The Heterogenous Effects of Copying: The Case of Recorded Music

David Blackburn*

First Draft: April, 2004 This Draft: June 1, 2006

Abstract

The availability of copies to consumers has competing effects on sales that are heterogenous across producers. First, there is a direct substitution effect on sales as copies replace originals and second, there is a penetration effect which increases sales, as the spread of the good makes it more well-known through the population. The first effect is strongest for ex ante well-known producers, while the second dominates for ex ante unknown producers. This phenomenon is examined within the recorded music industry and evidence shows that file sharing has had a strong distributional impact on sales. However, the dominance of sales by well-known artists leads to a large negative of copying at the industry level.

*National Economic Research Associates, david.blackburn@nera.com. A previous version of this paper was circulated under the title: "On-line Piracy and Recorded Music Sales." This version of the paper is a revision of the first chapter of my 2005 Harvard University PhD dissertation. 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 Neilsen 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[1]

1 Introduction

The widespread adoption of the internet has changes lives and industries in a large number of ways. One of the most discussed and controversial is by facilitating the quick and easy transfer of digital goods to and from consumers and firms, both by the growing importance of online merchants such as Amazon.com as well as through the introduction of file sharing, which has allowed internet users to upload and download digital copies of goods without the involvement of the ultimate producer of the good. This has generated intense debate on the impact of the availability of these copies on the sales of the original goods, and has most notably involved the recorded music industry and their battles to prevent digital music files from being exchanged online through services such as Napster and its successors.

The theoretical literature on the impacts of copying on sales is varied, and suggests that the precise effect will depend on the particulars of the market in question. In this paper, I identify the two main competing effects that the trading of copies has on the sales of originals; the exchange of copies generates knowledge about the existence (or quality) of a product, thus increasing sales of originals while at the same time serving as a substitute for sales, thus decreasing sales. Therefore, the net impact of the "sharing" of copies depends on the relative magnitudes of these two effects, and will depend on the characteristics of the particular *good* in question, and thus can and will vary *within* industries. I then turn to an empirical investigation of the effect that file sharing has had on the sales of recorded music, and take advantage of information on the ex ante popularity of artists to allow for

[1] Fuoco 2003

heterogenous effects of file sharing on sales. This allows me to avoid the problems inherent in assuming a common effect for all producers (artists) and to examine the distributional impacts that file sharing has had.

The empirical findings indicate that heterogenous effects do exist in the recorded music industry and are very important. As ex ante popularity increase, the positive effect of copies being traded is reduced while the negative effect is enhanced, suggesting a strong distributional impact of file sharing. Ex ante unknown artists have their sales increased as a result of their songs trading on file sharing networks, while ex ante popularity artists have their sales reduced. Thus, the distribution of sales is compacted by file sharing. However, because overall industry sales are dominated by sales of ex ante well-known artists, this compacting of the distribution is accompanied by a reduction in the overall level of sales. Using my estimates of the effect of file sharing, counterfactual exercises suggest that the lawsuits brought by the music industry in mid-2003 (which reduced the level of file sharing, at least temporarily) 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 online were reduced across the board by 30%, industry sales would have been approximately 10% higher in 2003.

2 A Quick History of File Sharing in the Music Industry

2.1 File Sharing

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 success, with a reported user base of over 20 million unique user accounts worldwide at its

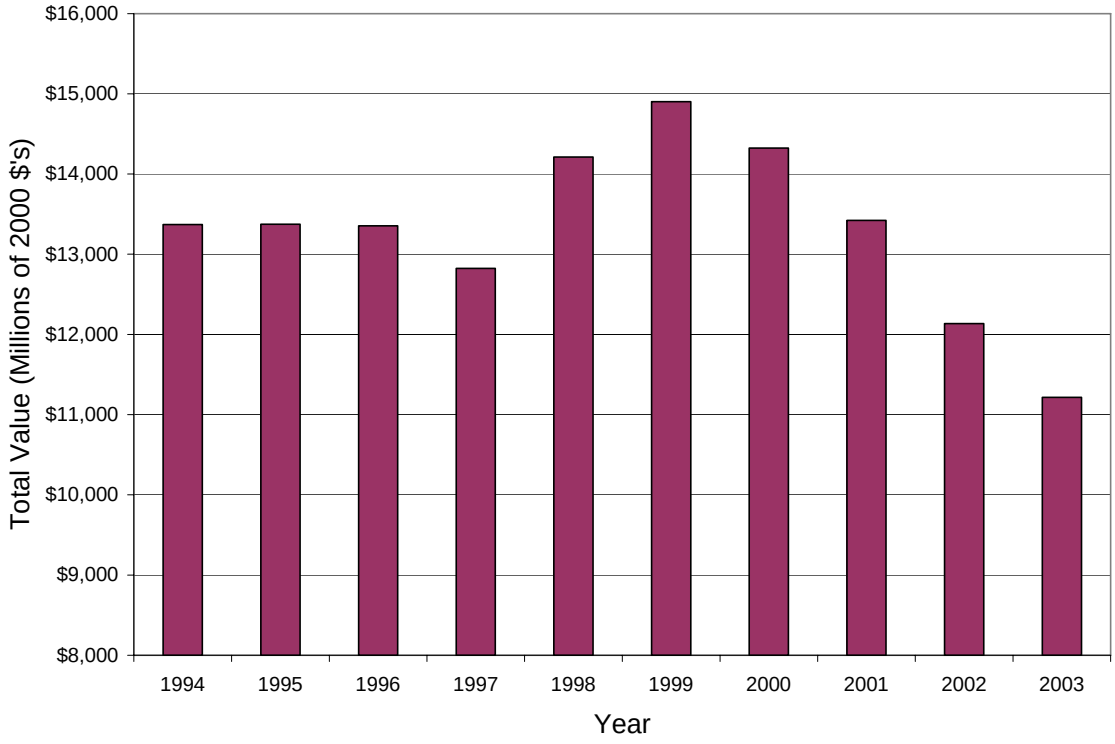


Figure 1: Total Real Value of Music Shipments by Year (RIAA 2003)

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. These lawsuits provide the basis for the instrumental variable approached used to identify the relationship between sales and file sharing. 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.[2]

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.[3] 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, Neilsen 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

[2]Since that time, the RIAA has continued to file lawsuits against heavy users of file sharing networks.

[3]Mortimer and Sorensen (2004) study the relationship between digitial distribution, both legal and illegal, and concert sales and pricing.

cross-country differences in broadband internet access[4] 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%.[5]

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

[4]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.

[5]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.

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

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

2.2 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 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[6]. 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

[6]Microsoft has recently announced plans for its own online MP3 distribution service.

4 million songs in the month after its launch[7]. 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 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. The availability of a copy, then, might allow potential customers to remove some of the uncertainty involved in purchasing an original by testing a copy before purchasing it. Thus, consumers will be more likely to buy goods that they learn they like better.[8] 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

[7] In the average sample week, approximately 11 million full albums were sold.

[8] And, of course, they will be less likely to buy originals that they like less.

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 a good leads to increased willingnesses to pay for other consumers. 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 copying. Blackburn (2006a) 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 increases 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.[9] Copying, then, has the ability to increase the share of potential consumers that are aware of a particular good. 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,[10] 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 users increases the share of

[9] In fact, it is surely false, as I myself am not aware of all the goods that I might purchase or download.

[10] Both of these first two routes for learning about a good are really just variants of the exposure effect.

the consumers who are aware of the good, thus raising the valuation of the average consumer.[11] Recent work by Hendricks and Sorensen (2006) has identified the importance of similar information spillovers when looking at the relationship between sales of current and past albums.

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 fraction of consumers aware of the good. Denote the quantity of units sold by a producer to be $Q(p(q_C), q_C, \theta(q_C))$, where p represents the price of the good, q_C represents the quantity of copies of the good, and θ represents the fraction of all consumers that are aware of the existence of the good. We are interested in the effect of changes in q_C on Q:

$$\underset{(?)}{\frac{dQ}{dq_C}} = \underset{(-)}{\frac{\partial Q}{\partial q_C}} + \underset{(+)}{\frac{\partial Q}{\partial \theta}\frac{\partial \theta}{\partial q_C}} + \underset{(?)}{\frac{\partial Q}{\partial p}\frac{\partial p}{\partial q_C}} \tag{1}$$

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

The first term, $\frac{\partial Q}{\partial q_C}$, is the direct substitution effect discussed above and is clearly negative. The second term above is $\frac{\partial Q}{\partial \theta}\frac{\partial \theta}{\partial q_C}$, which is the awareness effect. This effect is clearly positive, as discussed above, as the availability of additional copies should increase the fraction of the world aware of the good ($\frac{\partial \theta}{\partial q_C} \geq 0$) 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 the quantity of copies on sales as predicted by theory is ambiguous.

This final term, $\frac{\partial Q}{\partial p}\frac{\partial p}{\partial q_C}$, is a potential pricing effect, which is of ambiguous sign, as

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

$\frac{\partial Q}{\partial p}$ is likely negative, and the sign of the pricing response is unclear.[12] However, working within the short-run constraints of this paper, I assume that there is no price response from the music industry. That is, $\frac{\partial p}{\partial q_C} = 0$. 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 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 $\frac{dQ}{dq_C}$? However, theory is not yet out of ideas. There is reason to believe that the marginal effect of copying on sales will differ based on the ex ante popularity of the good. 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 good affect the relative magnitudes of the direct effect and the awareness effect? The negative-signed direct effect, $\frac{\partial Q}{\partial q_C}$, is surely decreasing (becoming stronger) in ex ante popularity. Put another way, the lost sales due to copying are surely greater if potential sales are greater. Similarly, the awareness effect is also decreasing

[12] Producers may wish to lower prices in order to recapture some of the demand lost to copies. On the other hand, they may want to raise prices if the demand that is lost to copies 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.

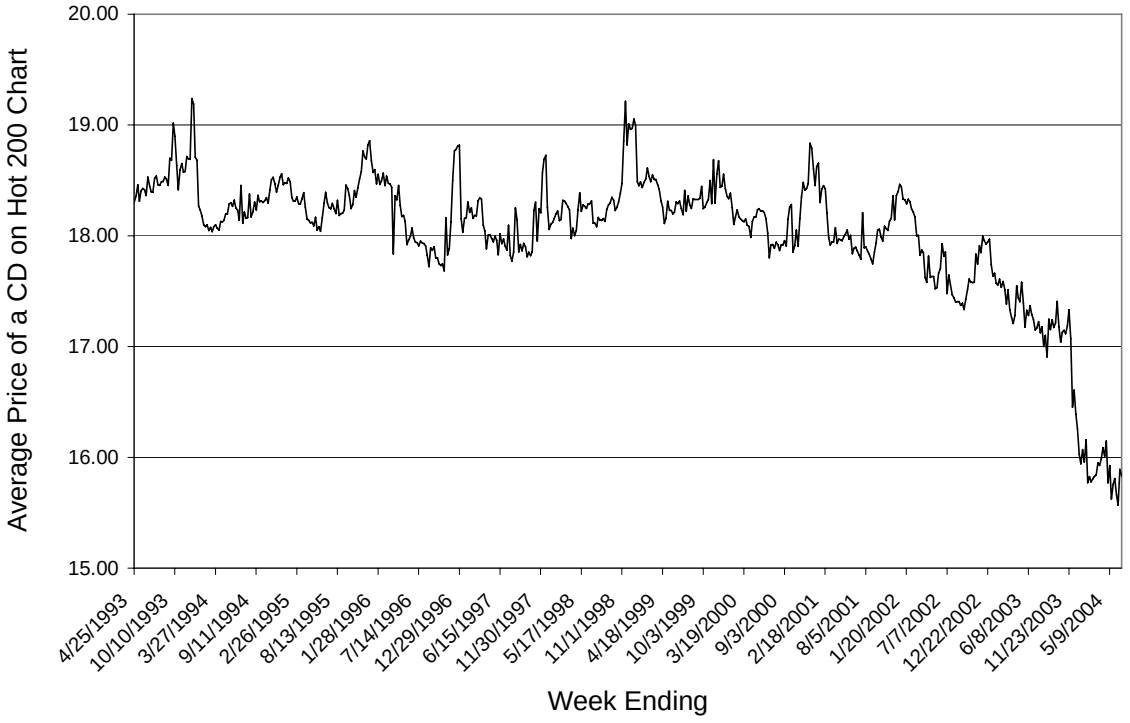


Figure 2: Average Price of a CD on Hot 200 Sales Chart by Week, 1993-2004

(becoming weaker) in the popularity of the good. This is intuitive– if the benefits of file sharing are essentially introducing new consumers to a good, the effect will be necessarily smaller if consumers are already aware of it.

Thus, it is clear that copying (file sharing) should have differential impacts on goods (artists) that are well-known to consumers ex ante versus goods (artists) which are relatively unknown ex ante. 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,[13] 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.

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 Neilsen SoundScan, which tracks retail sales of music and music video products throughout the United States. Neilsen 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.[14] 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.[15] This

[13]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.

[14]Throughout the timeframe of my data sample, these networks are the FastTrack network (Kazaa), Grokster, eDonkey, iMesh, and Overnet.

[15]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 maintence.

data is then reported weekly.[16]

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.

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[17] 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.[18] 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

[16] 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.

[17] A single artist is either a solo artist, such as Celine Dion, or a musical group, such as the Foo Fighters.

[18] See the Data Appendix for the complete details on how the sample was built.

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, 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 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[19] 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,[20] 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,

[19]The Billboard Hot 200 chart is released weekly and reports the ordinal ranking of albums at the national level.

[20]That is, back to September 1992.

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.

4.4 Measuring File Sharing

The primary variable used to measure the amount of file sharing activity for an album 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.

If data on the precise number of downloads in a week were available, I could estimate a model as follows:

$$q_{i,t}^{S} = \theta_0 + \theta_1 D_{i,t}^{FS} + \theta_2 W_{i,t} + \mu_{i,t} \tag{2}$$

where $q_{i,t}^{S}$ is the quantity of album i sold in week t, $D_{i,t}^{FS}$ is the number of downloads of album i in week t, $W_{i,t}$ is a vector of album and week characteristics for album i in week t, and $\mu_{i,t}$ is an error term. Due to the lack of data on $D_{i,t}^{FS}$, however, the model must be extended to include a determination of the unobserved number of downloads, $D_{i,t}^{FS}$:

$$D_{i,t}^{FS} = \pi_0 + \pi_1 C_{i,t}^{FS} + \pi_2 Y_{i,t} + \eta_{i,t} \tag{3}$$

where $C_{i,t}^{FS}$ is the (non-monetary) cost of acquiring a copy of album i in week t on file sharing networks, $Y_{i,t}$ is a vector of album and week characteristics for album i in week t, and $\eta_{i,t}$ is an error term. The cost of acquiring a copy of the album, as described in the preceeding paragraph, is a function of the availability of the album on file sharing networks:

$$C_{i,t}^{FS} = \delta_0 + \delta_1 q_{i,t}^{FS} + \delta_2 Z_{i,t} + \kappa_{i,t} \tag{4}$$

where $q_{i,t}^{FS}$ is the measure of the availability of album i in week t, $Z_{i,t}$ is a vector of album and week characteristics for album i in week t, and $\kappa_{i,t}$ is an error term.

Combining these equations yields the following basic estimation equation:

$$q_{i,t}^{S} = \alpha + \beta q_{i,t}^{FS} + \rho X_{i,t} + \varepsilon_{i,t} \tag{5}$$

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. In particular, $\alpha = \theta_0 + \pi_0 + \delta_0$, $\beta = \theta_1\pi_1\delta_1$, $\rho X_{i,t} = \theta_2 W_{i,t} + \theta_1\pi_2 Y_{i,t} + \theta_1\pi_1\delta_2 Z_{i,t}$, and $\varepsilon_{i,t} = \mu_{i,t} + \theta_1\eta_{i,t} + \theta_1\pi_1\kappa_{i,t}$. Given appropriate assumptions on the independence of the error terms, this equation can be estimated with standard techniques, and the interpretation of the estimated parameter ρ is as the marginal effect of album availability within file sharing networks on sales of the album. This marginal effect works through changes in the cost of acquiring a song through file sharing affecting the number of albums downloaded.

The particular variable used to measure song availability is constructed as follows. For each album, I construct a variable which takes the value of the number of copies of the most popular song from the album available on the file sharing networks that week. To illustrate, consider 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. Throughout the analysis, this is the variable used to measure file sharing activity.[21]

[21]This construction takes the stance on the substitutability of file downloads for album purchases that consumers equate an album to the most popular song on that album. However, to address concerns about this particular measurement, I created several other variables, described in the Data Appendix, that are used to verify that using the results are not driven by use of this measure. 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 Omitted Variable Bias and OLS Estimation

Simple OLS estimation of equation (5) is likely to result in estimates of β, the relationship between sales and file sharing activity, which are biased upwards due to omitted variable bias. The omitted variable in this case is the "quality" of the album in question. In particular, this causes an upward bias due to the fact that albums which are popular to buy are also likely to be popular to download. Thus, without controls for the "quality" of an album or instrumental variables to break this bias, both sales and file sharing activity are likely to be high for "good" albums and low for "bad" albums.

To highlight this problem, I begin by specifying and estimating the simple upwardly-biased OLS regressions. 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 in the literature will allow easy comparison with the work of others.

Table 2 presents the results of estimating equation (5), where both weekly album sales and the number of copies of the most popular song available in the week are expressed in logs.[22] Thus, the estimated coefficient $\widehat{\beta}$ is the estimated file sharing elasticity of sales. In the first column of Table 2, the equation is estimated without and album-week controls

[22] A Box-Cox transform was run to determine whether the relationship between sales and files shared is better specified in levels or in logs, following Godfrey and Wickens (1981). 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 to determine which functional form is appropriate. 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). Although both specifications are rejected in favor of an unconstrained Box-Cox specification, the log-log specification is used for its interpretability, and results are unchanged qualitatively when using levels or the uncontrained Box-Cox specification.

Table 2: OLS Estimation Results

	(1)	(2)
Log of Max # of Files Available	0.309	0.251
	[0.044]***	[0.057]***
Debut Week		0.923
		[0.057]***
# of Weeks Since Release		-0.116
		[0.006]***
# of Weeks Since Release (Squared)		0.001
		[0.0001]***
Christmas Week		0.635
		[0.065]***
Within 2 Weeks Before Christmas		0.472
		[0.073]***
Week After Christmas		-0.007
		[0.054]
Artist Appeared on TV During Week		0.130
		[0.061]**
Artist Appeared on TV During Previous Week		0.188
		[0.055]***
Album has #1 Song on Airplay Chart		0.663
		[0.192]***
Album has #2-#10 Song on Airplay Chart		0.595
		[0.186]***
Album has #11-#40 Song on Airplay Chart		0.480
		[0.107]***
Album has Song on Airplay Chart Below #40		0.415
		[0.100]***
Album Nominated for Grammy (2003)		0.055
		[0.277]
Album Wins Grammy (2003)		0.223
		[0.367]
Album Fixed Effects		YES
Constant	4.863	7.585
	[0.430]***	[0.289]***
Observations	7938	7938
R-squared	0.15	0.93

1. Dependent Variable is Log of Weekly Album Sales
2. Robust standard errors in brackets
3. * significant at 10%; ** significant at 5%; *** significant at 1%

$X_{i,t}$, and the resulting coefficient of 0.309 (indicating that a 10% increase in the # of files available online would increase album sales by 3%) demonstrates the importance of controlling for the "quality" of the album. By introducing a set of album-week controls, $X_{i,t}$, which includes controls for television appearances, airplay success, Grammy award nominations and wins, and time and holiday dummies, the effect of this omitted variable is reduced. Additionally, to the extent that the "quality" of an album is 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 including album-level fixed effects in $X_{i,t}$. The estimation results in the second column of Table 2 are the result of estimating equation (5) again, including this full set of controls. The estimated coefficient of 0.251 indicates that the omitted variable bias has been ameliorated to some extent, although the omitted quality variable is likely still causing the estimation of an artificially positive elasticity.

Note that all other variables have the expected sign when the full set of controls are used. Album sales are highest in the debut week, and then decay quadraticly after. Sales are higher during the Christmas shopping period, but return back to normal levels post-Christmas, and television appearances by the artist boost sales in the immediate future. As espected, albums with hit songs sell more, and although the differences in the point estimates are not all statistically significant, having a song higher on the airplay charts corresponds to higher album sales. The effects of being nominated for a Grammy and winning a Grammy have the expected sign, but the coefficients are not statistically significant from zero. As the signs and magnitudes of these coefficients change very little in other specifications, I will supress the further estimation of the coefficients for these variables in the results presented throughout the rest of the paper.

5.2 Instrumental Variables

In order to address the omitted variable bias that still exists in these estimates, I proceed with a two stage least squares approach, which exploits the timing of the RIAA lawsuits against consumers as instruments to identify the effect of file sharing activity on sales. As discussed in Section 2, the RIAA announced a plan to "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" on June 25, 2003 (RIAA 2003). The timing of this event provides the first instrument; as demonstrated below,

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. On September 8, 2003, the RIAA followed up on its plan and filed the first wave of lawsuits against file sharing users. A press release from the RIAA stated that their focus was on “major offenders who have been illegally distributing substantial amounts (averaging more than 1,000 copyrighted music files each)” (RIAA 2003). The timing of this event provides the second intrument used in estimation; in this case, the announcement that only heavy users were at risk resulted in the level of file sharing activity picking up again, as much of the reduction in activity resulting from initial announcemnt was countered by an increase in response to the implementation of the plan. This is likely due to the fact that the initial announcement did not specify a focus on only "major offenders," while the press release issued with the initial lawsuit filings indicated that only those with large numbers of shared files were at risk.[23]

The instruments are constructed as "jump" variables, which are equal to zero for all weeks prior to the event and equal to one for all weeks on or after the event. This creates a "Lawsuit Plan" variable, which takes the value of one starting with the week of June 25, 2003 and a "Lawsuit Implementation" variable, which takes the value of one starting with the week of September 8, 2003. There are thus 33 weeks for which "Lawsuit Plan" is equal to one and 22 weeks for which "Lawsuit Implemenation" is equal to one (and therefore 11 weeks between the announcement of the plan and the implemenation, for which "Lawsuit Plan" is equal to one while "Lawsuit Implementation" is equal to zero). The use of "jump" variables then allows for shifts in the mean level of file sharing activity resulting from the timing of the RIAA announcements. Further, when accounting for the effects of ex

[23]Gary Chamberlain has pointed out a more direct route for this change in file sharing activity: 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, consumers 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.

ante artist popularity on the relationship between file sharing and sales, the instruments are intereacted with ex ante popularity, thus allowing the shift in the mean level of file sharing activity to differ systematically by artist popularity.

Of course, it is necessary that the timing of the lawsuits is exogenous to the dependent variables in the primary regression in order for the instruments to be valid. And because 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, this is a potential concern. 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. While consumers may have been aware that lawsuits from the RIAA were a possibility, there is no evidence that anyone anticipating the exact timing of the announcement. Further, 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, even conditional on the initial announcement.

5.3 TSLS results with Homogenous Copying Effects

I begin by again estimating equation (5), but now I apply a two-staged least squares (TSLS) technique in order to exploit the exogenous changes in file sharing activity generated by the timing of the lawsuit announcement and implementation, but still without accounting for the impacts of heteregenous effects of copying on sales. The first column of Table 3 presents the results of the first stage regression. Only the coefficients for the excluded instruments (the timing of the lawsuit announcement and implementation) are presented, but

Table 3: TSLS Estimation with Homogenous Effects of Copying

	(1) First Stage	(2) Second Stage
Dependant Variable:	Log of Max # of Files Available	Log of Weekly Album Sales
Lawsuit Announcement Dummy	-0.466	
	[0.056]***	
Lawsuit Implemenation Dummy	0.116	
	[0.057]**	
Log of Max # of Files Available		-0.073
		[0.082]
Observations	7938	7938
R-squared	0.93	0.92

1. F-test statistic for excluded instruments in first stage regression is 64.98, with a p-value less than .0001
2. Robust standard errors in brackets
3. * significant at 10%; ** significant at 5%; *** significant at 1%

all other coefficients have the expected sign.[24] The effect of both the lawsuit announcement and the lawsuit implementation on the amount of file sharing activity are large and statistically signifcant.[25] In particular, the coefficent of -0.466 on the timing of the lawsuit announcement indicates that after the announcement of the RIAA's lawsuit strategy, the mean level of file sharing activity activity dropped by approximately 37%, as users were presumably scared away from file sharing due to the threat of being sued by the RIAA. However, upon the implementation of the lawsuits, file sharing activity regained a fraction of its loss from the announcement; the coefficient of 0.116 indicates that file sharing activity rose by 12% after the first lawsuits were filed, regaining slightly less than one-quarter of the loss from the initial announcement, on average.

[24]One 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 suggests 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. It has been suggested this result may be due to universities not being in session right before Christmas, but file sharing levels rebound in the week after Christmas when schools are still out of session, so that is unlikely to be the case.

[25]The F-statistic of nearly 65 indicates that no weak instrument concerns are warranted.

The second column of Table 3 presents the second-stage results of estimate of β, the file sharing elasticity of sales. The two-stage least squares estimation suggests that, without accounting for the possibility of heterogenous effects of copying on sales, file sharing has had essentially zero effect on sales. Although the coefficient is statistically insignificant from zero, if it is accepted as accurate, it suggests that eliminating 10% of files shared would increase album sales by only 0.7%. For the median artist-week in the data sample, whose weekly sales are 2,852 albums, this would increase album sales by only 20 albums per week.

This result is also entirely consistent with the results obtained by Oberholzer and Strumpf (2004), whose preferred estimate of the effect of file sharing on sales is a small, statistically insignificant effect, albeit a positive one. However, it is important to note the pitfalls associated with assuming a homogenous effect of copying on sales across producers (artists). As discussed in Section 3, it is extremely likely that the effects of copying on sales is heterogenous across artists, and thus assuming a constant effect of file sharing on sales forces this constant estimate to match the average effect. This would not be a problem (if the only concern was on the aggregate effect and not on any distributional issues), except for the fact that by using an album-week as the unit of observation (as is done both in this section and in Oberholzer and Strumpf (2004)), this average effect is weighted not by relative sales, but by the proportion of album-week appearances in the sample.[26]

A simple example highlights the issue. If there are two data points, one with a positive elasticity and the other with an equally sized, but negative elasticity, then assuming a constant effect yields an estimate of zero elasticity. If both of these producers have equal sales in the absence of copying, then this average effect is the aggregate effect. An increase in copying reduces the sales of one producer by the same amount it increase the sales of the other, resulting in a zero elasticity in the aggregate. However, if the producer with the neg-

[26] Liebowitz (2004) explains this point in finer detail.

ative elasticity has, say, double the sales of the producer with the positive elasticity, then the aggregate elasticity is not zero as the sales of producer with the negative elasticity will fall by twice as much (e.g. 100 units) as the sales of the producer with positive elasticity will rise (e.g. 50 units).

Thus, the naïve answer obtained above is not a reliable measure of either (a) the distributional effects of file sharing on the sales of albums or (b) the aggregate effect of file sharing on sales. In order to address this issue, then, I proceed by allowing the the marginal effect of file sharing on sales to vary across artists in a systematic way (to address the first concern) and by weighting these marginal effects by artist sales when constructing counterfactuals designed to estimate the aggregate effect of file sharing on sales (to address the second concern).

5.4 Time Trends

Additionally, there is one further issue with the estimation procedure which must be addressed. Unlike the sales variable, which is measured as weekly sales (a flow variable), file sharing activity is measured as the number of files available on file sharing networks during the week. This variable is a stock, rather than a flow, and is thus lower during the debut week(s) of an album because not enough time has passed for the peak level of the stock to be reached. As the typical pattern of sales is for sales to peak in the debut week and then fall from there, this 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 are decreasing, which would lead a finding of a negative relationship not due to underlying causality, but due to the time-path of each of these variables.

In order to control for this problem, I have included a flexible time trend to enter into

both stages of the TSLS 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. This should eliminate any artificial relationship due to the overall time-path of these two variables. Excluding this time trend results, as expected, in a much larger estimate, -4.19, of the elasticity between sales and file sharing activity. Further, performing the same estimation without a trend but excluding the first five weeks of an album's life makes the estimates too positive (1.11), as after the first few weeks of an album's life we have the opposite problem of both the number of files shared and sales decaying over time.

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. Including the life cycle trend corrects this problem.

5.5 Heterogenous Effects of Copying

I now 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.[27] This is done by creating a "continuous" definition of ex ante popularity as described in the Section 4.3, 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:

$$q_{i,t}^{S} = \alpha + \beta q_{i,t}^{FS} + \varphi P_i * q_{i,t}^{FS} + \rho X_{i,t} + \gamma_i + \varepsilon_{i,t} \tag{6}$$

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

Table 4: TSLS Results, Effects Differentiated by Artist Popularity

	(1) First Stage	(2) Second Stage
Dependant Variable:	Log of Max # of Files Available	Log of Weekly Album Sales
Lawsuit Announcement Dummy	-0.358	
	[0.077]***	
Lawsuit Announcement Dummy * Popularity	-0.001	
	[0.0006]**	
Lawsuit Implemenation Dummy	0.208	
	[0.087]**	
Lawsuit Implemenation Dummy * Popularity	-0.001	
	[0.0005]**	
Log of Max # of Files Available		0.445
		[0.211]**
Log of Max # of Files Available * Popularity		-0.005
		[0.002]**
F-Test on Excluded Instruments	36.29	
p-value	0.000	
Observations	7938	7938
R-squared	0.93	0.92

1. Robust standard errors in brackets
2. * significant at 10%; ** significant at 5%; *** significant at 1%
3. Supressed Controls include Debut Week, # of Weeks Since Release, # of Weeks Since Release (Squared), Christmas Week, Within 2 Weeks Before Christmas, Week After Christmas, Artist Appeared on TV During Week, Artist Appeared on TV During Previous Week, Album has #1 Song on Airplay Chart, Album has #2-#10 Song on Airplay Chart, Album has #11-#40 Song on Airplay Chart, Album has Song on Airplay Chart Below #40, Album Nominated for Grammy (2003), Album Wins Grammy (2003), and Album Fixed Effects

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 $\widehat{\delta}$ is expected to be negative. This relationship is then estimated by using timing of the RIAA's announcement and implemenation of their lawsuits as instruments in a TSLS set-up. Table 4 presents the results of this analysis.

Again, the effect of both the lawsuit announcement and the lawsuit implementation on the amount of file sharing activity are large and statistically signifcant. The coefficients in column (1) indicate that (for an artist of popularity zero, a "new" artist), files available online dropped by 30% after the lawsuit announcement, and regained nearly two-thirds of that

drop after the first lawsuits were filed. Furthermore, the effect of the lawsuit announcement was even stronger for more popular artists, to the point that the number of files available online for an artist with maximum popularity (200, or an artist would had a previous #1 album) dropped by over 42% after the announcement was made, and the implementation of the lawsuits had little, if any impact on the number of files available for these superstar artists.[28] Of course, for artists with intermediate levels of popularity, both the initial drop in files shared, and the rebound from the implementation are at intermediate levels.[29]

Now, the baseline elasticity of sales with respect to file sharing for an artist of zero popularity (a new artist) is 0.445, which is strong, although it has a 95% confidence interval lower bound of 0.03, 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 nearly 40%. 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 -0.508 with a standard error of 0.22, 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.[30] The marginal effect is significantly posi-

[28]This is presumably a result of consumers understanding that the RIAA was only focused on those "sharers" who were "major offenders" and intrepreting that to also mean that sharing songs from higher-selling artists would be more dangerous.

[29]As both the "Log of Max # of Files Available" and its interaction with popularity are endegenous variables, each one is instrumented for by the lawsuit announcement and implementation dummies. The first-stage results for the interaction between files available and popularity is suppressed because it is difficult to interpret and adds little to the discussion. However, the coefficeints have the same sign, and the F-test on the excluded instruments has a p-value below 0.0001.

[30]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 quantatative 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.

tive at the 10% level for popularities less than or equal to 50 (artists whose most successful previous album reached no higher than 151 on the Hot 200 chart) and significantly negative at the 10% level for popularities greater than or equal to 129 (artists whose most successful previous album reached at least 72 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 6 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 album-weeks in the dataset 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 100 albums per week were files

shared to be reduced by 10%. A similar calculation can be made for an artist of maximum popularity, whose weekly sales are 29,767. At the median level of sales for these artist, the estimate implies an increase in sales of 1,636 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.[31]

6 Implications in the Short Run

With the estimated effects of file sharing in hand, I now proceed to run 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. This is done by simply subtracting out the effect of the lawsuits from the first stage estimation done presented in Column 1 of Table 4 and then using the estimated effects of file sharing on sales from Col-

[31] 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* $\beta + \varphi\overline{P}$, 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 $(\beta + \varphi\overline{P})(\Delta\overline{q_i^{FS}})$.

umn 2 of Table 4 to get the change by album. Then, these individual effects are aggregated up to the market level using the weights described in the Appendix.

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 4.0% 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 11,029,473 per week in the absence of the reduction in file sharing caused by the lawsuit strategy. Using a rule of thumb of $5 of profit per compact disc,[32] this translates to an increase in industry profits of $2,205,895 per week, or $50.7 million over the 23 week period after the lawsuit strategy was announced to the public.

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 2 of Table 4, 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 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;[33] that is, I perform what is essentially an

[32]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.

[33]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 lawsuit announcement was to reduce file sharing by around 40% on average. Thus, the data should have no problem speaking to the effects of reducing file sharing by at least

Table 5: The Effect of Removing File Sharing on Industry Sales

% of File Sharing Activity Removed	% Increase in Industry Sales	File Sharing Elasticity of Sales
10%	2.0%	-0.20
20%	4.6%	-0.23
30%	7.9%	-0.26
40%	12.2%	-0.30
50%	17.9%	-0.36

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 5.

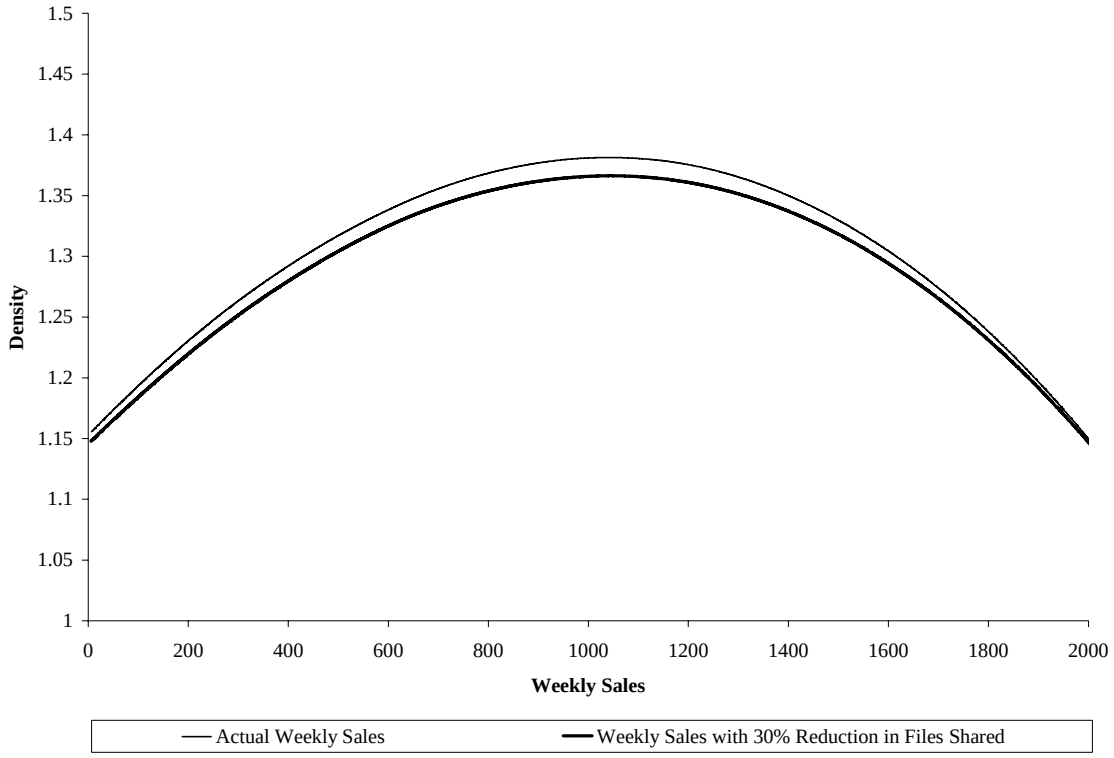
Focusing on the estimated elasticities, they suggest that removing file sharing would increase industry wide recorded music sales by anywhere from 20% to 36% relative to sales over the sample period.[34] As a point of perspective, recall from Figure 1, that real recorded music sales were approximately 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%, aggregate sales would increase by approximately 7.9%. According to Billboard magazine (Market Watch 2004), in 2003 industry-wide sales totaled approximately 660 million during the calender year. Thus, an increase in sales of 7.9% would amount to 52.1 million albums during the year. Returning to the estimate of $5 of variable profit per sale, this equates to approximately $261 million of additional profit in the calender year 2003.[35]

this proportion. However, going much beyond this level is unlikely to have much validity.

[34] Of course, extrapolating these elasticities that far is an out-of-sample exercise.

[35] A previous version of the paper presented the results of the counterfactual exercises considered in this

Figure 3: The Distribution of Actual Sales and Sales with 30% Fewer Files Shared, Part 1

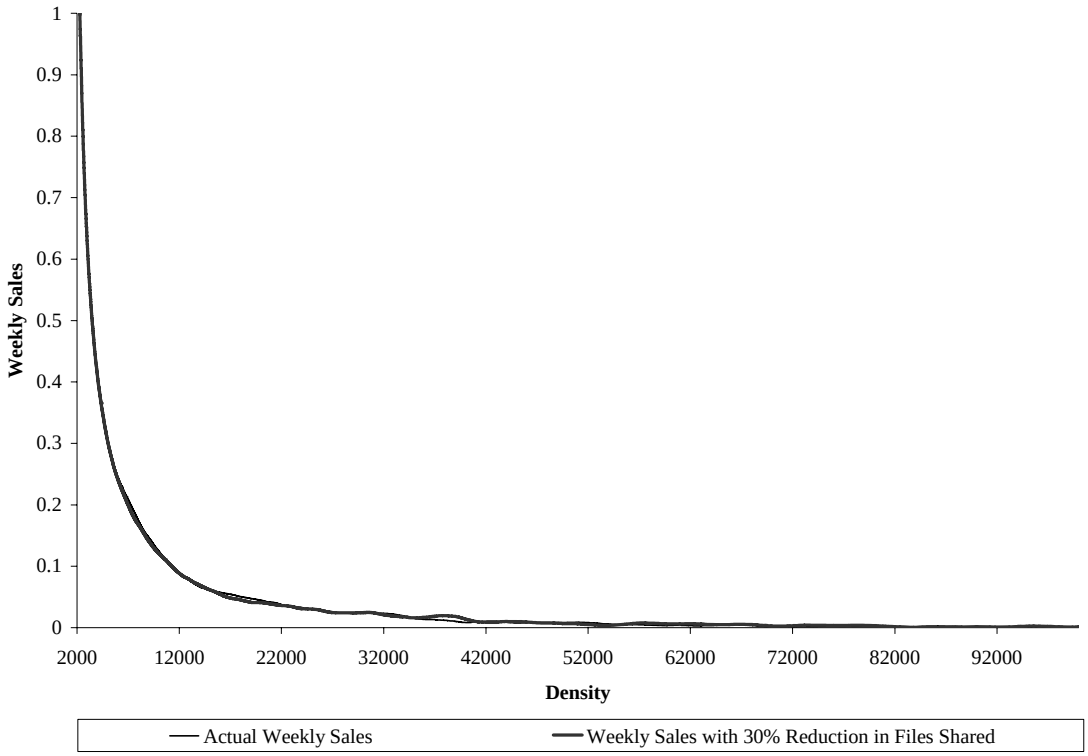


However, as highlighted in the previous discussions, this change has not been uniform across all artists. Rather, very different effects of file sharing on sales have been found for ex ante unknown artists relative to ex ante popular ones. Therefore, it is possible to investigate the effect that file 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.

I estimate the distribution of actual sales and sales in a counter-factual world with

section based on estimation of a multinomial logit demand system. The results presented here are very similar and thus the logit model estimation results have been supressed. They are available from the author upon requset.

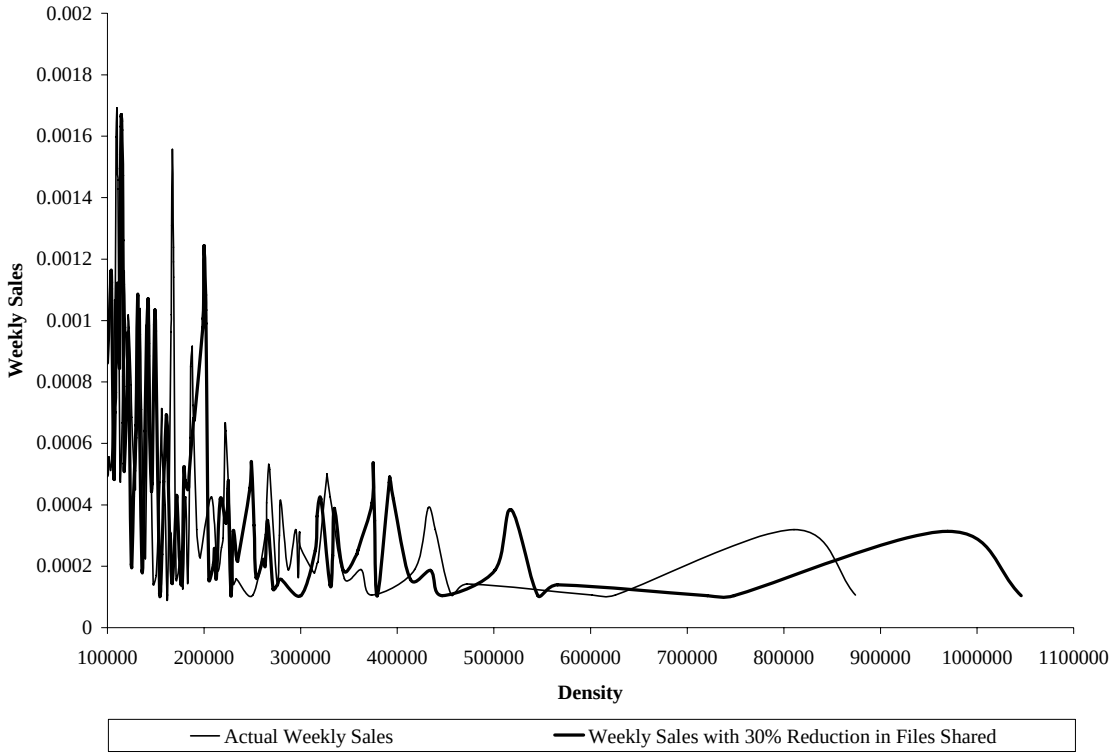
Figure 4: The Distribution of Actual Sales and Sales with 30% Fewer Files Shared, Part 2



30% less file sharing activity using an Epanechnikov kernel function, and "optimal" bandwidth.[36] 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. Therefore, I presente three graphs, each of which "zooms in" on a different segment of the distribution. Figure 3 focuses on the low end of the sales distribution. Here, it is clear that in a world with reduced levels of file sharing, the sales totals for artists at the low end of the distribution have been shrunk. In fact, it is not until weekly sales reach a total of more than 2,000 albums per week (approximately the 50% percentile of the distribution of weekly sales) that the two distributions match. The next figure, Figure 4, focuses on the distribution of the weekly

[36]The density is evaluated at all of the sales figures in the dataset. Also note that the density is not scaled so that the area below the curve equals one.

Figure 5: The Distribution of Actual Sales and Sales with 30% Fewer Files Shared, Part 3



sales between 2,000 and 100,000 albums per week. There is little difference in this part of the distribution between the actual distribution of weekly sales and that of one with reduced file sharing levels, likely due to the fact that the middle part of the distribution is sales distribution is made of artists for whom the estimated elasticity is close to zero. Finally, Figure 5 presents the top end of the sales distribution. The estimates here become "lumpy" due to the lack of data points in this area, but it is clear the distribution of sales when file sharing is reduced has been shifted rightwards, leading to large increases in sales for these artists.

Artists who are ex ante unknown, and thus most helped by file sharing, are those artists who sell relatively few albums, 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.[37]

7 Conclusions

This paper has investigated the short-run effects of copying on sales. In particular, I have exploited the timing of the music industry's strategy of suing file sharing network users to estimate these effects within the recorded music industry in the United States. Naïve estimates which do not allow for the effect of file sharing to differ systematically across artists yields results similar to those found elsewhere in the literature, suggesting that file sharing has not had a significant effect on the sales of recorded music.

[37] 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. Blackburn (2006b) takes a further look into this issues, and finds little evidence that labels have changed the way they develop artists.

Further inspection, however, reveals that it is unrealistic to believe that the effects of copying would be constant across all producers as the costs and benefits of copying differ with the ex ante popularity of the producer. This suggests 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.

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. In particular, I have found that the lawsuit strategy pursued by the RIAA increased industry profits by approximately $50.7 million over the 23 week period after the lawsuit strategy was announced to the public. A 30% reduction in the number of files available on file sharing networks in 2004 would have increased industry profits by $261 million.

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 question 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 Magainze* (2000), “Is Biz Poised For Renewed Price Wars?,” January 8, 2000.

[6] *Billboard Magainze* (2003), “Market Watch,” various issues.

[7] *Billboard Magainze* (2004), “Market Watch,” January 10, 2004.

[8] Blackburn, D. (2006a), “Network Effects and Copyright Enforcement," mimeo, May 2006.

[9] Blackburn, D. (2006b), “Developing Superstars,” mimeo, January 2006.

[10] Christman, E. (2003), “UMVD Expands Market-Share Dominance,” January 18, 2003, *Billboard Magazine*.

[11] Christman, E. (2004), “Ed Christman, UMG tops album share for fifth year,” January 17, 2004, *Billboard Magazine*.

[12] CNNMoney (July 19, 2000), "Napster: 20 million users," http://money.cnn.com/2000/07/19/technology/napster.

[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] 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.

[16] Grammy Awards (2004), website, htttp://www.grammy.com.

[17] Hendricks, K. and A. Sorensen. (2006), "Information Spillovers in the Market for Recorded Music," NBER Working Paper No. 12263.

[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 Entrepeneurship, Innovation, and Economic Growth*, edited by Gary Libecap, JAI Press.

[20] Liebowitz, S. (2003), "Pitfalls in Measuring the Impacts of File Sharing," mimeo.

[21] Milette, R. (2004), website, http://rym.waglo.com/wordpress/

[22] Mortimer, J.H. and A. Sorensen. (2004), "Digital Distribution and Demand Complementarities: Evidence from Recorded Music and Live Performances," mimeo.

[23] Nielsen SoundScan (2004), website, http://www.soundscan.com

[24] Oberholzer, F and K. Strumpf (2004), "The Effect of File Sharing on Record Sales: An Empirical Analysis." HBS mimeo.

[25] Recording Industry Association of America (2003), website, “Year End Marketing Reports,” http://www.riaa.com/news/marketingdata/yearend.asp

[26] Recording Industry Association of America (1999), website, “Legal Cases,” http://www.riaa.com/news/filings/pdf/napster/Napster_Complaint.pdf

[27] 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

[28] 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

[29] Recording Industry Association of America (2004), website, “Consumer Purchasing Trends,” http://www.riaa.com/news/marketingdata/pdf/2003consumerprofile.pdf

[30] 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 No. 10874.

[31] Schwartz, J. (2004), “A Heretical View of File Sharing,” April 5, 2004, P. C1, *The New York Times*.

[32] Standard and Poor's (2002), "Movies And Home Entertainment," *Industry Surveys*, November 2002.

[33] 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.

[34] Takeyama, L. (1997), "The Intertemporal Consequences of Unauthorized Reproduction of Intellectual Property," *Journal of Law and Economics*, October 1997, 40-2, Pp. 511-22.

[35] TVToMe (2004), website, http://www.tvtome.com.

[36] 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 Neilsen 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 Neilsen SoundScan claims that their sampling is correctly representative of the total market.[38] The data collected by Neilsen 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.[39] 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

[38] Although I can not independently verify this, the data collected by Neilsen SoundScan is universally accepted as accurately portraying the recorded music industry, and is essentially the only source for data on retail performance.

[39] 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 Neilsen 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 (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[40] and Grammy award wins for 2003 were constructed using data from www.grammy.com. Lastly, the Billboard charts for radio airplay were used to construct four dummy variables indicating various levels of radio airplay[41]. 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.

[40]The Grammy Awards are the recorded music industry's annual award shows, similar to the Oscars for motion pictures.

[41]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 pratically, 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.[42] 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

[42] In general, recorded music albums are released for sale on Tuesdays.

Table 6: 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 6 and 7 and is the primary source for analysis throughout the paper. Weekly sales data for the albums was then purchased from Neilsen 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.

Because the 197 albums in the sample were not randomly chosen from the potential

Table 7: 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%

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 6 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 Measuress 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 available 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

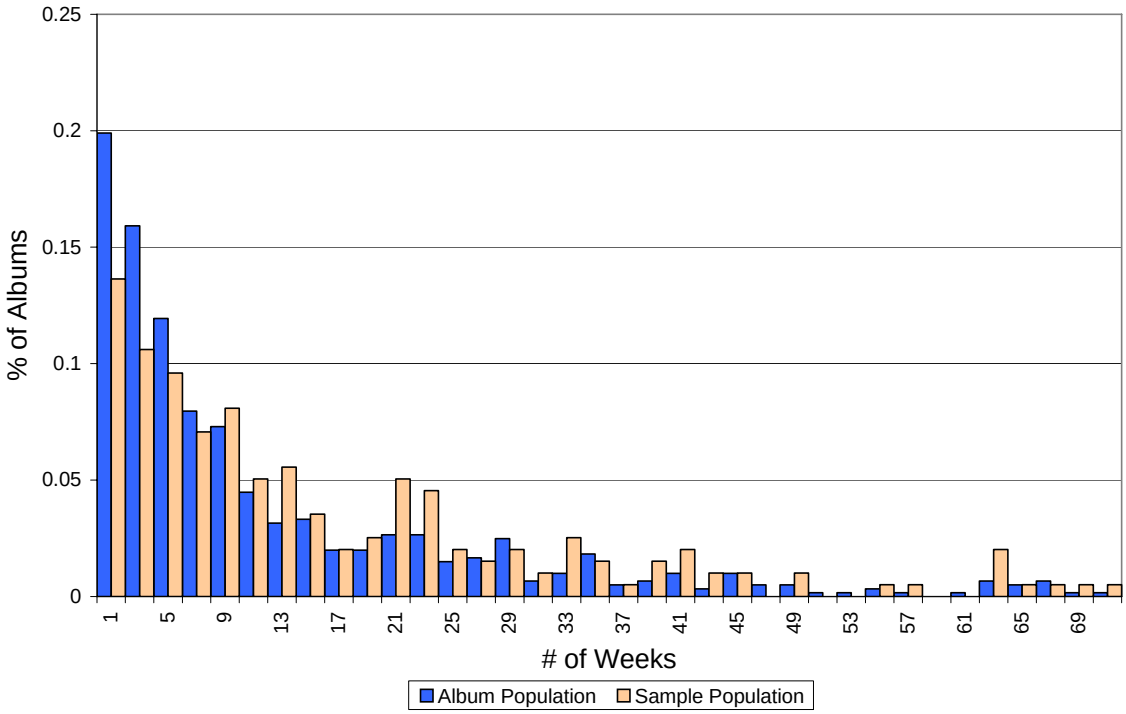


Figure 6: Total Number of Weeks on Hot 200 Sales Chart, for Sample and Population of Albums

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 6 and 7 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 7 graphs the percentage of all album sales in the US that are included in the data sample. This percentage is initally 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.

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,

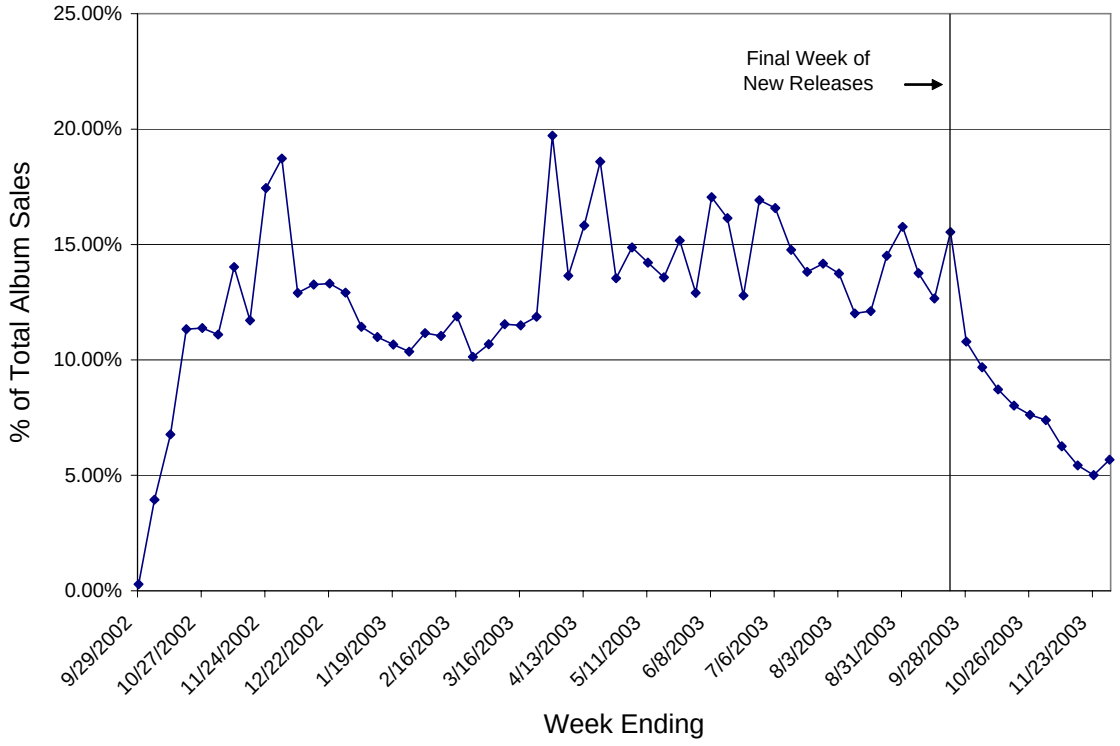


Figure 7: % of Total National Sales Data in Data Sample by Week

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.