Information and the Skewness of Music Sales

Ken Hendricks

University of Texas at Austin

Alan Sorensen

Stanford University and National Bureau of Economic Research

This paper studies the role of product discovery in the demand for recorded music. We show that releasing a new album causes a substantial and permanent increase in sales of the artist's old albums—especially if the new release is a hit. Patterns in these "backward spillovers" suggest that they result from consumers discovering the artist upon hearing the new release. To explore the implications of consumers' incomplete information, we estimate a simple, learning-based model of market demand. Our results imply that the distribution of sales is substantially more skewed than it would be if consumers were more fully informed.

I. Introduction

In cultural markets such as books, music, and movies, consumers face an overwhelmingly large and constantly growing choice set, as many new products flow into the market each week. However, only a small fraction of these products turn out to be profitable. Even among the profitable products, the distribution of returns is extremely skewed: a

We thank Dirk Bergmann, Greg Crawford, Steve Durlauf, Phillip Leslie, Marc Rysman, and Michael Whinston, as well as the editor and referees, for suggestions that prompted changes in the paper. We also had invaluable conversations with Don Engel, Michael Lopez, and Ralph Peer that helped clarify several institutional details about the music industry. Chris Muratore and Nielsen SoundScan were very helpful in providing the data, and Natalie Chun and Abe Dunn provided outstanding research assistance. We are responsible for any errors or shortcomings that remain in spite of all the thoughtful input.

[*Journal of Political Economy,* 2009, vol. 117, no. 2]

large share of total industry profit is claimed by a small number of very successful products. The skewness may simply reflect the products' relative qualities. However, it may also reflect a lack of information about the choice set: if consumers are unaware or poorly informed about most products, then market demand depends not only on their preferences but also on their knowledge of the product space and the process by which they obtain this knowledge. In entertainment industries, this process is driven in part by commercial success: consumers buy the products they hear about, and they hear about the products that other consumers buy. As a result, a product's success reinforces itself, causing the distribution of success across products to be more highly concentrated.

Understanding how consumers' lack of information about choice sets affects product market outcomes is important for various reasons. First, it represents a welfare loss to consumers who would prefer to buy less popular products if they knew about them. Second, the processes by which consumers learn about the choice set may affect product variety, for example, by tilting investment toward products with mass-market appeal instead of products targeted at narrower niche markets. Third, discovering a product in cultural markets typically leads consumers to learn about other, related products. For example, readers who liked a book will tend to seek out other books by the same author, and listeners who liked an album will tend to seek out other albums by the same artist. These information spillovers have important implications for investment in authors and artists, the structure of their contracts, and the lengths of their careers. Finally, the effect of consumer learning on the distribution of market returns is especially interesting given the recent rise of Internet technologies that dramatically lower the cost of information: our analysis sheds light on how the Internet will change the "shape" of demand in cultural markets.

In this paper we study these issues in the market for recorded music. We analyze music sales in the period just prior to the emergence of online markets, a time when consumers learned about albums primarily through radio play and purchased them mainly at brick-and-mortar stores. Scarce airtime and the desire of radio stations to get the largest possible audience created an informational bottleneck in which consumers listened to a relatively small fraction of albums, typically the most popular ones. Our objective is to quantify the extent to which albums "lost" sales because consumers may not have known about them. Our empirical strategy for addressing this issue is based on the effects of new album releases on sales of previous albums by the same artist. The promotional activity and radio airplay associated with a newly released album enhance consumer awareness about the artist and cause some consumers to discover and purchase the artist's past albums (which are referred to in the industry as "catalog" albums). We call this effect

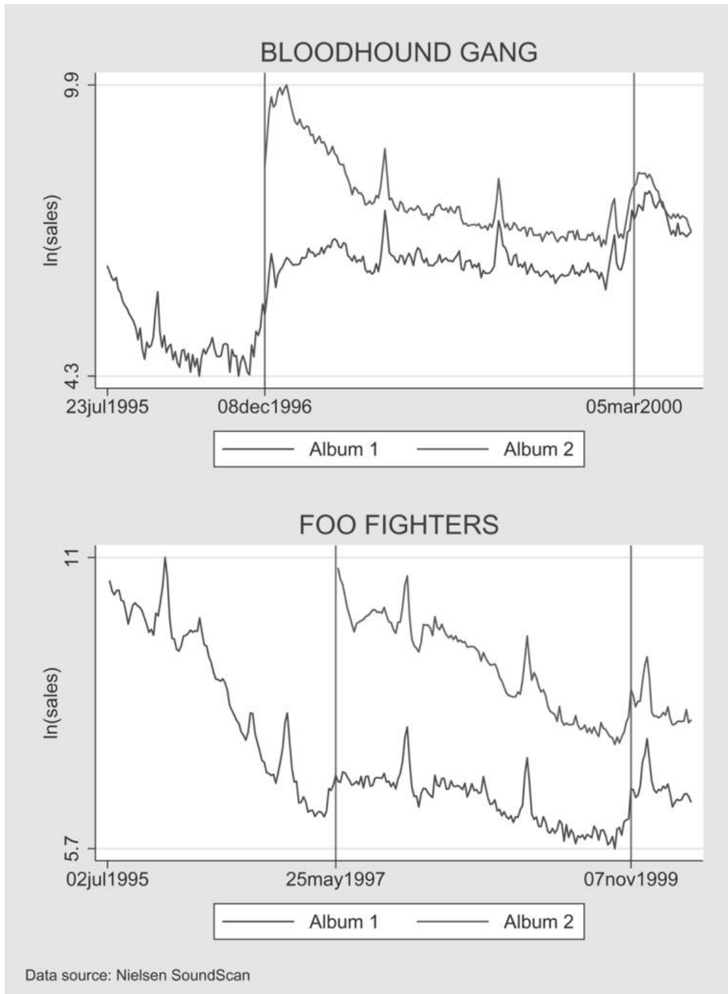

FIG. 1.—Album sales paths for two examples. These graphs show log(sales) over time (measured in weeks) for the artists' first and second albums. The vertical lines indicate the release dates of albums 2 and 3. The graphs illustrate the *backward spillover*: the release of a new album tends to cause a sales increase for previous albums by the same artist.

the *backward spillover*. In order to measure it, we constructed a data set consisting of weekly sales histories for a sample of 355 artists in the period 1993–2002. We observe sales separately for each of the artists' albums, and each artist in the sample released at least two albums (including a debut) during the sample period.

Figure 1 shows two clear examples of the backward spillover. The figure plots the logarithm of weekly national sales for the first and second albums of two popular recording artists, from the time of the artist's debut until 6 months after the artist's third release. The vertical lines in each graph indicate the release dates of the second and third albums. In the weeks surrounding the release dates, sales of catalog titles increased substantially. In the case of the Bloodhound Gang, a relatively obscure alternative rock band, the second album was considerably more popular than the first, and its release catapulted sales of the prior album to levels even higher than it had attained at the time of its own release, with the effect persisting for at least 3 years. For the Foo Fighters, a more popular hard rock band with a very successful debut album, the impact of the second release was somewhat less dramatic but still generated an increase in sales of the band's first album. In both examples, the backward spillover is significantly positive for both the second and third album releases.

The first part of our empirical analysis examines the variation in spillover sales in the weeks before and after the new album is released. We use an approach taken from the literature on treatment effects to measure the spillovers. The results confirm that the three patterns illustrated in figure 1 hold on average for artists in our sample. First, the increased sales of catalog albums start to appear roughly 4 weeks prior to the release of a new album and increase throughout the prerelease period. Second, the effect peaks in the week of the release and thereafter remains roughly constant as a percentage of sales for many months. Third, the spillovers are larger when the new release is a hit, and especially large when the new release is a hit and the catalog album was not. Finally, we also show that backward spillovers are smaller in an artist's home market (i.e., the city in which the artist began her career) even though sales are on average higher in the home market. These patterns suggest that spillovers result from changes in consumers' information. While our analysis does not rule out explanations based on changes in consumers' utility,[1] the patterns are most easily explained by consumers discovering artists from their new releases and learning about their catalog albums.

We pursue a more structural analysis of the album discovery explanation in the second part of our empirical analysis. We develop and estimate a model of market demand for catalog albums in the year following the release of a new album, focusing on total demand for the year rather than on weekly demand. The probability that a consumer purchases the catalog album in the first year of the new album release

[1] For example, preferences over an artist's albums could be supermodular (Becker, Grossman, and Murphy [1994] and Gentzkow [2007] are two interesting empirical studies of supermodular preferences), or preferences might depend on the artist's popularity and the new release could increase the artist's popularity. See Becker and Murphy (2000) and Brock and Durlauf (2001) for models with social effects in consumption.

is the product of two probabilities: the probability that she discovers the album during this period and the probability that she likes the album. The release of the new album is assumed to have no effect on consumers' preferences for the artist's catalog album, but it can increase the likelihood that consumers discover the catalog album. We specify a parametric function describing the probability of discovery, allowing that function to depend on first-year sales of the new album. Conditional on cumulative sales of the catalog album prior to the release of the new album, sales of the new album represent an exogenous shock to the probability of discovering the catalog album. This assumption allows us to empirically identify the parameters of the discovery function. We estimate the parameters using variation across artists in the spillover sales of second albums onto debut albums. We then use the estimated parameters to forecast the spillovers of the artist's third album onto her first and second albums and exploit these forecasts to construct tests of the model's underlying assumptions. On the basis of the results of these tests, we conclude that while other factors such as social effects may affect demand for albums, demand for catalog albums is driven largely by whether consumers know about them and the process through which they obtain this knowledge (i.e., radio play).

The primary motivation for estimating the discovery probability function is to conduct counterfactual analyses. Our main counterfactual consists of measuring the "lost" sales of debut albums due to consumers not discovering the album upon its release. Our results imply that while almost all consumers learn about an artist with a major hit, only 32 percent of consumers learn about an artist whose album achieves the median level of sales. This finding implies that if consumers were more fully informed, sales would have been substantially less skewed. For example, sales of the top artist in our sample would have exceeded the median artist's sales by a factor of 30 instead of the observed factor of 90. We also run a counterfactual that involves forecasting sales of second albums in the absence of a debut album (i.e., if the second albums had instead been the debut albums). We find that the difference between counterfactual sales and observed sales is large: collectively, the second albums in our sample sold 25 percent more than they would have if they had not been preceded by another album. We call this effect the *forward spillover*. It implies that contractual relationships between artists and record labels are complicated by a significant holdup problem and rationalizes the pervasive use of long-term contracts in the industry.

A recent experimental study by Salganik, Dodds, and Watts (2006) provides considerable support for our model and results. They created an artificial online music market in which thousands of participants arrived sequentially and were presented with a list of songs by unknown artists. Participants chose whether to listen to, rate, and download each song (for free), either with or without knowledge of the download decisions of previous participants. Participants who were shown the songs' popularity ranks tended to listen only to the most popular songs; however, the probability of downloading a song conditional on listening to it was roughly invariant to whether the participant was shown the song's popularity rank. In other words, participants tended to download the songs that others downloaded because they *listened* to the songs that others downloaded, not because their preferences were influenced by the popularity of the song.[2] The popularity rankings also substantially increased the inequality and unpredictability of the songs' download shares. Medium-quality songs had the most unpredictable download totals: "the best songs never did badly" and "the worst songs never did well," but any outcome was possible for songs in between. This is consistent with one of our main findings, which is that mid-range artists are the ones whose sales are most sensitive to the degree of information in the market.

We are not aware of prior empirical literature on information spillovers between products.[3] Goeree (2005) has estimated a structural model of demand for personal computers when consumers may be less than fully informed about the set of available products because of the rapid pace of technological change. Wernerfelt (1988), Choi (1998), and Cabral (2000) have developed theoretical models that study the impact of information spillovers on firms' decisions about whether to release new products under existing brand names. When consumers are uncertain about product qualities, the strong reputation of an existing product increases demand for new products sold under the same brand (the forward spillover), and the release of a high-quality new product can improve the brand image and boost sales of the existing product (the backward spillover).[4] In a sequel to this paper, Hendricks, Sorensen, and Wiseman (2009) use a variant of the herding models of Banerjee (1992), Bikhchandani, Hirshleifer, and Welch (1992), and Smith and Sørensen (2000) to develop a framework for studying demand for search goods such as music albums. Heterogeneous consumers can learn about their preferences for products from the purchasing decisions of other consumers and from costly search. The option to search prior to purchasing leads to different market dynamics and out-

[2] As the authors note, the experiment was not designed to test directly for social effects in consumption because the participants did not know each other. A detailed description of the experimental design is provided in Salganik (2007).

[3] Benkard's (2000) study of learning by doing in aircraft production shows that learning spills over across aircraft types, but we have not seen any empirical papers that analyze information spillovers on the demand side of a market.

[4] In Cabral's paper, e.g., the "feedback reputation effect" is exactly analogous to what we call the backward spillover.

comes than the standard herding models and yields testable predictions that are largely consistent with the results of this paper.

More broadly, our paper contributes to a growing literature about the impact of information provision on market outcomes. In markets with a large number of products whose quality is difficult to determine ex ante, a variety of mechanisms arise endogenously to provide information to consumers. These mechanisms are typically imperfect, however, and evaluating their impact on what gets sold (and, by extension, what ultimately gets produced) is an important objective for empirical research. Recent papers that address this general topic include Jin and Leslie (2003), which examines the effects of publicly posting restaurants' health inspection scores; Sorensen (2007), which analyzes the impact of published bestseller lists on the market for books; and Jin, Kato, and List (forthcoming), which studies the informational role of professional certifiers in the market for sports cards.

The paper is organized as follows. Section II describes the data and provides summary statistics. In Section III we use the data to measure the backward spillovers and document several stylized facts about the spillover. In Section IV we develop and estimate an album discovery model and describe the two counterfactual exercises aimed at revealing the quantitative impacts of consumer learning. Section V presents concluding remarks.

II. Data

Our data describe the album sales histories of 355 music artists who were active between 1993 and 2002. Weekly sales data for each artist's albums were obtained from Nielsen SoundScan, a market research firm that tracks music sales at the point of sale, essentially by monitoring the cash registers at over 14,000 retail outlets. SoundScan is the principal source of sales data for the industry and is the basis for the ubiquitous Billboard charts that track artist popularity. Various online databases were also consulted for auxiliary information (e.g., about genres and record labels) and to verify album release dates.

The sample was constructed by first identifying a set of candidate artists who released debut albums between 1993 and 2002, which is the period for which SoundScan data were available. Sampling randomly from the universe of such artists is infeasible, largely because it is difficult to find information on artists who were unsuccessful. Instead, we constructed our sample by looking for new artists appearing on Billboard charts. The majority of artists in our sample appeared on Billboard's Heatseekers chart, which lists the sales ranking of the top 25 new or ascendant artists each week.[5] A smaller number of artists were found because they appeared on regional New Artists charts, and an even smaller number were identified as new artists whose debut albums went straight to the Top 200 chart. This selection is obviously nonrandom: an artist must have enjoyed at least some small measure of success to be included in the sample. However, although the sample includes some artists whose first appearance on the Heatseekers list was followed by a rise to stardom, we note (and show in detail below) that it also includes many unknown artists whose success was modest and/or fleeting.[6]

Because our primary objective is to study demand responses to newly released albums, we restrict our attention to major studio releases. Singles, recordings of live performances, interviews, holiday albums, and anthologies or greatest hits albums are excluded from the analysis.[7] The resulting sets of albums were compared against online sources of artist discographies to verify that we had sales data for each artist's complete album history; we dropped any artists for whom albums were missing or for whom the sales data were incomplete.[8] Since timing of releases is an important part of our analysis, we also dropped a small number of artists with albums for which we could not reliably ascertain a release date.[9] Finally, we narrowed the sample to artists for whom we observe the first 52 weeks of sales for at least the first two albums; we then include an artist's third album in the analysis if we observe at least the first 52 weeks of sales for that album (i.e., we include third albums if they were released before 2002).

After all these filters were applied, the remaining sample contains 355 artists and 888 albums. The sample covers three broad genres of music: rock (227 artists), rap/rhythm and blues/dance (79 artists), and country/blues (49 artists). The artists in the sample also cover a broad

[5] Artists on the Heatseekers chart are "new" in the sense that they have never before appeared in the overall top 100 of Billboard's weekly sales chart; i.e., only artists who have never passed that threshold are eligible to be listed as Heatseekers.

[6] The weekly sales of the lowest-ranked artist on the Heatseekers chart are typically around 3,000, which is only a fraction of typical weekly sales for releases by famous artists who have graduated from the Heatseekers category.

[7] Greatest hits albums could certainly affect sales of previous albums—repackaging old music would likely cannibalize sales of earlier albums—but we are primarily interested in the impact of new music on sales of old music. Moreover, there are very few artists in our sample who actually released greatest hits albums during the sample period, making it difficult to estimate their impact with any statistical precision.

[8] The most common causes for missing data were that a single SoundScan report was missing (e.g., the one containing the first few weeks of sales for the album) or that we pulled data for the rerelease of an album but failed to obtain sales for the original release.

[9] For most albums, the release date listed by SoundScan is clearly correct; however, for some albums the listed date is inconsistent with the sales pattern (e.g., a large number of sales reported before the listed release date). In the latter case, we consulted alternative sources to verify the release date that appeared to be correct on the basis of the sales numbers. Whenever we could not confidently determine the release date of an album, we dropped it along with all other albums by the same artist.

TABLE 1
Summary Statistics

	N	Mean	Standard Deviation	Percentile .10	Percentile .50	Percentile .90
			A. Date of Release			
Album 1	355	May 13, 1996	102	Aug. 22, 1993	May 5, 1996	Feb. 28, 1999
Album 2	355	July 20, 1998	108	July 23, 1995	Aug. 2, 1998	May 27, 2001
Album 3	178	June 3, 1999	90	Oct. 13, 1996	Aug. 4, 1999	Aug. 5, 2001
			B. First-Year Sales			
Album 1	355	312,074	755,251	7,381	78,360	781,801
Album 2	355	367,103	935,912	10,705	55,675	951,956
Album 3	178	450,716	867,630	7,837	71,674	1,461,214
Overall	888	361,864	854,420	9,095	67,558	996,460
			C. First 4 Weeks/First Year			
Album 1	355	.121	.111	.016	.085	.265
Album 2	355	.263	.137	.086	.263	.441
Album 3	178	.305	.131	.134	.305	.500
Overall	888	.214	.148	.031	.198	.419
			D. Peak Sales Week			
Album 1	355	31.9	47.8	0	15	87
Album 2	355	7.83	23.1	0	0	28
Album 3	178	4.05	13.1	0	0	12
Overall	888	16.7	36.3	0	1	46
			E. Weeks between Releases			
Albums 1, 2	355	114	53.5	58	107	179
Albums 2, 3	178	111	46.7	58	104	169

range of commercial success, from superstars to relative unknowns. Some of the most successful artists in the sample are Alanis Morissette, the Backstreet Boys, and Shania Twain; examples at the other extreme include Jupiter Coyote, the Weakerthans, and Melissa Ferrick.

Table 1 summarizes various important aspects of the data. Panel A shows the distribution of the albums' release dates separately by release number. The median debut date for artists in our sample is May 1996, with some releasing their first albums as early as 1993 and others as late as 2000. There are 178 artists in the sample for whom we observe three releases during the sample period and 177 for whom we observe only two releases. Note that while we always observe at least two releases for each artist (because of the sample selection criteria), if we observe only two, we do not know whether the artist's career died after the second release or if the third album was (or will be) released after the end of the sample period. In what follows we will discuss this right-truncation problem whenever it has a material impact on the analysis.

Panel B of the table illustrates the considerable heterogeneity in sales across albums. For the period covered by our sample, production, marketing, and distribution costs for a typical album were in the ballpark of $500,000, so an album had to sell roughly 50,000 units (assuming a wholesale price of $10 per unit) in order to be barely profitable. Over half of the albums in our sample passed that threshold in the first year. However, although most of the albums in the sample were nominally successful, the distribution of success is highly skewed: as the table illustrates, sales of the most popular albums are orders of magnitude higher than sales of the least popular ones. For debut albums, for example, first-year sales at the 90th percentile are 10 times sales at the median and over 100 times sales at the 10th percentile.

The skewness of returns is even greater across artists than across albums, since artist popularity tends to be somewhat persistent. An artist whose debut album is a hit is likely to also have a hit with her second album, so absolute differences in popularity among a cohort of artists are amplified over the course of their careers. Across the artists in our sample, the simple correlation between first-year sales of first and second releases is 0.52. For second and third releases the correlation is 0.77. Most of an artist's popularity appears to derive from artist-specific factors rather than album-specific factors, but the heterogeneity in success across albums by a given artist can still be substantial.

Another interesting feature of the sales distributions is how little they differ by release number. To the extent that an artist's popularity grows over time, one might expect later albums to be increasingly successful commercially. However, while this pattern holds on average for albums 1–3, even for artists who ultimately have very successful careers it is often the case that the most successful album was the first.

Most albums' sales paths exhibit an early peak followed by a steady, roughly exponential decline. As indicated in panels C and D of table 1, sales typically peak in the very first week and are heavily front-loaded: a large fraction of the total sales occur in the first 4 weeks after release. Debut albums are an exception: first releases sometimes peak after several weeks, which presumably reflects a more gradual diffusion of information about albums by new artists. The degree to which sales are front-loaded increases with each successive release.

Seasonal variation in demand for music compact discs is substantial. Overall, sales are strongest from late spring through early fall, and there is a dramatic spike in sales during mid to late December. Not surprisingly, album release dates exhibit some seasonality as well. Table 2 shows the distribution of releases across months. Late spring through early fall is the most popular time to release a new album, and record companies appear to avoid releasing new albums in December or January. Albums that would have been released in late November or December are presumably expedited in order to capture the holiday sales period.

We define the release period of a new album as the time between its release date and the release date of the next album released by the

TABLE 2
Seasonality in Release Dates

Month	Percentage of Releases Occurring: Album 1 (N = 355)	Album 2 (N = 355)	Album 3 (N = 178)	Overall (N = 888)
January	3.94	3.10	3.37	3.49
February	8.17	4.23	3.93	5.74
March	13.24	9.58	11.80	11.49
April	9.01	8.45	8.99	8.78
May	11.83	9.01	7.30	9.80
June	7.61	12.68	6.74	9.46
July	8.45	9.01	10.11	9.01
August	11.55	9.58	10.67	10.59
September	7.32	11.27	11.80	9.80
October	12.39	10.70	16.29	12.50
November	5.92	11.83	6.74	8.45
December	.56	.56	2.25	.90

same artist. Panel E of table 1 provides information about the length of the release periods. The median release period for debut albums is more than 2 years, and the low end of the distribution is still more than 1 year. Figure 2 shows a more complete picture of the heterogeneity in release periods for adjacent albums. Note that we can compute time to next release only if there was a next release. If an artist's second album was released near the end of the sample period, we observe a third release only if the time to release was short. However, figure 2 shows that the distribution of elapsed time between albums 1 and 2 is clearly very similar to the distribution between albums 2 and 3, which suggests that the right-truncation problem is not very severe for third albums.[10]

In addition to the obvious right-truncation problem, our sample selection is likely to be biased toward artists whose success came early in their careers. For an artist to be selected into our sample, it must be the case that (*a*) the artist appeared on a Billboard chart between 1993 and 2002 and that (*b*) we have data on all the artist's CD sales, which means that the artist's first release must have come after January 1993. Taken together, these conditions imply that artists who hit a Billboard chart early in the sample period must have done so on their first or second album (otherwise we would have excluded them because of a lack of data on their previous releases). Moreover, of the artists debuting late in our sample period, only the ones with early success will make it into our sample, because only they will have appeared on a Billboard chart. So the selection pushes toward artists who start strong. While this means that our data will overstate the tendency of artists' successes to

[10] In a previous version of this paper we included fourth albums in the analysis. The right-truncation problem is much more salient for fourth albums.

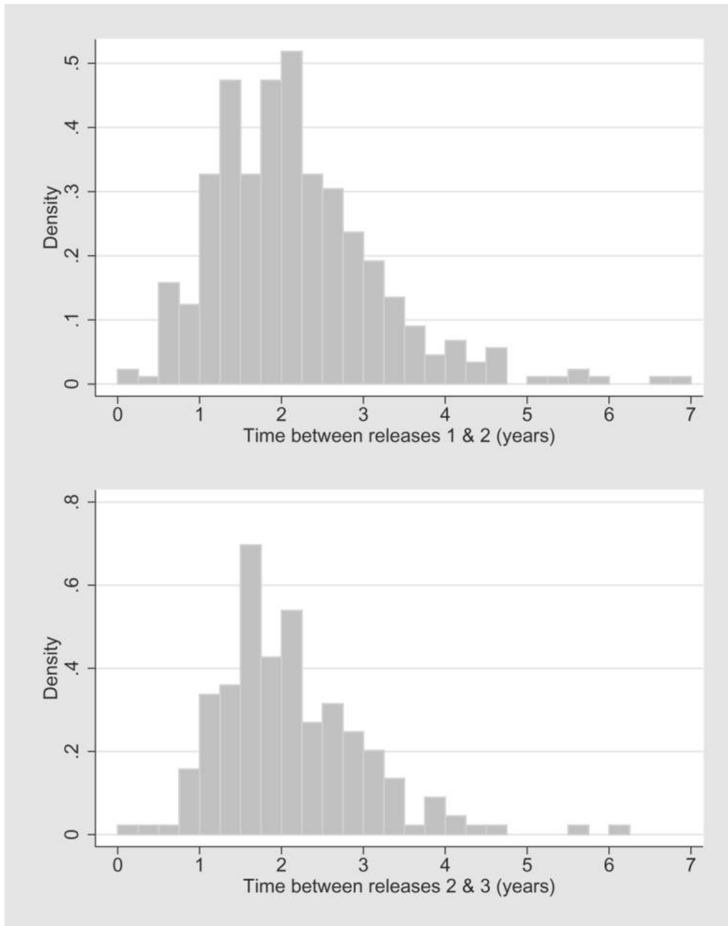

Fig. 2.—Distributions of elapsed time between releases. The upper panel plots the elapsed time between the releases of albums 1 and 2 by the 355 artists in our sample. The lower panel plots time between releases 2 and 3 for the 178 artists for whom we observe a third album.

come early in their careers, we do not see any obvious biases the selection will induce in the empirical analyses below. Moreover, a quick check of some out-of-sample data suggests that the selection bias is not very severe. We compiled a list of 927 artists who appeared on the Heatseekers chart between 1997 and 2002 but who are not included in our sample. Of these artists, 73 percent made it to the chart on their first or second album, as compared to 87 percent for the artists in our sample. The difference is qualitatively consistent with the selection problem described above, but we do not think that the difference is quantitatively large enough to undermine our main results.

III. Measuring the Spillovers

In this section we measure the backward spillovers and analyze how their magnitudes vary across artists. We use an empirical approach taken from the literature on treatment effects.[11] Our method exploits exogenous variation in albums' release times: a new album release by an artist is interpreted as the "treatment," and sales of "treated" artists are compared to the sales of control artists who have not yet released a new album. We follow the impact of a new release on sales of catalog albums for 39 weeks (13 pre- and 26 posttreatment) and refer to this period as the treatment "window."

A. Regression Model

In presenting the regression model, we focus on the first treatment episode: the release of album 2 and its impact on sales of album 1. Let y_{it}^{0} denote the log of album 1 sales of artist i in period t without treatment, and let y_{it}^{s} denote the log of album 1 sales in period t when artist i is in the sth period of treatment. For each artist, t indexes time since the debut album's release, not calendar time. By taking logs, we are implicitly assuming that treatment effects are proportional, not additive. There are two reasons for adopting this specification. One is that the distribution of album sales is highly skewed. The other is that the average treatment effect is likely to be nonlinear: a new release has a larger impact on total sales of catalog titles for more popular artists. By measuring the treatment effect in proportional terms, we capture some of this nonlinearity. However, it could bias our estimates of the treatment effects upward since proportionate effects are likely to be higher for less popular artists and there are many more of them. Proportionate effects may also be higher for popular artists who are treated later since their sales levels are likely to be a lot lower than popular artists who are treated earlier. We address these issues in discussing the results below.

Our objective is to estimate the average treatment effect on the treated (ATE) for each period of the treatment window. The ATE is simply the difference $y_{it}^{s} - y_{it}^{0}$. The main challenge in estimating the ATE is that, in each period, we observe only one outcome for each artist. Our approach to measuring this difference is to use the sales of not yet treated albums (i.e., albums whose artists have not yet released a second album) as the control group against which to compare sales of treated albums (i.e.,

[11] See Wooldridge (2002) for a summary.

albums whose artists have recently released a second album). Essentially, this approach assumes that for an album whose artist issues a new release at t, counterfactual sales (i.e., what sales would have been in the absence of the new release) can be inferred from the sales of all other albums at t for which there has not yet been a new release.

Our specific sampling and estimation procedure is as follows. Albums are included in the sample only until the last period of the treatment window: observations on sales after that window are not used in estimating the regressions. We adopt this approach to ensure that, at any given t, treated albums are being compared with not yet treated albums, rather than a mix of not yet treated and previously treated albums. Thus, the sample in period t includes artists who have not yet released a new album and artists who had a new release in periods $t-1$, $t-2$, …, or $t-S+1$ but excludes artists whose new release occurred prior to period $t-S+1$. Basically, we want the control group to measure what happens to sales over time before any new album is released.[12]

The regression model is as follows:

$$y_{it} = \alpha_0 + \alpha_i + \lambda_t + \sum_{m=2}^{12} \delta_m D_{it}^m + \sum_{s=-13}^{25} \beta_s I_{it}^s + \epsilon_{it}, \tag{1}$$

where α_i is an artist fixed effect, the λ_t's are time dummies, and the D^m's are month of the year dummies (to control for seasonality).[13] Here I_{it}^s is an indicator equal to one if the release of artist i's new album was s weeks away from period t, so β_s measures the new album's sales impact in week s of the treatment window ($t = 0$ corresponds to the first week following the new release). Intuitively, after accounting for time and artist fixed effects, we compute the difference in the average sales of album 1 between artists in treatment period s and artists who are not treated for each period and then average these differences across the time periods. The stochastic error, ϵ_{it}, is assumed to be heteroskedastic across i (some artists' sales are more volatile than others') and autocorrelated within i (random shocks to an artist's sales are persistent over time). The time dummies (λ_t) allow for a flexible decay path of sales, but implicitly we are assuming that the shape of this decay path is the same across albums. Although differences in the level of demand are captured by the album fixed effects, differences in the shapes of albums' sales paths are necessarily part of the error (ϵ).

Including separate indicators for successive weeks of treatment allows

[12] We believe that dropping posttreatment observations is the most appropriate approach, but it turns out not to matter very much: our estimates change very little if we include these observations.

[13] The results reported below are essentially unchanged if we control for seasonality with week of the year dummies instead of month of the year dummies.

us to check whether the new release's impact diminishes (or even reverses) over time, which is important for determining whether the effects reflect intertemporal demand shifts. We allow for a 39-week treatment window, beginning 13 weeks (3 months) before the release of the new album. The prerelease periods are included for two reasons. First, much of the promotional activity surrounding the release of a new album occurs in the weeks leading up to the release, and we want to allow for the possibility that the backward spillover reflects consumers' responses to these prerelease marketing campaigns. In some cases labels release singles from the new album in advance of the album itself, so that prerelease effects could also reflect advance airplay of the album's songs.[14] Second, including prerelease dummies serves as a reality check: we consider it rather implausible that a new album could have an impact on prior albums' sales many months in advance of its actual release, so if the estimated effects of the prerelease dummies are statistical zeros for months far enough back, we can interpret this as an indirect validation of our empirical model.

For the regression described above to yield consistent estimates of the treatment effect, the critical assumption is that the treatment indicators in a period are independent of the idiosyncratic sales shocks in that period. In other words, after controlling for time-invariant characteristics such as genre and artist quality that affect the level of sales in each period, we need the treatment (i.e., the release of a new album) to be random across artists. This is a strong but not implausible assumption. We suspect that the main factor determining the time between releases is the creative process, which is arguably exogenous to time-varying factors. Developing new music requires ideas, coordination, and effort, all of which are subject to the vagaries of the artist's moods and incentives. Nevertheless, the specific question for our analysis is whether release times depend on the sales patterns of previous albums in ways that album fixed effects cannot control.

One possibility is that release times are related to the *shape* of the previous album's sales path. For example, albums of artists who spend relatively more effort promoting the current album in live tours and other engagements will tend to have "longer legs" (i.e., slower decline rates) and later release times than albums of artists who spend more time working on the new album. To check this, we estimated Cox pro-

[14] One might wonder whether the relevant event is the release of the single or the release of the album. Although we have data on when singles were released for sale, this does not correspond reliably with the timing of the release on the radio. Radio stations are given advance copies of albums to be played on the air, and a given single may be played on the radio long before it is released for sale in stores. Moreover, even when a single has been released in advance of the album, the label's promotional activity is still focused around the release date of the album.

portional hazard models with time to release as the dependent variable and various album and artist characteristics included as covariates. Somewhat surprisingly, the time it takes to release an artist's new album is essentially independent of the success of the prior album (as measured by first 6 months' sales) and of its decline rate, after conditioning on genre.[15] These results seem to validate our assumption that release times are exogenous—at least with respect to the level and rate of change in the prior album's sales. However, subtle relationships between sales path shapes and release times may still exist. If so, the potential problem is that our regression controls only for the average rate of decline in album sales, so our estimates of the treatment effect will be biased if deviations from that average are systematically related to release times.

In order to address this issue, we can estimate the regression model of equation (1) using the first difference of log sales as the dependent variable; that is, we estimate

$$\Delta y_{it} = \tilde{\alpha}_0 + \tilde{\alpha}_i + \tilde{\lambda}_t + \sum_{m=2}^{12} \tilde{\delta}_m D_{it}^m + \sum_{s=-13}^{25} \tilde{\beta}_s I_{it}^s + \tilde{\epsilon}_{it}, \tag{2}$$

where $\Delta y_{it} \equiv y_{it} - y_{it-1}$. This model estimates the impact of new releases on the percentage rate of *change* (from week to week) in previous albums' sales. The advantage of this specification is that heterogeneity in sales levels is still accounted for (the first differencing sweeps it out), and the fixed effects, $\tilde{\alpha}_i$, now control for unobserved heterogeneity in albums' decline rates. Taking this heterogeneity out of the error term mitigates concerns about the endogeneity of treatment with respect to the shape of an album's sales path.

B. *Spillover Estimates*

We estimate the regressions in (1) and (2) separately for each of three treatments: the impact of the second and third releases on sales of the debut album and the impact of the third release on sales of the second album. In constructing the samples for estimating the regression, we impose several restrictions. First, we exclude the first 8 months of albums' sales histories in order to avoid having to model heterogeneity in early time paths. Recall that although most albums peak very early and then decline monotonically, for some "sleeper" albums we do observe accelerating sales over the first few months. By starting our sample at 8 months, we ensure that the vast majority of albums have already reached their sales peaks, so that the λ_t's have a better chance of controlling for the decay dynamics. A second restriction involves truncating

[15] A table showing the detailed results of this exercise is included in a previous version of this paper (Hendricks and Sorensen 2006).

the other end of the sales histories: we exclude sales occurring more than 4 years beyond the relevant starting point. This means that if an artist's second album was released more than 4 years after the first, then that artist is not included in the estimation of the impact of second releases on first albums; and (similarly) if an artist's third release came more than 4 years after the second, then that artist is excluded from the regressions estimating the impact of album 3 on album 2.

Because the number of coefficients being estimated is so large, we summarize the estimates graphically rather than present them in a table.[16] Figure 3 shows the estimated effects (i.e., the $\hat{\beta}_s$'s) from specification (1), along with 95 percent confidence bands, for each of the album pairs. The confidence bands are based on standard errors that were corrected for heteroskedasticity across artists and serial correlation within artists. As can be seen in the figure, the estimates of the effects for each of the weeks following the release of a new album are always positive, substantive, and statistically significant. Since the dependent variable is the logarithm of sales, the coefficients for specification (1) can be interpreted as approximate percentage changes in sales resulting from the new release. The largest spillover is between albums 2 and 1, with estimates ranging between 40 percent and 55 percent. The spillover of album 3 onto album 1 is smaller, with estimates ranging roughly between 20 and 38 percent, and the spillover of album 3 onto album 2 is roughly between 15 and 35 percent. Figure 4 shows estimates from specification (2) (the first-differenced model). The solid line plots the cumulative impact implied by the estimated weekly coefficients from the first-differenced model (2), and the dashed lines indicate the 95 percent confidence bands.[17] The implied effects are qualitatively and quantitatively very similar to those obtained in the undifferenced regressions, which we interpret as reassuring evidence that our results are driven by real effects, not by subtle correlations between current sales flows and the timing of new releases.[18]

In each treatment episode, the estimated impact of the new album 3 months prior to its actual release is statistically indistinguishable from

[16] Tables with a complete listing of coefficients and standard errors are available on the authors' Web sites.

[17] Because calculating the cumulative impact requires summing coefficients in this specification, the error associated with the cumulative effect at time t reflects the errors of all coefficients up to time t. That is, cumulating the estimates means that the errors cumulate too. Consequently, the confidence bands widen over time.

[18] We also checked the robustness of the estimates by splitting the sample in each treatment on the basis of the median treatment time. As expected, the patterns are the same but the estimated effects are smaller for the albums that are treated early and larger for albums treated later. (This pattern makes sense because our model assumes that the effects are proportional: albums treated later will tend to have lower sales flows at the time of treatment, so the proportional impact of the new release will tend to be larger than for albums with high sales flows.) The estimates are always strongly significant.

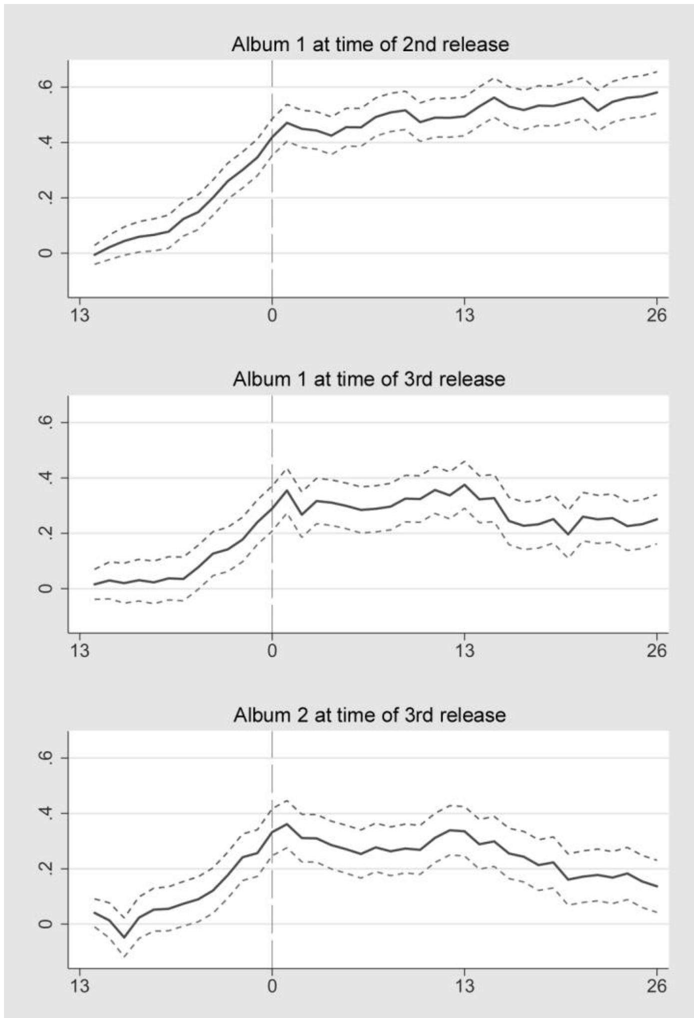

FIG. 3.—Time patterns of backward spillovers. These graphs plot the estimated coefficients from a regression of log sales on dummy variables indicating time relative to the release of the new album (eq. [1] in the text). So, for example, the coefficient at time 1 is the expected difference in sales of the catalog album during the first week of the new album's release. The dotted lines are 95 percent confidence intervals.

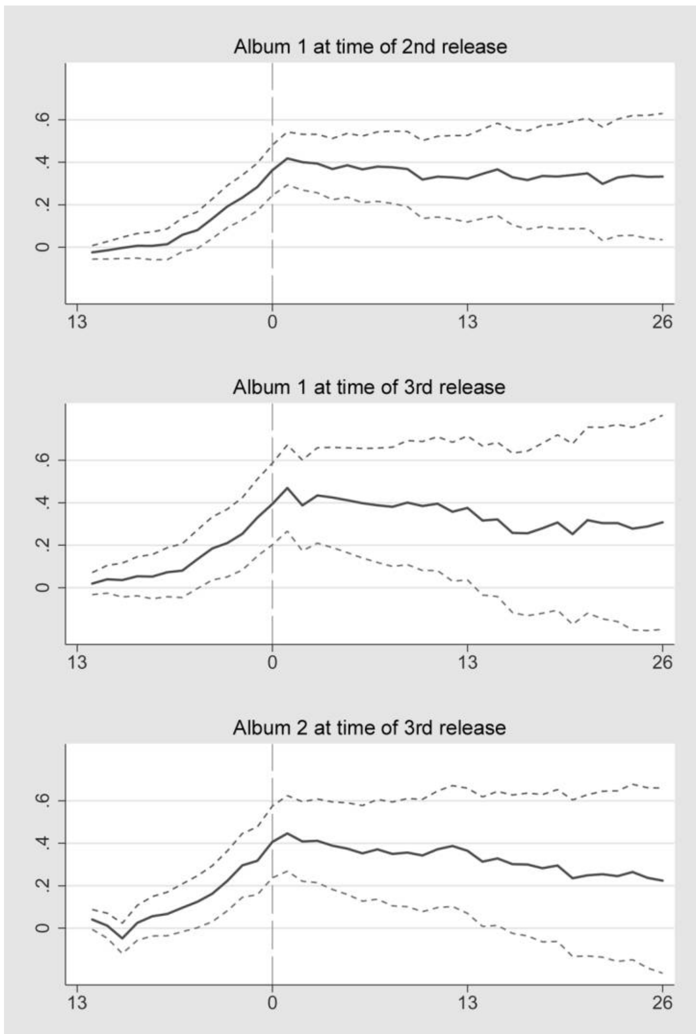

FIG. 4.—Time patterns of backward spillovers: first-differences model. These graphs are analogous to those in figure 1, except that the coefficients are estimated from the first-differenced model (eq. [2]). The confidence intervals (dotted lines) expand over time because the effect at time t is the sum of all coefficients up to time t; so the error associated with the cumulative effect at time t reflects the errors of all coefficients up to time t.

zero. As discussed above, this provides some reassurance about the model's assumptions: 3 months prior to the treatment, the sales of albums soon to be treated are statistically indistinguishable from those of control albums (after conditioning on album fixed effects and seasonal effects). In general, small (but statistically significant) increases start showing up 4–8 weeks prior to the new album's release, growing in magnitude until the week of the release ($t = 0$ in the table), at which point there is a substantial spike upward in sales.

The estimated effects are remarkably persistent: especially for the impact of album 2 on album 1, the spillovers do not appear to be transitory. If the spillover represents consumers who would have eventually purchased the catalog title anyway (i.e., even if the new album were never released), then the coefficients would decline and eventually would become negative. We have tried longer treatment windows. In some cases, the treatment effect does die out eventually, but in none of the cases does the treatment effect turn negative. It is important to note, however, that the increasing coefficients in some specifications do not imply ever-increasing sales paths, since the treatment effects in general do not dominate the underlying decay trend in sales. (In order to save space, the table does not list the estimated time dummies, which reveal a steady and almost perfectly monotonic decline over time.)

C. *Spillover Variation*

Although it is clear from our results that backward spillovers are significant, it is less clear why the spillovers occur. In this subsection we analyze variation in the magnitudes of the spillovers as a means of understanding their source. First, we split our sample on the basis of whether the albums were "hits" and examine how the backward spillover depends on the relative success of the new album vis-à-vis the catalog album. We define a hit as an album that sold 250,000 units or more in its first year; 30 percent of the albums in our sample meet this criterion.[19] We focus our attention on spillovers between adjacent albums and divide our sample into four categories: hits followed by hits, hits followed by nonhits, nonhits followed by hits, and nonhits followed by nonhits. We summarize the backward spillovers for each of the four categories in table 3.

The table is based on estimates of the regression model computed

[19] As a point of reference, the Recording Industry Association of America (RIAA) certifies albums as "gold" if they sell more than 500,000 units. Also, among the albums we categorize as hits, at least 90 percent had peak sales high enough to appear on Billboard's Top 200 chart (vs. less than 10 percent among those we categorize as nonhits).

TABLE 3
SPILLOVERS AND HITS

	Hit, Hit	Hit, Not	Not, Hit	Not, Not
	Album 1, Album 2			
Observations	53	45	34	206
Median number of weeks to release 2	108	124	101	104
Median weekly sales (album 1) prior to release	1,888	318	342	154
Median weekly decline around release	−.021	−.018	−.018	−.011
Estimated total change in sales	22,161	660	14,557	883
Percentage change in sales	42.7	7.2	148.5	17.6
Average of (sales before next release)/(first 4 years' sales)	.73	.85	.55	.62
	Album 2, Album 3			
Observations	49	13	12	99
Median number of weeks to release 3	105	117	95	103
Median weekly sales (album 1) prior to release	1,555	466	844	85
Median weekly decline around release	−.013	−.026	.004	−.010
Estimated total change in sales	19,884	1,110	20,788	687
Percentage change in sales	40.6	9.5	56.4	24.6
Average of (sales before next release)/(first 4 years' sales)	.73	.84	.59	.65

NOTE.—Hits are defined as albums that sold over 250,000 units nationally in the first year. Albums that did not clear this threshold are the "Not" albums (i.e., not hits). The estimated total changes and percentage changes in sales reflect increases over the 39-week treatment window.

separately for each subgroup.[20] These are then used to calculate the implied total change in sales for the "median" album. Specifically, we calculate the median weekly sales 14 weeks prior to the median release time and the median weekly decline over the 39 weeks that follow. (In these calculations, we use only albums whose artists have not yet released the next album, so that the median sales flows and median decline rates will not reflect any of the backward spillovers.) For example, in the group of 53 artists whose first two albums were both hits, the median time between the first and second releases is 108 weeks. Among first albums for which there was not yet a second release, the median weekly sales at week 94 ($= 108 - 14$) was 1,888, and the median decline rate over weeks 95–134 was 2.1 percent per week. So we take a hypothetical album, with weekly sales beginning at 1,888 and declining at 2.1 percent per week, and apply the percentage increases implied by our estimated

[20] We use the first-differences model in eq. (2). Some of the estimated sales increases are smaller if we estimate the model in levels, but the qualitative patterns are essentially the same.

coefficients. The predicted total increase in sales over the 39-week period is 22,161, or roughly $350,000 in additional revenues (using a retail price of $16 per unit).

The patterns in table 3 establish that the backward spillover is always larger when the new album is a hit, whether the previous album was a nonhit or a hit. The largest percentage increase occurs when a nonhit album is followed by a hit: for an artist whose second album was her first hit, we estimate that weekly sales of her first album more than double when the new album is released. The smallest increase occurs when a hit is followed by a nonhit. The same patterns hold when we examine the impact of the third release on the sales of album 2. The spillovers are large when the new album is a hit, but negligible otherwise. The numbers are slightly smaller than those for the previous album. An important lesson from table 3 is that although on average (across all types of albums) the backward spillovers are of modest economic significance, they are in fact quite large for the artists who matter: those who have hits or have the potential to produce hits.

In addition to splitting our sample to compare national sales across artists, we can also split the sample geographically to compare sales across markets for a given artist. An especially informative comparison is between an artist's home market (i.e., the city where the artist's career began) and other markets. Because new artists tend to have geographically limited concert tours—in many cases performing only in local clubs—artists in their early careers are more popular in their home markets.

We were able to determine the city of origin for 325 of the 339 artists included in the regression analyses summarized in figures 3 and 4; 268 of these artists originated in the United States, so we can observe sales in the home market and compare them to sales in other markets across the nation. SoundScan reports album sales separately for 100 designated market areas (DMAs), each one corresponding to a major metropolitan area such as Los Angeles or Boston. We determined each artist's city of origin and labeled the nearest DMA to be the artist's home market.[21] It is easy to verify that artists are indeed more popular in their home markets: over 80 percent of debut albums had disproportionately high sales in the artist's home market, meaning that the home market's share of national first-year sales was higher than the typical share for other artists of the same genre. On average, the home market's share of national sales was 8 percentage points larger than would have been pre-

[21] Roughly 20 percent of the artists are solo artists, and for these we were able to find only the city of birth—which is not necessarily the city in which the artist first began performing. However, it is plausible that solo artists are more well known in their birth cities than in other cities nationwide, even if they began their performing careers elsewhere. In any case, all our analyses deliver the same conclusions if we exclude solo artists.

TABLE 4
Sales and Spillovers in the Artist's Home Market

	$2 \rightarrow 1$	$3 \rightarrow 2$
Home market ($\hat{\psi}$)	.814	.647
	(.006)	(.008)
Home market × new release period ($\hat{\gamma}$)	−.105	−.104
	(.010)	(.013)
Observations	2,727,890	1,437,340
Number of artists	268	142

Note.—Estimates of the regression model described in eq. (3); the dependent variable is log sales. The term $\hat{\psi}$ measures the average difference in log sales between the artist's home market and other markets, and $\hat{\gamma}$ measures the average difference in the backward spillover in the artist's home market vs. other markets. Other coefficients are omitted to save space.

dicted on the basis of that market's share of overall sales within the artist's genre.

Are backward spillovers smaller in artists' home markets? Using the market-level data, we estimate a variant of the regression model in (1):

$$y_{imt} = \alpha_0 + \alpha_i + \sum_{g=1}^{4} \theta_{gm} G_i^g + \lambda_1 t + \lambda_2 t^2 + \psi H_{im} + \sum_{k=2}^{12} \delta_k D_{it}^k + \sum_{s=-13}^{26} I_{it}^s (\beta_s + \gamma H_{im}) + \epsilon_{imt}, \tag{3}$$

where y_{imt} is log sales of artist i's album in market m in week t, G_i^g is a dummy equal to one if artist i is in genre g (so the θ_{gm}'s are market × genre fixed effects), the D_{it}^k's are month of the year dummies, the I_{it}^s's are the treatment dummies, and H_{im} equals one if market m is artist i's home market. The key differences between this model and the one described in equation (1) are that (i) we use market-level sales data and control for heterogeneity in sales across markets using market × genre fixed effects,[22] (ii) we measure whether sales are on average higher in the artist's home market (i.e., the parameter ψ), and (iii) we allow the spillover effects to differ for home markets versus other markets (via the parameter γ).

Table 4 reports the results. The estimates of ψ confirm that on average sales are much higher in an artist's home market than in other markets. For the debut album, the coefficient of 0.814 implies that sales are over twice as high in the home market as in other markets, other things being equal. Notably, the home market advantage is smaller for later albums. Also, in spite of the fact that artists' albums are on average

[22] Note that we can alternatively include market × artist fixed effects. Doing so means that we cannot estimate ψ, the coefficient on H_{im}, because H_{im} is collinear with the market × artist effect for the home market. Adopting this specification yields results for all the other parameters that are virtually identical to those we report for the model with market × genre effects.

more successful in their home markets, the backward spillovers are on average smaller in home markets. The estimates of γ are similar across the album pairs, indicating that backward spillovers are 10–14 percentage points smaller in an artist's home market than in other markets.

D. *Summary*

The analysis of this section has established several facts about the backward spillover: (1) it starts to appear several weeks prior to the release of the new album and increases throughout the prerelease period; (2) it peaks in the week of the release and thereafter remains roughly constant as a percentage of sales, implying that the release of a new album generates permanent increases in demand for past albums, inducing purchases by customers who would not have otherwise purchased; (3) it is large and economically significant when the new release is a hit; (4) it is large when the catalog album was a hit but especially large (in percentage terms) when the catalog album was not a hit; and (5) it is smaller as a percentage of sales in the artist's home market, even though sales are on average substantially higher in the home market.

We do not have price data for the albums in our sample, but it is clear that these facts do not reflect price changes. Variation in price across titles and over time is very limited, and although discounts are occasionally "pushed down" to the retail level by distributors, these discounts are usually for new albums rather than catalog titles. According to two retail store managers with whom we had conversations, even when catalog albums are discounted, the timing of the sales is not systematically related to new releases by the same artist.[23]

The above facts are easily explained if the backward spillover results from changes in consumers' information, that is, if the new release causes some consumers to discover the artist and then purchase the artist's previous albums. The prerelease effects most likely reflect promotional activity and radio airplay that occur prior to the new album's release in stores.[24] The postrelease pattern can be explained by an arrival

[23] In a previous version we reported price data for a sample of CDs offered at a major online retailer. Comparing prices for three groups of albums—new releases, catalog titles by artists with new releases, and catalog titles by artists without new releases—we found that although new releases tended to be discounted, the price distributions for the other two groups were indistinguishable. Catalog titles by artists who recently released a new album were no more likely to be discounted than other catalog titles.

[24] As a simple check, we estimated the regressions separately for artists who released singles in advance of the new album (23 percent of the sample) vs. artists who released singles after the release of the album (16 percent of the sample). The prerelease effects on the catalog album's sales are much larger for the artists with advance singles: artists with pre–album release singles had roughly 30–40 percent sales increases of album 1 in the 3 weeks prior to the release of album 2, whereas artists with post–album release singles had increases of 5–20 percent. This pattern suggests that the prerelease sales increases

rate of consumers for the new release that is proportional to the stock of consumers who do not know about the catalog album, which falls over the release period. The magnitude of the spillover then depends on the capacity of the new release to draw consumers to the artist's market, which in turn depends on how well the new release does and on the number of consumers who do not know about the artist. The spillover is larger when the new release is a hit because consumers are more likely to discover the artist. The variation in the spillover of a hit release with the success of the catalog album reflects the interaction of two effects: the number of consumers who do not know about the artist is lower when the catalog was a hit, but the probability of purchase conditional on artist discovery is higher. The fact that spillovers are smaller in artists' home markets can be attributed to home markets having a larger percentage of consumers who are already familiar with the artist. Also, the fact that the home market advantage is smaller for later albums is consistent with the notion that awareness of the artist becomes less geographically concentrated as the artist's career progresses.

Other explanations of the backward spillover are certainly possible. For example, it is possible that new releases directly affect consumers' preferences for catalog albums. We will return to discuss alternative explanations in Section IV.C. In the next section, however, we pursue a more formal examination of the album discovery explanation and its implications.

IV. An Album Discovery Model

In this section, we develop and estimate a model of album discovery. The model is based on the simple idea that consumers must know about an album before they can purchase it. Hence, the probability that a consumer purchases an album is the product of two probabilities: the probability that she likes the album conditional on discovering the album and the probability that she discovers the album. In the population, these probabilities correspond to the proportion of consumers who discovered the album and the proportion of these consumers who liked the album. Obviously, neither of these proportions is directly observable in the data. Our model assumes that preferences are fixed and discovery is random, so the proportion of consumers who liked the album conditional on discovery differences out when comparing sales of the same album across release periods of later albums. We specify a functional form for the probability that a consumer discovers an album during a

come primarily from consumers who discover the artist as a result of prerelease promotion and airplay.

release period. This function determines the proportion of consumers who discover the album in each release period. The parameters of the function are estimated using data on the impact of an artist's second release on sales of the debut album. On the basis of the estimates, we forecast the 3–1 and 3–2 spillovers and construct tests of the model. We then use the estimated discovery probability function to conduct counterfactual analyses.

A. *Model*

Each artist releases three albums, $k = 1, 2, 3$, sequentially in release periods $t = 1, 2, 3$. In contrast to the previous section, where we analyzed weekly sales dynamics, in this section we are interested in cumulative sales of an album during a release period. Our primary focus is on the debut album and demand for the debut album in release periods 2 and 3. In each of these periods, there are three types of consumers: those who discovered the debut album and purchased it in a prior release period, those who discovered the album and have not yet purchased it, and those who do not know about the album. We refer to the first two groups as "informed" and the last group as "uninformed." In what follows, we assume that a nonpurchase by an informed consumer means that she does not care for the album and will not buy it in later periods. Given this assumption, the backward spillover must be generated by the uninformed consumers. The size of the spillover is determined by (i) the number of uninformed consumers, (ii) the proportion of these consumers who discover the debut album as a result of the new release, and (iii) the proportion of these consumers who like the debut album enough to buy it. These are the key unobservables in the model.

The assumption that most of the spillovers are generated by consumers who have not yet discovered the album seems to us to be a good approximation. Albums are search goods: at low cost, consumers can learn their preferences for an album. They typically learn about an album for free by hearing selected songs played on the radio. Upon hearing the songs, most consumers know whether they like it enough to buy the album or dislike it enough to not buy the album.[25] Those who remain undecided can always learn more by listening to the album online or from friends or at listening posts in record stores. Thus, the binary nature of information in our model—either you know your pref-

[25] The fact that consumers rarely experience ex post regret provides further support for the view that music albums can be modeled as search goods. In a survey of music buyers conducted by Rob and Waldfogel (2006), respondents reported being "disappointed from the start" for only 9.5 percent of the albums they purchased. The level of ex post regret seems to be higher for books or movies, which are arguably more like experience goods.

erences for an album or you do not—seems like a reasonable simplification. We discuss the implications of relaxing this assumption in the next subsection.

The backward spillover arises because the release of a new album generates information about the artist and leads some of the uninformed consumers to discover the artist's catalog albums. In particular, the likelihood that an uninformed consumer i discovers the debut album in period t is assumed to depend on the sales of the album released in period t. The rationale for this is that consumers learn about music primarily through radio airplay,[26] and most radio airplay is devoted to songs from an artist's newest album.[27] Since radio stations want to play albums that are popular, they allocate playing time on the basis of albums' expected sales. Formally, let $I^i_{1,t}$ denote a binary variable that is equal to one if consumer i learns about album 1 in period t (conditional on not having learned in any prior period). We write the probability of discovery as

$$\Pr\{I^i_{1,t} = 1\} = \frac{ae^{bS_{t,t}}}{(1-a) + ae^{bS_{t,t}}},$$

where $S_{t,t}$ is sales of the new release in period t.[28] If the new album is expected to sell very little, then playing time is very low and the learning probability is a. As sales get very large, the probability converges to one, at a rate that depends on the parameter b.[29] We assume that $\{I^i_{1,t}\}$ are independent random variables across consumers and release periods; the law of large numbers then implies that proportions of consumers in a large population converge to the associated probabilities.

[26] In the Soundata National Music Consumer Survey conducted in 1994 by the National Association of Recording Merchandisers, consumers were asked what motivated their recent music purchases, and the most common response was having heard the music on the radio. A more recent survey in 2006 (reported by the Associated Press [2006]) produced a similar finding: 55 percent of consumers said they learn about new music primarily from FM radio.

[27] For example, we checked the Billboard Hot 100 Airplay chart for July 17, 1999, and found that 74 of the 75 listed songs were from the respective artist's newest album. However, even if the release of a new album leads to increased promotion or airplay of the artist's old albums, this is consistent with our assumption as long as the increase is proportional to expected sales of the new album.

[28] For simplicity, we assume that radio stations have perfect foresight, so expected sales are just equal to actual sales.

[29] Notice that we are assuming that the parameters of the discovery probability (a and b) are the same in every release period. We make this assumption primarily for convenience. However, we tested this assumption by estimating models in which a was allowed to be different in period 1 than in period 2 or in which b was allowed to be different in period 1 than in period 2. In neither case could we reject the hypothesis that the parameter is equal across periods.

The proportion of informed consumers for album 1 in period t accumulates according to the equation

$$q_{1,t} = q_{1,t-1} + (1 - q_{1,t-1}) \frac{ae^{bS_{1,t}}}{(1-a) + ae^{bS_{1,t}}}, \tag{4}$$

where $q_{1,t}$ denotes the proportion of consumers who know their preferences for album 1 at the end of period t. For period 1, we set $q_{1,0} \equiv q_0$, so q_0 is interpreted as the baseline awareness of the artist prior to her debut.

The probability that consumer i purchases album 1 in period t conditional on discovering the album is simply denoted p_1. We make the critical assumption that a consumer's utility for album 1 does not change over the release periods. Because the choice set is changing over the release periods, this assumption requires preferences to be additive across albums. Additivity is a strong assumption, but it is testable, as we explain below. Notice that we are also implicitly assuming that the discovery probability is independent of the consumer's preferences.

Since preferences are assumed not to change across release periods, spillover sales reflect changes in the number of informed consumers: $(q_{1,t} - q_{1,t-1})N$, where N is the number of potential consumers for the album. Appealing to the law of large numbers, sales of album 1 in period $t > 1$ are given by

$$S_{1,t} = p_1(q_{1,t} - q_{1,t-1})N. \tag{5}$$

Sales of album 1 in its own release period are simply $S_{1,1} \equiv p_1 q_{1,1} N$. Since $q_{1,1}$ is a function of $S_{1,1}$ (as indicated in eq. [4]), sales are reinforcing: higher sales lead to more consumers discovering the album, which further increases sales.[30] This, along with the fact that p_1 is unobservable, makes it difficult to estimate the parameters of the model using only data on sales of the album in its own release period.

Instead, we estimate the parameters of the model using the spillover of album 2 onto album 1. Specifically, our estimation exploits the comparison of album 1 sales in release period 2 to album 1 sales in release period 1. There are two main reasons for this approach. First, sales of the new album in release period 2 shift consumer awareness ($q_{1,2}$) exogenously (conditional on sales of album 1 in period 1). Second, the

[30] It is straightforward to show that a solution (i.e., fixed point) to this sales relationship always exists, that there are either one or three solutions (generically), and that the minimum and maximum solutions are increasing in album quality (i.e., p_k). Multiple equilibria can arise because of the logistic learning curve and the lack of coordination among radio stations in choosing playing time. However, for the learning curve we estimate below, it turns out that the fixed point is unique for every album.

comparison allows us to eliminate unobservable p_1 from the model: substituting $p_1 = S_{1,1}/q_{1,1}N$ into equation (5), we obtain

$$\frac{S_{1,2}}{S_{1,1}} = \frac{q_{1,2} - q_{1,1}}{q_{1,1}} = \frac{1 - q_{1,1}}{q_{1,1}} \frac{ae^{bS_{2,2}}}{(1-a) + ae^{bS_{2,2}}}. \tag{6}$$

Notice that N, the (unknown) number of potential consumers for album 1, is also eliminated in this step. The remaining unobservable is $q_{1,1}$, the fraction of consumers who know their preferences for album 1 in release period 1. When we substitute for this variable (using eq. [4]) and take logs, the spillover equation becomes

$$\log \frac{S_{1,2}}{S_{1,1}} = \log\left[\frac{(1-q_0)(1-a)}{q_0(1-a) + ae^{bS_{1,1}}}\right] + \log(a) + bS_{2,2} - \log(1 - a + ae^{bS_{2,2}}). \tag{7}$$

In examining the backward spillovers, we noticed that their magnitude appears to decrease as a function of time between releases. To accommodate this feature of the data, we make an ad hoc modification to equation (7) that allows for depreciation. We assume that mean utility declines over time at a rate that is common across albums, writing the purchase probability for a consumer who learns her preferences for album 1 in period $t > 1$ as

$$p_{1,t} = p_1 e^{-\gamma T_{1,t}}, \tag{8}$$

where $T_{1,t}$ is the length of time between the release of album 1 and the release of album t. This specification allows consumers to have a taste for "newness" and for the spillover to decline as a function of time between releases. From equation (8), the spillover equation that we take to the data is

$$\log \frac{S_{1,2}}{S_{1,1}} = \log\left[\frac{(1-q_0)(1-a)}{q_0(1-a) + ae^{bS_{1,1}}}\right] - \gamma T_{1,2} + \log(a) + bS_{2,2} - \log(1 - a + ae^{bS_{2,2}}) + \eta, \tag{9}$$

where η is the error term.

It will be convenient to standardize the length of the period over which to measure albums' sales. We calculate sales over a 1-year period: $S_{1,1}$ is measured as first-year sales of album 1, and $S_{1,2}$ is measured as cumulative sales of album 1 during the first year of release 2. The definition of a release period as 1 year is long enough for the sales dynamics to have run their course: almost anyone who was going to learn about the new release and buy it before the release of the next album will have done so within the first year. It introduces some measurement error into the model, since the fraction of informed consumers at the end of an album's first year is not the same as the fraction of informed consumers at the time of the next release. However, the error is small: on average, first-year sales represent 85 percent of cumulative sales at the time of the next release.[31] Time between releases ($T_{1,2}$) is measured from the end of the first year of album 1's release to the beginning of the first year of album 2's release.

Given the above definition of release periods, the estimating equation generated by our model is essentially a nonlinear regression of spillover sales ($S_{1,2}$) on first-year sales of the catalog album ($S_{1,1}$) and first-year sales of the new release ($S_{2,2}$). Straightforward estimation methods will yield unbiased estimates of the learning parameters as long as the error term (η), which represents approximation errors and unobserved shocks to spillover sales, is orthogonal to $S_{1,1}$ and $S_{2,2}$. This condition is almost certainly true for $S_{1,1}$, since the debut album's first-year sales are predetermined at the time of the new album's release, and typically the time between release periods is many months. Indeed, by the time the new album is released, sales of the catalog album are typically flat. It is less obvious that η is orthogonal to $S_{2,2}$, since unmeasured promotional activities that occur after the new release (radio airplay, television appearances, concert tours, etc.) will tend to increase sales of both the new and catalog albums. However, as long as these activities are a direct result of the new album release, they represent exactly the effects $S_{2,2}$ is supposed to capture as a proxy. Only promotional activities that would have occurred irrespective of the new release would be problematic from an econometric standpoint, since they could generate a spurious positive correlation between $S_{1,2}$ and $S_{2,2}$. We suspect that this issue is unimportant, however, since airplay and other promotions almost always focus on the artist's new release.[32]

[31] An alternative approach would be to let the release periods be artist-specific, measuring $S_{1,2}$ as cumulative sales of album 1 all the way up to the time of album 2's release. However, it turns out not to make a meaningful difference, precisely because album sales after the first year are so low. We also tried defining release periods as consisting of the first 6 or 18 months (instead of 12) after an album's release and found that our main results were largely unaffected.

[32] Another possible endogeneity problem that we considered is the timing of Christmas effects. All sales variables include a Christmas effect, because we measure sales over a 1-year period. However, if the new album is released just prior to Christmas, the holiday sales spike may be larger (for both the new album and the catalog album) than it would be for an album released many months after Christmas, which could generate a spurious positive correlation between $S_{1,2}$ and $S_{2,2}$. This does not appear to be empirically important, however: including controls for the season of the new album's release does not meaningfully change our estimates of the learning parameters in eq. (9).

TABLE 5
Estimated Parameters of Album Discovery Model

	National Sales (1)	DMA-Level Sales: Home Market (2)	DMA-Level Sales: Nonhome Markets (3)
“Baseline” awareness: q_0	.180	.305	.143
	(.049)	(.052)	(.045)
Learning function parameters:			
a	.161	.284	.144
	(.072)	(.121)	(.061)
b	.065	.112	.098
	(.013)	(.027)	(.018)
Time between releases: γ	.701	.567	.697
	(.073)	(.086)	(.086)
Observations	311	247	247
R^2	.288	.214	.274

Note.—Asymptotic standard errors are in parentheses.

B. *Results*

Before reporting our parameter estimates, we explain briefly how they are identified by the data. Our estimate of b is driven by the sensitivity of the backward spillovers to sales of the new album. As shown in table 3 above, backward spillovers are significantly larger when the new release is successful, so we should expect a positive estimate of b. If instead spillovers were invariant to the success of the new release, then we would expect our estimate of b to be close to zero. We also found that sometimes backward spillovers occur even when the new release sells very little. The observed magnitude of backward spillovers in such cases identifies a, the baseline flow of learning. If a were zero, we would expect backward spillovers only when the new release is successful; the higher a is, the larger the spillovers will be even in instances in which the new release is a dud. The q_0 parameter is identified by the average magnitudes of backward spillovers. If q_0 is near zero, the model allows for large spillovers that may depend on sales of the new album (through the a and b parameters); if instead q_0 approaches one, the model predicts very small spillovers no matter how strong the sales of the new album. Notice that unlike the a parameter, q_0 does not interact with sales of the new album. This is what allows the two parameters to be separately identified, in spite of serving similar purposes in the model. Finally, γ is identified by the extent to which spillovers tend to be smaller when the time between releases is longer. (This pattern in the data is precisely what motivated the inclusion of the γ term.)

Column 1 of table 5 reports nonlinear least squares estimates of equation (9) based on national sales.[33] Our estimate of q_0 implies that on average 18 percent of potential buyers are aware of an artist before she releases her debut album. Although this number may seem somewhat high, it is not implausible given that most artists tour extensively (playing small concerts in clubs or performing as the opening act for a larger band) before ever releasing an album. In columns 2 and 3 of the table, we use the DMA-level data to estimate equation (9) separately in the artists' home markets versus nonhome markets, respectively. Not surprisingly, the estimate of q_0 is considerably higher in the home market: since most artists begin their careers by playing small concerts in their home cities, we should expect a higher level of baseline awareness in the home market. We read this comparison as offering basic support for our interpretation of the model's parameters. Of course, it also suggests that there may be other interesting sources of heterogeneity in learning across markets. In the discussion that follows, however, we simply focus on the estimates based on national sales, leaving market-level heterogeneity as an issue to explore more fully in future work.

Figure 5 illustrates the discovery function defined by column 1 of table 5. Its shape is determined by the parameters a and b, with a representing the baseline learning rate and b representing the rate at which learning increases with sales. Initially, the probability of discovery increases at an increasing rate as a function of sales, but eventually the function becomes concave and the fraction of informed consumers approaches one. The inflection point is at 2.54 million sales. As noted above, the logistic learning curve can potentially give rise to multiple equilibria; however, this turns out to be irrelevant given our parameter estimates.[34]

Our estimates imply that learning is nearly complete for artists with extremely successful albums. For example, an artist whose debut album sells 10 million copies—which would classify it as a huge hit and earn it the RIAA "diamond" award—would be known to 99 percent of consumers.[35] At the other end of the success spectrum, the majority of consumers remain uninformed: if a debut album sells fewer than 500,000 copies in the first year, our estimates suggest that only a third

[33] The sample size is slightly smaller than we used to obtain the estimates shown in figs. 3 and 4. In this model we are aggregating sales over time and defining the sales periods to be 1 year. The difference in sample size arises because we exclude artists for whom the 1-year release periods overlapped, i.e., for whom the new release came between 8 and 12 months after the previous release.

[34] Owing in particular to the relatively high estimated values of q_0 and a, for the albums in our sample the relationship described by $S_{1,t} = p_1 q_{t,t} N$ has only one fixed point.

[35] This need not mean 99 percent of all consumers, but rather 99 percent of the relevant population of consumers, which presumably consists of those consumers who could potentially be exposed to information about the album. The model implicitly defines the size of the market by the point at which there can be no backward spillovers.

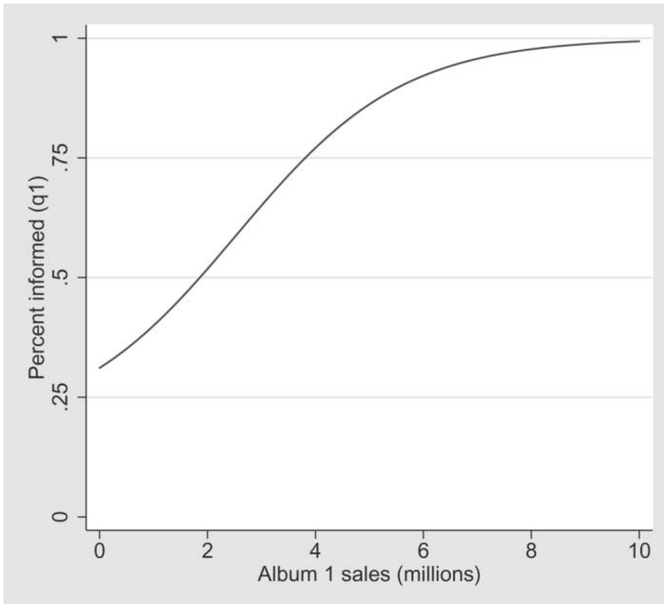

FIG. 5.—Estimated album discovery function. This graph shows our estimate of equation (4) for the debut album, that is, the fraction of the relevant population of consumers who discover album 1, as a function of its sales. For an album with sales near zero, we estimate that roughly 30 percent of consumers will know about it. For an album with sales over 8 million, we estimate that roughly 98 percent of consumers know about it.

of potential consumers will have discovered the artist in that year. Note that the most successful debut album in our data set sold roughly 8.2 million copies in its first year, so the graph shown does not extrapolate far beyond the range of the data.

The estimated value of a suggests that with each new album released, at least 16 percent of previously uninformed consumers will learn their preferences for the catalog album even if the new album has zero sales. Because learning is cumulative in our model, this implies (somewhat counterintuitively) that an artist could become a household name by releasing a long sequence of very low-quality albums. However, the numbers imply that such an artist would need to release 13 such albums before 90 percent of consumers would become aware of the first album. By contrast, a successful artist can become famous with only two or three hit albums. For example, after a sequence of three "triple-platinum" albums (sales of 3 million each), 94 percent of consumers would know their preferences for the first such album.

As mentioned above, the model assumes that preferences are additive. In particular, preferences for a given album do not depend on the existence or characteristics of other albums, even by the same artist. We can test this assumption by looking at the spillover sales of album 1 after the release of album 3. If discovery is uncorrelated with preferences and preferences do not change over time, additivity implies that the fraction of new consumers who buy album 1 in period 2 is the same as the fraction who buy album 1 in period 3 (controlling for age effects as indexed by $\hat{\gamma}$). Specifically, the relationship between the two spillovers is given by

$$\log \frac{S_{1,2}}{S_{1,3}} = \hat{\gamma}(T_{1,3} - T_{1,2}) + \log\left(\frac{q_{1,2} - q_{1,1}}{q_{1,3} - q_{1,2}}\right). \tag{10}$$

Given parameter estimates ($\hat{q}_0$, $\hat{a}$, $\hat{b}$, and $\hat{\gamma}$), we can calculate the fraction of consumers who knew their preferences for album 1 at the ends of periods 1, 2, and 3 (i.e., $\hat{q}_{1,1}$, $\hat{q}_{1,2}$, and $\hat{q}_{1,3}$, respectively) and plug these in to the right-hand side of equation (10) to obtain a prediction for log $(S_{1,2}/S_{1,3})$. Comparing our predictions to what we observe in the data, we find that the differences are on average positive. (The average difference is .291, with a standard error of .089.) In other words, we tend to slightly overpredict the size of the 3-on-1 spillover. This suggests that, contrary to our assumption, preferences might be subadditive (i.e., albums by the same artist are substitutes for one another). It might also indicate the presence of a weak selection effect: the consumers who are most likely to buy an album may tend to hear it sooner, leading to a decline over time in the probability of purchase conditional on becoming informed. We return to a discussion of these issues in subsection C below.

1. Forward Spillovers

In studying the backward spillover, we focused on the set of uninformed consumers who discovered album 1 in periods 2 and 3 from hearing the new albums released in those periods. We turn now to the set of informed consumers (i.e., consumers who discovered album 1 in period 1) and ask whether they are more likely to know about album 2 when it is released. The probability that these consumers—particularly those who liked album 1—learn about album 2 is potentially much higher than for consumers who do not know about the artist. If this is the case, then sales of album 2 will depend on the success of album 1.

Testing for the presence of a forward spillover is complicated by the fact that sales of album 2 depend not only on the stock of informed consumers but also on the probability of purchase, which is a function of unobservable preferences. However, we can use the same trick as above to difference out the unobserved preference probability for album

2, p_2, this time using the spillover sales of album 2 in period 3. When we substitute album 2 for album 1 in equations (4) and (5), the spillover equation for album 2 in period 3 is given by

$$\log \frac{S_{2,3}}{S_{2,2}} = \log\left[\frac{(1-q_{2,1})(1-a)}{q_{2,1}(1-a)+ae^{bS_{2,2}}}\right] - \gamma T_{2,3} + \log(a) + bS_{3,3} - \log(1-a+ae^{bS_{3,3}}), \tag{11}$$

where $T_{2,3}$ is measured from the end of the first year of album 2's release to the beginning of the first year of album 3's release. Here $q_{2,1}$ is the baseline awareness of album 2 at the beginning of period 2 (analogous to q_0 from the spillover equation for album 1 in period 2). To test for the presence of a forward spillover, we use our estimates $\hat{a}$, $\hat{b}$, and $\hat{\gamma}$ (as reported in table 5) to compute a separate estimate of initial awareness ($q_{2,1}$) for each album and then check to see if these estimates depend positively on the success of the artist's first album. That is, for each artist's second album, we solve equation (11) for $q_{2,1}$ and then see if these $q_{2,1}$'s are positively correlated with $S_{1,1}$.

The results of this test clearly indicate the presence of a forward spillover. The average implied initial awareness ($q_{2,1}$) is .362, which is double the baseline awareness we estimate for debut albums. Also, the estimates are positively correlated with sales of the artist's debut album, $S_{1,1}$. (The simple correlation is .201 and is statistically significant at the 1 percent level.)

In fact, we can take this test one step further and ask whether the fraction of consumers who are aware of album 2 at the beginning of period 2 ($q_{2,1}$) is simply equal to the fraction of consumers who learned about album 1 in period 1 ($q_{1,1}$). This would mean that consumers who discovered album 1 in period 1 and bought it are on the lookout for the artist's second album and know about it when it is released. Consumers who discovered album 1 in period 1 and did not like it are not interested in buying album 2. Comparing the values of $q_{2,1}$ implied by the 3-on-2 spillover (eq. [11]) to the values of $q_{1,1}$ estimated from the 2-on-1 spillover (eq. [9]),[36] we fail to reject that the two are equal. (The average difference is .015, with a standard error of .019.)

With the assumption that $q_{2,1} = q_{1,1}$, we can use our estimates of the learning model to calculate the sales that artists' second albums would have garnered in the absence of a forward spillover, that is, if the second

[36] We calculate the estimate of $q_{1,1}$ as

$$\hat{q}_{1,1} = \hat{q}_0 + (1-\hat{q}_0)\frac{\hat{a}e^{\hat{b}S_{1,1}}}{1-\hat{a}+\hat{a}e^{\hat{b}S_{1,1}}}.$$

albums had instead been the debut albums. Abusing notation somewhat, let $\tilde{S}_2$ denote the counterfactual sales of album 2 if it had instead been the debut album, and let $\tilde{q}_{1,1}$ be the fraction of consumers who would have discovered album 2 in that case. Then

$$\tilde{S}_2 = p_2\tilde{q}_{1,1}N = S_{2,2}\frac{\tilde{q}_{1,1}}{q_{2,2}}, \tag{12}$$

where the second equality follows from the fact that $S_{2,2} = p_2q_{2,2}N$. The equation states that we can estimate counterfactual sales of album 2 by simply rescaling the observed sales of album 2 by a factor equal to the ratio of $\tilde{q}_{1,1}$ to $q_{2,2}$.

Since $\tilde{q}_{1,1}$ is itself a function of $\tilde{S}_2$, we calculate $\tilde{S}_2$ by finding the root of equation (12), substituting in our estimates of a, b, and q_0 wherever those parameters appear.[37] The results imply that forward spillovers have a substantial impact on sales. The median difference between $S_{2,2}$ (observed album 2 sales) and $\tilde{S}_2$ (predicted sales of album 2 if it had instead been the debut album) is 16,450. The largest differences, which occur for artists whose first albums were big hits and whose second albums were smaller hits, are over 1 million. We estimate that the artists in our sample collectively sold 29.48 million more units on their second albums than they would have if those albums were debuts—a difference of roughly 25 percent.

2. "Lost Sales" and the Skewness of Sales

How many sales does an album lose as a result of consumers' lack of information about that album? Our estimates allow us to calculate counterfactual sales under the assumption that all consumers are fully informed about a given album, with the consumers' information about all other albums in the choice set held fixed. Specifically, letting $\tilde{S}_{1,1}$ denote counterfactual (complete information) sales, for each album we calculate

$$\tilde{S}_{1,1} = p_1N = \frac{S_{1,1}}{\hat{q}_{1,1}},$$

[37] The equation for $\tilde{q}_{1,1}$ is

$$\tilde{q}_{1,1} = \hat{q}_0 + (1-\hat{q}_0)\frac{\hat{a}e^{\hat{b}\tilde{S}_2}}{1-\hat{a}+\hat{a}e^{\hat{b}\tilde{S}_2}},$$

and the equation for $q_{2,2}$ is

$$q_{2,2} = \hat{q}_{1,1} + (1-\hat{q}_{1,1})\frac{\hat{a}e^{\hat{b}S_{2,2}}}{1-\hat{a}+\hat{a}e^{\hat{b}S_{2,2}}}.$$

(The assumption that $q_{2,1} = q_{1,1}$ is imposed in the second equation.)

TABLE 6
COUNTERFACTUAL SALES UNDER FULL INFORMATION

Percentile	Artist and Title	Observed Sales	Sales If $q_{1,1} = 1$	Difference
Max	Alanis Morissette, *Jagged Little Pill*	8,204,835	8,378,094	173,259
.90	Coolio, *It Takes a Thief*	802,380	2,113,075	1,310,695
.75	Queen Pen, *My Melody*	302,254	902,572	600,318
.50	Coal Chamber, *Coal Chamber*	90,449	284,108	193,659
.25	Wild Colonials, *Fruit of Life*	27,323	87,095	59,772
.10	Prince Paul, *Psychoanalysis . . .*	8,193	26,232	18,039
Min	Oleander, *Shrinking the Blob*	883	2,832	1,949

NOTE.—Compares first-year sales of debut albums to the model's prediction of sales if consumers had been fully informed about the album (i.e, if $q_{1,1}$ had been equal to one for that album).

where the second equality follows from the fact that observed sales $S_{1,1}$ are equal to $p_1 q_{1,1} N$, and we calculate $\hat{q}_{1,1}$ from equation (4) using our estimates of the learning parameters ($\hat{q}_0$, $\hat{a}$, and $\hat{b}$, as reported in table 5). The idea is very simple: for an album that sold 100,000 units in its release period, if we estimate that $q_{1,1}$ was 33.3 percent for that album (i.e., only a third of consumers knew about it), then the implied counterfactual sales for that album would be 300,000.

The results of these calculations are summarized in table 6. Our estimates imply that albums at the very top are not substantially undersold. Almost all consumers learn about a major hit (such as Alanis Morissette's *Jagged Little Pill*) in its release period, and the counterfactual is only a small change from reality. Albums at the bottom end of the success spectrum are also not undersold, but for a different reason. Even though most consumers are unaware of these albums, the albums' qualities are sufficiently low that sales would be minimal even if everyone were fully informed. By contrast, we estimate that moderately successful (but sub-superstar) albums are substantially undersold, in the sense that many would-be buyers remain uninformed about such albums. For all but the very top artists, album sales would have doubled or even tripled if every consumer had been aware of the album, and in absolute terms these differences would have been very large for the moderately successful artists.

Of course, one might argue that some of these potential sales would have occurred when later releases by the artist generated new information. However, consumer utility for an album appears to decline over time, so delays in discovery times are costly. Furthermore, substantial learning takes place only if one of the later releases is a major hit. If we look at sales of debut albums in the 4 years following their release years, we find that the typical artist recovers only 25 percent of the "lost sales" implied by our counterfactual analysis.[38] For artists who had a

[38] In making this calculation we restricted our attention to the 190 artists for whom we observe at least 5 complete years of sales data for the first album.

major hit (defined as an album selling over 2 million units in its first year) on a subsequent release, typically 45 percent of the lost sales are recovered. For example, the debut album from Coal Chamber (noted in table 6) sold fewer than 1 million units in its first year but over 3 million units after that, largely because the band's second release was a major hit. So the eventual sales of the debut album were even greater than what our model predicted for counterfactual sales. In contrast, Queen Pen's debut album sold over 3 million units in its first year but less than a quarter million after that, perhaps because her second release was not very successful. Our data indicate that most artists experience a fate similar to Queen Pen's.

A central implication of our finding that lesser-known artists' albums are undersold is that commercial success in this industry is more concentrated than it would be in a world in which consumers were more fully informed. More popular albums are more widely promoted, so more consumers know about them, and popularity is self-reinforcing. The fact that artists release multiple albums serves only to amplify the skewness in sales: popular first albums are likely to be followed by popular second albums, and the effect of the consequent backward spillovers will tend to increase sales disproportionately for albums that were already popular.[39] Our model does not merely measure this direct effect of spillovers on album sales; instead, it uses the spillover as a way of estimating the extent to which sales are dependent on information.

Exactly how much of the observed skewness in album sales can be attributed to consumers' lack of information about the choice set of albums? Our model cannot address this question directly unless we assume that preferences are additive. This approximation is acceptable when the counterfactual consists of adding an album to the actual choice sets of uninformed consumers, but it is not very plausible when the counterfactual consists of giving all consumers the choice set of all albums. Nevertheless, our estimates of counterfactual sales do provide an interesting benchmark. Figure 6 plots the distribution of counterfactual first-year sales ($\tilde{S}_{1,1}$, as defined above) in comparison to the observed sales of debut albums ($S_{1,1}$). The counterfactual distribution of sales is still quite skewed, but it is substantially less concentrated than the distribution of actual sales. The Gini coefficient is .647, as opposed to .724 for actual sales.[40] Among the artists in our sample, fewer than half (48 percent) sold more than 100,000 units of their debut albums; but our estimates imply that if consumers had been fully informed, then

[39] The fact that artists release multiple albums also means that the distribution of success across artists will be even more skewed than the distribution across albums.

[40] As a point of reference, the income distribution in the United States has a Gini coefficient of around .47.

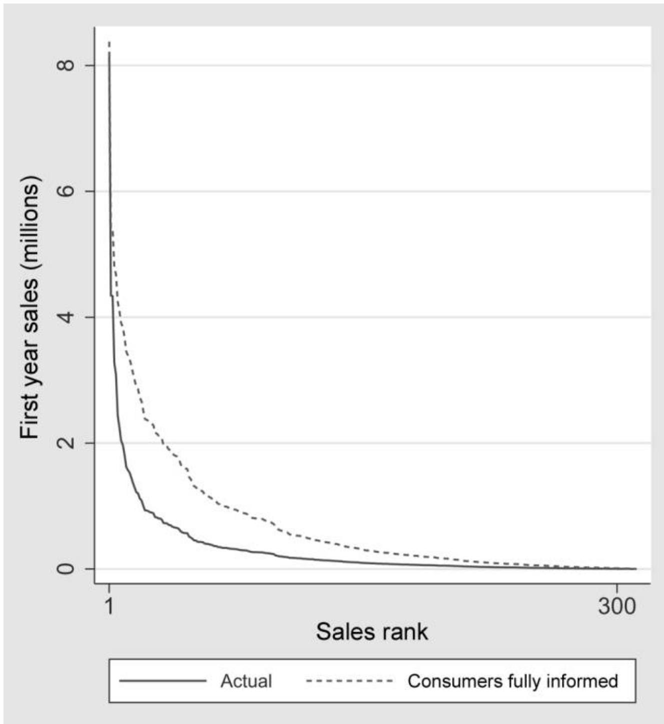

FIG. 6.—Counterfactual sales distribution for debut albums. The solid line shows the actual first-year sales of the top 300 debut albums in our sample plotted against sales rank. The dotted line indicates our estimate of what this plot would look like if all consumers were fully informed and preferences were additive across albums (i.e., no substitution effects).

nearly three-quarters (72 percent) would have met or exceeded this threshold.

We suspect that allowing for substitution effects would make the flattening of the distribution even more pronounced. To understand why, note that making consumers fully informed about all albums would have two opposing effects on the sales of each album. The direct effect would be an increase in sales resulting from more consumers knowing about the album. The indirect effect would be a decrease in sales due to consumers becoming more aware of competing albums. For the most successful of albums, the direct effect is small (most consumers know about these albums already), so the indirect effect is likely to dominate. For the least successful albums, the increased competition would come mostly from albums of higher quality, so sales would likely decrease for these albums too. For albums in the middle (i.e., moderately successful albums), the direct effect is large (since we estimate that consumer awareness is still somewhat low for such albums), and the indirect effect would be small (since they were competing against the hit albums already anyway, and exposure to competition from lower-quality albums should not have a large impact). Overall, therefore, we think that allowing for substitution effects would only accentuate the change pictured in figure 6: mid-range albums would gain, and their gains would come at the expense of albums at the top and bottom of the success distribution.

For most of our sample period, the counterfactual represented in figure 6 was unrealistic. Consumers were restricted to purchasing albums from brick-and-mortar retailers such as Wal-Mart and Tower Records and to learning about these albums from radio play and music videos on television. Physical inventory costs forced the brick-and-mortar retailers to carry only a small fraction of the available albums, and the costs of providing airtime forced radio and television stations to play songs from an even more limited selection of albums. However, the counterfactual does come close to describing the online market that has developed over the past 5 years. Digital retailers such as Rhapsody offer millions of tracks at essentially zero storage cost, compared to the 60,000 available at stores such as Wal-Mart. Powerful search engines combined with an array of sophisticated recommendation and ranking tools allow consumers to find music they like at much lower search costs. Anderson (2006) compared album sales offline with track sales online and found that the online demand curve is indeed much flatter. In the offline market, the top 1,000 albums make up nearly 80 percent of the total market, but in the online market the top 1,000 account for less than a third of the market. Over half of the online market consists of albums beyond the top 5,000 carried by brick-and-mortar retailers. The analysis does not account for possible differences in preferences between offline versus online consumers, but it does suggest that market demand has moved in the direction predicted by our model.[41]

C. *Alternative Models*

Two key features of our album discovery model that we exploit in the estimation are that (1) spillovers from new releases are determined solely by changes in the number of consumers who know about the catalog album, and (2) in each release period, consumers who discover

[41] A recent paper by Brynjolfsson, Hu, and Simester (2006) also shows some evidence that Internet marketing can make the distribution of sales less concentrated. They show that the sales distribution for a midsize clothing retailer is significantly less skewed for its Internet sales channel than for its catalog channel, and they argue that the difference reflects the lower costs of product search on the Internet.

the album are a random sample from the population. The latter feature is crucial since it allows us to difference out the unobserved purchasing probability of a randomly selected consumer, estimate the parameters of the discovery probability function from any one of the three spillovers (i.e., 2–1, 3–1, or 3–2), and use the results to test our model.

One alternative would be a model in which learning is more gradual and nuanced. Instead of making the simplifying assumption that consumer knowledge is binary, we could instead assume that consumers update their preferences gradually as they hear more songs from an album on the radio. In this model, consumers may know about an album but still need to hear more songs from the artist before deciding whether they will like it enough to buy it. Spillovers occur when a new release generates information that convinces a significant proportion of these consumers that the album is worth purchasing. If higher sales is "good news" about album quality, then each consumer's probability of purchasing the catalog album will be an increasing function of sales of the new release. We could specify a functional form for this probability—similar to the one that we specified for the probability of discovery—and estimate its parameters using, say, the 2–1 spillover. The main problem in taking this model to the data, however, is that consumers who did not purchase the catalog album in prior release periods are not a random sample: they are less likely to buy the album than a randomly selected, uninformed consumer. While it would be possible in principle to model the selection process, the resulting model would be much more complicated than the one we estimated above. More important, the selection issue would make our results highly sensitive to the functional forms chosen for the distributions of signals and tastes.

Other models could attribute the backward spillovers to changes in utility rather than changes in information. If consumers have supermodular preferences over albums by the same artist, for example, then a new release increases the utility of the artist's catalog albums. Consumers who were previously not willing to buy a catalog album may do so when they can consume it together with the new album.[42] Another form of consumption complementarities is social effects: a consumer's utility from an album may depend on the number of other consumers purchasing the artist's albums.[43] A new release that sells well could increase utility (and hence sales) of catalog albums if (*a*) the social effects

[42] The complementarity could be interpreted as a characterization of fans: e.g., when consumers listen regularly to an artist's music, they become accustomed to it or invested in the image associated with it and therefore more likely to purchase more music from that artist. Such complementarities would be similar to those modeled by Becker et al. (1994) to describe cigarette addiction and by Gentzkow (2007) to describe consumption of online and print editions of a newspaper.

[43] See Becker and Murphy (2000) and Brock and Durlauf (2001) for insightful overviews of models with social effects.

operate to some extent at the artist level and (*b*) the social effects associated with the new album exceed the social effects generated by the catalog albums themselves when they were released. The implications of these models are similar to those of a nonbinary learning model: sales of the new release affect the purchasing probabilities of consumers who have not yet purchased the catalog album, and the latter are a selected sample. In this case, the magnitude of the spillover and its variation across release periods depends on the structure of consumer preferences and how these preferences are distributed in the population.

We cannot definitively reject these alternative models using aggregate sales data. We have tried estimating versions of these models and have found that they have difficulty rationalizing the variation in spillovers across artists and release periods. For example, the selection effect makes it difficult to obtain significant spillover from later releases of successful artists. Part of the appeal of the binary learning model (aside from its plausibility) is that it provides a unified model of spillovers across artists and album pairs that, somewhat surprisingly, fits the data very well. This finding suggests to us that while gradual learning, album complementarities, and social effects may play a role, the main factor determining demand for albums is whether consumers know about them and the process through which they obtain this knowledge (i.e., radio play).

Preference-based models also have different implications about how changes in the market environment will affect the distribution of album sales. For example, as explained in the previous section, recent evidence suggests that Internet technologies have led to a flattening of the distribution of music sales. This is a natural implication of the album discovery model. By contrast, even though a model based on complementarities between albums could possibly explain the backward spillovers, it would not have direct implications about the impact of Internet technologies that facilitate the flow of information. In a model based on social effects, the impact of Internet music technologies would be ambiguous; but intuitively one might expect such a model to predict an increase in skewness, since the Internet makes it much easier for consumers to observe each others' purchases and coordinate on what is popular.

V. Conclusion

We have shown that the release of a new album generates substantial, persistent increases in the sales of previous albums by the same artist. The evidence strongly suggests that these backward spillovers are generated by changes in information: a new album release causes some consumers to discover artists and albums about which they were previously uninformed. Cross-sectional variation in the spillovers allows us to make quantitative inferences about the importance of product discovery and its impact on market outcomes. Estimates of our model imply that the distribution of sales is substantially more skewed than it would be if consumers were more fully informed. In particular, mid-range artists' albums are dramatically undersold (to the tune of hundreds of thousands of units) relative to what sales would have been if consumers were fully informed. We also find that nondebut albums benefit from a large forward spillover, selling tens of thousands more units than they would without the information generated by the prior albums.

During the period covered by our data, radio airplay was the primary channel through which consumers learned about albums. Scarce airtime and the desire of radio stations to get the largest possible audience created an informational bottleneck, with playlists focusing on a small fraction of popular new albums. As a result, sometimes consumers did not buy the right albums or at the right time. The backward spillover reflects consumers correcting initial "mistakes" and buying the right albums at a later time. But in many cases they cannot make these corrections. Artists whose early albums are mediocre are likely to have their careers truncated. For example, if an artist's first album is only moderately successful, her label may decline to produce any future albums, even though our findings suggest that with full information the artist may have eventually become a success. This of course does not mean that all unsuccessful artists are potential stars. But it does suggest that some potential stars' careers may be truncated because consumers were unaware of their music.

On the supply side, the spillovers create potential distortions in investment. Because investments in new albums yield returns on future albums, the recording rights for a new album need to be bundled with recording rights for future albums. Otherwise, these returns will not be fully captured by the investing label: other labels can free-ride and selectively bid for new albums by artists whose previous albums did well, allowing the artist to capture some of those investment returns. This is the familiar holdup problem. It reduces the willingness of the label to invest in a new album, leading to underinvestment (and possibly no investment) in that album. Long-term contracts resolve the holdup problem, and indeed virtually all contracts between artists and labels during our sample period are initially long-term contracts.[44]

[44] The RIAA and the American Federation of Television and Radio Artists have repeatedly lobbied Congress to end long-term contracting, as was done in the movie industry in the 1940s (see Terviö 2004). Our results suggest that this policy would likely lead to significant inefficiencies: fewer albums would be produced, and a higher proportion of the albums would be by established artists.

In practice, however, artists may not be able to commit to a long-term contract. Contract terms are almost always renegotiated after an artist has a successful album, with the artist threatening to strategically withhold or delay new recordings or (with the help of a lawyer) to get out of the recording contract altogether.[45] The backward spillover helps resolve the artist's inability to commit. Efficient investment in the new album requires the rights on catalog albums to be bundled with the rights on the new album. Otherwise, the label that owns the recording rights to the new album will not internalize the impact of its investment on sales of the catalog. The incumbent label, which owns the recording rights to the catalog, will have an advantage in bidding for the rights to the new album: it is willing to invest more and pay more for those rights than an outside label. Thus, the backward spillover tends to lock in the artist and explains why artists in our sample almost always stay with their incumbent labels.[46]

These arguments suggest that while spillovers give rise to potential distortions in how much a label invests in a given artist, these distortions are largely resolved through contracts. Thus, the main distortion on the supply side is not on how much gets invested in a given artist, but rather on product variety, that is, how many and which types of artists get investments from record labels. Our results suggest that albums of artists who targeted more narrowly defined niche markets were disproportionately undersold. As a result, the labels would have been less likely to invest in these artists than in artists with broader appeal. The recent development of Internet technologies for sharing and sampling music has largely eliminated the information bottleneck, at least for online consumers. As the online market grows, we expect spillovers to become smaller, long-term contracts to become less important, variety to increase,[47] and the distribution of success to flatten.

More generally, our results illustrate the importance of product discovery in markets with frequent inflows of new products. Other entertainment industries such as books, movies, and video games are obvious examples of such markets.[48] In these markets, potential consumers typ-

[45] A more thorough description of contracting practices is provided in a previous version of this paper (Hendricks and Sorensen 2006). Our understanding of these practices is based largely on conversations with Don Engel, one of the more successful lawyers who specializes in renegotiating contracts. (His press pseudonym is Busta Contract.)

[46] In our sample, fewer than 10 percent of artists ever switched between major labels, and most of the observed switches were due to termination by the incumbent label.

[47] Anderson (2006) reports that the number of new albums released grew from 44,000 titles in 2004 to 60,000 titles in 2005 and attributes that growth largely to the ease with which artists can now reach consumers.

[48] The highly publicized success of the author Dan Brown provides a clear illustration of backward spillovers and learning in the market for books. In 2003, Brown published a novel that was wildly successful, and its success catapulted one of his earlier novels (initially published in 2000) onto bestseller lists as well, even though the earlier novel had previously sold very few copies and was generally unknown.

ically observe only what other consumers buy, so they draw inferences about a product's quality only from the knowledge of its overall popularity. Consumers in these markets are more likely to discard their own information and follow the herd.[49] But in music, when other consumers buy an album, the songs on that album get played more frequently on the radio, generating information to consumers about their own preferences for the album. As a result, consumers are less likely to make mistakes (in the ex post sense) in buying albums than, say, books, and hit albums are less likely to be oversold than best-selling novels.[50] Another example is personal computers: Goeree (2005) argues that the rapid pace of technological change in computers leads consumers to be less than fully informed about the set of available products. Our findings indicate that the distribution of success in these markets may be very different from what it would be in a world with more fully informed consumers.

References

Anderson, C. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More.* New York: Hyperion.

Associated Press. 2006. "Details about the AP/Rolling Stone Poll on Downloading Music." February 2. http://surveys.ap.org/data/Ipsos/national/2006/2006-01-26%20ROLLING%20STONE%20Topline%20results.pdf.

Banerjee, A. 1992. "A Simple Model of Herd Behavior." *Q.J.E.* 107 (3): 797–817.

Becker, G., M. Grossman, and K. Murphy. 1994. "An Empirical Analysis of Cigarette Addiction." *A.E.R.* 84 (3): 396–418.

Becker, G., and K. Murphy. 2000. *Social Economics: Market Behavior in a Social Environment.* Cambridge, MA: Harvard Univ. Press.

Benkard, C. L. 2000. "Learning and Forgetting: The Dynamics of Aircraft Production." *A.E.R.* 90 (4): 1034–54.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1992. "A Theory of Fads, Fashions, Custom, and Cultural Change as Informational Cascades." *J.P.E.* 100:992–1026.

Brock, W., and S. Durlauf. 2001. "Interactions-Based Models." In *Handbook of Econometrics*, vol. 5., edited by J. Heckman and E. Leamer. Amsterdam: North-Holland.

Brynjolfsson, E., Y. Hu, and D. Simester. 2006. "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales." Working paper, Sloan School Management, Massachusetts Inst. Tech.

Cabral, L. M. B. 2000. "Stretching Firm and Brand Reputation." *Rand J. Econ.* 31 (4): 658–73.

Choi, J. P. 1998. "Brand Extension as Informational Leverage." *Rev. Econ. Studies* 65:655–69.

Gentzkow, M. 2007. "Valuing New Goods in a Model with Complementarity: Online Newspapers." *A.E.R.* 97 (3): 713–44.

[49] See Banerjee (1992) and Bikhchandani et al. (1992) for examples of herding models.

[50] This may partly explain why book sales are much more skewed than music sales. See Sorensen (2007) for some evidence and discussion of the skewed distribution of sales for hardcover fiction.

Goeree, M. S. 2005. “Advertising in the U.S. Personal Computer Industry.” Working paper, Claremont McKenna Coll.
Hendricks, K., and A. Sorensen. 2006. “Information Spillovers in the Market for Recorded Music.” Working Paper no. 12263, NBER, Cambridge, MA.
Hendricks, K., A. Sorensen, and T. Wiseman. 2009. “Observational Learning and the Demand for Search Goods.” Working paper, Stanford Univ. http://www.stanford.edu/~asorense.
Jin, G., A. Kato, and J. List. Forthcoming. “That’s News to Me! Information Revelation in Professional Certification Markets.” *Econ. Inquiry*.
Jin, G., and P. Leslie. 2003. “The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards.” *Q.J.E.* 118 (2): 409–51.
Rob, R., and J. Waldfogel. 2006. “Piracy on the High C’s: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students.” *J. Law and Econ.* 49 (April): 29–62.
Salganik, M. 2007. “Success and Failure in Cultural Markets.” PhD diss., Princeton Univ.
Salganik, M., P. Dodds, and D. Watts. 2006. “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market.” *Science* 311:854–56.
Smith, L., and P. Sørensen. 2000. “Pathological Outcomes of Observational Learning.” *Econometrica* 68 (2): 371–98.
Sorensen, A. 2007. “Bestseller Lists and Product Variety: The Case of Book Sales.” *J. Indus. Econ.* 55 (4): 715–38.
Terviö, M. 2004. “Mediocrity in Talent Markets.” Working paper, Univ. California, Berkeley.
Wernerfelt, B. 1988. “Umbrella Branding as a Signal of New Product Quality: An Example of Signalling by Posting a Bond.” *Rand J. Econ.* 19:458–66.
Wooldridge, J. 2002. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press.