REGULAR ARTICLE

The impact of internet file-sharing on the purchase of music CDs in Canada

George Barker[1] · Tim Maloney[2]

Abstract This paper re-examines data from a survey commissioned by Industry Canada on the effects of internet peer-to-peer (P2P) file sharing activities on music purchasing behaviour. The survey was designed to "inform Industry Canada's policy development work" (Quote from project Description from Industry Canada's website at http://www.ic.gc.ca/eic/site/ic1.nsf/eng/01464.html downloaded 23 January 2012) and the need for copyright law reform in Canada in light of the technological innovation posed by P2P file sharing. The *Journal of Evolutionary Economics* published a study of the Industry Canada data by Andersen and Frenz (AF) in 2010 which claimed to show "… no association between the number of P2P files downloaded and CD album sales (Andersen and Frenz 2010 ibid p 374)," and "… that P2P file-sharing is not to blame for the decline in CD markets. Music markets are not simply undermined by free music downloading and P2P file-sharing (Ibid p375)." Our paper corrects a number of fundamental errors in this analysis of AF, in particular the fact AF biased their results by excluding from their analysis the group of consumers who had completely stopped purchasing CDs (potentially because of P2P activity) prior to 2005. This is the very group who were most responsive, or likely to have substituted P2P downloading for CD purchases. We use longitudinal analysis of how reported changes in P2P downloading by individuals related to their reported changes in CD demand between 2004 and 2005 to better test the hypothesis of whether P2P downloading may reduce CD demand. Contrary to AF's results we find negative and generally statistically significant partial correlations between CD purchases and P2P downloads under a number of specifications and sample definitions. The range of these estimated correlations is between −0.047 and −0.061. This implies that a 10 % growth in P2P downloads is associated with between a 0.47 and 0.61 % decline in CD purchases. Our estimated

✉ George Barker
  george.barker@anu.edu.au

  Tim Maloney
  tim.maloney@aut.ac.nz

[1]   ANU College of Law, Australian National University, Canberra ACT 0200, Australia

[2]   Department of Economics, Faculty of Business and Law, Auckland University of Technology, Auckland, New Zealand

relationships between CD sales and P2P downloads are broadly consistent with market sales data up to the time of the Industry Canada survey, unlike AF.

Keywords Copyright law · Economics · Internet peer-to-peer (P2P) file sharing Music · Downloads

JEL Classification K11 · K29

1 Introduction

As pointed out by economist Blaug (2005), despite Schumpeter's emphasis on innovation, "oddly enough he never discussed, or even mentioned, intellectual property rights." [1] This might seem odd given Schumpeter maintained that the essence of capitalism is economic change driven by innovative activities. But does this imply he thought intellectual property rights (IPR) didn't matter for innovation? Mark Blaug addresses why Schumpeter neglected IPR, concluding that,

> it was the rise of property rights economics in the 1970s, and especially the 1980s, that finally tied together the old labels of patents, copyrights and trademarks in one label of IPR, giving rise to our question that would simply have made no sense to anyone writing in 1942. The economic rationale for patents, copyrights and trademarks was itself transformed by the property right approach, stemming from Coase's objection to Pigovian welfare economics and the emergence of law and economics as a distinct disciplinary subject. [2]

As the former Director General of WIPO, Kamil Idris, has written, recent economists, such as William Baumol, have attempted,

> to pick-up on Schumpeter's work … [and] have postulated that an IP system is to be regarded as an important factor influencing the behavior of the entrepreneur in encouraging innovators, applying the innovation, introducing it into the economy, and marketing the product in a creative or innovative way. [3]

Recent work by Deborah Spar[4] further suggests, that the history of technology and economic development reveals a four-phase cycle, which involves clear interactions with the legal system including: (i) a phase marked by technological discovery; (ii) a phase in which there is technological disruption; (iii) a transitory phase marked by the lack of clear rules and enforcement as society wrestles with the implications of the technology; and, (iv) a final phase in which well-defined property rights re-emerge to establish clear market rules – (a discovery-disruption-anarchy-property cycle).

---

[1] Blaug (2005) at p.69
[2] *Id.*, p.72. Also see S.G. Medema (1995).
[3] Idris (2005), p.27
[4] Deborah Spar, *Ruling the Waves: From the Compass to the Internet, a History of Business and Politics along the Technological Frontier*, 2001, p.14.

While the expansion of the technological frontier briefly creates a stage of "creative anarchy" that allows "pirates" to operate "almost without restriction", Professor Spar argues that this stage - at least in Western economies - was soon replaced by clear rules that allowed entrepreneurs to flourish. For Spar, the lesson for policy makers was to ensure that the initial period of creative anarchy was expeditiously replaced with a clearly defined and well-functioning set of property rights:

> [Entrepreneurs] must have property rights … and some means of enforcing them. Without property rights, rival claims over the new market can rapidly disintegrate into chaos, stunting commercial development as it did in the early days of the telegraph and broadcast radio. … The establishment of property rights is one of the most crucial events along the technological frontier. It allows the market to unfold in a predictable way, and gives pioneers a hefty dose of ownership and security. Most important, perhaps, the creation of property rights also marks the difference between pioneers and pirates, between those whose claim on new technology is legitimate and those whose is not.[5]

No doubt, we are observing profound changes in markets, business and commerce as a result of the technological revolution associated with Information and Communications Technology (ICT). In a short space of time, this revolution has given birth to the rapid spread of the internet and the growth of new forms of exchange. But this technological revolution has relied on law generally, and IPR in particular, to flourish. In many respects, its emergence is a by-product of the many patents, copyrights, and trademarks that have supported the required investment in, and commercialisation of, underlying inventions that make ICT possible (including the IP underlying electricity, telecommunications, and computers industry, and the associated hardware and software they depend on).

Given the output of the new so-called digital economy is created, stored, transacted and consumed in digital form, we are now moving into the phase when copyright law in particular may be critical to the further growth of investment and exchange on the internet. The recent growth of so called internet piracy is thus seen by some as posing a threat to this, requiring the strengthening of copyright law enforcement. P2P file sharing for example, described in greater depth in the next section enables downloading from the internet of digital files (music, film, e-books, video games and software) for free, in breach of copyright - so called digital piracy. It would appear hard for creative digital works to compete with pirated copies. Perhaps when viewed in the broad frame of technological and economic history offered by Spar then, the Internet may simply be in the early phase of a familiar cycle that has happened many times before according to Spar, requiring a move from piracy to enforcement of a clearly defined and well-functioning set of property rights.

As many economists in the law and economics, and property rights tradition have indeed underscored, the raison d'etre of copyright enforcement in the first place (and IPR enforcement generally) is to respond to the so called "Free Rider" problem with creative goods, which may indeed be manifest in internet piracy. Free riders use or benefit from copyright material without paying for it. Such free riding reduces the

---

[5] Spar, *supra*, pp. 371, 374.

ability of the copyright owner to earn an adequate return on their investment, and ultimately acts as a disincentive for the creation of further original material, and weakens incentives to disseminate. As noted by Harold Demsetz,

> Efficiency calls for controls or institutional changes that limit the degree of free riding if these changes are less costly to implement than is the resource misallocation costs created by free-rider behavior. Goods whose existence requires significant costs of creation would seem to demand more legal restraints on entry.[6]

Copyright law may thus not involve a trade-off or exacerbate scarcity as some have suggested recently (Lemley (2005), p1055: Lessig (2004)) but rather alleviate it, bringing forth more and new information and creative goods to meet the expanding demand for them.[7]

The foregoing suggests the importance of empirical work examining the implications of technological innovations on the internet, like P2P file sharing, for institutional and legal arrangements like copyright – and vice versa. For this reason this paper re-examines data from a survey commissioned by Industry Canada which was designed precisely to "inform Industry Canada's policy development work"[8] in the context of copyright law reform in Canada with a particular focus on the technological innovation called P2P file sharing.[9] The Journal of Evolutionary Economics published a study of this Industry Canada data in 2010 by Andersen and Frenz (AF) which claimed to show "… no association between the number of P2P files downloaded and CD album sales,"[10] and "… that P2P file-sharing is not to blame for the decline in CD markets. Music markets are not simply undermined by free music downloading and P2P file-sharing."[11] This tended to question the predictions of the economics literature on copyright outlined above, and suggest the irrelevance of calls to strengthen copyright law in the digital age, in that the main example of its breach in the case of free music downloading and P2P file sharing on the internet did not appear to hurt the music industry.

We re-examine the Industry Canada data to isolate the effects of internet peer-to-peer (P2P) file sharing activities (so called music piracy) on music purchasing behaviour. In what follows, we first provide a brief background on the music file sharing debate and the importance of the question whether P2P file sharing adversely affects legitimate sales, then outline the nature of the survey data, and then summarise and critique the analyses by Andersen and Frenz which suggested that P2P file-sharing activity could act as both a substitute and complement to music purchases in CD markets. The empirical results

---

[6] Harold Demsetz, "Creativity and the Economics of the Copyright Controversy", 6(2) *Review of Economic Research on Copyright Issues* 5 at p.11 (2009).

[7] See Landes and Posner (2003), who argue that "information is a scarce good, just like land" (p. 374).

[8] Quote from project Description from Industry Canada's website at http://www.ic.gc.ca/eic/site/ic1.nsf/eng/01464.html downloaded 23 January 2012

[9] See description of P2P file sharing in the next section

[10] Andersen and Frenz 2010 ibid p 374

[11] Ibid p375

of the authors purportedly showed that P2P downloading had no statistically measurable impact on CD sales.

We identify several key issues with this previous work by Andersen and Frenz. Of particular concern was that that the authors used only a sub-sample of the survey results, excluding those who did not purchase CDs in 2005. We believe that this biased the results by excluding from this analysis those individuals who had completely stopped purchasing CDs, possibly due to increased P2P activity prior to 2005 (i.e., the very group who may have been most responsive or most likely to have substituted P2P downloading for CD purchases). Of further concern is the fact that these previous studies examined the cross-sectional relationship between the P2P downloading and CD purchases. The problem with this approach is that both P2P downloading and CD demand may be high (or low) because both variables are influenced by a latent third factor (such as preferences for music). As a result, any positive correlation between the variables may be mistaken as a causal association. Furthermore, we provide support for our contention that the instrumental variable approach employed by Andersen and Frenz was ill-advised because their chosen instruments were inappropriate in the context of an underlying demand system for music from these different sources.

We then present our own empirical analysis which uses the same variables on P2P downloading, CD purchases and covariates as this earlier study. However, we make five critical amendments to this previous study:

1. We include in our sample the group (of over 400 individuals) that had zero CD purchases in 2005. This group had been excluded from the 2010 study by Andersen and Frenz because it was claimed that they "… may never have been active in CD purchasing." We also test the impact of the exclusion of respondents who provided inconsistent or non-useable answers to questions regarding P2P downloading and CD purchases. Furthermore, we exclude four survey respondents who stated that they had downloaded more 400 music tracks per month in either year 2004 or 2005 as outliers.

2. We test the impact of the use of weights employed in the previous regression analysis. Because these weights were developed for use with the full sample ($n=$ 2100), there is some doubt whether or not these weights continue to be appropriate for much smaller sub-samples.

3. We do not use the price variable included by Andersen and Frenz, because it is unlikely to be indicative of any true variation in market prices facing different individuals in this sample. It was based on information solicited from respondents on the average prices that they paid for the CDs that they purchased. Rather than representing actual variation in CD prices faced by consumers at a given point in time, this variable is more likely to represent differences in either the type of CDs purchased (e.g., latest hits or obscure older music) or where these CDs were purchased (e.g., speciality music stores, discount department stores or second-hand sales).

4. We use a "first-differences" estimation approach that examines how a change in P2P downloads is related to a change in CD purchases for the same individual over two consecutive years. This differencing eliminates fixed effects, where both CD demand and P2P downloading may be high (or low) for an individual because both behaviours are influenced by latent or difficult to measure variables – like love of music.

5.   We finally control for the fact that an individual's consumption of CDs, cannot fall
     below zero. To do this we estimate a lagged demand structure that is equivalent to a
     first-differences model (subject to certain restrictions), but which enables us to use
     a "Tobit" estimation technique that allows for censored data.

2 A brief background on P2P music file sharing

File sharing first came to attention when Napster began operations in mid-
1999. Its purpose was to allow music files to be copied and shared among
strangers for free. Napster worked by storing music files on a central server
and allowing people access to copy or download them for free and in breach
of the owner's copyright. Napster quickly grew until it was shut down by a
preliminary injunction granted to the recording industry in February 2001.[12]
With Napster's closure, numerous other file-sharing programs emerged, partic-
ularly in a new form called peer-to-peer (P2P) file sharing. Unlike Napster,
P2P file sharing was not based on a central server and instead allowed one
computer on the Internet to search for and access files on the hard drives of
other computers that were connected to the Internet. Successful lawsuits
against companies engaging in P2P file sharing around the world (including
Grokster, StreamCast,[13] and Kazaa[14]) established they could be sued for
inducing copyright infringement. Kazaa, for example, settled with the music
industry in 2006 and is now understood to be running under licence as a legal
music subscription service by Atrinsic.

   BitTorrent websites use a more sophisticated technology that makes it easier
for users to find and download large files such as movies. BitTorrent sites,
such as Vancouver-based isoHunt, collect, categorize, index, and make avail-
able BitTorrent files for download. In December 2009, a U.S. federal district
court found isoHunt liable for massive copyright infringement. The court cited
unrebutted evidence that 95 % of the files traded through isoHunt's sites are
likely infringing. In his judgment, United States District Court Judge Stephen
Wilson explains that these new sites are no different than those P2P file-
sharing sites previously mentioned:

> "Defendants' technology is nothing more than old wine in a new bottle. Instead
> of logging into a proprietary network in order to download files from each others'
> computers, Defendants' users access Defendants' generally-accessible website to
> download those files." [Columbia Pictures Indus., Inc. v. Fung, 2009 WL
> 6355911 (C.D. Cal. Dec. 21, 2009) at 9].

---

[12] See A&M Records v. Napster, Inc., 239 F.3 d 1004 (9th Cir.) 2001.

[13] See MGM Studios, Inc. v. Grokster, Ltd. 545 U.S. 913 (2005), a United States Supreme Court decision in
which the court unanimously held that defendant P2P file-sharing companies Grokster and StreamCast (maker
of Morpheus file-sharing software) could be sued for inducing copyright infringement for acts taken in the
course of marketing file-sharing software. The plaintiffs were a consortium of twenty-eight of the largest
entertainment companies (led by Metro-Goldwyn-Mayer studios).

[14] See Universal Music Australia Pty Ltd v. Sharman License Holdings Ltd [2005] FCA 1242 (5 September
2005), available at http://www.austlii.edu.au/au/cases/cth/federal_ct/2005/1242.html.

The companies engaged in providing file sharing technology make millions of dollars by selling advertising. They often use incentives to encourage users to upload the most popular files to their sites in order to increase traffic and create better value for advertisers.[15] Furthermore, they do not pay the artists for the creative content they are giving away for free. The fact that the so called "pirates" make a lot of money is suggested by the fact that Kazaa's owners were able to settle their legal issues by paying $100 million in reparations as well as agreeing to close down the pirate file-sharing service.[16]

Any individual on a file-sharing network can make available any file on his or her hard drive to all other members of the file-sharing network. Stan Liebowitz noted:

> "The term "file sharing" is actually something of a misnomer, however. Individuals do not "share" the files that move back and forth on the Internet. They do not experience these files together nor are they likely to ever meet or even know one another. Nor do they lend or trade the files among one another, since the files are not borrowed or given back. A more appropriate term might be "anonymous file copying," since that reflects what actually occurs. The end result of file sharing is that individuals who do not own and have not purchased a particular song or movie can nevertheless obtain that song or movie from unknown third parties." (2006, p. 5)

Because music files are easily compressed, relatively small, and very popular, they have been the most common creative content to be the subject of file sharing. However, with new technologies, like the BitTorrents described above, and as Internet transmission speeds increase, file sharing is now also affecting full-length movies, e-books, and computer programs. But does file sharing affect legitimate sales? If not it may not affect investment and therefore the supply of creative works. This is the question we seek to address. In what follows we first briefly review the result of past studies on the effect of file sharing on music sales which indicates that Andersen and Frenz results are an outlier. We then turn to directly examine the Industry Canada data to see what it may have actually shown.

3 The academic literature on the impact of file sharing

A recent comprehensive review of academic articles published in reputable journals on the effects of file sharing on music sales has concluded that the vast majority of these studies all show file sharing caused harm (Liebowitz 2014). Ten published articles all show that file sharing reduced sales, these include Hong (2007, 2013), Liebowitz (2006, 2008), Michel (2006), Peitz and Waelbroeck (2004), Rob and Waldfogel (2006), Waldfogel (2010) and Zentner (2005, 2006). Two further

---

[15] See isoHunt decision.
[16] See http://www.mtv.com/news/1537233/kazaa-settles-with-music-industry-for-100-million-promises-to-go-legit/

unpublished papers also found that file sharing reduced music sales (see Blackburn (2004) and Zentner (2009)).

The cumulative decline in revenues since Napster arrived in 1999 has been 62 % in real terms from 2000 to 2013. Liebowitz (2014) concluded that the majority of all studies support a conclusion that the entire decline in sound recording sales can be explained by the growth in file-sharing.

By comparison there are in fact no articles published in reputable academic journals that find a positive impact of file-sharing on sound recording sales, and only two published articles that do not find harm. The two articles that failed to find harm were by Oberholzer-Gee and Strumpf (2007) and the second by Andersen and Frenz (2010). There are many reasons to question the results from the Oberholzer-Gee and Strumpf (2007) study (see Liebowitz 2007a, 2007b, 2010). It is not possible to replicate the Oberholzer-Gee and Strumpf results, as they do not make their data available for further analysis. Discounting Oberholzer-Gee and Strumpf paper therefore as not verifiable and suspect, the only paper that does not find harm from unauthorized file sharing is Andersen and Frenz.

Therefore, this paper takes the survey data from Canada used by Andersen and Frenz and tests the validity of their results. The survey was funded by Industry Canada in order to support its policy decisions regarding the copyright regime in Canada. The survey used a large nationally representative sample, and was designed to measure the extent to which peer to peer (P2P) file-sharing activities act as substitutes or complements to music purchases. This Industry Canada Survey dataset is therefore relatively unique, being the only dataset known to us based on a nationally representative sample that contain person specific CD and P2P quantities. The fact that the survey was a large nationally representative sample commissioned by Government that sought to directly address the question does file sharing displace sales, suggests it is a dataset that should have be given considerable weight and attention.

4 The Canadian survey and the resulting data

The survey was conducted by Decima Research for Industry Canada in 2006. Telephone interviews were conducted with 2100 randomly selected Canadians (15 years and older) across the country between April and June 2006.

The questionnaire contained two general types of questions on current music acquisition behaviour and the different methods used for acquiring music (e.g., buying CDs and P2P downloads). Respondents were first asked whether or not they acquired music by a particular method (buying CDs, copying MP3 files from friends, ripping music tracks from CDs, or downloading music from pay-sites, promotional websites, private internet websites or P2P file sharing networks). Follow-up questions asked the same individuals to estimate the number of units that they may have acquired by a particular method.

Table 1 provides descriptive statistics of some of the key variables taken from this survey. Slightly less than one-half of respondents claimed to have downloaded free music tracks from P2P file sharing networks in 2005. Of those who had engaged in this activity, the mean number of tracks downloaded per month was slightly under 30 in 2005. The mean number of monthly music tracks downloaded for these same

Table 1 Descriptive Statistics

Variable Description	Mean	Standard deviation	Minimum	Maximum
Downloaded Music Tracks from P2P Websites in 2005? (1=Yes, 0=No)	0.476	0.500	0	1
If Yes, Average Monthly Music Tracks Downloaded in 2005	29.982	59.952	0	500
If Yes, Average Monthly Music Tracks Downloaded in 2004	19.814	44.202	0	500
Purchased Music CDs in 2005? (1=Yes, 0=No)	0.797	0.402	0	1
Number of Music CDs Purchased in 2005	8.645	11.826	0	100
Number of Music CDs Purchased in 2004	8.908	11.813	0	150
Perceived Price of CDs Purchased in 2005	$17.723	$4.075	$1	$45
Perceived Price of CDs Purchased in 2004	$17.669	$4.288	$2	$45
Proportion Female	0.507	0.500	0	1
Proportion 24 Years Old or Less	0.391	0.488	0	1
Proportion 55 Years Old or More	0.156	0.363	0	1
Proportion with Less Than a High School Diploma	0.241	0.428	0	1
Proportion with More Than a High School Diploma	0.577	0.494	0	1
Proportion Currently a Student	0.280	0.449	0	1
Proportion with Household Income Less Than $20,000	0.102	0.303	0	1
Proportion with Household Income More Than $60,000	0.318	0.466	0	1
Proportion with Internet Skills Described as 'Very Skilled' or 'Skilled'	0.507	0.500	0	1
Proportion with Internet Skills Described as 'Not Very Skilled' or 'Not at All Skilled'	0.196	0.397	0	1

Data from the Decima Research for Industry Canada Survey 2006 (n=2100 individuals)

individuals was slightly under 20 in the previous year 2004. Thus, there is evidence of a substantial increase in downloaded tracks for the average person who engaged in this activity.

The survey also asked respondents about their CD purchases in both years. Nearly 80 % of individuals said that they had purchased music CDs in 2005. Across all respondents, the mean number of music CDs purchased annually declined slightly from 8.908 in 2004 to 8.645 in 2005. This difference is not statistically significant at conventional test levels. It should be noted that there were systematic differences in the way information was solicited on CD purchases and P2P downloads. All respondents were asked about CD purchases in 2004, while only respondents who downloaded P2P music in 2005 were asked about similar outcomes in 2004. We assume that the rationale for the different treatment of these variables was based on the belief that only someone downloading in 2005 could have downloaded tracks in 2004. There is no way to test the validity of this assumption in the data available from this survey. This forces us to make an assumption about P2P downloading activities in 2004 for a sizable proportion of the sample. Either we assume that the missing information implies no downloading activities in 2004 (i.e., both 2004 and 2005 P2P downloads were zero), or that the missing information means that these observations

are invalid for looking at the change in downloading activities relative to changes in CD purchases in the same years. We test the robustness of our results by employing these alternative sample definitions.

Respondents were also asked about the average price that they paid for music CDs in both years. Nothing was done to limit this information to particular types of CDs (e.g., new releases or bargain bin castoffs) or particular outlets (e.g., premium music stores, discount department stores or second-hand dealers). This makes it difficult to assume that these variables reflect true variation in the CD prices facing consumers. This is evidenced by the substantial range in average CD prices in this sample from a minimum of $1 to as high as $45. We argue that for the purposes of this analysis that it would have been far better to use data from other sources (e.g., due to location in different geographic markets) to indicate the true price variation facing consumers.

Information on a variety of other background variables are available in this survey. For example, we have categorical data on the age, education, household income and internet skill level of the respondent. Individuals were asked to rate their own internet skill levels as being 'very skilled', 'skilled, somewhat skilled', 'not very skilled', 'not at all skilled' or 'don't know'. A higher proportion of respondents self-assessed their internet skills as being in the highest two categories (0.507) compared to the lowest two categories (0.196). This variable will feature prominently in our later analysis, because Andersen and Frenz used these data as instrumental variables in their regression analysis.

5 Overview of Andersen and Frenz (2010)

Andersen and Frenz purport to use the data described above to measure the extent to which P2P file-sharing activities act as substitutes or complements to CD music purchases. The authors claim to adopt an "evolutionary economics approach" in their paper. Although the availability of P2P file-sharing might be expected to lead to a reduction in CD music sales, it is hypothesised that the opposite might be true due to "market creation" (e.g., hearing before buying) and "market segmentation" (e.g., downloading single tracks) effects.

To test these effects, Andersen and Frenz initially used a single-equation regression model that specified the number of CDs purchased by respondents as the dependent variable, and regressed this against the respondents P2P downloads as an independent variable, plus a number of other variables collected in the survey, including respondent's age, income etc. The equation they estimated looked something like this:

$$ln q_i^{CD} = \alpha_0' + \alpha_1' ln p^{CD} + \alpha_2' ln q_i^{P2P} + \alpha_3' ln y_i + X\gamma + u_i^{CD} \qquad (1)$$

where the matrix X includes an alternating set of covariates. The emphasis is on the estimate of the slope coefficient on the natural logarithm of the quantity of P2P downloads. If $\alpha_2' < 0$, then an increase in P2P downloads was interpreted as leading to a net decrease in the quantity of CDs purchased (i.e., P2P downloads and CD purchases are net substitutes). If $\alpha_2' > 0$, then the two music sources are net complements.

Andersen and Frenz admitted that there was a possibility of "… an issue of simultaneity in any estimations linking CD sales and P2P file-sharing" (p.729, 2010). In other words, both CD purchases and P2P downloads are choice variables for the individual, and unless all of the determinants of this behaviour are included among covariates in this regression model, the estimated coefficients may be biased and inconsistent. To overcome this issue, Andersen and Frenz employed an instrumental variable (IV) estimation procedure. As always, the key with this approach is to find a valid set of instruments. These must be variables that are correlated with the endogenous explanatory variable but uncorrelated with the disturbance term in the regression equation with CD purchases as the dependent variable. The authors chose to use a set of categorical measures on self-reported personal internet skill levels for this purpose.

Based on this approach, Andersen and Frenz concluded that they could find "… no association between the number of P2P files downloaded and CD album sales,"[17] claiming therefore that "… this paper show (sic) that P2P file-sharing is not to blame for the decline in CD markets. Music markets are not simply undermined by free music downloading and P2P file-sharing."[18] The authors claimed that they found evidence of both a negative market substitution and positive market creation effect. These effects largely offset one another, and the net result was that there was no statistically significant relationship between P2P file downloads and CD music sales.

This published paper in 2010 by Andersen and Frenz is actually a revised version of an earlier study for Industry Canada produced in 2007, entitled "Don't blame the P2P file-sharers: the impact of free music downloads on the purchase of music CDs in Canada",[19] which used the same data. In the earlier study, the key claim emphasized by Andersen and Frenz was even more counter-intuitive and highly controversial, being that "… our analysis of the Canadian P2P file-sharing subpopulation suggests that there is a strong positive relationship between P2P file-sharing and CD purchasing."

6 Weaknesses in the Andersen and Frenz studies

There are a number of substantial issues with the methodology used in these earlier studies utilizing the Industry Canada dataset. As a result, the conclusions from this prior analysis are highly suspect. In what follows we briefly summarise the five problems with this analysis.

1. Sample Definitions

    Andersen and Frenz excluded from their statistical analysis all answers to the survey provided by participants who said they did not purchase CDs in 2005. There seems to be little justification for this sample restriction which may have substantially biased their subsequent results. Andersen and Frenz excluded over

---

[17] Andersen and Frenz 2010 ibid p 374
[18] Ibid p375
[19] Andersen and Frenz (2007)

400 of the 2100 survey respondents (nearly 1/5th of initial sample). They provided the following explanation for this decision:

"When investigating the effect of P2P file-sharing, the paper includes all respondents also active in CD markets. We did not include people who did not report any activity in those markets, because the group may never have been active in CD purchasing. In this case, they can of course not be blamed for the downfall of the CD market." (Andersen and Frenz 2010, p.726)

Basic econometric practice is to not throw away data like this simply because a variable is censored at zero. Doing so will almost inevitably bias the resulting coefficient estimates. In this case, Andersen and Frenz were excluding from their analysis the group of consumers who had completely stopped purchasing CDs (potentially because of increased P2P activity) prior to 2005. This is the very group of individuals who were most likely to have been responsive to substitution possibilities in these alternative forms of music acquisition.

Moreover, Andersen and Frenz did this without testing the validity of their assumption that the excluded group "… may never have been active in CD purchasing." When we tested their assumption with additional data from the same survey that they ignored in their analysis, we found that nearly one-third (31.6 %) of respondents who had purchased no CDs in 2005 reported purchasing a positive numbers of CDs in 2004. This finding clearly runs counter to their claim that 2005 non-purchasers would never have purchased CDs in the past. Their own dataset refutes this claim, but they failed to explore or test their assumption.

Unlike Andersen and Frenz, we present our results starting with the fullest possible dataset containing valid information obtained from the respondents (e.g., we include data from respondents who said that they did not purchase CDs in 2005). Following conventional econometric practice, we adopt a Tobit estimation procedure in this case to allow for the censoring of CD purchases at zero in 2005. Also unlike Andersen and Frenz, we begin by including observations from respondents who declared initially that they purchased CDs or downloaded music tracks from P2P file sharing networks, but subsequently provided a zero response when they were asked about the number of purchases or downloads in the previous year.[20] Andersen and Frenz eliminated these observations from the outset claiming that the answers were inconsistent and that the data were therefore invalid. This does not seem obvious to us. The question on the ways in which individuals acquire music makes an early reference to the year 2005, but ends with the following request: "Please indicate whether you obtain music in the following ways: …" before listing the various options. It seems possible, therefore, that respondents could indicate that they 'generally' purchase music CDs or engage in P2P downloading even if they didn't do so in the previous year. We test the importance of this sample

---

[20] See Andersen and Frenz (2010, p.726)

inclusion decision later in our analysis by re-estimating our models without these dubious 'zero' observations.

This definition of alternative samples also allows us to consider another vexing question for our analysis. We will rely on *changes* in music acquisition via CD purchases and P2P downloads between 2004 and 2005. Unfortunately, the survey questionnaire did not treat questions in these two areas symmetrically. All respondents were asked about their CD purchases in both years, but only individuals who said they engaged in P2P downloading were asked about this activity in 2004. The implicit assumption here is that someone who wasn't downloading music in 2005 from P2P file sharing networks wouldn't have been engaged in this same activity in 2004. We simply have no information on P2P downloads in 2004 for this subsample. We decided to treat this information in two distinct ways. We first follow the rationale of this questionnaire design and assume that missing information on P2P downloads in 2004 are equivalent to a zero response. In other words, for individuals who didn't download P2P tracks in 2005 we assume that the same was true for 2004. We test the importance of this assumption by removing all observations on individuals who were not downloading music tracks from P2P network sharing sites in 2005 and re-estimating our regressions.

Finally, we also discovered evidence of potential 'outliers' in this dataset. Outliers are logical candidates for exclusion from our analysis to the extent that these observations may have been mis-recorded by the survey administrator, or overstated by survey participants. There is little evidence of outliers in annual CD purchases. The maximum values are 100 in 2005 (10 observations out of 1741) and 150 (one observation out of 1741) and 100 (6 observations out of 1741) in 2004. However, there is some evidence of outliers in P2P downloads. There are, in particular, four observations with P2P downloads greater than 400 per month in either 2004 or 2005 which are clearly well beyond the number of monthly downloads of other respondents in this survey who engage in this behaviour. Remember that this is in terms of *monthly* downloads. This implies that these four individuals are responsible for at least 4800 music downloads per year. Moreover, the exclusion of just these four observations has a substantial impact on the estimated relationship between CD purchases and P2P downloads. Our initial results reported in this paper are based on the exclusions of these four outliers. We subsequently report the effects of re-introducing these four observations into the estimation.

2. Use of Weights

In their 2007 study, Andersen and Frenz note, in relation to their work, that "All following analyses will use weighted data to be representative of gender, age and regional distributions with respect to the Canadian population."[21] They further describe the weights that they use in their analysis as follows:

---

[21] See Andersen and Frenz (2007, p.17)

Sampling weights were constructed in order to scale the number of observations to match the actual Canadian population according to Statistics Canada 2001 Census data…….the weight attached to each survey response is the inverse of the probability of being included in the sample divided by the sample proportion. For instance, if the true proportion of female downloaders under the age of 25 living in Quebec is 1.1 % of the population, and the sample proportion is 4.5 %, then the applied weight to this segment is 0.24… The observations in the survey are scaled up to match the Canadian population. In total there are 2100 observations in the sample that represents a population of around 24 million.[22]

Andersen and Frenz are not as explicit about whether or not they use these weights in their 2010 article. The problem is that given that they excluded the 246 respondents who initially declared that they were P2P downloaders but subsequently did not provide a non-zero response when asked about the volume of their file-sharing and then excluded more than 400 people whose CD demand was zero, the use of the original weights based on a full sample without recalculation could be considered inappropriate.

Indeed, even if one only excludes from any statistical analysis those survey answers that were not useable, the sample size is reduced to around 1760 individuals (or around 80 % of the original sample). Given that only a non-representative subsample is useable, the use of weights in our regression analysis is somewhat contentious. We assume that Andersen and Frenz did not use weights in their 2010 article, and adopt a similar approach in our analysis. However, we report the results from regression analysis with these weights among the set of robustness checks of our findings.

3. Absence of Valid Price Variables

In order to estimate a demand equation (or demand system) and the price elasticities, we need some variation in the relevant market prices faced by the respondents in our sample. Andersen and Franz relied on the self-reported "perceived" CD prices by individuals in this cross section. This is a serious problem with this paper. As they note, there are a variety of places where CDs can be purchased and a range of prices that are charged in these outlets (e.g., speciality music stores, lower-priced supermarkets, on-line vendors, second-hand sales). In addition, CD prices can vary by the type of music purchased (e.g., latest hits, older music, obscure bargain-basement recordings). However, these are approximately the same range of prices that everyone faces. The problem is that the price that people report paying could itself be a function of where, what and how many CDs they buy. Suppose, on average, larger volume buyers have more of an incentive to seek out lower-cost vendors. In regressing the quantity of CDs purchased against the perceived price, we might get a negative relationship, but this would <u>not</u> necessarily be an estimate of the own-price elasticity. The causality may be reversed. What is

---

[22] Ibid p.17

needed here is true variation in actual market prices (maybe by geographic location) in this cross section. This information is not available in this dataset.

Furthermore, there is no information in this dataset on the relevant prices for P2P downloads from file sharing networks. This is partly because there are no direct charges for downloading, and partly because there is no available information on the indirect costs of this activity. Even if these downloads are ostensibly 'free', one could argue that there are implicit costs associated with downloading music files from these sites. For example, there could be a cost of transferring data over the internet, the cost of equipment necessary for gaining internet access and storing data files, and the expected cost of being detected and prosecuted for illegal P2P downloads. Unfortunately, the dataset available for this analysis does not contain the potential components of an implicit price for P2P downloads. Conceptually, such costs could vary substantially across individuals due to their personal circumstances and location. We will see in the next section how accurate information on the variation in true CD and P2P prices could be used in this analysis, and what happens when they are unavailable to the researchers.

4. Additional Hypotheses

Andersen and Frenz (2010) identified a number of hypotheses that they sought to test in their regression analysis. We've already exposed the inherent difficulties is estimating the own-price elasticity of demand for CDs in the absence of true exogenous variation in this price variable. Similar concerns could be raised about some of the other hypotheses mentioned in this study. For example, the authors regress CD purchases against a self-reported perception that the 'album was too expensive'. Although they estimate a negative relationship between these variables, this is hardly convincing evidence of price sensitivity in demand behaviour. Stating that a price is 'too high' could be a mere rationalisation of the decision not to purchase CDs. In other words, the causality could be reversed. Just as concerning is the fact that nearly two-thirds of the original sample were eliminated in this regression estimation, because this explanatory variable was only defined for individuals who engaged in P2P file sharing. To truly estimate this own-price elasticity of demand, there is no alternative to truly exogenous variation in the actual prices faced by consumers.

The authors also stated additional hypotheses that P2P downloading may have positive effects on CD sales through market creation ('hearing before buying') and market segmentation ('not wanting to buy a whole album'). To test these hypotheses, they included individual self-reports on these variables in their regressions. They found evidence of a positive and significant relationship between CD sales and the proportion of music tracks downloaded from P2P networks that occurred because the individual wanted to hear the songs before buying. They found no evidence of any measurable relationship between CD sales and the tracks downloaded from P2P networks because he or she didn't want to buy the whole album. Again, issues arise over the direction of causality in this relationship, and the substantial non-random reduction in sample size for this analysis. It's difficult to interpret these results. Would an exogenous increase in the desire to hear

music before buying lead to more CD purchases? Or are individuals who purchase unusually large numbers of CDs more likely to prefer to hear these tracks before buying?

In the end, the overall relationship between CD purchases and P2P downloads for the entire sample should provide the best evidence of the *net effect* of all of these factors. The increased prevalence of P2P internet sites could stimulate demand for CD purchases. Yet, higher volumes of P2P music downloads could reduce CD sales. Our goal is to consider how these data could best be used to isolate the possible net relationship between these variables.

5. Model Specification

Andersen and Frenz (2010) don't specify a true demand system for the two goods: CD purchases and P2P downloads. A conventional demand system approach would model demand for the two commodities as separate but interdependent equations. We believe that it is preferable to start with this general demand system in order to motivate the eventual equation that is estimated in this study. This structure shows that cross-price elasticities would be the customary way of measuring the relationship between the demands for the two commodities. However, in the absence of valid price variables for these goods in this situation, we need to adopt a more ad hoc approach and directly estimate the relationship between P2P downloads and CD purchases. One important lesson learned from this demand system approach is that variables that might be incorporated in the relevant prices for these commodities would be questionable instruments for this analysis.

Assume that an individual with a given income (y_i) has a well-defined preference ordering over quantities of these two goods (q_i^{CD} and q_i^{P2P}) and faces relevant market prices (explicit or implicit) for these commodities (p^{CD} and p^{P2P}). For convenience, we could write the general demand system in a log-log form:

$$ln q_i^{CD} = \alpha_0 + \alpha_1 ln p^{CD} + \alpha_2 ln p^{P2P} + \alpha_3 ln y_i + \delta_i^{CD} + \varepsilon_i^{CD} \qquad (2)$$

$$ln q_i^{P2P} = \beta_0 + \beta_1 ln p^{P2P} + \beta_2 ln p^{CD} + \beta_3 ln y_i + \delta_i^{P2P} + \varepsilon_i^{P2P} \qquad (3)$$

where α_1 and β_1 are the own-price demand elasticities for CDs and P2P downloads, respectively, α_2 and β_2 are the cross-price demand elasticities, and α_3 and β_3 are the income elasticities for the two goods. Individual-specific, time-invariant factors in both equations are captured by δ_i^{CD} and δ_i^{P2P} (e.g., tastes for the two sources of music). Disturbance terms (ε_i^{CD} and ε_i^{P2P}) capture all other aspects of the demand for these two commodities.

In a cross-sectional setup, such as the one used in Andersen and Frenz, it would be virtually impossible to estimate these own and cross-price elasticities. This is because we generally assume that all consumers face the same set of prices at a given time. Although there could be some variation in prices at a

point in time because consumers operate in distinct markets (e.g., due to geographic location), these are not the same as the individual price variation reported by respondents in this survey (often due to the type of music purchased and the outlets where these CDs were purchased). In addition, there is no information in this dataset on the different indirect costs across respondents that could comprise the implicit price of P2P downloads. This is why there are no individual subscripts on the prices in Eqs. (2) and (3). This means that these own and cross-price elasticities cannot be directly estimated in this cross section. As a result, these price responses would be lumped into either the constant or disturbance terms in these equations.

This is an important consideration because the cross-price elasticity in the CD demand equation would be the typical way of estimating the degree of 'substitutability' between the two sources of music. In other words, as the effective price of P2P downloads decreases over time, to what extent does it decrease the quantity of CDs demanded? Are the two good substitutes (i.e., $\alpha_2 > 0$)? As noted, however, Andersen and Frenz estimated a demand function for CDs, where the quantity of P2P downloads appeared directly on the right-hand side of this equation (see Eq. (1) earlier). This is essentially a 'partial correlation analysis', where both quantities are endogenous, choice variables for the consumer. The authors measure the correlation between these quantities while holding constant the average price paid for CDs (self-reported), household income (sometimes imputed) and various other explanatory variables included the X matrix in Eq. (1) (e.g., quantities of other entertainment goods purchased (DVDs, videogames, cinema tickets, concert tickets), level of interest in music, dummy variables for age categories, region, gender, student status).

The biggest problem associated with this approach is that unobserved factors in the disturbance term of this equation may be correlated with quantity of P2P downloads. This could lead to a biased estimate of the coefficient. For example, suppose that latent preferences for P2P downloads are also associated with a greater demand for CD purchases. The result would be an estimated coefficient that is biased upward. In other words, it could appear that there is no statistical relationship between CD and P2P downloads, when the true relationship is clearly negative.

6. Instrumental Variable Approach

We challenge the validity of the instruments used by Andersen and Frenz on a number of grounds. The authors claimed to have an overidentified model, because they had a series of five dummy variables on self-reported internet skill levels (ranging from very skilled to not at all skilled). They viewed these four dummy variables as multiple instruments, which motivated the use of the Sargan and Basmann tests for exogeneity. Since they were unable to reject the null hypothesis of exogeneity in either case, they claimed that these were valid instruments. A fundamental question is whether or not this model is truly overidentified. The spirit of these exogeneity tests is that researchers have competing instruments, because each one could be used independently to produce results through two-stage least-squares (TSLS). There is no way of knowing a priori how to choose among these alternative

estimators. This isn't the case in this situation. There is just a single source of information for these instruments: a self-reported assessment of personal internet skills. However, because this information is categorical, the authors correctly use multiple dummy variables to capture these different skill levels. Although in some sense these might be considered multiple instruments, they stem from the same single data source and all of the dummies should appear in the equation at the same time to capture the full range of internet skills. We would argue that this model is exactly identified, and that it is therefore impossible to test for the exogeneity in this system. If this were not the case, then no model would ever be exactly identified. We could always allow for some nonlinearity in the instrument and treat the model as overidentified. In fact, in this case, it appears from the results reported by AF that the relationship between P2P downloads and internet skill levels is reasonably linear (i.e., P2P downloads increase at approximately the same rate as we move from the lowest to the highest internet skill categories).

Furthermore, to be classified as 'valid instruments' these internet skill variables need to be correlated with the number of P2P downloads (i.e., they're relevant), but not the number of CDs purchased (i.e., they're also exogenous). In other words, internet skills cannot be a determinant of CD demand behaviour. This assumption is difficult to justify in the context of the general demand system for the two sources of music shown by Eqs. (2) and (3). Although no information is available in this dataset to infer what the implicit price of P2P downloads might be for an individual, this cross-price effect is a fundamental part of this overall demand specification. Since internet skill levels are potentially a key component of this implicit price, there is no justification from a general demand perspective for the validity of self-reported internet skills as valid instruments. Internet skill levels, operating through the implicit price of P2P downloads, could potentially have direct effects on the quantities of P2P downloads and CD purchases. In fact, the latter mechanism (the cross-price effect) is precisely the main research question addressed in this paper. Thus, allowing for the likelihood of a substitution effect essentially precludes the possibility of using internet skill levels as valid instrumental variables.

Thus, we challenge the validity of this IV approach on theoretical grounds. Internet skills would constitute at least part of shadow price P2P downloads, and this cross-price effect should be included in well-specified CD demand function. As a result, indicators of internet skill levels are inappropriate instruments. Furthermore, given that this is fundamentally an exactly-identified model, there is no legitimate way of statistically testing this exogeneity assumption.

7 Our approach

We seek to address the issues discussed in the previous section with the available data for this study starting with the equation used by Andersen and Frenz. The core problem we want to address is that demand for CDs and P2P may move together to the extent that people who have greater preferences for music might both demand more CDs and

engage in more P2P downloading. Thus, latent preferences for CDs and P2P downloads (i.e., δ_i^{CD} and δ_i^{P2P} in Eqs. (2) and (3)) may be positively correlated across the sample. Suppose that this problem is related to person-specific, time-invariant factors. Furthermore, with data on CD purchases and P2P downloads from two consecutive years (2004 and 2005), we can modify the 'partial correlation' specification of Andersen and Frenz with the following equations:

$$\boldsymbol{lnq_{i2004}^{CD} = \alpha_0' + \alpha_1' lnp_{2004}^{CD} + \alpha_2' lnq_{i2004}^{P2P} + \alpha_3' lny_{i2004} + X\gamma + \delta_i^{CD} + \varepsilon_{i2004}^{CD}} \quad (4)$$

$$\boldsymbol{lnq_{i2005}^{CD} = \alpha_0' + \alpha_1' lnp_{2005}^{CD} + \alpha_2' lnq_{i2005}^{P2P} + \alpha_3' lny_{i2005} + X\gamma + \delta_i^{CD} + \varepsilon_{i2005}^{CD}} \quad (5)$$

Taking the differences between Eqs. (5) and (4), assuming that there is no variation in the change in prices across individuals between years, recognising that there is no information on household income in the earlier period, and assuming that all of the variables in the vector X are time invariant, we would get Eq. (6) below[23]:

$$\boldsymbol{\Delta lnq_i^{CD} = \pi_0' + \alpha_2' \Delta lnq_i^{P2P} + \alpha_3' lny_{i2005} + \Delta\varepsilon_i^{CD}} \quad (6)$$

Differencing eliminates both the observed and unobserved determinants of demand that are time invariant. This differencing removes at least one possible source of bias in estimating the partial correlation coefficient between CD purchases and P2P downloads.[24]

The above equation in first differences can be estimated directly using the available dataset. Alternatively an equivalent form can be estimated that takes the difference between Eqs. (5) and (4), but expands out the difference terms and re-arranges the equation to place only CD demand in 2005 on the left hand side as follows:

$$\boldsymbol{lnq_{i2005}^{CD} = \alpha_d' + \alpha_1' lnq_{i2004}^{CD} + \alpha_2' \Delta lnq_i^{P2P} + \alpha_3' lny_{i2005} + \Delta\varepsilon_i^{CD}} \quad (7)$$

The key restriction that make Eq. (7) the same as Eq. (6) is that $\alpha_1' = 1$. The latter, more general specification also can be used more naturally in a Tobit estimation technique that allows for the censored data on individual CD purchases in 2005 that by definition cannot fall below zero.

8 Our results

We use in our analysis the same dependent variables and basic covariates as Andersen and Frenz. However, we make four key amendments to their analysis:

1. We include in our analysis the group of respondents excluded by Andersen and Frenz that did not purchase CDs in 2005. We further modify the relevant sample by

---

[23] If the γ coefficients vary between the years, then all of the covariates could be included in the first-difference estimation depicted in Eq. (6).

[24] The differences in the logs could alternatively be written as logs of the ratios (i.e., $\Delta lnq_i^{CD} = ln\left(q_{2005}^{CD} / q_{2004}^{CD}\right)$ and $\Delta lnq_i^{P2P} = ln\left(q_{2005}^{P2P} / q_{2004}^{P2P}\right)$. The coefficient can be interpreted as the relationship in the rate of increase between the two variables.

excluding four outliers who stated that they downloaded more than 400 music tracks per month in either year. In the end, to test the robustness of our results to the contentious issues regarding the legitimacy of observations on CD purchases and P2P downloads we use alternate sample inclusion criteria.

2. Due to the uncertainty regarding the appropriateness of using sample weights for narrower sub-samples, we test the importance of the weights used by Andersen and Frenz by reporting regression with and without these weights. This will indicate how robust our results are to the decision to use the full-sample weights in this analysis.

3. We exclude the self-reported average CD prices paid by respondents in our regression specifications. We also exclude some other covariates used by Andersen and Frenz (e.g., quantities of other entertainment goods purchased (DVDs, videogames, cinema and concert tickets) because there is little justification for their inclusion in a well-specified demand equation.

4. A first-difference estimation approach is used to eliminate any fixed effects, and this source of possible bias in estimating the relationship between changes in CD purchases and P2P downloads. In addition, we alter this approach slightly to estimate an autoregressive specification where CD purchases in 2005 serves as the dependent variable and CD purchases in 2004 serves as one of the independent

Table 2 Ordinary least-squares regression results on the full sample

	No Other Covariates	Complete Set of Covariates
Constant	0.012	-0.189**
	(0.016)	(0.090)
$\Delta ln q_i^{P2P}$	-0.050**	-0.054***
	(0.020)	(0.020)
R^2	0.004	0.030
Root MSE	0.651	0.649
F-Statistic	6.54	1.61
P-Value for F-Statistic	0.011	0.015
Sample Size	1760	1760

Notes: All data used in this regression analysis were taken from the 2006 survey questionnaire provided by Industry Canada. Robust standard errors are reported. This broad sample includes observations where the respondent reported buying CDs or downloading P2P music tracks in a given year, but subsequently said that these quantities were zero. Furthermore, this sample assumes that P2P downloads in 2004 were zero if the individual reported no downloads in 2005 (unless the respondent downloaded P2P tracks in 2005, similar information was not solicited from the individual in 2004). The regression results reported in the last column in this table come from a specification that includes explanatory variables on income, age, gender, student status, region, education, internet skill level and interest in music, including dummy variables for missing data on these covariates. The results on these other covariates are not included in this table, but are available from the authors.

Data from the Decima Research for Industry Canada Survey 2006

Dependent variable is the difference in natural logarithms of CD purchases in 2005 and 2004

*** Significantly different from zero at 1 % level using two-tailed t test

** Significantly different from zero at 5 % level using two-tailed t test

* Significantly different from zero at 10 % level using two-tailed t test

variables. This is a slightly more general form of the first-difference specification. A Tobit estimation procedure is used to allow CD purchases in 2005 to be censored at zero. This is the technique that Andersen and Frenz should have used in dealing with this censoring issue in their analysis, rather than simply excluding these non-participants from the sample.

9 First-difference results

The regression results using our first-difference specification for the full sample are summarised in Table 2. There are 1760 individual observations in this initial sample. Observations are included where individuals reported buying music CDs or downloading P2P music tracks, but subsequently said that the numbers of purchases or downloads was equal to zero in 2005. These are considered to be valid 'zero' observations for this particular sample. We also set P2P downloads equal to zero in 2004 if the individual reported that they did not engage in P2P downloading behaviour in 2005.

Other than the quantities of CD purchases and P2P downloads, none of the other variables (including personal income) were available for both 2004 and 2005. This means that changes in all of the other covariates between the years could not be included in the regression model. The dependent variable is the individual difference in the natural logarithms of the quantities of CDs purchased in 2005 and 2004.[25] Other than the constant term, parameter estimates are reported for only the difference in the natural logarithms of the quantities of P2P downloads in an average month in 2005 and 2004. This is the essence of our first-difference specification.

Two sets of regression results are reported in this Table 2. The first column reports results from a regression that excludes any other covariates, and includes only the difference in log P2P downloads between 2005 and 2004 as the sole regressor. The estimated coefficient on the change in log P2P downloads is −0.050. This estimated correlation coefficient is marginally significantly different from zero at 1.1 % level (i.e., the *p*-value). This estimated correlation has the following interpretation. Holding any individual-specific, time-invariant factors constant (as a result of the differencing between 2005 and 2004), we estimate that a 10 % growth in P2P downloading is associated with, on average, a 0.50 % decline in the growth rate in CD purchases.

The second column in Table 2 reports results from a regression that includes a full set of covariates. These other explanatory variables include the income, age, gender, student status, regional location, educational qualifications, internet skill levels and interest in music. These are similar to the covariates used by Andersen and Frenz. For brevity, we exclude the results on these control variables in this table. These other covariates can be justified in this regression

---

[25] Like Andersen and Frenz we add a value of one to the quantities of CD purchases and P2P downloads before taking natural logarithms. This means that the log value of someone with CD purchases or P2P downloads is zero.

specification if these factors are largely time invariant, but coefficients on these variables are allowed to be different in 2004 and 2005. In this case, we are allowing the estimated relationship between P2P downloads and CD purchases to vary systematically with these personal attributes.

We find that the correlation between changes in CD purchases and P2P downloads is slightly larger in magnitude with the inclusion of these other covariates. The estimated coefficient is −0.054, and significantly different from zero at a 0.7 % level. This suggests that a 10 % rise in the growth in P2P downloading is associated with an average decline of 0.54 % in the growth rate in CD purchases. The point estimates for these coefficients with and without the inclusion of the full set of covariates are not statistically different from one another at conventional test levels.

To test the robustness of these regression results for changes in both the definition in the sample and the specification of the regression, we summarise the results from a series of auxiliary regressions. These robustness checks are partly intended to make our analysis more comparable to that of Andersen and Frenz. These results can be directly compared to the estimated partial correlations between the growth rates in P2P downloads and CD purchases of −0.050 and −0.054 reported in Table 2:

1. We previously mentioned the issue of population weights that could be used in this analysis. The results reported in Table 2 are unweighted (presumably similar to the approach used in Andersen and Frenz in their 2010 article). When we used Weighted Least Squares (similar to their 2007 study), the estimated coefficients were fairly similar to those reported above (−0.048 with no other covariates, and −0.054 with the complete set of covariates). Again, these estimated correlations were statistically significant at conventional test levels.
2. We removed four observations on extremely high levels of P2P downloading activity. These were individuals who had downloaded an average of more than 400 music tracks per month in either year. Including these potential outliers in the regression estimation reduces these partial correlations to between −0.029 and −0.037. These estimated coefficients are no longer significantly different from zero at conventional test levels.[26]
3. Excluding the zero observations on CD purchases in 2005 reduced the sample size by 17.9 %, and resulted in estimated correlations between −0.032 and −0.046. These estimated coefficients are all statistically significant at better than a 10 % level.
4. Adding the difference in the natural logarithms in perceived prices for music CDs to the regressions resulted in estimated correlations that ranged between −0.014 and −0.028. These estimated correlations are no longer statistically significant at conventional test levels.

---

[26] Our preferred specification is clearly to exclude these four observations. All four of these individuals reported downloading an average of exactly 500 music tracks per month in one or both of the years. These are approximately 20-times the number of monthly downloads of those who engaged in this behaviour in our sample. If these four observations are not legitimate outliers, then the regression results should be minimally affected by the exclusion of just 0.2 % of the original sample. The fact that the relationship between P2P downloads and CD purchases is weaker when these observations are included suggests that the behaviour of these high-downloading individuals is fundamentally different from the rest of the population.

5. Excluding observations on individuals who were never asked directly about their P2P downloading activities in 2004 reduced the sample size by 55.7 %. This excluded group consisted of individuals who were not downloading music tracks in 2005. Excluding this group had virtually no impact on the regression results reported in Table 2. These partial correlations range from −0.048 to −0.051, and all of them are significantly different from zero at either 5 % or 10 % levels. None of the point estimates on these correlations are significantly different from each other at conventional test levels.

These auxiliary regressions based on different sample definitions and regression specifications produce some variation in the estimated correlations between the growth rates in P2P downloads and CD purchases. However, all of these point estimates have a negative sign. Statistical significance disappears when we re-introduce the four observations on potential outliers in P2P downloads (in excess of 400 music tracks per month), and include the self-reported change in the average price paid for CDs between the years (which itself had no significant effects on the growth in CD purchases). Thus, when our estimation approaches more closely mirror those of Andersen and Frenz, we

Table 3 Maximum likelihood Tobit regression results on the full sample

	No Other Covariates	Complete Set of Covariates
Constant	0.018	−0.132
	(0.051)	(0.107)
lnq_{i2004}^{CD}	0.947***	0.903***
	(0.021)	(0.023)
Δlnq_i^{P2P}	−0.047**	−0.061***
	(0.023)	(0.024)
Pseudo R^2	0.324	0.341
F-Statistic	1022.0	77.4
P-Value for F-Statistic	0.000	0.000
Sample Size	1760	1760

Notes: All data used in this regression analysis were taken from the 2006 survey questionnaire provided by Industry Canada. Robust standard errors are reported. This broad sample includes observations where the respondent reported buying CDs or downloading P2P music tracks in a given year, but subsequently said that these quantities were zero. Furthermore, this broad sample assumes that P2P downloads in 2004 were zero if the individual reported no downloads in 2005 (unless the respondent downloaded P2P tracks in 2005, similar information was not solicited from the individual in 2004). The regression results reported in the last column in this table come from a specification that includes explanatory variables on income, age, gender, student status, region, education, internet skill level and interest in music, including dummy variables for missing data on these covariates. The results on these other covariates are not included in this table, but are available from the authors.

Data from the Decima Research for Industry Canada Survey 2006

Dependent Variable is the Natural Logarithms of CD Purchases in 2005

*** Significantly different from zero at 1 % level using two-tailed t test

** Significantly different from zero at 5 % level using two-tailed t test

* Significantly different from zero at 10 % level using two-tailed t test

produce estimated correlations between P2P downloads and CD purchases that are often statistically insignificant.

10 Tobit results

Furthermore, to test whether the censored data on CD purchases matters for our results, we used maximum likelihood Tobit estimation using an autoregressive specification which includes the lagged dependent variable on the right-hand side of our regression equation. The regression results for the full sample are displayed in Table 3. Note that this specification is more general than the first-difference approach used earlier. It is identical to the previous approach if the coefficient on the lagged dependent variable is exactly equal to one. The key is that Tobit allows us to explicitly account for the censoring of 2005 CD purchases at zero.

The dependent variable for the regression results summarised in Table 3 is the natural logarithm of the quantity of CDs purchased in 2005 for an individual. Again, regression results are reported with and without the inclusion of other covariates. The results in columns 1 come from a regression that excludes other covariates, and includes only the lagged value of CD purchases in 2004 along with the difference in the natural logarithms of P2P downloads between 2005 and 2004 as regressors. The results in column 2 include the full set of covariates in this estimation. Again, for brevity, we exclude from this table the results on these control variables.

Without any other covariates in the regression specification, the estimated coefficient on the difference in log P2P downloads is −0.047. This estimated correlation is significantly different from zero at 3.9 % level. This result suggests that if CD purchases were <u>not</u> censored, a 10 % growth in P2P downloads would be associated with a decline in the growth in CD purchases by 0.47 %.

With the inclusion of the other covariates in this Tobit estimation, the estimated coefficient on the difference in log P2P downloads increases in both magnitude and statistical significance. This estimated correlation coefficient is −0.061, and statistically significant at a 0.9 % level. This result suggests that if CD purchases were <u>not</u> censored, a 10 % growth in P2P downloads would be associated with a decline in the growth of CD purchases by 0.61 %.[27] The point estimates for these coefficients with and without the inclusion of the full set of covariates are not statistically different from one another at conventional test levels.

Again, to test the robustness of these regression results to alternative sample definitions and regression specifications, we summarise the results from a series of auxiliary regressions. These results can be directly compared to estimated partial

---

[27] Supplementary analysis shows that this point estimate and significance level are nearly identical if we compute the marginal effects for *censored* CD purchases. The estimated marginal effect is again −0.061 for the growth in P2P downloads, if we allow CD purchases to be censored at zero.

correlations between the growth rates in P2P downloads and CD purchases of −0.047 and −0.061 reported in Table 3:

1. When we used the population weights, the point estimates on the coefficients on the difference in the log of P2P downloads were fairly similar to those reported above (−0.041 with no other covariates, and −0.058 with the complete set of covariates). However, these estimated correlations were statistically insignificant at conventional test levels in this weighted regression analysis.
2. Including the four potential outliers on high levels of P2P downloads in our regression estimation reduced these partial correlations to between −0.031 and −0.045. These estimated coefficients are no longer significantly different from zero at conventional test levels.
3. Excluding the zero observations on CD purchases in 2005 reduced the estimated correlations to between −0.031 and −0.043. These estimated coefficients are all statistically significant at better than a 10 % level.
4. Adding the difference in the natural logarithms in perceived prices for music CDs to the regressions resulted in estimated correlations that ranged between −0.013 and −0.026. These estimated correlations are no longer statistically significant at conventional test levels.
5. Excluding observations on individuals who were never asked directly about their P2P downloading activities in 2004 had little impact on the regression results reported in Table 3. These partial correlations range from −0.048 to −0.051, and all of them are significantly different from zero at either 5 or 10 % levels. None of the point estimates on these correlations are significantly different from each other at conventional test levels.

These auxiliary regressions produced some variation in the estimated correlations between the growth rates in P2P downloads and CD purchases. However, all of these point estimates have a negative sign. As before, statistical significance disappears when we use weighted regression analysis, re-introduce the four observations on potential outliers in P2P downloads (in excess of 400 music tracks per month), and include the self-reported change in the average price paid for CDs between the years (which itself had no significant effects on the growth in CD purchases in this Tobit estimation). Again, when our estimation approaches more closely mirror those of Andersen and Frenz, we often find estimated correlations that are statistically insignificant.

In comparing the regression results using OLS and Tobit estimation, we prefer the latter specification. This is a more flexible functional form which does not assume the coefficient on the lagged dependent variable is equal to one (i.e., that $\alpha_1' = 1$ in Eq. 7). In the regression results reported in Table 3, we can reject this null hypothesis that $\alpha_1' = 1$ in both regressions at better than a 5 % significance level. The Tobit specification allows the annual purchase of CDs to be censored at zero, and more than one-sixth of the sample report no purchases of CDs in 2005. However, what maybe just as important, is that both regression procedures produce results that are more or less consistent with one another. Both estimation techniques produce estimated coefficients that are consistently negative, relatively small in magnitude and generally significantly different from zero.

11 Conclusion

All of our regression results reported in Tables 2 and 3 with our preferred unrestricted sample show a consistent negative association between P2P downloading and CD demand. These partial correlation coefficients are statistically different from zero at conventional levels. The point estimates of the correlations across all specifications range between −0.047 and −0.061. These coefficient estimates are fairly robust to changes in estimation techniques (Ordinary Least-Squares, weighted regression analysis or maximum likelihood Tobit), the inclusion or exclusion of other covariates, and even alternative sample inclusion criteria. Overall, these results imply that a 10 % increase in P2P downloads is associated with a reduction in CD demand of between approximately 0.47 and 0.61 %.

This finding of negative and generally statistically significant correlations between P2P downloads and CD purchases directly contradicts the much cited and controversial conclusion of Andersen and Frenz in the only other paper to study these Industry Canada data. They claimed that the data showed "… no association between the number of P2P files downloaded and CD album sales,"[28] claiming therefore that "… this paper show (sic) that P2P file-sharing is not to blame for the decline in CD markets. Music markets are not simply undermined by free music downloading and P2P file-sharing."[29]

In this paper we have corrected for two fundamental errors in previous analysis by Andersen and Frenz leading to their contentious conclusion. First, we corrected for the fact that the authors may have biased their results by excluding from their analysis the group of consumers who had completely stopped purchasing CDs (potentially because of P2P activity) by 2005. This is the very group who were most responsive, or likely to have substituted P2P downloading for CD purchases. Second, we controlled for the fact that the *level* of an individual's demand for CDs, and the *level* of an individual's P2P downloading may be correlated simply because they are both affected by the same third factor, such as preferences for music, so that high (or low) levels of CD demand is likely to be associated with high (low) levels of P2P demand. Such a positive association between the level of demand and level of P2P downloading may have led Andersen and Frenz to mistakenly conclude they had found evidence of a positive market creation effect, because the authors regressed the *level* of individuals CD demand against the *level* of individuals P2P downloading. Instead, we focused on the *changes* in CD demand and *changes* in P2P downloading, using the same survey data available to Andersen and Frenz. By focusing on a longitudinal analysis of how the *change* in individual P2P downloading behaviour might be related to the *change* in CD demand, we were better able to isolate the true association between CD demand and P2P downloading behaviour.

Finally, it's notable that our estimated relationships between CD sales and P2P downloads are broadly consistent with the market data up to the time of the Industry Canada survey. As noted, file sharing first came to attention when Napster

---

[28] Andersen and Frenz 2010 ibid p.374
[29] Ibid p.375

began operations in mid-1999, and free music downloads then grew substantially between 1999 and 2006. During this period from 1999 to 2006, CD unit sales fell by around 20 % in Canada.[30] Extrapolating from our linear estimates, CD sales would be expected to decline by between 4 and 6 % for every doubling (i.e., 100 % increase) in P2P downloads. Thus, we would expect that a 20 % decline in CD sales between 1999 and 2006 would be associated with between a 333 and 500 % increase in P2P downloads. The latter is roughly consistent with the substantial increase in P2P downloading behaviour between 1999 and 2006. However, there are two words of caution here. First, we have estimated a linear relationship using data from two consecutive years. The relationship might be quite different (i.e., nonlinear) for much larger changes in music acquisition behaviour over longer time horizons. Second, we don't know what the 'counterfactual' is in this case. In other words, CD sales may have otherwise increased substantially between 1999 and 2005 without the possibility of P2P downloads.

References

Andersen B, Frenz M (2007) The impact of and P2P file-sharing on the purchase of music: A study for Industry Canada. Retrieved from http://www.ic.gc.ca/eic/site/ippd-dppi.nsf/eng/h_ip01456.html

Andersen B, Frenz M (2010) Don't blame the P2P file-sharers: the impact of free music downloads on the purchase of music CDs in Canada. J Evol Econ 20(5):715–740

Blackburn D (2004) On-line piracy and recorded music unpublished working paper. Economics Department, Harvard University

Blaug M (2005) Why did Schumpeter neglect intellectual property rights?. 2(1) Review of Economic Research on Copyright Issues 69 at p.69

Demsetz H (2009) Creativity and the economics of the copyright controversy. Rev Econ Res Copyr Issues 6(2) 5 at p.11

Hong SH (2007) The recent growth of the internet and changes in household-level demand for entertainment. Inf Econ Policy

Hong SH (2013) Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes. J Appl Econ 28(2):297–324

Idris K (2005) WIPO, intellectual property: a power tool for economic growth. p.27

Landes WM, Posner RA (2003) The economic structure of intellectual property law. Harvard University Press, Cambridge

Lemley MA (2005) Property, intellectual property, and free riding. Texas Law Rev 83(1031):2004–2005

Lessig L (2004) Free culture: how big media uses technology and the law to lock down culture and control creativity. Penguin, New York

Liebowitz SJ (2006) File-sharing: creative destruction or just plain destruction? J Law Econ 49:1–28

Liebowitz SJ (2007a). A comment on Oberholzer-gee and Strumpf paper on file-sharing. Retrieved from http://ssrn.com/abstract=1017418

Liebowitz SJ (2007b). How reliable is the Oberholzer-Gee and Strumpf paper on file-sharing? Retrieved from http://ssrn.com/abstract=1014399

Liebowitz SJ (2008) Testing file-sharing's impact on music album sales in cities. Manag Sci 54:852–859

---

[30] CD unit sales fell by 19 % from 2000 to 2006 according to Music Canada statistics see http://www.musiccanada.com/statistic.aspx

Liebowitz SJ (2010) The Oberholzer-Gee/Strumpf file-sharing instrument fails the laugh test. Retrieved from http://ssrn.com/abstract=1598037 and http://dx.doi.org/10.2139/ssrn.1598037

Liebowitz S (2014) How much of the decline in sound recording sales is due to file-sharing? J Cult Econ, forthcoming, DOI 10.1007/s10824-014-9233-2 (see SSRN-id1598037.pdf)

Medema SG (1995) Introduction. In SG Medema (ed.), The legacy of Ronald coase in economic analysis

Michel NJ (2006) The Impact of digital file sharing on the music industry: an empirical analysis. Topics Econ Anal Policy 6(1) Article 18. http://www.bepress.com/bejeap/topics/vol6/iss1/art18

Oberholzer-Gee F, Strumpf K (2007) The effect of file sharing on record sales: an empirical analysis. J Polit Econ 115(1):1–42

Peitz M, Waelbroeck P (2004) The effect of internet piracy on music sales: cross-section evidence. Rev Econ Res Copyr Issues 1(2):71–79

Rob R, Waldfogel J (2006) Piracy on the high c's: music downloading, sales displacement, and social welfare in a sample of college students. J Law Econ 49(1):29–62

Spar D (2001) Ruling the waves: from the compass to the internet, a history of business and politics along the technological Frontier, NY: Harcourt

Waldfogel J (2010) Music file sharing and sales displacement in the iTunes era. Inf Econ Policy 22(4):306–314

Zentner A (2005) File sharing and international sales of copyrighted music: an empirical analysis with a panel of countries. Topics Econ Anal Policy 5, art. 21, pp. 1–15. http://www.bepress.com/bejeap/topics/vol5/iss1/art21

Zentner A (2006) Measuring the effect of music downloads on music purchases. J Law Econ 49(1):63–90

Zentner A (2009) Ten years of file sharing and its effect on international sales of copyrighted music: an empirical analysis using a panel of countries. Available at SSRN: http://ssrn.com/abstract=1724444