P2P’s Impact on Recorded Music Sales

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

Koleman Strumpf
UNC Chapel Hill
cigar@unc.edu

1 Introduction

Peer-to-Peer (P2P) networks have markedly weakened the protection for certain forms of intellectual property (IP). Virtually every form of information which can be transformed into a digital file is currently shared and downloaded on networks such as KaZaA and eDonkey. A leading question in economics is whether strong protection for IP is needed to ensure adequate returns to innovation (Plant, 1934; Boldrin and Levine, 2003).

In this paper we focus on the impact of P2P on the sales of recorded music as a test case for the proposition that loosening IP protection markedly damages the initial owner. We match a dataset of over one million downloads collected during the last third of 2002 to weekly US record sales for over five hundred albums. To establish causality we instrument for downloads using technical features of P2P networks, such as internet congestion, which are plausibly exogenous to sales. We find there is little robust evidence that P2P has significantly hurt record sales. This is the first empirical analysis of the effect of file sharing on music sales using actual data collected from a P2P network.

2 Data

The download data were collected from a pair of OpenNap servers which operated continuously over September through December of 2002. One server was linked through a hub to a subnetwork which averaged seven other servers. The latter network had about five thousand simultaneous users which is roughly the connected set on KaZaA (Dotcom Scoop, 2001; giFT-FastTrack CVS Repository, 2003). Our accompanying paper provides more details on these data, and shows that the distribution of downloaded files is representative of other P2P networks and does not vary as the network size grows (Oberholzer-Gee and Strumpf, 2004).

From this data we used a Perl script to extract audio files which were downloaded by clients with a US I.P. address. After omitting duplicate and interrupted downloads, 260,889 files satisfy this criteria. These files were matched to a database of 680 music albums which is a representative sample of top-selling albums in eight music genres (Alternative, Hard Music, Jazz Current, Latin, R&B Current, Rap Current, Country, and Soundtracks) and from charts for three categories which may be uniquely impacted by P2P (Top Current, New Artists, and Catalogue). In total 47,709 audio downloads could be matched to tracks on these albums.

For each album we have weekly Nielsen SoundScan sales data. There are 10,093 album-weeks with complete data.

3 Methodology

The main specification we consider is,

$$(1) \qquad S_{it} = \gamma D_{it} + \boldsymbol{\omega t} + \nu_i + \mu_{it}$$

where S_{it} is observed sales for album i in week t, and D_{it} is the number of downloads. $\boldsymbol{\omega t}$ is a polynomial time trend (of order six), ν_i is an album fixed effect, and μ_{it} is the error term. The parameter of interest is γ. The main difficulty in estimating (1) is unobserved heterogeneity. Album sales decay at different rates following their release and are differentially impacted by the surge in demand during the holiday season. More intuitively, unobserved time-varying popularity drives both downloads and sales. Since we expect Corr(D_{it},μ_{it})>0, OLS estimates of γ will have a positive bias (a formal model of downloads which highlights this result is presented in Oberholzer-Gee and Strumpf, 2004).

To establish causality we use an instrumental variables approach. The instruments Z_{it} should influence downloads but should not directly influence sales. The Z_{it} provide an exogenous source of variation in downloads. More formally, we first estimate the equation,

$$(2) \qquad D_{it} = Z_{it}\delta + \boldsymbol{\omega_2 t} + v_{2i} + \mu_{2it}$$

and then use the fitted D_{it} in (1). The identifying assumptions are that the instruments are relevant (δ is statistically significant in (2)) and valid ($Z_{it}\perp\mu_{it}$). The orthogonality condition is tested in the context of an overidentified model in the Results section.

We use instruments which should influence the user cost of downloading and vary over time (Oberholzer-Gee and Strumpf, 2004 provide a theoretical justification and discuss data sources). One set of instruments is based on internet weather/congestion. The idea is that congestion reduces downloading because it implies greater time costs for users. A variety of congestion measures are used including website load time (Keynote Consumer 40 Index), ping times between over three hundred internet locations (Internet End-to-End Performance Measurement, IEPM), and the fraction of Internet2 backbone traffic from file sharing (Internet2 Netflow Statistics). There is variability across time in these measures. A second instrument is the fraction of German school children on holidays. This is relevant since in our data US users download 16% of their files from Germany. Most file sharers in Germany are youth who access the internet from home. School holidays provide a surge of supply in files as the kids access the P2P networks. The greater supply makes it easier for US users to find and also to download files. Note that the school holidays are not simply contemporaneous with US holidays or the Christmas period. The final instrument is based on the set of competing albums. As albums are released, supply of related music is crowded out perhaps reflecting limits in storage space. We consider the set of albums within the same genre to be competitors. Because of concerns about endogenous release dates, we consider the distribution of a non-market characteristic (mean album time). Notice that this third measure varies over time (due to new album releases) and album (since it is genre-specific and the album in question is excluded from the calculation).

4 Results

The estimates relating downloads to album sales are presented in **Table 1**. To fix ideas, column (I) is an OLS estimate of equation (1) without album fixed effects. The parameter on downloads is positive and statistically significant. Column (II) adds fixed effects which markedly reduces the parameter on downloads (though it is still positive and significant). The fixed effects absorb the album-invariant unobserved heterogeneity, and so the reduction in the download parameter reflects the bias discussed in the Methodology section.

Columns (III) and (IV) consider the more appropriate system of equations (1) and (2), where instruments for downloads are used. Column (III) uses just the German school vacation instrument, which has the expected positive relationship with downloads. After instrumenting, downloads no longer have a statistically significant effect on record sales. Column (IV) repeats the estimates using the full set of instruments which again have the anticipated effect in the first stage (for example, greater internet congestion is associated with fewer downloads). Having multiple instruments allows for an overidentification test. Using the Sargan test we cannot reject a null of 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 sales equation. In this richer specification, downloads continue to have an insignificant effect on sales.

We next consider the robustness of the estimates. Column (V) first differences the data, since inference and consistency is compromised in the presence of non-stationary. The number of downloads continues to have no statistically discernible effect on sales. Column (VI) omits observations from December, since it is possible that downloads are less substitutable for purchases during the gift-giving holiday season. There is still no statistically significant effect of file sharing on sales.

We finally address two other potential criticisms of the approach (in the interest of brevity, the estimates themselves are omitted). One argument is that music consumers are now segmented into two populations, purchasers and downloaders. Each group obtains music from exactly one source, so downloaders get all of their music from P2P and never purchase albums. Under this view sales will decline as more individuals opt out of purchases, but our relatively short-term data might not detect such as secular trend. To address this, we take advantage of the large growth of P2P over our observation period (the number of simultaneous users on the largest network grew by roughly 50%). An implication of this is that our sampling rate declines over time because the servers for which we have data handle a limited number of users and growth in the file sharing community is managed by additional server capacity (which we do not observe). We test this hypothesis by scaling up the number of downloads in our sample (based on the number of KaZaA users each week), so that they reflect growth in the file sharing community. If anything, scaling downloads in accordance with the "two separate groups" hypothesis increases our estimates of the effects of file sharing (the parameter on downloads lies in the range 0.15-0.18, depending on the specification, though it continues to be statistically insignificant).

A second criticism is that we force the effect of downloads to be immediate, whereas it could occur weeks later. For example, an individual might download tracks from an album today which he did not intend to buy until a week or two later. To address this point, we estimate models where one to three lags of sales are included (under the Koyck model, this is equivalent to including all possible lags of downloads). These dynamic panel models are estimated using first-differenced GMM, and finite-sample corrections of standard errors are applied. Still we continue to find no statistically significant relationship between downloads and sales.

5 Discussion

We find little evidence for the claim that file sharing has materially impacted the sale of music albums. To put our results in perspective, consider the most pessimistic estimates which are in column (IV) of Table 1. Some simple calculations indicate that the parameter estimate implies that file sharing reduced annual sales by 2 million albums. This is not a large quantity for an industry which ships over 800 million albums per year.

More broadly, our estimates suggest the recent decline in sales— album purchases have fallen by 139 million from 2000 to 2002—is not primarily the result of file sharing. This implication is plausible for several reasons. First, theoretically P2P could boost sales. This is because downloads allow users to learn about new albums or even music genres. Similarly, there could be no effect if users are time-rich but cash-poor and would not have purchased their downloaded music even in the absence of P2P. Second, several other factors could explain the sales drop. Some candidates include a poor economy, a shift in demand to competing products like video games or DVDs, an end of a period where consumers repurchased albums they owned on older media like vinyl and cassettes, and the consolidation in radio (radio listenership fell by 7% between 1999 and 2003). Third, other products which are widely downloaded, including software, video games, and movies, have not experienced a sales decline.

Our results have potential implications for the social welfare impact of P2P networks. File sharing should not markedly influence the incentives to create and sell music, since we do not find a large impact on sales (in Oberholzer-Gee and Strumpf, 2004 we find there is a very small negative effect on low selling albums, but because such albums are rarely downloaded the total effect is not economically important). At the same time, file sharing has led to a vast increase in music consumption. This helps reduce the deadweight loss associated with oligopolist pricing in this largely five-firm industry. In the case of recorded music at least, weakening IP protection does not appear to have markedly hurt property owners and may be welfare enhancing through increases in consumption.

References

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

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

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.

Oberholzer-Gee, Felix and Koleman Strumpf (2004). “The Effect of File Sharing on Record Sales An Empirical Analysis.” UNC Chapel Hill working paper.

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

Table 1. Estimates of Equations (1) and (2)

	(I)	(II)	(III)		(IV)		(V)		(VI)	
									w/o holiday sales	
	sales	sales	1^{st} stage downloads	2^{nd} stage sales	1^{st} stage downloads	2^{nd} stage sales	1^{st} stage downloads	2^{nd} stage Δ sales	1^{st} stage downloads	2^{nd} stage Δ sales
# downloads	1.193	0.281		-0.001		-0.014				
	(0.022)**	(0.025)**		(0.195)		(0.175)				
Δ # downloads								0.088		0.129
(instrumented)								(0.49)		(0.236)
German kids on			0.670		0.366		0.370		0.038	
Vacation (million)			(0.054)**		(0.123)**		(0.113)**		(0.099)	
Internet Consumer 40					-1.122		-0.820		0.897	
Performance Index					(0.347)**		(0.273)**		(0.393)*	
Internet average					-0.184		-0.164		-0.168	
roundtrip time (ms)					(0.059)**		(0.048)**		(0.073)*	
Internet std deviation					0.135		-0.332		-0.052	
roundtrip time (ms)					(0.079)**		(0.149)*		(0.077)	
Internet2 net flow:					-0.260		0.102		-1.743	
% file sharing					(0.069)**		(0.065)		(0.288)**	
Mean album time					0.126		0.156		0.189	
"other" albums					(0.043)**		(0.086)		(0.084)*	
Polynomial time trend of degree six	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Album Fixed Effects?	no	yes	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	285.836	-4.043
	(5.470)**	(3.753)**	(1.602)**	(3.799)**	(17.652)*	(3.821)**	(25.172)	(0.62)	(53.172)**	(16.532)
P										
Observations	10093	10093	10093	10093	9991	9991	9320	9320	6616	6616
Prob *F*>0 on excluded instruments			0.000		0.000		0.000			0.000
Sargan test (p-value)					0.1715			0.586		0.6807
R-squared	0.23	0.03	0.029	0.005	0.0139	0.0104	0.029	0.01	0.04	0.01

Dependent variables are album sales (1,000s) and # downloads at the 1^{st} stage. Specification (V) and (VI) estimate the model in first differences. The first-stage instruments in these models are also first-differenced. Model (VI) excludes the last four weeks of data (December sales) to see if the holiday shopping season influences our results. The Sargan statistic is an overidentification test. 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