Profit Leak? Pre-Release File Sharing and the Music Industry*

Robert G. Hammond[†]

February 27, 2012

*I thank Stanley Liebowitz, Stephen Margolis, Julian Sanchez, and Koleman Strumpf for comments; the staff at Nielsen Media & Entertainment (especially Sean Bradley) for assistance with sales data; and the Research Innovation Grant fund for financial support. The sales data presented in this paper are proprietary and were purchased from Nielsen Media & Entertainment under a non-disclosure agreement.

[†]Department of Economics, North Carolina State University. Contact: robert_hammond@ncsu.edu.

Profit Leak? Pre-Release File Sharing and the Music Industry

Abstract

It is intuitive that the potential for buyers to obtain a good without remuneration can harm the producer of the good. I test if this holds empirically in the music industry using data from an exclusive file-sharing website that allows users to share music files using the BitTorrent protocol. These are the most reliable file-sharing data available because they are from a private tracker, which is an invitation-only file-sharing network. To overcome the endogeneity of the number of downloads, I exploit variation in the availability of sound recordings in file-sharing networks to isolate the causal effect of file sharing on music sales. The results strongly suggest that the aggregate effect of file sharing is precisely estimated to be extremely small. The direction of the effect is positive but its magnitude indicates that file sharing does not increase or decrease aggregate sales in the music industry. I conclude with an investigation of the distributional effects of file sharing on sales and find that file sharing benefits more established and popular artists at the expense of newer and smaller artists. These results are consistent with recent trends in the music industry.

JEL classification: L82, K42, C26

Keywords: Sound recordings, intellectual property rights, distributional effects of file sharing, instrument exogeneity

1 Introduction

How does the existence of a black market affect outcomes in the formal market? The market of interest here is recorded music, a market that has featured prominently in the literature on intellectual property rights following the rise of file sharing. Revenues in the music industry fell 10.9% in 2010 to $6.9 billion, continuing a near-decade-long decline in revenue (RIAA, 2011a). Simultaneously, unauthorized sharing of sound recordings in file-sharing networks continues to increase, with 1.5 billion file-sharing web searches performed in 2009 (BBC News, 2010). The coincidence of these two phenomena has sparked a growing literature investigating the causal relationship between a sound recording's popularity in file-sharing networks and its sales in retail markets. I build on this literature with a paper that focuses on the sales of music albums, as opposed to individual tracks. An album-oriented approach is appropriate given that albums are the main source of music-industry revenue and are an increasing share of the digital market as well (Segall, 2012). While song sales play an increasingly important role in the music industry, albums remain the largest source of revenue, comprising 75.4% of total music-industry revenue in 2010 (up from 72.4% in 2008) (RIAA, 2011a). As a result, album sales should be the focus of the policy discussions that surround the music industry's legal campaign to protect its intellectual property rights against members of file-sharing networks.

I focus on pre-release file sharing, in which file sharers download sound recordings that are not yet publicly available. Industry representatives have referred to the pre-release period as "the most sensitive time of an album's sales cycle" and stated that "curbing pre-release piracy is a particular priority for the recording industry" (IFPI, 2011). Further, pre-release file sharing has been the focus of legal action against file-sharing websites in the United Kingdom (Juskalian, 2009) and Sweden (Lindenberger, 2009). Pre-release files are often made available "by music industry insiders – such as radio DJs, employees of music magazine publishers, workers at CD-manufacturing plants, and retailers who frequently receive advance copies of music" (Billboard, 2009). While a recent paper has focused on pre-release file sharing in the movie industry (Ma et al., 2011), this is the first paper that focused on the pre-release sharing of music files. To study this phenomenon, I must overcome the fact that file sharing is endogenous in its determination of an album's sales because an album that is popular in file-sharing networks will also be popular in retail markets. I address this endogeneity by exploiting exogenous variation in the date at which an album first became available in file-sharing networks and how widely available the album was prior to its official release date. These pre-release characteristics are correlated with an album's popularity in file-sharing networks but are uncorrelated with the album's sales in retail markets.

I use data on the date at which albums first become available in file-sharing networks, what is known as the album's "leak" date, and the album's ease of availability in file-sharing networks. These availability characteristics are positively correlated with an album's popularity in file-sharing networks because albums that "leak" earlier have a longer period during which they are available in file-sharing networks and are therefore more heavily downloaded. I argue that the leak date and availability are uncorrelated with an album's popularity in retail markets because a leak can be thought of a "crime of opportunity" that occurs at some stage of the CD production or album marketing process. According to one executive: "Leaks come from all over. Sometimes songwriters, sometimes press" (Wolk, 2007). If this common view is correct, then the variation in an album's leak date can be taken as plausibly exogenous, providing the basis for an appropriate instrument. Previous work has used a variety of instrumental variables to deal with this endogeneity, to varying degrees of success. Oberholzer-Gee and Strumpf (2007) use the presence of vacations among high-school students, which may allow students more time for file sharing. Zentner (2006), Rob and Waldfogel (2006), and Liebowitz (2008) use measures of high-speed Internet penetration, which may facilitate faster sharing of files. Blackburn (2006) and Bhattacharjee et al. (2007) use a dummy variable for days in their sample after the Recording Industry Association of America's (RIAA) announcement that it would pursue legal action against users in file-sharing networks, which may reduce file sharing.

The data used in this paper have two key advantages over that in the previous literature. First, given the illicit nature of file-sharing networks, it is difficult to obtain data on file sharing. My data source provides remote access to the activities of the largest network within the BitTorrent protocol, which is the largest component of the broad class of protocols that are jointly referred to file-sharing networks. BitTorrent accounted for between 27 and 55% of all Internet traffic in 2008 (Ipoque, 2009). Second, given the dispersed nature of file-sharing networks, it is difficult to make statements that are global to the entire file-sharing community using data that come only from a particular file-sharing application. My data source is the largest private tracker specializing in sound recordings as measured by the number of files in the site's database (565,277 albums from 440,573 artists, which have been downloaded 66.6 million times as of December 2011). A private tracker is an invitation-only file-sharing network, which implies that private trackers are small relative to the size of public trackers, a fact that is evident from the relatively small number of downloads per album on the tracker in question. However, I will argue that private trackers are of particular interest for understanding the effects of file sharing because they play a unique role in the initial appearance of sound recordings in both private and public file-sharing communities.

Theoretical work in the law and economics literature has shown that file sharing need not reduce the profits of incumbent producers. Peitz and Waelbroeck (2006) focus on the role of sampling, specifically that products are differentiated and consumers are uncertain which variety of the product they prefer. File sharing allows a consumer to sample several products at zero marginal cost. Willingness to pay can be higher under the certainty of demand that sampling affords than under taste uncertainty, even though some consumers will not purchase following sampling. The insights of Peitz and Waelbroeck (2006) also highlight the importance of an album's leak date if some consumers download in file-sharing networks only because albums are available earlier there than in retail markets. Takeyama (1994) focuses on markets with network externalities, where a consumer's willingness to pay for a good is increasing in the number of other consumers of the good. She shows that file sharing allows a firm to price discriminate between low-valuation consumers (who download and pay zero) and high-valuation consumers (who pay the market price). When network effects are strong, such price discrimination can raise profits and increase social welfare. Sound recordings exhibit network externalities due to bandwagon effects (Leibenstein, 1950) but whether externalities in this market are large is an empirical question.

Empirical work on file sharing has become increasingly popular (e.g., a special issue in the *Journal of Law and Economics* (Liebowitz, 2006)). While a number of product markets have been studied in this literature (such as Qian (2008) who studies counterfeit and authentic shoes produced in China), most empirical work uses data on the music industry because of the coincidence of its decline with the ascent of file-sharing networks. Unlike the present paper, Zentner (2005) and Liebowitz (2008) do not have data on file sharing of the music in their sample; instead, they use proxies that are correlated with file sharing and thus provide a less reliable estimation of its effect on sales. Gopal et al. (2006) use survey data, which are less precise than field data if the illegal nature of the activity in file-sharing networks leads some respondents to answer untruthfully.

I find that file sharing does not increase or decrease sales in the music industry but that this aggregate impact masks important distributional effects. The results suggest that established/popular artists benefit from file sharing, while new/small artists do not. In particular, file sharing redistributes sales toward artists in the pop genre and away from artists in niche genres such as indie music. Further, the effects of file sharing are meaningfully positive for artists who have had an album sell at least 100,000 units but not for artists who have not. Likewise, the effects are meaningfully positive for artists who have released more than three albums but not for newer artists. These distributional effects are likely to have important consequences for the ability of the music industry to continue to produce sound recordings that cater to a large variety of tastes among heterogeneous music consumers.

I now discuss the types of file-sharing networks that generate the data that are used.

2 Background on Private BitTorrent Trackers

There are numerous summaries of the birth and growth of file sharing and its relationship with changes in the music industry. See Juskalian (2009) for an overview in the popular press and Liebowitz (2006) for a more rigorous discussion. In lieu of a providing a similar discussion, I overview private BitTorrent trackers and then provide generic information on the tracker that served as the source of the file-sharing data used here. BitTorrent is a protocol for peer-to-peer network sharing and has become the most widely used method of sharing files following its initial release on July 2, 2001 (Cohen, 2008). BitTorrent differs from previous file-sharing services such as Napster in that BitTorrent files are shared by several hosts concurrently, which is more efficient than a single host at a time. Network users access a file in .torrent format with a BitTorrent client (e.g., Azureus/Vuze (azureus.sourceforge.net) or μTorrent (www.utorrent.com)), allowing the user's computer to establish a connection to the network to download the files and subsequently share the files with other users that access them.

Within a BitTorrent network, a download is defined as a completed transfer of the files included in the .torrent file to a user's computer. Once users have gained access to the .torrent file, they are called leechers while transferring the files to their computer and seeders after completing the transfer and remaining connected to the file-sharing network. Users who completely leech the files but do not seed are warned or punished if such "hit-and-run" leeching is detected; individual file-sharing communities differ in the extent to which such behavior is punished. In any case, there are no technological constraints that require users to seed the files that they have leeched and file-sharing communities have developed community norms and various enforcement mechanisms to encourage seeding. The main such mechanism is a user's share ratio, or ratio, which equals the amount of data that the user has uploaded divided by the amount of data that the user has downloaded. Since this is the ratio of the amount that is shared with leechers relative to the amount that was leeched, a user's ratio is expected to be above some minimum threshold to remain in good standing. The most restrictive file-sharing communities warn low-ratio users to raise their ratio within a certain interval of time and then ban users who fail to do so (Juskalian, 2009).

BitTorrent users search for files in one of two ways. First, users can search public trackers such as isoHunt (www.isohunt.com) or TorrentSpy (www.torrentspy.com), though note that well-known public trackers are often shut down by criminal or civil complaints. Files that are available on public trackers are often accessed indirectly by users who use standard search engines and limit their search results to include only .torrent files. For example, a user who searches Google for "Metallica filetype:torrent" will find (according to Google, as of this writing) "about 355,000 results" that link to .torrent files that include sound recordings from the band Metallica. A second way that users search for files is private trackers, which mimic public trackers but restrict entry to users that have met some criteria. Generally, users gain access to a private tracker by receiving an invitation from an existing user of the tracker, that is, an existing member of the particular community. Typically, the developers of a tracker begin the tracker in a beta period by limiting access to a small group of friends or fellow members of some other file-sharing community. The community expands through invitation periods, where existing members receive invitations that extend membership to others. Good file-sharing etiquette (such as maintaining a good ratio on other trackers) is often considered a prerequisite to receiving an invitation to a private tracker (Frucci, 2010).

Members of a given private tracker have an incentive to join other private trackers because private trackers tend to specialize in a certain category of files, such as music, movies, video games, animation cartoons, etc.[1] Members also seek out invitations to join other private trackers that tread much of the same ground (e.g., music trackers that have similar genre coverage). To discuss why this happens, consider all private trackers that specialize in music. A highly anticipated album may first appear in file-sharing networks on any of these private trackers. One of the first users of that tracker who downloads the album will upload it onto the other private trackers of which she is a member. The reason for doing this is that an uploader of an album receives a large increase in her ratio and a user's ratio is a metric of her stature in the file-sharing communities (including, again, a threshold for continued membership).[2]

There are a number of ways that files appear on file-sharing networks. For sound recordings, albums, singles, or individual tracks are first converted into an electronic format and then uploaded. Music that appears on file-sharing networks prior to its release is said to have leaked and leaks are a key part of my empirical strategy. How leaks happen is a contentious issue but, as discussed in the introduction, they are widely viewed as a "crime of opportunity" (Wolk, 2007). I will refer to the source of my file-sharing data in generic terms without naming the tracker or providing its web address. The need for anonymity should not be surprising given the illicit nature of the activities of the tracker's users. The BitTorrent tracker is question is a leading private tracker, among the largest private trackers that specialize in music files. It began after OiNK (formerly, www.OiNK.cd) was shut down by police in England and the Netherlands (Baker, 2007). OiNK was described as "the world's biggest source of pirated pre-release albums" (Harris, 2007). The tracker in question can be considered OiNK's closest descendant, though other private trackers were also developed following OiNK's demise. Various statistics support the interpretation that the tracker in question is an important part of the universe of private trackers that make music files available for downloading and sharing. Not surprisingly, market share or related formal measures of the relative importance of the tracker in question are scarcely available.

For details on the tracker in question, consider a snapshot as of December 2011: since its launch in the fourth quarter of 2007, the tracker had 66.6 million downloads of 565,277 albums

[1] See en.wikipedia.org/wiki/Comparison_of_BitTorrent_sites for several examples of public and private trackers, including their area of specialization.

[2] As an example, consider users 1-4, where 1 uploads an album and 2-4 download the album sequentially in ascending order. User 1 seeds to 2 but then 1 and 2 seed to 3, etc. This many-hosts aspect of BitTorrent is its key advantage over earlier file-sharing services. If users 2-4 download the album simultaneously, then they can upload parts of the album to each other, which is another advantage of BitTorrent in that files are split into smaller pieces. Reciprocal seeding and leeching is possible because pieces are shared with simultaneous leechers in a different order.

from 440,573 artists. The tracker has 148,465 users and, at a given moment, 9.012 million seeders compared to 173,049 leechers. In October 2011, there were around 4,000 newly registered members and a similar number of members that were disabled for rules violations (most of whom failed to maintain a minimum ratio). Imprecise information is available about the country of origin for the tracker's users but (based on the user's IP address) approximately 80,000 users were from the United States, 11,000 from Canada, and 8,000 from Great Britain. Next, in descending order, are Sweden, Australia, Russia, and the Netherlands. The operating system breakdown of these users is 64.0% Windows, 27.0% Mac, and 9.0% Linux/other, while the browser breakdown is 43.6% Firefox, 34.7% Chrome, 10.1% Safari, 4.8% Internet Explorer, and 6.8% other. It is clear from these operating system and browser statistics that the users of the tracker in question, and the users of file-sharing networks in general, are not representative of the population as a whole.

To check the representativeness of these data relative to other file-sharing networks, I compare the leak date in my data to the earliest appearance of the same albums on public BitTorrent trackers, specifically isoHunt, that have significantly more users but play a secondary role in the initial leaking of albums in the pre-release period.[3] A clear pattern is found: albums appear on private trackers first and are then soon uploaded on public trackers. For specific examples, consider the highest-selling albums in these data. Taylor Swift's *Speak Now* leaked on the tracker in question on October 22, 2010 and leaked publicly the next day on October 23. The same holds for Susan Boyle's *The Gift* (November 3, 2010 on the tracker in question versus November 7 publicly) and Jackie Evancho's *O Holy Night* (November 18, 2010 versus November 20). Further, the "private tracker first, public soon thereafter" pattern also holds for moderately popular albums such as Cake's *Showroom of Compassion* (December 31, 2010 versus January 3, 2011) and Amos Lee's *Mission Bell* (January 11, 2011 versus January 26) as well as less popular albums such as Stone Sour's *Audio Secrecy* (September 3, 2010 versus September 4) and Tuck From Hell's *Thrashing* (November 14, 2010 versus November 24). Several albums leaked on the same day privately and publicly, such as Bryan Adams's *Bare Bones* (October 21, 2010), Drake's *Thank Me Later* (June 2, 2010), and Nicki Minaj's *Pink Friday* (November 17, 2010).[4] As a result, I consider the tracker

[3] IsoHunt was a large, public file-sharing network during the sample period but has faced numerous lawsuits and is likely to be shut down. To ensure a global comparison, I also compare the tracker in question to the results from a Google search of the artist and album with the "filetype:torrent" qualifier.

[4] The main exception to the pattern is that a few albums leaked much later on public trackers than on the private tracker in question, including Kula Shaker's *Pilgrims Progress* (June 1, 2010 versus July 21) and Newsboys's *Born*

in question to be representative of the larger file-sharing universe.

3 Album-Level Data

These data contain all albums that were released between May 2010 and January 2011 (including re-releases of limited-release albums). By choosing all new albums, I avoid any potential biases in sample selection as long as the sample period is representative. Importantly, I include the fourth-quarter holiday shopping period as this period features a spike in music demand. There are 1,095 albums in the data set from 1,075 artists. The albums in the data set cover a variety of genres (see Table 1). Genre designations are derived from the BitTorrent tracker in question, where users vote on genre "tags" for each artist and for each album individually. I assign each album to the genre for which it received the most votes, aggregating related sub-genres. For example, metal albums are those whose most-voted genre matches either hard rock, metal, or punk, while country albums match either Americana, bluegrass, or country. This "crowdsourced" genre categorization has the advantage of representing the genre perceptions of listeners of the music.

An alternative source of genre categorization is available from Nielsen SoundScan, according to the chart on which the album is listed. I prefer the crowdsourced genre categorization because the SoundScan categorization provides little variation in that most albums are listed on the rock charts and this results in 60.9% of these albums being consider rock albums. When albums are not on the SoundScan rock chart, the crowdsourced and SoundScan categorizations of their genres generally agree. Some genres may require further explanation. The dance genre encompasses music from electronica to hip-hop that is centered around nightclub dancing. The indie genre takes its name from music that is recorded, produced, and released independently from major recording labels but the name has less relevance in the increasingly conglomerated music industry, where few artists operate in complete independence from large music corporations (Christman, 2011).

Further, these 1,095 albums were released by a variety of recording labels. In particular, 37.1% of the albums were released by the "Big 4" labels and their subsidiaries (EMI: 4.8% of these 1,095

Again (April 4, 2010 versus July 12). An example of an album that is hard to evaluate for this comparison is Eminem's *Recovery* because a "fake" copy of the album (i.e., unmastered and unfinished) leaked significantly before the album's official release. It is not obvious when the first real copy of the album appeared on isoHunt because fake torrents are not labeled as such. There is less of a problem with fake torrents on the tracker in question (or private trackers generally) because private trackers are quick to remove fake torrents.

albums, Sony Music: 7.8%, the Universal Music Group: 15.0%, and the Warner Music Group: 9.5%).[5] These albums will be designated as major-label albums, where major labels are typically defined as owning a distribution channel. Other albums are recorded and produced independently from major labels but are distributed by major labels. I designate these albums as major-label-distribution albums and they comprise 22.4% of the data set. The third label designation used here is independent-label albums, 40.6% of these albums. Independent labels that re-occur in these data include E1 Music (the largest independent recording label in the U.S.): five albums, Epitaph Records: eleven albums, Merge Records: six albums, Nuclear Blast: six albums, Vanguard Records: five albums, and Yep Roc Records: seven albums.

Other album-level covariates that are included in these data are the number of albums that the artist released prior to the album in these data, broken down by level of sales. I use these data to construct a variable for the total number of previous albums and the ratio of albums that sold at least 100,000 units to the total number of previous albums. I refer to the latter variable as an artist's (ex ante) popularity index, where a value of zero implies that none of the artist's previous albums sold at least 100,000 units, while a value of one implies that all of the artist's previous albums sold at least 100,000 units. Next, I include a dummy variable equal to one if the album was sold with a bonus DVD, which often includes live footage or documentary footage of the album's creation. Finally, I control for whether the album was re-released following an earlier, limited release of the album. Albums are often re-released when their initial release was on a smaller label with a limited distribution channel but the album achieved sufficient interest to warrant re-release.[6]

3.1 File-Sharing Data

The data collection process began on May 25, 2010, which is a Tuesday because new albums are released on Tuesdays in the U.S. I collected data on each album released that day by searching the BitTorrent tracker in question to obtain the following data, if the album had leaked: the day,

[5] Universal purchased EMI in November 2011 but I will discuss these labels as separate because they were separate firms during my sample period. The above shares of the data set are generally in line with market shares based on sales: EMI: 9.6% of U.S. music sales in 2005, Sony: 25.6%, Universal: 31.7%, and Warner: 15.0%. The smaller shares reflect the difference between ranking based on share of albums released (as in the text above) versus ranking based on share of sales.

[6] Re-released albums (2.4% of the data) are fundamentally different than the majority of the albums in data set and may be better treated separately. I control for re-release but all results are robust to excluding re-released albums or to interacting the re-release dummy variable with the number of downloads.

hour, and minute that the album leaked; the number of cumulative downloads of the album; the number of current seeders; and the number of current leechers. If the album had not leaked, then no data were available. On each successive Tuesday, I repeated the data collection process for the albums that were released that day and collected the number of cumulative downloads, current seeders, and current leechers for the albums that were released in the previous weeks. I followed each album for five weeks (i.e., through the fourth Tuesday following its release), which I argue is sufficiently long because majority of an album's downloads occur prior to or around its release. In particular, 65.6% of an album's downloads in the first month occur by the end of the first week following its release. Further, the median share by the end of the first week is 80.0% and the 75^{th} percentile is 91.6%. I define total downloads as the cumulative number of downloads during the period prior to and in the first four week following release.

The date that an album leaked will be referred to as its length and is the number of days that an album leaked before (if positive) or after (if negative) its release date. Of the 1,095 albums in the data set, 991 (90.5%) of the albums leaked and 655 (59.8%) of the albums leaked prior to their release date. Given that an album leaked, the median album leaks 3.7 days prior to its release date. The mean album leaks 7.7 days prior to its release date but this figure is inflated by re-released albums (2.4% of the data), which are outliers because they typically leaked around the date of their previous, limited release. The 25^{th} percentile of length is -1.7 or 1.7 days after release, while the 75^{th} percentile is 13.8 or just under two weeks prior to release. In words, most albums in the data set leak in the two weeks prior to or soon after their release date.

3.2 Sales Data

For each album in the data set, I purchased sales data from Nielsen SoundScan for each album in each week for each of the first six weeks following its release. I argue that six weeks are sufficient because, as with downloads, the majority of an album's sales occur by the end of the second week following its release: 38.5% of an album's sales in the first six weeks occur by the end of the first week following its release, while 55.8% occur within the first two weeks. I define total sales as the cumulative number of sales during the first six weeks following release. Importantly, these sales data also include digital sales. Digital album sales are the fastest growing segment of the music industry (8.4% higher in 2011 than 2010) and now comprise a majority of music sales (Segall, 2012).

4 Instrumental Variables Estimation

The econometric approach follows a generalized method of moments (GMM) instrumental variables (IV) estimation. In particular, I consider the model of Mullahy (1997), who considers moment conditions from a restriction of $E[y|X] = \exp(X\beta)$, where y is the outcome variable, X is a vector of regressors, β is a coefficient vector, and the error term enters multiplicatively. The resulting IV Poisson estimator is implemented in Stata 11.2 via *ivpois* (Nichols, 2007).

In this section, I discuss two possible instruments and the degree of which each are appropriate. An instrument must be strong (correlated with the endogenous variable, file sharing) yet must itself be exogenous (uncorrelated with the outcome variable, sales); neither instrument a priori dominates the other. The first instrument, an album's length, is the number of days that an album leaked before (if positive) or after (if negative) its release date. Length should be strongly correlated with downloads because albums that leak earlier have a longer period during which they are available in file-sharing networks and are therefore more heavily downloaded. Also, if (as I have argued above) leaks are "crimes of opportunities," then length will be exogenous in the determination of sales. A great deal of anecdotal evidence is available to support the notion that leaks occur in a sufficiently random way as to insure identification (Wolk, 2007; Crosley, 2008; Levine, 2008; New Musical Express, 2008; Williams, 2009).

The second instrument, an album's ratio, is the ratio of seeders of the album to leechers of the album at its release date. I add one to the number of leechers to handle albums that did not leak early and albums with only seeders at the release date (i.e., ratio $= \frac{\#\text{seeders}}{\#\text{leechers}+1}$). This definition generates a ratio of zero for albums that did not leak early and takes a maximum value of 1,099.5 seeders for every leecher, for the album *Kiss Each Other Clean* by the band Iron & Wine. Ratio should be strongly correlated with downloads because it affects the availability and the speed at which users can download the album. Also, an album's ratio is plausibly exogenous because it is influenced by the file-sharing etiquette of the album's downloaders. More specifically, ratio depends on whether or not users who downloaded the album remain connected to the file-sharing tracker and make the downloaded album available for continued leeching by other users. The factors that explain seeding behavior are a function of features of the file-sharing network to a much greater extent than they are a function of the artist or the album itself.

For more detail on the ratio instrument, consider the ratios of select high-selling albums in these data: Kanye West (188.9 seeders for every leecher) versus the smaller ratios for Lil Wayne (65.8) and Nicki Minaj (26.3) within the rap genre; Michael Jackson (109.7) versus Josh Groban (65.0) and Katy Perry (27.6) within pop; the Zac Brown Band (112.0) versus Keith Urban (28.3) and Sugarland (16.5) within country; and finally, the Black Eyed Peas (53.7) and Rihanna (40.3) versus Justin Bieber (26.0) within dance music. These ratios, and those shown in Table 1, defy any pattern of ratios across genres or conventional wisdom concerning which albums are popular in file-sharing networks. Instead, the ratio instrument determines an album's availability on the tracker in question and this is the only apparent channel through which ratio affects an album's popularity in either file-sharing networks or retail markets.

Both the length and ratio instrument rely on characteristics of the tracker in question during the period between the leak date and the release date. What if members of file-sharing networks simply download an album whenever it becomes available? In particular, what if file-sharing behavior is not affected by how many days it is available prior to its release date or how many seeders of the album there are? Oberholzer-Gee and Strumpf (2005, Appendix D) consider and reject this argument by documenting empirically that file sharers are impatient and quickly lose interest in an album. As a result of the immediate nature of music consumption, file sharers exhibit a high degree of impatience, implying that an album's leak date and its file-sharing availability (i.e., its ratio) can be an important determinant of its popularity in file-sharing networks. Further, I document that these instruments are meaningfully related to file sharing in these data.

Before presenting the estimation results, I highlight the conclusion of the discussion of instrumental validity in Section 5.3. The main results in the next section use only the ratio instrument because there is solid econometric support for its validity. However, I then show that the choice of instrument does not affect the statistical or quantitative significance of the results.

5 Does File Sharing Reduce Album Sales?

5.1 Summary Statistics and First-Stage Results

First, Table 1 gives an overview of the data set, organized by genres from most to least common in these data. Recall that I use a crowdsourced genre categorization from votes of the users of the BitTorrent tracker in question in order to most accurately represent the genre perceptions of listeners of the music. Table 1 provides an example observation from each genre, chosen arbitrarily. For each, I show the number of downloads, length (number of days between leak date and release date), ratio of seeders to leechers at release date, and sales of the album. These results suggest that downloads are correlated with both (1) sales (i.e., downloads appear to be endogenous) and (2) the two potential instruments: length and ratio.

Next, consider the first-stage regression results of the ratio instrument (ratio of seeders to leechers at the album's release date) on the number of times that an album was downloaded. First notice in Table 2 that the adjusted R^2 is sufficiently high at 0.445. More importantly, the first-stage $F = 89.1$ strongly rejects the null of weak identification (p-value = 0.00) and is well above the rule-of-thumb that F should exceed 10 for a strong instrument. The results suggest that one additional seeder for every leecher leads to 6.2 additional downloads, an effect that is highly statistically significant. The full set of album and artist-level controls are included but few appear to have a meaningful effect on downloads. While a full set of genre-classification dummies are included, only dance, indie, and rap/hip-hop music are statistically more popular in file-sharing networks than the baseline "Other" genre that includes genres that rarely appeared in these data, such as classical, comedy, and world music. The only other statistically significant finding is that albums released by the label Universal are more-heavily downloaded, an effect that does not have a clear explanation but does not appear to be driven by outliers. Before moving to the main results, note that an IV approach is needed because, as expected, downloads are found to be endogenous with a Wu-Hausman F statistic of 30.59, which rejects the null of exogeneity (p-value = 0.00).

5.2 Main Results

Table 3 presents the main results using the ratio of seeders to leechers at release date as an instrumental variable to isolate the causal impact of file sharing on sales in the music industry. The regressor of interest is shown in two forms: the number of downloads prior to the album's release date and the total number of downloads. The outcome variable is similarly shown in two forms: the number of sales in the first week following the album's release and the total number sales. The results show little variation across these four specifications, especially for the coefficient of interest on downloads.

The effect of file sharing on sales is precisely estimated to be extremely small. The direction of the effect is positive but its magnitude is small: one additional download results in between 5 and 15 additional sales. The economic significance of this finding strongly indicates that file sharing does not increase or decrease sales in the music industry. Specifically, use Model (4) to consider a substantial increase in the total number of downloads and its effect on the total number of sales. Moving from the 25^{th} percentile of downloads to the 75^{th} percentile of downloads is associated with a 7.3% increase in sales when measured at the mean of 36,888 sales. This effect requires a massive increase in an album's popularity in file-sharing networks, which demonstrates that a reasonably sized increase in downloads will have very little effect on sales.

Using Model (4) for concreteness, further results from Table 3 suggest that artists with more previous albums and artists whose previous albums were more popular produce better-selling albums. This latter result is especially unsurprising and suggests that an artist with two previous albums, both of which sold at least 100,000 units (i.e., popularity of 1.0), gains 13,649.3 sales relative to an artists with only one of two that met this threshold (i.e., popularity of 0.5). Albums that included a bonus DVD are not found to sell more or less than standard releases. Re-released albums sell meaningfully less than first-release albums, perhaps because these albums reached a considerable fraction of their primary audience in their initial release.

Characteristics of the label that distributed the album are strongly predictive of sales, with major-label albums outselling major-label-distribution albums (though not statistically so for every label) and major-label-distribution albums outselling independent-label albums (the omitted group). These results are in line with industry conventional wisdom in that independent-label artists who albums sell well are often signed by major labels for their next release (Christman, 2011). Within major labels, the results provide a ranking of Sony, Universal, Warner, then EMI but pairwise comparisons do not reveal statistically significant differences. The clearest implication here is the dominance of artists who are affiliated with major labels over artists who are not. This is not surprising.

Finally, I include the full set of genre categorizations in each model. The results are shown in a separate table for clarity but are from a single regression per model. Table 4 suggests that dance, folk, holiday, indie, and jazz music undersell the other category, while no genres significantly outsell the other category. This result depends on the inclusion of the index of artist ex ante popularity, shown in the previous table as robustly predictive of sales. If I exclude popularity, there is a much clearer (in a statistical sense) ranking of genres that upholds convention wisdom (e.g., pop and rap/hip-hop outsell the other category). This indicates that, once the model controls for past sales, genres are less predictive of sales than discussion in the popular press may suggest. Note though that the removal of the popularity index does not affect the statistical or economic significance of downloads on sales.

5.3 Instrumental Validity

In this section, I present strong support for the appropriateness of the ratio instrument and then compare the main results to results that use alternative instruments. To elaborate of the discussion of the first stage in Section 5.1, the ratio instrument is strongly correlated with downloads, alleviating concerns about underidentification or weak identification. The Cragg-Donald minimum eigenvalue statistic is 577.2, which soundly rejects the null of weak identification because it is well above the critical value of 16.4.

In this application, it is perhaps more important to test if the ratio instrument is itself endogenous. There is a growing literature on detecting endogenous instruments and I follow the recent work of Caner and Morrill (2011). The authors develop an approach for testing the relationship of an endogenous regressor with the outcome variable (where the true value of the coefficient is β_0) that simultaneously tests the correlation of the instrument and the unexplained component of the outcome variable (where the true value of the correlation is ρ_0). I present the 95% joint confidence intervals from the Caner and Morrill test in Figure 1, following their recommendation of searching over a range of $\rho_0 \in [-0.3, 0.3]$. The shaded area in this figure indicates combinations of β_0 and ρ_0 that cannot be rejected. Since I am interested in the value of β_0, I focus on the values of ρ_0 at which the main result no longer holds. In other words, how large a value of ρ_0 is needed to find that downloads are negatively and significantly related to sales?

An atheoretic way to roughly approximate ρ_0 is as follows: estimate sales using the number of downloads as well as all control variables, predict the sample residuals, and estimate the sample correlation between these residuals and the instrument: the ratio of seeders to leechers at release date. While this approach is not informative on whether the instrument is itself endogenous, it does guide me in looking at plausible values of ρ_0 when interpreting the results in Figure 1. The estimated correlation is 0.04, which is statistical indistinguishable from zero (p-value = 0.24). Based on this, the results in Figure 1 suggest that reasonably sized violations of perfect exogeneity of the instrument do not overturn the main result: the confidence intervals for the effect of downloads on sales includes or is bounded above zero. Only for a $\rho_0 > 0.16$ is the effect of downloads negative and statistically significant. In summary, only implausibly large violations of perfect exogeneity of the ratio instrument in which ratio is positively correlated with the unexplained component of sales would overturn the main result, leading me to conclude that the previous section's results are robust to concerns about the instrument.

Next, I consider whether alternative instruments provide the same conclusion of no quantitatively large effect of downloading on sales. First, I consider the length instrument described earlier: the number of days that an album leaked before (if positive) or after (if negative) its release date. Another instrument that I consider is the total number of seeders of other albums at the album's leak date. The population of seeders should be strongly correlated with downloads because it is a function of the thickness of this file-sharing network at the time the album in question appeared in the network. Also, because seeder population excludes the album in question, it is plausibly exogenous. Finally, I use two dummy variables: one for if the album leaked early (i.e., prior to its release date) and another for if the album leaked at all. It is possible that variation in the amount of time that an album leaked prior to its release date (length) is less important than simply whether or not the album was available at all prior to its release date, which argues that a leak-early dummy is a better instrument. The same argument taken further argues for the leak-ever dummy.

In Table 5, I present an alternative version of Model (4) from Table 3 with different instruments, each regression containing a single instrument. Column (1) matches the previous Model (4) but now shows the marginal effects from a log-transformed sales variable from a standard GMM IV model, instead of the less-standard GMM IV Poisson model. This demonstrates that the choice of the Poisson specification, while theoretically preferred to a log transformation (Wooldridge, 2002, p. 645), does not alter the results. Columns (2) – (5) use the following instruments: (2) length, (3) the number of seeders of other albums at the album's leak date, (4) dummy equal to 1 if the album leaked early, and (5) dummy equal to 1 if the album leaked at all. Column (2) includes fewer observations because only albums that ever leaked have information on their length and thus the instrument is missing for the 9.5% of albums that never leaked.

Note three points. First, the results are insensitive to the choice of instrument among these five alternatives in that the coefficient of interest remains statistically significant but quantitatively small. Other results from the log-transformed model generally match those from the Poisson model from Table 3 and different instruments do not affect the results from these control variables either. Second, the weak-instruments test statistic (Kleibergen-Paap rk Wald F statistic) suggests that ratio and the leak-early dummy are both strong instruments, while the leak-ever dummy is reasonably strong. In contrast, length and the seeder population instrument both appear to be weak. Third, while both ratio and the leak-early dummy are strong instruments, ratio is the preferred instrument from a Caner and Morrill (2011) bias-corrected test that accounts for the potential for non-exogeneity of the instrument. In words, there is strong econometric support that the ratio instrument is both strong and exogenous.

5.4 Panel Data

As a final robustness check, I exploit within-album variation using a fixed-effects model of how week-to-week variation in downloads is related to week-to-week variation in sales. A panel-data approach is advantageous because it does a better job of handling album-level unobservables than the main specification and thus provides cleaner identification. On the other hand, the policy discussions surrounding file sharing and sales concern the falling *level* of sales rather than *changes* in sales that are addressed using a fixed-effects model. At a minimum, how changes in sales depend on changes in downloads provides an interesting robustness check for the results on the levels of sales and downloads.

The panel-data results are in Table 6. No additional album or artist-level controls are included because these regressors do not vary week-to-week and are controlled for with album fixed effects. The instrumental variables used here are the ratio of seeders to leechers during the week in question and the first lag of this ratio. Tests for weak instruments indicate that the contemporaneous seeder/leecher ratio alone is weak (F = 1.9, p-value = 0.17) but, together with its first lag, the two instruments are not weak (F = 25.2, p-value = 0.00). Because evidence suggests that serial correlation is present, I present autocorrelation-consistent standard errors via the Bartlett kernel and a bandwidth of two. Neither the choice of kernel or bandwidth matters in the sense that the coefficient on downloads does not change by more than one-thousandth and remains statistically significant across alternatives. The results in Table 6 are consistent with the main results, suggesting that one additional download leads to 10 additional sales. As before, this positive effect is statistically significant but quantitatively small.

6 The Distributional Effects of File Sharing

Heterogeneity in the effects of file sharing are an important consideration: the effect of file sharing on sales is believed by industry insiders to be a function of an artist's previous sales history (Crosley, 2008; Youngs, 2009). There are two potential patterns that we may expect to emerge. Under the first hypothesis, artists with no proven track record of high sales may benefit from file sharing because it can generate "buzz" and build anticipation of the album to grow the artist's fan base (Peters, 2009). In contrast, established/popular artists may experience only the negative aspects of file sharing from the loss of sales. Under the second hypothesis, newer and smaller artists may be disproportionately hurt by file sharing, as is often claimed by representatives of the music industry, such as the International Federation of the Phonographic Industry (IFPI, an interest group that represents the music industry worldwide). "The music industry's greater loss of revenues due to piracy is having an impact on the success of new artists as investment comes under pressure. Consequently, fewer new acts are also breaking into the top selling charts" (IFPI, 2011).[7] While the music industry may have an incentive to overemphasize the harm to less established artists because they are more sympathetic than wealthy superstars, it is not clear a priori whether file sharing affects new/small artists differently than established/popular artists and, if so, which group benefits relatively.

Table 7 shows genre-specific results that match Model (4) from Table 3 from fifteen different regressions. Each regression includes only albums from a given of the fifteen genre in these data. Only the coefficient of interest (the effect of total downloads on total sales) is shown but all regressors from Table 3 are included in the model. To discuss, I split genres into three categories: genres where the effect of file sharing on sales in small (i.e., less than 10 additional sales per download):

[7] As stated by the RIAA: "artist rosters have been significantly cut back. The successful partnership between a music label and a global superstar – and the revenue generated – finances the investment in discovering, developing, and promoting the next new artist. Without that revolving door of investment and revenue, the ability to bring the next generation of artists to the marketplace is diminished – as is the incentive for the aspiring artist to make music a full time professional career." (RIAA, 2011b).

alternative, dance, folk, indie, jazz, and other; genres where the effect moderately large (i.e., between 10 and 100 additional sales per download): country, metal, rap, and rock; and a genre where the effect large (i.e., more than 100 additional sales per download): pop.[8] The most concrete way to interpret these results is to compare the genres where the effects are estimated to be essentially zero to the pop genre, which is the only genre that generates results consistent with precisely-estimated, large effects of downloading on sales. Table 7 then says that file sharing benefits pop albums at the expense of albums in niche genres such as indie music.

Next, Table 8 breaks the main results across more and less popular artists, while Table 9 breaks the main results across more and less established artists. The former comparison uses the index of artist ex ante popularity, based on sales of previous albums, and compares artists who have never had an album sell at least 100,000 units (i.e., popularity index of 0) in Column (1) to artists who have had an album reach that threshold (i.e., popularity index greater than 0) in Column (2). The latter comparison use the number of previous albums from the artist and compares artists with fewer than three previous albums in Column (1) to artists with at least three in Column (2). The findings in Table 8 suggest that the benefits of file sharing are larger for more popular artists than for less popular artists, with a point estimate that is four times larger ($t = 3.52$, p-value $= 0.00$). The findings in Table 9 suggest that the benefits of file sharing are larger for more established artists than for newer artists, with a point estimate that is nearly three times larger ($t = 2.53$, p-value $= 0.01$).

Finally, in Table 10, I re-estimate the main model after weighting the regression by an artist's past sales. The previous finding of a positive effect only for more popular artists suggests that the aggregate effect for the music industry should be larger than the effect from Section 5.2 because artists with more past sales are likely to sell more of their most-recent album, making these artists a larger share of the industry as a whole. As a result, weighting by past sales should increase the size of the positive but quantitatively small effect found in the main result. The weighted effect is indeed larger: one additional download leads to 37 additional sales when the regression is weighted by the artist's past sales.[9]

[8]While they are shown in the table for completeness, I do not discuss the genres whose sample sizes are too small to impart much confidence: blues (44 albums), gospel (13 albums), holiday (19 albums), and soul/R&B (42 albums).

[9]The construction of the weights is explained in the notes to Table 10. I essentially weight each album by the number of units that the artist sold previously. The weights are not exactly equal to past sales because I only have data on the number of the artist's past albums that met one of several sales thresholds and not the exact sales of

I discuss the implications of these results in the next section, where the findings are reconciled with the current state of the music industry.

7 Where this Leaves the Music Industry

I isolate the causal effect of file sharing on sales in the music industry by exploiting exogenous variation in the date at which an album first became available in file-sharing networks and how widely available the album was prior to its official release date. The rapid growth of file-sharing networks coupled with the rapid decline of profits in the music industry has sparked a great deal of interest in research on the connection between the two phenomena. I find that file sharing does not have a large, negative effect on sales. In no specification, using no instrument, do I find any evidence to the contrary. A slightly positive effect of file sharing on sales is consistent with Oberholzer-Gee and Strumpf (2007) and a quantitatively small effect is consistent with both Oberholzer-Gee and Strumpf (2007) and Blackburn (2006). In contrast, Liebowitz (2005) reviews the literature on file sharing and concludes that "the majority of studies find results supportive of the thesis that file sharing is causing harm."

Further in Section 6, I find that file sharing has benefited established/popular artists more so than new/small artists. The primary paper in the previous literature that finds distributional effects of file sharing between more and less popular artists is Blackburn (2006), who finds that file sharing is beneficial for less popular artists and harmful for more popular artists. Why do I find contrastingly that file sharing *benefits* more popular artists more so than less popular artists? My contention is that the file-sharing data that I use offer several advantages relative to those of Blackburn, which may explain much of the discrepancy. First, his data do not contain information on the number of downloads, only the number of files that are available. As a result, Blackburn can only discuss the availability and not the popularity of music in file-sharing networks. My data contain information on both the availability and popularity of music, which allows me to ask how an increase in the number of downloads affects sales. Second, Blackburn's file-sharing proxy is a stock variable, which is more difficult to correlate to the flow of sales, as opposed to my flow of downloads. Third, the instruments that Blackburn uses are dummy variables that jump from zero

those previous albums.

to one after the RIAA announced plans to pursue legal action against file sharers. I argue that my continuous instruments concerning pre-release file-sharing availability offer both econometric and theoretical improvements.[10]

Most importantly, the contrasting results of the present paper and Blackburn (2006) should be reconciled with recent trends in the industry, especially the trends since Blackburn's paper in 2006. I argue that these trends are consistent with file sharing having distributional effects that benefit established/popular artists. This is consistent with claims from representatives of the music industry. According to the IFPI, the cumulative sales of debut albums in the global top 50 fell by 77% between 2003 and 2010, substantially more than the 28% fall for non-debut albums. The share of debut albums in the global top 50 sales was 27% in 2003 but only 10% in 2010 (IFPI, 2011). In contrast, Leeds (2005) reports that artists on independent labels benefit from the Internet, focusing on the role of social networking and blogs in creating buzz for independent artists. He cites increasing market shares for independent labels as of 2005 but this trend did not continue to the present period. Consistent with the evidence that is presented in Section 6, Christman (2011) tabulates market shares by label type and finds falling market shares for independent-label albums (from 12.9% in 2007 to 12.5% in 2011) and for major-label-distribution albums (from 21.5% in 2007 to 18.7% in 2011). The only market share that saw an increase over this period was for major-label albums, which increased from 65.6% in 2007 to 68.2% in 2011.

While I have focused on the short-run consequences of file sharing on sales and the distribution of sales between new/small artists and established/popular artists, the long-run effects are equally important. To understand how a shift toward more established artists will affect the trajectory of the music industry, one must conjecture how major and independent labels will respond to the increasingly top-heavy landscape that is predicted by these findings. It is arguable that one should expect increasing concentration of recording and distribution labels and it would be worthwhile to investigate how much of the increased concentration that has already occurred can be explained by file sharing. While Waldfogel (2011) presents evidence that suggests that the quality of new recorded music has not fallen since the rise of file sharing, it is not clear what path we should expect

[10] The work of Mortimer et al. (2010) is related to that of Blackburn (2006) in that they use the same file-sharing data source. Mortimer et al. (2010) classify cities into high and low downloading cities and find that new/small artists benefited from their proxies for file sharing from increased concert revenue, while established/popular artists saw no effect on their concert revenues.

as we move further from the period in the music industry before file sharing existed.

References

Baker, L. (2007). Police Pull Plug On 'OiNK' Pre-Release Music Piracy Giant. *New Zealand Herald*, October 24. http://www.nzherald.co.nz/technology/news/article.cfm?c_id=5&objectid=10471743.

BBC News (2010). Music File-Sharer 'OiNK' Cleared of Fraud. http://news.bbc.co.uk/2/hi/uk_news/england/tees/8461879.stm.

Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R., and Telang, R. (2007). The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts. *Management Science*, 53(9):1359–1374.

Billboard (2009). Pre-Release Pirates Face the Music. http://www.billboard.com/bbcom/news/article_display.jsp?vnu_content_id=1002113928.

Blackburn, D. (2006). The Heterogenous Effects of Copying: The Case of Recorded Music. http://www.davidjhblackburn.com.

Caner, M. and Morrill, M. S. (2011). A New Paradigm: A Joint Test of Structural and Correlation Parameters in Instrumental Variables Regression When Perfect Exogeneity is Violated. http://www4.ncsu.edu/~msmorril/research.html.

Christman, E. (2011). What Exactly Is An Independent Label? Differing Definitions, Different Market Shares. *Billboard*, July 18. http://www.billboard.biz/bbbiz/industry/indies/what-exactly-is-an-independent-label-differing-1005281802.story.

Cohen, B. (2008). BitTorrent Protocol 1.0. *BitTorrent.org*, January 10. http://www.bittorrent.org/beps/bep_0003.html.

Crosley, H. (2008). Album Leaks: In Through the Out Door. *Billboard*, July 19.

Frucci, A. (2010). The Secret World of Private BitTorrent Trackers. *Gizmodo*, February 19. http://gizmodo.com/5475006/the-secret-world-of-private-bittorrent-trackers.

Gopal, R. D., Bhattacharjee, S., and Sanders, G. L. (2006). Do Artists Benefit from Online Music Sharing? *Journal of Business*, 79(3):1503–1533.

Harris, C. (2007). Music File-Sharing Site OiNK Shut Down Following Criminal Investigation. *MTV News*, October 23. http://www.mtv.com/news/articles/1572554/20071023/story.jhtml.

IFPI (2011). International Federation of the Phonographic Industry Digital Music Report 2011. www.ifpi.org/content/library/DMR2011.pdf.

Ipoque (2009). Internet Study. http://www.ipoque.com/resources/internet-studies/.

Juskalian, R. (2009). 10 Years After Napster, Online Pirates Alive and Well; Some Websites are Even Exclusive Clubs for Sharing Music and Videos. *USA Today*, June 24:5B. http://www.usatoday.com/tech/news/2009-06-23-online-files-napster_N.htm.

Leeds, J. (2005). The Net Is a Boon for Indie Labels. *New York Times*, December 27. http://www.nytimes.com/2005/12/27/arts/music/27musi.html?pagewanted=all.

Leibenstein, H. (1950). Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. *Quarterly Journal of Economics*, 64(2):183–207.

Levine, R. (2008). Despite Leaks Online and File Sharing, Lil Wayne's New CD Is a Hit. *New York Times*, June 18. http://www.nytimes.com/2008/06/18/arts/music/18wayne.html.

Liebowitz, S. J. (2005). Pitfalls in Measuring the Impact of File-Sharing on the Sound Recording Market. *CESifo Economic Studies*, 51(2-3):435–473.

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

Liebowitz, S. J. (2008). Testing File-Sharing's Impact on Music Album Sales in Cities. *Management Science*, 54(4):852.

Lindenberger, M. A. (2009). Internet Pirates Face Walking the Plank in Sweden. *Time Magazine*, February 20. http://www.time.com/time/business/article/0,8599,1880981,00.html.

Ma, L., Montgomery, A., Singh, P. V., and Smith, M. D. (2011). The Effect of Pre-Release Movie Piracy on Box-Office Revenue. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1782924.

Mortimer, J. H., Nosko, C., and Sorensen, A. (2010). Supply Responses to Digital Distribution: Recorded Music and Live Performances. Working Paper 16507, National Bureau of Economic Research. http://www.nber.org/papers/w16507.

Mullahy, J. (1997). Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior. *Review of Economics and Statistics*, 79(4):586–593.

New Musical Express (2008). New Metallica Album "Death Magnetic" Leaks. http://www.nme.com/news/metallica/39463.

Nichols, A. (2007). IVPOIS: Stata Module to Estimate an Instrumental Variables Poisson Regression via GMM. *Statistical Software Components*. http://ideas.repec.org/c/boc/bocode/s456890.html.

Oberholzer-Gee, F. and Strumpf, K. (2005). The Effect of File Sharing on Record Sales: An Empirical Analysis. Working paper version, http://www.unc.edu/~cigar/papers/FileSharing_June2005_final.pdf.

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

Peitz, M. and Waelbroeck, P. (2006). Why the Music Industry May Gain from Free Downloading–The Role of Sampling. *International Journal of Industrial Organization*, 24(5):907–913.

Peters, M. (2009). Leak Builds "Biltz!". *Billboard*, March 28.

Qian, Y. (2008). Impacts of Entry by Counterfeiters. *Quarterly Journal of Economics*, 123(4):1577–1609.

Recording Industry Association of America (2011a). 2010 Year-End Shipment Statistics. http://www.riaa.com/keystatistics.php.

Recording Industry Association of America (2011b). What is Online Piracy? http://www.riaa.com/physicalpiracy.php.

Rob, R. and Waldfogel, J. (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.

Segall, L. (2012). Digital Music Sales Top Physical Sales. http://money.cnn.com/2012/01/05/technology/digital_music_sales/index.htm?hpt=hp_t3.

Takeyama, L. N. (1994). The Welfare Implications of Unauthorized Reproduction of Intellectual Property in the Presence of Demand Network Externalities. *Journal of Industrial Economics*, 42(2):155–166.

Waldfogel, J. (2011). Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster. Working Paper 17503, National Bureau of Economic Research. http://www.nber.org/papers/w17503.

Williams, P. (2009). Safeguarding Unreleased Material Is Getting Tougher. *Music Week*, August 29.

Wolk, D. (2007). Days of the Leak. *Spin*, July 31. http://www.spin.com/articles/days-leak.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Youngs, I. (2009). Bands "Better Because of Piracy". *BBC News*, June 12. http://news.bbc.co.uk/2/hi/8097324.stm.

Zentner, A. (2005). File Sharing and International Sales of Copyrighted Music: An Empirical Analysis with a Panel of Countries. *Topics in Economic Analysis & Policy*, 5(1):21.

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

Table 1: Music Genres in the Data Set

		Example Observation							
				Downloads		Instruments		Sales	
Genre	Share	Artist	*Album*	Release	Total	Length	Ratio	First Week	Total
Alternative	14.7%	Kings of Leon	*Come Around Sundown*	3,895	4,322	16.5	311.4	184,099	378,367
Dance	10.9%	Ke$ha	*Cannibal*	512	774	4.6	52.8	74,217	226,987
Indie	9.5%	Arcade Fire	*The Suburbs*	4,192	7,606	7.0	291.2	156,079	297,586
Pop	9.3%	Susan Boyle	*The Gift*	35	113	5.7	13.5	317,895	1,684,400
Rock	9.2%	Tom Petty	*Mojo*	325	713	3.0	192.0	125,126	228,097
Metal	8.7%	Ozzy Osbourne	*Scream*	339	535	4.1	92.5	81,493	165,639
Country	7.9%	Taylor Swift	*Speak Now*	1,135	1,526	3.2	65.4	1,046,718	2,147,103
Folk	5.5%	Ray LaMontagne	*God Willin' . . .*	242	1,014	3.5	117.0	64,162	148,938
Jazz	4.6%	Fourplay	*Let's Touch the Sky*	13	17	27.5	6.0	2,704	10,438
Rap/Hip-Hop	4.6%	Eminem	*Recovery*	6,874	8,140	13.6	136.9	741,413	1,825,307
Blues	4.0%	Eric Clapton	*Clapton*	149	331	1.1	28.0	47,382	105,943
Soul/R&B	3.8%	Jamie Foxx	*Best Night of My Life*	124	191	4.0	18.3	143,657	249,859
Holiday	1.7%	Mariah Carey	*Merry Christmas II You*	36	174	5.2	8.0	55,447	289,461
Gospel	1.2%	Natalie Grant	*Love Revolution*	0	14	−22.1	0.0	12,467	24,311
Other	4.5%	Gaelic Storm	*Cabbage*	23	37	3.2	13.0	5,783	12,113

Notes: Albums are categorized into a genre according to votes by users of the BitTorrent tracker in question. Length is interpreted as the number of days that an album leaked before (if positive) or after (if negative) its release date. Ratio is the ratio of seeders to leechers of the album at its release date.

Table 2: First-Stage Regression Results

	(1)
Ratio of Seeders to Leechers at Release Date	6.227
	(0.663)***
Number of Previous Albums	1.169
	(0.961)
Artist Popularity Index	161.191
	(110.090)
Includes Bonus DVD	-38.852
	(104.601)
Re-released Album	38.064
	(109.690)
Label=EMI	-67.551
	(85.408)
Label=Sony	-3.355
	(63.303)
Label=Universal	215.406
	(71.881)***
Label=Warner	34.041
	(73.367)
Major-Label Distribution	2.943
	(47.862)
Dance	405.803
	(97.048)***
Indie	189.553
	(78.739)**
Rap/Hip-Hop	841.552
	(251.922)***
Constant	-80.119
	(44.262)*
Observations	1095
Adjusted R^2	0.444

Notes: Robust standard errors are in parentheses; $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% level, respectively. Only the genres whose coefficients were statistically significant are shown in the table but the full set of genre dummy variables are included. The omitted genre in these results is the "Other" category that includes genres that rarely appeared in these data, such as classical, comedy, and world music.

Table 3: Instrumental Variables Regression Results

	Sales in First Week		Sales, Total	
	(1)	(2)	(3)	(4)
Downloads at Release Date	0.008		0.016	
	(0.002)***		(0.003)***	
Downloads, Total		0.005		0.011
		(0.001)***		(0.002)***
Number of Previous Albums	0.083	0.072	0.190	0.171
	(0.020)***	(0.015)***	(0.042)***	(0.035)***
Artist Popularity Index	11.454	10.352	28.718	26.427
	(1.110)***	(0.886)***	(2.846)***	(2.429)***
Includes Bonus DVD	-0.715	-0.337	-1.739	-0.994
	(0.844)	(0.791)	(1.978)	(1.929)
Re-released Album	-7.819	-6.521	-16.669	-14.143
	(1.497)***	(1.263)***	(3.562)***	(3.180)***
Label=EMI	4.758	4.476	10.877	10.379
	(0.883)***	(0.776)***	(2.014)***	(1.830)***
Label=Sony	7.907	6.816	18.968	16.760
	(1.126)***	(0.873)***	(2.766)***	(2.290)***
Label=Universal	5.973	5.514	14.804	13.871
	(0.881)***	(0.803)***	(2.337)***	(2.228)***
Label=Warner	6.693	5.592	14.987	12.724
	(1.145)***	(1.053)***	(2.627)***	(2.474)***
Major-Label Distribution	3.431	2.976	7.915	6.980
	(0.588)***	(0.490)***	(1.397)***	(1.204)***
Observations	1095	1095	1095	1095

Notes: Each model uses an IV Poisson estimation. Average marginal effects are shown along with delta-method standard errors.

Table 4: Instrumental Variables Results for Genre

	Sales in First Week		Sales, Total	
	(1)	(2)	(3)	(4)
Alternative	-1.536	-1.642	-4.513	-4.991
	(1.462)	(1.333)	(3.756)	(3.549)
Blues	-0.832	-0.990	-2.638	-3.028
	(1.835)	(1.632)	(4.284)	(3.939)
Country	1.350	1.415	3.176	3.140
	(1.463)	(1.369)	(3.744)	(3.587)
Dance	-3.482	-3.376	-8.893	-9.078
	(1.630)**	(1.501)**	(4.000)**	(3.825)**
Folk	-2.913	-3.304	-7.622	-8.614
	(1.636)*	(1.419)**	(4.119)*	(3.695)**
Gospel	1.296	1.403	3.605	3.629
	(2.102)	(1.921)	(5.371)	(5.017)
Holiday	-4.342	-3.960	-4.100	-3.942
	(1.858)**	(1.699)**	(4.652)	(4.335)
Indie	-3.542	-3.779	-9.897	-10.529
	(1.522)**	(1.413)***	(3.835)***	(3.683)***
Jazz	-3.436	-3.048	-6.312	-5.847
	(1.594)**	(1.458)**	(4.063)	(3.826)
Metal	0.278	0.413	-1.208	-1.033
	(1.526)	(1.403)	(3.775)	(3.575)
Pop	1.101	1.187	2.389	2.526
	(1.529)	(1.420)	(3.919)	(3.756)
Rap/Hip-Hop	0.338	0.325	-0.355	-0.654
	(1.556)	(1.449)	(3.956)	(3.785)
Rock	0.523	0.659	-0.349	-0.171
	(1.639)	(1.531)	(4.033)	(3.848)
Soul/R&B	1.676	1.784	2.701	2.747
	(1.551)	(1.431)	(3.933)	(3.722)
Observations	1095	1095	1095	1095

Notes: These results are from the same four models in the previous table, broken into two tables for ease of presentation. Average marginal effects are shown along with delta-method standard errors.

Table 5: Results with Alternative Instrumental Variables

	(1)	(2)	(3)	(4)	(5)
Downloads, Total	0.005	0.022	0.017	0.014	0.029
	(0.001)***	(0.007)***	(0.004)***	(0.002)***	(0.006)***
Number of Previous Albums	0.130	0.088	0.085	0.093	0.039
	(0.017)***	(0.037)**	(0.027)***	(0.020)***	(0.036)
Artist Popularity Index	17.237	12.771	12.594	13.425	7.812
	(1.158)***	(4.187)***	(3.061)***	(2.102)***	(4.565)*
Includes Bonus DVD	-0.201	-1.943	-0.940	-0.807	-1.701
	(1.000)	(2.186)	(1.614)	(1.426)	(2.887)
Re-released Album	-4.569	-4.658	-4.087	-4.173	-3.591
	(0.929)***	(2.784)*	(2.104)*	(1.790)**	(3.805)
Label=EMI	5.745	6.180	5.313	5.390	4.867
	(0.969)***	(2.101)***	(1.481)***	(1.307)***	(2.632)*
Label=Sony	7.688	10.071	8.478	8.336	9.291
	(1.066)***	(2.055)***	(1.422)***	(1.305)***	(2.303)***
Label=Universal	6.008	3.587	3.675	4.093	1.272
	(0.759)***	(2.088)*	(1.581)**	(1.188)***	(2.424)
Label=Warner	5.722	3.523	3.314	3.745	0.834
	(0.824)***	(2.553)	(1.680)**	(1.329)***	(2.772)
Major-Label Distribution	3.328	3.945	3.318	3.320	3.308
	(0.479)***	(1.358)***	(0.916)***	(0.800)***	(1.650)**
Observations	1095	991	1095	1095	1095
F statistic	88.168	9.425	10.998	113.684	33.707
P-value	0.000	0.002	0.001	0.000	0.000

Notes: Each regression has a single instrument as follows: (1) the ratio of seeders to leechers at release date, (2) length of time that the album leaked prior to release date, (3) the number of seeders of other albums at the album's leak date, (4) dummy equal to 1 if the album leaked early, and (5) dummy equal to 1 if the album leaked at all. The F statistic is the first-stage Kleibergen-Paap rk Wald statistic that rejects the null of weak identification when the p-value is below 0.05. Each model uses a GMM IV estimation with Log($Sales$) as the dependent variable. To convert back into terms of $Sales$ rather than logs, average marginal effects are shown along with delta-method standard errors.

Table 6: Panel-Data Results for Sales by Week

	(1)
Downloads by Week	0.010
	(0.004)**
Observations	4380
F statistic	25.181
P-value	0.000

Notes: No other regressors are included in the fixed-effect model as they do not vary over time.

Table 7: Heterogeneous Effects by Genre

	(1) Downloads, Total
Alternative	0.006
	(0.002)***
Blues	0.025
	(0.006)***
Country	0.060
	(0.021)***
Dance	0.004
	(0.001)***
Folk	0.002
	(0.003)
Gospel	-0.153
	(0.138)
Holiday	0.078
	(0.039)**
Indie	0.003
	(0.001)***
Jazz	0.008
	(0.004)**
Metal	0.024
	(0.005)***
Other	0.003
	(0.004)
Pop	0.109
	(0.063)*
Rap	0.020
	(0.009)**
Rock	0.025
	(0.019)
Soul	0.100
	(0.083)

Notes: These results are from fifteen different regressions, each of which follows Model (4) from Table 3. Each regression includes only albums from the displayed genre. Only the coefficient of interest (the effect of total downloads on total sales) is shown but all regressors from Table 3 are included in the model.

Table 8: Heterogeneous Effects by Popularity Level

	(1) Less Popular	(2) More Popular
Downloads, Total	0.006	0.028
	(0.001)***	(0.008)***
Number of Previous Albums	0.108	0.392
	(0.034)***	(0.148)***
Artist Popularity Index		110.252
		(16.964)***
Includes Bonus DVD	0.014	-5.958
	(1.212)	(12.373)
Re-released Album	-4.422	-86.761
	(1.200)***	(38.543)**
Label=EMI	2.255	46.769
	(0.838)***	(11.496)***
Label=Sony	6.446	73.272
	(1.061)***	(12.496)***
Label=Universal	5.729	48.846
	(1.110)***	(10.126)***
Label=Warner	5.005	58.849
	(1.457)***	(12.758)***
Major-Label Distribution	2.079	38.868
	(0.479)***	(11.902)***
Observations	690	405

Notes: Each regression follows Model (4) from Table 3. Column (1) includes only albums by artists where none of the artist's previous albums sold at least 100,000 units (i.e., popularity index of 0). Column (2) includes only albums by artists where some of the artist's previous albums sold at least 100,000 units (i.e., popularity index greater than 0).

Table 9: Heterogeneous Effects by Number of Previous Albums

	(1) Fewer Previous Albums	(2) More Previous Albums
Downloads, Total	0.006	0.017
	(0.002)***	(0.004)***
Number of Previous Albums	1.409	0.218
	(0.347)***	(0.052)***
Artist Popularity Index	3.686	65.801
	(0.946)***	(6.460)***
Includes Bonus DVD	0.602	1.639
	(1.212)	(4.341)
Re-released Album	-4.551	-43.101
	(1.217)***	(7.075)***
Label=EMI	2.797	18.090
	(0.880)***	(4.329)***
Label=Sony	8.308	19.652
	(1.167)***	(3.865)***
Label=Universal	6.763	18.095
	(1.364)***	(3.429)***
Label=Warner	6.176	18.326
	(1.679)***	(4.340)***
Major-Label Distribution	2.571	12.420
	(0.637)***	(3.106)***
Observations	541	554

Notes: Each regression follows Model (4) from Table 3. Column (1) includes only albums by artists with fewer than three previous albums. Column (2) includes only albums by artists with three or more previous albums.

Table 10: Results Weighted by an Artist's Past Sales

	(1)
Downloads, Total	0.037
	(0.010)***
Number of Previous Albums	0.959
	(0.224)***
Artist Popularity Index	148.585
	(28.977)***
Includes Bonus DVD	11.774
	(22.373)
Re-released Album	-124.395
	(24.012)***
Label=EMI	106.551
	(21.312)***
Label=Sony	119.349
	(24.756)***
Label=Universal	94.946
	(23.218)***
Label=Warner	111.605
	(28.800)***
Major-Label Distribution	89.075
	(22.549)***
Observations	1095

Notes: The model uses a GMM IV estimation with Log(*Sales*) as the dependent variable. To convert back into terms of *Sales* rather than logs, average marginal effects are shown along with delta-method standard errors. The results are comparable with Model (1) in Table 5, which also uses a GMM IV estimation. The weights that are used are constructed as follows: add one to the weight for each previous album from the artist that sold less than 1,000 units, then add the lower bound of the sales interval for each of the artist's previous albums that fell in each of the following intervals: 1,000-10,000, 10,000-100,000, 100,000-1,000,000, and 1,000,000-above. As an example, rap/hip-hop artist Nappy Roots released the album *The Pursuit of Nappyness*, after releasing four previous albums, with exactly one album in each of the above four sales intervals. This observation takes a weight of 1,111,000, which equals $1 \times 1,000 + 1 \times 10,000 + 1 \times 100,000 + 1 \times 1,000,000$.

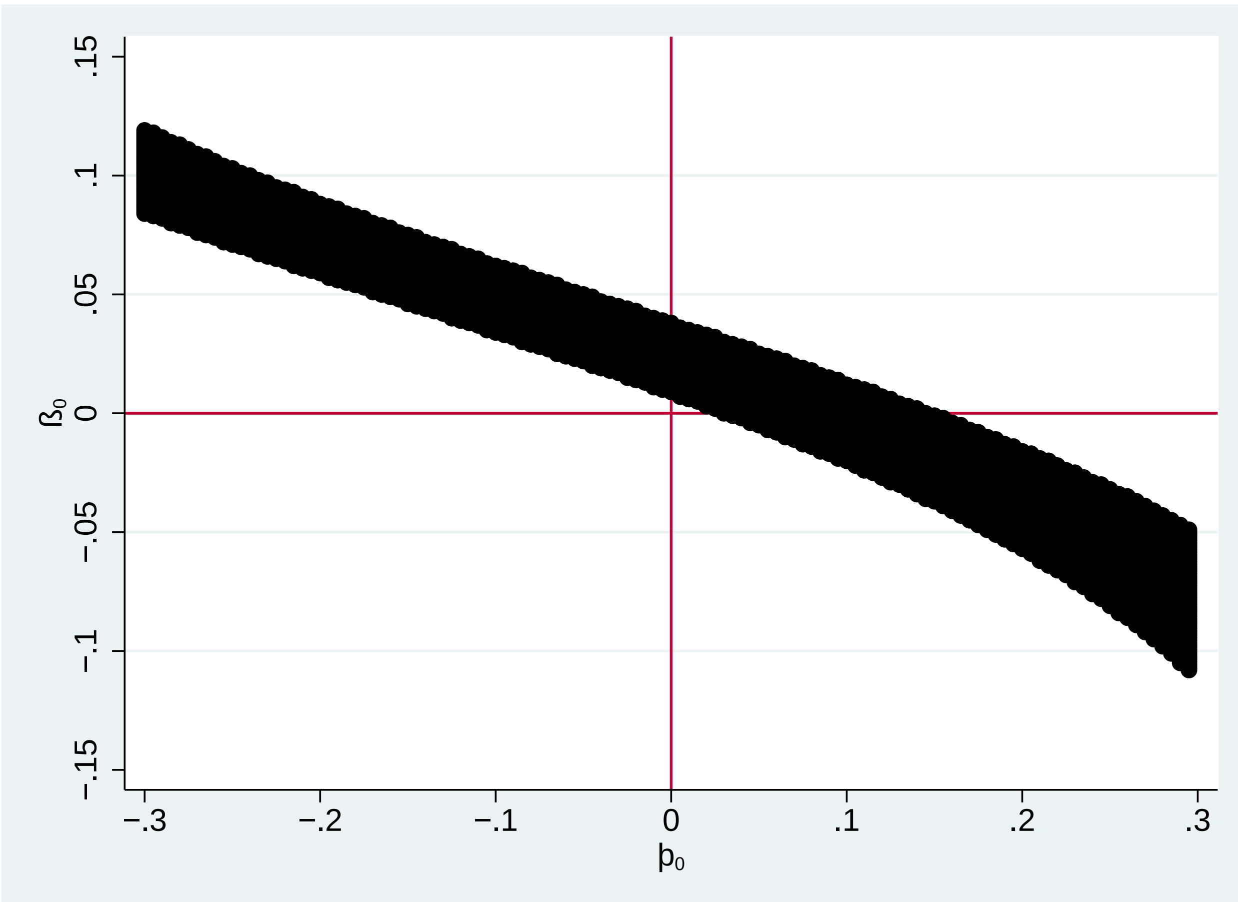

Figure 1: Joint Confidence Intervals for Testing Instrumental Validity

Notes: Following Caner and Morrill (2011), the shaded area indicates combinations of ρ_0 and β_0 that cannot be rejected at the 95% confidence level, where ρ_0 is the correlation between the ratio instrument and the unexplained component of sales and β_0 is the effect of downloads on sales.