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

Felix Oberholzer
Harvard Business School
foberholzer@hbs.edu

Koleman Strumpf
UNC Chapel Hill
cigar@unc.edu

March 2004

Abstract

A longstanding economic question is the appropriate level of protection for intellectual property. The Internet has drastically lowered the cost of copying information goods and provides a natural crucible to assess the implications of reduced protection. We consider the specific case of file sharing and its effect on the legal sales of music. A dataset containing 0.01% of the world's downloads is matched to U.S. sales data for a large number of albums. To establish causality, downloads are instrumented using technical features related to file sharing, such as network congestion or song length, as well as international school holidays. Downloads have an effect on sales which is statistically indistinguishable from zero, despite rather precise estimates. Moreover, these estimates are of moderate economic significance and are inconsistent with claims that file sharing is the primary reason for the recent decline in music sales.

*We thank Shane Greenstein and participants at the 2004 AEA meeting for valuable comments and suggestions. We also acknowledge Sarah Woolverton for her tireless efforts to improve the quality of our song matching algorithm and Christina Hsiung Chen for research assistance. The CMJ Network, Nathaniel Leibowitz, and Nevil Brownlee generously provided us with auxiliary data. 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 gratefully acknowledged.

I. Introduction

File sharing has become one of the most common on-line activities. File sharing occurs in networks which allow individuals to share, search for, and download files from one another. A key property of these networks is that sharing files is largely non-rivalrous because the original owner retains his copy of a downloaded file. This makes the cost of sharing quite low. Moreover, there are network externalities, since more individuals imply a greater selection of files.

These features fueled the dramatic growth of file sharing, particularly of copyrighted music recordings. While few participated in file sharing prior to 1999 (the founding year of the now defunct Napster), there were more than three million simultaneous users sharing over a half a billion files on the most popular network (FastTrack/KaZaA) in 2003. Each week there are more than one billion downloads of music files alone. Participation in file sharing has also grown. Over 60 million Americans above the age of twelve have downloaded music (Ipsos-Reid, 2002b). File sharing is heavily skewed to youth. While a majority of Americans under eighteen have downloaded and half of those are heavy users, only a fifth of those aged 35-44 have downloaded files (Edison Media Research, 2003). Among U.S. adults at least eighteen years old, the number of down-loaders has about doubled since 2000 (Pew Internet Project, 2000 and 2003). Because physical distance is largely irrelevant in file sharing, individuals from virtually every country in the world participate.

There is tremendous interest in understanding the economic effects of file sharing. As file sharing becomes easier and faster, a greater variety of information goods, including movies and software, are likely to be downloaded. The effects of such downloads are likely to parallel the experience to date with sales of recorded music. According to the RIAA (2002), the number of CD's shipped in the U.S. fell from 940 million to 800 million--or 15%--between 2000 and 2002 (though shipments continued to rise during the

first two years of popular file sharing, 1999-2000). The record industry has claimed this decline is due to file sharing.[1]

Such causality, however, is unclear. While file sharing significantly reduces the financial cost of obtaining music, it has an ambiguous theoretical effect on record sales. Participants could substitute downloads for legal purchases, thus reducing sales. 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 to discuss music in file server chat rooms. This learning may promote new sales. Other mechanisms have ambiguous effects. Individuals may use file sharing to sample music, which will increase or decrease sales depending on whether they like what they hear. 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 because music tracks have gained a new use, sharing with others. 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. That is, file sharing primarily serves to increase total music consumption.[2]

With no clear theoretical prediction, the effect of file sharing on sales is an empirical question. To address this topic, one route is to ask individuals how downloading influences their purchase behavior. In an on-line survey of actual file sharers, users

[1]These quotes, from the heads of the main industry lobbies, broadly summarize the record labels' 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" (Cary Sherman, RIAA president, *USA Today*, 18 September 2003).

> "Internet piracy means lost livelihoods and lost jobs, not just in record companies but across the entire music community. For those who think the 10.9% first half sales fall in 2003 does not speak for itself, look at the other evidence. Artist rosters have been cut, thousands of jobs have been lost, from retailers to sound engineers, from truck drivers to music journalists." (Jay Berman, IFPI chairman, IFPI Network Newsletter, December 2003).

[2]Many of these issues have been broadly discussed in the literature. File sharing might also independently change revenue through its influence on prices (see Bakos, et al. 1999, Takeyama, 1994, and Varian 2000).

acknowledged both crowd-out and learning effects.[3] While 65% of users say downloading led them to not purchase an album, 80% claim they bought at least one album after first sampling it on a file sharing network. The net effect is reported to be positive. According to the survey, file trading led the average user to purchases an additional 8 albums. While these results are suggestive, there is a concern that users might overstate their additional purchases to make their file sharing behavior appear more favorable.

Rather than relying on surveys, this study uses observations of actual file sharing behavior to assess the impact of downloads on sales. We analyze a large file sharing dataset which includes 0.01% of the world's downloads from the last third of 2002. We focus on users located in the U.S. Their audio downloads are matched to the album they were released on, for which we have concurrent U.S. weekly sales data. This allows us to consider the relationship between downloads and sales. To establish causality, we instrument for downloads using technical features related to file sharing (such as network congestion or song length) and international school holidays, both of which are plausibly exogenous to sales. We are able to obtain relatively precise estimates because the data contain over ten thousand album-weeks.

We find that file sharing has only had a limited effect on record sales. OLS estimates indicate a positive effect on downloads on sales, though this estimate has a positive bias since popular albums have higher sales and downloads. After instrumenting for downloads, most of the impact disappears. This estimated effect is statistically indistinguishable from zero despite a narrow standard error. The economic effect is also small. Even in the most pessimistic specification, five thousand downloads are needed to displace a single album sale. We also find that file sharing has a differential impact across sales categories. For example, high selling albums actually benefit from file sharing. 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

[3]This survey was conducted on a file sharing server, described in more detail later in the paper, over 11/23/02-12/2/02. 159 users completed the survey. To the best of our knowledge this is the first survey conducted while individuals are engaged in downloading, so the appropriate population is targeted.

likely individuals who would not have bought the album even in the absence of file sharing.

Our results have broader applications beyond the specific case of file sharing. A longstanding question is whether strong protection for intellectual property is necessary to ensure innovation. Economic research on the relevant role for 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). This point is also linked to new growth theory where information spillovers from innovation have a central role (Romer, 1990). A key question in this literature is the extent to which diminishing protection reduces the returns for the initial innovator. We provide specific evidence on this point for the case of a single industry, recorded music. File sharing markedly lowers the protection which copyrighted music recordings enjoy, so the impact on sales is a natural test of the need for protecting intellectual property.

The outline of the remaining 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 the econometric approach and identification strategy are discussed. Section VI presents the results, and the last section discusses the implications of this work. Appendix A provides evidence that our sample of downloads is representative of the overall universe of downloads, and Appendix B presents a model of downloads and purchases which underlies our econometric strategy.

II. The Literature

Empirical research on file sharing and record sales has been inconclusive, primarily, we believe, due to data limitations.[4] The leading study to date is Liebowitz (2003). Liebowitz tries to explain annual trends in national sales using a wide variety of possible factors including the macro-economy, demographics, changes in recording format and

[4]A related empirical literature examines the incentives for contributing to internet-based public goods and the resulting free-rider problem (Dempsey, et al., 1999; Adar and Huberman, 2000).

listening equipment, prices of albums and other entertainment substitutes, and changes in music distribution. He finds these factors cannot fully explain the decline in sales from 1999-2002 and therefore concludes that file sharing has reduced aggregate sales. By gauging the effect of other factors, Liebowitz (2003) helps to put bounds on the potential negative effect of file sharing on sales. Our paper complements this aggregate analysis because it uses micro-level, panel data (the sale and downloads of particular albums) to make relatively precise estimates of the impact of file sharing on music purchases.

Another set of papers uses phone surveys or Internet panels to determine if individuals who download also purchase fewer music albums.[5] A general difficulty with these studies is that they do not consider the appropriate counterfactual, namely purchase behavior in the absence of file sharing. While down-loaders may purchase fewer records, this could simply reflect a lower willingness to pay which would always lead such individuals to purchase fewer records. An additional problem is the accuracy and the population sample of the data. Those who agree to have their Internet behavior discussed or monitored are unlikely to be representative of all Internet users.[6]

A third approach is to see how geographic variability in correlates of downloading, such as the availability of high-bandwidth Internet access, influences record sales at local stores (Fine, 2000). Unfortunately, such correlates also allow for easier access to on-line purchases of albums which will not be reflected in the local sales data.

[5]These are primarily industry studies which have mixed conclusions about the effect of file sharing. These surveys include Pew Internet Project (2000), Forrester (2002), IFPI (2002), Ipsos-Reid (2002a), Jupiter Media Metrix (2002), Edison Media Research (2003), Neilsen//NetRatings (2003). Liebowitz (2002) reviews and critiques earlier industry studies used in the Napster trial (*A&M Records, Inc., et al. vs. Napster, Inc.*).

[6]With phone data individuals are likely to incorrectly self-report their downloading, since it is currently considered illegal. Internet panels rely on individuals who willingly agree to have all of their internet behavior monitored, and such individuals are not likely to be representative of those who engage in illegal behavior. Our survey of file sharers discussed in the introduction mitigates this sample selection.

A recent academic paper, Zentner (2003), uses a mail survey. Unfortunately, the sample omits a crucial demographic (those under 16 years old, who are among the most active users of file sharing and heaviest purchasers of music) and does not contain information about the intensity of downloads or music purchases (which makes it difficult to draw inferences about the total impact of file sharing on record sales). The data itself is also subject to the criticisms of phone surveys listed above.

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

III. File sharing Networks

This section provides background on the basic mechanics of file sharing. 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. Clients, servers, and peers are connected in peer-to-peer (P2P) networks. In our discussion we refer to individuals on P2P as users.

Figure 1 illustrates the three basic P2P architectures. A *centralized P2P network* has individual clients log into a central server. The server serves much like an Internet search engine in that it keeps a real-time index of all files being shared and handles all search requests from clients (the server does not store files, but only maintains their characteristics and host client). 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 (the server plays no role in the transfer). This is the structure of Napster and its open-source descendant, OpenNap. A *decentralized P2P network* has no central server, and every node acts a peer. Each peer is connected to some small number of other peers, and some set of connections interconnect any peer pair. A peer's search requests are sent to neighboring peers which in turn propagate it to their neighbors (the request terminates after some number of hops). Positive matches are sent back though the intermediate peers, though transfers occur directly between the nodes as with centralized P2P. This is the structure of Gnutella and Freenet. A *hybrid P2P network* is an intermediate case. A few nodes are designated as super-nodes, and the remaining peers connect to a single super-node. Super-nodes act like central servers, keeping indices of shared files of their peers and handling all search requests. Each super-node is also connected to a subset of

other super-nodes, and it passes search requests along to these neighbors. File transfers are handled directly between peers. This is the structure of the FastTrack (KaZaA, iMesh, Grokster), eDonkey, and WinMX networks.

Since at least 2002, several P2P networks including examples of each basic architecture have been running simultaneously. These networks operate largely autonomously, so file sharing activity on one is mainly independent from the others. There are several reasons for this proliferation of structures, though legal issues relating to copyright infringement are likely the primary factors.

The size of these networks varies substantially, but during our fall 2002 study period they were all quite large. The largest network was FastTrack (hereafter FastTrack/KaZaA) which grew from 2.5 million to 3.5 million simultaneous users over September to December 2002. On FastTrack/KaZaA there were typically more than 500 million files holding 5 Petabytes of data available at any time. The second largest network was WinMX, which had about 1.5 million simultaneous users in 2002. Even the smaller networks are fairly large. OpenNap had at least 25,000 simultaneous users sharing over 10 million files. Note that Napster did not operate during our study period.

IV. Data

A. Overview

We use three types of data in this study. Server logs for two OpenNap servers allow us to observe what files users search for and what they download. Weekly album-level sales data come from Nielsen SoundScan (2002), which 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 from a variety of publications. For each of the 680 albums in our data set, we collected the titles of the individual tracks, information on performing artists and track time from AllMusic.com (2003), an on-line media guide published by Alliance Entertainment Corp. We form indicators for whether the album has a track

which is receiving heavy media attention in each week. Our indicator for frequent commercial radio play is based Billboard's (2002) "Top 50 Airplay," for heavy MTV rotation based on the top twenty-five ranks listed in Radio & Records (2002), and for widespread college radio play based on the top twenty ranks listed in CMJ Networks (2002). We also form weekly indicators for whether the artist is on tour based on concert dates from the weekly trade publication Pollstar (2002).

B. File Sharing Data and Album Sample

1. Overview

Our file sharing data was collected from OpenNap, a centralized P2P network. We have records for two servers, which operated continuously for seventeen weeks from 8 September to 31 December 2002. During this time most high school and college students, primary users of file sharing (Ipsos-Reid, 2002ab; Pew Internet Project, 2003), had access to broadband connections at school. The study period also includes the holiday shopping season when about half of all CDs are sold.

The servers were connected to T-3 lines which provided actual Internet transmission speeds of several megabits per second for both uploads and downloads. The high-speed connections ensured that a large number of search requests and downloads could be handled in real time. The information on file transfers is collected as part of the usual log files which the servers generate, and most users were not actively aware that they were being monitored. Search lines describe what users are looking for, and transfer lines give the location of the file that is being transferred as well as the name of the file, which includes information on the artist and the song. Typical examples are:

```
[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)
```

There are three important institutional features of OpenNap. First, there are several independent servers in the network, and clients are typically simultaneously logged into

many of them. As a result, the set of files available to users is quite large (in many cases the entire OpenNap network). In this sense, OpenNap resembles a hybrid P2P architecture as clients search across and download from several servers. Second, several software clients are used. 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. This means that our data overlap with the larger networks. Third, many servers are 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 a hybrid P2P. One of our servers was part of a sub-network of servers.[7]

An important question is whether our sample is representative of data on all P2P networks. We present here a brief overview of this point, and relegate the full discussion of this point to **Appendix A**. While we are unaware of any database spanning the universe of downloads,[8] we were able to compare downloads on our servers with a large sample from FastTrack/KaZaA, the leading network at the time. It is not possible to reject a null that the two download samples are drawn from the same population. We also find that the availability of titles are highly correlated on the two networks. The resemblance of the files on the networks is intuitive. First, the users are likely to be similar. Many of the clients in our data are from the WinMX network, which is one of the most popular networks and has a similar architecture as FastTrack/KaZaA. Second, there are few technical reasons relating to network architecture or the user experience which would drive differences. The portion of the OpenNap network where our data come from have many features of hybrid P2P, as we discussed earlier. Finally it is worth stressing that the relatively small size of the OpenNap network does not in itself cause problems. So long as the sample is representative (and in the absence of scale-effects), our estimates can be used to gauge the impact of total downloads on sales.

[7]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.

[8]Bigchampagne.com monitors some behavior on a variety of networks, but their full database is not public.

In our analysis, we focus on downloads because they most accurately capture what users want to hear among the set of available files. Downloads are the relevant measure that can potentially crowd out record sales, since these are the files users actually obtain.[9] To ensure only relevant files would be included, we analyze downloads which are in standard audio formats (MP3/MP2, OGG, ALBW, AU, AIF, WAV, WMA/WMP, MID/MIDI). We also restrict the analysis to downloads by clients in the U.S. The server logs include the I.P. address for each client (see the example above where the I.P. is partially masked). We mapped the I.P.'s to countries using a monthly updated database.

2. File Sharing Data: Descriptive Statistics and Matching Algorithm

A strength of our data is its size and span. Over the sample period we observe 1.75 million file downloads or roughly ten per minute.[10] This is about 0.01% of all the downloads in the world.[11] A significant majority of the downloads were music files. U.S. users accounted for about one third of the downloads (and the data contain about 0.01% of all music downloads by U.S. users). The breadth of file availability is also quite large, and at any time there are an average of 3 million files containing 100 Terabytes which are accessible. These data were shared by and made available to an average of 5,000 simultaneous users on the servers. This is similar to the user-base which a KaZaA user would see.[12]

A useful overview of our data is presented in **Figures 2-3** and Table 1. Figure 2 presents

[9]The alternatives to downloads are less desirable. Most searches go unfulfilled due to a lack of supply, and the queries themselves are often unrefined and difficult to match with specific music tracks. Shared files could have been legally purchased or might be an old download which is related to old, not current, sales.

[10]There were over 50 million searches or more than three hundred per minute.

[11]At the end of 2003, roughly one billion songs are downloaded per week (*Wall Street Journal*, 19 November 2003) or 17 billion file downloads during our seventeen week sample. This overstates the world-wide number of downloads during our observation period, since file sharing has a high growth rate (the number of simultaneous users on the FastTrack/KaZaA grew by over a third from mid-2002 through the end of 2003, and the number of world-wide downloads likely increased at about the same rate, *Ad Age*, 28 July 2003). During February 2001, at Napster's peak, about half a billion songs were downloaded per week (Romer, 2002).

[12]KaZaA nominally has millions of users, but the hybrid P2P architecture means that each user only has access to the files of a limited number of other users. 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 Dotcom Scoop, 2001 and giFT-FastTrack CVS Repository, 2003), resulting in simultaneous access to about 5,000 other computers. Our totals reflect users and files for the entire sub-network which one of our servers was on. The file totals include videos and may include multiple copies of some music titles.

the distribution of users across countries for our sample period. For the purpose of this figure, we define a user as log-in and log-out for a particular username plus I.P. address. 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. Figure 3 shows the distribution of downloads across countries. A download is defined as a transfer of a unique file name between a unique pair of clients, and the country is based on the I.P. address of the downloading client. This map mirrors the user distribution in Figure 2, with a wide range of countries represented and a U.S. share of 36% (the distribution of upload countries is quite similar). Table 1 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.[13]

Table 2 shows that interaction among file sharers transcends geography and language. While the left panel indicates that U.S. users downloaded almost half of their files from other U.S. users, the remainder comes from a diverse range of countries including Germany, Italy, and Brazil. The five percent of downloads not covered in this top fifteen list are spread out over almost every other country in the data. The right panel shows that the distribution for files uploaded from U.S. users follows a similar pattern.

User behavior in our data is also interesting. Over the entire sample period, the average user is observed on only two days, indicating large turnover in the user-base. During these two days, the user makes 17 downloads. There is quite a bit of heterogeneity, with one user observed during seventy-one days and downloading over five thousand files.

Table 3 reports the weekly number of unique downloads by users in the U.S. during our study period. Over the 17 weeks, U.S. users downloaded 260,889 audio files. We use a Perl program to match each transfer line to a set of popular albums containing 10,271 songs (the generation of the album sample is described below.) The approach we use 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 artist names in our set of albums.

[13]For example Italy has a much higher share of users than Spain despite a comparable rate of software piracy. More formally, software piracy does not have statistically significant effect in explaining file sharing. Only GDP has a large and positive economic effect when the last four covariates listed in Table 1 are regressed on the file sharing user share.

The list of artists against which we compare text strings 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%. The matching rate is fairly stable across our study period (see Table 3).

3. Album Sample

The list of albums in our sample is a subset of titles which were sold in U.S. stores in the second half of 2002. We start building our sample using Nielsen SoundScan (2002) charts for eight different genres of music: 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), and Top Soundtracks (100). Taken together, these eight genres made up 81.8% of all CD sales in the United States in 2002. The charts are published on a weekly basis, and we include an album if it appears on any chart in any week during the second half of 2002. There are 1,476 such albums. From this set, we draw a stratified random sample of 500 albums. To reflect the different music styles, we set the sample share of a genre equal to its fraction of CD sales in 2002. In the final sample of 500 titles, these shares are 29% R&B, 23% Alternative, 15% Rap, 13% Country, 7% Soundtrack and 4% for each of the categories Hard Music, Jazz and Latin. Within each genre, we randomly selected the individual titles. Random sampling is obviously important for the validity of our measures.

In addition to the genre-based charts, we also drew random samples from three charts that are of particular interest from a file sharing perspective. Top Current (200) is a list of

best-selling albums. New Artists (150) can shed light on the effect of file sharing on new talent, and Catalogue Albums (200) shows how older releases fare. Our final sample of 680 albums includes 80 titles from the Top Current, 50 from the New Artist and 50 from the Catalogue charts.

Table 4 reports sales data for the sample. The mean of sales for these albums during our observation period is 151,786 copies, ranging from 71 copies to 3.5 million copies. One way to assess the effect of random sampling is to compare the number of sales for albums included in the sample with total sales for each genre. For example, our sample represents 42% of all sales in the Catalogue category. Across all categories, 44% of sales are represented in the sample. A second measure is a comparison between sample sales and overall sales in the U.S., which is given in Table 5. Overall, our sample albums represent about a third of overall sales and this value is stable across weeks.

V. Empirical Strategy

A. Econometrics

Our goal is to measure the effect of file sharing on sales. We present a model of purchase and download behavior in **Appendix B** and highlight here the key implications. The simplest approach is to estimate simple pooled models of the form,

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

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 and difficult-to-measure album characteristics. For example, the popularity of a particular band is likely to drive both file sharing and sales, implying a positive bias on the estimated γ (see Appendix B for details and also a justification for the linear specification).

Making use of the fact that we observe sales and downloads for 17 weeks, we can control for album-specific time-invariant characteristics by estimating the fixed effects model,

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

In this specification, ν_i is an album fixed effect, t denotes time in weeks, and the summation allows for a flexible time effect.[14] 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. For example albums sales decay at very different rates following their release, and the pick-up in sales during the holiday season (see Table 5) might well vary by album. This type of unobserved heterogeneity can still bias our estimates of γ in specification (2).

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,

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

Valid instruments, Z_{it}, influence file sharing but are uncorrelated with the second stage errors, μ_i or μ_{it}. The model in Appendix B points out that shifters of download costs are candidates for instruments, since they influence downloads but typically have no direct influence on sales. 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.[15] An advantage of our instruments, which are discussed 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 (see the discussion in Nevo, 2001).

B. Instruments

[14]We consider a polynomial time trend of degree six, though our results below are virtually identical if we instead include week fixed effects. The advantage of having a polynomial rather than fixed effect is that we can use environmental variables to instrument for downloads, which are discussed below.

[15]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.

To identify the impact of file sharing on sales, exogenous shifts in downloads are needed. We consider several cost shifting instruments, which in terms of the model in Appendix B should influence α_q. These instruments stem from particular features of the file sharing infrastructure, and our identifying assumptions are that they directly influence downloads and are otherwise orthogonal to sales. We develop specific arguments for each instrument along these lines in the discussion below. As further justification of our assumptions, over-identifying tests are presented in the Results section. We utilize three types of instruments to capture a wide variety of forces which influence downloads but are not related to unobserved album popularity: album-specific instruments which are fixed over time, time-specific instruments which equally impact all released albums, and finally a time-varying and album-specific instrument. The availability of panel data is clearly central to our approach.

The first class of instruments are album-specific but time invariant. We consider album average and minimum track length which can affect download costs but are not typically related to album popularity. There is a one-for-one relationship between song length and the size of the resulting digital file: longer songs result in bigger files. Song lengths vary widely in our sample, from as short as a few seconds to as long as forty minutes (the mean is four minutes and the standard deviation is a minute and a half). Bigger files take longer to download. Not only is there the file transfer, but downloads are often interrupted.[16] Since interruptions are more likely with bigger files, downloads increase at a faster than linear rate with file size. Actual download time can vary from a few minutes up to an hour based on the size of the file.[17] We therefore expect an album's average track length to be negatively related to the number of downloads. There is a similar logic for using an album's minimum track length, which we expect to be positively related to

[16]Even with widespread access to broadband services, downloads are interrupted quite frequently. In our server logfiles, we observe repeated attempts of individual users to download the same song because these attempts result in multiple transfer lines. While we have 260,889 unique U.S. audio transfers in our logs – these are the basis for our analysis – the total number of U.S. audio transfer attempts is 549,870, with the bulk of the difference consisting of interrupted transfers.

[17]We have confirmed this on the FastTrack/KaZaA network. While a 5M file--the size of a typical music track--downloaded in eight minutes, a 15M file took forty-five minutes (these values are for a high speed university connection, and download times can be much longer on a slower dial-up connection). Download time is roughly proportionate to file size so long as the transfer is progressing, but there are often time gaps when transfers are interrupted or terminated.

downloads.[18] Song length has little relationship with popularity, with some top-selling albums consisting entirely of short tracks and others having mainly longer numbers. And even a sophisticated label which would like to strategically set track length to influence downloads is constrained because commercial radio play, a primary driver of sales, is devoted almost exclusively to three to five minute songs.

A second class of instruments are time-varying but at each moment have a relatively uniform impact on all albums. In this class our instruments are network traffic conditions (congestion should increase download costs for all individuals and thus decrease transfers) and exogenous shifts in the supply of albums (changes in participation from individuals outside of the treatment population). We aggregate these measures to a weekly frequency.

Several measure of congestion throughout the Internet are considered. The four "Internet weather" measures we consider are: The Consumer 40 Performance Index, which is 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).[19] These variables should reflect the delays a typical P2P user faces. For example, the IEPM measure is based on typical roundtrip times between a wide range of internet locations and so should be linked to P2P download speed. Similarly, a high share of file sharing traffic on the backbone will delay downloads. Note that the internet congestion measures have the advantage that they should influence download time not just in the file sharing network we study, but also all others. Hence the measures should be related to total downloads in

[18]Many albums contain very short tracks, typically introductions by the artist, which are unlikely to be downloaded for reasons of benefit and cost. On the cost side, these tracks are difficult to find on P2P networks because they all have similar titles (often "Intro," "Outro," or "Skit") and searches for these titles result in large numbers of false matches. On the benefit side, it is close to impossible to know whether or not an "Intro" is worth downloading because these tracks are not played on the radio. In addition, if downloading carries fixed cost per search, per-minute enjoyment is lower for shorter tracks. As the shortest track gets longer, it becomes more likely that it is a real song as opposed to a spoken introduction.

[19]These variables are highly correlated with other congestion indices. We also considered measures of local server congestion (rejected connections on our OpenNap servers), OpenNap network congestion (ping times to all active servers), DNS server lag time (described in Brownlee, et al., 2001), and other measures of Internet-wide congestion (packet loss rates, average throughput rates, and total traffic flows). The estimates below are similar when these alternative instruments are used.

the universe of P2P networks. Still all of these congestion measures are plausibly exogenous to music sales. For example, while a quarter of internet backbone packet traffic during our observation period is from file sharing applications, a majority of traffic is due to activities like data transfer, measurement, or unidentified (non-file sharing) packets.

Our measure of supply shift is based on the earlier observation that U.S. users download a majority of their files from non-domestic users. In particular, Table 2 shows that in our sample one out of every six U.S. downloads is from Germany. Shifts in participation of German users would influence download costs in the U.S. by altering the available supply of albums. German teens, the primary participants in file trading, tend to go on-line at home (Niesyto, 2002 documents that 87% of German students access the Internet at home, while only about a third regularly access the Internet at school). A candidate instrument would exogenously shift the population at school. Our instrument is the percentage of German kids on vacation due to German school holidays, which exhibits a surprising variability over time. German holidays produce a supply shock of files, making it easier for U.S. users to download music when many German kids sit at their computers. Our instrument is time-varying 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.[20] Our instrument is based on the total population of schoolchildren, though the estimates below are largely unaffected if we use the number of older children and youth (Sekundarbereich I&II.) Finally, there is little reason to believe this variable is endogenous. While German school holidays are potentially linked to downloads in the target population of U.S. users, these dates or the number of German kids who are off from school should not have an independent effect on American CD sales.

A final type of instrument is an album-specific and time-varying cost shifter. We consider the time length of albums in the same music category, which should influence the availability of tracks (and thus the cost of search and download) for the album in

[20]Data on the timing of the school periods was taken from Agentur Lindner (2004). The Kultusministerkonferenz publishes data on the number of German children and youth in school (Statistische Veröffentlichungen der Kultusministerkonferenz, 2002.)

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. So to be specific, a hip-hop fan is less likely to share some rap song when related artists have recently released an album (we observed just such a crowd-out of songs on Nelly's Nellyville album when the 8 Mile Soundtrack was released). 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. We argued earlier that song length is not related to album popularity, and yet it still varies over time due to the continual release of new albums. It is also album-specific since the album in question is excluded from the distribution.[21]

C. Further Econometric Issues

Given our relatively large number of time periods, time series concerns are relevant. In particular we consider issues related to the use of dynamic panel data. A potential concern with equations (2) and (3) is that our data may be non-stationary. This would imply the usual problems of spurious regression, inconsistency, and difficulties with inference (Baltagi, 2001).

Shocks, such as additional radio play or media exposure, can have persistent effects and continue to effect sales or downloads weeks after their occurrence. In fact, the *t*-test for unit roots in heterogeneous panels developed by Im, Pesaran and Shin (2003) cannot reject non-stationarity for our sales data series. Non-stationarity of the dependent variable leads to biased estimates (Evans and Savin, 1981). To address this issue, we estimate our model in first differences after which both sales and downloads are stationary.

[21]We also considered various interactions between the album-specific and time-specific instruments. Intuitively these could be reasonable instruments, since (for example) network congestion should be more of an issue for albums with long tracks than ones with short tracks. Nonetheless the interactions are not significant predictors of downloads and so are excluded from our analysis.

We further explore explores the importance of dynamics in our data by allowing the disturbance in (2) to be first-order autoregressive, $\mu_{it} = \rho\mu_{it-1}+\eta_{it}$ where η_{it} is white noise.

VI. Results

A. Cross-Tabulations and Validity Issues

We start describing our results by taking a closer look at file sharing activities. Table 6 reports frequencies of downloads of songs in our sample. The average song is downloaded 4.6 times over the study period. Downloading is heavily concentrated on a limited number of songs. For the sum of all weeks, the median number of downloads of a particular song is 0, the 75th percentile is 2, the 90th percentile is 11, and the 95th percentile is 22. The most popular song among our users is "Lose Yourself" from the 8 Mile Soundtrack, which was downloaded 1,258 times. Aggregated up to the album level (Table 7), users downloaded 70 songs from the average album in our sample. The 8 Mile Soundtrack, the album in our sample that sold the most copies during the observation period, was also the most popular among file sharers. 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.

As one would expect, songs from Top Current chart are most frequently downloaded (Table 8). Songs in this category average 17.2 downloads over our sample period (as opposed to 4.4 for Catalogue albums and 0.3 for Jazz, the least downloaded category.) The patterns are similar at the album level, with 277 downloads for Top Current albums and only 4 for Jazz. Mann Whitney test statistics in Table 8 confirm that Top Current albums are significantly more frequently downloaded than any other category.

Songs from higher selling albums are downloaded more frequently (Table 9). In the top quartile of sales, albums average 200 downloads. In the bottom category, the mean number of downloads is only 11. As Table 9 shows, the mean number of downloads

increases at a rate that is less than proportional to the rate of increases in sales.[22] More generally, while downloads and sales are both quite concentrated downloads are a bit more dispersed. In our sample of albums and during our observation period, the weekly top selling albums accounts for 7.6% of total sales while the weekly most downloaded albums accounts for 5.2% of all downloads. Similarly, the weekly top ten account for 31.5% of total sales and 25.7% of all downloads. More to the point, the top ten selling albums over the observation period account for 22.4% of sample sales while these same albums are only 15.5% of total downloads. The greater concentration of sales suggests that, contrary to popular opinion, individuals are not just downloading top hits. And more generally, the similar pattern of concentration is anecdotal evidence that common factors drive downloads and sales, and so serves as motivation for our instrumental variables approach.

Two other issues need to be discussed before turning to the main estimates. An important question is whether scale-effects influence the distribution of downloads. If the kinds of albums downloaded are systematically different on small rather than large networks, it will be difficult to make inferences about the aggregate effect of downloads from our sample. **Appendix A** provides both intuitive and empirical evidence suggesting such scale-effects do not seem to be particularly strong. The second issue involves time aggregation. We use high frequency data, and so it is possible that downloads can influence sales many periods later. For example an individual may decide today to download and not purchase some album, but he might delay his download until a later week if it is currently costly to access the file due to congestion or availability issues. Alternatively an individual could download an album and decide he wants to purchase it, but he does not go to the store until some later week. This suggests the stock of previous downloads might have important dynamic effects. To address this issue we estimated a distributed lag model with seven lags of sales. The main conclusions we draw from the estimates below are robust to this change.

B. Pooled Sample Models

[22]Table 10 shows the relationship between release dates and the number of downloads is less clear cut. Songs on recently released albums (during Summer 2002) are as likely to be downloaded as older albums (released prior to 11/9/2001).

Table 11 reports the results for specification (1), which pools sales and downloads in all weeks. Model (I) controls for the music category an album belongs to. We find that downloads increase sales, which is not unexpected given our concerns about the likely endogeneity of file sharing. Relative to Top Current albums, the omitted category, sales in all other genres are significantly lower. Model (II) in Table 11 presents 2SLS estimates for the pooled data. The time invariant instrumental variables have the expected signs. Longer tracks are less likely to be downloaded, while the minimum track time bears a positive relationship to the number of downloads. Instrumenting for the number of downloads increases our point estimate of the effect of downloads on sales. However, since our instruments are not particularly strong, we next explore the robustness of this result in a setting where we make use of the panel nature of our data.

C. Panel Data Models

In Table 12, we report results for specification (2). The simplest specifications are OLS with a polynomial time trend (model I) or with a time trend and album fixed effects (model II). While we continue to find a positive effect of downloads on sales, the relationship is much weaker in the fixed effects model. This indicates that unobserved time-invariant album characteristics such as popularity biased our pooled OLS estimates upward.

The next two sets of estimates instrument for downloads (we cannot use the time invariant instruments from the last section because album fixed effects are included in both stages). We first use the German holiday instrument (model 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 US. A one standard deviation increase in children off from school increases the number of observed downloads by about one fifth of a standard deviation (2.4 downloads). More importantly, once we instrument for downloads, the estimated effect of file sharing on sales is quite small (slightly negative) and statistically indistinguishable from zero. We next add as instruments the Internet congestion measures and non-sales characteristics of competing albums (model IV). The additional instruments have the expected first stage signs, i.e. greater congestion or ease

of acquiring competing albums reduce downloads. The instruments satisfy the standard test.[23] In this richer model downloads have a more negative effect on sales, but the effect continues to be statistically indistinguishable from zero.

The remaining two models in Table 12 account for the dynamic panel data issues discussed in Section VC. The first issue is the non-stationarity of sales. We estimate in first-differences (model V), since then sales and downloads are stationary. The full set of first-differenced instruments is used, and the over-identification test indicates the first-differenced variables remain valid instruments. We continue to find that the number of downloads has no statistically discernible effect on sales, though the parameter is now positive. The last specification (model VI) allows for an AR(1) error term in the sales equation. As before, the number of downloads is assumed to be endogenous and is instrumented for with the full set of instrumental variables. After taking first differences, we cannot reject a null of stationarity (see the small estimate for ρ and the Baltagi-Wu test). And again we find that file sharing activities do not have a statistically significant effect on sales.

The statistical insignificance of the point estimate notwithstanding, how large an effect is the estimated reduction in sales? NPD's MusicWatch Digital, an industry market tracking service, estimates that users in the U.S. download 0.8bn music files every month from file sharing networks (Crupnick, 2003). Applied to our study period, this implies that each matched file transfer in our data set corresponds to roughly 71,000 transfers in the entire United States. Focusing on the most negative point estimate (model IV in Table 12), it would take 5,000 downloads to reduce the sales of an album by one copy. After annualizing this would imply a yearly sales loss of 2m albums, which is virtually rounding error (total U.S. CD sales were 803m in 2002). To provide a point of reference, aggregate sales declined by 139m from 2000 to 2002. Given that the estimated effect of downloads is even smaller in model (III) and positive (but still economically small following a similar calculation as above) in models (V) and (VI), there is little evidence in our results that file sharing has a marked negative impact on sales.

[23]This specification is overidentified, so we report a Sargan-type overidentification test 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. We cannot reject the null.

From an industry perspective, it is particularly interesting to know how the effect of file sharing varies by the album popularity. For major labels, a few successful acts contribute the lion share of sales and profits. In Table 13, we ask how the effect of file sharing varies across commercially more or less successful albums. We do this by separately estimating our preferred specification, instrumented downloads in first-differences as in Table 12 model (V), for various sales quartiles.[24] The parameter coefficients indicate there is only a modest impact of file sharing on the low selling quartiles. The effect grows stronger as we move to higher selling categories. For the top quartile, downloads have a relatively large positive effect (150 downloads increase sales by one copy) though this is estimated rather imprecisely. These results are also inconsistent with the argument that file sharing is reducing sales of commercially important albums.

We perform a similar analysis to study if the effects of downloads vary by music category. We estimate our preferred model using sub-samples of alternative, hard, jazz, Latin, R&B, rap, country and soundtrack albums. We find no statistically discernible effect of file sharing on sales for all these individual categories.

*** note: we are still checking the robustness of this result!***

Finally, we consider a robustness check on the estimates in Table 14. It is widely believed that promotion of albums in media or through tours boosts sales. The growing visibility might also increase downloads. We therefore include our measures of such "advertising" in both the first and second stage estimates using our preferred first-difference specification.[25] Model (I) of Table 14 includes indicators for whether the album has a song which was on heavy MTV rotation or made the Billboard list of widespread commercial radio play. As we would expect, MTV play increases both sales and downloads: heavy rotation increases weekly U.S. downloads by about 300,000 and weekly sales by 6,000. Radio play has a similar effect on sales (the negative effect on

[24]It is inappropriate to run a single equation where the instrumented downloads are interacted with various sales ranking indicators. While the download variable has been purged of the endogenous popularity component, the rankings have not. This means the estimated parameter on downloads will have a bias which grows more positive as the sales ranking increases.

[25]In the interest of brevity, we omit results using college radio play (CMJ Networks, 2002) which appears to have a negative impact on our outcomes. However, this likely reflects the relative obscurity of albums played on college stations.

downloads is in part due to the collinearity with the MTV indicator). More importantly, the impact of downloads on sales continues to be small and statistically indistinguishable from zero. This result remains in model (II) when an indicator for touring is included (the negative parameter on tours in the sales equation reflects the lag between an album release and the tour). These estimates point out that the record labels and artists themselves, through media promotion and touring, are important drivers of downloads.

VII. Conclusion

We find that file sharing has no statistically significant effect on purchases of the average album in our sample. Moreover, the estimates are of rather modest size when compared to the drastic reduction in sales in the music industry. At most, file sharing can explain a tiny fraction of this decline. This result is plausible given that movies, software, and video games are actively downloaded, and yet these industries have continued to grow since the advent of file sharing. While a full explanation for the recent decline in record sales are beyond the scope of this analysis, several plausible candidates exist. These alternative factors include poor macroeconomic conditions, a reduction in the number of album releases, growing competition from other forms of entertainment such as video games and DVDs (video game graphics have improved and the price of DVD players or movies have sharply fallen), a reduction in music variety stemming from the large consolidation in radio along with the rise of independent promoter fees to gain airplay, and possibly a consumer backlash against record industry tactics.[26] It is also important to note that a similar drop in record sales occurred in the late 1970s and early 1980s, and that record sales in the 1990s may have been abnormally high as individuals replaced older formats with CDs (Liebowitz, 2003).

Our results can be considered in a broader context. A key question is the impact of file sharing (and weaker property rights for information goods) on societal welfare. To make such a calculation, we would need to know how the production of music responds to the

[26]There is a movement to boycott music sales from the major labels., as discussed at `http://www.boycott-riaa.com/` and `http://www.dontbuycds.org/`.

presence of file sharing. Based on our results, we do not believe file sharing will have 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. For these albums, we find that the impact of file sharing on sales is likely to be positive, leaving the ability of major labels to promote and develop talent intact. Our estimates indicate that less popular artists who sell few albums are most likely to be negatively affected by file sharing. (Note, however, that even for this group the estimated effect is statistically insignificant.) Even if this leads record labels to reduce compensation for less popular artists, it is not obvious this will influence music production. This is because the financial incentives for creating recorded music are quite weak. Few of the artists who create one of the roughly 30,000 albums released each year in the U.S. will make a living from their sales because only a few albums are ever profitable.[27] In fact, only a small number of established acts receive contracts with royalty rates ensuring financial sufficiency while the remaining artists must rely on other sources of income like touring or other jobs (Albini, 1994; Passman, 2000). Because the economic rewards are concentrated at the top and probably fewer than one percent of acts ever reach this level (Ian, 2000), altering the payment rate should have very little influence on entry into popular music.

If we are correct in arguing that downloading has little effect on the production of music, then file sharing probably increases aggregate welfare. 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. File sharing lowers the price and allows an apparently large pool of individuals to enjoy 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.

[27]Major label releases are profitable only after they sell at least a half million copies, a level only 113 of their 6,455 new albums reached (Ordonez, 2002). 52 records account for 37% of the total sales volume (Ian, 2000). Twenty-five thousand new releases sold less than one thousand copies in 2002 (Seabrook, 2003).

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, the network is largely composed of WinMX clients which formed the second largest file sharing community among U.S. users during our sample 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.[28] The OpenNap network has a similar structure, particularly the sub-network associated with one of our 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 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.[29] This is near the average user base of our server which is on the sub-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

[28]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 Dotcom Scoop (2001), giFT-FastTrack CVS Repository (2003), and Buchanan (2003).

[29]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 Dotcom Scoop, 2001 and giFT-FastTrack CVS Repository, 2003).

whole sample (N=144) indicating that the availability of common and rare songs move in tandem in the two networks.[30]

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, purportedly based on monitoring behavior on a broad range of file sharing networks (they do not reveal whether their list is based on shares, searches, or downloads). Over our seventeen week sample period, two-thirds of our top ten downloads also appear in the BigChampagne top ten list.[31]

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).[32] 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 OpenNap networks are effectively as big as the larger FastTrack/KaZaA or WinMX. 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.

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

[31]There were 42 unique tracks from our album list which received a top ten rank in BigChampagne over our sample period. 28 of these tracks were in the top ten downloads during at least one week of our data. 28 of our top ten most downloaded tracks

[32]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.

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 <u>popularity of an album</u> 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 popular or equivalently their population distribution is shifted to the right. α_{qi} measures the <u>cost of downloading an album</u> 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 show download cost shifters are appropriate instruments. A cost drop increases downloads and increases purchases *iff* the feedback effect from downloads is positive. 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 $\phi(\alpha_V)$.

References

Adar, Eytan and Bernando Huberman (2000). "Free Riding on Gnutella." *First Monday*. 5:10. http://firstmonday.org.

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

Albini, Steve (1994). "The Problem With Music." MAXIMUMROCKNROLL. http://www.arancidamoeba.com/mrr/problemwithmusic.html.

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

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

Baltagi, Badi (2001). *Econometric Analysis of Panel Data*. Chichester: John Wiley & Sons, Ltd.

Baltagi, Badi and Ping Wu (1999). "Unequally Spaced Panel Data Regressions with AR(1) Disturbances." *Econometric Theory*. 15: 814-823.

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

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

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

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.

Brownlee, Nevil, kc Claffy, and Evi Nemeth. "DNS Root/gTLD Performance Measurements." CAIDA working paper. San Diego Supercomputer Center.

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.

Crupnick, Russ (2003). *Digital Music In Perspective: A Behavioral View*. NPD Group. Interview at http://www.insidedigitalmedia.com, 31 December 2003.

Dempsey, Bert, Debra Weiss, Paul Jones, and Jane Greenberg (1999). "A Quantitative Profile of A Community of Open Source Linux Developers." SILS Technical Report TR-1999-05.

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

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

Evans, G. B. A., Savin, N. E. (1981). Testing for Unit Roots. *Econometrica* 49, 753-779.

Fine, Michael (2000). "SoundScan Study on Napster Use and Loss of Sales." http://www.riaa.com/news/filings/pdf/napster/fine.pdf.

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.

Ian, Janis (2000). "From the Majors To the Minors." http://www.janisian.com/article-from_the_majors_to_the_minors.html.

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.

Im, Kyung So, M. Hashem Pesaran, Yongcheol Shin (2003). "Testing for Unit Roots in Heterogeneous Panels." *Journal of Econometrics* 115: 53-74.

Internet2 Netflow Statistics (2004). *Internet2 NetFlow: Weekly Reports*. http://netflow.internet2.edu/weekly/. Abilene NetFlow Nightly Reports.

Ipsos-Reid (2002a). "File Sharing and CD Burners Proliferate (12 June 2002)." Tempo: Researching the Digital Landscape. http://www.ipsos-na.com/dsp_tempo.cfm.

Ipsos-Reid (2002b). "Americans Continue to Embrace Potential of Digital Music (5 December 2002)." Tempo: Researching the Digital Landscape. http://www.ipsos-na.com/dsp_tempo.cfm.

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

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

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.

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 (2002). "Policing Pirates in the Networked Age." *Policy Analysis*. Number 438. http://www.cato.org/pubs/pas/pa438.pdf.

Liebowitz, Stan (2003). "Will MP3 downloads Annihilate the Record Industry? The Evidence so Far." In *Advances in the Study of Entrepreneurship, Innovation, and Economic Growth*, edited by Gary Libecap, JAI Press.

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

Neilsen//NetRatings (2003). "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

Ordonez, Jennifer (2002). "Pop Singer Fails to Strike A Chord Despite the Millions Spent By MCA." *Wall Street Journal*. 26 February 2002.

Passman, Donald (2000). *All You Need To Know About The Music Business*. New York: Simon & Schuster.

Pew Internet Project (2000). "Downloading Free Music: Internet Music Lovers Don't Think It's Stealing." http://www.pewtrusts.com/pubs/.

Pew Internet Project (2003). "Music Downloading, File-Sharing and Copyright." http://www.pewtrusts.com/pubs/.

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

Pollstar (2002). *POLLSTAR--The Concert Hotwire*. Fresno, CA: Pollstar.

Radio & Records (2002). *R&R: The Industry's Newspaper*. LA: Radio & Records, Inc.

RIAA (2002). *The Recording Industry Association of America's 2002 Yearend Statistics*. http://www.riaa.com.

Romer, Paul (1990). "Endogenous Technological Change." *Journal of Political Economy*. 98: S71-S102.

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

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

Seabrook, John (2003). "The Money Note: Can The Record Business Survive?" *The New Yorker*. 7 July. 42-55.

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.

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

Zentner, Alejandro (2003). "Measuring the Effect of Online Music Piracy on Music Sales." University of Chicago working paper.

Table 1 – The Geography of File Sharing (numbers in %)

Country	Share of users	Share of downloads	Share World Population	Share World GDP	Share World Internet Users	Software Piracy Rate
United States	30.9	35.7	4.6	21.2	27.4	23
Germany	13.5	14.1	1.3	4.5	5.3	32
Italy	11.1	9.9	0.9	2.9	3.2	47
Japan	8.4	2.8	2.0	7.2	9.3	35
France	6.9	6.9	1.0	3.1	2.8	43
Canada	5.4	6.1	0.5	1.9	2.8	39
United Kingdom	4.1	4.0	1.0	3.1	5.7	26
Spain	2.5	2.6	0.6	1.7	1.3	47
Netherlands	2.1	2.1	0.3	0.9	1.6	36
Australia	1.6	1.9	0.3	1.1	1.8	32
Sweden	1.5	1.7	0.1	0.5	1.0	29
Switzerland	1.4	1.5	0.1	0.5	0.6	32
Brazil	1.3	1.4	2.9	2.7	2.3	55
Belgium	0.9	1.2	0.2	0.6	0.6	31
Austria	0.8	0.6	0.1	0.5	0.6	30
Poland	0.5	0.7	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 2 – U.S. Download and Upload Locations. Shares of Top 15 Countries (in %)

Users in the U.S. Download from		Users in the U.S. Upload to	
United States	45.1	United States	49.0
Germany	16.5	Germany	8.9
Canada	6.9	Canada	7.9
Italy	6.1	Italy	5.7
United Kingdom	4.2	France	4.7
France	3.8	United Kingdom	4.2
Japan	2.5	Australia	2.2
Netherlands	1.9	Spain	2.0
Spain	1.8	Japan	1.8
Sweden	1.8	Netherlands	1.6
Brazil	1.2	Sweden	1.5
Norway	0.9	Brazil	1.3
Switzerland	0.9	Belgium	1.0
Australia	0.8	Switzerland	1.0
Poland	0.7	Mexico	0.6

Note: The U.S. share in the two columns does not match because the number of downloads and uploads involving U.S. users are different.

Table 3 – Downloads and Matched Songs

		Number of downloads in server log files	Number of songs matched to downloads	% downloads matched to song in sample
all weeks		260,889	47,709	18.29
week 1	week of 8 September	2,164	442	20.41
week 2	week of 15 September	1,347	144	10.68
week 3	week of 22 September	12,051	2,239	18.58
week 4	week of 29 September	15,742	3,050	19.38
week 5	week of 6 October	8,922	1,695	18.99
week 6	week of 13 October	12,534	2,681	21.39
week 7	week of 20 October	8,688	1,530	17.61
week 8	week of 27 October	5,967	1,130	18.93
week 9	week of 3 November	4,468	811	18.16
week 10	week of 10 November	20,936	4,273	20.41
week 11	week of 17 November	29,755	5,813	19.54
week 12	week of 24 November	29,284	5,824	19.89
week 13	week of 1 December	23,914	4,304	18.00
week 14	week of 8 December	26,404	4,345	16.45
week 15	week of 15 December	22,820	2,979	13.05
week 16	week of 22 December	19,428	3,461	17.82
week 17	week of 29 December	16,465	2,989	18.15

Note: Numbers in the Table only include audio files which are downloaded by users located in the U.S. Multiple downloads of the same file by a client from one other client (reflecting an interruption or disconnection) are only counted once. The downloads are matched to tracks in our sample of albums (=10,271 tracks.)

Table 4 – Sample Sales(1,000s) by Category

	obs	Mean sales	Std dev	Min	Max	Proportion of sales in original charts
Full sample	680	151.786	363.541	0.071	3498.496	0.44
Catalogue	50	49.754	42.606	0.235	239.502	0.42
Alternative	117	125.589	141.238	9.746	844.727	0.65
Hard	19	29.796	24.003	2.962	93.942	0.35
Jazz	21	23.975	70.276	0.083	325.919	0.41
Latin	21	28.321	34.698	3.702	138.242	0.96
New artists	50	16.508	13.627	0.318	56.915	0.55
R&B	146	49.472	75.445	2.002	500.805	0.21
Rap	77	39.483	62.658	1.027	315.445	0.30
Current	80	792.547	741.119	4.236	3498.496	0.39
Country	66	92.012	137.191	0.071	701.880	0.64
Soundtrack	33	47.411	83.159	5.032	346.569	0.39

Note: Proportion of sales in original charts compares sales of albums included in our samples to total sales in the Billboard chart from which the random sample was drawn. A comparison to overall US sales is provided in Table 5. 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 5 – Sample Sales by Week

		Sales of albums in sample (# copies)	% of total album sales in U.S.
all weeks		104,002,856	35.9
week 1	week of 8 September	3,661,568	30.7
week 2	week of 15 September	3,078,103	32.2
week 3	week of 22 September	3,409,499	33.7
week 4	week of 29 September	3,911,991	35.8
week 5	week of 6 October	4,111,011	36.5
week 6	week of 13 October	3,676,026	34.4
week 7	week of 20 October	4,048,804	32.7
week 8	week of 27 October	3,809,819	32.4
week 9	week of 3 November	5,003,957	37.2
week 10	week of 10 November	5,384,753	33.2
week 11	week of 17 November	5,789,505	30.9
week 12	week of 24 November	6,684,465	33.0
week 13	week of 1 December	9,929,928	36.5
week 14	week of 8 December	7,353,564	36.9
week 15	week of 15 December	10,046,509	34.5
week 16	week of 22 December	13,618,747	36.0
week 17	week of 29 December	10,484,607	35.3

Table 6 – Number of downloads per song

	Number of songs in sample	Mean number of downloads	Std dev	Min	Max
all weeks	10271	4.645	21.462	0	1258
week 1	10271	0.043	0.446	0	17
week 2	10271	0.014	0.176	0	7
week 3	10271	0.218	1.274	0	66
week 4	10271	0.297	1.451	0	35
week 5	10271	0.165	0.953	0	34
week 6	10271	0.261	1.419	0	60
week 7	10271	0.149	1.040	0	47
week 8	10271	0.110	0.748	0	39
week 9	10271	0.079	0.636	0	45
week 10	10271	0.416	3.250	0	250
week 11	10271	0.566	3.612	0	260
week 12	10271	0.567	2.932	0	155
week 13	10271	0.419	2.705	0	140
week 14	10271	0.423	2.409	0	104
week 15	10271	0.290	1.703	0	80
week 16	10271	0.337	1.952	0	86
week 17	10271	0.291	1.534	0	56

For the sum of all weeks, the median number of downloads of a particular song is 0, the 75^{th} percentile is 2, the 90^{th} percentile is 11, and the 95^{th} percentile is 22.

Table 7 – Number of Downloads per Album

	Number of albums in sample	Mean number of downloads	Std dev	Min	Max
all weeks	680	70.162	158.628	0	1799
week 1	680	0.654	2.476	0	34
week 2	680	0.209	1.027	0	12
week 3	680	3.287	9.824	0	120
week 4	680	4.491	12.380	0	136
week 5	680	2.491	8.105	0	90
week 6	680	3.938	11.477	0	124
week 7	680	2.254	6.904	0	72
week 8	680	1.663	5.146	0	48
week 9	680	1.194	3.588	0	53
week 10	680	6.278	19.061	0	349
week 11	680	8.547	23.303	0	368
week 12	680	8.560	21.262	0	253
week 13	680	6.331	16.852	0	285
week 14	680	6.385	16.056	0	164
week 15	680	4.387	11.198	0	116
week 16	680	5.096	13.433	0	104
week 17	680	4.396	12.867	0	180

For the sum of all weeks, the median number of downloads per album is 16, the 75^{th} percentile is 63, the 90^{th} percentile is 195, and the 95^{th} percentile is 328. For 147 albums, there are zero downloads.

Table 8 – Downloads by Genre

	# songs (# albums) in sample	Mean # of downloads	Std dev	Min	Min	Mann-Whitney
			Song level			
Catalogue	714	4.361	10.370	0	152	13.152**
Alternative	1707	7.021	18.153	0	312	11.432**
Hard	270	4.830	8.684	0	52	7.454**
Jazz	261	0.333	0.920	0	7	17.324**
Latin	309	0.550	2.927	0	28	19.122**
New artists	711	0.609	7.039	0	184	26.664**
R&B	2249	1.635	7.680	0	159	33.382**
Rap	1227	0.920	4.887	0	82	30.750**
Current	1342	17.182	51.286	0	1258	
Country	913	1.974	6.382	0	128	21.213**
Soundtrack	568	1.673	5.301	0	61	19.304**
			Album level			
Catalogue	50	62.280	103.114	0	680	5.698**
Alternative	117	102.436	122.794	0	674	4.969**
Hard	19	68.632	82.899	0	264	3.791**
Jazz	21	4.143	4.542	0	13	6.682**
Latin	21	8.095	26.344	0	121	6.578**
New artists	50	8.660	33.097	0	229	9.045**
R&B	146	25.542	56.494	0	433	10.275**
Rap	77	14.855	24.487	0	119	9.458**
Current	80	277.807	333.935	2	1799	
Country	66	27.303	51.649	0	344	8.202**
Soundtrack	33	28.788	36.611	0	185	6.288**

Mann Whitney test statistics are for the null that the current downloads, which have the largest mean, are from the same population as the other genres. This hypothesis is rejected for all comparisons.
** 1% level of significance

Table 9 – Downloads by Sales – Album Level

	Obs	Mean # of downloads	Std dev	Min	Max	Mann-Whitney
1st quartile: mean 7,330 copies [up to 36,066 copies]	170	10.812	38.060	0	402	-14.223**
2nd quartile: mean 21,619 copies [up to 132,654 copies]	170	21.882	52.401	0	433	-12.375**
3rd quartile: mean 60,371 copies [up to 603,308 copies]	170	47.694	55.331	0	264	-8.270**
4th quartile: mean 517,747 copies [max 11,176,209 copies]	170	200.259	265.370	1	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 10 – Downloads by Release Date – Album Level

	Obs	Mean # of downloads	Std dev	Min	Min	Mann-Whitney
1st quartile [prior to 11/9/2001]	170	63.647	99.661	0	680	0.483
2nd quartile [prior to 6/26/2002]	173	60.081	131.884	0	980	-2.427*
3rd quartile [prior to 9/25/2002]	180	51.611	120.788	0	706	-3.209**
4th quartile [prior to 12/18/2002]	157	109.592	246.423	0	1799	

Earliest release date is 12/8/1983. Mann Whitney test statistics are for the null that the 4th quartile with the most recent release dates comes from the same population as the other sales quartiles.
** 1% level of significance * 5% level of significance

Table 11 – Downloads and Album Sales

	(I)	(II)	
		1st stage	2nd stage
	sales	# downloads	sales
# downloads	1.071		1.467
	(0.194)**		(0.567)**
Alternative	-479.066	-175.820	-409.633
	(65.146)**	(19.510)**	(95.749)**
Hard	-538.641	-205.246	-455.824
	(67.644)**	(34.606)**	(113.377)**
Jazz	-475.369	-270.465	-367.020
	(71.022)**	(33.377)**	(144.448)**
Latin	-475.257	-277.996	-367.836
	(69.383)**	(34.065)**	(140.549)**
R&B	-472.798	-247.962	-372.982
	(68.450)**	(18.845)**	(131.685)**
Rap	-471.338	-253.844	-367.008
	(68.825)**	(21.955)**	(136.926)**
Country	-432.146	-258.409	-332.966
	(69.879)**	(22.663)**	(132.146)**
Soundtrack	-478.338	-244.529	-379.746
	(69.505)**	(28.110)**	(131.656)**
New artists	-487.675	-267.350	-381.290
	(69.290)**	(24.657)**	(139.305)**
Catalogue	-511.878	-214.646	-428.407
	(67.536)**	(24.499)**	(116.300)**
Mean track time on album		-0.199	
		(0.096)*	
Minimum track time on album		0.228	
		(0.089)**	
Constant	494.905	294.229	384.916
	(69.709)**	(21.793)**	(144.543)**
# Observations	680	673	673
Adjusted R^2 (uncentered R^2)	0.599	0.275	0.577 (0.640)
Partial R^2 instruments (Prob $F>0$)		0.018 (0.037)	
Sargan overid test χ^2 p-value			0.149

Dependent variables are album sales (1,000s) and # downloads at the 1st stage. 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. We also tested the orthogonality conditions for each individual instrument using the difference-in-Sargan statistic, which is the difference of the Hansen-Sargan statistic of the unrestricted and the restricted equations (see Baum, Schaffer and Stillman, 2003). The null is that both the restricted and unrestricted equations are well-specified. We cannot reject the null for the reported specification. Robust standard errors are in parentheses.
** 1% level of significance * 5% level of significance

Table 12 – Panel Analysis - Downloads and Album Sales

	(I) sales	(II) sales	(III) 1st stage # downloads	(III) 2nd stage sales	(IV) 1st stage # downloads	(IV) 2nd stage sales	(V) 1st stage # downloads	(V) 2nd stage Δ sales	(VI) 2nd stage Δ sales
# downloads	1.193	0.281		-0.001		-0.014			
	(0.022)**	(0.025)**		(0.195)		(0.175)			
Δ # downloads (instrumented)								0.088	0.038
								(0.49)	(0.05)
German kids on vacation (million)			0.670		0.366		0.370		
			(0.054)**		(0.123)**		(0.113)**		
Internet Consumer 40 Performance Index					-1.122		-0.820		
					(0.347)**		(0.273)**		
Internet average roundtrip time (ms)					-0.184		-0.164		
					(0.059)**		(0.048)**		
Internet std deviation roundtrip time (ms)					0.135		-0.332		
					(0.079)**		(0.149)*		
Internet2 net flow: % file sharing					-0.260		0.102		
					(0.069)**		(0.065)		
Mean album time "other" albums in musical genre					0.126		0.156		
					(0.043)**		(0.086)		
Polynomial time trend of degree six	yes	yes	yes	yes	yes	yes	yes	yes	yes
Album Fixed Effects?	no	yes	yes	yes	yes	yes	yes	yes	yes
Constant	19.199	21.671	4.889	21.888	37.720	22.043	-2.588	-7.342	-0.292
	(5.470)**	(3.753)**	(1.602)**	(3.799)**	(17.652)*	(3.821)**	(25.172)	(0.62)	(0.12)
ρ									0.023
Observations	10093	10093	10093	10093	9991	9991	9320	9320	8649
Prob $F>0$ on excluded instruments			0.000		0.000		0.000		
Sargan test (p-value)					0.1715			0.586	0.593
Baltagi-Wu LBI									2.710
R-squared	0.23	0.03	0.029	0.005	0.0139	0.0104	0.029	0.01	0.0188

Dependent variables are album sales (1,000s) and # downloads at the 1st stage. Specification (V) and (VI) estimate the model in first differences. Specification (VI) models the disturbance term as first-order autoregressive. In this model, the polynomial time trend is replaced with weekly indicators. ρ is the estimated coefficient on the AR(1) disturbance. The Baltagi and Wu (1999) test for unbalanced panels is for the null that $\rho=0$. We cannot reject the hypothesis. For an explanation of the Sargan overidentification test, see Note for Table 11. Robust standard errors are in parentheses. Album-weeks prior to the release date are excluded from the sample

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

Table 13 – Downloads and Album Sales – Effects by Sales

	(I) 2nd stage Δ sales for 1st quartile sales	(II) 2nd stage Δ sales for 2nd quartile sales	(III) 2nd stage Δ sales for 3rd quartile sales	(IV) 2nd stage Δ sales for 4th quartile sales
Δ # downloads	-0.005	0.051	0.084	0.468
(instrumented)	(0.009)	(0.021)*	(0.030)**	(0.307)
Polynomial time trend of degree six	yes	yes	yes	yes
Constant	-0.226	1.578	3.301	45.159
	(0.268)	(0.612)*	(1.890)	(75.373)
Observations	2243	2397	2388	2388
Prob *F*>0 on excluded instruments	0.000	0.000	0.000	0.000
Sargan test (p-value)	0.1749	0.2914	0.2628	0.4404

Robust standard errors in parentheses
* significant at 5%; ** significant at 1%

Table 14 – Downloads and Album Sales – Role of Radio, TV, and Touring

	(I)		(II)	
	1st stage # downloads	2nd stage sales	1st stage # downloads	2nd stage sales
Δ # downloads		0.012		-0.037
(instrumented)		(0.172)		(0.200)
Video shown on MTV	3.686	5.724	4.811	7.324
Top 25 video	(0.726)**	(1.869)**	(0.752)**	(2.060)**
Song is on Billboard's	-1.399	5.390	-1.525	5.518
Top 50 Airplay	(0.691)*	(1.692)**	(0.712)*	(1.753)**
Band is on tour this			0.471	-1.826
week			(0.657)	(1.595)
German kids on	0.361		0.650	
vacation (million)	(0.123)**		(0.201)**	
Internet Consumer 40	-1.118		-1.249	
Performance Index	(0.347)**		(0.358)**	
Internet average	-0.186		-0.307	
roundtrip time (ms)	(0.059)**		(0.089)**	
Internet std deviation	0.140		0.208	
roundtrip time (ms)	(0.079)		(0.087)**	
Internet2 net flow:	-0.261		-0.133	
% file sharing	(0.069)**		(0.098)	
Mean album time "other"	0.128		0.144	
albums in musical genre	(0.043)**		(0.044)**	
Album Fixed Effects?	yes	yes	yes	yes
Polynomial time trend of degree six	yes	yes	yes	yes
Constant	37.268	20.567	52.043	19.762
	(17.628)*	(3.726)**	(20.485)**	(3.748)**
Observations	9991	9991	9399	9399
Prob *F*>0 on excluded instruments	0.000		0.000	
Sargan test (p-value)		0.183		0.209
R-squared	0.0182	0.1114	0.0176	0.0595

Dependent variables are album sales (1,000s) and # downloads at the 1st stage. For an explanation of the Sargan overidentification test, see Note for Table 11. Robust standard errors are in parentheses. Album-weeks prior to the release date are excluded from the sample
* significant at 5%; ** significant at 1%

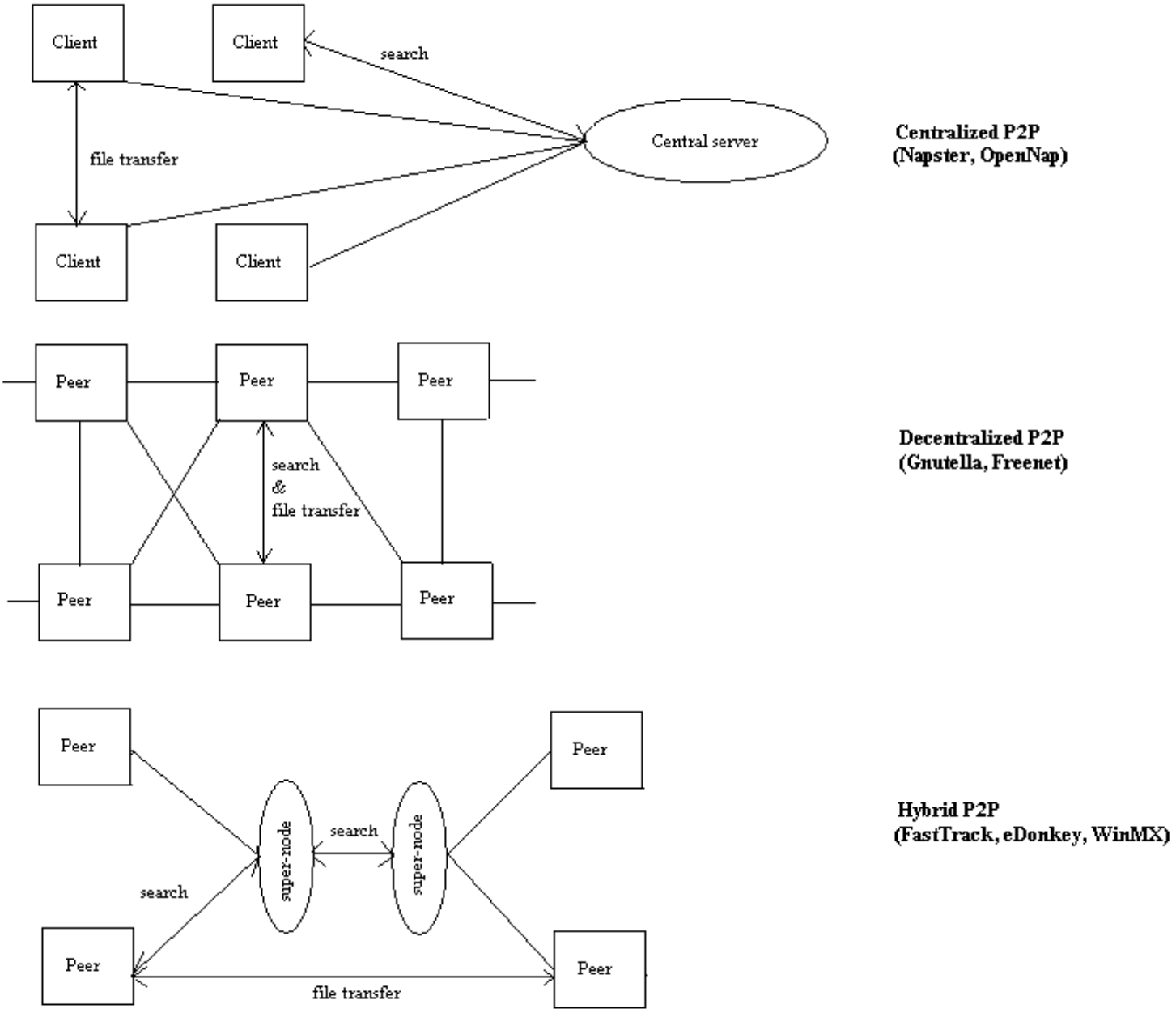

Figure 1: P2P Architectures

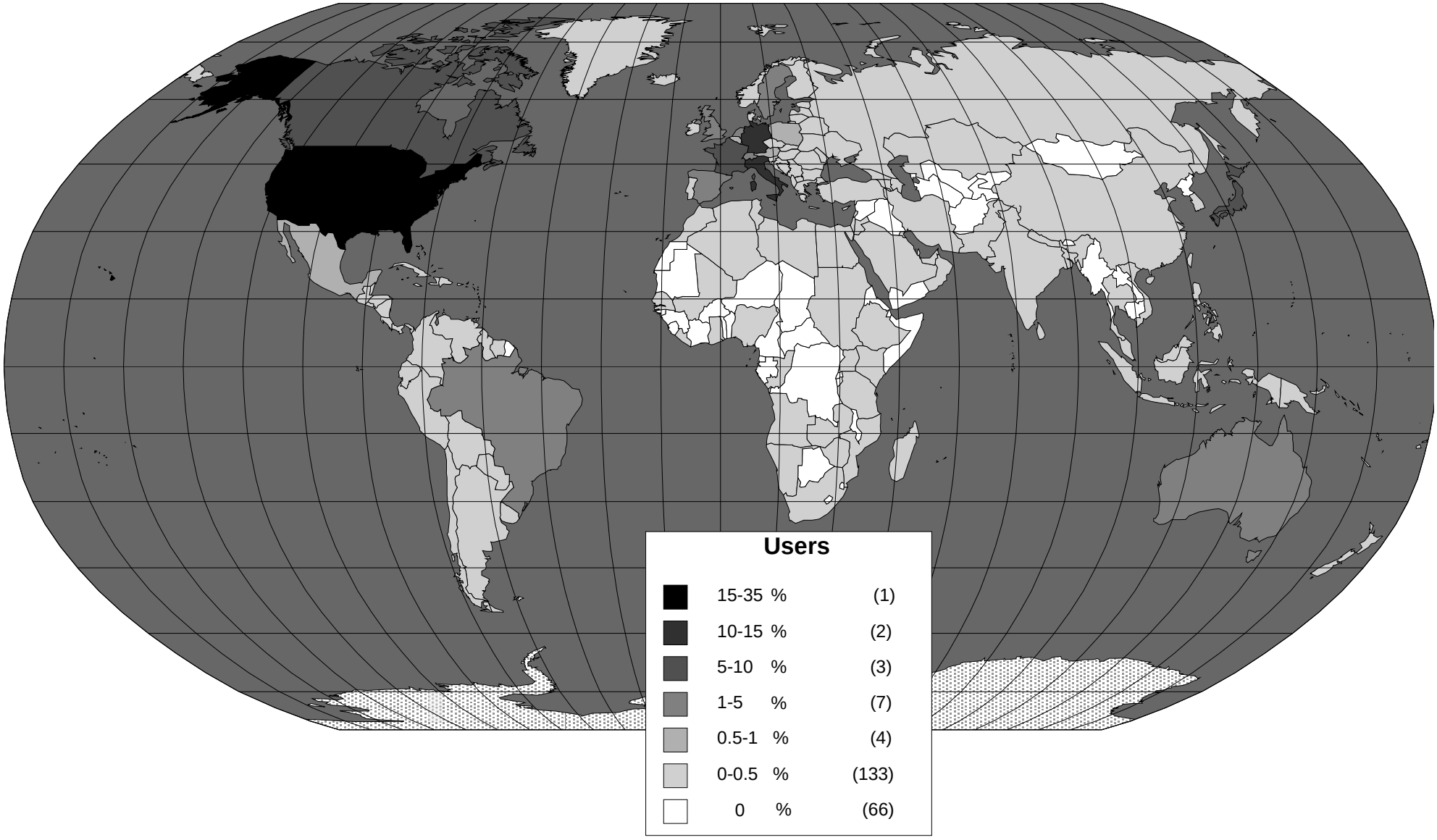

Figure 2: Distribution of Users (Unique log-ins) by Country

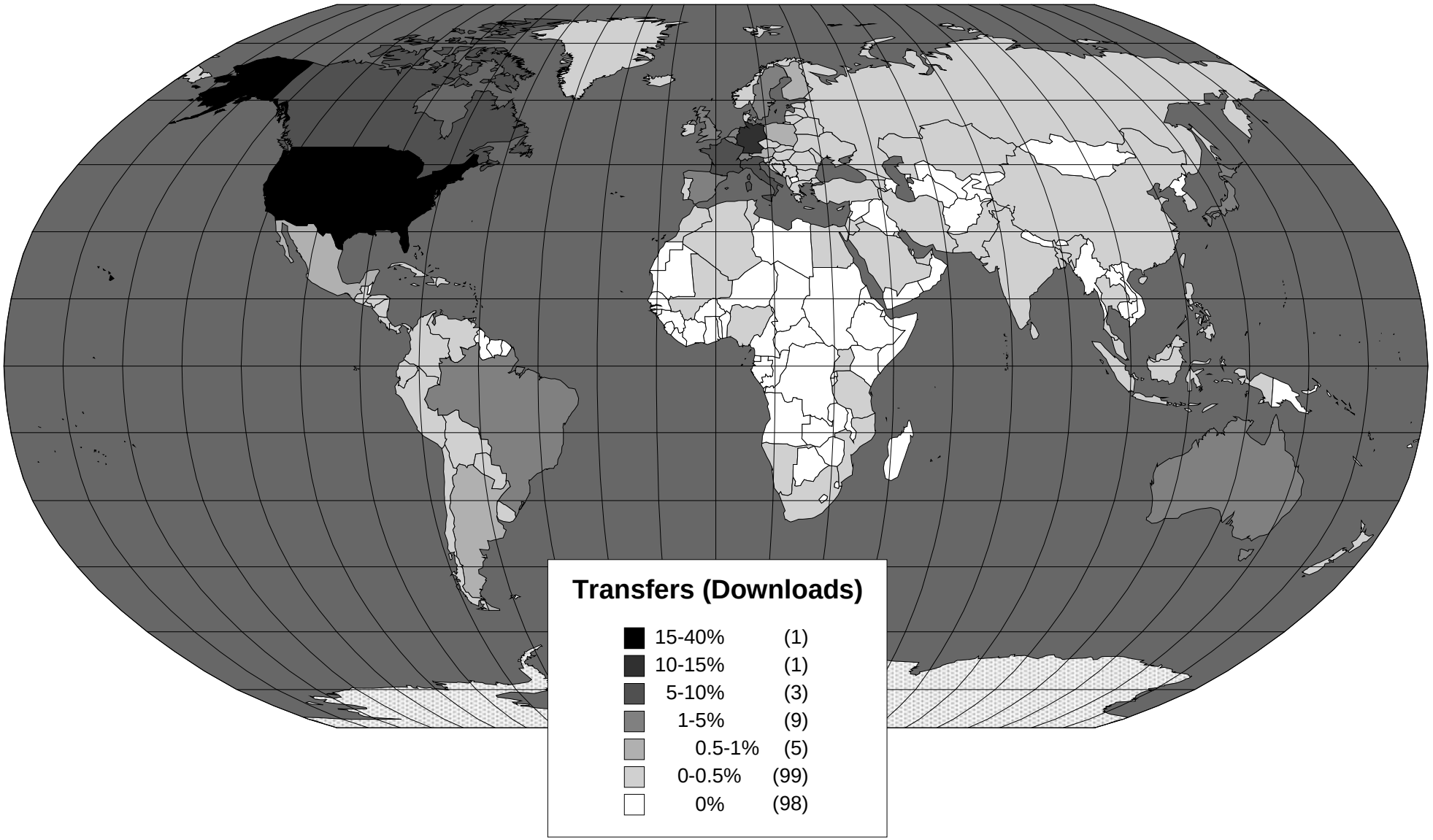

Figure 3: Distribution of Downloads by Country