

# On-line Piracy and Recorded Music Sales

David Blackburn\*  
Harvard University

First Draft: April, 2004 This Draft: December, 2004

## Abstract

Ever since the introduction of Napster, the impact of file sharing on the music industry has been the focus of intense debate. The availability of songs on file sharing networks has two competing effects on sales that are likely to vary across artists. First, there is a direct substitution effect on sales as some consumers download rather than purchase music. Second, there is a penetration effect which increases sales, as the spread of an artist's works helps to make the artist more well-known throughout the population. The first effect is strongest for ex ante well-known artists, while the second is strongest for ex ante unknown artists. Thus file sharing reduces sales for well-known artists relative to unknown artists. Taking account of this heterogeneity in estimating the effect of file sharing provides strong evidence of this distributional effect. Additionally, I find a large aggregate negative effect on sales not apparent in previous work that failed to account for the differential impacts on more and less well-known artists. The overall negative impact of file sharing arises because aggregate sales are dominated by sales of well-known artists. Using my estimates of the effect of file sharing, counterfactual exercises suggest that the lawsuits brought by the RIAA have resulted in an increase in album sales of approximately 2.9% during the 23 week period after the lawsuit strategy was publicly announced. Furthermore, if files available on-line were reduced across the board by 30%, industry sales would have been approximately 10% higher in 2003.

---

\*Department of Economics. Email: [blackb@fas.harvard.edu](mailto:blackb@fas.harvard.edu). An updated version of the paper may be found at <http://www.economics.harvard.edu/~dblackbu/papers.html>. I would like to thank Mariana Colacelli, Jan De Loecker, David Evans, Kate Ho, Joy Ishii, Larry Katz, Bryce Ward, and participants at the Harvard IO Workshop and the 2004 International Industrial Organization Conference for helpful suggestions. Special thanks to Gary Chamberlain, Julie Mortimer, and Ariel Pakes for their advice and encouragement. Additionally, I am indebted to Eric Garland and Adam Toll at BigChampagne and Rob Sisco at Nielsen SoundScan for providing access to themselves and their data, without which this project would have been impossible. I stake sole claim to any remaining errors.

"(Napster) helped me on this first album because nobody knew about it. It made it easier for people to know about the music. Once you get successful and you get another album, you want to start safeguarding it."

*-Josh Kelley, Hollywood Records Recording Artist<sup>1</sup>*

## **1 Introduction**

Ever since the introduction of Napster gave consumers the ability to trade digital music files across the internet, the impact of file sharing on the sales of music has been the focus of intense debate. To some, file sharing has been the root cause for the recent decline in the size of the music industry, while others contend that file sharing need not necessarily cause sales to fall. This ambiguity concerning even the signs of the effect stem from the fact that economic theory can not tell us whether the potential positive effects of file sharing are stronger or weaker than the potential negative effects; thus the question of whether or not file sharing helps or hurts the sales of recorded music is an empirical question. Nevertheless, there is reason to reexamine the theoretical underpinnings of this debate. In particular, while the discussion up to this point has been implicitly assuming that the effect of file sharing is uniform across artists, economic theory tells us that this is not true. Therefore, the previous literature examining these effects has been misguided. In this paper, I reexamine the economic theory of copying and file sharing and investigate what the effects of file sharing have been on the recorded music industry, how they differ across artists, and what the impacts are on the distribution of outcomes in the industry in the short run.

File sharing burst into the public consciousness in May of 1999, with the release of the software program Napster, which provided a simple to use interface with which consumers of music could share and download digital copies of songs. Napster became a huge suc-

---

<sup>1</sup>Fuoco 2003

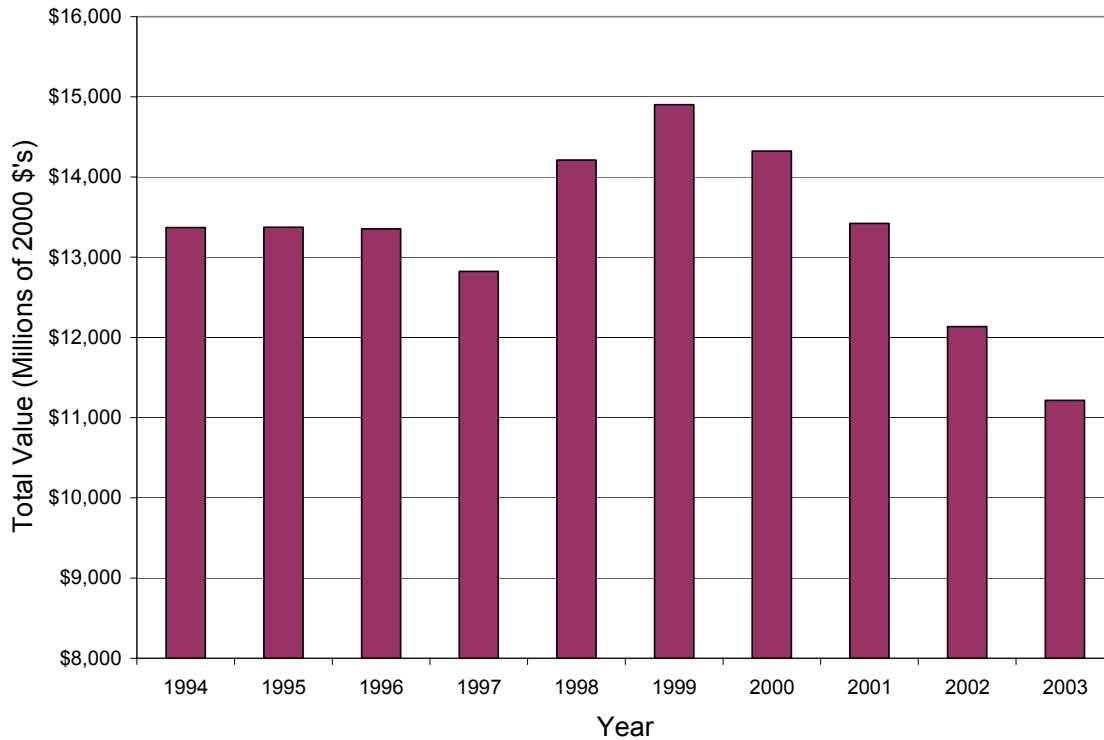


Figure 1: Total Real Value of Music Shipments by Year (RIAA 2003)

cess, with a reported user base of over 20 million unique user accounts worldwide at its peak, with routinely more than 500,000 unique IP addresses connected at any time (CNN-Money 2000). Up to the introduction of Napster, the recorded music industry in the United States was experiencing a huge period of growth, as Figure 1 demonstrates. However, the gains made in the years prior to 1999 quickly disappeared and industry sources were quick to attribute this decline to the rapidly increasing popularity of Napster. As a result, the Recording Industry Association of America (RIAA) in December of 1999 filed suit in U.S. District Court (RIAA 1999) to have Napster dismantled. This began a long line of lawsuits which resulted in the end of Napster, although file sharing has continued on many other networks since then.

As file sharing continued to grow in size and scope, and the industry continued to see

declines in sales over a period of several years, the RIAA turned in 2003 towards suing individual participants in file sharing networks. On June 25, 2003, the RIAA announced publicly that it would “begin gathering evidence and preparing lawsuits against individual computer users who are illegally offering to ‘share’ substantial amounts of copyrighted music over peer-to-peer networks” (RIAA 2003). Not surprisingly, this announcement caused a substantial drop in file sharing activity, as many consumers presumably became concerned about the risk of being sued for potentially thousands of dollars. The RIAA then followed through with their threats against consumers, filing the first wave of lawsuits against file sharing users on September 8, 2003. The RIAA focused their attention on “major offenders who have been illegally distributing substantial amounts (averaging more than 1,000 copyrighted music files each)” (RIAA 2003). This focus on “major offenders” meant that many casual users who initially abandoned file sharing for fear of being sued returned to the file sharing networks.<sup>2</sup>

Despite the very public debate about the effects of file sharing on the sale of recorded music, previous work on this relationship is relatively sparse and here I provide a quick summary.<sup>3</sup> The first attempt at measuring the effect of file sharing on sales was contracted by the RIAA for their lawsuit against Napster. In this study, Nielsen SoundScan applied what amounts to a difference-in-differences estimator to measure the changes in music sales between 1997 and 2000 in areas around college campuses and areas not around college campus. They found much larger drops in sales in the areas around college campuses, attributing this change to the effects of Napster (Fine 2000). More recent analyses have been done by Zentner (2004) and Oberholzer and Strumpf (2004), and have come to conflicting results. Zentner, using a panel of European country-level data argues that by exploiting

---

<sup>2</sup>Since that time, the RIAA has continued to file lawsuits against heavy users of file sharing networks.

<sup>3</sup>Mortimer and Sorensen (2004) study the relationship between digital distribution, both legal and illegal, and concert sales and pricing.

cross-country differences in broadband internet access<sup>4</sup> as well as some individual-level survey data, he is able to determine that the usage of file sharing networks reduces the probability of purchasing music by 30%.<sup>5</sup>

Oberholzer and Strumpf's recent paper has received the most attention, including a lengthy article in the New York Times (Schwartz 2004) concerning their results. Using album-level data on sales and file sharing activity similar to that used in this paper, they contrastingly find that file sharing has had no statistically significant effect on the sales of music. While this result has garnered a lot of attention, and the ire of the RIAA, there are outstanding questions regarding their ability to control for the simultaneity of sales and file sharing activity (Liebowitz 2004).

As discussed above, it is important to note that the effects of file sharing on sales of recorded music are extremely unlikely to be consistent across artists, and therefore it is vital to identify these differences to get an accurate representation of the effects. In particular, the effect of file sharing on sales depends on the ex ante popularity of the artist in question. Artists who are unknown can benefit from the awareness created by the spread of their music to a greater extent than ex ante well-known artists can, and similarly are less likely to lose sales to downloads, as they start with less sales.

I use a data set combining data on national-level sales data with data on file sharing activity over more than 60 weeks between September 2002 and November 2003 combined with various artist-level controls which are used to differentiate among groups of artists. The time frame of this data allows me to use the changes in the behavior of consumers on file sharing networks that stem from these lawsuits launched by the RIAA to address the endogeneity between file sharing activity and sales. This identification then allows me to

---

<sup>4</sup>Broadband internet access is potentially important for the use of file sharing networks, as it can greatly increase the speed at which files can be downloaded.

<sup>5</sup>A recent NBER working paper from Rob and Waldfogel (2004) finds that among a sample of college students, each album download reduces purchases by about 0.2.

estimate how file sharing has impacted the sales of recorded music.

The results suggest that not accounting for the heterogeneity of ex ante artist popularity results in inaccurate estimates of the effect of file sharing on sales. By allowing for the effect of file sharing on market outcomes to vary with ex ante artist popularity, I find that file sharing has had strong distributional impacts on the sale of albums. The effect of file sharing on sales becomes more negative as ex ante artist popularity increases, resulting in a point estimate of an elasticity between file sharing activity and sales of approximately -0.5 for the ex ante most popular artists in the data set. Further, the results show that file sharing has dramatically altered the distribution of outcomes among artists, and that the aggregate effect of file sharing on sales is quite strongly negative. Counterfactual exercises suggest that the lawsuits taken by the RIAA that have curbed file sharing resulted in a 2.9% increase in albums sold during the 23 week period after the lawsuit plan was publicly announced, increasing industry profits by \$37 million. Finally, the estimates suggest that a 30% across-the-board reduction in the number of files shared would have resulted in an additional 66 million albums sold in 2003, an increase of approximately \$330 million in profits.

## **2 The Recorded Music Industry**

The recorded music industry is one which is extremely concentrated both horizontally and vertically, with the top five recording distributors combining to distribute over 80% of all album sales in the United States in both 2002 and 2003 (Christman 2003, 2004). The same five companies also own virtually all significant record labels. These “Big Five” companies, Universal Music Group (UMG), Warner/Elektra/Atlantic (WEA), Sony, Bertelsmann Music Group (BMG), and Electric and Musical Industries (EMI), then have tremendous market power in the signing of artists, the release of albums, and the distribution of the albums. Table 1 presents aggregate market share data for total album sales in 2002 and

Table 1: Market Shares of Big Five Firms  
Recording Company Market Shares, 2002-2003

Company	Market Share 2002	Market Share 2003
UMG	28.9%	28.1%
WEA	15.9%	16.4%
BMG	14.8%	15.5%
SONY	15.7%	13.7%
EMI	8.4%	9.7%
TOTAL (BIG FIVE)	83.7%	83.4%
Independents	16.4%	16.7%

Notes:

1. Source: Christman (2003, 2004)
2. Totals may not add up to 100% due to rounding error

2003, the two years in the data sample.

Albums are typically produced in the following manner. First, an artist, who is represented by a manager, is signed to multi-year recording contract by a record label, with compensation consisting of an up-front payment and then royalties from the sales of albums, generally between 5% and 13% of the retail price of the album (Standard and Poor's 2002). An album is then produced in one of the label's recording studios, printed onto a compact disc by the production arm of the owner recording company, and distributed by the distribution arm of said company. Thus, in addition to the tight horizontal concentration illustrated above, the path from artist to consumer is essentially completely vertically integrated. The typical distribution cost to retailers of an album hovers around \$10 and a baseline industry figure is that the record company makes somewhere on the order of \$5 per album sold (Billboard 2000), depending on the album specifics.

Meanwhile, distribution channels have also changed greatly since file sharing and the internet started to cause changes in the industry. In 1999, 51% of albums were sold in retail music stores and 34% in "other stores." By 2002 and 2003 the share of sales in music stores had dropped to approximately 35%, with over 50% sold in "other stores" (RIAA 2004). Additionally, by 2003, fully 5% of all music sales occurred through the internet,

a figure that has continued to grow (RIAA 2004). The general consensus in the industry is that this shift is a movement towards sales through large electronics chains such as Best Buy and Circuit City, as well as mass merchants such as Wal-Mart and away from small, localized music stores and chains. While this change has occurred over this five year period, the shares are essentially stable in 2002 and 2003 (36.8% to 33.2% for music stores and 50.7% to 52.8% for other stores), which is important when analyzing the market during 2002 and 2003, as is done in this paper.

Finally, the end of 2003 and the beginning of 2004 have seen the roll out of several new distribution channels utilizing legal MP3 downloads on a subscription, single track, or full album basis, starting with iTunes for Windows in October of 2003, and currently including offerings from Rhapsody, MusicMatch, Roxio's revamped Napster service, and even Walmart.com, among many others<sup>6</sup>. Only iTunes was active at any noticeable level during the sample period, and then only for the final several weeks of the sample period. According to Apple press releases (2003a, 2003b), iTunes for Windows sold approximately 4 million songs in the month after its launch<sup>7</sup>. While this is not an insignificant amount, all attempts to control for this change in the empirical specifications that follow fail to identify any effect that iTunes has had on either CD sales or file sharing behavior during the sample period and therefore the introduction of iTunes is ignored throughout the rest of the paper.

### **3 Fixing Ideas**

The question of how file sharing affects the sales of recorded music in the short run is a primarily empirical question. Theory presents economists with multiple possible answers. Here, I summarize the possibilities and examine how the effects of different explanations

---

<sup>6</sup>Microsoft has recently announced plans for its own online MP3 distribution service.

<sup>7</sup>In the average sample week, approximately 11 million full albums were sold.

might mesh together. The most immediate story, and the story favored by the RIAA, is that downloads are a direct substitute for sales. Thus, the availability of a song or album on a file sharing network simply allows some consumers who would have purchased the album otherwise to download the music instead, leading to a loss in sales. However, it has also been suggested that file sharing might have positive effects on the sales of records. There are two main arguments concerning how sales might be increased by file sharing.

The first is what was originally coined the exposure effect by Liebowitz (1982). The exposure effect refers to the ability of consumers to sample a good before purchasing it. File sharing, then, might allow potential customers to remove some of the uncertainty involved in purchasing a full album of music by sampling more songs from the artist than they would otherwise be able to do. Thus, consumers will be more likely to buy music from the artists that they learn they like better.<sup>8</sup> Much of the attention of file sharing proponents focuses on this angle, with any number of web sites offering many claims of experimentation leading to purchase. Recent work by Anantham & Ben-Shoham (2004) has taken a formal theoretical look at this claim in regards to other markets and find that while the exposure effect may increase sales, the conditions under which this would be are somewhat restrictive.

The second argument focuses on network effects, where the fact that some portion of the population consumes or listens to the music leads to increased willingnesses to pay for other consumers. That is, if some consumers are listening to the music of an artist, then other consumers start to like that artist better solely for that reason. Liebowitz (2004) provides an in-depth discussion of the ability of network effects to exist in the market for recorded music, concluding that while it is possible that network effects through file sharing may increase sales of recorded music, it seems very unlikely that this effect is strong. Nevertheless, the literature on copying and network effects, in particular Takeyama (1994, 1997), suggests that network effects can strongly mitigate the negative effects of

---

<sup>8</sup>And, of course, they will be less likely to buy music from artists that they like less.

copying. Blackburn (2003) further demonstrates that firms with more mature products would prefer less copying and firms with new products would prefer relatively more, in line with the findings in this work.

I propose an alternative route through which copying (file sharing) might increase sales, which is more of a hybrid of the two stories above than a new route. Both stories above are implicitly assuming that all consumers are aware of all albums which they might purchase. This is extremely unlikely to be true.<sup>9</sup> Copying, then, has the ability to increase the share of potential consumers that are aware of a particular album or artist. Consumers may learn about previously unknown albums through various routes— either by hearing a downloaded song at a friend’s house or at a party, by hearing their music on the radio or on television,<sup>10</sup> or through word-of-mouth or news programs, all of which become more likely if consumers who download music actively listen to it. Thus, ignorant consumers become more likely to discover previously unknown artists as knowledgeable consumers download (or purchase). This awareness effect is essentially a network effect— however, rather than increasing the valuation of individual consumers, the increased number of listeners increases the share of the consumers who are aware of the artist, thus raising the valuation of the average consumer.<sup>11</sup>

There are, then, essentially two competing effects of copying on sales, one positive and one negative. In what follows below, I illustrate a simplified example highlighting these two effects which allows me to discuss how the relative sizes of these effects will differ based on the ex ante popularity of the artist. Denote the quantity of albums sold by an artist to be  $Q(p(q_{FS}), q_{FS}, \theta(q_{FS}))$ , where  $p$  represents the price of the album,  $q_{FS}$  represents the quantity of downloads from the album, and  $\theta$  represents the fraction of all consumers that

---

<sup>9</sup>In fact, it is surely false, as I myself am not aware of all the different music that I might purchase or download.

<sup>10</sup>Both of these first two routes for learning about an album are really just variants of the exposure effect.

<sup>11</sup>Technically, the awareness effect could be thought of as raising the valuation of previously ignorant consumers from negative infinity to some finite value.

are aware of the existence of the album. We are interested in the effect of changes in  $q_{FS}$  on  $Q$ :

$$\frac{dQ}{dq_{FS}} = \frac{\partial Q}{\partial q_{FS}} + \frac{\partial Q}{\partial \theta} \frac{\partial \theta}{\partial q_{FS}} + \frac{\partial Q}{\partial p} \frac{\partial p}{\partial q_{FS}} \quad (1)$$

(?)
(-)
(+)
(?)

I now discuss each of the terms above, in an attempt to sign the effect of file sharing on record sales.

The first term,  $\frac{\partial Q}{\partial q_{FS}}$ , is the direct substitution effect discussed above and is clearly negative. The second term above is  $\frac{\partial Q}{\partial \theta} \frac{\partial \theta}{\partial q_{FS}}$ , which is the awareness effect. This effect is clearly positive, as discussed above, as file sharing should increase the fraction of the world aware of the album ( $\frac{\partial \theta}{\partial q_{FS}} \geq 0$ ) and greater awareness leads to greater sales ( $\frac{\partial Q}{\partial \theta} \geq 0$ ). There is still one potentially important effect remaining, but regardless of the sign of the remaining term, the overall sign of the marginal effect of file sharing on sales as predicted by theory is ambiguous.

This final term,  $\frac{\partial Q}{\partial p} \frac{\partial p}{\partial q_{FS}}$ , is a potential pricing effect, which is of ambiguous sign, as  $\frac{\partial Q}{\partial p}$  is likely negative, and the sign of the pricing response is unclear.<sup>12</sup> However, working within the short-run constraints of this paper, I assume that there is no price response from the industry. That is,  $\frac{\partial p}{\partial q_{FS}} = 0$ . Figure 2 examines the average real list price of a compact disc (CD) appearing on the Billboard Hot 200 album sales chart over time for the last ten years, including the data period which starts in September 2002 and continues through November 2003. An analysis of these mean prices reveals that there has been a slight, consistent downward trend in the real price of a CD over the past five years. More alarmingly, however, there appears to be a structural break occurring over the last 10 weeks of my data set that indicates that record labels may have finally begun responding to file sharing with

---

<sup>12</sup>Record companies may wish to lower prices in order to recapture some of the demand lost to file sharing. On the other hand, they may want to raise prices if the demand that is lost to file sharing comes from consumer with low willingness-to-pay. But no matter what the sign of the pricing effect is, the overall effect would remain ambiguous.

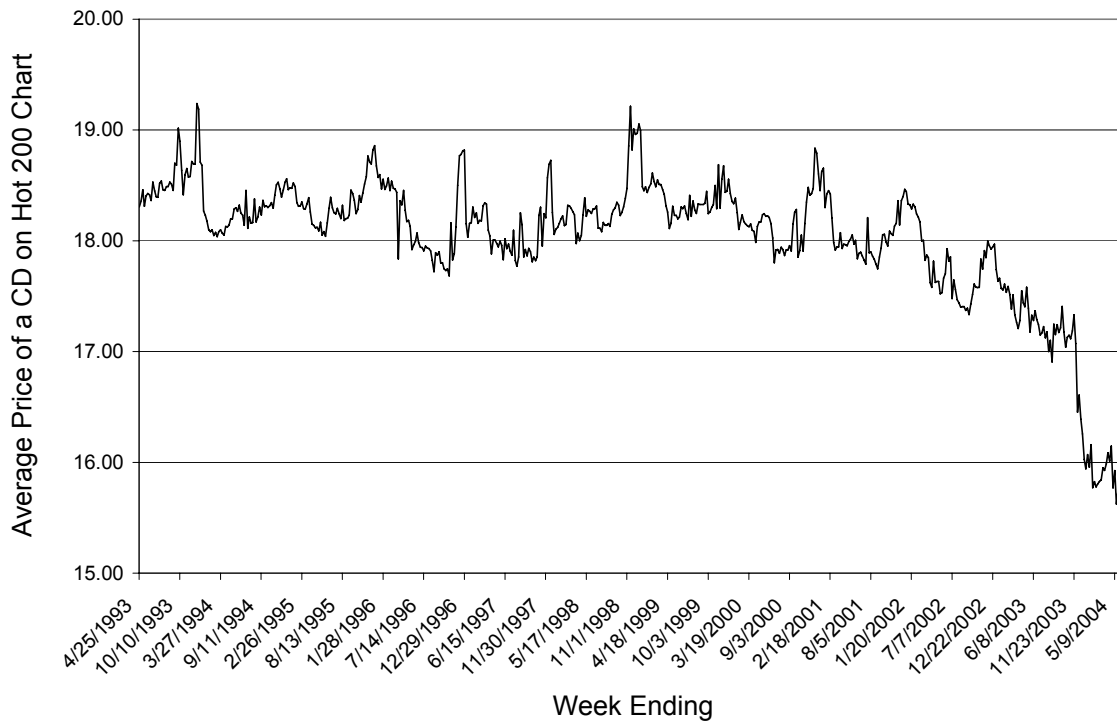


Figure 2: Average Price of a CD on Hot 200 Sales Chart by Week, 1993-2004

pricing strategies. This substantial drop in the average price of a CD corresponds with the announced policy of Universal Music Group (UMG), one of the “Big Five,” to reduce CD prices across the board on all their releases. In order to avoid the endogeneity problem for this change, for which I have no believable instruments, I remove the last 10 weeks of data from my sample, in order to maintain a consistent market set-up throughout and maintain my short run assumption.

Thus, we return to the empirical question: What is the sign of  $\frac{dQ}{dq_{FS}}$ ? However, theory is not yet out of ideas. There is reason to believe that the marginal effect of file sharing on sales will differ based on the ex ante popularity of the artist. Continuing to assume that there is no price response to file sharing, only the first two terms terms above come into play. How does the popularity of the artist affect the relative magnitudes of the direct effect

and the awareness effect? The negative-signed direct effect,  $\frac{\partial Q}{\partial q_{FS}}$ , is surely decreasing (becoming stronger) in the ex ante popularity of the artist. Put another way, the lost sales due to file sharing are surely greater if potential sales are greater. Similarly, the awareness effect is also decreasing (becoming weaker) in the popularity of the artist. This is intuitive—if the benefits of file sharing are essentially introducing new consumers to an artist, the effect will be necessarily smaller if consumers are already aware of the artist.

Thus, it is clear that file sharing should have differential impacts on artists that are well-known to consumers ex ante versus artists who are relatively unknown ex ante. The positive effect of file sharing should be stronger for relatively unknown artists while the negative effect should be correspondingly weaker. In light of this, it is naïve to believe that file sharing has either been “good” or “bad” for recording artists in general. As discussed above, the previous literature focusing on the effects that file sharing has had on the music industry has either implicitly or explicitly assumed that there is an effect common to all artists. Rather file sharing has distributional consequences for the industry, in addition to the average overall effect that has been the focus both in the courtroom and in academics. File sharing makes it harder for very popular acts to sell more and more records,<sup>13</sup> while consequently making it easier for new and previously unknown artists to break through. These distributional effects, in addition to any immediate short-run impacts, thus have potentially very important implications for the long-run development of artistic talent and distribution of outcomes for artist, labels, and consumers.

---

<sup>13</sup>The last album to sell even 7 million copies in one year was 'N Sync's “No Strings Attached,” which sold 9.9 million copies in 1999, just as file sharing was born.

## 4 A Look at the Data

### 4.1 Data Sources

The Data Appendix provides a detailed discussion of the complete set of variables and data sources used throughout. Here I provide a quick summary of the most important aspects of the data. The data for the analysis undertaken in this paper come primarily from two sources. Data on album sales come from Nielsen SoundScan, which tracks retail sales of music and music video products throughout the United States. Nielsen SoundScan obtains their data from point-of-sale cash registers at over 14,000 outlets in the United States, including retail stores, mass merchants, and on-line stores, and reports it weekly.

Data on the file sharing activity for albums come from BigChampagne, which tracks all visible file sharing activity on the 5 largest file sharing networks.<sup>14</sup> BigChampagne collects their data by using the search features inherent in file sharing networks to investigate what files are being shared by each user seen on the network. They then use this information to determine what fraction of network users are sharing particular songs on an album.<sup>15</sup> This data is then reported weekly.<sup>16</sup>

Finally, I build other album-level control variables from various sources in order to control for any observable week to week variation in the quality of an album. This includes data on radio airplay for songs from the album, television appearances by the artist, and Grammy award nominations and wins.

---

<sup>14</sup>Throughout the timeframe of my data sample, these networks are the FastTrack network (Kazaa), Grokster, eDonkey, iMesh, and Overnet.

<sup>15</sup>Fractions are reported rather than totals because the total number of users “seen” each week fluctuates due to internet congestion affecting BigChampagne’s web servers as well as routine server maintenance.

<sup>16</sup>Additionally, BigChampagne also records all search requests that it sees that are sent out over the file sharing network and reports the fraction of all searches that correspond to a particular artist, track, or album. This is a less exact measure of interest in a particular song, however, as a user searching for a copy of a song by an artist may search for it without even entering the name of the song. For example, I could search for Faith Hill’s song “Cry” by simply searching for “Faith Hill” and selecting the appropriate file that appears in the search results. For this reason, I focus on the number of files shared as the main variable of interest for file sharing activity.

## 4.2 The Data Sample

Throughout the empirical analysis that I conduct, I will consider a recorded music album to be the unit of analysis, and observations will be album-week pairs. Albums were chosen from the set of all albums containing new material by a single artist<sup>17</sup> released between September 24, 2002 and September 16, 2003, inclusive. Due to data availability limitations for the file sharing data, 197 albums were able to fit the criteria for inclusion in the data sample.<sup>18</sup> While file sharing and sales data were available through February 8, 2004, due to the structural change in pricing that occurred at the end of November 2003, I use data only through the week ending November 30, 2003. This results in a full sample of 197 albums and 7,938 album-weeks.

It is also possible that other structural components of the industry have changed during this time period in response to file sharing. In particular, it could be that firms started to adopt new strategies concerning the release of albums or the signing and development of new acts. These changes would be much harder to detect, but I have found no evidence that record labels have acted on changing traditional patterns of album development before UMG's price change at the end of 2003. Therefore, I proceed with my analysis confident in my choice of time frame, in which the short run is defined as above, leaving a total of 62 weeks of data.

Finally, there is the issue of the non-randomness of the albums chosen to be in the data sample, which raises potential questions about the similarity between the data sample and the full population of albums. While sales data is not available for the albums not in the sample, it is possible to compare the total Billboard chart performance of the two sets of albums. As detailed in the Data Appendix, it appears that the sample of albums for which file sharing data is available is slightly more successful than the general album population,

---

<sup>17</sup>A single artist is either a solo artist, such as Celine Dion, or a musical group, such as the Foo Fighters.

<sup>18</sup>See the Data Appendix for the complete details on how the sample was built.

though not by a large amount. Weights can be constructed to match the distribution of chart performance for the sample to that of the population. Thus, in what follows, I apply weights when aggregating up from individual albums to the market level in order to properly represent aggregate effects.

### **4.3 Measuring File Sharing and Artist Popularity**

In order to differentiate the effects of file sharing on artists based on ex ante popularity, I use data taken from Billboard's Hot 200 chart<sup>19</sup> in order to build a measure of ex ante popularity. Using Hot 200 chart positions for the previous 10 years prior to the start of my sample,<sup>20</sup> I record the peak position obtained by a previous album from the artist. This peak position is then transformed into a continuous measure of ex ante popularity, defined as 201 minus the peak position of the artist in the past ten years. Thus, for example, Faith Hill, whose album "Breathe" was the number one album on the Hot 200 chart on September 11, 1999 is categorized as having a popularity of 200. Artists who have never had an album appear on the Hot 200 chart are given a popularity of 0. This classification system provides an objective measure of ex ante popularity, which is based on the market success of the artist in the past. In general, when referring to an artist whose popularity index is 0, I will simply call them "new" artists. Ex ante well-known artists have high popularity indices, while ex ante unknown artists have lower levels of the popularity index. For comparative static exercises, increasing artist popularity has the effect of increasing the popularity index variable. I also performed robustness checks to verify that other possible measures of ex ante popularity do not modify the results.

The primary variable used to summarize the amount of file sharing activity for an al-

---

<sup>19</sup>The Billboard Hot 200 chart is released weekly and reports the ordinal ranking of albums at the national level.

<sup>20</sup>That is, back to September 1992.

bum is the number of copies of songs from an album that are available on the file sharing networks. To construct this variable, I take the reported fractions of file sharing network users that are sharing a particular song and multiply by the size of the file sharing network that week, as measured by the average number of users logged in during the week, using data provided by Robin Millete (2004).

Ideally, I could use data at the artist or song level on the actual number of downloads during a week, rather than the number of copies of the song available on-line. However, this data is not available and thus the number of copies of a song that are available on the network is used. This serves as proxy for the “cost” of downloading a song, because in the structure of peer to peer file sharing networks, a file is simultaneously downloaded from multiple users and then reassembled on the downloader’s machine. Thus, more copies on the network means that the song can be downloaded more quickly. Additionally, because there are so many different ways of searching for a particular song, album, or artist on file sharing networks, more copies on a network suggest that it may take less time to search for the track, as different copies will be named (and thus found by the search engine) in different ways, again causing the download process to take less time to complete.

For each album, I construct a variable which takes the value of the number of copies of the song that is most prevalent on the file sharing networks that week. To illustrate, imagine there is an album with only two songs, “Popular Song” and “Unpopular Song.” If in a given week there are 10,000 copies of “Popular Song” available on the file sharing networks, and only 200 copies of “Unpopular Song” available, the variable measuring the maximum number of copies available would receive a value of 10,000. This construction is taking the stance on the substitutability of file downloads for album purchases that consumers equate an album to the most popular song on that album. Throughout the analysis, this is the variable used to measure file sharing activity. However, to address concerns about this particular measurement, I created several other variables, described in the Appendix,

that are used to verify that using the most shared song does not drive the results. Although unreported, using the other measures does not qualitatively change the results, and thus I proceed to estimation using the number of shared copies of the most shared song as the variable of interest.

## 5 Estimation

### 5.1 Linear Reduced Form Estimation

I begin by specifying and estimating several linear reduced form regressions to pin down the impact of file sharing on music sales. In these regressions, I am implicitly assuming that each album is a monopolistic market, and I treat each album-week pair as an observation for the market for that album in particular. While this assumption of monopolistic markets is ignoring relationships across albums that might exist, the simplicity gained by such an approach is useful. Additionally, following a specification similar to the ones used previously in the literature will allow me to view easily how the results fit with previous work. In particular, the reduced form specifications below mirror those of Oberholzer and Strumpf (2004), whose results are striking in that their estimate of the overall effect of file sharing on sales suggests virtually no effect.

To frame what follows, I begin by using simple OLS estimation as a means of highlighting the issues involved with estimating the effect of file sharing on sales. The simplest model of the relationship between sales and file sharing discussed in Section 3 is a pooled model of the form:

$$q_{i,t}^S = \alpha + \beta q_{i,t}^{FS} + \rho X_{i,t} + \varepsilon_{i,t} \quad (2)$$

where  $q_{i,t}^S$  is the quantity of album  $i$  sold in week  $t$  (possibly expressed in logs),  $q_{i,t}^{FS}$  is the measure of file sharing activity for album  $i$  in week  $t$  (also possibly expressed in logs),  $X_{i,t}$

is a vector of album and week characteristics for album  $i$  in week  $t$ , and  $\varepsilon_{i,t}$  is an error term.<sup>21</sup> The problem with estimating  $\beta$  comes from omitted variables bias, as there are relevant variables such as the “artistic merit” of the album that are not observed. As this “quality” is likely also relevant for determining the amount of file sharing that occurs, the  $\varepsilon_{i,t}$  is positively correlated with  $q_{i,t}^{FS}$  and the OLS estimate of  $\beta$  is biased upward.

If artistic merit is assumed constant over time, then it is possible to exploit the panel nature of the data set to correct for the bias caused by its omission by using album fixed effects in the estimation. This allows for a model such as:

$$q_{i,t}^S = \alpha + \beta q_{i,t}^{FS} + \rho X_{i,t} + \gamma_i + \nu_{i,t} \quad (3)$$

where the  $\gamma_i$ 's are album level fixed effects. This can potentially address a large fraction of the omitted bias as the fixed effect captures time-invariant album merit. However, there is still likely to be week to week unobserved variation in the “quality” of an album that will still lead to an upward bias in the estimate of  $\beta$ . This will be the case if there are any events that change consumer’s valuations of albums that can not be observed.

Table 2 presents the results of these OLS specifications, using the number of copies of the most popular song available on the networks from an album as the measure of file sharing activity. The upper table presents the results of the OLS estimation in levels and the lower table presents it when measuring both sales and files shared in logs.

The first column of each table presents the simple correlation between file sharing and sales. As expected, this is significantly positive as a result of the omitted bias. Moving across the tables from left to right displays the estimated coefficient  $\hat{\beta}$  as we start to add album- and artist-level controls. Not surprisingly, adding album level controls brings down

---

<sup>21</sup>The set of time dummies includes a specific dummy for the initial week of release and also a second-order polynomial in the number of weeks since release. A more flexible specification for decay over time fails to affect the results, and while coefficients are often significant until fifth order polynomials, the coefficients are economically insignificant. Thus, I continue with the simple quadratic polynomial.

Table 2: OLS Estimation Results  
Dependent Variable: Weekly Sales

	(1)	(2)	(3)	(4)	(5)	(6)
Max # of Files Available	0.074 [0.009]***	0.083 [0.010]***	0.081 [0.009]***	0.037 [0.014]***	0.039 [0.014]***	-0.107 [0.031]***
CONTROLS:						
Time Variables	NO	YES	YES	YES	YES	YES
Holiday Variables	NO	NO	YES	YES	YES	YES
Television Variables	NO	NO	YES	YES	YES	YES
Airplay Dummies	NO	NO	NO	YES	YES	YES
Grammy Awards	NO	NO	NO	NO	YES	YES
Fixed Effects	NO	NO	NO	NO	NO	YES
Observations	7938	7938	7938	7938	7938	7938
R-squared	0.06	0.26	0.3	0.35	0.35	0.5

	Dependent Variable: Log of Weekly Sales					
	(1)	(2)	(3)	(4)	(5)	(6)
Log of Max # of Files Available	0.309 [0.044]***	0.335 [0.046]***	0.333 [0.045]***	0.234 [0.038]***	0.226 [0.038]***	0.251 [0.057]***
CONTROLS:						
Time Variables	NO	YES	YES	YES	YES	YES
Holiday Variables	NO	NO	YES	YES	YES	YES
Television Variables	NO	NO	YES	YES	YES	YES
Airplay Dummies	NO	NO	NO	YES	YES	YES
Grammy Awards	NO	NO	NO	NO	YES	YES
Fixed Effects	NO	NO	NO	NO	NO	YES
Observations	7938	7938	7938	7938	7938	7938
R-squared	0.15	0.53	0.54	0.62	0.63	0.93

1. Robust standard errors in brackets

2. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%

3. Time Variables include a non-linear decay trend and a dummy for the week of release, Holiday Variables include dummies for the two weeks before, the week of, and the week after Christmas, Television Variables include dummies for appearances on TV during the week, or the week prior, Airplay Dummies are as described in the text, and Grammy Award variables include dummies for nominations and wins in 2002 as well as nominations in 2003.

the estimate of  $\beta$  by introducing more information on the co-movements of sales and file sharing activities. Column (5) presents the full pooled estimation of Equation (2). The estimated coefficient of .037 in the levels specification, which is known to be biased upward, suggests that every 26 additional files available on-line increases the week's sales of the album by one unit. For comparison's sake, Oberholzer and Strumpf's pooled model suggests that every additional 66 downloads increase sales by one unit. While downloads are most certainly not the same as file availability, these two results are comparable. The log specification, for which the added controls do less to reduce the upward bias, suggests a file sharing elasticity of sales of .23; that is, reducing file sharing by 10% would decrease sales by 2.3%. We can compare this elasticity to the elasticity implied by the linear

specification, which (at the mean) is .26. Thus, we see that the two specifications are essentially identifying the same relationship. For the artist with the median level of weekly sales, these estimates suggest that a 10% reduction in files shared would decrease sales by approximately 70 albums per week.

Finally, column (6) presents the results of the full OLS fixed effects approach of Equation (3). When measuring sales and file sharing activity in levels, there are now enough controls to move the point estimate of  $\beta$  below zero; while when measured in logs, the omitted variable bias is still strong enough to leave the point estimate positive. Regardless, it is unreasonable to suspect that all of the bias can be corrected by these observable album-level covariates and fixed effects. It is more striking that the sign of the two effects differs after adding in artist-level fixed effects, with the linear OLS specification finding a negative relationship between sales and file sharing while the log specification continues to capture a positive relationship. This difference is puzzling; however, as will be seen in Table 4, controlling for the endogeneity of the error term through a two-staged least squares technique eliminates this contrast. The only specification in which this sign difference occurs is an OLS specification that includes artist fixed effects, and so while this result is interesting, I view it as a mere curiosity, which is eliminated by the TSLS approach.

Before continuing, I run a Box-Cox transform to determine whether the relationship between sales and files shared is better specified in levels or in logs, following Godfrey and Wickens (1981). This functional form test embeds both the level-level and log-log specification in a larger, Box-Cox transform that can be estimated with maximum likelihood techniques. Within this framework, it is possible to use the log likelihoods generated in estimation to test the restrictions imposed by either a level-level or a log-log specification (or others) to ascertain which functional form is appropriate for the relationship in question. For the relationship between sales and files shared, the restrictions imposed by a log-log specification result in a much higher log likelihood than the level-level functional form does

(-68,966 as opposed to -91,131). Thus, while I will carry along the levels specification for completeness and comparison's sake, due to the much stronger fit of the model specified in logs, I will focus solely on the log-log specifications.<sup>22</sup> It is worth noting, then, that the coefficient  $\beta$  is the elasticity of sales with respect to file sharing activity in a log-log specification.

### 5.1.1 TSLS Estimation

Now, in order to account more completely for the omitted variable bias that still exists in these estimates, I proceed with a two stage least squares approach and use the timing of the RIAA lawsuits against consumers. Of course, it is necessary that the timing of the lawsuits is exogenous to the dependent variables in my primary regression in order for the variable to be valid. And this may be a concern as the lawsuits are clearly an industry-wide response to what is perceived at least to be a reduction in sales as a result of file sharing activity. However, while the existence of the lawsuits is clearly not exogenous to the phenomenon in question, the exact timing regarding both the announcement of the plan to sue consumers and the eventual implementation of the suits is a random shock to the behavior of consumers.<sup>23</sup> Thus, I instrument for the amount of file sharing using dummy variables indicating that the RIAA's plan has been announced or that the RIAA's plan has been implemented. Both of these events, which occur during the weeks of June 25, 2003 and September 9, 2003 respectively, are important because while the first scared off many potential consumers from file sharing networks, the second actually brought some consumers back in. Table 3 presents the results of the first stage regression, where the dependent variable is the amount of file sharing activity, as measured by the number of

---

<sup>22</sup>Both specification are technically rejected in favor of an unconstrained Box-Cox specification. However, this flexible model is more difficult to interpret and does not change the results of the paper.

<sup>23</sup>In its announcements leading up to the first round of lawsuits, the RIAA never announced a target date or timeline for the lawsuits to begin, so the timing of the lawsuits themselves is also essentially random from the point of view of consumers.

Table 3: TSLS First Stage Results

Dependent Variable:	(1)	(2)
	Log of Max # of Files Shared	Max # of Files Shared
Dummy for Weeks After Lawsuit Plan Announced	-0.458 [0.056]***	-25,356 [4,448]***
Dummy for Weeks After Lawsuits are Implemented	0.098 [0.057]*	12,358 [4,577]***
Debut Week	-0.582 [0.055]***	-18,076 [2,589]***
Christmas Week	-0.114 [0.055]**	-11,214 [3,720]***
Less Than 2 Weeks Before Christmas	-0.160 [0.058]***	-14,744 [3,759]***
The Week After Christmas	0.055 [0.080]	3,100 [4,479]
#1 Radio Airplay Song	0.531 [0.127]***	107,978 [50,503]**
#2 - #10 Radio Airplay Song	0.352 [0.099]***	40,079 [20,422]*
#11 - #40 Radio Airplay Song	0.236 [0.082]***	19,405 [12,187]
#41 - #75 Radio Airplay Song	0.144 [0.067]**	5,577 [7,011]
Observations	7938	7938
R-squared	0.9	0.93

1. Robust standard errors in brackets
2. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%
3. Non-reported controls include non-linear album age trend, Grammy nominations and wins, and TV appearance dummies for the current and previous weeks.

copies of the most prevalent song on the album.

The results show that the announcement of the RIAA’s plan to sue the users of file sharing software individually rather than their previous strategy of going after the network providers is shown to have, as expected, a strongly negative effect on the number of files available for download from file sharing networks. In Column (1), the number of files shared is measured in logs, whereas in Column (2), it is measured in levels. The point estimate of -0.46 from the log specification suggests a decrease in mean file sharing activity (relative to trend) of approximately 40% after the lawsuit announcement. Also as expected, there is a positive (and smaller in absolute value) point estimate associated with

the implementation of the lawsuits, raising file sharing up again approximately 10%. This indicates that file sharing regained approximately one-fourth of the drop caused by the threat of lawsuits when it became apparent to consumers that the threat was less than anticipated.<sup>24</sup> Though not reported in the table, an F-test on the joint significance of the two excluded instruments results in an F-statistic of 57.88, which is significant below even the .1% level and suggests that the instruments are sufficiently strong that I need not worry about weak instrument concerns. The levels specification provides similar results, with the announcement causing a drop in the mean number of files shared of 25,356, which amounts to a mean decline of 32%, with an increase of 12,358 files associated with the implementation of the first round of lawsuits, and an F-stat of 195.75 for the joint significance of the excluded instruments.

Another item of note from the first-stage results is that the Christmas holiday period is associated with large reductions in file sharing. This is a noteworthy result, as the holiday period is also associated with large increases in sales. While this alone does not tell a complete story, it is reasonable to believe that during holiday periods much of the consumption of music is done in the form of purchasing gifts for friends and relatives. The fact that sales and file sharing activity move in different directions seems to suggest that consumers are perhaps willing to download a song or album as a substitute for a purchase for themselves, but are unwilling (or unable) to give as a gift an album that has been downloaded rather than purchased.

Additionally, unlike sales, file sharing activity as measured by the number of files available is lower during the debut week of an album. This is not surprising because while iden-

---

<sup>24</sup>I argue above that this would be expected because it became clear that the RIAA would only sue users who shared large quantities of songs, although Gary Chamberlain has pointed out a more direct route: even as of July 2004, the RIAA has only sued approximately one thousand users out of a total of anywhere from 5 to 8 million. Thus, consumer might rationally have decided to accept the risk, given the probability of being sued appears to be very small. Of course, it should be pointed out that the cost of being sued by the RIAA is also potentially quite large.

tification comes off of movements in the number of files available on file sharing networks, the measure used is, in fact, a stock and not a flow, and thus in early weeks not enough time has passed for the peak level of the stock to be reached. This also poses a potential problem for the estimation strategy: in the early weeks of sales, the stock of files available on the file sharing networks is generally still growing, while weekly sales figures tend to decrease. This relationship is a result of comparing a stock and a flow and not the underlying relationship between downloads and sales. This problem would likely bias my estimates downward, because during early weeks files shared increase while sales decrease, implying a false negative relationship.

In order to control for this time trend, I allowed for a flexible time trend to enter into both stages of the estimation. Although suppressed, the flexible trend suggests that a simple quadratic polynomial in time is sufficient to capture the effect, as more flexible polynomials provided neither additional explanatory power nor a change in the estimated effect of file sharing on sales. The identification of the relationship between sales and files shared is then coming from the deviations relative to this trend. This can be seen in columns (1) and (2) of Table 4,<sup>25</sup> which reports the estimation in the absence of such a trend. In this case, the estimated effects are seen to be significantly more negative, as we would predict. Including the life cycle trend corrects this problem.

Table 4 presents the results of the second stage estimation. Three specifications are presented. The first is comparable to the estimate presented in column (1) of the lower table of Table 2, the second is the full fixed effects model lacking the quadratic time trend discussed above, while the specification presented in column (3) is the result of a full fixed effects model comparable to column (6) of Table 2. The estimates are striking. First the

---

<sup>25</sup>Further, performing the same estimation without a trend but excluding the first several weeks of an album's life makes the estimates too positive, as after the first few weeks of an album's life we have the opposite problem. Reassuringly, returning the trend into the regression specification while leaving out the first few weeks of an album's life restores the previous estimates. Thus, I proceed without concern over this potential problem.

Table 4: TSLS Second Stage Results

	TSLS Second Stage Results					
	(1)	(2)	(3)	(4)	(5)	(6)
<b>Dependent Variable:</b>	<b>Log of Weekly Sales</b>			<b>Weekly Sales</b>		
Log of Max # of Files Available	-3.103 [0.970]***	-4.187 [0.928]***	-0.073 [0.082]			
Max # of Files Available				-0.455 [0.173]***	-0.164 [0.043]***	-0.036 [0.033]
<b>CONTROLS:</b>						
Time Variables	NO	NO	YES	NO	NO	YES
Holiday Variables	NO	YES	YES	NO	YES	YES
Television Variables	NO	YES	YES	NO	YES	YES
Airplay Dummies	NO	YES	YES	NO	YES	YES
Grammy Awards	NO	YES	YES	NO	YES	YES
Fixed Effects	NO	YES	YES	NO	YES	YES
Observations	7938	7938	7938	7938	7938	7938

1. Robust standard errors in brackets
2. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%
3. Time Variables include a non-linear decay trend and a dummy for the week of release, Holiday Variables include dummies for the two weeks before, the week of, and the week after Christmas, Television Variables include dummies for appearances on TV during the week, or the week prior, Airplay Dummies are as described in the text, and Grammy Award variables include dummies for nominations and wins in 2002 as well as nominations in 2003.
4. Excluded Instruments: Indicators for lawsuit announcement, lawsuit implementation, and interactions with artist popularity

estimated elasticity of -3.1 and -4.2 in the specifications without a quadratic time trend are extremely unrealistically negative; however, this is likely due to the pitfalls discussed in the preceding paragraph and is thus not a cause for concern. The importance of including the time trend to account for the structure of the data can be seen by examining the aggregate elasticity of sales for file sharing in the full model, which is -.07, although statistically insignificant.

The two-stage least squares estimation suggests that on an aggregate level, file sharing has had approximately zero effect on sales; if there is an effect, it is very small. The estimated elasticity suggests that eliminating 10% of files shared would increase sales of recorded music by only 0.7%. Again considering the median artist, this 10% reduction in files shared increases sales by only 20 albums per week. This result is consistent with the estimates of Oberholzer and Strumpf (2004), who focus their attention on a small, statistically insignificant effect, albeit a positive one.

However, *it is extremely important to realize that these estimates are likely incorrect.* In addition to ignoring competition effects between albums, the specifications used assume a constant effect of file sharing on sales for all artists, and thus forces this constant estimate to match the average effect across albums. Worse, by using an album-week as the unit of observation, this average effect is weighted not by relative sales, but by the proportions of album-week appearances in the sample.<sup>26</sup> Thus, the naïve answer obtained above is not a reliable measure of the effects of file sharing on sales.

### 5.1.2 Effect of Artist Popularity

Therefore, I allow for the possibility that these estimated effects may be badly specified and remove the assumption of a consistent effect across albums. To allow for different marginal effects of file sharing on sales across artists, I now interact the effect of file sharing on sales with a measure of the ex ante popularity of the artist.<sup>27</sup> This is done by creating a “continuous” definition of ex ante popularity as described in the previous section, defined as 201 minus the highest Hot 200 chart position attained by the artist in the past ten years. This construction of ex ante popularity then defines a regression of the form:

$$q_{i,t}^S = \alpha + \beta q_{i,t}^{FS} + \varphi P_i * q_{i,t}^{FS} + \rho X_{i,t} + \gamma_i + \varepsilon_{i,t} \quad (4)$$

where  $P_i$  is popularity index of artist  $i$ . Thus, the marginal effect of file sharing on sales is given by  $\beta + \varphi P_i$ . Recall that the discussion above suggests that the marginal effect of file sharing on sales is more positive for less well-known artists than for star artists, so the estimated coefficient  $\hat{\delta}$  is expected to be negative. Table 5 presents the results from estimating equation (4).

---

<sup>26</sup>Liebowitz (2004) explains this point in finer detail.

<sup>27</sup>Another way to do this, of course, would be to simply estimate on each album separately. This however, would lead to sample size problems as well as losing the ability to use all the data to help pin down the life

Table 5: TSLS Results, Effects Differentiated by Artist Popularity  
**Two Stage Least Square Results, Effects Differentiated by Popularity**  
**Dependent Variable: Log of Weekly Sales**

	(1)	(2)
Log of Max # of Files Available	-0.073 [0.082]	0.473 [0.225]**
Log of Max # of Files Available * Popularity Index		-0.005 [0.002]**
Observations	7938	7938

1. Robust standard errors in brackets
2. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%
3. The full set of control variables (time trends, holidays, television, airplay, Grammy awards, and album-level fixed effects) is included in both specifications.
4. Excluded Instruments: Indicators for lawsuit announcement, lawsuit implementation, and interactions with artist popularity

Column (1) simply reproduces the estimate resulting from imposing equal marginal effects across artists, imposing  $\varphi = 0$ . In column (2), we see that the baseline elasticity of sales with respect to file sharing for an artist of zero popularity (a new artist) is 0.47, which is strong, although it has a 95% confidence interval lower bound of 0.02, so it is not certain that the effect is this large. Nevertheless, this suggests that new and relatively unknown artists may find file sharing very beneficial, as doubling the amount of file sharing activity for an album from a new artist would increase sales by 38%. More striking, however, is that as predicted this estimated elasticity gets smaller as the artist's ex ante popularity is increased, eventually reaching a point estimate of -0.54, with a standard error of 0.248, for an artist with a popularity index of 200, which, recall, means that the artist had a #1 album in the ten year period prior to the sample. This effect is significant at the 5% level and indicates that artists who ex ante were well-known are, in fact, harmed by file sharing.<sup>28</sup> For reference, the marginal effect is significantly positive at the 10% level for popularities

---

cycle trend of an album, which as discussed previously, is very important to the identification.

<sup>28</sup>Other definitions of artist popularity can also be considered, and should yield similar results. In particular, using the top radio airplay position for an artist in the year prior to album release as a popularity index yields similar quantitative results, though with less power. Presumably, this is because there is less variation in the radio airplay charts (which rank 75 positions) than in the album sales charts. Additionally, conditioning the analysis on the album's debut week and using debut week sales as a measure of popularity also yields similar results.

less than 54 (artists whose most successful previous album reached no higher than 147 on the Hot 200 chart) and significantly negative at the 10% level for popularities greater than 130 (artists whose most successful previous album reached at least 71 on the Hot 200 chart).

This result highlights the problem of simultaneously taking an album as the unit of observation and imposing that all albums are subject to the same effects of file sharing. When the effects of file sharing are forced to be equal for all artists and albums, the “average” effect of file sharing is essentially zero. However, it is wrong to conclude that there is no effect of file sharing on sales. Quite to the contrary, file sharing has large effects on the sales of albums, and in a way that has significant distributional impacts for the sales of records.

Perhaps more importantly for the industry as a whole, the zero average effect is not just misleading when considering the effects on individual artists, but also leads to incorrect answers when calculating what the aggregate effect on sales is. It is incorrect to consider the marginal effect on an ex ante unknown artist to be as important as the marginal effect for a popular artist. As Table 9 in Appendix A reports, the mean sales figures are very different for artists with different ex ante popularity. The average sales week for an artist with a popularity index of zero in the sample is 7,792 while the average sales week for an artist with a popularity index of more than 180 is 12,002. Thus, treating the impact on a “star” artist as equal to the impact on a “new” artist for the sake of aggregating the effect across all artists is wrong. That is, while the average elasticity across albums may be zero, the aggregate impact on sales is clearly not. In fact, artists who are ex ante more popular sell more albums, and so on an aggregate industry level, the negative effects of file sharing will outweigh the positive effects of file sharing.

In particular, the point estimates imply that the median “new” artist, whose weekly sales are 2,163 albums, would see a decrease in weekly sales of 101 albums per week were files

shared to be reduced by 10%. A similar calculation can be made for an artist of maximum popularity. At the median level of sales for these artist, the estimate implies an increase in sales of 490 albums per week if file sharing were to be reduced by 10%. This stark contrast between the magnitudes of the effects for artists of varying levels of popularity highlights the importance of this heterogeneity in estimating the aggregate effects of file sharing.<sup>29</sup>

It is possible to perform some back of the envelope calculations using these estimates of the file sharing elasticity of sales to determine what the impact of file sharing has been on an aggregate level. First, I calculate that the RIAA's public announcement of its lawsuit strategy and the subsequent implementation of the first round of lawsuits had the effect of increasing sales by approximately 12% during the weeks after the lawsuit strategy was announced.<sup>30</sup> This 12% increase implies that aggregate industry sales were 1,376,000 greater on average every week during the 23 weeks remaining in the sample after the announcement was made. Using a rule of thumb of \$5 of profit per compact disc,<sup>31</sup> that equates to a nearly \$7 million increase in industry profits each week, or nearly \$160 million over the whole 23 week period. This effect is very large and indicates why the RIAA would be willing to sue individual consumers despite the negative press associated with the strategy of suing their own customers. However, one should be careful with this estimate, as the linear reduced form estimation here fails to account for any competition effect among artists, which is an important omission.

A similar calculation can be made for estimating the total effect of file sharing on sales.

---

<sup>29</sup>This problem is not merely an artifact of using a logarithmic specification for sales and file sharing. Even in a linear specification, the proper sales-based average effect is *not*  $\beta + \varphi \bar{P}$ , as the average change in sales is  $\frac{1}{n} \sum_i (\beta + \varphi P_i) (\Delta q_i^{FS})$  which does not simplify to  $(\beta + \varphi \bar{P}) (\Delta q_i^{FS})$ .

<sup>30</sup>This is done by simply subtracting out the effect of the lawsuits from the first stage estimation done presented in Table 3 and then using the estimated effects of file sharing on sales from Table 4, and then aggregating up to market level numbers using the appropriate weights.

<sup>31</sup>This figure is taken from an analysis of CD pricing presented in Billboard magazine (2000). According to their analysis, a \$17 compact disc yields a \$10.75 wholesale price. Of this \$10.75, approximately half can be attributed to variable costs, while the other half is either deemed as profits or attributed to what appears to be fixed costs.

To estimate the aggregate effect of a 30% reduction in file sharing across the board,<sup>32</sup> I simply subtract out the effect of the deleted files from the second stage estimation in Table 4 and then aggregate up to market level numbers using the appropriate weights. The estimated effect of such an across-the-board reduction in file sharing is to increase aggregate sales by 15%. Again, while these calculations were useful for placing the analysis inside the framework of the previous literature, they do not take into account competition effects across albums, and so the effects of file sharing will be overstated in these estimates. Therefore, attempting to quantify these effects on an aggregate scale is best done with a model that can capture some of the competition effects among sales, which are precluded from the linear reduced form analysis.

## **5.2 Allowing for Competition Effects**

While the results above are illustrative of the effects that file sharing have had on recorded music sales and allow for comparison with previous results in the literature, the assumptions implicit in treating each album as a separate market severely limit the econometrician from analyzing counterfactuals regarding file sharing, including calculations on the cost of file sharing to the industry. In particular, by assuming that all albums are independent from each other, the linear reduced form estimation above precludes any competition effects through which weaker or stronger competition in a particular week might influence the sales of an album. This is particularly troubling when performing counterfactuals, which would then ignore the effects of competition in calculations on how sales might change if file sharing behavior were changed. By using a simple multinomial logit demand system, I can better capture the effects of competition among albums.

---

<sup>32</sup>The reason for estimating the effect of a 30% reduction in files shared is to keep the counterfactual exercise from projecting the estimated elasticities too far out of sample to remain reliable.

### 5.2.1 The Multinomial Logit Model

The linear reduced form estimated above precludes any album competition effect by assuming that the cross elasticities are all equal to zero. That is, by construction we assume that  $\frac{\partial q_i}{\partial q_j^{fs}} = 0$  for all  $i \neq j$ , and therefore changes in the amount of files available for a particular album do not affect the quantity sold for any other album. This is a restrictive assumption that is important when estimating the effects of the RIAA's lawsuits on sales. For example, proceeding under this assumption fails to account for the fact that if less file sharing for Celine Dion's songs makes consumers more likely to buy an album from Celine Dion, it will make them less likely to buy an album from other artists. When aggregating the effects of reducing file sharing on many artists simultaneously, these neglected competition effects accumulate, causing the estimation to result in overstated effects on sales.

A more general linear reduced form analysis than the ones already estimated could potentially allow directly for these effects. However, to allow for the full interaction between the sales of albums and the quantity of files shared for all albums, it would be necessary to include right-hand side parameters for all of these quantities for all albums. Thus, such a procedure suffers from the curse of dimensionality and my data is unable to handle such a job, even after imposing simplifying restrictions such as symmetry among albums.

The multinomial logit model first presented by McFadden (1973) solves this dimensionality problem by supposing a particular discrete choice form for the problem. Doing so allows for the effect of substitution among albums and thus the feedback effects that are left out by the restricted linear reduced form estimated in the previous section. As will be illustrated below, estimation based off of the logit model imposes a particular structure of substitution that can be estimated easily. But this simplicity is not free; the substitution patterns of the logit model are such that all goods that have the same market share have the same substitution patterns. Nevo (2000) provides a detailed description of the costs

and benefits of the various models of demand that have arisen to overcome this restriction. Due to the nature of the data set, which provides only weekly, national level observations, the estimation undertaken here is incapable of handling these more advanced procedures. Instead, I appeal to the simple logit model as a way to allow for substitution among albums parsimoniously.<sup>33</sup>

Despite the fact that the multinomial logit model is based on optimization of an individual utility function, the specific model is unrealistic, assuming, among other things, homogenous consumers. Instead, the model should be viewed as a reduced form which allows for richer interactions among competing albums in a parsimonious way. Define the utility of consumer  $n$  from *purchasing* album  $i$  in week  $t$  to be:

$$u_{n,i,t} = \delta_{i,t} + \xi_{i,t} + \epsilon_{n,i,t} = \mu_i + \lambda X_{i,t} + \theta q_{i,t}^{FS} + \xi_{i,t} + \epsilon_{n,i,t} \quad (5)$$

where  $\mu_i$  is the underlying quality of album  $i$ ,  $X_{i,t}$  again is a vector of album and week characteristics, and  $q_{i,t}^{FS}$  is a measure of the file sharing activity of album  $i$  in week  $t$ .  $\xi_{i,t}$  is an (unobserved) propensity to like album  $i$  in week  $t$ , which can be thought of as an album-week deviation from the observable components of quality which is common to all consumers, and  $\epsilon_{n,i,t}$  is an idiosyncratic error term. It is important to note the functions of the two error terms.

First,  $\xi_{i,t}$  is specific to an album-week and does not vary over consumers. This term captures the effect of unobservable events that increase or decrease the utility gained from purchasing album  $i$  in week  $t$ . The second error term,  $\epsilon_{n,i,t}$ , is the only term in equation (5) that differs by consumer. This term captures idiosyncratic taste among consumers for goods and is assumed to be independently and identically distributed. This assumption is

---

<sup>33</sup>It has been suggested that allowing for a more non-parametric specification for files shared in the simple TSLS model could help to capture the proper relationship. However, this fails to be the case. Even second-order polynomials in file sharing are not statistically significant, and thus they cannot capture these additional complications.

unlikely to hold in practice, as consumer tastes for albums are likely both correlated across albums (I have a taste for rap music rather than rock music) and correlated within album across time (if I purchase Celine Dion's album this week, I am not likely to buy it next week as well). However, as discussed above, the nature of the data does not allow for the relaxation of these assumptions in estimation. Given the eclectic tastes for popular music, the first correlation across albums may not be a large problem; the logit model imposes that if Britney Spears and Eminem have the same market share, they have the same substitution patterns. In practice, this is largely true, at least to a first approximation.

The second concern is stronger, and worth considering. Identifying dynamic substitution patterns such as this require a much richer data set, and is unnecessary for the needs of this analysis.<sup>34</sup> Instead, it should be emphasized again that use of an individual optimization problem only serves as a richer reduced form, rather than a specific representation of individual choice. Interpreted this way, the specification is not a model of individual choice and this particular concern is mitigated, as the decay in album utility over time in  $X_{i,t}$  will capture much of this problem.

Also, I specify that the amount of file sharing for an album directly affects the consumer's utility from purchasing a copy of the album. Ideally, this would not be necessary and I could capture the effect of file sharing on sales by specifying consumer utility from both downloads and purchases explicitly, and work from there. Unfortunately, there are two problems with this route. The first is that discrete-choice models of the style presented here require all goods to be substitutes at both an individual and aggregate level— thus, the specification of the model would assume the answer to the question of interest. A recent paper by Gentzkow (2004), however, has expanded the discrete-choice literature to allow consumers to choose multiple goods in an econometrically tractable way, so this problem

---

<sup>34</sup>Robustness checks that allow for time-dependence by using lagged dependent variables as explanatory variables demonstrate that the estimated effects remain essentially unchanged.

could be overcome. Regrettably, there is a more concrete difficulty with such an approach in this case— I do not observe downloads, and so it is impossible to use a model that is based on consumers choosing between album sales and downloads. Instead, by allowing the amount of file sharing activity for an album to affect directly the utility from purchasing an album, I apply the flavor of a more complete model within the constraints of this analysis. If file sharing is complementary to sales, the estimated coefficient  $\mu$  will be positive— file sharing makes albums more attractive to consumers. If file sharing is substituting away from sales, the estimated coefficient will be negative.<sup>35</sup>

To complete the model, I assume that consumers have the option to choose no album, but rather select an outside good, denoted by  $i = 0$ , whose utility is normalized in each week to zero except for the idiosyncratic error, so that

$$u_{n,0,t} = \epsilon_{n,0,t} \quad (6)$$

In each period, consumers choose the single good from the set of available options (all albums and the outside good) that yields the highest utility. Following McFadden’s seminal paper (1973) and the literature that grew from it, I assume that the  $\epsilon_{n,i,t}$  idiosyncratic error term is distributed according to a Type I extreme value distribution, which allows for easy analytical integration over consumers. Doing so, and then aggregating over consumer choices to calculate market shares yields the following standard multinomial logit market shares:

$$s_{i,t} = \frac{\exp(\delta_{i,t} + \xi_{i,t})}{1 + \sum_{k \in J_t} \exp(\delta_{k,t} + \xi_{k,t})} \quad (7)$$

where  $J_t$  is the set of albums that are available for sale in week  $t$ .<sup>36</sup> The quantity sold for

---

<sup>35</sup>This model is mathematically very close to a nested-logit specification in which each album forms a nest, and then after choosing an album the consumer decides between an album purchase or download. However, a nested logit model of this form would, of course, require data on downloads as well as force downloads to be aggregate substitutes for sales.

<sup>36</sup>Keep in mind that as I am not working with data on albums available for sale, albums that are outside of

album  $i$  in week  $t$  is then  $s_{i,t}M_t$ , where  $M_t$  is the market size in week  $t$ . The market size is assumed throughout to be equal to the population of the United States, which is interpolated linearly from data reported by the U.S. Census Bureau. Taking the natural log of the shares of the goods and the standard rearrangement yields the following relationship for market shares:

$$\ln(s_{i,t}) - \ln(s_{0,t}) = \delta_{i,t} + \xi_{i,t} = \mu_i + \lambda X_{i,t} + \theta q_{i,t}^{FS} + \xi_{i,t} \quad (8)$$

The left-hand side of equation (8), which is simply the mean utility of good  $i$  relative to the outside good that is implied by the data, can be calculated empirically. The market share of album  $i$  in week  $t$  is  $s_{i,t} = \frac{q_{i,t}}{M_t}$ , the quantity of sales as a fraction of the population of the United States in week  $t$ ; that is, the share of the U.S. population that purchased album  $i$  in week  $t$ , under the logit assumption that all consumers buy at most one album per week. Analogously, the market share of the outside good is the share of the population of the United States that does not purchase an album (an inside good), and thus  $s_{0,t} = \frac{M_t - \sum_{k \neq 0} q_{k,t}}{M_t}$ . With these calculations in hand, it is possible to estimate equation (8) through the standard linear regression techniques, given appropriate assumptions on the distribution of  $\xi_{i,t}$ . I proceed under the assumption that  $\xi_{i,t}$  has the same distribution as  $\varepsilon_{i,t}$  in equations (2) and (3). Specifically the omitted variable of artist “quality” creates a positive correlation between  $\xi_{i,t}$  and  $q_{i,t}^{FS}$  which necessitates again the use of an instrumental variables technique.

There are three components to the mean utility specified in equation (8) in addition to the error term:  $\mu_i$ , the time-invariant fixed quality of the album,  $\lambda X_{i,t}$ , the effect of album and week characteristics that affect quality, and  $\theta q_{i,t}^{FS}$ , which measures the effect on the quality of the album that arises from the amount of file sharing.

The first term is simply an album-specific fixed effect which captures all of the specifics

---

the data sample are treated as part of the outside good and thus not available for sale.

of an album that do not change over its life. This includes such characteristics as the artist releasing the album, the release label, pre-release advertising, and any unobservable but real album characteristics such as artistic merit or pop culture significance. As in the reduced form models above, this fixed effect term goes a long way towards addressing the biases associated with estimating the effects of file sharing on sales (or more specifically, relative market share).

The second term,  $\lambda X_{i,t}$ , captures the effects of time-varying album or artist characteristics as well as potential week characteristics, such as holidays and the general decay of the utility from purchasing the album over time. Again, I impose a quadratic time decay for simplicity, so that  $\lambda X_{i,t}$  can be written as

$$\lambda X_{i,t} = \lambda_1(t - \tau_i) + \lambda_2(t - \tau_i)^2 + \lambda_3 W_{i,t} \quad (9)$$

where  $\tau_i$  is the release date of album  $i$  so that  $(t - \tau_i)$  are the number of weeks since release, and  $W_{i,t}$  are the other album- and week-level covariates which affect the mean quality of an album in a particular week, including Christmas time dummy variables, airplay status, and indicators for television appearances and Grammy award nominations and wins. The decay of mean utility over time is extremely important to capture, as most albums follow a pattern in which weekly sales peak in the debut week, and drop after that. As Figure 3 illustrates, the quadratic decay specification is a good approximation of the actual decay rate of implied mean utility.<sup>37</sup>

Finally, the last term,  $\theta q_{i,t}^{FS}$ , captures the effects of file sharing on sales. As above, I will estimate both a pooled effect as explicitly specified in Equation (8) as well as an effect that differs across ex ante popularity. In the case of estimating a differential effect,  $\theta$  should be

---

<sup>37</sup>The sales of recorded music appear to follow decay patterns and seasonality patterns similar to those of motion pictures. Einav (2002) provides an excellent analysis of the seasonality in the market for motion pictures. His conclusions for motion pictures regarding gross and net seasonality are very likely applicable to the market for recorded music as well, though that analysis is beyond the scope of this paper.

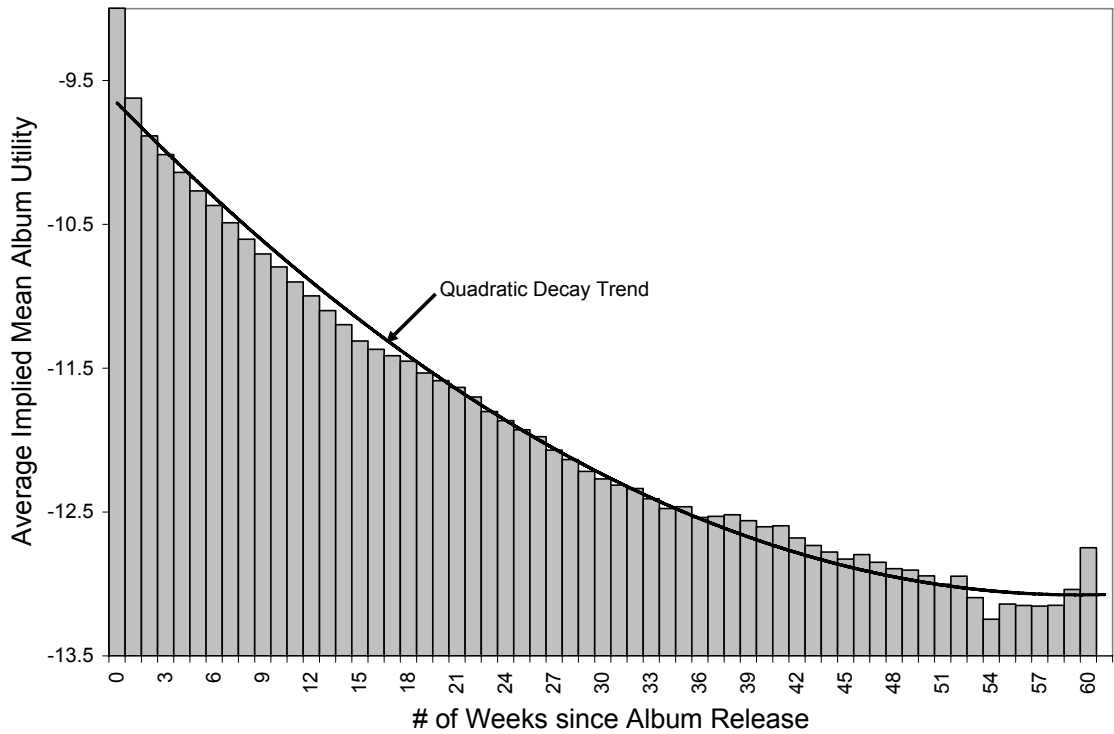


Figure 3: Average Implied Mean Album Quality by # of Weeks Released

subscripted by  $i$ , as it varies by artist, and instead of  $\theta q_{i,t}^{FS}$  should be written as

$$\theta_i q_{i,t}^{FS} = (\theta_1 + \theta_2 P_i) q_{i,t}^{FS} \quad (10)$$

with the popularity index  $P_i$  defined as detailed in the previous section. The estimated marginal effects of file sharing on sales are also defined analogously, though now I estimate the marginal effect of file sharing not on sales, but instead on the implied mean album utility relative to the outside good.

Table 6: Logit Demand System Results  
 Dependent Variable: Implied Album Utility

	(1)	(2)	(3)
Max # of Files Available (in 100,000s)	0.174 [0.103]*	-0.108 [0.140]	1.013 [0.153]***
Max # of Files Available (in 100,000s) * Popularity Index			-0.00750 [0.001]**
Specification	OLS	IV	IV
Observations	7938	7938	7938

1. Robust standard errors in brackets
2. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%
3. The full set of control variables (time trends, holidays, television, airplay, Grammy awards, and album-level fixed effects) is included, and all variables have the expected sign.
4. Excluded Instruments: Indicators for lawsuit announcement, lawsuit implementation, and interactions with artist popularity

### 5.2.2 Results in Multinomial Logit

Although the multinomial logit model is already defining a non-linear relationship between quantity sold and the amount of file sharing, it is still prudent to examine the right specification for the analysis of the relationship between the implied mean album utility and the amount of file sharing. Analogously to the test used in the restricted reduced form, I use a Box-Cox transform to test again between specifying the number of shared files in levels or logs, under equation (8). This test rejects both the levels and logs specification for the number of shared files, although the likelihood is slightly higher for the levels specification. While it is possibly useful to apply the estimated Box-Cox transform throughout the analysis, I will use instead the simple levels specification for clarity. Proceeding, then, with this specification, Table 6 presents the results of estimating Equation (8).

The estimates in Table 6 are strikingly similar to those found in the previous section. The pattern of the marginal effect of file sharing activity on sales in the reduced form model is the same as the pattern of the marginal effects of file sharing on implied product quality in a multinomial logit demand setting. This is important because it indicates that the restrictions imposed by the logit in order to allow for substitution between albums are

not fundamentally changing the estimated effect of file sharing on sales. Note also that the maximum number of files shared is now measured in units of 100,000 files in order to scale the estimates to be similar in magnitude to the previous results.

Column (1) provides the upwardly biased OLS estimate of the marginal relationship between the number of files shared and the implied mean utility of an album. Although the estimate is significant at only the 10% level, it indicates that an additional 10,000 copies available on file sharing networks would increase the implied utility of an album by 0.017, which is only .2% of the average album quality of -11.5 (for comparison's sake, recall that the outside good is fixed to have a utility of 0). This is hard to interpret and it is more instructive to instead translate the marginal effect of mean utility into a file sharing elasticity of sales. Using the algebra of the multinomial logit demand system above, the marginal effect of the number of files available on-line,  $q_{i,t}^{FS}$ , on the market share of album  $i$  is:<sup>38</sup>

$$\frac{\partial s_{i,t}}{\partial q_{i,t}^{FS}} = s_{i,t}(1 - s_{i,t})(\theta_1 - \theta_2 P_i) \quad (11)$$

and thus the marginal effect of file sharing depends on both the artist's popularity as well as the market share of the album.<sup>39</sup> Using the data in Appendix A and noting that the average market size in the sample is 290 million, this relationship implies that for an album with the mean amount of sales and file shared, the file sharing elasticity of sales is 0.14.

However, we know that this effect is positively biased, and thus turn towards the two-stage least squares estimates. Column (2) presented the pooled estimate, in which the marginal effect of file sharing on album quality is restricted to be the same for all albums. The TSLS procedure causes the estimated effect to become negative, although it is statis-

---

<sup>38</sup>Note that from Appendix A, the average market share for an album in the sample is 3.6e-5. The average market share for a new album is 3.2e-5, and for an album whose artist previously had a top 20 album is 5.4e-5.

<sup>39</sup>Again, see Nevo (2000) for a complete discussion of the substitution patterns in the multinomial logit demand system and the models derived from it. In particular, the common objection to the logit model is that elasticities differ across goods only through the market share of the good. Of course, by allowing for heterogeneous marginal effects, I do allow the elasticities to differ across albums in this dimension.

tically insignificant. Nevertheless, an analogous calculation of the elasticity for an album with mean levels of file sharing and sales is  $-.08$ . This again is consistent with Oberholzer and Strumpf's conclusions of no effect, in the statistical sense, of file sharing on sales when we don't allow for effects to differ across artists.

Column (3) presents the marginal effects of file sharing on implied mean utility when this marginal effect is again allowed to differ according to the artist's ex ante popularity, as specified in equation (10). In this case, we see a positive and statistically significant effect of files shared on sales for a new artist and an effect that becomes more negative as the artist's ex ante popularity increases. Using the same procedure for calculating elasticities, but now using the mean levels of files shared and sales for a new artist, the file sharing elasticity of sales for a new artist is equal to  $.67$ , which is slightly higher than the estimate obtained in the log-linear reduced form. Contrastingly, for an artist of the highest possible ex ante popularity of 200, indicating a #1 album in the 10 years prior to the sample period, the estimated marginal effect of increasing files shared by 10,000 on market share is  $-0.049$ , with a standard error of  $0.021$ , which is significant at the 5% level. Using the mean data for an artist of popularity of 200, the estimated elasticity is  $-0.82$ , which is also slightly stronger than found in the log-linear reduced form. These elasticities make the distributional effect of file sharing clear: file sharing is reducing the sales of ex ante popular artists while redistributing some of these lost sales to smaller, less well known artists.

Recall that the reason for using the logit demand system is that it allows for competition effects between albums, which are important to consider in counterfactual exercises. In particular, if removing file sharing raises the quality of one album, it reduces sales of all other albums, since all albums are competing against each other. Within the multinomial logit model, the file sharing cross-elasticity of market share is:

$$\frac{\partial s_{i,t}}{\partial q_{j,t}^{FS}} = -s_{i,t}s_{j,t}(\theta_1 - \theta_2 P_j), i \neq j \quad (12)$$

which is of opposite sign of  $\partial s_{j,t}/\partial q_{j,t}^{FS}$ , highlighting the fact that if the availability of my songs on a file sharing network hurts my sales, then they help your sales and vice versa. Again, this is particularly important because when considering the effects of reducing file sharing, the reduction in file sharing for large selling, ex ante popular artists increases the implied utility of the album. As these albums are the biggest sellers, the increases in their sales have a large crowding out effect on other albums, which tempers the gains made from reducing file sharing. This effect is absent when considering each album as its own monopoly market as the simple reduced form estimation does. Failure to account for this effect leads to potentially grossly overstated and misleading calculations regarding aggregate changes in file sharing behavior. Therefore, in the next section I use the estimation from this multinomial logit demand system to examine the counterfactual worlds in which there are market-level changes in the amount of file sharing activity.

## 6 Implications for the Industry in the Short Run

With the estimated effects of file sharing in hand, I now proceed to rerun the two counterfactual exercises. I begin by estimating the change in the sales of recorded music sales as a result of the lawsuits put forth by the RIAA. Recall that on June 25, 2003, the RIAA announced that it would begin monitoring file sharing networks and taking legal action against users of these networks. This announcement had the effect of reducing file sharing activity across the board, which according to the estimates above, suggests a change in the pattern and level of sales in the industry.

To estimate the effect that the announcement regarding the lawsuit strategy had, as well

as the implementation of the first round of lawsuits themselves, is a relatively straightforward exercise. By exploiting both stages of the two stage least squares procedure, it is a simple calculation to determine what the level of file sharing would have been in the absence of the lawsuit plan by simply subtracting out the effect of the lawsuit plan and its implementation from the first stage estimates of file sharing activity. Then, using the coefficient estimates from Column (3) of Table 6, these changes in file sharing activity are translated into changes in the implied quality of albums, which are in turn translated into changes in the sales of the albums, which can be aggregated up to market level changes using the appropriate weights.

Doing this reveals that as a result of the lawsuit strategy followed by the RIAA against users of file sharing networks, album sales increased by 2.9% over the 23 weeks in the data sample after the strategy was announced. During this period, actual record sales in the U.S. were an average of 11,470,652 albums per week, based on national level data reported by Billboard magazine (2003) each week, and thus would have been 11,147,378 per week in the absence of the reduction in file sharing caused by the lawsuit strategy. Again using a baseline of \$5 markup per CD, this translates to an increase in industry profits of \$1,616,370 per week, or \$37 million over the 23 week period after the lawsuit strategy was announced to the public. Note that, as expected, this increase in profits is much smaller than the number obtained using the simple reduced form estimates (approximately \$160 million), which fail to account for the effect of competition among albums.

Similarly, the data can be used to understand what effects eliminating file sharing across the board would have had. In particular, again using the estimates from Column (3) of Table 6, it is possible to estimate how industry wide sales would change if file sharing were scaled back further, or didn't exist at all by simply subtracting out the effects that file sharing have had on the implied mean utility of albums and aggregating these effects. However, simply removing all file sharing from the estimation above in many cases takes the data far out of

Table 7: The Effect of Removing File Sharing on Industry Sales

% of File Sharing Activity Removed	% Increase in Industry Sales	File Sharing Elasticity of Sales
10	1.8%	-0.18
20	5.3%	-0.26
30	10.3%	-0.34
40	17.0%	-0.43
50	25.7%	-0.51

Note: Chart performance based weights used to aggregate sales across industry. Point Estimates used to calculate effects come from Table 6

sample, and so the usual caveats apply.

To estimate the industry-wide effect of reducing file sharing, I perform only calculations that are arguably within the data space or slightly outside of it. I calculate the effect of removing 10% of file sharing across the board, and then continue removing another 10% until 50% of file sharing has been removed;<sup>40</sup> that is, I perform what is essentially an experiment of “deleting” 10% of files at a time uniformly across artists from file sharing networks. The industry-wide effects are then calculated as above, using the estimates of the effects of file sharing on mean album utility to calculate changes in album-purchase utilities and then translating those changes into album sales, which are aggregated up to the industry level. The effects of these changes to the quantity of file sharing are reported in Table 7.

We see that the non-linearity of the multinomial logit manifests itself in a way such that as the data is taken more and more out of sample, the aggregate file sharing elasticity of sales is increasing. Focusing on the estimated elasticities, they suggest that removing file sharing would increase industry wide recorded music sales by anywhere from 18% to 51% relative to sales over the sample period.<sup>41</sup> As a point of perspective, recall from Figure 1,

---

<sup>40</sup>Why stop at 50%? As discussed previously, taking the estimates too far out of sample is problematic, and as shown in Table 3, the effect of the lawsuits was to reduce file sharing by around 40%. Thus, the data should have no problem speaking to the effects of reducing file sharing by at least this proportion. However, going much beyond this level is unlikely to have much validity.

<sup>41</sup>Of course, extrapolating these elasticities that far is an out-of-sample exercise.

that real recorded music sales were approximately 30% lower in 2003 (the primary sample period) than they were in the industry's peak year of 1999, and approximately 65% lower than it would have been in 2003 had the industry stayed on its 1997-1999 trend. In this light, these estimates of the aggregate effect file sharing has had on sales are reasonable. If file sharing were to be reduced across the board by 30%, sales would increase by just over 10%. According to Billboard magazine (Market Watch 2004), in 2003 industry-wide sales totaled approximately 660 million during the calendar year. Thus, an increase in sales of 10% would amount to 66 million albums during the year. Returning to the estimate of \$5 of variable profit per sale, this equates to \$330 million of additional profit in the calendar year 2003.

However, as highlighted in the previous discussions, this change has not been uniform across all artists. Rather, very different effects of file sharing on sales have been found for ex ante unknown artists relative to ex ante popular ones. Therefore, it is possible to investigate the effect that file sharing has had on not only the level of industry sales, but also on the distribution of sales in the industry. Again, given the estimated effects from the previous section, it is a simple exercise to perform the same calculations at the album level, and rather than aggregate them up to the industry level total, instead focus on the distributional changes.

Due to scale issues, it is difficult to see the effects of file sharing on the distribution of sales through a plot of a kernel density estimation; however, Table 8 presents some key percentiles which reveal what effect file sharing is having on the distribution of sales in the industry.

In the counterfactual world with 30% file sharing, the lower 75% of the distribution of sales is shifted further to the left, while the top of the distribution increases its sales. This is what should be expected given the estimates from above. Artists who are ex ante unknown, and thus most helped by file sharing, are those artists who sell relatively few al-

Table 8: The Distribution of Actual Sales and Sales with a 30% Reduction in Files Shared

Percentile	Sales	
	Actual Sales	Without File Sharing
1%	73	70
5%	170	166
10%	281	277
25%	757	745
50%	2852	2851
75%	10110	9831
90%	26531	26934
95%	45255	47357
99%	133983	165054

bums, whereas artists who are harmed by file sharing and thus gain from its removal, the ex ante popular ones, are the artists whose sales are relatively high. Thus, the existence of file sharing, in addition to the aggregate mean effect discussed above, has a clear distributional impact on the sales of recorded music: the distribution becomes more skewed and the peak of the distribution is shifted leftward.

This conclusion leads to further questions regarding the impacts that file sharing has had and will have on the recorded music industry. In particular, if file sharing essentially shifts sales away from established acts toward unknown acts, this has potentially very important implications for how talent is developed and distributed in the industry. As with the simple short-run effects of file sharing on sales, the direction of the impact is not ex ante clear. While one might guess that increasing the sales of new acts would lead to more investment in developing new talent, it is also possible that the investment in new acts is done as a fishing expedition to find artists who will sell millions of records. File sharing is reducing the probability that any act is able to sell millions of records, and if the success of the mega-star artists is what drives the investment in new acts, it might reduce the incentive to invest in new talent.<sup>42</sup> This is, at its heart, an empirical question which is left to future work.

---

<sup>42</sup>This is consistent with the reduction in the sales numbers of the top selling albums over the past several years, as well as with the decline in the number of gold and platinum album certifications in the same time period.

## 7 Conclusions

This paper has investigated the short-run effects of file sharing on the sales of recorded music in the United States, during a time period in which it appears that the industry had not yet begun to change market strategies. I exploit the timing of an industry strategy of suing file sharing network users to estimate these effects. Naïve estimates which do not allow for the effect of file sharing to differ systematically across artists yields results similar to those found in the literature previously, suggesting that file sharing has not had a significant effect on the sales of recorded music.

Further inspection, however, reveals that it is unrealistic to believe that the effects of file sharing are constant across all artists as the costs and benefits of file sharing differ with the ex ante popularity of the artist. This suggest that ex ante unknown artists are likely to see more positive overall effects of file sharing than ex ante popular artists are. By adopting an estimation procedure which allows for the effect to vary according to measures of artist popularity, I find that file sharing has had strong effects on the sales of music. In particular, new artists and ex ante relatively unknown artists are seen to benefit from the existence of their songs on file sharing networks, while ex ante popular artists suffer for it.

And while the average effect across artists is essentially zero, the average effect on sales is not zero, as more popular artists not surprisingly tend to have higher sales. Thus, this paper finds that file sharing has had large, negative impacts on industry sales and that the RIAA's strategy of suing individual file sharing users has led to reduced file sharing activity and sizeable increases in sales.

Furthermore, the differential effect of file sharing on the sales of artists of different levels of ex ante popularity has led to a dramatic shift in the distribution of sales among artists, as new and less popular artists are now selling more records while star artists have seen their sales shrink, compacting the distribution of outcomes. It remains an open ques-

tion, left for future work, what effect this distributional change has had or will have on the investment in new talent and the distribution of returns to that talent in the recorded music industry.

## References

- [1] Anantham, S. and A. Ben-Shoham (2004), "Quality Uncertainty and Monopolistic Pricing," Harvard University mimeo.
- [2] Apple.com (2003), website, "Apple Launches iTunes for Windows," October 16, 2003, <http://www.apple.com/pr/library/2003/oct/16itms.html>
- [3] Apple.com (2003), website, "iTunes Sells 1.5 Million Songs During Past Week; Five Times Napster's First Week Downloads," November 6, 2003, <http://www.apple.com/pr/library/2003/nov/06itunes.html>
- [4] BigChampagne.com (2004), website, <http://www.bigchampagne.com>.
- [5] *Billboard Magazine* (2000), "Is Biz Poised For Renewed Price Wars?," January 8, 2000.
- [6] *Billboard Magazine* (2003), "Market Watch," various issues.
- [7] *Billboard Magazine* (2004), "Market Watch," January 10, 2004.
- [8] Blackburn, D. (2002), "Complementarities and Network Externalities in Casually Copied Goods," *Estudios de Economia*, June 2002, 29-1, Pp 71-88.
- [9] Christman, E. (2003), "UMVD Expands Market-Share Dominance," January 18, 2003, *Billboard Magazine*.
- [10] Christman, E. (2004), "Ed Christman, UMG tops album share for fifth year," January 17, 2004, *Billboard Magazine*.
- [11] CNNMoney (July 19, 2000), "Napster: 20 million users," <http://money.cnn.com/2000/07/19/technology/napster>.

- [12] Einav, L. (2003), "Gross Seasonality and Underlying Seasonality: Evidence from the U.S. Motion Picture Industry," forthcoming, *RAND Journal of Economics*.
- [13] Fine, M (2003), "Report of Michael Fine on Napster and Loss of Sales," <http://www.riaa.com/news/filings/pdf/napster/fine.pdf>.
- [14] Fuoco, C. (2003), website, AMG Biography for Josh Kelley, <http://www.allmusic.com>.
- [15] Gentzkow, M (2004), "Valuing New Goods in a Model with Complementarity: Online Newspapers," Harvard University mimeo.
- [16] Godfrey, L. G. and M. R. Wickens (1981), "Testing Linear and Log-Linear Regressions for Functional Form," *The Review of Economic Studies*, July 1981, 48-3, Pp. 487-496.
- [17] Grammy Awards (2004), website, <http://www.grammy.com>.
- [18] Liebowitz, S. (September 1982a), "Durability, Market Structure And New-used Goods Models," *American Economic Review*, September 1982a, 72-4, Pp. 816-824.
- [19] Liebowitz, S. (2003), "Will MP3 downloads Annihilate the Record Industry? The Evidence so Far," in *Advances in the Study of Entrepreneurship, Innovation, and Economic Growth*, edited by Gary Libecap, JAI Press.
- [20] Liebowitz, S. (2003), "Pitfalls in Measuring the Impacts of File Sharing," mimeo.
- [21] McFadden, D. (1973), "Conditional Logit analysis of Qualitative Choice Behavior," in P. Zarembka, eds., *Frontiers of Econometrics*, New York, Academic Press.
- [22] Milette, R. (2004), website, <http://rym.waglo.com/wordpress/>

- [23] Mortimer, J.H. and A. Sorensen. (2004), "Digital Distribution and Demand Complementarities: Evidence from Recorded Music and Live Performances," mimeo.
- [24] Nevo, A. (2000), "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics & Management Strategy*, Winter 2000, 9-4, Pp. 513-548.
- [25] Nielsen SoundScan (2004), website, <http://www.soundscan.com>
- [26] Oberholzer, F and K. Strumpf (March 2004), "The Effect of File Sharing on Record Sales: An Empirical Analysis." HBS mimeo.
- [27] Recording Industry Association of America (2003), website, "Year End Marketing Reports," <http://www.riaa.com/news/marketingdata/yearend.asp>
- [28] Recording Industry Association of America (1999), website, "Legal Cases," [http://www.riaa.com/news/filings/pdf/napster/Napster\\_Complaint.pdf](http://www.riaa.com/news/filings/pdf/napster/Napster_Complaint.pdf)
- [29] Recording Industry Association of America (2003), website, "Press Release: Recording Industry To Begin Collecting Evidence And Preparing Lawsuits Against File "Sharers" Who Illegally Offer Music Online," June 25, 2003, <http://www.riaa.com/news/newsletter/062503.asp>
- [30] Recording Industry Association of America (2003), website, "Press Release:Recording Industry Begins Suing P2P File Sharers Who Illegally Offer Copyrighted Music Online," September 8, 2003, <http://www.riaa.com/news/newsletter/090803.asp>
- [31] Recording Industry Association of America (2004), website, "Consumer Purchasing Trends," <http://www.riaa.com/news/marketingdata/pdf/2003consumerprofile.pdf>

- [32] Rob, R. and J. Waldfogel (November 2004), "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students." NBER Working Paper #10874.
- [33] Schwartz, J. (2004), "A Heretical View of File Sharing," April 5, 2004, P. C1, *The New York Times*.
- [34] Standard and Poor's (2002), "Movies And Home Entertainment," *Industry Surveys*, November 2002.
- [35] Takeyama, L. (1994), "The Welfare Implications of Unauthorized Reproduction of Intellectual Property in the Presence of Demand Network Externalities," *Journal of Industrial Economics*, June 1994, 42-2, Pp. 155-66.
- [36] Takeyama, L. (1997), "The Intertemporal Consequences of Unauthorized Reproduction of Intellectual Property," *Journal of Law and Economics*, October 1997, 40-2, Pp. 511-22.
- [37] TVToMe (2004), website, <http://www.tvtome.com>.
- [38] Zentner, A. (2003), "Measuring the Effect of Online Music Piracy on Music Sales," University of Chicago Working Paper.

## Appendix A Data Appendix

### A.1 A Description of the Data Sources

As discussed in the main text, the data for the empirical analysis undertaken in this paper comes primarily from two sources. Data on album sales come from Nielsen SoundScan, which tracks retail sales of music and music video at over 14,000 outlets in the United States, including retail stores, mass merchants, and on-line stores. These 14,000 outlets correspond to approximately 90% of total U.S. music industry sales and Nielsen SoundScan claims that their sampling is correctly representative of the total market.<sup>43</sup> The data collected by Nielsen SoundScan is the primary source of data on retail success for the industry, and is used as the source for the weekly Billboard music charts, published each week in Billboard magazine.

Data on the file sharing activity for albums come from BigChampagne, which tracks all visible file sharing activity on the 5 largest file sharing networks.<sup>44</sup> While it is impossible to know how representative this sample is, it is very likely that BigChampagne's coverage of the file sharing world is both wide and representative, as there is very little file sharing activity taking place on smaller networks, presumably due to the large network effects involved in file sharing communities. Additionally, Oberholzer and Strumpf (2004), in using data from a small file sharing network, find that the distribution of activity across artists and genres within smaller file sharing networks is very similar to that of the larger networks. BigChampagne was founded in 2000, and as one of the first and most comprehensive companies to track file sharing activity, its data is the most widely used of its kind in the music industry. According to BigChampagne, their data is used by many major recording labels

---

<sup>43</sup>Although I can not independently verify this, the data collected by Nielsen SoundScan is universally accepted as accurately portraying the recorded music industry, and is essentially the only source for data on retail performance.

<sup>44</sup>Again, throughout the timeframe of my data sample, these networks are the FastTrack network (Kazaa), Grokster, eDonkey, iMesh, and Overnet.

as well as radio stations and media outlets (such as Entertainment Weekly and E! Entertainment Television) to monitor file sharing. Both Entertainment Weekly and E! Entertainment Television use BigChampagne data to publish weekly file sharing charts.

While Nielsen SoundScan collects and processes sales data on essentially all albums that are for sale in stores, BigChampagne does not process data for all albums and all artists that are in stores. While as much raw data as possible is collected, it is not all processed into file sharing numbers for every album. Instead, BigChampagne processes the data at the request of their clients.

I build other album covariates from a variety of other sources. Using data from the website [www.tvtome.com](http://www.tvtome.com) (2004) which reports the guests scheduled to appear on various television shows, I construct dummy variables to control for the week and the week after an artist appears on either The Tonight Show with Jay Leno, Late Night with David Letterman, the Oprah Winfrey Show, Saturday Night Live, or the Superbowl Halftime Show. These television appearances are selected because they are the major national broadcasts on which artists may appear, and industry knowledge suggests that these promotional appearances tend to increase sales. Variables designating Grammy award nominations<sup>45</sup> and Grammy award wins for 2003 were constructed using data from [www.grammy.com](http://www.grammy.com). Lastly, the Billboard charts for radio airplay were used to construct four dummy variables indicating various levels of radio airplay<sup>46</sup>. To the extent that radio stations' maximize over the number of listeners (which would presumably maximize advertising revenue), it should be the case that radio airplay is a good proxy for the "quality" of a song or artist during a period. That is, these variables are used to help capture the week-to-week fluctuations in consumer tastes for different albums.

---

<sup>45</sup>The Grammy Awards are the recorded music industry's annual award shows, similar to the Oscars for motion pictures.

<sup>46</sup>The dummy variables are: Having the #1 radio airplay song, having a song between #2 and #10, having a song between #11 and #40, and appearing on the chart at or below #41.

## A.2 A Detailed Description of the Data Sample

To construct the sample of albums for the analysis, I first restrict the sample to albums that have had enough commercial success to have appeared for at least one week on the Billboard magazine Hot 200 album sales chart (the Hot 200). In most weeks, national sales of 5,000 albums will place an artist at the bottom of the Hot 200 chart. While this immediately removes very small albums from the analysis, it is necessary in order to focus attention on albums for which file sharing data is potentially available. Furthermore, I have restricted the analysis to albums that are composed of new material by a single artist. This is done to eliminate albums that contain work by multiple artists, such as movie soundtracks, as the success of the album and the searches and shared files that correspond to the album may be due to multiple artists. This would also cause difficulty in determining the ex ante popularity of the artist associated with the album. Additionally, I focus on albums consisting of new material only both because most file sharing (and radio airplay) is concerned with new material and also because knowledge of previously released material has already been dispersed throughout consumers and thus it would be more complicated to assess what impact file sharing has had. More practically, file sharing data was only available to me for the newest releases and is arranged by artist; focusing on these albums maximizes data availability.

Albums were then considered only if they were released after September 24, 2002 due to the fact that the file sharing data that was available to me started approximately 2 weeks before that time, except in rare cases. From that initial release date, I focus my attention on albums released within the next year so that the latest album release in the sample is September 16, 2003.<sup>47</sup> Finally, albums which are classified by Nielsen SoundScan as being "gospel" records are also excluded from the sample, as the weekly sales numbers for these

---

<sup>47</sup>In general, recorded music albums are released for sale on Tuesdays.

Table 9: Summary Statistics for Weekly Sales, by ex ante Artist Popularity

	ALL	Popularity=0	Popularity = [1,100]	Popularity = [101,180]	Popularity = [181,199]	Popularity = 200
Mean Weekly Sales	11,516	7,792	5,051	7,198	12,002	29,767
SD Weekly Sales	32,446	16,645	8,514	13,932	38,693	59,768
Minimum	7	7	78	79	42	71
10% Percentile	281	207	293	278	303	651
25% Percentile	757	567	845	741	756	2,296
Median	2,851	2,163	2,071	2,577	2,768	9,073
75% Percentile	10,110	7,802	6,811	7,246	11,740	28,776
90% Percentile	26,530	19,373	11,215	20,130	26,634	74,096
Maximum	874,137	297,381	122,400	213,728	874,137	601,516
# Albums	197	76	12	35	49	24
# Album-Weeks	7938	3055	492	1330	2002	1059
% of Albums	100%	39%	6%	18%	25%	12%
% of Album Weeks	100%	38%	6%	17%	25%	13%

albums include sales numbers from the Christian Booksellers Association. Because no other album's sales include numbers from this market, these albums have been excluded. This left a potential sample of 602 albums. Using this list of 602 albums, I was able to collect some form of file sharing data from BigChampagne for 197 of them. This sample of 197 albums is described in more detail in Tables 9 and 10 and is the primary source for analysis throughout the paper. Weekly sales data for the albums was then purchased from Nielsen SoundScan from the week of release up through the week ending February 8, 2004, resulting in 9,908 unique album-weeks in the broadest possible sample. Albums range from 21 to 71 weeks of data, depending on how early in the sample the album was released.

However, in order to maintain the ability to focus solely on short-run effects rather than industry-level responses in the long run, I am forced to exclude the final nine weeks of data from the sample, resulting in a sample that ends on November 30, 2003 and includes 7,938 unique album-weeks. As discussed in the main text and illustrated in Figure 2, there appears to be a structural change in the choice of album prices starting at that time. Therefore, the data sample consists of album data between the weeks of September 29, 2002 and November 30, 2003, leaving a total of 62 weeks of data.

Table 10: Summary Statistics for Weekly Shared Files, by ex ante Artist Popularity

	ALL	Popularity=0	Popularity = [1,100]	Popularity = [101,180]	Popularity = [181,199]	Popularity = 200
Mean Weekly Shared Files	77,904	66,064	48,249	57,721	69,705	166,683
SD Weekly Shared Files	111,206	101,081	63,875	95,150	91,491	157,609
Minimum	0	0	0	0	0	1,990
10% Percentile	1,080	740	530	609	2,431	23,080
25% Percentile	7,725	5,211	1,441	4,075	11,816	43,647
Median	32,260	24,854	7,208	21,430	30,379	107,076
75% Percentile	93,984	72,970	94,752	77,733	79,823	253,968
90% Percentile	232,262	226,788	143,985	152,178	207,523	413,920
Maximum	788,895	556,248	249,896	788,895	519,746	726,455
# Albums	197	76	12	35	49	24
# Album-Weeks	7938	3055	492	1330	2002	1059
% of Albums	100%	39%	6%	18%	25%	12%
% of Album Weeks	100%	38%	6%	17%	25%	13%

Because the 197 albums in the sample were not randomly chosen from the potential sample of 602 albums, one might be concerned about how similar the data sample is to full population of albums. While sales data is not available for the albums not in the sample, it is possible to compare the total Billboard chart performance of the two sets of albums. Figure 4 shows a histogram of the number of weeks that an album appears on the Billboard Hot 200 chart between the date of release through February 8, 2004. It appears that the sample of albums for which file sharing data is available is slightly more successful than the general album population, though not by a large amount. While this difference is small, I reweight observations to equalize the distribution of chart success for the sample to that of the full population. This reweighting never changes any qualitative results, nor does it cause even moderate quantitative changes. Therefore, in the analysis, I apply weights only when aggregating up from individual albums to the market level in counterfactual exercises.

### A.3 A Detailed Description of the Measure of File Sharing

As discussed in the main text, the primary variable used to summarize the amount of file sharing activity for an album is the number of copies of songs from an album that are avail-

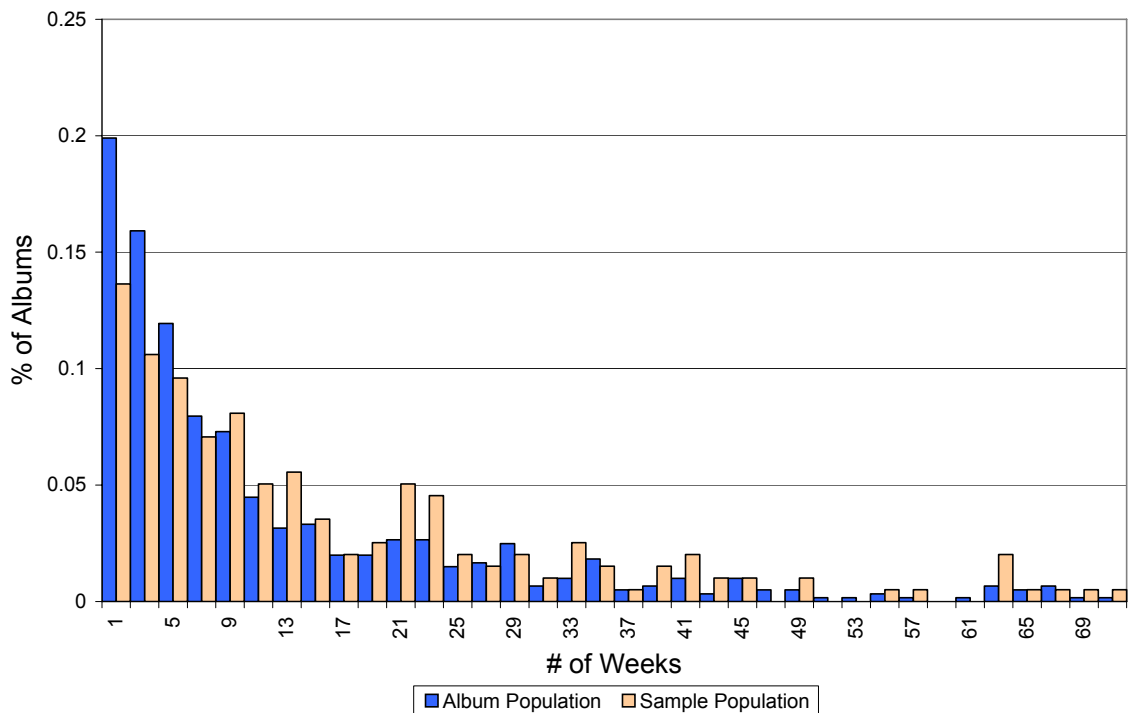


Figure 4: Total Number of Weeks on Hot 200 Sales Chart, for Sample and Population of Albums

able on the file sharing networks. This is constructed by taking the reported fractions of file sharing network users that are sharing a particular song and multiply by the size of the file sharing network that week, as measured by the average number of users logged in during the week, using data provided by Robin Millete (2004). After transforming the BigChampagne data measured in percentages of users into the total number of files available, it is still left to take song-level data and map it into album-level data. This is done in several ways to ensure the robustness of the results.

In addition to calculating the number of shared copies of the most shared file on an album, as detailed in the main text, I construct several other measures of file sharing in order to provide robustness to the results.

Specifically, I follow a similar procedure in which I map an album into the song which has the median number of copies available on the networks. This construction is then taking a weaker stance on substitutability, imposing that the relevant file sharing that matters for an album is half of the tracks on an album. I choose to focus on the median number of copies rather than the mean (or even the total) number of copies of songs for several reasons. Primarily, this is due to the fact that many albums in the sample have “songs” that are not really songs at all— rather they are 20 or 30 second spoken introductions or similar. Additionally, albums vary in the number of songs that they contain, and so a total would be very misleading for some albums. Using medians helps to mitigate the effects of short versus long albums and undesirable tracks to some extent. Finally, I also use the least popular song on the file sharing networks as another robustness check.

Before moving on, it is worthwhile to take a quick look at some summary statistics for some of the data’s most important variables. Tables 9 and 10 provide detailed summary statistics for the weekly sales and weekly files shared for artists of different ex ante levels of popularity. In particular, there is a clear, obvious pattern in the data, where ex ante more popular artists have both larger sales numbers as well as larger amounts of file sharing than ex ante unknown artists. New artists have mean levels of sales and file sharing comparable to artists of medium ex ante popularity, though with greater variance, representing the increased breakout potential, as well as failure potential, of new artists. Furthermore, new artists not surprisingly also sell fewer albums and experience lower levels of file sharing than star and superstar artists do.

Finally, Figure 5 graphs the percentage of all album sales in the US that are included in the data sample. This percentage is initially very small, as early weeks contain very few albums, but rises quickly so that the sample contains between 10% and 20% of total industry sales for the majority of the sample, before fading off again at the end of the sample after albums stop entering the data in week 52 out of the total of 62 weeks.

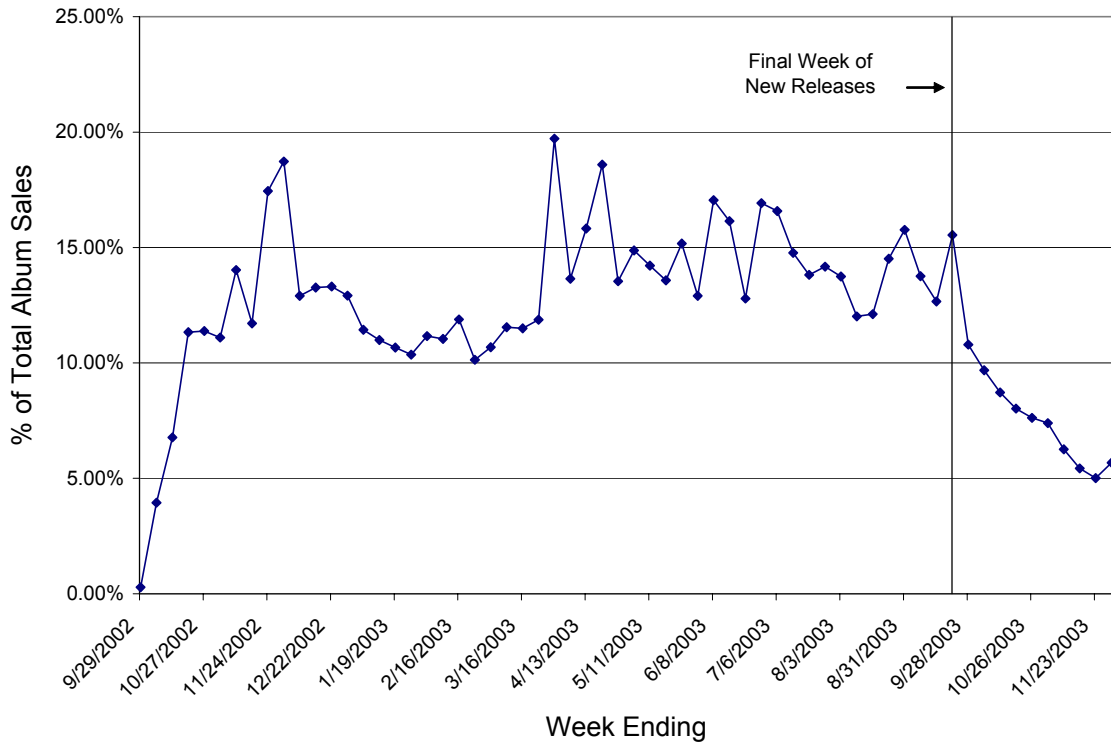


Figure 5: % of Total National Sales Data in Data Sample by Week

The final data issue worth mentioning is that BigChampagne does not track file sharing activity for all songs on every album in the dataset. In particular, for 69 of the 197 albums, file sharing data is available only for one song on the album. When using the maximum number of copies of a song, I will consider the data for the one song tracked by BigChampagne to be the maximum across all songs on the album. While it is possible that in some cases this may not be true, it is likely that these instances are rare. Because BigChampagne tracks albums and songs at the request of their clients, if only one song on a particular album is tracked, it is reasonable to assume that the clients care primarily about the single song, and thus that consumers also care primarily only about that song. When using other variables as a robustness check, I remove these albums from the dataset, as the one song tracked is almost certainly neither the median nor the least available track.