Sales Displacement and Streaming Music: Evidence from YouTube

R. Scott Hiller*

Abstract

In this paper I exploit the removal of Warner Music content from YouTube in January 2009, and its restoration in October 2009, as a plausible natural experiment to investigate the impact of online content availability on album sales. I find that the blackout on YouTube had both statistically and economically significant positive effects on Warner albums, which are quickly moderated as top-selling albums are dropped from our sample. Results also show that Pareto imputed sales would lead to a qualitatively different conclusion in our analysis.

JEL Classification: D12, L82, O33

Keywords: online content, sales displacement, copyright policy, YouTube

*Fairfield University, 1073 North Benson Rd, Fairfield, CT 06824, USA. E-mail: rhiller@fairfield.edu

1 Introduction

Whether free access to online music displaces album sales has been a controversial subject among academics, policymakers, and practitioners. For instance, the 1995 Digital Performance Right in Sound Recordings Act (DPRA) created a public performance right for sound recordings that are transmitted by satellite radio and, increasingly today, Internet companies. In contrast, the DPRA exempted over-the-air broadcasters from paying for their use of the sound recordings on the assumption that the broadcasts have promotional effects.[1] Hence, today's licensing negotiation between online music providers and labels often leads to disagreements because of the inconsistent treatment and lack of definitive evidence. This paper aims to examine whether and how much the digital content services displace album sales.

A number of authors have examined the effects of consumer piracy on album sales. This paper is different from these in a couple of dimensions. First, we are not investigating illegal, peer-to-peer file sharing activities but we focus on YouTube, a legal channel that pays licensing fees to record labels. According to research firm NPD group, file sharing has in fact been declining since 2005, and music sales increased year over year in 2009 with digital accounting for 40% of sales. Further, in this period streaming services such as Spotify, iTunes Radio, and Google Play Music did not yet exist or were not operating in the U.S. while YouTube dominated the online multimedia market. Thus, this study aims to shed light on the size of sales displacement, which can help narrow the differences in opinions between contracting parties.

I exploit the removal of Warner content from YouTube for a nine-month period (which I call a "blackout" in this paper) and find a substantial treatment effect from the blackout using a sample of Billboard top 200 albums. Specifically, using a nine-month window before and after the blackout, the removal of Warner content from YouTube seems causally associated with an increase of 6591 units per week per album in the Billboard top 200 sample, 2551 units when I exclude weekly top 10 albums, and 1717 units when I exclude top 25 albums. Using a simple theoretical framework to interpret my results, I argue that there are likely substantial sales displacement for highly ranked albums, but these results are not inconsistent with the presence of substantial promotional effect

[1] This exemption, however, has been criticized in light of technological developments and alternative sources of music. The U.S. executive branch has supported an equal treatment of terrestrial and online music services (Peters, 2007; Department of Commerce, 2013).

from YouTube exposure for relatively lower ranked albums.

2 Related Literature

Starting from the Napster case (*A&M Records, Inc. v. Napster, Inc.*), sales displacement effects from online file-sharing services have been the subject of a number of studies. While I do not intend to survey this literature here (see, e.g., Liebowitz (2006) and Waldman (2013) for surveys), many of the pioneering works on music piracy made use of either survey of individuals on their past consumption (e.g., Rob and Waldfogel, 2006; Zentner, 2006; Andersen and Frenz, 2010; Waldfogel, 2010; Hong, 2013) or city/country-level panel data that often make use of variation in broadband penetration (e.g., Hui and Png, 2003; Peitz and Waelbroeck, 2004; Liebowitz, 2008; Zentner, 2010). The main difference of my treatment effect study using micro-level data is that I can control for the observed and unobserved heterogeneities to make more plausible inferences.

To my knowledge, there are only a few papers using album-level, actual sales data to investigate the effects of file sharing activities (Blackburn, 2004; Oberholzer-Gee and Strumpf, 2007; Hammond, 2013). In all three papers, the authors have a measurement of albums available on file-sharing networks and use an instrumental variable approach to address the omitted variable bias. That is, Blackburn (2004) uses RIAA lawsuits against consumers; Oberholzer-Gee and Strumpf (2007) use German students on vacation; and Hammond (2013) uses pre-release file sharing activities as an instrument. Regardless of validity of these instruments, which has been criticized in Liebowitz (2010), there are reasons to suspect that the effect of file sharing and that of legal channels such as YouTube would be different, and understanding the latter is the focus of this paper.

I am by no means the first to examine the effect of legal content distribution on sales (see also Waldfogel, 2009). In particular, Danaher et al. (2010) use the removal of NBC content from Apple's iTunes Store and its restoration as a natural shock to the supply of legitimate digital content and find that the removal is causally associated with a more than 10% increase in BitTorrent activity for NBC's content but no change in NBC's DVD sales (imputed from sales rank at Amazon.com). This analysis is similar in style, but the mechanisms are different because they look at whether users who are no longer able to purchase content (at iTunes Store) would be more inclined to make another legal purchase (at Amazon), while I examine whether users who can no longer view content free (on YouTube) would be inclined to purchase (either digital or physical) albums.[2]

In an independently developed paper, Kretschmer and Peukert (2013) study the same subject matter using European data. They exploit the fact that due to an ongoing royalty dispute between YouTube and rights holders' association in Germany, a large fraction of videos that contain music cannot be accessed in Germany while much of the same content is easily accessible in other European countries. They focus on the effects of this cross-country variation in YouTube content on top 300 songs and albums sold on the iTunes Store, where unit sales are imputed from ranks using the Pareto distribution. They conclude that the promotional effect of online music is big enough to offset sales displacement. While the details of these two papers are different, I think that their results are complementary to this paper.[3]

Lastly, I note here that analysis is confined to the Billboard top 200 sample and thus the results need not generalize to those outside of the top 200. That is, YouTube enables a vast array of user-generated content and may indeed bring substantial promotional benefits to emerging or independent artists. Waldfogel (2012) assembles comprehensive data on albums released between 1980 and 2010 and finds some evidence that Internet radio increases the number of albums consumers are aware of and an increasing number of albums find commercial success without substantial radio play.[4] I abstract from the supply side (or long-tail) effect of YouTube, which need to be taken into account for societal effects of free online distribution. However, licensing agreement between established labels and online services would be equally important.

3 Background Information

YouTube was launched in November 2005 as a video sharing website. The site grew rapidly, and Reuters reported in 2006 that YouTube was the leader in Internet video content with 29% share of the U.S. multimedia market and 20 million unique viewers per month. According to data published

[2]Another difference is that while I focus on the impact of content removal on relatively newly released albums, Danaher et al. (2010) remove all recent television episodes because NBC did not sell then-current season content on iTunes prior to the removal.

[3]One potential caveat on data in Kretschmer and Peukert (2013) is that they use only digital sales (rank) information from the iTunes Stores. This seems to assume a high correlation between digital and CD sales while my sales data include both.

[4]Bourreau et al. (2013) find that the number of new releases can increase without having higher overall sales. Thus, a strong sales displacement effect at the top can be consistent with online content services having some promotional effect for lesser-known artists.

by market research company comScore in 2010, YouTube's market share in online video content was 43.1% followed by Hulu (3.5%). Further, 84.8% of the total U.S. Internet audience viewed online video, where 144.1 million viewers watched 14.6 billion videos on YouTube (101.2 videos per viewer). At least since 2010, the web information company Alexa ranks YouTube as the third most visited website on the Internet, behind Google and Facebook. Hence, a media blackout on YouTube would have an effect that is large enough to show statistically.

To be precise, YouTube started as a platform to upload, view and share home-made, user-generated videos, but soon the site contained many unauthorized clips of copyrighted music. Users could upload videos in an unlimited number and anyone could watch free. In this early period, YouTube was basically an on-demand radio, where almost every song a user wanted to hear could be found. Because YouTube did not review videos before they were posted, it was left to copyright owners to issue a takedown notice pursuant to the Digital Millennium Copyright Act. However, in June 2007 Google, having acquired YouTube in November 2006, began resolving copyright infringement claims that characterized YouTube's early days, both through licensing deals with major content providers and a copyright management system called Content ID.[5]

YouTube entered into a revenue-sharing partnership with content providers as early as 2006, and all the major labels had licensing agreements with YouTube in 2007. However, labels were disappointed with those agreements having included a small fee for every video watched and a small share of the advertising revenue, leading up to renegotiation of terms. In late December 2008, when it was time for licensing renewals, multiple press releases reported that Warner and YouTube failed to agree to terms on a new licensing deal. YouTube began to remove Warner music videos, both professionally made music videos and amateur material that included Warner content. Apparently, most (if not all) of Warner music on YouTube was pulled or muted promptly, which the Electronic Frontier Foundation called the 'January Fair Use Massacre.'

The removal of Warner content came at a time when the other three majors (Universal Music, Sony BMG, and EMI) were also up for renegotiating their licensing deals with YouTube, which were to expire soon. Interestingly, YouTube and the remaining three labels reached renewal agreements with no reported disputes or content removal. On 29 September 2009, YouTube and Warner

[5]When a video is uploaded, Content ID checks it against reference libraries of copyrighted audio and video material and alerts copyright owners whenever any part of their content went up on YouTube. Owners can then choose to remove the content or sell ads and share the revenue with YouTube.

announced that they finally reached an agreement and Warner's artists (both the full catalog and user-generated content containing Warner acts) were returning to YouTube. This created a nine-month blackout period during which it was extremely unlikely to find Warner content on YouTube, providing a natural experiment as YouTube had licensing deals with all the major labels (including Warner) for almost two years before and after the blackout.[6]

The validity of the difference-in-difference approach rests on the common trends assumption, and researchers must be concerned about pre-existing differences in levels as well as changes between the periods. Figure 1 shows the aggregate weekly sales of the four major labels and independent labels nine months before and after the blackout (which is shaded). Warner Music in fact has the largest average sales closely followed by Universal Music. The overall time trends do not seem to differ noticeably at least among the top three labels, and except for the holiday season the trends also appear to be stationary. Here, to ameliorate the concerns of differential trends, I will interact each of my control variables with the blackout to control for the differential changes in sales driven by shifts in the effects of observable covariates.

Once concern is that the licensing agreements between YouTube and labels are not public information, making it rather difficult to rule out the possibility that there are some omitted variables that drive the changes in our outcomes before and after the blackout. However, a person familiar with the situation told a reporter that YouTube and Warner were close to an agreement until the last moment, when Warner suddenly changed its terms. In response, Google made the move to remove the label's content, which was said to be unexpected and drew considerable press coverage.[7] Hence, I believe that Google's intent was to appear tough on Warner to prevent the other three majors from behaving opportunistically in the upcoming licensing renegotiation. Indeed, there were no reported disputes between Google and the other labels.

When Warner's content came back on YouTube after nine months, Warner had secured more rights to the advertising revenue. The 2009 annual report for Warner said, "We reached a new and expanded agreement with YouTube, which has brought our music videos back to this important service. We have established a platform on YouTube that provides us with greater monetization op-

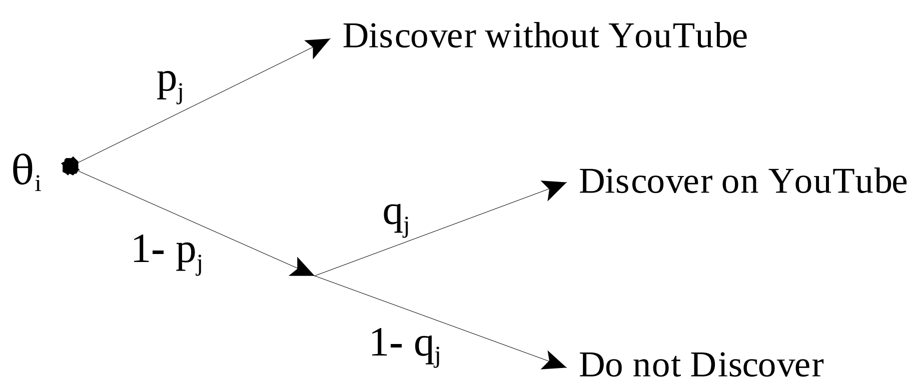
[6]This case stands in contrast to the 2007 Viacom case, where there was no prior agreement; that is, users who posted Viacom's programming did so without authorization. In the Warner case, users were able to legitimately post Warner's contents before their videos were suddenly taken down.

[7]This explanation is consistent with the fact that Google did not say it was taking the music down at Warner's request. See http://allthingsd.com/20081220/warner-music-group-disappearing-from-youtube-both-sides-take-credit/.

portunities with premium brand advertisers and drives commerce through links to artist websites." It thus appears that their contractual dispute was mainly over monetization on the video sharing website rather than reflecting changes in Warner Music's marketing and/or pricing strategies. I could not find any evidence that Warner chose different sales or promotion strategies during the blackout, but I note here that the researcher cannot know for sure.

4 Model

I provide a simple (static) model that captures the consumer's purchase decision and derive the expression for the treatment effect of interest, that is, effect of blackout on YouTube. There is a continuum of consumers characterized by a preference parameter θ_i which follows some distribution $F(\theta)$. For each album j, consumer i discovers the album without YouTube with probability p_j. If the consumer had not discovered the album through other media, then the consumer may discover the album with probability q_j on YouTube. For simplicity and without loss, I normalize $q_j = 1$ which means that p_j is the probability that a consumer discovers album j with a channel other than YouTube, conditional on being informed of the album.



When a consumer discovers an album (with or without YouTube), the consumption decision depends on the availability of content on YouTube. If the content is available on YouTube, then the consumer compares utility from buying an album to watching on YouTube. The utility of the former is $\theta_i v_a - c$, where v_a is the value of an album and c is the price; the utility of the latter is $\theta_i v_y$, where v_y is the value of watching on YouTube and $v_a > v_y$. It follows that consumer i will purchase the album if and only if θ_i is above a certain threshold $\bar{\theta} = \frac{c}{(v_a - v_y)}$. Hence, if all content is available on YouTube, the demand for album j is $D_j = 1 - F(\bar{\theta})$ in normal times. That is, conditional on being informed, it does not matter how consumers discover the album.

Now suppose that some albums (js) are not available on YouTube ("blackout"). For such albums, consumers can only discover the album with probability p_j. Further, the consumption decision is whether or not to purchase the album. Assuming that the outside option yields zero utility, the demand for such albums is $D_j = p_j[1 - F(\hat{\theta})]$, where the new threshold is $\hat{\theta} = c/v_a$. For those albums (ks) that are still available on YouTube, there can be potential substitution effects from the albums that become unavailable on YouTube. That is, substitution may occur because some consumers who do not discover an album without YouTube may buy other albums that they do discover without YouTube. I denote the fraction of consumers substituting an album k for j as ξ_{jk}.

Then the demand for albums during blackout are given by

i) $D_j = p_j[1 - F(\hat{\theta})]$ for album j that becomes unavailable on YouTube.

ii) $D_k = 1 - F(\bar{\theta}) + \sum_j \xi_{jk}(1 - p_j)[1 - F(\bar{\theta})]$ for album k that remains available on YouTube.

Subtracting the demand $D_j = 1 - F(\bar{\theta})$ in normal times from i) and ii) and then taking the difference of the two, I get the following treatment effect from the blackout:

$$-\sum_j \xi_{jk}(1 - p_j)[1 - F(\bar{\theta})] - [1 - F(\bar{\theta})] + p_j[1 - F(\hat{\theta})].$$

In an ideal world, a definition of sales displacement effect is $F(\bar{\theta}) - F(\hat{\theta})$, that is *conditional on being informed of an album* how many would-be sales are lost due to YouTube watching. Notice that this expression is always positive given $\hat{\theta} < \bar{\theta}$.

Observing that the treatment effect expression is equal to sales displacement effect when $p_j = 1$, I come to the conclusion that the treatment effect estimates would underestimate the true sales displacement effect when $p_j < 1$. The intuitive reason is that when p_j is small consumers depend heavily on YouTube exposure for discovery of the album so that album sales decrease if it is not available on YouTube. In other words, YouTube's promotional effect and the sales displacement effect pull the treatment in opposite directions, and finding which dominates is the empirical question. Although the first term in the above expression reduces the estimate of the treatment effect, I think that it is unlikely that large amounts of substitution (ξ_{jk}) will occur because of the blackout as albums are often imperfect substitutes, depending considerably on preferences, genre, and band.

On the other hand, the model predicts that when the probability p_j that consumers become aware of an artist's album from a channel other than YouTube increases the treatment effect from blackout on YouTube also increases. The issue is finding a good empirical proxy for p_j. There is some anecdotal evidence that artists and labels make considerable effort in promoting an album before the debut week and most of the album's contents are not available on YouTube before the debut week. Therefore, I think that the best available proxy for consumers' awareness of an album is the Billboard rank of the album in its debut week. That is, I will examine the treatment effect interacted with debut week's rank to render support for or against this prediction of the model. However, note that I cannot determine the scale of this proxy because p_j is in theory between 0 and 1.

5 Data

The sample for this study is based on the Billboard 200, the U.S. industry standard for album sales. The Billboard 200 is a ranking of the 200 highest-selling music albums from any genre. The chart in this period is based solely on sales (physical and digital combined) of albums in the United States. To reiterate, this is a restricted sample, and thus I do not claim that the findings in this paper generalize to those not making it into this chart. However, I believe the main issue here is to get accurate variation in sales. That is, the Billboard rankings omit any sales information making it impossible to determine, for instance, if the number one album this week sold the same volume as number one from the previous week.

I obtained access to the weekly sales data for the Billboard 200 albums from Nielsen SoundScan, which is the official basis for the Billboard charts rankings. Looking at the data, there is considerable variation both within and across weeks. On average, the 200th ranked album has just 1.5% of the sales accruing to the top ranked album in a given week. Importantly, top sales also vary considerably across weeks (see Figure 1). Because of the proprietary nature, grid sizes are removed from Figure 1. In Section 7, we show that Pareto imputation cannot capture much of the variation at the top.

I could not expand the sample further below top 200 albums because it would be too costly to obtain sales data.

There are 1796 albums from 1449 artists in Table 1 which includes a nine-month period before and after the blackout. I drop on average 10-20 entries from every week because there are typically non-music albums and albums that are compilations (thus no artist fixed effects) in the top 200. For each album, I construct the following variables: $TWsales$ is this week's sales, that is, the number of albums sold in a given week; *Wkson* is the cumulative number of weeks on the Billboard chart, and *Wksonsq* is the square of *Wkson*. Because demand for a new album (pre-orders) often builds up before the premiere week, I create a variable, *Firstweek*, that indicates the first week of each album in our sample.

The data includes both new and catalog albums, so I create an indicator for new albums: *Firstalbum* is an indicator for the first album by an artist found in my SoundScan database. Because the database contains the top 200 albums beginning in 2004, this variable indicates that an album is the first to make this chart since at least 2004. To control for the effect of previous album sales (e.g., Hendricks and Sorensen, 2009), I include *Previousalbumduration* and *Previousalbumsales*, which are the length of time an album has been on the Billboard chart and total sales for the last album an artist placed on the Top 200. These are equal to zero if they had no previous album on the charts since 2004.

I then match albums with genre and label information based on the Discogs.com database which is exceptionally extensive. I manually coded all major/indie labels: majors being either directly one of the majors, or a subsidiary of one of the majors. This involved following the path for each sublabel. If the label was under a major, the indicator for that label was coded one. For some labels this involved finding a website or article about them. If I could find no connections anywhere, I labeled them as Indy. If Discogs.com labeled them as self-release or if the only releases under the label were of the artist, then I labeled it self-release. There are 14 standard genres as listed by Discogs.com.

The data are merged with a separate radio chart ranking. The source of data on radio airplay is the weekly USA Airplay Top 200.[8] Billboard also has a radio chart listing the 75 most aired

[8] The radio chart ("The most played tracks on USA radio stations") is available at http://www.charly1300.com/usaairplay.htm.

songs of the week, but I preferred the USA Airplay Top 200 because of its broader coverage. Notice that the radio chart is for songs, while my sales data is for albums. In many instances a song will be played extensively before the album is released. Therefore, when mapping each song to the album-week in my data, I matched airplay for the weeks where an album was on the chart taking into account the fact that an album can appear on the radio chart several weeks before the album chart.

I do not have data on airtime minutes, but I think that using the radio rank may be less of a concern because the total number of slots for songs in radio stations is relatively fixed over time. Specifically, I construct the following variables: *Lastweekradiorank* is the last week's chart ranking for matched songs for all album-weeks; *Wksonradio* is the number of weeks on the chart (and zero if not on the chart); *Weeksinceradio* is the number of weeks since an album had a song on the radio chart last (zero if active on the chart). These variables capture the chart duration, but allow for differential effects. Finally, *Noradio* is an indicator for albums that have never had a song on the radio chart.

6 Empirical Results

Although the primary contribution of this paper is on the sales displacement effect, it is worth clarifying the relationship between a sales displacement effect and promotional effect to help understand my empirical specification below. First of all, YouTube is a legal distribution channel which contained a large variety of official as well as user-generated music videos. As previously elaborated, it was unique in its capacity during my sample period and unlike file sharing there was no need to download the content which reduced costs substantially. Further, YouTube suggested related music videos, exposing multiple songs by the same artists, genre, or period. Hence, there was a good chance that consumers could learn about unknown albums and match with their preferences (see, e.g., Gopal et al., 2006; Peitz and Waelbroeck, 2006).

Existing papers often regard and interpret the sales displacement effect as stronger than any promotional effect. However, it is clear from the above-mentioned papers that the promotional effect applies more strongly to lesser-known artists. Because my treatment effect is closer to (less than) the true sales displacement effect when p_j is near (further away from) 1, one way to gauge the extent of the promotional effect is to sequentially drop top-selling albums from the samples. In this way, I can present both the economic impact of the blackout episode as well as the onset of the promotional effect at the same time (although the latter evidence would be necessarily indirect). I think that this is more effective than taking a log transformation of sales because doing so would suppress significant variation at the top as shown in the next section, although I replicate the baseline specification with a log transformation to show the similarities.

Figure 1 shows the average weekly album sales by label during the 9 month period before and after the blackout, represented by the shaded area. Clearly Warner albums on average sold more than any other labels' throughout the sample period. Although I am agnostic about the reason for this superior sales, at least the labels' average sales do not exhibit reversals during the sample period. As is standard, I include label fixed effects (as well as week fixed effects) to control for these pre-existing differences. Of course, a major issue in difference-in-difference analyses is to look out for pre-existing trends. Here, the gap between Warner's and other labels' weekly average sales does not seem to show any noticeable trends either before or after the blackout, and I also interact our controls with the blackout period.

The first specification is the standard difference-in-difference model:

$$Y_{it} = \beta\, Warnereffect_{it} + \boldsymbol{X}_{it}\boldsymbol{\delta} + Week_t + Artist_i + \epsilon_{it},$$

where Y_{it} is artist i's album sales in week t (*twsales*) and ϵ_{it} is i.i.d.

The data set is an unbalanced panel of artists because no single album lasts for the entire sample period. That is, I stack the observations by artist instead of album because some artists have multiple albums and I can exploit more within artist variation this way. Specifically, there are 302 artists (out of 1449) with multiple albums over the sample period.[9] If I use album fixed effects, then the estimates are similar but a little less precise, and of course any constant album characteristics would not be identified.

$\boldsymbol{X}_{it}$ is the set of album characteristics that were previously described. For instance, this includes statistics based on radio chart as well as genre and label indicators. Note that these can vary within artist if the artist had more than one album over time. To account for the unobserved heterogeneity

[9]There are in fact some exceptional cases where more than one album by the same artist show up in the same chart week, in which case I aggregated the variables.

and seasonality, I include artist fixed effects and week dummies, and to control for any changes in the relationship between the covariates and outcome variables, I interact variables in $\boldsymbol{X}_{jt}$ with the indicator for the blackout period.

The main coefficient of interest is β. *Warnereffect* is an indicator for the artist-weeks with albums released by Warner (or its subsidiaries) during the blackout, which runs from the first week of January 2009 to the last week of September 2009. For instance, if Warner released an album five weeks prior to the blackout and it charted for 15 weeks, then *Warnereffect* would be zero for the first five weeks, and one for the ensuing ten. Many albums do fall entirely within or outside of the blackout period.[10]

Table 2 shows the main results where column (1) uses the full sample, column (2) drops top 10 albums, column (3) drops top 25 albums, and column (4) drops top 50 albums.[11] The estimates of *Warnereffect* show how content removal from YouTube affected Warner albums during the blackout. The results are twofold. The first is that although the samples are limited to Billboard top 200 albums, Warner albums on average sold larger quantities ranging up to several thousand units per week compared to comparable non-Warner albums during the blackout, after attributes of the albums were controlled for. The second is that the treatment effect tapers off relatively quickly as I drop top sellers from the sample. Thus, there seems to be significant sales displacement for top artists but a promotional effect seems to balance displacement at a relatively high level of sales.

The treatment effect for top artists is both statistically and economically significant. For instance, suppose that the blackout on YouTube had a causal impact on sales by 4,000 units per week for a top album which seems reasonable. Then a rough back-of-the-envelope calculation indicates 4,000 $\times$ \$12 (the average album price) $\times$ 20 (the average *Wkson*) $\approx$ \$1 million of lost sales for a top album. If we additionally assume that a label has say, 40 top albums per year, then the total lost sales become \$40 million per year. Note, however, that the sales displacement from YouTube is not necessarily a problem as the licensing deal is a contractual agreement between both sides, but it does seem to have a nontrivial revenue implication that needs to be considered in licensing

[10] Our preferred specification is to exploit variation both within and across albums, but the results are similar if I only use albums that cross either of the two blackout thresholds. Further, my specifications using a three or six month window yield stronger results.

[11] Robust standard errors are clustered at the artist level in all tabulations as sales are likely correlated across albums by artist. I also bootstrapped standard errors following Bertrand et al. (2004), which does not really change qualitative results. If anything, the estimate on *Warnereffect* often becomes more significant.

negotiation between labels and video sharing sites.

The results on other parameters seem sensible. The estimate on the first week indicator shows that sales are particularly high in the premiere week. The negative coefficient on the first album indicator suggests that new albums (either an artist's debut album or a new album in at least four years' time) tend to sell relatively small quantities holding constant other parameters. These findings are consistent with Hendricks and Sorensen (2009) where consumers are less likely to be aware of new artists even with the moderate levels of success associated with being on the Billboard 200 chart. On the other hand, I do not find significant effects from an artist's previous album although a longer chart duration of previous album seems to be negatively correlated and previous album sales are positively correlated with current album sales.

I also find that the number of weeks but not the latest rank on the radio chart is positively associated with album sales. Although the number of weeks on the Billboard chart has a quadratic shape (*Wkson* and *Wksonsq*), it is mostly negative given the average chart duration. The latter captures the fact that sales tend to naturally decrease over chart week while the former indicates some positive effects of radio play, especially since I start counting radio weeks before an album may appear on the Billboard 200 chart. The number of weeks since the last appearance on the radio chart are negatively associated with sales and albums with no radio rank tend to have lower sales, all else equal. All these seem sensible, but they are correlative effects so I cannot attest to the causal effects of radio exposure.

Table 3 replicates the specification in Table 2, with a log transformation of sales as the dependent variable rather than the sales levels. Results show that a log specification looks qualitatively similar to that in the baseline levels model. Levels are used for the remainder of the specifications to provide an idea of the impact of the net displacement, but log transformations provide similar results in each case.

The second specification is an interaction model:

$$Y_{it} = \beta_1 \, Warnereffect_{it} + \beta_2 \, Warnerdebutrank_{it} + \boldsymbol{X}_{it}\boldsymbol{\delta} + Week_t + Artist_i + \epsilon_{it},$$

where *Warnerdebutrank* is *Warnereffect* interacted with chart debut week's rank.

The hypothesis is that the treatment effect increases with the probability that consumers are aware of an album from a channel other than YouTube, and I use an album's debut rank as a proxy for this parameter because the album must be ubiquitous in popular culture to debut high, and an album that debuts higher has had more prerelease promotion. Table 4 shows that as the rank of the debut week decreases the treatment effect is indeed lessened. That is, the widely-promoted, higher debuting albums get a boost from the blackout because YouTube was displacing sales, and the lower debuting albums lose some of the promotional power from the blackout on YouTube. I think that popular culture and album promotion are predetermined factors and also note that the treatment effect estimates do not seem to be particularly sensitive when I include this interaction term.

7 Imputation

When sales data is either unavailable or costly to obtain, researchers have often imputed sales quantities using more readily available data on sales rank. Since I have actual sales data, in this section I briefly discuss the performance of the common imputation method (i.e., Pareto distribution). Pareto distribution is a type of power law, specifically $s_{it} = \alpha r_{it}^{\beta}$ where s_{it} is sales quantity and r_{it} is the sales rank of product i in time t, which has allowed researchers to infer the sales when they do not have data.[12] The two parameter values (α and β) can be time-varying; however, as long as researchers have a reasonable estimate, they can back out sales quantities from sales rank.

The parameters α and β thus need to be estimated first using some sparse or imperfect measures of actual sales together with rank. We follow Waldfogel (2012) where these parameters are estimated using Billboard rankings and RIAA sales certifications. With certifications, researchers cannot see any weekly results, but can see cumulative sales at certain benchmarks. For instance, albums are certified at the half million, million, two million, and ten million sales points.[13] Thus, the researcher can estimate a version of the Pareto distribution knowing how many weeks an album was ranked on the Billboard chart, the ranking in each week on that chart, and the certification of the albums

[12] Brynjolfsson et al. (2003) and Chevalier and Goolsbee (2003) apply a power law to sales data. Both papers estimate sales volume of books sold on Amazon.com and provide supporting evidence that this is a good approximation. Waldfogel (2012) applies a similar method to albums.

[13] Certification levels from the Recording Industry Association of America (RIAA) are available at http://www.riaa.com/goldandplatinum.php?content_selector=criteria

that reach the sales thresholds.

To be more specific, because the certification only allows the researcher to view the sparse cumulative sales of the album to that point, the equation to be estimated is $S_{i\tau} = \sum_{t=0}^{\tau} \alpha r_{it}^{\beta}$, where S_{it} is the cumulative total sales of album i in week τ. An estimate of α and β can be obtained given sufficient data in some time interval. In this way, Waldfogel (2012) estimates that the slope parameter β is around -0.65 in the 2000s for the Billboard 200. For my analysis, I estimate β using the same specification above and only RIAA certification thresholds (as is available to anyone with only certification data) for albums in my SoundScan database (2004-2012) and I found similar estimates where $\beta = -0.69$.

Notice that α and β are jointly estimated but researchers need not use the α estimate if they have superior information that can guide the selection of α.[14] Intuitively, holding β constant, the choice of α determines the level of the Pareto distribution, so it can influence whether the imputed sales quantities fit underlying data better at the top, middle, or bottom rank. Here, the economically important issue is to capture the variation at relatively higher ranks. Since α serves as an estimation of the sales of the top album, for α I use the actual average sales of the number one ranked albums for each year in my sample. If imputed sales using this data is problematic, then sales imputed with less information must be as well.

Table 5 uses the imputed sales and replicates Table 2 with the exact same sample. Imputed sales data seem to contain considerable noise and is unable to capture meaningful variation in actual sales data. Although the sign of the treatment effect is the same, the estimates are substantially lower than those reported in Table 2 and they are not statistically significant. It is worth noting that album sales have declined in recent year so the number of certification data points has also declined. Imputed data may be more reliable in a cross-sectional analysis, but it can lead to a potentially biased conclusion in panel-data analysis.[15] Hence, researchers must be cautioned when applying Pareto imputation without some finer level information.

[14] For instance, Waldfogel (2012) uses yearly aggregate shipment data to set α as the ratio between the sum of the total imputed sales and the aggregate shipment. This relies on the assumption that the sales of albums that never appear on the Billboard 200 are negligible.

[15] The imputation assumes all weeks in a year have the same parameters. The biggest problem with imputed data is that it cannot capture both the level and decay of the actual sales of highly ranked albums which can vary greatly depending on week and release.

8 Conclusion

Digital performance rights are growing in importance as a source of remuneration for artists, and online distribution channels such as YouTube are becoming a prime music-listening portal with little to no cost required on consumers. While previous studies have provided some evidence on the effect of file sharing services, systematic evidence on the effect of legal online channels has been lacking in the literature. I have helped to fill this gap by using a difference-in-difference approach and provided some evidence that a widely-used video sharing services like YouTube can have a substantial sales displacement effect on top albums while at the same time providing benefits of wider exposure to potential consumers.

Finally, I caution that the promotional and displacement effect of 'noninteractive' services (such as Pandora) may well be different from that of interactive services where listeners have control over the content. There are in fact considerable policy debates on the streaming services like Pandora that have opted for statutory licensing rather than making direct licensing agreements with labels while larger companies such as Apple and Google have sought to negotiate with labels. A direct deal which allows for more functionality would require more revenue sharing with rights holders. That is, my results should not be applied without qualification to streaming services which have sought to lower royalty payments.[16]

References

[1] Andersen, B., and M. Frenz (2010). "Don't Blame the P2P File-Sharers: the Impact of Free Music Downloads on the Purchase of Music CDs in Canada," *Journal of Evolutionary Economics* 20(5): 715–740.

[2] Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119(1): 249–275.

[3] Blackburn, D. (2004). "The Heterogenous Effects of Copying: The Case of Recorded Music," Mimeo.

[16] The Internet Radio Fairness Act, introduced in September 2012, sought to move the "willing buyer, willing seller" standard for online music streaming to the one used for satellite and cable radio outlets like Sirius XM and Music Choice by requiring the Copyright Royalty Judges to consider factors including whether the licensing costs will have a "disruptive impact" on the industry.

[4] Bourreau, M., M. Gensollen, F. Moreau, and P. Waelbroeck (2013). "“Selling Less of More?” The Impact of Digitization on Record Companies," *Journal of Cultural Economics* 37: 327–346.

[5] Brynjolfsson, E., Y. Hu, and M. Smith (2003). "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* 49: 1580–1596.

[6] Chevalier, J., and A. Goolsbee (2003). "Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com," *Quantitative Marketing and Economics* 1(2): 203–222.

[7] Danaher, B., S. Dhanasobhon, M. Smith, and R. Telang (2010). "Converting Pirates without Cannibalizing Purchasers: The Impact of Digital Distribution on Physical Sales and Internet Piracy," *Marketing Science* 29(6): 1138–1151.

[8] Department of Commerce. (2013). "Copyright Policy, Creativity, and Innovation in the Digital Economy," Green Paper.

[9] Gopal, R., S. Bhattacharjee, and L. Sanders (2006). "Do Artists Benefit from Online Music Sharing?" *Journal of Business* 79: 1503–1533.

[10] Hammond, R. (2013). "Profit Leak? Pre-Release File Sharing and the Music Industry," *Southern Economic Journal* (forthcoming).

[11] Hendricks, K., and A. Sorensen (2009). "Information and the Skewness of Music Sales," *Journal of Political Economy* 117(2): 324–369.

[12] Hong, S.-H. (2013). "Measuring the Effect of Napster on Recorded Music Sales: Difference-in-Differences Estimates under Compositional Changes," *Journal of Applied Econometrics* 28(2): 297–324.

[13] Hui, K., and I. Png (2003). "Piracy and the Legitimate Demand for Recorded Music," *Contributions to Economic Analysis & Policy* 2(1): Article 11.

[14] Kretschmer, T., and C. Peukert (2013). "Video Killed the Radio Star? Online Music Videos and Digital Music Sales," Mimeo.

[15] Liebowitz, S. (2006). "File Sharing: Creative Destruction or Just Plain Destruction?" *Journal of Law and Economics* 49(1): 1–28.

[16] ——— (2008). "Testing File-Sharing's Impact by Examining Record Sales in Cities," *Management Science* 54(4): 852–859.

[17] ——— (2010). "The Oberholzer-Gee/Strumpf File-Sharing Instrument Fails the Laugh Test," Mimeo.

[18] Oberholzer, F., and K. Strumpf (2007). "The Effect of File Sharing on Record Sales. An Empirical Analysis," *Journal of Political Economy* 115(1): 1–42.

[19] Peitz, M., and P. Waelbroeck (2004). "The Effect of Internet Piracy on CD Sales: Cross-Section Evidence," *Review of Economic Research on Copyright Issues* 1(2): 71–79.

[20] ———, and ——— (2006). "Why the Music Industry May Gain from Free Downloading," *International Journal of Industrial Organization* 24(5): 907–913.

[21] Peters, M. (2007). Statement of Marybeth Peters, the Register of Copyrights, before the Subcommittee on Courts, the Internet, and Intellectual Property, Committee on the Judiciary, United States House of Representatives.

[22] 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," *Journal of Law and Economics* 49(1): 29–62.

[23] Waldfogel, J. (2009). "Lost on the Web: Does Web Distribution Stimulate or Depress Television Viewing?" *Information Economics and Policy* 21: 158–168.

[24] ——— (2010). "Music File Sharing and Sales Displacement in the iTunes Era," *Information Economics and Policy* 22(4): 306–314.

[25] ——— (2012). "And the Bands Played On: Digital Disintermediation and the Quality of New Recorded Music," Mimeo.

[26] Waldman, M. (2013). "Piracy of Intellectual Property: Past, Present, and Future," *Review of Economic Research on Copyright Issues* 10(2): 1–26.

[27] Zentner, A. (2006). "Measuring the Effect of File Sharing on Music Purchases," *Journal of Law and Economics* 49(1): 63–90.

[28] ——— (2010). "Ten Years of File Sharing and Its Effect on International Physical and Digital Music Sales," Mimeo.

Table 1: Summary statistics - weekly observations

Variable	Mean	Std. dev.
TWsales	14492.473	32567.778
Wkson	20.711	29.123
Firstweek	0.119	0.324
Firstalbum	0.229	0.42
Previousalbumduration	17.321	28.298
Previousalbumsales	567.062	1183.021
Wksonradio	4.142	7.467
Lastweekradiorank	29.791	50.108
Weeksinceradio	0.739	4.416
Noradio	0.562	0.496
EMI	0.098	0.298
Sony	0.177	0.382
SonyBMG	0.076	0.264
Universal	0.242	0.428
Warner	0.148	0.355
Indy	0.242	0.428
Self-release	0.017	0.131
N	17314	

The date range for this is from April of 2008 to June 2010.

Table 2: Sales regressions - baseline model

	(1) Full Sample	(2) Drop Top 10	(3) Drop Top 25	(4) Drop Top 50
Warnereffect	6591.39**	2551.82**	1717.82**	397.94
	(3135.94)	(1132.24)	(705.80)	(455.84)
Wkson	-645.05***	-195.24***	-96.06***	-41.39***
	(99.98)	(26.81)	(13.37)	(7.09)
Wksonsq	2.46***	0.67***	0.27***	0.10***
	(0.64)	(0.17)	(0.08)	(0.03)
Firstweek	30287.79***	3723.08***	682.47**	-494.74***
	(2462.28)	(599.72)	(326.56)	(177.10)
Firstalbum	-10196.97**	-1483.40	-1742.45**	-637.74
	(4278.77)	(1131.85)	(716.51)	(399.22)
Previousalbumduration	26.63	-26.78	-20.97***	-3.10
	(92.31)	(17.76)	(7.15)	(5.45)
Previousalbumsales	-0.98	0.10	0.29**	0.23**
	(1.61)	(0.35)	(0.13)	(0.09)
Wksonradio	191.84**	106.48***	76.30***	31.44***
	(94.51)	(32.67)	(21.63)	(11.89)
Lastweekradiorank	-28.90	-2.53	-3.22	-2.14
	(22.48)	(5.10)	(2.70)	(1.74)
Weeksinceradio	-456.78***	-173.45***	-123.62***	-78.83***
	(94.54)	(39.85)	(37.77)	(21.96)
Noradio	-6313.53**	-1281.99	-121.01	-309.77
	(2967.20)	(1009.78)	(525.75)	(312.33)
N	17314	16373	15006	12693

Interaction of the above ten covariates with the blackout as well as genre, label, and week dummies are not shown for brevity. Standard errors clustered by artists are in parentheses.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 3: Sales regressions - baseline replicated with log transformation

	(1) Full Sample	(2) Drop Top 10	(3) Drop Top 25	(4) Drop Top 50
Warnereffect	0.24*	0.22**	0.18**	0.076
	(0.13)	(0.10)	(0.094)	(0.067)
Wkson	-0.020***	-0.013***	-0.0088***	-0.0046***
	(0.0019)	(0.0014)	(0.0012)	(0.00086)
Wksonsq	6.3e-05***	4.0e-05***	2.3e-05***	9.4e-06***
	(1.1e-05)	(6.7e-06)	(4.8e-06)	(3.4e-06)
Firstweek	0.73***	0.33***	0.13***	-0.021
	(0.030)	(0.032)	(0.028)	(0.020)
Firstalbum	-0.31***	-0.19**	-0.17***	-0.13***
	(0.10)	(0.078)	(0.062)	(0.048)
Previousalbumduration	-0.0014	-0.0025***	-0.0019***	-0.00096
	(0.0017)	(0.00091)	(0.00067)	(0.00065)
Previousalbumsales	-0.000016	0.000020	0.000034**	0.000025**
	(0.000033)	(0.000021)	(0.000014)	(0.00001)
Wksonradio	0.0097***	0.0094***	0.0077***	0.0037***
	(0.0021)	(0.0018)	(0.0015)	(0.0013)
Lastweekradiorank	-0.00049	-0.00034	-0.00035*	-0.00021
	(0.00032)	(0.00024)	(0.00020)	(0.00019)
Weeksinceradio	-0.028***	-0.022***	-0.019***	-0.014***
	(0.0051)	(0.0047)	(0.0044)	(0.0033)
Noradio	-0.13*	-0.085	-0.031	-0.063*
	(0.073)	(0.060)	(0.046)	(0.037)
N	17314	16373	15006	12693

Interaction of the above ten covariates with the blackout as well as genre, label, and week dummies are not shown for brevity. Standard errors clustered by artists are in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 4: Sales regressions - interaction model

	(1) Full Sample	(2) Drop Top 10	(3) Drop Top 25	(4) Drop Top 50
Warnereffect	6269.96*	2340.69**	1591.74**	351.81
	(3409.04)	(1189.32)	(726.94)	(429.51)
Warnerdebutrank	-213.67***	-71.49***	-35.84***	-22.49***
	(40.28)	(14.31)	(10.15)	(5.20)
Wkson	-649.94***	-198.06***	-98.20***	-43.04***
	(102.90)	(27.81)	(13.83)	(7.31)
Wksonsq	2.53***	0.69***	0.29***	0.11***
	(0.68)	(0.19)	(0.09)	(0.03)
Firstweek	30747.72***	3931.97***	784.99**	-443.09**
	(2472.38)	(590.70)	(322.55)	(176.43)
Firstalbum	-9489.51**	-1299.74	-1654.50**	-604.96
	(4326.66)	(1121.91)	(710.40)	(400.66)
Previousalbumduration	34.21	-24.52	-20.18***	-2.65
	(92.11)	(17.64)	(7.02)	(5.41)
Previousalbumsales	-0.81	0.15	0.32**	0.24***
	(1.60)	(0.34)	(0.13)	(0.09)
Wksonradio	181.09*	102.18***	74.14***	31.15***
	(96.15)	(32.71)	(21.53)	(11.94)
Lastweekradiorank	-31.46	-3.34	-3.47	-2.28
	(22.61)	(5.12)	(2.70)	(1.75)
Weeksinceradio	-434.42***	-166.78***	-121.25***	-77.55***
	(95.26)	(38.52)	(37.77)	(22.03)
Noradio	-6138.52**	-1261.59	-133.53	-313.11
	(2940.09)	(994.17)	(523.03)	(315.11)
N	17314	16373	15006	12693

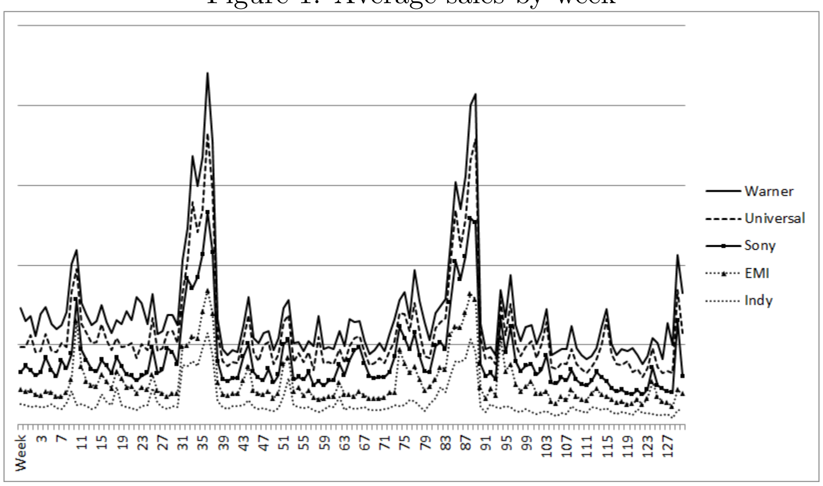
Interaction of the above ten covariates with the blackout as well as genre, label, and week dummies are not shown for brevity. Standard errors clustered by artists are in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 5: Sales regressions - imputed data

	(1) Full Sample	(2) Drop Top 10	(3) Drop Top 25	(4) Drop Top 50
Warnereffect	1342.34	1515.04	1040.97	-1320.62
	(3860.09)	(2361.66)	(2323.21)	(2346.42)
Wkson	-630.05***	-211.51***	-83.93*	12.55
	(98.46)	(49.13)	(46.03)	(59.12)
Wksonsq	2.60***	0.92***	0.44**	0.24
	(0.65)	(0.30)	(0.22)	(0.26)
Firstweek	23496.71***	4512.67***	2024.46	2014.54
	(1854.90)	(1306.05)	(1478.13)	(1593.35)
Firstalbum	-9942.28*	-2892.68	-4758.97	-4479.94
	(5961.91)	(4343.31)	(4294.16)	(4420.60)
Previousalbumduration	77.26	-25.15	-19.66	-3.50
	(103.16)	(39.79)	(36.60)	(42.51)
previousalbumsales	-1.31	0.89	0.58	1.05
	(1.57)	(0.93)	(0.77)	(0.81)
Wksonradio	54.95	43.97	57.41	101.89
	(188.05)	(164.14)	(177.29)	(189.10)
Lastweekradiorank	-46.70*	-22.77	-22.35	-6.56
	(24.16)	(17.08)	(19.14)	(19.04)
Weeksinceradio	-92.52	284.32	503.27	749.02
	(759.76)	(755.56)	(810.49)	(984.35)
Noradio	-9631.20**	-3791.90	-1162.17	159.88
	(4439.06)	(3261.33)	(3098.32)	(2935.04)
N	17314	16373	15006	12693

Interaction of the above ten covariates with the blackout as well as genre, label, and week dummies are not shown for brevity. Standard errors clustered by artists are in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Figure 1: Average sales by week



(Grid sizes have been removed due to proprietary data.)