USC FBE APPLIED ECONOMICS WORKSHOP
presented by Koleman Strumpf
FRIDAY, March 25, 2005
1:30 pm – 3:00 pm, **Room: HOH-601K**

The Effect of File Sharing on Record Sales
An Empirical Analysis*

Felix Oberholzer-Gee
Harvard Business School
foberholzer@hbs.edu

Koleman Strumpf
UNC Chapel Hill
cigar@unc.edu

December 2004

Abstract

For industries ranging from software to pharmaceuticals and entertainment, there is an intense debate about the level of protection for intellectual property that is necessary to ensure innovation. In the case of digital information goods, web-based technologies provide a natural crucible to assess the implications of reduced protection because these technologies have drastically lowered the cost of copying information. In this paper, we study the impact of file-sharing technologies on the music industry. In particular, we analyze if file sharing has reduced the legal sales of music. While this question is receiving considerable attention in academia, industry and in Congress, we are first to study the phenomenon employing data on actual downloads of music files. We match 0.01% of the world's downloads to U.S. sales data for a large number of albums. To establish causality, we instrument for downloads using data on international school holidays and technical features related to file sharing. Downloads have an effect on sales which is statistically indistinguishable from zero. Moreover, our estimates are of moderate economic significance and are inconsistent with claims that file sharing can explain the decline in music sales during our study period.

*We thank Bharat Anand, Shane Greenstein, Alan Krueger, Tom Mroz, Alan Sorensen, Joel Waldfogel, and Steven Wildman. We also received valuable comments from participants at the 2004 AEA meeting and seminar participants at Cornell University, Duke University, the Federal Communications Commission, the Federal Trade Commission, Harvard Business School, the NBER University Research Conference, SUNY-Buffalo, the University of Arizona, the University of Florida, the University of North Carolina, the University of Michigan, the University of Pennsylvania Law School, and the University of Zurich. This project would not have been possible without the assistance of several individuals and organizations. MixMasterFlame and the FlameNap network shared P2P data with us, and BigChampagne LLC, the CMJ Network, Nathaniel Leibowitz, and Nevil Brownlee generously provided auxiliary data. We thank Keith Ross and David Weekly for assistance in understanding the KaZaA, OpenNap, and WinMX search protocols and database indices. Sarah Woolverton's tireless efforts to improve the quality of our song matching algorithm and Christina Hsiung Chen's research assistance is also appreciated. Oberholzer-Gee gratefully acknowledges the financial support of the George F. Baker Foundation. Aural support from Massive Attack, Sigur Ros and The Mountain Goats is appreciated.

I. Introduction

File sharing is now one of the most common on-line activities. More than 60 million Americans have downloaded music and the number of file sharers continues to grow rapidly (Karagiannis et al., 2004). File sharers use networks of computers to search for and download files from one another. Sharing files is largely non-rivalrous because the original owner retains his copy of a downloaded file. The low cost of sharing and significant network externalities are key reasons for the dramatic growth in the size of the file sharing community. While few participated in file sharing prior to 1999, the founding year of the original Napster, in 2004 there were more than nine million simultaneous users on the major peer-to-peer (P2P) networks. Because physical distance is largely irrelevant in file sharing, individuals from virtually every country in the world participate.

There is great interest in understanding the economic effects of file sharing, in part because the music industry was quick to blame file sharing for a recent decline in sales.[1] Between 2000 and 2003, the number of CD's shipped in the United States fell by 20% to 750 million units (RIAA, 2003a). Claiming that file sharing was the culprit, the recording industry started suing thousands of individuals who share files. The industry also asked the Supreme Court to rule on the legality of file sharing services, a question which critically hinges on the "market harm" caused by the new technology. Congress is currently considering a number of measures designed to counter the perceived threat of file sharing.

While concerns about P2P are widespread, the theoretical effect of file sharing on record sales and industry profits is ambiguous (Bakos et al, 1999; Takeyama, 1997; Varian, 2000). Participants could substitute downloads for legal purchases, thus reducing sales. The inferior sound quality of downloads and the lack of features such as liner notes or cover art perhaps limit such substitution. Alternatively, file sharing allows users to learn about music they would not otherwise be exposed to. In the file sharing community, it is a common practice to browse the files of other users and discuss music in file server chat rooms. This learning may promote new sales. Other mechanisms proposed in the theoretical literature have unclear effects on sales.

[1]Cary Sherman, President of the RIAA, sums up the industry position: "There's no minimizing the impact of illegal file-sharing. It robs songwriters and recording artists of their livelihoods, and it ultimately undermines the future of music itself, not to mention threatening the jobs of tens of thousands" (*USA Today*, 18 September 2003).

Individuals can use file sharing to sample music, which will increase or decrease sales depending on whether users like what they hear (Shapiro and Varian, 1999). The availability of file sharing could change the willingness to pay for music, either decreasing it due to the ever present option of downloading or increasing it through network effects and the greater ease of sharing a purchased album (Takeyama, 1994; Liebowitz, 1985). Finally, it is possible there is no effect on sales. File sharing lowers the price of music, which draws in low-valuation individuals who would otherwise not have purchased albums. In a recent survey of college students, Rob and Waldfogel (2004) find that albums purchased in the store were valued at $15.91. In contrast, respondents' willingness to pay for albums they downloaded was only $10.66, a value below the average purchase price of a CD.

With no clear theoretical prediction, the effect of file sharing on sales is an empirical question. Most of what we know about the effects of file sharing is based on surveys. The evidence is mixed. File sharers generally acknowledge both sales displacement and learning effects, and it is unclear if either effect dominates the other. Rather than relying on surveys, this study is the first to use observations of actual file sharing behavior of a large population to assess the impact of downloads on sales. Our dataset includes 0.01% of the world's downloads from the last third of 2002. We match audio downloads of users in the United States to a representative set of commercially relevant albums for which we have concurrent weekly sales, resulting in a database of over ten thousand album-weeks. This allows us to directly consider the relationship between downloads and sales. To establish causality, we instrument for downloads using international school holidays and technical features related to file sharing such as network congestion or song length, all of which are plausibly exogenous to sales.

We find that file sharing has only had a limited effect on record sales. After instrumenting for downloads, the estimated effect of file sharing on sales is statistically indistinguishable from zero. The economic effect of the point estimates is also small. A typical prediction of our models is that file sharing increased sales by less than 10%. And even our most pessimistic prediction indicates file sharing displaced less than 1% of albums per year for the entire music industry. Drawing on our most precise estimates, we can statistically reject the null that even a quarter of the recent sales decline stems from file sharing. At the same time we can never reject the hypothesis that downloads have no effect on overall sales. These results continue to hold

after accounting for the dynamics of consumer choices, omitting data from the holiday shopping season, allowing the impact of downloads to vary by album popularity, and after controlling for the long-term growth in the number of file sharing users. In total the estimates indicate that the sales decline over 2000-2002 was not primarily due to file sharing. While downloads occur on a vast scale, most users are likely individuals who in the absence of file sharing would not have bought the music they downloaded.

Exploiting temporal and spatial variation in the intensity of file sharing, we also provide quasi-experimental evidence on the effect of P2P on music sales. For instance, we document that the share of sales during the summer months when fewer students have access to high-speed campus Internet connections have not changed as a result of P2P. Similarly, sales did not decline more precipitously in the Eastern Time zone of the United States where P2P users can more conveniently download files provided by the large European file-sharing community.

The experience of recorded music is not unique. Newspaper circulation, for example, does not seem to decline following the creation of a free online edition. The print circulation growth rate for the New York Times was comparable to the growth rate of the Wall Street Journal over 1996 to 2003, although the Times created a free version of its paper in 1996 while the Journal offers a pay site (see also Belden Associates, 2003). An in-house New York Times study has even found their website has a small positive effect on paid circulation (Nisenholtz, 2002).

Our results have broader implications beyond the specific case of file sharing. A longstanding question in economics concerns the level of protection for intellectual property that is necessary to ensure innovation (Posner, 2002). Economic research on the role of patents and copyrights likely began with the critique in Plant (1934) and continues today in the debate between Boldrin and Levine (2003) and Klein, et al. (2002). We provide specific evidence on the impact of a change in property rights for the case of a single industry, recorded music. The file sharing technology available in 2002 had markedly lowered the protection that copyrighted music recordings enjoyed, so it is interesting to analyze if this reduced protection adversely affected sales. For our study period, we fail to detect a significant impact.

The outline of the remainder of the paper is as follows. The next section provides an overview of the empirical literature. Section III describes the mechanics of file sharing. The data are

discussed in Section IV. Next we describe the econometric approach. Section VI presents the results, and the last section discusses the implications of this study.

II. The Literature

Empirical research on file sharing and record sales has been limited and inconclusive, primarily, we believe, due to shortcomings with the data. Most of what we know about the effect of file sharing on sales is based on phone surveys. There are numerous industry studies which arrive at a diverse range of conclusions. For instance, Forrester Research (2002) and Jupiter Media Metrix (2002) find neutral or positive effects, while the International Federation of the Phonographic Industry (2002), Edison Media Research (2003) and Forrester Research (2004) document a sales displacement. A general difficulty with these studies is that they compare the purchases of individuals who download files with the purchases of those who do not. While downloaders may in fact buy fewer records, this could simply reflect a selection effect. File sharing is attractive to those who are time-rich but cash-poor, and these individuals would purchase fewer CDs even in the absence of P2P networks.

A handful of academic studies relies on micro data to address the issue of unobserved heterogeneity among file sharers.[2] Rob and Waldfogel (2004) study the survey responses of a convenience sample of U.S. college students. For hit albums which sold more than 2 million copies since 1999, they find no relationship between downloading and sales. Expanding the set of albums to include all music the students acquired in 2003, downloading five albums displaces the sale of one CD. This difference is interesting. One interpretation is that piracy does not affect hit albums but hurts smaller artists. It is also possible that file sharing had less of an effect on sales in earlier years. After instrumenting for downloads with the school the students attend – everyone at Penn has broadband access while this is not true for the other schools – the resulting estimates are too imprecise to draw any firm conclusions. Zentner (2004) employs European survey data to study the relation between file sharing and sales. Using measures of Internet

[2]There are also studies based on aggregate time series data, which consider correlations in music sales and file sharing activity. These works are compromised by the role of confounding third factors like changes in consumer tastes, the quality of music, the macroeconomy, or the availability of other entertainment products.

sophistication and access to broadband as instruments, Zentner finds some displacement. Unfortunately, neither the Rob and Waldfogel study nor Zentner's work allow inferences about the total impact of file sharing on record sales because neither paper studies a representative sample of file sharers. Zentner also lacks information about the number of titles that individuals download or purchase.

Finally, Blackburn (2004) examines the relationship between the supply of digital files and sales, using the timing of the RIAA lawsuits as an instrument. Unlike broadband access or Internet sophistication, this instrument has the advantage of not being a choice variable. Restricting the estimated effect to be identical across albums, Blackburn (2004) concludes that increases in the supply of files had no impact on sales. In models that allow the supply of files to differentially affect popular and less popular artists, Blackburn's results imply that removing 50% of the files from P2P services would have increased sales by more than 25%. In contrast to Rob and Waldfogel (2004), popular artists are most negatively affected in this study. One concern with this paper is that it is difficult to know how the supply of files is related to downloading. For instance, the sales numbers for new releases often peak in the first few weeks, while the stock of files on P2P networks continues to increase, leading to a mechanical negative relationship between downloads and sales. More generally, the number of shared files reflects both past sales and past downloads. As a result, regressing the supply of files on purchases does not provide an estimate of the impact of P2P even in the presence of album-specific time trends. A second concern involves the identification strategy. Blackburn presumes there was an exogenous reduction in file sharing immediately following the announcement of the RIAA lawsuits. But the announcements mainly discouraged participation by marginal users, who used file sharing to sample music, and had little effect on hardcore users, who were more likely to substitute downloads for purchases (NPD, 2003). This means the estimated effect of shared files on sales has a negative bias.

Our approach differs from the current literature in that we directly observe file sharing. Our results are based on a large and representative sample of downloads, and individuals are generally unaware that their actions are being recorded.

III. File sharing Networks

To better understand the data we collected, it is useful to review the basics of P2P technology. File sharing relies on computers forming networks which allow the transfer of data. Each computer (or node) may agree to share some files and has the ability to search for and download files from other computers in the network. Individual nodes are referred to as clients if they request information, servers if they fulfill requests, and peers if they do both. Our data come from the OpenNap network, an open-source descendant of Napster. OpenNap is an example of a centralized P2P network which has individual clients log into a central server. The server returns to a client a set of potential matches for its search, after which the client may initiate a transfer directly from the host client.

During our study period in the fall of 2002, P2P networks were already quite large. FastTrack (which includes the popular KaZaA service) had grown to 3.5 million simultaneous users by December 2002. Typically, more than 500 million files holding 5 Petabytes of data were available on FastTrack/KaZaA at any time. The second largest network was WinMX, which had about 1.5 million simultaneous users in 2002. Even the smaller networks were fairly large. OpenNap had at least 25,000 simultaneous users sharing over 10 million files. Napster no longer operated in the fall of 2002.

IV. Data

We use two main data sources for this study. Logs for two OpenNap servers allow us to observe what files users download. Weekly album-level sales data come from Nielsen SoundScan (2003). SoundScan tracks music purchases at over 14,000 retail, mass merchant and on-line stores in the United States. Nielsen SoundScan data are the source for the well-known Billboard music charts. We complement download and sales data with information on song names and track times from AllMusic.com (2003), an on-line media guide published by Alliance Entertainment Corp. To develop our instruments, we rely on a large number of additional data sources which we discuss below.

A. File Sharing Data

Our file sharing data was collected from two OpenNap servers, which operated continuously for seventeen weeks from 8 September to 31 December 2002. The servers were connected to T-3 lines which provided actual Internet transmission speeds of several megabits per second for both uploads and downloads, which ensured all client requests could be handled. The information on file transfers is collected as part of the usual log files which the servers generate. An excerpt of a typical log file is:

```
[2:53:35 PM]: User evnormski "(XNap 2.2-pre3, 80.225.XX.XX)" logged in
[2:55:31 PM]: Search: evnormski "(XNap 2.2-pre3)": FILENAME CONTAINS "kid rock devil"
   MAX_RESULTS 200 BITRATE "EQUAL TO" "192" SIZE "EQUAL TO" "4600602" "(3 results)"
[3:02:15 PM]: Transfer: "C:\Program Files\KaZaA\My Shared Folder\Kid Rock –Devil
   Without A Cause.mp3" (evnormski from bobo-joe)
```

The last two lines in the log file show user "evnormski" downloading the song "Devil Without a Cause" by Kid Rock from user "bobo-joe". Information on downloads are the building blocks of our analysis. We focus on downloads because these are the files users actually obtain and they can potentially displace sales. Over the sample period we observe 1.75 million file downloads. This is about 0.01% of all downloads in the world.[3] A significant majority of downloads were music files. We restrict the analysis to audio files by clients in the U.S. The server logs include the I.P. address for each client which we use to identify our users' home country.

An important question is whether our sample is representative of data on all P2P networks. We present here a brief overview of this point. An Appendix A that addresses this question in greater detail is available from the authors. While we are unaware of any database spanning the universe of music downloads, we were able to compare downloads on our servers with a large sample from FastTrack/KaZaA, the leading network at the time. We find that the availability of titles is highly correlated on the two networks and cannot reject a null that the two download samples are drawn from the same population. The resemblance of the files is not surprising. Individuals in our data are similar to those on the most popular networks because the user experience is quite similar and many individuals employ software which allows them to simultaneously participate on several networks. We also find no evidence that network size

[3]At the end of 2003, roughly one billion songs are downloaded per week (*Wall Street Journal*, 19 November 2003). During February 2001, at Napster's peak, about half a billion songs were downloaded per week (Romer, 2002).

influences the distribution of downloads, and the effective number of files available to users is comparable on our servers and on FastTrack/KaZaA.

B. Sales Data

Ideally we would like to relate downloads of every track on every existing album to sales, but this is infeasible since millions of titles are available. Instead we focus on a sample which is a subset of albums sold in U.S. stores in the second half of 2002. The sample is representative of all commercially relevant albums and the releases from the major record labels,[4] allowing us to draw meaningful inferences about P2P's impact on the music industry as a whole.

The sample is drawn from a population of albums on 11 charts produced by Nielsen SoundScan (2003): Alternative Albums (a chart with 50 positions), Hard Music Top Overall (100), Jazz Current (100), Latin Overall (50), R&B Current Albums (200), Rap Current Albums (100), Top Country Albums (75), Top Soundtracks (100), Top Current (200), New Artists (150) and Catalogue Albums (200). The charts are published on a weekly basis, and we include an album in the population if it appears on any chart in any week during the second half of 2002. The original population is extensive (2,282 albums) and includes many poor-selling albums. For instance, our data include two albums which sold fewer than 100 copies during our study period and the 25th percentile of sales in our data is only 12,493 copies.[5] Thus while we study the commercially most relevant music, it would be wrong to think of our population as a set of superstar albums. Including many poor-selling titles also mitigates concerns about sample selection. From this population of albums, we draw a genre-based, stratified random sample of 680 releases. To reflect the popularity of different music styles, we set the sample share of a genre equal to its fraction of CD sales in 2002.[6] Within each genre, we randomly selected

[4]The genre charts we sample from made up 81.8% of all CD sales in the United States in the last third of 2002. This is virtually identical to the 2002 share of 83.6% for the Big Five record companies, and 97% of the albums on the annual version of these charts were released on RIAA-associated labels. On an annualized basis, the total number of albums on the charts matches up quite closely with the seven thousand releases from major labels for 2002 (RIAA, 2003b)

[5]A typical measure of album success is gold certification which occurs at sales of half a million copies.

[6]Albums can appear on more than one chart because some charts (e.g., New Artists, Top Current) comprise many musical styles. For sampling purposes, we grouped all albums by style; a Rap album on the Top Current list is grouped with all other Rap albums during the sampling process. In the descriptive statistics, we classify albums by their original charts.

individual titles. Random sampling is obviously important for the validity of our analysis (Kish, 1987).

The average album in the resulting sample sold 143,096 copies during our study period. Table 1 reports sales statistics for the full sample and for individual categories. Across all categories, 44% of population sales are represented in the sample. We conducted a two-sample Kolmogorov-Smirnov test to compare the distribution of sales on the original charts and in our sample. We are unable to reject the null that sample sales are representative of the population of all albums on the Billboard charts (p=0.991). We also reject this null comparing each of our 11 original charts with the sample sales for that particular chart (p>0.539 for all 11 charts.) The sample is also representative across time, with little variation in the share of popular sales present in each week.

In order to compare sales and downloads, we match the 260,889 audio files which U.S. users successfully transferred during our study period to the 10,271 songs on the 680 albums in our sample. The matching procedure is hierarchical in that we first parse each transfer line, identifying text strings that could be artist names. These text strings are then compared to the artist names in our set of albums. The list of artists contains the name on the cover and up to two other performing artists or producers that are associated with a particular song. For example, the track "Dog" on the B2K album "Pandemonium" is performed by Jhene featuring the rapping of Lil Fizz. For "Dog," B2K, Jhene and Lil Fizz are recognized as artists. Once an artist is identified, the program then matches strings of text to the set of songs associated with that particular artist. For both artists and songs, we allow matching on substrings ("Snoop Dog" matches "Snoop Dogg"), and we ignore punctuation marks such as apostrophes that are often ignored in the names of files. Using this algorithm, we match 47,709 downloads in the server log files to our list of songs, a matching rate of about 18%.

C. Descriptive Statistics

As this is one of the few data sets that allows us to directly observe P2P users, we describe our data in some detail to convey a sense of what individuals do on P2P networks. A first stylized fact is that file sharing is truly global in nature. **Figure 1** presents the distribution of users across countries for our sample period. While over ninety percent of users are in developed countries, a

total of 150 countries are represented in the data. U.S. users represent 31% of the sample. Table 2 shows the top countries in terms of users and downloads. As the data indicate, there is only a loose correlation between user share and other country covariates such as Internet use or the software piracy rate. Column 3 in Table 2 confirms that interactions among file sharers transcend geography and language. Only 45.1% of all files downloaded by U.S. users come from computers in the U.S., the remainder comes from a diverse range of countries including Germany (16.5%), Canada (6.9%) and Italy (6.1%). The five percent of downloads not covered in this list are spread out over almost every other country in the data.

While file sharing activities are dispersed geographically, only a limited number of songs are transferred with any frequency. Table 3 shows the average song is downloaded 4.6 times over the study period, but the median number of downloads is zero.[7] Although our sample is representative of all commercially relevant music in the second half of 2002, it is striking to see that more than 60% of the songs in our sample are never downloaded. The most popular song among our users is "Lose Yourself" from the 8 Mile Soundtrack, which was downloaded 1,258 times (2.6% of all matched downloads). Aggregated up to the album level, users downloaded 70 songs from the average album in our sample. The 8 Mile Soundtrack, which was the second highest selling album in our sample during the observation period, was the most popular among file sharers and downloaded 1799 times. For the sum of all weeks, the median number of downloads per album is 16, the 75th percentile is 63, the 90th percentile is 195, and the 95th percentile is 328. Both downloads and sales closely follow a power-law (pareto) distribution.

File sharing is limited to a select number of songs and most of these songs come from just a few charts. Tracks on the Top Current chart ("Billboard 200") are most frequently downloaded. Downloads from this chart alone make up 48% of all file transfers. Another 25% come from the "Alternative" category. The remaining 9 charts are not particularly popular among file sharers (see Table 3). In view of the low cost of sharing files and sampling music on P2P, it would seem reasonable to expect users to seek out a great variety of songs representing many musical styles. But this is not the case. P2P downloads closely resemble the play lists of Top 40 radio stations. What is being transferred are mostly the hits of the week. As a result, it is not surprising that songs from higher-selling albums are downloaded more frequently (Table 4). In the top quartile

[7]The 75th percentile of downloads per song is 2, the 90th percentile is 11, and the 95th percentile is 22.

of sales, albums average 200 downloads. In the bottom category, the mean number of downloads is only 11. As Table 4 shows, the mean number of downloads increases at a rate that is less than proportional to the rate of increases in sales. While downloads and sales are both quite concentrated, it is worth pointing out that downloads are a bit more dispersed.[8] Still the similar pattern of concentration is evidence that common factors drive downloads and sales, a key concern for the development of our empirical strategy.

V. Empirical Strategy

A. Econometrics

Our goal is to measure the effect of file sharing on sales. An Appendix B with a formal model of purchase and download behavior is available from the authors. Here, we highlight the key implications of this model. The simplest approach is to estimate pooled models of the form,

$$(1) \qquad S_i = X_i\beta + \gamma D_i + \mu_i ,$$

where i is the album, S_i is observed sales, X_i is a vector of album characteristics and D_i is the number of downloads. This is generally inappropriate because the number of downloads is likely to be correlated with unobservable album-level heterogeneity. As the descriptive statistics suggest, the popularity of a particular band is likely to drive both file sharing and sales, implying the parameter of interest γ will be estimated with a positive bias.

Making use of the fact that we observe sales and downloads for seventeen weeks, we can control for album-specific time-invariant characteristics by estimating the fixed effects model,

$$(2) \qquad S_{it} = X_{it}\beta + \gamma D_{it} + \sum_s \omega_s t^s + \nu_i + \mu_{it} .$$

In this specification, ν_i is an album fixed effect, t denotes time in weeks, and the summation

[8] Over the entire observation period, the across-album Herfindahl index is 0.010 for sales and 0.009 for downloads. The weekly top-selling albums account for 7.6% of total sales while the weekly most-downloaded albums account for 5.2% of all transfers. Similarly, the weekly top ten accounts for 31.5% of total sales and 25.7% of all downloads.

allows for a flexible time effect.[9] While the fixed effects in this specification address some concerns, there is good reason to believe that album-specific time-varying unobservables μ_{it} might be critical in our application because album sales decay at very different rates.

We address this latter issue by instrumenting for D_i in both (1) and (2). That is, for the panel data approach we substitute into (2) the fitted value of downloads from

$$(3) \qquad D_{it} = Z_{it}\delta + X_{it}\beta_2 + \sum_s \omega_{2s} t^s + \nu_{2i} + \mu_{2it}\,.$$

Valid instruments, Z_{it}, influence file sharing but are uncorrelated with the second stage errors, μ_i or μ_{it}. Shifters of download costs are candidates for instruments because they influence downloads but often have no direct influence on sales (this is formalized in Appendix B). Our instruments are in the spirit of the differentiated products literature, where the problem is correlation between prices and unobserved product quality. To break this link, Berry (1994) and Bresnahan, et al. (1997) suggest using cost shifters and characteristics of competing firms as instruments for prices.[10] An advantage of our instruments, which we discuss below, is that they stem from factors not relevant to purchase decisions, and so do not rely on the common but potentially problematic assumption that product characteristics are exogenous (Nevo, 2001).

B. Instruments

We use three types of instruments to capture a wide variety of forces which influence the cost of downloading music files: album-specific instruments which are fixed over time, time-specific instruments which have a similar impact on all released albums, and finally time-varying and album-specific instruments. Only the time variation for the last two groups of instruments is used for identification, since these variables are included in specifications with album fixed effects. The availability of panel data is clearly central to our approach. Some of the

[9]We consider a polynomial time trend of degree six, and in supplemental estimates we include a full set of week fixed effects.

[10]We avoid many of the econometric complications of this literature because our model focuses on within-product choices (purchase or download) rather than between-product choices (which album to purchase). In particular multiple albums may be consumed, so our endogenous covariate, downloads, enters the demand function in a relatively simple manner. We can apply instrumental variables directly to the demand equation, rather than the transformation laid out in Berry (1994). See Section D of Appendix B for details.

instruments also presume a scarcity of supply. In fact there are queues even for popular tracks, since users limit the number of simultaneous uploads from their collection.

We first describe the time-invariant instruments used in the pooled models. The first two instruments are motivated by database and information theory (Salton, 1989; Shannon, 1948). To download a song, a user's search query must match a shared file. Popular file sharing software is rather rigid in determining matches.[11] Unless both the searcher and sharer agree on the naming convention, no match will occur. This two-sided search problem suggests some candidate instruments. We first consider the median number of "misspellings" in an album's song titles. Many song titles have unconventional spellings which can make it more difficult to find these songs. We use MS Word's spell checker to determine the number of unconventional spellings. A related instrument is based on the length of the song title. Very long strings are less likely to be found because users often truncate names. For example many song titles have parentheses at the end or beginning, and if the sharer omits these the search request will be unfulfilled. As a result, longer titles are less likely to be matched. Our instrument is the number of words in the shortest song title on an album. There is little reason to suppose these two features of song naming convention have any direct impact on sales.

The next album-specific time-invariant instrument is the length of songs, a proxy for the time it takes to download the file. Song lengths vary widely in our sample, from as short as a few seconds to as long as forty minutes. Transfers of bigger files take longer and are more likely to be interrupted. During our study period, roughly every other file transfer remains incomplete, in part because fewer than a third of U.S. homes with Internet access had a broadband connection (Nielsen//NetRatings, 2003a). We therefore expect an album's average track length to be negatively related to the number of downloads.[12] There is a similar logic for using an album's minimum track length, which we expect to be positively related to downloads.[13] These are likely

[11]For example, "lose yourself," the name of a popular song, would typically return over a thousand results, but mistyping even one character (such as "lose yourse;f") or omitting part of a word ("lose yours") returned zero results.

[12]Actual download time (exclusive of the initial queuing period) can vary from a few minutes up to an hour based on the size of the file (we documented this pattern which is also reported in Gummadi et al, 2003).

[13]Many albums contain very short tracks, typically introductions by the artist, which are unlikely to be downloaded because they may not hold great intrinsic interest and because they often have similar titles ("Intro," "Outro," or "Skit") and are difficult to find on P2P networks. As the shortest track gets longer, it becomes more likely that it is a real song as opposed to a spoken introduction.

to be valid instruments as song length has little relationship to popularity, with some top-selling albums consisting entirely of short tracks and others having mainly longer numbers. Our last time invariant instrument is the number of song titles on an album which also appear on other albums by the performing artist. Such songs are more widely available for download, since there are more potential sources sharing the song.[14]

A second class of instruments are time-varying factors which have a fairly uniform impact on all albums. The first such instrument is based on an exogenous supply shift. Recall that one out of every six U.S. downloads comes from Germany, a relationship which is also documented in the authoritative BigChampagne P2P database (New Media, 2004).[15] Our instrument is the number of German kids on vacation due to German school holidays. The holidays produce an exogenous supply shock of files, making it easier for U.S. users to download music. We believe the holiday variation is related to U.S. downloads because German teens, the primary participants in file trading, tend to go on-line at home and not in school (Niesyto, 2002). These kids can stay up later when out of school allowing them to engage in file sharing during the peak U.S. hours (early evening, est). Vacations also provide more overall time for file sharing. Supporting this intuition, we find empirically that the number of German kids on vacation is positively correlated with the number of files uploaded from Germany to the U.S. (ρ=0.42). In a simple regression model including a polynomial time trend of degree six, the number of German kids on vacation is a significant predictor of the number of files uploaded from Germany to the United States (p=0.011). The vacation variable varies over time because the sixteen German Bundesländer (states) start their academic year at different points in time. In addition, German kids have typically two to three weeks of fall vacation and the timing of this recess also varies by Bundesland (Agentur Lindner, 2004; Kultusministerkonferenz, 2002). There is little reason to believe this variable is endogenous.

The other time-varying instruments are four "Internet weather" measures which are indicators of Internet congestion and the cost of downloading: The Consumer 40 Performance Index, which is

[14]This instrument would not be valid if more popular tracks are released on multiple albums. It is not obvious if this is the case because many repeated titles are from older artists and older albums, both of which tend to be less popular. For all instruments used in this study, we will report overidentification tests in the Results section.

[15]The intuition for this link is that connection speed, and not physical distance, matters in P2P. During our study period Germany had the highest rate of high speed internet access—over half of homes had a broadband or ISDN connection-- and the largest internet population in Europe (Nielsen//Netratings, 2002)

based on access times to popular websites (Keynote, 2004); the average and the standard deviation of ping times in the Internet End-to-end Performance Measurement (IEPM) measured in milliseconds (IEPM, 2004); and finally the fraction of Internet2 backbone traffic that is due to file sharing (Internet2 Netflow Statistics, 2004). These variables reflect the delays a typical P2P user faces. For example, the IEPM average measures typical roundtrip times for data packets between a wide range of Internet locations. Similarly, a high share of file sharing traffic on the backbone will delay downloads. The IEPM standard deviation is included to account for possible intra-week retiming of downloads: controlling for the mean, there will be more low-congestion periods conducive to downloads when there is greater dispersion in ping time. All of these congestion measures are plausibly exogenous to music sales and the unobserved popularity of any particular album. This is because the traffic related to any given album is miniscule; overall congestion variation is mainly driven by the more volatile internet traffic sources which are not related file sharing.

The final two instruments are album-specific and time-varying cost shifters. These are useful because they provide identification even if a full set of week and album fixed effects are included. We first consider the mean time length of albums in the same music category, which should influence the availability of tracks for the album in question. The idea is that users tend to supply files of a similar genre, and there is some crowd-out in supply stemming from limits in storage space. This crowd-out varies over time, since new competing albums are continually being released. For instance, a hip-hop fan is less likely to share some rap song when related artists have recently released an album.[16] Note we are not presuming individuals delete the older track, but rather that they archive them on a media (like a CD) which is not shared. Since the timing of release dates may be a function of the unobserved album popularity, the number of competing albums cannot be used directly. Instead we focus on the distribution of track times on other albums in the same genre. A second time-varying album-specific instrument is the interaction between the number of German students on vacation and an indicator for whether an album's performing artist is on tour in Germany that week. Tours spur local interest and sales of an album, and they seem likely to create a positive supply shock of downloadable files. This [tour × vacation] instrument should not directly be related to U.S. sales because the promotional

[16]We observed one such crowd-out of songs on Nelly's Nellyville album when the 8 Mile Soundtrack was released.

effect of tours is local and will not spill across the Atlantic. Also the decision to tour largely reflects popularity in Germany (since the U.S. concert season is heavily weighted towards summer), and the timing of fall and winter concerts typically reflects idiosyncratic features like venue availability and weather. We present summary statistics for all our instruments in Table 5. Each of the measures we use exhibits noticeable variation.

VI. Results

Before turning to the estimates, it is instructive to consider the dynamics of downloads and sales. **Figure 2** shows the weekly time series of sales and purchases for two albums. The "Superstar" album was largely ignored in file sharing networks until it became available for sale in week ten of our sample. This suggests it is the publicity associated with an official release which drives downloads as well as sales. Notice also the rapid decay in sales and downloads, which highlights the importance of using high-frequency data. The bottom panel shows trends for a "Sleeper" album. Here a rise in downloads leads a sales increase in the last weeks of the data, suggesting again that similar forces drive downloads and sales.

A. Pooled Sample Models

Table 6 reports the results for specification (1), which pools sales and downloads in all weeks. Among the strengths of this approach is that it imposes minimal dynamic structure, and it allows downloads to influence sales with a long lag or lead. Model (I) is a simple regression with controls for an album's music category and the number of weeks since its release.[17] There is no explicit control for prices which have virtually no cross-sectional variation. The OLS estimates show that downloads are positively correlated with sales, though this is likely the result of unobserved album heterogeneity. The remaining columns of Table 6 present 2SLS estimates for the pooled data. Columns (II) include only the misspelling instrument, for which we have a strong prior of exogeneity to sales. The first-stage estimates show that, as anticipated, more unconventional spelling significantly discourages file sharing. This effect is economically

[17]In the unreported genre fixed effects, Top Current albums have significantly higher sales and downloads than each of the other categories. The differences between any other two categories are small.

meaningful, with a one standard deviation increase in the instrument reducing downloads by one seventh of its mean. Instrumenting noticeably reduces the second-stage effect of downloads on sales, which remains positive but is no longer statistically significant. Columns (III) add the additional time-invariant instruments. They have the expected effect on downloads. At the bottom of the Table, we report a Sargan-type overidentification test. We cannot reject a null that the instruments are valid and correctly excluded from the second stage. Adding the additional instruments has only a modest effect on the point estimate of downloads in the second stage, but the parameter is more precisely estimated. Nonetheless, the effect is still not significantly different from zero. This result is robust to omitting individual instruments in the first stage.

The predicted impact on sample sales is reported at the bottom of the Tables throughout the Results section. The impact is the difference between predicted sales and the fitted value when downloads are set at zero (for industry-wide impact calculations, see Section F.)[18]

B. Panel Data Models

In Table 7 we report results for specification (2), where the unit of observation is the album-week. A simple OLS specification with a polynomial time trend of degree six is in column (I). A model with a time trend and album fixed effects is given in (II). While we continue to find a positive effect of downloads on sales, the relationship is much weaker in the fixed effects model, indicating that unobserved time-invariant album characteristics impart a positive bias on the pooled OLS estimates.

The remaining estimates in Table 7 instrument for downloads. We again begin by using the one instrument for which we have the greatest confidence in exogeneity, the number of German kids on school holidays (columns III). The first stage estimates indicate that, as expected, increases in the number of German kids on vacation lead to a larger number of downloads in the U.S. A one standard deviation increase in children off from school boosts weekly album downloads by 2.4, which is slightly more than half of the mean. Once we instrument for downloads, the estimated effect of file sharing on sales is quite small (slightly negative) and statistically indistinguishable

[18]For the linear pooled models from specification (1) this is calculated as $\sum_i S_i(D_i)-S_i(0) = \gamma \times \sum_i D_i$. For the linear panel models from specification (2), this is $\sum_t \sum_i S_{it}(D_{it})-S_{it}(0) = \gamma \times \sum_t \sum_i D_{it}$. The formulae naturally generalize to the non-linear models.

from zero. We next consider a specification which adds the remaining time-varying instruments: the Internet congestion measures, the non-sales characteristics of competing albums, and the German tour-school holiday interaction (columns IV). The latter two variables are of particular interest since they vary across albums as well as over time and so provide an additional source of identification. The new instruments have the expected first-stage signs. Greater congestion and ease of acquiring competing albums reduce downloads while volatility in congestion or international promotion increase downloads. The Sargan test at the bottom of the table indicates the instruments satisfy the standard validity criterion. In this richer model downloads have a small positive effect on sales, but the effect continues to be statistically indistinguishable from zero. As in the pooled models, the results are robust to omitting individual instruments.

The models in Table 7 only allow for a contemporaneous effect of downloads on sales, which seems rather restrictive. It is quite possible that downloads crowd out sales at a later point in time. In Table 8, we address this issue by studying the effect of several weeks of downloads on sales and by estimating GMM models. Downloads are highly correlated across time which prevents us from including downloads in past weeks as individual covariates. Instead, we study the effect of a weighted sum of current and past downloads on current sales. Downloads are instrumented using the full set of instruments reported in Table 7. Our formal measure is the weighted stock of current and previous weekly downloads, $D_t^{Stock} = \sum_{s\geq 0}\delta_s \times D_{t-s}$. The weights δ_s are chosen in a grid search that minimizes the unexplained fraction of the variance in our sales equation subject to $\delta_s \geq \delta_{s+1}$. The weights $(\delta_0, \ldots, \delta_T)$ are (1,1,0.05) when we allow two lags in model (I), and (1,1,0.1,0.05) when we allow three lags in model (II). It is interesting to note that the weights which best fit our data give much importance to downloads in the current and previous week, while downloads further back in the past do not appear to heavily influence sales. We continue to find small and statistically insignificant positive effects for the weighted sum of three weeks of downloads (column I) and four weeks of downloads (column II).

Models (III) and (IV) in Table 8 use the Generalized Method of Moments estimator developed by Arellano and Bond (1991). These GMM models are more general than the previous specifications in the sense that we do not need to make any explicit assumptions about the appropriate lag structure. The lags of sales that are included on the right-hand side account for any effect that past downloads might have had on the sale of CDs. The model is estimated in

first differences. We instrument for past sales using suitable lags of their own levels.[19] For this type of model, the two-step estimates of the standard errors tend to be downward biased (Blundell and Bond, 1998). We correct standard errors using the two-step covariance matrix derived by Windmeijer (2000). Arellano-Bond tests for autocorrelation are applied to the first-difference equation residuals. Second-order autocorrelation would indicate that some lags of the dependent variable which are used as instruments are endogenous, but the tests reveal no such problem. As in our previous estimates, we find a small and statistically insignificant effect of file sharing on sales using either the level or stock of downloads.[20]

Our previous models constrained the effect of downloads on sales to be identical for all releases. In Table 9, we relax this assumption in several ways. We first explore the idea that the effect varies by artist popularity. We do this by interacting the download variable with three measures of popularity: the number of previous releases by the same artist, the Billboard ranking of the artist's last release, and the best ever Billboard ranking by the artist. We cannot directly include these time-invariant popularity measures because the models are estimated with album fixed effects in both stages. But the interaction term [downloads × popularity] varies by week. The coefficient on this term measures how downloads' impact on sales changes with artist popularity. To make it easier to interpret the results, Billboard ranks are coded as [201 – actual rank] so that larger numbers indicate greater popularity.[21] We estimate these models using both contemporaneous downloads and the weighted sum of downloads, and we employ the full set of instruments from Table 7.

The estimates in Table 9 provide some evidence that the effect of file sharing varies by popularity. While the download terms are typically positive, the estimated coefficients on the three types of popularity interactions in columns (I)-(VI) are all negative, indicating that the

[19]The formal model is,

$$S_{it} = \alpha S_{i,t-1} + X_{it}\beta + \gamma D_{it} + \sum_s \omega_s t_s + \nu_i + \mu_{it}.$$

The lagged sales term soaks up any delayed effect of downloads, regardless of how far in the past they occurred (taking a Koyck transformation yields a specification with infinite lags of downloads on the right hand side). Estimating in first differences purges the fixed effects. We instrument for the first-differenced $S_{i,t-1}$ which are now endogenous.

[20]In model (IV), the download stock weights ($\delta_0,\ldots, \delta_T$) are (1,1,0.35,0.1).

[21]Specifically, the interaction terms are [downloads × # of prior releases] and [downloads × band had Billboard ranking × (201−Billboard rank)].

effect of downloads on sales is less positive for more popular artists. However, the joint effect of the download and the popularity interaction terms is never statistically significant (see hypothesis test at the bottom of the Table.) More importantly, the predicted effect on sample sales remains positive and small in all specifications. This is in part because downloads have a small effect on even the top-selling album and a positive effect on less popular titles, as we describe in more detail in Section F.

An alternative measure for the popularity of a band is the frequency with which its songs are downloaded. We study the link between this measure of popularity and sales in a specification where downloads enter as a spline. Column (VII) presents estimates for a four-spline model. The knots are located so that each segment has a quarter of the fitted download values. In this estimate, none of the parameters are statistically significant. Albums with few downloads appear most likely to benefit from file sharing, while moderately popular albums appear to be hurt. The large standard errors are partly the result of the collinearity between the download terms. Statistical insignificance notwithstanding, the predicted impact on sample sales remains positive and the economic effect on sales is quite small for even the top selling album.

From a social welfare point of view, it is particularly important to understand how file sharing affects monetary incentives for new artists. In columns (VIII) and (IX) of Table 9 we interact downloads with an indicator which takes on a value of one if the release is an artist's first CD. The measured effect is economically and statistically insignificant. Finally, we investigate whether the effect of downloads on sales varies by the number of popular songs on a CD. As we documented earlier, most file sharers obtain just a few tracks from a CD. One might suspect that P2P is a fairly good substitute for CDs with only one or two popular songs. We calculate a Herfindahl index for each album-week as a measure of concentration of downloads. The index is included in both the first and the second stage in model (X) of Table 9. There is weak evidence that albums with more concentrated downloads sell somewhat better, but this effect is smaller at higher volumes of downloads. Both effects are individually and jointly insignificant. This is unsurprising. Individuals download only a few tracks from each album, which they are unlikely to value at the full purchase price.

C. "Drop-out" Hypothesis

A possible explanation for our inability to find a statistically significant relationship between file sharing and sales is that file sharers and consumers who purchase music are in fact two completely separate groups. According to this hypothesis, growth in the file sharing community does in fact displace sales but we cannot identify this effect in our previous models because our data do not reflect the growing number of file sharers.

There are three responses to this conjecture. First, it is inconsistent with what we know about consumer behavior. The premise underlying the "drop-out" hypothesis is that file sharers no longer buy CDs. However, every survey we are aware of, including the industry studies listed in the literature section, indicates that downloaders, even heavy ones, continue to purchase legal CDs. We corroborated these findings with our own survey of individuals who were engaged in file sharing (described in Oberholzer-Gee and Strumpf, 2004). Ninety percent reported that they recently purchased a CD, a value reaching one hundred percent among the most active downloaders.

Secondly, we can test the "drop-out" hypothesis directly by controlling for the increasing number of users. An implication of the hypothesis is that our download sampling rate declines over time because the servers for which we have data handle a limited number of users.[22] Growth in the file sharing community, however, is managed by additional server capacity which we do not observe. If we account for this growth we should find a negative relationship between downloads and sales, since the "drop-outs" are replacing purchases with transfers. We address this issue by scaling up the number of downloads in our sample to reflect the growth in the file sharing community. We use the number of FastTrack/KaZaA users as a proxy for the rate of growth in overall file sharing. Because the number of users increased by over a third over our observation period, there should be enough growth to detect a drop-out effect. Table 10 reports the estimates for the linear model (with the German vacation instrument and the full set of instruments), for a model with a stock of previous downloads (the weighted sum of four weeks), and for the dynamic GMM model. In all the models downloads still do not have a significant

[22]Our servers did have some slack capacity at almost all times, so there was space, for instance, for the additional international users when they wanted to log-in. However, there was not enough slack capacity to handle the proportionately much larger expansion of the file sharing community over the entire observation period.

effect on sales after the scaling of downloads. A third approach to testing the drop-out hypothesis is to compare the long-run sales growth of individual genres of music. We describe the results of this experiment at the end of section E.

D. Robustness Tests

To further corroborate our results, we perform a large number of robustness checks, some of which we report in Table 11 (detailed results for all tests are presented in Oberholzer-Gee and Strumpf, 2004). First, we investigate the importance of the holiday season when many consumers purchase CDs as gifts. It is possible that downloads are less substitutable for sales during this period due to the reluctance to give a burned CD as a present. Note this is also an argument against the idea that file sharing is the main cause of the sales decline since purchases are heavily concentrated in the holiday season. Still it is straightforward to test for this effect. In Table 11, column (I), we exclude the December data from our sample and continue to find that there is no statistically significant effect of file sharing on sales. These estimates also suggest that the imbalanced panel, due to the entry of albums during the observation period, does not drive the estimates.

A further issue is the role of promotion. Individuals might be primarily interested in getting heavily promoted tracks, so there should be greater substitution between downloads of such songs and album purchases. However, when we omit downloads of popular "singles" tracks (as measured by presence on the weekly "Billboard Top Ten Singles" list, the "Billboard Top 50 Airplay" or the MTV Video Playlist) the estimated effects are largely unchanged (Table 11, column II). In unreported results, we find similar estimates when we explicitly control for media promotion or touring.

In the remaining columns of Table 11, we pursue a different identification strategy and rely exclusively on time-varying, album-specific instruments. This allows us to replace the polynomial time trend with week fixed effects which means we fully control for time-varying heterogeneity, such as shifts in aggregate demand, which impact all albums. The causal effect is identified from album-specific changes over time in the instruments. In column (III) we include only the [German vacation × tour] instrument, for which we have the most confidence in exogeneity. In column (IV), we include three additional instruments: the track times of

competing albums, which we had introduced before, and two new German chart instruments: the current ranking on the weekly German Top 100 singles chart for the best-selling track on an album (coded as 101 minus rank), and the current number of weeks the track was on these charts (Musikmarkt, 2002). The latter is included since the charts evolve slowly, so a highly-ranked song will be downloaded less if it has been available for many weeks (due to fatigue in demand). Information from German charts can potentially be used as an instrument because tracks on these charts are more popular in Germany and should be in greater supply. Obviously, there is a concern that these chart positions might be endogenous. Note, however, that this instrument is included along with album fixed effects, so it is the timing of the chart rankings in Germany that are identifying downloads. In addition there are clear differences in tastes between the U.S. and Germany, which lead to differences in the dynamics of the popularity of tracks. For example, European-based performers tend to have highly-ranked singles in Germany earlier than in the U.S. These differences are particularly accentuated outside the very top ranked songs, and we consider a relatively extensive ranking of songs.

In models (III) and (IV) in Table 11, which include both album and week fixed effects, file sharing bears no statistically significant relationship to sales. The Sargan test gives no indication that the German chart instruments are endogenous. Omitting the two charts instruments does not qualitatively change our results, although the resulting estimates are less precise.

In additional robustness tests, we explore alternative specifications of the models presented in the paper (see Oberholzer-Gee and Strumpf, 2004 for details). For instance, to address the possibility of non-stationarity we estimate the panel models in first differences, using the full set of first-differenced instruments. In these models, we continue to find that the number of downloads has no statistically discernible effect on sales. To reduce the importance of outlier albums with a large number of sales, we use log(Sales) as the dependent variable. The impact on sales continues to be insignificant both in specifications with contemporaneous downloads and in specifications with the stock of downloads. Next, we re-estimate our panel models sales quartile by sales quartile, again using the full set of instruments. In these models, there is a negative but insignificant effect for all quartiles. The effect is estimated rather precisely for the bottom three quartiles and much less precisely for the top quartile. We also allow the effect of downloads to vary across music genres. This is important because music styles which cater to older age

groups could be less impacted by file sharing. We find little evidence for this claim, since all but one of the [download × genre] interactions are statistically insignificant.

Finally, we consider the possibility that downloads have spillover effects between albums. It is possible that downloads reduce sales through an indirect mechanism, causing sales to shift away from competing albums towards the one being downloaded. This inter-album substitution would not be detected in the main specifications and could compromise our calculation of the aggregate effect of downloads on sales. However, after allowing such spillovers we find that albums have higher sales when there are more downloads of tracks on other albums in its genre.

E. Quasi-experimental Evidence

In addition to the evidence presented so far, our data also allow us to study the impact of P2P on sales in a quasi-experimental context. In particular, we can examine how album sales respond to relatively exogenous variation in file sharing intensity during certain times of the year, in certain geographic areas, and across music genres.

The first experiment involves variation over time. The number of file sharing users in the U.S. drops fifteen percent over the summer (estimated from BigChampagne, 2004) because college students are away from their high-speed campus Internet connections. If downloads crowd out sales, we should observe that the share of albums sold in the summer increases following the advent of file-sharing. We consider a differences-in-differences approach and compare the share of summer sales in the period prior to file sharing (the control group) with sales following the introduction of file sharing (the treatment group). We calculate the share of album sales occurring in May-September using weekly data from Nielsen SoundScan (2003). We find that the introduction of widespread file-sharing has had virtually no impact on summer sales. In the four years (1995-1998) preceding the introduction of Napster, the average share of summer sales was 37.0% with a range of 36.4-37.8%. During the more recent period of extensive file-sharing (1999-2003), the average share of summer sales was 37.2% with a range of 35.9-37.8%. Using a differences-in-differences approach, the elasticity of sales with respect to file sharing users is

−0.01, indicating that file sharing displaces few album sales.[23] Similarly, we find that the relative size of Christmas sales slightly decreases after 1998 despite the reduction in file sharing activity and the lower substitution between downloads and purchases during this period.

A second experiment considers variation across space. Recall that U.S. users download over a third of their music files from Western European countries such as Germany and Italy. Due to time zone differences, such transfers are far easier for east rather than west coast users. This is because the peak period for file sharing, 7pm to 3am, overlaps between Western Europe and the east coast (which have a six hour time difference), but not between Europe and the west coast (which have a nine hour difference). So east coast users can draw on a larger base of files from international users than west coast users. Therefore if file sharing has a large negative effect on record sales, sales during the file sharing era should decrease more on the east coast than the west coast.

For the period 1998-2002, we obtained total album sales for the one hundred and one largest "Designated Market Areas" (Nielsen SoundScan, 2003). Each of these areas was classified as east coast, west coast, or other based on their time zone. We then calculate annual sales by time zone from these values. Despite the differences in availability of files, sales have not noticeably varied across the country. In 1998, the last year in the pre-P2P period, the share of album sales in the eastern time zone was 43.9%. This share has hardly moved since then and over 1999-2002 the mean was 43.5% and the range was 42.7-44.0%. This suggests some common national factors, rather than file-sharing, are driving sales trends.

A final experiment, which also provides a test of the "drop-out" hypothesis, is to see whether download intensity influences long-run sales growth after explicitly controlling for trends in music format popularity. The model for the period 1999-2003 (the lifetime of file sharing) is,

$$(4) \qquad \text{Sales Growth}_g = \alpha + \gamma \times \text{Downloads}_g + \lambda \times \text{Listenership}_g + e_g$$

[23] ε = (%Δ in summer sales share)/(%Δ in summer file-sharing users) where the difference is taken between the pre- and post-file-sharing periods. Using mean annual sales, %Δ in summer sales share=0.2% and %Δ in summer file-sharing users=−15%. The elasticity calculation likely overstates the impact of the file-sharing, since the reduction in summer file-sharing is concentrated among the heaviest downloaders (college students).

where g indicates genre, Sales Growth_g is the percentage growth in sales over 1999-2003, Downloads_g are measures of genre-specific download intensity from our data (which we presume to be representative of other time periods), and Listenership_g is the genre-specific radio listenership growth rate (Arbitron, 2004). Since downloading is relatively concentrated across genres (Table 3), the "drop-out" hypothesis predicts a greater sales reduction for genres which are popular on file sharing networks. The estimated γ is not statistically significant using either downloads levels or downloads relative to purchases (for example using mean downloads per album and controlling for genre sales levels, the estimated γ=0.02 with a standard error of 0.23).

The results of all three quasi experiments are consistent with our earlier findings. Looking at variation in downloading intensity that is either due to geography, seasonality or the genre of music, we find no evidence that the advent of P2P technologies caused the recent slump in music sales.

F. Economic Impact of File Sharing

The statistical insignificance of the point estimates in all our models notwithstanding, how large an effect did P2P have on CD sales in 2002? This effect can be readily determined from our estimates (see note 18). The predicted effect for our album sample over the observation period is listed below each specification. The values listed at the bottom of Tables 6-11 indicate that the effect is generally quite small. While sample sales are about 100m, file sharing changes sales by less than 13m in all of the instrumented specifications. The most negative effect is a displacement of less than 1m sales.

We can also use the non-linear estimates in Table 9 to determine the distributional impact on different kinds of albums. While the estimates in columns (I)-(VI) imply that more popular albums tend to be hurt, this effect is never large. Using column (III) which is the most pessimistic estimate, downloads depress sales of the top-selling album (Dixie Chick's "Home") by 300,000 copies which is less than a tenth of its sales over the observation period. This effect is even smaller in the more general spline estimate in column (VII), which implies the Dixie Chicks release gains 9,000 copies over the study period. Album sales are largely unaffected under the spline, with an effect that ranges from −22,000 to 26,000 copies over the study period.

What is of greater interest than the effect of file sharing on sample sales is to determine the annual aggregate impact (AAI) on total album sales. Recall that the sample was constructed to be representative of the population of commercially relevant albums. We can make inferences about the impact on the entire industry after scaling up the number of album sales and after annualizing the observation period. A simple calculation shows the AAI is 5.04 times the sample effect discussed above.[24] Focusing on the most negative point estimate (column X in Table 9), the annual industry sales loss due to file sharing is 3 million copies. This is virtually rounding error given that sales in 2002 were 803m CD albums (RIAA, 2003a).

Our estimates can also be used to assess two leading hypotheses regarding file sharing's impact on album sales. In particular we would like to consider the "RIAA hypothesis" that downloads are responsible for the 80m reduction in CD sales in 2002 (RIAA, 2003a). We can also test the hypothesis that downloads have no effect on sales. To do this we take into account whether the hypothesized value falls outside of the 95% confidence interval around each point estimate of the AAI. The only uncertainty in our AAI calculation concerns γ, the estimated impact of downloads on sales. Thus the confidence bands are the AAI point estimate plus or minus twice the (scaled up) standard errors of γ. For example the lower bound of the estimate in Table 7 column (IV) is,

$$\sum_t\sum_i(D_{it}\times 5.04\times 1000)\times(\gamma-2\times se(\gamma))=200m\times(0.012-2\times 0.170)=-70m.$$

In general we can reject the RIAA hypothesis for all of the instrumented panel estimates in Tables 7-10. (At a lower confidence level we can reject the RIAA hypothesis for the pooled estimates in Table 6 and when week fixed effects are included in Table 11). Our estimates become more precise if we relax the assumption that file sharing exclusively affects this week's sales via this week's downloads. Based on the more realistic models of consumer behavior in Table 8, where we analyze the weighted sum of downloads and specifications with lagged sales, we can make sharper statements. For instance, for the dynamic panel estimates in Table 8

[24]AAI = (Effect of file sharing on sample sales over observation period) × (population sales/sample sales) × (file sharing activity over year/file sharing activity in observation period). From our sales data, the ratio (population sales/sample sales) is 2.27. The second ratio is (File sharing activity over year/file sharing activity in observation period) = 2.22. The latter calculation is based on weekly file sharing traffic rates over the 2002 calendar year on the Internet2 backbone (Internet2 Netflow Statistics, 2004) and the monthly average number of U.S. file sharing users (BigChampagne, 2004). Note that the second conversion factor is close to a naïve correction based simply on time, (52 weeks in year/17 weeks in observation period)=3.06.

column (III) we can reject the hypothesis that file sharing causes even *a quarter* of the 2002 sales reduction. Alternatively, in the instrumented specifications we can *never* reject the hypothesis that downloads have no effect on sales. There is little evidence in our estimates that file sharing is the main culprit behind the recent decline in CD album sales.

VII. Conclusion

Using detailed records of transfers of digital music files, we find that file sharing has no statistically significant effect on purchases of the average album in our sample. In specifications that identify the effect of file sharing on sales relatively precisely, we reject the hypothesis that file sharing is responsible for the majority of lost sales. This result is plausible given that album sales have increased in the most recent past while the growth of file sharing has continued unabated.[25] Also, a recent internal study at a major record company is reportedly consistent with the results presented here (Economist, 2004).

Although a full explanation for the decline in record sales is beyond the scope of this analysis, several plausible candidates exist. A first factor is poor macroeconomic conditions. While reported incomes had not declined in any year since 1953, they fell in both 2001 and 2002. A second factor is the change in how music is distributed. Between 1999 and 2003 a fifth of music sales shifted from record stores to more efficient discount retailers such as Wal*Mart. Half of the RIAA's reported decline in CD shipments can be linked to the resulting reduction in store inventories. A third factor is the ending of a period of atypically high sales, when consumers replaced older music formats with CDs.

Perhaps more important than these developments in the music industry is the growing competition from other forms of entertainment. A shift in entertainment spending towards recorded movies alone can largely explain the reduction in music sales. The sales of DVDs and VHS tapes increased by over $5 billion between 1999 and 2003. This figure more than offsets the $2.6 billion reduction in album sales since 1999. The shift in spending in part reflects a sharp change in relative prices: since 1999 CD prices increased 10% while DVD prices decreased by

[25]From fall 2003 to summer 2004, albums sales increased by over 5%.

20%, and the price of DVD players fell by 60%. Consumers also spent more on videogames, where spending increased by 40%, or $3 billion, between 1999 and 2003, and on cell phones. Teen cell phone use alone tripled between 1999 and 2003. As a result of the growing competition, there is some indication that consumers spend less time listening to music. For instance, radio listenership rates fell 7% over 1999-2003.

The advent of the new P2P technologies can be considered in a broader context. A key question is how social welfare changes with weaker property rights for information goods. To make such a calculation, we would need to know how the production of music responds to the presence of file sharing. Based on our results, we do not believe file sharing had a significant effect on the supply of recorded music. Our argument is twofold. The business model of major labels relies heavily on a limited number of superstar albums. Although we find some evidence that top albums sell fewer copies as a result of P2P, the economic impact is small, less than 10% of sales even for the most popular releases. It appears unlikely that superstars would exit the industry – or produce albums of lower quality – in response to these changes. On the other hand, our estimates indicate that less popular artists who sell few albums are most likely not affected or perhaps even positively affected by file sharing, leaving the incentives to enter the industry unchanged or perhaps even improved. (Note that while we base these comments on predicted changes in sales, the estimated effects are statistically insignificant for both popular and less successful artists.)

If we are correct in arguing that downloading has had little effect on the production of music – and ignoring possible price effects for complements such as concerts (Krueger, 2004) – we believe that file sharing likely increased aggregate welfare (see also Rob and Waldfogel, 2004). The limited shifts from sales to downloads are simply transfers between firms and consumers. And while we have argued that file sharing imposes little dynamic cost in terms of future production, it has considerably increased the consumption of recorded music. The sheer magnitude of this activity, the billions of tracks which are downloaded each year, suggests the added social welfare from file sharing is likely to be quite high.

The P2P technology available in 2002 had lowered the protection of digital information goods quite drastically. Yet, this reduction apparently did not reduce the legal sales of recorded music,

suggesting there is room for experimentation with weaker property rights for entertainment products. Note, however, that our results do not imply that property rights are unimportant for the production and distribution of music. The impact of future, more powerful technologies that allow consumers to share copyrighted goods at even lower cost is unknown at this time. File sharing was still fairly cumbersome for many consumers in 2002. A majority of households did not have access to broadband and more than half of file transfers in our sample remained unsuccessful. As such, it is a limited experiment that we study in this paper. Whether the production of new music can remain a vibrant business with even weaker property rights than we observe in 2002 is an interesting question for future research.

References

Agentur Lindner (2004). http://www.agentur-lindner.de/special/schulferien/index.html.

Allmusic.com (2003). *All Music Guide*. http://www.allmusic.com.

Arbitron (2004). *Format Trends Report*. http://wargod.arbitron.com/scripts/ndb/fmttrends2.asp.

Arellano, Manuel and Stephen Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies* 58(2): 277-97.

Bakos, Yannis, Brynjolfsson, Erik and Lichtman, Douglas (1999). "Shared Information Goods." *Journal of Law and Economics*. 42: 117-156.

Baum, Christopher, Mark Schaffer, Steven Stillman (2003). "Instrumental Variables and GMM: Estimation and Testing." *Stata Journal*. 3-1: 1-31.

Belden Associates (2003). *Belden Sales and Site Survey Summary: Q1-2 2003 Report*. http://www.beldenassociates.com.

Berry, Steven (1994). "Estimating Discrete-Choice Models of Product Differentiation." *Rand Journal of Economics*. 25: 242-262.

BigChampagne (2004). "Average Simultaneous U.S. Users: August 2002-November 2004." personal correspondence.

Billboard (2002). *Billboard Magazine*. Billboard Pub. Co. Cincinnati.

Blackburn, David (2004). "On-line Piracy and Recorded Music Sales." Harvard University working paper.

Blundell, Richard and Stephen Bond (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics* 87(1): 115-43.

Boldrin, Michele and David Levine (2003). "Perfectly Competitive Innovation." UCLA working paper.

Bresnahan, Timothy, Scott Stern, and Manuel Trajtenberg (1997). "Market Segmentation and the Sources of Rents from Innovation: Personal Computers in the late 1980s." *Rand Journal of Economics*. 28: S17-S44.

Buchanan, J. (2003). "The WinMX Peer Network (WPN)." `http://homepage.ntlworld.com/j.buchanan/`.

CMJ Networks (2002). *CMJ RADIO 200*. Personal communication from Mike Boyle.

comScore Networks (2003). "File Sharing in the comScore Panel." Personal communication from Graham Mudd.

Dotcom Scoop (2001). "Internal RIAA legal memo regarding KaZaA, MusicCity & Grockster." `http://www.dotcomscoop.com/article.php?sid=39`.

Economist (2004). "Music's Brighter Future." 28 October 2004.

Edison Media Research (2003). *The National Record Buyers Study III*. Sponsored by Radio & Records. `http://www.edisonresearch.com`.

Forrester (2002). "Downloads Save the Music Business." `http://www.forrester.com`.

giFT-FastTrack CVS Repository (2003). "The FastTrack Protocol." `http://cvs.berlios.de/cgi-bin/viewcvs.cgi/gift-fasttrack/giFT-FastTrack/PROTOCOL?rev=1.6&content-type=text/vnd.viewcvs-markup`.

Gummadi, Krishna, Richard Dunn, Stefan Saroiu, Steven Gribble, Henry Levy, and John Zahorjan (2003). "Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload," Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP-19).

IEPM (2004). *Internet End-to-end Performance Measurement (IEPM)*. Calculated from

SLAC PingER data available at `http://www-iepm.slac.stanford.edu/`.

IFPI (2002). *Recording Industry in Numbers 2001*. International Federation of Phonographic Industry.

Internet2 Netflow Statistics (2004). *Internet2 NetFlow: Weekly Reports*. `http://netflow.Internet2.edu/weekly/`. Abilene NetFlow Nightly Reports.

Jupiter Media Metrix (2002). "File Sharing: To Preserve Market Value Look Beyond Easy Scapegoats." `http://www.jupiterresearch.com`.

Karagiannis, Thomas, Andre Broido, Nevil Brownlee, kc claffy, and Michalis Faloutsos (2004). "Is P2P dying or just hiding?" Presented at Globecom 2004 in November-December 2004. `http://www.caida.org/outreach/papers/2004/p2p-dying/`.

Keynote (2004). *The Keynote Consumer 40 Internet Performance Index*. `http://www.keynote.com/solutions/performance_indices/consumer_index/consumer_40.html`.

Kish, Leslie (1987). *Statistical Design for Research*. New York: John Wiley & Sons.

Klein, Benjamin, Andres Lerner, and Kevin Murphy (2002). "The Economics of Copyright 'Fair Use' in a Networked World." *American Economics Association: Papers and Proceedings*. 92: 205-208.

Krueger, Alan (2004). "The Economics of Real Superstars: The Market for Rock Concerts in the Material World." http://www.irs.princeton.edu/pubs/pdfs/484.pdf.

Leibowitz, Nathaniel Aviv Bergman, Roy Ben-Shaul, Aviv Shavit (2002). "Are File Swapping Networks Cacheable? Characterizing P2P Traffic." Expand Networks working paper. Presented at the 7th International Workshop on Web Content Caching and Distribution (WCW).

Liebowitz, Stan (1985). "Copying and Indirect Appropriability: Photocopying of Journals." *Journal of Political Economy*. 93: 945-957.

Liang, Jian, Rakesh Kumar, and Keith Ross (2004). "Understanding KaZaA." working paper.

Musikmarkt (2002). "Deutschland Single-Charts." http://musikmarkt.lw-t1.thuecom-medien.de/content/charts/history.php3?jahr=2002.

Nevo, Aviv (2001). "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica*. 69: 307-342.

New Media (2004). "File-sharing Still Soaring, Despite Suits." 13 July 2004. http://69.20.6.242/news2004/Jul04/Jul12/2_tues/news4tuesday.html.

Nielsen//NetRatings (2002). "Deutschland bei Breitbandzugängen nur Mittelmaß." http://www.netratings.com.

Nielsen//NetRatings (2003a). "Broadband Access Grows 59 Percent, While Narrowband Use Declines." http://www.netratings.com.

Neilsen//NetRatings (2003b). "More Than One in Five Surfers Download Music (8 May 2003)." http://www.nielsen-netratings.com/.

Nielsen SoundScan (2003). http://home.soundscan.com/about.html.

Niesyto, Horst (2002). *Digitale Spaltung - digitale Chancen: Medienbildung mit Jugendlichen aus benachteiligenden Verhältnissen*. Mimeo, Pädagogische Hochschule Ludwigsburg.

Nisenholtz, Martin (2002). "The Death of Free Content?" Speech at Jupiter Media Forum 2002. http://nytdigital.com/learn/jupiter_200203181.pdf.

NPD (2003). "RIAA Lawsuits Appear To Reduce Music File Sharing, According To The NPD Group." http://www.npd.com.

Oberholzer-Gee, Felix and Koleman Strumpf (2004). "The Effect of File Sharing on Record Sales: Supplemental Details and Estimates." working paper.

Plant, Arnold (1934). "The Economic Aspects of Copyright in Books." *Economica*. 1: 167-195.

Posner, Richard (2002). "The Law & Economics of Intellectual Property." *Daedalus*. 131,2: 5-12.

RIAA (2003a). *The Recording Industry Association of America's 2003 Yearend Statistics*. http://www.riaa.com.

RIAA (2003b). *RIAA Market Data: The Cost of a CD*. Archived copy from the Internet archive, http://web.archive.org/web/20030416004543/http://www.riaa.com/MD-U.S.-7.cfm.

Rob, Rafael and Joel Waldfogel (2004). "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students." NBER working paper 10874.

Romer, Paul (2002). "When Should We Use Intellectual Property Rights?" *American Economic Review: Papers and Proceedings*. 92: 2. 213-216.

Salton, Gerald (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, Mass: Addison-Wesley.

Sandvine (2003). "Regional Characteristics of P2P: File Sharing As A Multi-Application, Multi-National Phenomenon." White Paper. http://www.sandvine.com.

Shannon, C. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*. 27: 379-423, 623-656.

Shapiro, Carl and Hal Varian (1999). *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard Business School Press.

Statistische Veröffentlichungen der Kultusministerkonferenz (2002). Nummer 162 vom August.

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

Takeyama, Lisa (1997). "The Intertemporal Consequences of Unauthorized Reproduction of Intellectual Property." *Journal of Law & Economics*. 40: 511-22.

Varian, Hal (2000). "Buying, Sharing and Renting Information Goods." *The Journal of Industrial Economics*. 48: 473-488.

Windmeijer, Frank (2000). A finite sample correction for the variance of linear two-step GMM estimators. Institute for Fiscal Studies, IFS Working Papers: W00/19.

Zentner, Alejandro (2004). "Measuring the Effect of Music Downloads on Music Purchases." University of Chicago working paper.

Table 1 – Sample Sales (in thousands) by Category

	Obs	Mean sales	Std dev	Min	Max
Full sample	680	143,096	344,476	74	3,430,264
Catalogue	50	46,833	40,031	219	223,085
Current Alternative	117	118,599	130,257	9,210	785,747
Hard Music Top Overall	19	28,304	22,103	2,945	86,416
Jazz Current	21	21,940	62,522	86	290,026
Latin	21	27,590	35,840	3,143	153,209
New artists	50	15,816	13,635	319	61,673
R&B	144	46,512	67,050	2,151	457,338
Rap	76	39,307	61,278	1,069	324,426
Top Current ("Billboard 200")	83	744,022	710,054	4,092	3,430,264
Top Current Country	66	87,839	130,096	74	669,575
Top Soundtrack	33	44,920	79,264	1,788	318,538

Note: These figures only include sales over our seventeen week observation period. Most of the top-selling albums are classified as "Current" for the purposes of this table

Table 2 – The Geography of File Sharing (numbers in %)

Country	Share of users	Share of downloads	Users in U.S. download from (%)	Users in U.S. upload to (%)	Share World Population	Share World GDP	Share World Internet Users	Software Piracy Rate
United States	30.9	35.7	45.1	49.0	4.6	21.2	27.4	23
Germany	13.5	14.1	16.5	8.9	1.3	4.5	5.3	32
Italy	11.1	9.9	6.1	5.7	0.9	2.9	3.2	47
Japan	8.4	2.8	2.5	1.8	2.0	7.2	9.3	35
France	6.9	6.9	3.8	4.7	1.0	3.1	2.8	43
Canada	5.4	6.1	6.9	7.9	0.5	1.9	2.8	39
United Kingdom	4.1	4.0	4.2	4.2	1.0	3.1	5.7	26
Spain	2.5	2.6	1.8	2.0	0.6	1.7	1.3	47
Netherlands	2.1	2.1	1.9	1.6	0.3	0.9	1.6	36
Australia	1.6	1.9	0.8	2.2	0.3	1.1	1.8	32
Sweden	1.5	1.7	1.8	1.5	0.1	0.5	1.0	29
Switzerland	1.4	1.5	0.9	1.0	0.1	0.5	0.6	32
Brazil	1.3	1.4	1.2	1.3	2.9	2.7	2.3	55
Belgium	0.9	1.2	0.5	1.0	0.2	0.6	0.6	31
Austria	0.8	0.6	0.6	0.4	0.1	0.5	0.6	30
Poland	0.5	0.7	0.7	0.5	0.6	0.8	1.1	54

Notes on country covariates:
Shares of users and downloads is from the file sharing dataset described in the text. All other statistics are from *The CIA World Factbook* (2002, 2003), except the software piracy rates which are from the *Eighth Annual BSA Global Software Piracy Study* (2003). All values are world shares, except the piracy rates are the fractions of business application software installed without a license in the country. All non-file sharing data are for 2002 except population which is for 2003.

Table 3 – Downloads by Genre

	# songs (# albums) in sample	Mean # of downloads	Std dev	Min	Max
			Song level		
All genres	10271	4.645	21.462	0	1258
Catalogue	714	4.361	10.370	0	152
Alternative	1707	7.021	18.153	0	312
Hard	270	4.830	8.684	0	52
Jazz	261	0.333	0.920	0	7
Latin	309	0.550	2.927	0	28
New artists	711	0.609	7.039	0	184
R&B	2249	1.635	7.680	0	159
Rap	1227	0.920	4.887	0	82
Current	1342	17.182	51.286	0	1258
Country	913	1.974	6.382	0	128
Soundtrack	568	1.673	5.301	0	61
			Album level		
All genres	680	70.162	158.628	0	1799
Catalogue	50	62.280	103.114	0	680
Alternative	117	102.436	122.794	0	674
Hard	19	68.632	82.899	0	264
Jazz	21	4.143	4.542	0	13
Latin	21	8.095	26.344	0	121
New artists	50	8.660	33.097	0	229
R&B	146	25.542	56.494	0	433
Rap	77	14.855	24.487	0	119
Current	80	277.807	333.935	2	1799
Country	66	27.303	51.649	0	344
Soundtrack	33	28.788	36.611	0	185

Table 4 – Downloads by Sales – Album Level

	Obs	Mean # of downloads	Std dev	Min	Max	Mann-Whitney
1st quartile: mean 7,235 copies [up to 12,493 copies]	170	11.358	38.472	0	402	- 14.067**
2nd quartile: mean 21,022 copies [up to 31,115 copies]	170	20.929	52.082	0	433	-12.431**
3rd quartile: mean 57.940 copies [up to 100,962 copies]	170	48.088	55.223	0	264	-8.187**
4th quartile: mean 486,184 copies [max 3,430,264 copies]	170	200.270	265.369	0	1799	

Mann Whitney test statistics are for the null that the 4th quartile with the highest sales comes from the same population as the other sales quartiles.
** 1% level of significance

Table 5 – Summary Statistics for Pooled and Panel Models

	# obs	mean (std dev)	min	max
		pooled models		
Weeks since release of album	675	59.206 (174.268)	-33	1689
Median # of misspellings	680	0.042 (0.194)	0	1
Mean track time on album (secs)	673	242.337 (87.285)	102.35	2100
Minimum track time on album (secs)	673	133.744 (93.797)	4	1800
# of songs on album which appear on other album	680	0.794 (2.450)	0	25
# of words of song with fewest words	680	1.156 (0.375)	1	3
		panel models		
German kids on Vacation (million)	10093	9.855 (3.576)	0	12.491
Internet Consumer 40 Performance Index	10093	22.989 (0.739)	21.35	24.1
Internet average Roundtrip time (ms)	10093	159.187 (4.366)	152.638	169.296
Internet std deviation roundtrip time (ms)	10093	33.381 (2.915)	31.022	40.911
Internet2 net flow: % file sharing	10093	28.301 (12.999)	9.916	57.936
Mean album time "other" albums (secs)	9991	243.254 (22.754)	211.442	315.539
Band on tour in Germany	10093	0.003 (0.053)	0	1
# previous releases	10093	6.718 (15.574)	0	194
Billboard rank previous album (calculated as 201 minus rank)	10093	61.136 (82.314)	0	200
Best Billboard rank ever (calculated as 201 minus rank)	10093	83.548 (89.994)	3	200
First release	10093	0.186 (0.389)	0	1
HHI downloads	10093	2460.231 (3671.814)	0	10000
Rank of single on German charts	10093	1.576 (10.268)	0	100
# weeks single is on German charts	10093	0.199 (1.413)	0	24

The precise definition, construction, and intuition for the instruments is provided in Section V.B. of the text.

Table 6 – Pooled Sample - Downloads and Album Sales

	(I)	(II)		(III)	
	Sales	1st stage downloads	2nd stage sales	1st stage downloads	2nd stage sales
# downloads	0.992		0.205		0.176
	(0.182)**		(0.554)		(0.276)
Weeks since release	-0.033	0.013	-0.023	-0.017	-0.023
of album	(0.018)	(0.021)	(0.012)	(0.027)	(0.011)*
Median # of		-50.380		-36.838	
Misspellings		(17.078)**		(15.640)*	
Mean track time				-0.169	
on album				(0.068)*	
Minimum track time				0.216	
on album				(0.078)**	
# of songs which				15.084	
appear on other album				(6.529)*	
# of words of song				-30.710	
with fewest words				(10.250)**	
Genre Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Constant	468.339	280.233	687.044	301.440	695.242
	(66.607)**	(36.853)**	(144.245)**	(38.015)**	(108.017)**
# Observations	675	675	675	673	673
Adjusted R^2	0.58	0.28	0.49	0.34	0.48
Partial R^2 instruments		0.005		0.091	
(Prob F>0)		(0.073)		(0.000)	
Sargan (p-value)				0.240	
Predicted impact on sample sales (1,000s)	47,345		9,786		8,376

Dependent variables are album sales (in 1,000s) and # downloads at the 1st stage. All models include ten Billboard indicators for the musical genre of an album. The Hansen-Sargan overidentification test is for the joint null hypothesis that the excluded instruments are valid, i.e., uncorrelated with the second-stage error term, and that they are correctly excluded from the estimated equation (see Baum, Schaffer and Stillman, 2003). Robust standard errors are in parentheses.

** 1% level of significance * 5% level of significance

Table 7 – Panel Analysis - Downloads and Album Sales

	(I)	(II)	(III)		(IV)	
	Sales	Sales	1st stage downloads	2nd stage sales	1st stage downloads	2nd stage sales
# downloads	1.193	0.281		-0.001		0.012
	(0.022)**	(0.025)**		(0.195)		(0.170)
German kids on Vacation (million)			0.670		0.365	
			(0.054)**		(0.123)**	
Internet Consumer 40 Performance Index					-1.118	
					(0.347)**	
Internet average Roundtrip time					-0.184	
					(0.059)**	
Internet std deviation Roundtrip time					0.135	
					(0.079)	
Internet2 net flow: % file sharing					-0.259	
					(0.069)**	
Mean album time "other" albums					0.124	
					(0.043)**	
Kids vacation × band on tour in Germany					0.452	
					(0.168)**	
Polynomial time trend	Yes	Yes	Yes	Yes	Yes	Yes
Album Fixed Effects?	No	Yes	Yes	Yes	Yes	Yes
Constant	19.199	21.671	4.889	21.888	38.159	21.380
	(5.470)**	(3.753)**	(1.602)**	(3.799)**	(17.647)*	(3.711)**
Observations	10093	10093	10093	10093	9991	9991
Prob $F>0$ on excluded instruments			0.000		0.000	
Sargan test (p-value)					0.247	
R-squared	0.23	0.75	0.74	0.76		0.76
Predicted impact on sample sales (1,000s)	52,488	12,376		-52		507

Dependent variables are album sales (1,000s) and # downloads at the 1st stage. Robust standard errors are in parentheses. For the fixed-effects models, the reported R-squared is the sum of the explained within-variance and the fraction of the variance that is due to the fixed effects. Album-weeks prior to the release date are excluded from the sample.
** 1% level of significance * 5% level of significance

Table 8 – Dynamic Panel Analysis - Downloads and Lagged Album Sales

	(I) 2nd stage sales	(II) 2nd stage sales	(III) GMM Δ sales	(IV) GMM Δ sales
weighted ∑ of 3 weeks of downloads (instrumented)	0.020 (0.094)			
weighted ∑ of 4 weeks of downloads (instrumented)		0.018 (0.094)		
Δ downloads			0.091 (0.091)	
Δ weighted ∑ of 4 weeks of downloads				0.032 (0.134)
lagged sales			1 lag	3 lags
Polynomial time trend?	Yes	Yes	Yes	Yes
Album Fixed Effects?	Yes	Yes	No	No
Constant	59.037 (25.232)*	114.361 (61.542)		
Observations	8649	7982	8649	7321
Arellano-Bond test for AR(1) in first differences: Pr > z			0.221	0.476
Arellano-Bond test for AR(2) in first differences: Pr > z			0.557	0.502
R-squared	0.92	0.93		
Predicted impact on sample sales (1,000s)	2.05×1,591	2.15×1,387	3,598	2.45×2,538

Dependent variables are album sales (1,000s) and # downloads at the 1st stage. The number of downloads is instrumented using the full set of instruments listed in Table 7. The weighted sum of *n* weeks of downloads includes the current week. The weights are chosen in a grid search which minimizes the unexplained fraction of the variance in our models. The standard errors in models (I) and (II) are bootstrapped using 1,000 replications. Models (III) and (IV) use the Generalized Method of Moments estimator developed by Arellano and Bond (1991). Standard errors are corrected using the two-step covariance matrix derived by Windmeijer (2000). Arellano-Bond tests for autocorrelation are applied to the first-difference equation residuals. Second-order autocorrelation would indicate that some lags of the dependent variable which are used as instruments are endogenous. The tests reveal no such problem. Album-weeks prior to the release date are excluded from the sample.
** 1% level of significance * 5% level of significance

Table 9 – Panel Analysis - Downloads and Popularity

Dependent variable	(I) 2nd stage Sales	(II) 2nd stage Sales	(III) 2nd stage Sales	(IV) 2nd stage Sales	(V) 2nd stage Sales	(VI) 2nd stage Sales	(VII) 2nd stage Sales	(VIII) 2nd stage Sales	(IX) 2nd stage Sales	(X) 2nd stage Sales
Variable that is instrumented	d'loads this week	∑ d'loads 4 weeks	d'loads this week	∑ d'loads 4 weeks	d'loads this week	∑ d'loads 4 weeks	d'loads this week	d'loads this week	∑ d'loads 4 weeks	d'loads this week
# downloads	0.035	0.020	0.133	0.092	0.138	0.093		-0.010	0.018	-0.010
(instrumented)	(0.162)	(0.095)	(0.162)	(0.098)	(0.167)	(0.100)		(0.167)	(0.095)	(0.173)
# downloads ×	-0.005	-0.000								
# previous releases	(0.007)	(0.002)								
# downloads ×			-0.002	-0.001						
BB rank last album			(0.002)	(0.000)*						
# downloads ×					-0.002	-0.001				
best BB rank ever					(0.001)	(0.000)*				
# downloads (instr)							0.508			
1st spline							(0.558)			
# downloads (instr)							0.336			
2nd spline							(0.807)			
# downloads (instr)							-0.685			
3rd spline							(0.553)			
# downloads (instr)							-0.068			
4th spline							(0.217)			
# downloads ×								0.054	-0.004	
first release								(0.172)	(0.095)	
HHI downloads										0.063
/ 1,000										(0.098)
# downloads ×										-0.001
HHI / 1,000										(0.020)
Polynomial time trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Album Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	22.081	114.347	22.035	116.028	22.063	115.394	21.599	22.027	114.352	21.867
	(4.412)**	(61.546)*	(4.049)**	(61.523)	(4.236)**	(61.530)	(4.797)**	(4.380)**	(61.547)	(3.834)**
Observations	9991	7982	9991	7982	9991	7982	9991	9991	7982	9991
R-squared	0.76	0.94	0.76	0.94	0.76	0.94	0.76	0.76	0.94	0.76
H_0: main effect+inter-action = 0 (Prob > F)	0.865	0.834	0.483	0.353	0.486	0.356		0.869	0.904	0.944
Predicted impact on sample sales (1,000s)	928	2.15× 1,395	392	2.15× 1,224	83	2.15× 832	5,265	1	2.15× 1,457	-564

Dependent variables are album sales (1,000s). In odd-numbered models, we instrumented for the # downloads at the 1st stage using the full set of instruments from Table 7. In even-numbered models, the weighted sum of 4 weeks of downloads is the instrumented variable (see Table 8 for details). Billboard ranks (BB ranks) are coded as 201 minus the actual rank. For model (VII), predicted downloads are used to calculate the predicted impact on sample sales (this is done to ensure the album is assigned to the same spline segment in both the estimates and in the fitting exercise). In models (IX) and (X), the HHI is included in the first stage. HHI is computed using the share of downloads for each song on the album during each week. Standard errors are bootstrapped using 1,000 replications. Album-weeks prior to the release date are excluded from the sample.
** 1% level of significance * 5% level of significance

Table 10 – Robustness Check with Scaled Downloads – Testing the "Drop-out" Hypothesis

	(I)		(II)		(III)	(IV)
	1st stage downloads	2nd stage sales	1st stage downloads	2nd stage Sales	2nd stage sales	GMM
Scaled downloads		-0.001		0.020		
		(0.152)		(0.131)		
Scaled and weighted					0.025	
∑ of 4 downloads					(0.070)	
Δ scaled and weighted						0.054
∑ of 4 downloads						(0.118)
German kids on	0.857		0.513			
Vacation (million)	(0.073)**		(0.165)**			
Internet Consumer 40			0.189			
Performance Index			(0.058)**			
Internet average			-1.572			
Roundtrip time(ms)			(0.467)**			
Internet std deviation			-0.256			
roundtrip time (ms)			(0.079)**			
Internet2 net flow:			0.148			
% file sharing			(0.107)			
Mean album time			-0.315			
"other" albums			(0.093)**			
Kids vacation × band			0.572			
on tour in Germany			(0.227)*			
Lagged sales						3 lags
Constant	6.531	21.889	48.383	21.376	117.649	
	(2.157)**	(3.801)**	(23.758)*	(3.709)**	(60.874)	
Observations	10093	10093	9991	9991	7982	7321
Sargan test (p-value)			0.246			
AB test for AR(1)						0.503
AB test for AR(2)						0.492
R-squared	0.75	0.76	0.77	0.76		
Predicted impact on sample sales (1,000s)		-55		1,145	2.15×2,794	2.45×5,263

Dependent variables are album sales (1,000s) and scaled downloads at the 1st stage. Downloads are scaled to reflect the growth of KaZaA users over the sample period. We use 22 data points on the number of KaZaA users in the period from 9/9/2002 to 2/4/2003 to fit a fractional polynomial trend in the number of users. The model explains 85% of the variation. The weights for the sum of 4 weeks of downloads are identical to the weights in Table 8. For notes on the GMM model, see also Table 8. For the fixed-effects models, the reported *R*-squared is the sum of the explained within-variance and the fraction of the variance that is due to the fixed effects. Album-weeks prior to the release date are excluded from the sample

** 1% level of significance * 5% level of significance

Table 11 – Robustness Checks

	(I)	(II)	(III) week FE		(IV) week FE	
	w/o holiday sales	w/o singles	1st stage downloads	2nd stage sales	1st stage downloads	2nd stage sales
# downloads	0.139	0.012		0.277		0.184
(instrumented)	(0.257)	(0.170)		(0.881)		(0.253)
Kids vacation × band			0.457		0.460	
on tour in Germany			(0.167)**		(0.168)**	
Mean album time					0.126	
"other" albums					(0.043)**	
Rank of single on					0.014	
German charts					(0.002)**	
# weeks single is on					-0.028	
German charts					(0.011)*	
Polynomial time trend	Yes	Yes	No	No	No	No
Week fixed effects	No	No	Yes	Yes	Yes	Yes
Album Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Constant	17.621	21.380	1.365	12.598	-29.246	12.861
	(5.481)**	(3.711)**	(0.397)**	(1.549)**	(10.441)**	(1.029)**
Observations	7321	9991	10093	10093	9991	9991
Prob $F>0$ on excluded instruments	0.000	0.000	0.000		0.000	
Sargan test (p-value)	0.785	0.247			0.257	
R-squared	0.85	0.76	0.75	0.76	0.75	0.76
Predicted impact on sample sales (1,000s)	4,186	507		12,198		8,079

Dependent variables are album sales (1,000s) and # downloads at the 1st stage. Model (I) excludes the last four weeks of data (December sales) to see if the holiday shopping season influences our results. Model (II) excludes downloads for hit singles from the analysis. A hit single is a song that appeared either on the "Billboard Top Ten Singles" list, the "Billboard Top 50 Airplay" or the MTV Video Playlist. In specifications (I) and (II), downloads are instrumented using the full set of instruments from Table 7. Robust standard errors are in parentheses. The reported *R*-squared is the sum of the explained within-variance and the fraction of the variance that is due to the fixed effects. Album-weeks prior to the release date are excluded from the sample

** 1% level of significance * 5% level of significance

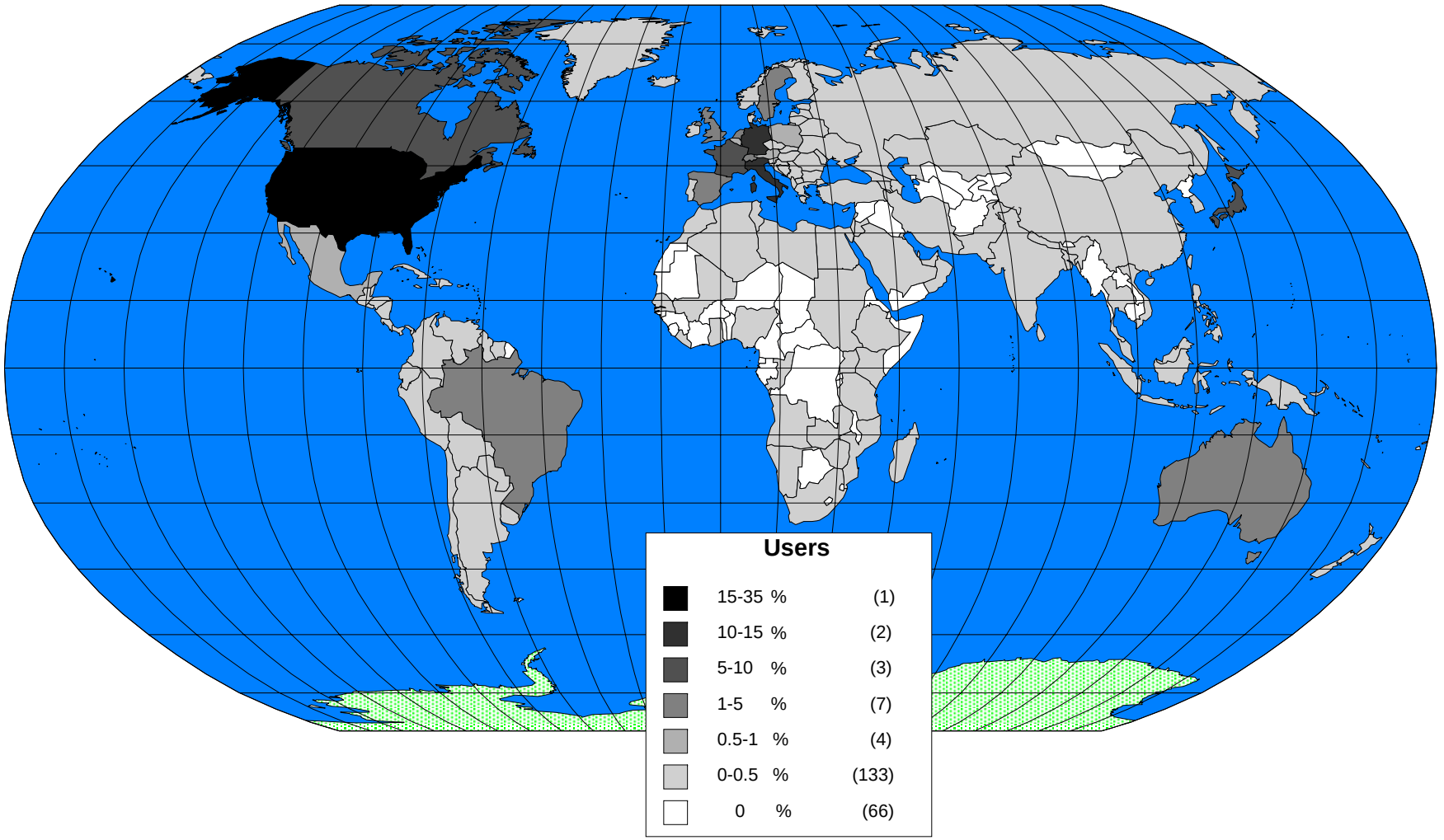

Figure 1: Distribution of Users (Unique log-ins) by Country

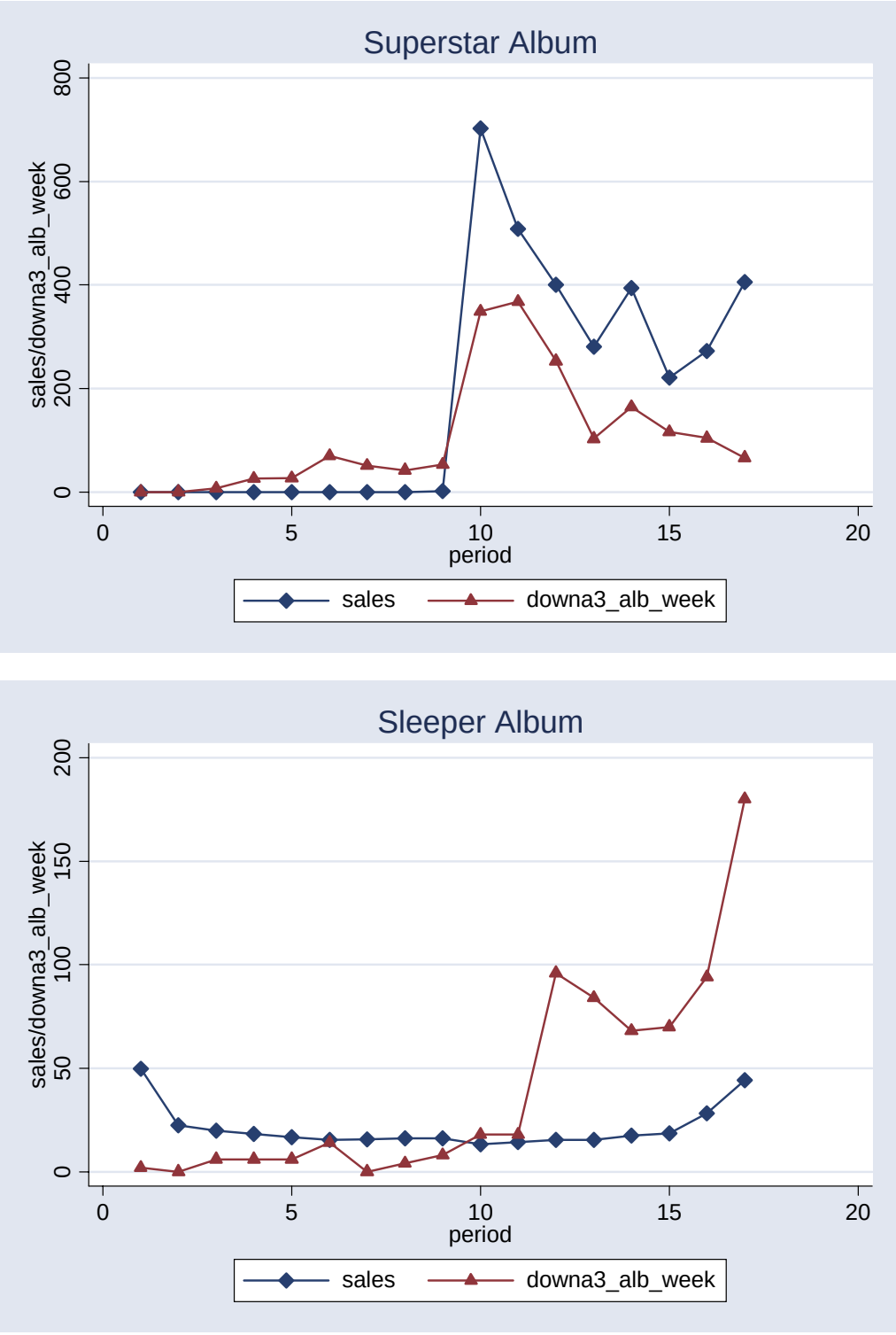

Figure 2: Dynamics of Downloads and Albums Purchases
(by week, sales in thousands)

Additional Material for Reviewers
(not intended for publication)

Appendix A: Data Issues

A. Validity of Data Sample

Our inferences about the effect of file sharing on record sales would be invalid if we had an unrepresentative sample of downloads. However, there are several reasons why this should not be true. We first discuss the intuition for why we expect our downloads to be representative and then present quantitative evidence on this point.

First, our data overlap with that on larger networks. The OpenNap network is largely composed of clients which are simultaneously accessing WinMX and FastTrack/KaZaA (In our data roughly a third of the clients use the WinMX software. These users simultaneously log into and search both the WinMX and OpenNap networks. About a tenth use mldonkey which allows for simultaneous searches of FastTrack/KaZaA, eDonkey and OpenNap.). WinMX and FastTrack/KaZaA were the two largest file sharing communities among U.S. users during the study period. According to comScore Networks, which tracks the on-line behavior of over one million representative Internet users, roughly one-fifth of the active file sharing home computers in the U.S. during our sample period used the WinMX software. The KaZaA share of users was about two-thirds (comScore Networks, 2003). These networks also have a similar relative share of Internet2 backbone traffic over November-December 2002 (authors' calculations based on Internet2 Netflow Statistics, 2004) as well as of North American bandwidth use (Sandvine, 2003). Also, the main text points out that WinMX has a substantial share of world file sharing.

Second, the technical nature of searching and downloading is similar across the main networks. For example the WinMX network architecture is quite similar to the larger FastTrack/KaZaA network, with user nodes sending search requests through one of a large number of super-nodes spread throughout the network.[26] OpenNap has a similar structure as these hybrid networks. Multiple OpenNap servers are often linked together in a sub-network. This architecture allows a client to interact with those logged onto another server in the sub-network, much as they do on WinMX and FastTrack/KaZaA. One of our servers was part of a sub-network of servers.[27] Another feature promoting similarity with hybrid P2P is the presence of independent OpenNap sub-networks. Since most clients simultaneously log into multiple sub-networks, users search across and download from multiple servers In addition, the user experience is comparable in the different networks. In all cases the user first logs in, then enters text into a search box to locate

[26]In both networks, the super-nodes (or primary connections in the WinMX parlance) typically host roughly a hundred user peers. The super-nodes are inter-connected, and a user's search requests are propagated only to users on a few nearby super-nodes. That is, not all files available on the overall network are available under either KaZaA or WinMX. For additional details, see Buchanan (2003), Dotcom Scoop (2001), giFT-FastTrack CVS Repository (2003), and Liang, et al (2004).

[27]There were on average seven servers on the network which had a devoted hub to handle server-to-server communications. As with the hybrid P2P, searches were passed to all servers and downloads occur directly between clients. Our records include all searches on the network and all downloads where at least one user is logged onto our server.

files, and downloads files directly from another peer/client. Downloads speeds appear to be relatively similar.

Third, the effective size of the networks are comparable. This is important because of the possibility of network externalities, e.g. larger networks should make rarer files easier to find. While KaZaA nominally has millions of users, the hybrid P2P architecture means each user only has access to the files of about five thousand users.[28] This is near the average user base of our server which is on the sub-network. And since most OpenNap users are simultaneously logged into multiple servers, the set of files available to users is quite large (in many cases the entire OpenNap network).

Fourth, we explicitly compared song availability on our OpenNap servers with the FastTrack/KaZaA network. Each week during the second half of our sample period, we recorded the number of available copies of 15-20 songs drawn from currently popular tracks on the Billboard 100 (Billboard, 2002), recently released "indie" albums on the CMJ chart (CMJ Networks, 2002), and upcoming releases. To ensure comparability, the networks were searched simultaneously. The correlation coefficient is 0.62 over the whole sample (N=144) indicating that the availability of common and rare songs move in tandem in the two networks.[29]

Fifth, we considered whether our most popular downloads were also common in other file sharing networks. To do this, we compared the top ten downloads each week in our data with the concurrent list from `http://www.bigchampagne.com`. BigChampagne generates their own weekly top lists, a "TopSwaps" index which is based on a variety of activities (searches, shared files, and downloads) which are passively monitored on the leading P2P networks. Over our seventeen week sample period, two-thirds of our top ten downloads also appear in the BigChampagne top ten list.

The final piece of evidence is the most convincing. We received a large sample of downloads on FastTrack/KaZaA from a P2P caching firm, Expand Networks (Leibowitz, et al., 2002).[30] This allows us to directly compare whether our sample of downloads is comparable to that on FastTrack/KaZaA using the standard test of homogeneity. Our two samples each include over twenty-five thousand downloads, and we are able to identify 1789 unique tracks. The resulting Pearson χ^2 statistic is 1824.1. This indicates that we cannot reject a null that both were drawn from the same population with almost any confidence level.

B. Scale-Effects in Downloading

An important question is whether the size of a file sharing network influences the type of music which is downloaded. For example, one might argue that larger networks allow individuals to find rarer tracks which are unavailable on smaller networks. We make two arguments that this concern is not a serious barrier. First, it is important to recall that even our relatively small

[28]In KaZaA one to two hundred peers connect to a super-node, which in turn is connected to about twenty-five other super-nodes (see the last three references in the last note).

[29]The correlations are also large and positive for each of the three categories of albums in the sample.

[30]As with the OpenNap data, the file sharers in the Expand sample were unaware that their actions were being monitored. The data was collected during January-February 2003, which we matched to records from one of our OpenNap servers.

OpenNap networks are effectively as big as the larger FastTrack/KaZaA or WinMX (see Section A). This is because hybrid P2P limits the effective set of users one can search to a small subset of the entire network (see the discussion in the last sub-section).

A second piece of evidence comes from our data. We have observations from two servers, one which is part of a network of other servers and another which is standalone and has a user base which is roughly an order of magnitude smaller. If there are scale-effects, then the distribution of downloads should be different on the two servers. Looking at the distribution for the 680 albums over all weeks, the resulting Pearson χ^2 statistic is 737.21. We cannot reject the null of homogeneous distributions at the 95% confidence level.

Appendix B: Model

A. Setup

Consider a stylized model of downloading and purchase behavior. Suppose that each individual values music but faces some acquisition costs. There is population heterogeneity in these values and costs. Individuals first decide whether to download and then later whether to purchase.

In particular, let:

- $V_{ij} \geq 0$ be the value of purchased album $i = 1,...,N$ for individual $j \in \mathbb{R}^+$.
- $D_{ij} \equiv \gamma V_{ij}$ be the value of downloaded album i for individual j. Presumably $0 \leq \gamma \leq 1$ since downloads are inferior to the original album (lower sound quality, no liner notes, and perhaps remorse at not compensating the artist) though all that is needed is $\gamma \geq 0$.
- $p > 0$ be the cost of a purchased album (presumed to be constant since album prices rarely vary)
- $q_{ij} > 0$ be the monetized cost of downloading album i for individual j. This cost stems from time spent searching for and downloading the album. q_{ij} varies across individuals (due to different value of time or the speed of Internet connection) and albums (since some albums are longer and hence take more time to download).

Preferences are assumed to be separable over the goods. Given a single outside good which serves as the numeraire, after substituting the budget constraint the utility function of individual j is,

(A1) $U_j = \sum_i \mathbb{1}_{ij}(\text{purchase}) \cdot (V_{ij} - p) + \mathbb{1}_{ij}(\text{download}) \cdot (\gamma V_{ij} - q_{ij})$

where $\mathbb{1}_{ij}(.)$ is an indicator that the individual bought or downloaded album i.

Individuals face a sequence of discrete choices. First they must decide whether to download any of the albums, and then whether to purchase any of them (the discount factor is near unity since these decisions occur at nearly the same time). These are discrete choices in that each album can be downloaded or purchased once or not at all.

We presume the values of the albums and the costs of downloads are independent. The population density of values for album i is $V_i \sim f(V_i, \alpha_{Vi})$ and the population distribution is $F(V_i, \alpha_{Vi})$. The population density of costs for album i is $q_i \sim g(q_i, \alpha_{qi})$ and the population distribution is $G(q_i, \alpha_{qi})$. The $\alpha_{\cdot}$ terms parameterize the distributions. α_{Vi} measures the popularity of an album which is viewed in terms of first order stochastic dominance: $F(V, \alpha_{VA}) \leq F(V, \alpha_{VB})$ (with a strict inequality for at least one V) when $\alpha_{VA} > \alpha_{VB}$. That is, albums with higher values of α_{Vi} are more valuable in aggregate or equivalently their population distribution is shifted to the right. α_{qi} measures the cost of downloading an album and is defined analogously: $G(q, \alpha_{qA}) \leq G(q, \alpha_{qB})$ (with a strict inequality for at least one q) when $\alpha_{qA} > \alpha_{qB}$.

B. Preliminary Result

To fix ideas, we first consider the case where preferences are independent across downloads and purchases. That is, we ignore the possibility of crowd-out or learning. From (A1) an individual purchases *iff* $V_{ij} > p$ and downloads *iff* $\gamma V_{ij} > q_{ij}$, and so aggregate values are,

(A2) Total Purchases of album $i \equiv \int_{q>0}(1-F(p,\alpha_{Vi}))g(q,\alpha_{qi})dq = 1-F(p,\alpha_{Vi})$

(A3) Total Downloads of album $i \equiv \int_{q>0}(1-F(q/\gamma,\alpha_{Vi}))g(q,\alpha_{qi})dq$

These equations yield the first result.

Result 1. *More popular albums have higher total downloads and total purchases, even if there is no feedback between purchases and downloads.*

Proof:
Consider album A and a less popular album B, $\alpha_{VA}>\alpha_{VB}$, which both have the same cost distribution, $\alpha_{qA}=\alpha_{qB}\equiv\alpha_q$. From (A2),

(A4) Purchases(A) – Purchases(B) = $F(p,\alpha_{VB})-F(p,\alpha_{VA}) > 0$

where the inequality follows from first order stochastic dominance. From (A3),

(A5) Downloads(A) – Downloads(B) = $\int_{q>0}(F(q/\gamma,\alpha_{VB})-F(q/\gamma,\alpha_{VA}))g(q,\alpha_q)dq > 0$

where the inequality again follows from first order stochastic dominance.

□

This highlights the problem with simply regressing downloads on purchases: both are endogenously determined by popularity, so OLS will yield a spurious positive relationship.

C. Main Model

More generally downloads should influence purchases (we continue to presume there is no spillover between albums). The effect of downloads is modeled as a shift in the α_{Vi}:

(A6) $\alpha'_{Vi} \equiv \alpha_{Vi}$ following a download = $\phi(\alpha_{Vi})$

where $\phi(.)$ is a weakly monotone increasing function, $\alpha_{VA}>(<)\alpha_{VB} \rightarrow \phi(\alpha_{VA})>(<)\phi(\alpha_{VB})$. (A6) allows downloads to increase or decrease the popularity of an album (and hence purchases), and for this effect to vary by the ex ante popularity: $\alpha'_{Vi}\geq\alpha_{Vi}$ or $\alpha'_{Vi}\leq\alpha_{Vi}$ and this relationship may vary with the level of α_{Vi}. The only restriction is that downloading does not change the ranking of album popularity, e.g. $\phi(.)$ is an order-preserving function.

A <u>modified definition of album popularity</u> is also used: when $\alpha_{VA}>\alpha_{VB}$, then we presume $f(V,\alpha_{VA})\geq f(V,\alpha_{VB})$ (with a strict inequality for at least one V) $\forall V\geq p$. That is, a more popular album (with a higher α_{Vi}) has a greater mass of individuals at every value which could lead to purchases. More popular albums have a thicker right tail in their density of values. This is typically a stronger condition on the density than stochastic dominance.

We presume individuals download myopically. That is, they do not take into account the potential for learning (the shift from α_{Vi} to α'_{Vi}) when making their downloading decision.

The positive correlation of purchases and downloads from Result 1 still holds in this more general framework. For example consider albums A and B with $\alpha_{VA}>\alpha_{VB}$ and $\alpha_{qA}=\alpha_{qB}\equiv\alpha_q$ The change in download equation (A5) in the proof of Result 1 is unaffected. The change in purchases equation is,

(A7) Purchases(A) – Purchases(B) | Downloads have feedback

$$= \int_{V>p}((f(V,\phi(\alpha_{VA}))-f(V,\phi(\alpha_{VB})))G(\gamma V,\alpha_q)+(f(V,\alpha_{VA})-f(V,\alpha_{VB}))(1-G(\gamma V,\alpha_q)))dV > 0$$

where the first term is for individuals who download ($\gamma V_{ij}>q_{ij}$) and the second is for those who do not download ($\gamma V_{ij}<q_{ij}$). The inequality follows from the modified definition of popularity and the monotonicity of $\phi(.)$. Again the intuition is that album popularity drives both downloads and purchases.

The main objective of the paper is to understand the shape of $\phi(\alpha_{Vi})$, which shapes the effect of downloads on purchases. This cannot be measured from simply regressing downloads on purchases due to the positive correlation result. Instead it suggests using instruments, variables which shift downloads but have no direct effect on purchases. A natural instrument is the download costs parameter, α_{qi}.

Result 2. *Download costs influence purchases only though their effect on downloads. Download costs reduce album downloads.*

Proof:
Consider album A and a more costly to download album B, $\alpha_{qA}>\alpha_{qB}$, which both have the same popularity distribution, $\alpha_{VA}=\alpha_{VB}\equiv\alpha_V$. From (A3),

(A8) Downloads(A) – Downloads(B)

$$= \int_{q>0}(g(q,\alpha_{qB})-g(q,\alpha_{qA}))F(q/\gamma,\alpha_V)dq$$
$$= -\gamma^{-1}\int_{q>0}(G(q,\alpha_{qB})-G(q,\alpha_{qA}))f(q/\gamma,\alpha_V)dq < 0$$

where the second equality is from integration by parts and the inequality again follows from first order stochastic dominance. After separately integrating the downloading and non-downloading populations, the change in purchases equation is,

(A9) Purchases(A) – Purchases(B) | Downloads have feedback

$$= \int_{V>p}(G(\gamma V,\alpha_{qA})-G(\gamma V,\alpha_{qB}))(f(V,\phi(\alpha_V))-f(V,\alpha_V))dV$$

In the absence of feedback effects, $\phi(\alpha_V)=\alpha_V$, purchases are identical for the two albums (or simply see (A2)).

□

Asides:

- While the proof compares two albums, the equations can equivalently be interpreted as a comparison of the same album at two moments in time when its cost of downloading differ.
- After allowing for feedback, higher download costs increases (decreases) purchases *iff* downloading decreases (increases) album sales. That is, (A9) is positive *iff* $\phi(\alpha_V)<\alpha_V$ (this follows since costs are increased--so the first term in the integral is negative—and an application of the modified popularity definition—so the second term is negative when $\phi(\alpha_V)<\alpha_V$).

Result 2 shows download cost shifters are appropriate instruments. A cost drop increases downloads and (*iff* the feedback effect from downloads is positive) increases purchases. The

opposite holds for a cost hike. With enough data we can ascertain the shape of $\phi(\alpha_V)$ for a wide range of popularity levels.

D. Functional Form for the Estimation Equation

A final issue is the appropriate functional form for the estimates. We argue that a linear equation relating aggregate sales to downloads is appropriate. To see this, we first write the expressions for downloads and purchases of some album,

(A10) $\text{Downloads} = \int_{V>0} f(V,\alpha_V)G(\gamma V,\alpha_q)dV$

and,

(A11) $\text{Purchases} = (1-F(p,\alpha_V)) + \int_{V>p}(f(V,\phi(\alpha_V))-f(V,\alpha_V))G(\gamma V,\alpha_q)dV$

These can be can be combined to give,

(A12) Purchases

$$= (1-F(p,\alpha_V))+\int_{V>p} f(V,\phi(\alpha_V))G(\gamma V,\alpha_q)dV+\int_{0>V>p} f(V,\alpha_V)G(\gamma V,\alpha_q)dV-\int_{V>0} f(V,\alpha_V)G(\gamma V,\alpha_q)dV$$

$$\equiv \text{Purchases}_{\text{NoDownloads}}(p,\alpha_V) + \Psi(p,\gamma,\alpha_V,\phi(\alpha_V),\alpha_q) - \text{Downloads}(\gamma,\alpha_V,\alpha_q)$$

The first term on the bottom row measures total purchases in the absence of downloads, and is independent of the download cost parameter α_q. The remaining two terms reflect the effect of downloads. (A12) shows that it is roughly appropriate to use a linear specification in the estimates. It also highlights our instrument strategy. An exogenous shift in the distribution of download costs, as measured by α_q, influences downloads and, recalling the discussion after Result 2, will increase or decrease purchases based on the shape of $\phi(\alpha_V)$.