw actionable inferences from popularity information. This effect applies to inferences about both music and blog quality, leading to our first hypothesis.

Hypothesis 1A (H1A). *Music sampling is positively associated with both blog and music popularity.*

We turn now to the dynamics of social influence and examine how the effect of music/blog popularity

changes over time. When a song first comes out, there is little word-of-mouth information available, quality uncertainty is high, and consumer choice is driven by popularity information alone. This setting is conducive to the formation of information cascades, where other individuals' actions (as summarized by popularity measures) are likely to be the primary driver of sampling behavior. Over time, consumers get exposed to other sources of information and word of mouth, which tends to break up any information cascades. In this sense, the longer a song has been out, the influence of popularity is correspondingly less important. That is, over time consumers rely less on observational learning as a way of discovering music or inferring its quality—consistent with the substitute information argument of Chen et al. (2010). We therefore hypothesize that

Hypothesis 1B (H1B). *The association between sampling and both blog and music popularity is stronger for newly released songs compared with previously released songs*.

We turn now to the differential effects of blog popularity for mainstream versus niche music. Using the search versus experience good classification discussed above, we note that mainstream music is more like a search good, whereas niche music is closer in nature to experience goods. This is because consumers might already be aware of mainstream music through its play on the radio and other mainstream media, but they are less likely to have had prior exposure to niche music. Blogs catering more to niche music are likely to have proportionately less following, so there would be a higher level of quality uncertainty in the mind of the average user. Therefore, we predict that blog popularity is a more informative proxy for the quality of blogs catering to niche music as compared to those offering mainstream music.

This argument is also supported by the literature on word-of-mouth effects. Specifically, Bearden and Etzel (1982) and Senecal and Nantel (2004) make a clear case that personal influence is more important for experience goods. In our context, this means that influence associated with popular blogs is more important for niche music sampling compared with mainstream music sampling, leading to the following hypothesis.

Hypothesis 2A (H2A). *The association of sampling with blog popularity is stronger in the tail compared with the body of song sales distribution*.

Finally, we discuss the differential effects of music popularity for mainstream versus niche music. We would expect niche music to have greater quality uncertainty in the mind of the user compared with mainstream music. Accordingly, we expect that music popularity, at the margin, is a relatively more important signal of quality for niche music. Also, in the conceptual framework of King and Balasubramanian (1994), "other-based preference formation" (relying on the choices of others for making one's own choices) is more salient for experience goods compared with search goods.

On the other hand, it has been argued in the word-of-mouth literature that informational influence is less important for high-preference heterogeneity services (Price et al. 1989): if everybody has different taste, then knowing what another consumer has chosen does not influence one's own choice. To the extent that preferences for niche music are more heterogeneous, we should expect peer influence to be less important for niche music as compared with mainstream music. This effect works in the opposite direction to the OL argument above, and so it is an empirical question as to which effect is dominant. For the sake of hypothesis testing, we posit the following.

Hypothesis 2B (H2B). *The association of sampling with music popularity is stronger in the tail compared with the body of the sales distribution*.

3. Data and Empirical Specification

3.1. Data

We combined data from three major sources: The Hype Machine (a music blog aggregator), Amazon.com, and Nielsen SoundScan. The Hype Machine (THM)[4] is one of the leading music blog aggregators, tracking MP3s that are posted on more than 1,200 music blogs (at the time we received our data) and posting them on THM's website. When THM posts the songs on its site, it adds a "listen" link that allows users to listen to, but not download, the entire song that was posted on the corresponding blog post. The THM posting also provides basic information about the song, including the track title and artist, as well as links to the track on iTunes, the corresponding album on Amazon, and back to the original blog post. THM has provided us with data for the full set of songs that were posted between July 1, 2006, and August 31, 2006. For each of these songs, THM has provided the total click-throughs on the "listen" link, the "Amazon" link, the "iTunes" link, and the "post" link from the date the song was posted until the date we received the data, November 14, 2006. THM also provided the basic description of the songs that are posted on the various music blogs (i.e., track title and artist), the date the song was posted on the music blog and the date it was saved to THM, the Technorati score, and number of del.icio.us bookmarks for the blog.

From the links that THM provided to Amazon.com pages, for the album corresponding to each track

[4] http://www.hypem.com.

posted on THM, we were able to extract the Amazon Standard Identification Number (ASIN) and obtain the following data as of November 15, 2006: each album's Amazon sales rank, average value of all the customer reviews, number of customer reviews, the value of the last 100 reviews (from which we calculated the standard deviation of the customer reviews), release date, and record label data.[5] We categorized each album by record label based on whether it is "independent" or not, determined by whether the label was part of the Recording Industry Association of America (one of the generally accepted methods of classifying the labels), and categorized the albums by the corresponding genre as listed on allmusic.com.

We were also able to obtain radio play and sales data for both albums and songs from Nielsen SoundScan. We obtained these data by taking a random sample of 2,500 song titles from our complete data set, where Nielsen was able to match approximately 1,800 songs with data from its database. Specifically, Nielsen provided us the weekly nationwide "spins," i.e., the number of times the song had been played on the radio anywhere in the United States. In addition, for all of the albums corresponding to the tracks in our data set, we obtained weekly total album unit sales data from Nielsen SoundScan from the date of release of the album until April 2007. Nielsen SoundScan sales data compile both online and off-line album sales, and Nielsen is the data source used by Billboard music charts. We also obtained song-level sales data for a subset of the tracks in our THM data set. We were told by Nielsen that song sales is primarily composed of digital downloads, and therefore song sales ought to correlate well with the online social interactions that we study.

Finally, we obtained information from Billboard charts on artists associated with the songs in our data set. Specifically, we extracted data from the "Billboard Top 100 Artists of the Year" for the years 2002 through 2006 and the "Billboard Hot 100 All-Time Top Artists," which is a list of the top-selling artists since 1958; this information was used to measure artist reputation in our data set.

After combining the data from each of these data sources, we constructed a cross-sectional data set for subset of songs posted on music blogs between July 1, 2006, and August 31, 2006, with total sampling measured on November 14, 2006; total radio play and song sales measured at the end of the week corresponding to November 14, 2006; and the Amazon.com data measured on November 15, 2006. Our final sample is drawn from 281 unique blogs, comprising songs from 1,088 unique albums in 24 genres. This sample has both song-level data (the listen and click-through data from THM and the song sales and radio play data from Nielsen SoundScan) and album-level data (the customer review and rank data from Amazon and album-level sales from Nielsen SoundScan).

The first two columns of Table 1 provide the labels and descriptions of the variables we use in this study. These definitions are self-descriptive, but a few key variables warrant additional explanation. *Sampling* is the click-throughs to the "listen" link on THM. *BlogPop* is the total number of del.icio.us bookmarks for a given blog from a given blog, and *MusicPop* is measured by the total number of track-level sales of a given song. *RadioPlay* is the total number of times a song has been played on the radio. *AlbumSales* are the sales of the album that correspond to the track that was posted on THM. *MusicPop, RadioPlay*, and *AlbumSales* are each cumulative measures, measuring the total of each activity during the time period we are studying (July 1, 2006, until November 14, 2006). *DaysRel* is the number of days since the release of the album; *DaysPost* is the number of days since the track was posted on THM. *RecentRel* is a dummy variable indicating whether the song was posted within 10 weeks of its release, to control for potential bursts in sampling activity at the time of album release, where 10 weeks was chosen as an appropriate timeframe based on previous research indicating that albums remain on the Billboard charts for an average of 10 weeks after release (Bhattacharjee et al. 2007). *RevNum*, *RevVal*, and *RevStdDev* are the number, average valence, and standard deviation of customer reviews posted on Amazon.com for the album corresponding to the track, respectively. *ArtistRep* is a dummy variable, defined as 1 if the song's artist was on the Billboard Top 100 Artists of the Year between 2002 and 2006 or if the artist was on the Hot 100 All-Time Top Artists, which are the top-selling artists since 1958, and 0 otherwise.

To address the research questions we are interested in, we first need to define the long tail in music. Previous research examining the long tail for book sales defines the tail as the sales in the offerings outside the capacity of a typical brick-and-mortar store (Brynjolfsson et al. 2006, Anderson 2006). Although this number has been established in studies done on books (100,000), a concrete number has not been established for music. It is well known that Walmart, which is one of the largest retailers of music, carries up to 5,000 music albums in its store; thus, we define the "tail" of the music sales distribution as those albums having an Amazon.com rank greater than 5,000 and the "body" as those albums having an Amazon.com rank of less than 5,000. As a robustness check, we also analyze alternative partitions of the distribution into

[5] Recall that the THM click-through data cover the period from the time the song was posted on THM through November 14, 2006. We collected data from Amazon as soon after this date as possible.

Table 1 Summary Statistics

Variable	Description	Full sample	Body[a]	Tail[a]
Sampling	Total number of click-throughs to the MP3 link	417.14 (620.60)	578.47 (806.07)	256.54 (265.34)
BlogPop	Blog popularity, represented by the number of del.icio.us bookmarks	55.28 (116.41)	50.80 (91.27)	59.72 (136.84)
MusicPop	Total unit song sales	8,289.16 (60,127.31)	14,538.45 (84,479.86)	2,061.06 (5,655.32)
RadioPlay	Total number of times the song was played on the radio	1,328.97 (9,571.65)	2,327.33 (13,414.47)	336.23 (1,381.37)
DaysPost	Number of days since the song was posted on THM	104.54 (17.75)	104.93 (18.22)	104.15 (17.27)
DaysRel	Number of days since the release of the album	1,643.58 (1,767.58)	1,513.95 (1,709.67)	1,772.63 (1,815.19)
RevVal	Average valence of customer reviews on Amazon	4.35 (0.43)	4.37 (0.40)	4.33 (0.46)
RevNum	Number of customer reviews on Amazon	102.26 (174.38)	151.43 (209.25)	53.32 (110.98)
RevStdDev	Standard deviation of the last 100 customer reviews on Amazon	0.97 (0.32)	1.01 (0.27)	0.93 (0.37)
AlbumSales	Album unit sales since MP3 posted on THM	22,590.96 (84,328.50)	42,559.96 (11,5916.92)	2,712.01 (5,423.38)
SalesRank	Amazon sales rank	14,784.04 (23,715.32)	1,805.10 (1,321.93)	27,704.44 (28,038.92)
RecentRel	Dummy variable = 1 if song posted within 10 weeks of release date; 0 otherwise	0.18 (0.38)	0.20 (0.40)	0.16 (0.36)
ArtistRep	Dummy variable = 1 if artist has "high" reputation; 0 otherwise	0.13 (0.34)	0.22 (0.42)	0.05 (0.21)
Indie	Dummy variable = 1 if independent label; 0 otherwise	0.32 (0.44)	0.26 (0.44)	0.38 (0.49)
Tail	Dummy variable = 1 if music is in the tail; 0 otherwise	0.50 (0.50)		
N		1,762	880	882

[a]The "body" is defined as albums with Amazon sales rank $\leq$ 5,000, whereas the "tail" consists of albums with Amazon sales rank $>$ 5,000. Standard deviations are in parentheses.

body and tail based on cutoffs of 2,000 and 10,000, respectively.

When we examine the summary statistics in Table 1, we see that the average sales rank in the body is roughly 1,800 compared to almost 28,000 in the tail. The distinction becomes even more evident when we see that average song sales (*MusicPop*) in the tail are much lower (but have considerable variation) than in the body; similarly, the average *AlbumSales* are much higher in the body than in the tail. *BlogPop* has approximately the same average value and standard deviation in the full sample, the body, and the tail. The average *RevNum* in the body is much larger than the average *RevNum* in the tail, whereas the mean of *RevVal* is high and similar across the two groups. Whereas 38% of albums in the tail are independent albums, only 26% of the albums in the body are independent. The average of *Sampling* is lower in the tail. *DaysPost* and *DaysRel* are approximately the same across the subsamples. Roughly 22% of the artists in the body have an established reputation, whereas only 5% of the artists of albums in the tail do.

3.2. Empirical Specification

We are primarily interested in understanding the relationship between sampling and music/blog popularity. Thus, our dependent variable is *Sampling*, and the key explanatory variables are blog popularity *BlogPop* and music popularity *MusicPop*. These are linked together in the following equation, where for any track i,

$$\begin{aligned}\log(Sampling_i) &= \alpha_0 + \alpha_1 \log(BlogPop_i) + \alpha_2 \log(MusicPop_i) \\ &\quad + \alpha_3 RevVal_i + \alpha_4 \log(RevNum_i) \\ &\quad + \alpha_5 RevStdDev_i + \alpha_6 \log(DaysPost_i) \\ &\quad + \alpha_7 \log(DaysRel_i) + \alpha_8 RecentRel_i \\ &\quad + \alpha_9 Indie_i + \alpha_{10} ArtistRep_i + \sum_{l=1}^{L} \delta_l Genre_{il} + \varepsilon_i, \quad (1)\end{aligned}$$

Table 2 Correlation Matrix

	Log(*Sampling*)	Log(*BlogPop*)	Log(*MusicPop*)	Log(*RadioPlay*)	Log(*Sales*)	Log(*SalesRank*)	*RevVal*	Log(*RevNum*)	*RevStdDev*	Log(*DaysPost*)	Log(*DaysRel*)	*ArtistRep*	*RecentRel*	*Indie*	*Tail*
Log(*Sampling*)	1	0.079***	0.425***	0.189***	0.443***	−0.444***	−0.056**	0.333***	0.146***	−0.007	−0.073***	0.182***	0.064***	0.101***	−0.373***
Log(*BlogPop*)		1	−0.028	−0.043*	0.003	−0.001	0.020	−0.046*	0.028	−0.002	−0.048**	−0.040*	0.047**	−0.013	−0.013
Log(*MusicPop*)			1	0.739***	0.465***	−0.477***	−0.063***	0.346***	0.120***	0.012	0.068***	0.268***	−0.068***	−0.230***	−0.379***
Log(*RadioPlay*)				1	0.329***	−0.317***	−0.068***	0.212***	0.093***	0.014	0.071***	0.024***	−0.022	−0.216***	−0.253***
Log(*Sales*)					1	−0.814***	−0.097	0.471***	0.248***	0.095***	−0.354***	0.331***	0.325***	−0.134***	−0.665***
Log(*SalesRank*)						1	−0.027	−0.568***	−0.162***	−0.008	0.101***	−0.337***	−0.087***	0.187***	0.814***
RevVal							1	−0.099***	−0.673***	−0.018	0.164***	−0.065***	−0.041*	0.093***	−0.044*
Log(*RevNum*)								1	0.333***	0.016	0.381***	0.408***	−0.300***	−0.315***	−0.460***
RevStdDev									1	0.010	−0.105***	0.136***	−0.021	−0.070	−0.120***
Log(*DaysPost*)										1	0.069***	0.114***	0.000	−0.041**	−0.017
Log(*DaysRel*)											1	0.173***	−0.676***	−0.322***	0.085***
ArtistRep												1	−0.099***	−0.223***	−0.255***
RecentRel													1	0.074***	−0.051**
Indie														1	0.133***
Tail															1

Note. These results are based on a total of 1,762 observations.
***, **, and *Denote significance at 1%, 5%, and 10%, respectively.

Table 3 Hausman Specification Test

Efficient under H0	Consistent under H1	Statistic	*p*-Value
OLS	2SLS	94.15	$p < 0.01$

where $Genre_{il}$, for $l = 1, \ldots, L$, are dummy variables so that $Genre_{il} = 1$ if track i is of genre l and 0 otherwise; the other variables are as described in Table 1.

In the above specification, one might suspect the endogeneity of *BlogPop* and *MusicPop* because of missing variables jointly correlated with the dependent variable *Sampling*. One example of such a missing variable might be song quality because sampling and both popularity measures are likely to be increasing in song quality. Song reviews on sites such as Amazon and iTunes could serve as a useful measure of song quality, but unfortunately, song-level reviews are only available for a very small fraction of songs in our data set. We are able to obtain album-level review data, which should be correlated with song quality, although imperfectly. We include album review data in the specification (*RevVal*, *RevNum*, and *RevStdDev*), but this may not remove endogeneity completely.

We conducted the Hausman specification test to help choose the correct estimation method, where we compare ordinary least squares (OLS) with two-stage least squares (2SLS) estimation. Table 3 presents the results from the Hausman specification test, comparing OLS with 2SLS, where the latter allows for the endogeneity of *BlogPop* and *MusicPop* (along with *Sampling*), which indicates that 2SLS is preferred to OLS. The endogenous variable *BlogPop* is instrumented by the variable *Technorati*,[6] measured as the number of in-links to a given blog (which is highly correlated with *BlogPop* but not correlated with *Sampling*). *MusicPop* is also endogenous and is instrumented by the variable *AlbumSales*.[7] The set of instrumental variables also includes all other exogenous variables in Equation (1). Finally, White's (1980) test indicated significant heteroskedasticity, so we use heteroskedasticity-adjusted standard errors throughout.

4. Empirical Results

4.1. Hypothesis Tests

In this section we present the results of tests for our hypotheses. Table 4 presents results comparing the

[6] A blog with more links to or from other blogs (i.e., having a higher Technorati score) has a higher blog popularity in the music blogosphere, which should be correlated with blog popularity within THM but will not necessarily drive sampling within the THM site; this is supported by the correlations mentioned.

[7] Finding an instrument for *MusicPop* is difficult. We use *AlbumSales* because it is correlated to sales of songs within the album; however, albums sales is less likely to drive song-level sampling directly.

Table 4 Regression Results for Full Sample Based on OLS and 2SLS Estimation

	OLS	2SLS
Intercept	4.002***	1.241
	(0.687)	(0.875)
Log(*BlogPop*)	0.067***	0.070***
	(0.015)	(0.022)
Log(*MusicPop*)	0.194***	0.551***
	(0.013)	(0.041)
RevVal	0.020	0.039
	(0.068)	(0.082)
Log(*RevNum*)	0.241***	0.086***
	(0.022)	(0.031)
RevStdDev	−0.099	−0.048
	(0.104)	(0.128)
Log(*DaysPost*)	−0.033	0.004
	(0.120)	(0.145)
Log(*DaysRel*)	−0.128***	−0.045
	(0.027)	(0.034)
Indie	0.014	0.215***
	(0.051)	(0.063)
RecentRel	0.193**	0.296***
	(0.080)	(0.093)
ArtistRep	0.121	−0.084
	(0.078)	(0.095)
N	1762	1762
Adjusted R^2	0.299	0.247

Notes. Variables are as defined in Table 1. The dependent variable is Log(*Sampling*), and the results correspond to the OLS and 2SLS estimation of Equation (1), with tail set at the Amazon sales rank of 5,000. Log(*BlogPop*) and Log(*MusicPop*) are treated as endogenous variables, and the instruments are Log(*Technorati*) and Log(*AlbumSales*) as well as all other exogenous independent variables. Heteroskedasticity-adjusted standard errors are in parentheses.

***, **, and *Denote significance at 1%, 5%, and 10%, respectively.

OLS and 2SLS estimation methods for the full sample. We see that the signs and significance of the coefficients are largely consistent. Given the results of the Hausman test reported in Table 3, we continue our analysis using only 2SLS for our subsample estimation. Focusing on the 2SLS column, we see that the estimated coefficients are generally consistent with our prior expectations: *BlogPop* and *MusicPop* are both positive and significant, providing support for H1A. Looking at the other coefficient estimates, we see that *RevNum* is positive and significant, as is *RecentRel*, confirming that a song that is posted closer to its album release day is sampled more. We also see that *Indie* is positive and significant, indicating that songs released by independent labels are on average sampled more than are songs released by major labels. Interestingly, *RevVal*, *DaysPost*, *DaysRel*, and *ArtistRep* are not significant.

To examine H1B, Table 5 presents results for a sample split based on a song's life cycle, partitioning the data set based on whether the song was posted within 13 weeks of the album's release date (the so-called "shallow releases" in the music industry) or outside 13 weeks of the album's release date ("deep releases" in the music industry). We see that H1B is supported by the results. Specifically, the *MusicPop* coefficient is larger for shallow releases compared with deep releases, and the difference is significant ($p < 0.01$). The point estimate of *BlogPop* is positive and significant for both subsamples. The coefficient is larger in magnitude for shallow releases compared with deep releases, consistent with H1B, but the difference is not significant.

Now we turn to the estimation results of regression Equation (1) for the body and tail subsamples, presented in Table 6. We find support for both H2A and H2B. That is, both *BlogPop* and *MusicPop* have a stronger association with sampling in the tail compared with the body ($p < 0.01$). More specifically, the marginal effect of blog popularity is a stronger determinant of music sampling for niche music (music in the tail) than for mainstream music (music in the body). Similarly, the marginal effect of music popularity is a stronger determinant of music sampling

Table 5 Regression Results for Recent vs. Older Album Releases

	Posted ≤ 13 weeks of album release (shallow releases)	Posted > 13 weeks of album release (deep releases)
Intercept	−3.678	2.929***
	(2.380)	(0.945)
Log(*BlogPop*)	0.089*	0.060**
	(0.049)	(0.025)
Log(*MusicPop*)	0.926***	0.465***
	(0.150)	(0.039)
RevVal	0.243	−0.009
	(0.179)	(0.098)
Log(*RevNum*)	−0.265**	0.128***
	(0.127)	(0.032)
RevStdDev	0.481	−0.147
	(0.308)	(0.150)
Log(*DaysPost*)	1.284***	−0.144
	(0.478)	(0.152)
Log(*DaysRel*)	−0.788***	−0.077**
	(0.302)	(0.033)
Indie	−0.086	0.207***
	(0.167)	(0.067)
ArtistRep	−0.513	−0.006
	(0.419)	(0.091)
N	365	1397
Adjusted R^2	0.256	0.243

Notes. Variables are as defined in Table 1. The dependent variable is Log(*Sampling*), and the results correspond to the 2SLS estimation of Equation (1), with tail set at Amazon sales rank of 5,000. Log(*BlogPop*) and Log(*MusicPop*) are treated as endogenous variables, and the instruments are Log(*Technorati*) and Log(*AlbumSales*) as well as all other exogenous independent variables. Heteroskedasticity-adjusted standard errors are in parentheses.

***, **, and *Denote significance at 1%, 5%, and 10%, respectively.

Table 6 Regression Results for Body and Tail Subsamples

	Body (Amazon rank $\leq$ 5,000)	Tail (Amazon rank $>$ 5,000)
Intercept	0.408	1.271
	(1.159)	(1.837)
Log(*BlogPop*)	0.044	0.110**
	(0.029)	(0.044)
Log(*MusicPop*)	0.476***	0.885***
	(0.043)	(0.136)
RevVal	0.201	−0.168
	(0.129)	(0.142)
Log(*RevNum*)	0.124***	0.134**
	(0.044)	(0.056)
RevStdDev	0.021	−0.268
	(0.209)	(0.212)
Log(*DaysPost*)	0.095	−0.109
	(0.177)	(0.292)
Log(*DaysRel*)	−0.050	−0.129**
	(0.043)	(0.065)
Indie	0.288***	0.246**
	(0.081)	(0.125)
RecentRel	0.440***	0.149
	(0.121)	(0.189)
ArtistRep	−0.076	−0.208
	(0.099)	(0.308)
N	880	882
Adjusted R^2	0.227	0.082

Notes. Variables are as defined in Table 1. The dependent variable is Log(*Sampling*), and the results correspond to the 2SLS estimation of Equation (1), with tail set at the Amazon sales rank of 5,000. Log(*BlogPop*) and Log(*MusicPop*) are treated as endogenous variables, and the instruments are Log(*Technorati*) and Log(*AlbumSales*) as well as all other exogenous independent variables. Heteroskedasticity-adjusted standard errors are in parentheses.

***, **, and *Denote significance at 1%, 5%, and 10%, respectively.

for niche music than for mainstream music. Together, these results indicate that the popularity of both music and blogs is a particularly important signal of quality for niche music, where there is more uncertainty compared with mainstream music.

4.2. Robustness Checks

We start by considering alternative partitions of body and tail to make sure our comparative results are not an artifact of the criterion used for the distinction. To do this, we consider two different sets of partitions, one based on a threshold of 2,000 for the Amazon rank and the other using a cutoff of 10,000. The results of the 2SLS estimation for these subsamples are reported in Table 7. As can be seen, the results for both sets of subsamples are consistent with H2A and H2B.

One might be concerned that because of the unequal sampling of music blogs, the most popular blogs might be skewing our results. We can test this issue by eliminating the blogs with extremely high popularity. We do this by reestimating Equation (1) after removing the top 10% and top 25% most popular blogs as measured by *BlogPop*. The results for both of these reduced samples (not reported for the sake of brevity) are largely consistent with our prior findings.

5. Discussion and Conclusions

In this study, we integrated the research on observational learning and the long tail to understand how consumers make music consumption decisions, particularly looking at how these decisions are made differently for mainstream and niche music. We propose that because of the quality uncertainty associated with niche music, observational learning would be a stronger driver of consumption of niche music (in the tail) compared with mainstream music (in the body). Indeed, our empirical results establish that both blog popularity and music popularity have a stronger association with sampling in the tail compared with the body of music sales distribution. We also find that product life cycle plays a role in the relationship between popularity information and consumption in that music popularity information becomes less important the longer the music has been on the market, though we see a tendency for blog popularity to be more influential for older music compared with newer music.

Given our findings, a natural and interesting question that follows is how this opportunity for music discovery through social media (not only traditional media, i.e., the radio) ultimately affects music sales. We are not able to provide a definitive analysis of this important question because of limitations in our data set: we only have music blog sampling data from one site (The Hype Machine), which is only a fraction of overall blog sampling. However, we have found that sampling on THM and overall blog buzz[8] during this time period are positively and significantly correlated, thus indicating that the THM audience may be representative of the broader population of online music consumers;[9] i.e., THM sampling might be a reasonable proxy for overall blog sampling. With this caveat in mind, Table 8 provides the results of a 2SLS regression that relates song sales to radio play and THM sampling (both treated as endogenous), along with several control variables. Looking at the full sample, we see that both radio play (traditional media) and

[8] Blog buzz is calculated by the number of blogs that blogged about a given song, as indicated by Google Blog Search, during the same time period as our sampling data from THM.

[9] For the songs in our data set, we examined the correlation between sampling on THM and overall buzz about the songs in blogs on the Internet and found that the correlation between THM sampling and blog buzz is 0.357 ($p < 0.01$). This suggests that the intensity of sampling on THM is a good proxy for the overall buzz about the album in the blogosphere.

Table 7 Robustness to Alternative Partition of Body and Tail

	Body (Amazon rank ≤ 2,000)	Tail (Amazon rank > 2,000)	Body (Amazon rank ≤ 10,000)	Tail (Amazon rank > 10,000)
Intercept	−0.972	0.354	0.292	0.812
	(1.581)	(1.347)	(1.046)	(2.757)
Log(*BlogPop*)	0.051	0.079**	0.061**	0.107*
	(0.041)	(0.031)	(0.026)	(0.063)
Log(*MusicPop*)	0.568***	0.732***	0.497***	1.147***
	(0.058)	(0.086)	(0.039)	(0.269)
RevVal	0.197	0.063	0.191	−0.305
	(0.192)	(0.106)	(0.108)	(0.203)
Log(*RevNum*)	0.164**	0.098**	0.081**	0.227***
	(0.068)	(0.042)	(0.039)	(0.079)
RevStdDev	0.023	−0.003	−0.014	−0.541*
	(0.366)	(0.160)	(0.177)	(0.302)
Log(*DaysPost*)	0.163	−0.002	0.102	−0.065
	(0.239)	(0.210)	(0.164)	(0.422)
Log(*DaysRel*)	−0.046	−0.095**	−0.024	−0.229***
	(0.060)	(0.047)	(0.038)	(0.097)
Indie	0.327***	0.253***	0.287***	0.303
	(0.123)	(0.091)	(0.073)	(0.188)
RecentRel	0.638***	0.147	0.467***	−0.127
	(0.174)	(0.126)	(0.109)	(0.231)
ArtistRep	−0.161	0.017	−0.056	−0.219
	(0.129)	(0.176)	(0.098)	(0.459)
N	525	1237	1146	616
Adjusted R^2	0.230	0.129	0.238	0.042

Notes. Variables are as defined in Table 1. The dependent variable is Log(*Sampling*), and the results correspond to the 2SLS estimation of Equation (1), with tail set at Amazon sales rank of 2,000. Log(*BlogPop*) and Log(*MusicPop*) are treated as endogenous variables, and the instruments are Log(*Technorati*) and Log(*AlbumSales*) as well as all other exogenous independent variables. Heteroskedasticity-adjusted standard errors are in parentheses.

***, **, and *Denote significance at 1%, 5%, and 10%, respectively.

sampling (social media) are associated with incremental song sales. The point estimates for the body/tail subsamples differ slightly from each other, but the differences are not statistically significant.

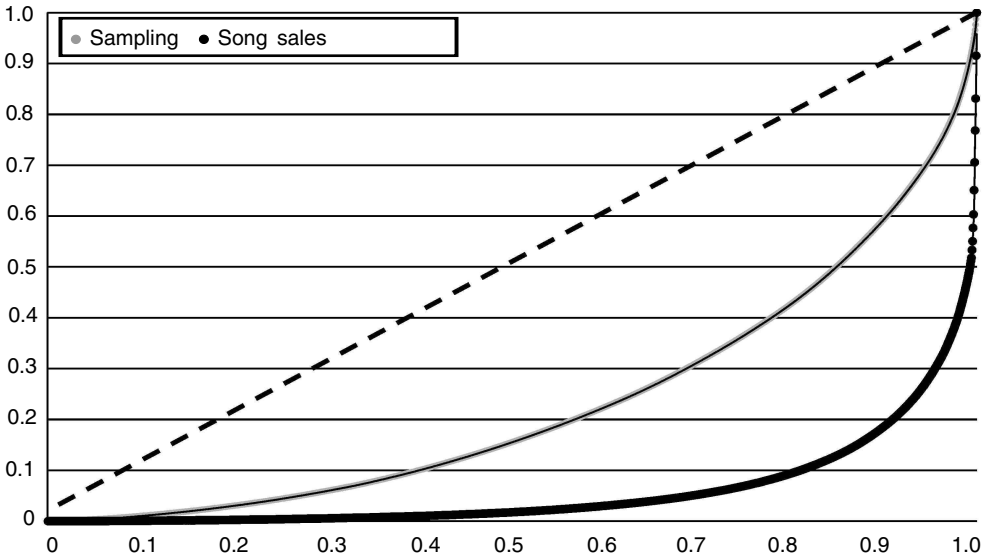
These findings have implications for artists, music blogs, and music communities such as The Hype Machine. Given that the goals of these entities are to increase consumption of music, particularly music that may not be discovered otherwise, the results of our study provide straightforward recommendations for engaging social media and driving consumption. For example, artists that do not have an established reputation (and therefore are likely to have music in the tail of the distribution) will benefit from the endorsement from a popular music blog; this is likely to increase consumption through sampling. Music blogs benefit from being popular because consumers will seek out their recommendations; thus, engaging their users in such a way they bookmark and recommend their blog is important. Finally, taking THM as an example, the goal of many of these communities is about "music discovery." We have seen that blog popularity drives sampling of otherwise less known music (in the tail) as well as older music; displaying the blog popularity information could help increase the diversity of consumption even more.

Table 8 2SLS Estimation of Song Sales

	Full sample	Body (Amazon rank ≤ 5,000)	Tail (Amazon rank > 5,000)
Intercept	0.415	2.058***	−0.321
	(0.571)	(0.967)	(0.795)
Log(*RadioPlay*)	0.454***	0.446***	0.445***
	(0.011)	(0.015)	(0.017)
Log(*Sampling*)	0.692***	0.712***	0.535***
	(0.059)	(0.077)	(0.096)
RevVal	0.024	−0.200	0.180
	(0.081)	(0.145)	(0.107)
Log(*RevNum*)	0.064**	0.106**	−0.049
	0.031	(0.049)	(0.043)
RevStdDev	−0.074	−0.401	0.182
	(0.129)	(0.244)	(0.151)
Log(*DaysRel*)	0.114***	0.043	0.241***
	(0.035)	(0.048)	(0.049)
Indie	−0.118*	−0.096	−0.094
	(0.068)	(0.093)	(0.095)
RecentRel	−0.236***	−0.321**	−0.162
	(0.096)	(0.130)	(0.136)
ArtistRep	0.077	0.008	0.439**
	(0.085)	(0.096)	(0.191)
N	1762	880	882
Adjusted R^2	0.639	0.661	0.515

***, **, and *Denote significance at 1%, 5%, and 10%, respectively.

Figure 3 Lorenz Curves: Music Blog Sampling vs. Song Sales



Note. The Gini coefficient, calculated as the proportion of the area between the 45° line and the Lorenz curve to the total area under the 45° line, is equal to 0.53 for music blog sampling and 0.88 for song sales.

Putting together all of our results, we can distill out a few key observations. First, new and fast-growing social media in the form of music blogs are exposing consumers to a far wider range of music than do traditional media such as the radio (see Figure 1). Second, music blog sampling is characterized by more equality of consumption compared with the distribution of song sales, as shown in the Lorenz curves in Figure 3. That is, consumers are more willing to sample music in the long tail, but they are less willing to purchase such songs. Finally, our preliminary results (see Table 8) with song sales as the dependent variable suggest that music blog sampling is associated with higher song sales—both in the body and in the tail. This reminds us of the empirical analysis of Tucker and Zhang (2007), who show that online popularity information at an online retailer simultaneously resulted in both the "steep tail" (i.e., popular products become more popular) and "long tail" (i.e., increased sales of niche music) phenomena. This is plausible in our context, with our results implying that social media (i.e., music blog sampling) not only expand sales of popular songs but also bring new consumers into the market by exposing them to niche music, something that is not feasible in the realm of traditional media (i.e., radio play). These results provide evidence that social media are a driver of the long tail in sampling, as they provide access to a larger variety of music and enable consumers to find and consume this more easily. Given that our results also show preliminary evidence that sampling drives song sales, this suggests that social media could ultimately lead to a long-tailing of music sales as well. We admit this latter conclusion is somewhat speculative at this point, and it would be fruitful for further research to examine the issue in more concrete terms.

This work does have some limitations, overcoming which will provide some directions for future research. First, we have sampling data at a single point in time, so this precludes an analysis of how demand shifts over time are related to blogging and sampling activity. The lack of data over time also prevents an analysis of peer effects in sampling, wherein sampling itself could be driven by prior sampling behavior by other consumers. Second, as always, finding the appropriate instrumental variables for the endogenous variables is challenging; given the data that we have access to, we have used the best instrumental variables available and have provided tests to demonstrate the robustness of our results. Third, as discussed above, we have data from only one online music community, which limits the conclusions that we can draw on the relationship between blogging and sales. We also observe only songs that have been blogged about and do not observe the counterfactual; conducting an analysis where songs that both are and are not blogged about are observed could provide additional insights into the role of observational learning on consumption decisions. Finally, the generalizability of our results outside of the music domain is an open question and could serve as a fruitful direction for further research.

Acknowledgments

The order of the authors is alphabetical and they contributed equally. The authors thank Anthony Volodkin from The Hype Machine for generously sharing his music blog aggregator data and NielsenSoundscan for providing essential

music sales data. One of the authors also acknowledges a generous dissertation fellowship from CalIT2, sponsored by the Emulex Corporation. Finally, the authors are grateful for helpful comments and suggestions from the seminar participants at Conference on Information Systems and Technology (CIST) 2007, Workshop on Information Systems and Economics (WISE) 2007, and International Symposium of Information Systems (ISIS) 2007.

References

Anderson, C. 2004. The long tail. *Wired* (October) 170–177.

Anderson, C. 2006. *The Long Tail*. Hyperion Books, New York.

Anderson, L. R., C. A. Holt. 1997. Information cascades in the laboratory. *Amer. Econom. Rev.* **87**(5) 847–862.

Banerjee, A. V. 1992. A simple model of herd behavior. *Quart. J. Econom.* **107**(3) 797–817.

Bearden, W. O., M. J. Etzel. 1982. Reference group influence on product and brand purchase decisions. *J. Consumer Res.* **9**(9) 183–194.

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

Bikhchandani, S., D. Hirshleifer, I. Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *J. Econom. Perspect.* **12**(3) 151–170.

Brynjolfsson, E., Y. Hu, M. D. Smith. 2006. From niches to riches: The anatomy of the long tail. *Sloan Management Rev.* **47**(4) 67–71.

Chellappa, R. K., B. Konsynski, V. Sambamurthy, S. Shivendu. 2007. An empirical study of the myths and facts of digitization in the music industry. Presentation 2007 Workshop Information Systems Economics (WISE), Montreal.

Chen, Y., Q. Wang, J. Xie. 2010. Online social interactions: A natural experiment on word of mouth versus observational learning. Working paper, University of Florida, Gainesville.

Dellarocas, C. 2003. The digitization of word of mouth: Promises and challenges of online feedback mechanisms. *Management Sci.* **49**(10) 1407–1424.

Godes, D., D. Mayzlin. 2004. Using online conversation to study word of mouth communication. *Marketing Sci.* **23**(4) 545–560.

Godes, D., D. Mayzlin. 2009. Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Sci.* **28**(4) 721–739.

Kelman, H. C. 1961. Processes of opinion change. *Public Opinion Quart.* **25**(1) 57–78.

King, M. F., S. K. Balasubramanian. 1994. The effects of expertise, end goal, and product type on adoption of preference formation strategy. *J. Acad. Marketing Sci.* **22**(2) 146–159.

Nelson, P. 1970. Information and consumer behavior. *J. Political Econom.* **78**(2) 311–329.

Price, L. L., L. F. Feick, R. A. Higie. 1989. Preference heterogeneity and coorientation as determinants of perceived informational influence. *J. Bus. Res.* **19**(3) 227–242.

Salganik, M. J., P. S. Dodds, D. J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**(5762) 854–856.

Sanneh, K. 2004. A draining week in the indie-music spotlight. *New York Times* (October 18), http://www.nytimes.com/2004/10/18/arts/music/18band.html.

Senecal, S., J. Nantel. 2004. The influence of online product recommendations on consumers' online choices. *J. Retailing* **80**(2) 159–169.

Tucker, C., J. Zhang. 2007. Long tail or steep tail: A field investigation into how popularity information affects the distribution of customer choices. Working paper, MIT Sloan School Working Paper 4655-07, Cambridge, MA.

Tucker, C., J. Zhang. 2009. How does popularity information affect choices? A field experiment. *Management Sci.* **57**(5) 828–842.

White, H. 1980. A heteroskedasticity-consistent covariance matrix-estimator and a direct test for heteroskedasticity. *Econometrica* **48**(4) 817–838.