Measuring the Effect of Digital Technology on the Sales of Copyrighted Goods: Evidence from Napster[*]

Seung-Hyun Hong
Department of Economics
University of Illinois
hyunhong@ad.uiuc.edu

January 2007

Abstract

This paper measures the effect of file sharing on equilibrium record sales. Previous research has suffered from little direct information on who downloaded music and how much their music expenditures changed. This paper uses the introduction of Napster and exploits two household-level data sets. I begin with a difference-in-differences approach, treating the presence of Napster as a technological event that only Internet users experienced. To account for compositional changes in Internet users as well as the likely relationship between music expenditure and the propensity to adopt the Internet, I improve upon nonparametric matching methods. Because of confounding factors due to other new online activities, I additionally combine two datasets. Taken together, the results suggest that file sharing explains 20% of total sales decline, which is driven by households with children aged 6-17.

Keywords: file sharing, record sales, difference-in-differences, compositional changes, nonparametric matching, data combination

JEL classification: C31, L82, L86, O34

[*] I am grateful to Frank Wolak, Liran Einav, and Peter Reiss for invaluable insights and encouragement. I also thank Susan Athey, Tim Bresnahan, Ben Malin, Mark Jacobsen, Roger Noll, Azeem Shaikh, Alan Sorensen, Ed Vytlacil, and participants of IO Workshop and Applications Workshop at Stanford University as well as the structural lunch at Stanford Business School for helpful discussion and comments. I appreciate the UCLA Internet Project and Jeffrey Cole for generously providing data. Financial support was provided by Stanford Institute for Economic Policy Research. All remaining errors are my responsibility. Contact information: Department of Economics, University of Illinois at Urbana-Champaign, 470F Wohlers Hall, 1206 S. Sixth St., Champaign, IL 61820

> "Over the past three years shipments of recorded music in the U.S. have fallen by an astounding 31%. And worldwide, the music industry has shrunk from a $40 billion industry in 1999 down to a $26 billion industry in 2002. . . . The root cause for this drastic decline in record sales is the astronomical rate of music piracy on the Internet." (Cary Sherman (2003), President, Recording Industry Association of America, before the Senate Judiciary Committee).

1 Introduction

The widespread use of digital technology has allowed consumers to reproduce digital versions of copyrighted songs, movies, books, and computer software inexpensively. As a result, recent years have witnessed a substantial increase in digital copying activities, and producers of copyrighted music and movies have claimed that digital copying is largely responsible for their considerable loss in revenues. The coincidence between the introduction of Napster and the start of the ongoing slump in record sales appears to attest to this alleged economic harm (see Figure 1). Nevertheless, theoretical studies on copying suggest ambiguous predictions for the effects of digital copying on the sales of copyrighted goods.[1] In fact, a recent empirical study finds an insignificant relationship between file sharing and music album sales (Oberholzer and Strumpf 2007).

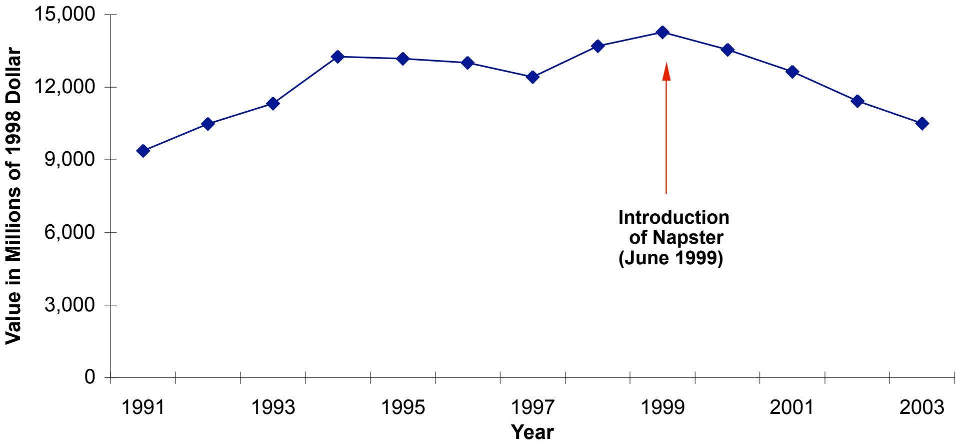
This paper investigates the impact of digital copying technology on the sales of copyrighted goods by examining the effect of Napster on record sales. Specifically, I analyze changes in U.S. household-level music expenditure between the pre- and post-Napster periods[2] with two goals. The first objective is to quantify the effect of file sharing[3] on equilibrium music expenditures during the Napster period. File sharing is likely to have reshaped music demand and supply, and therefore the equilibrium quantity of recorded music sold. I attempt to measure this change in equilibrium recorded music expenditures. The second objective of this paper is to identify the demographic groups responsible for the sales decline attributable to file sharing. Because of substantial heterogeneity in music preferences and in the costs associated with file sharing,

[1] See, e.g., Besen and Kirby (1989), Takeyama (1994), and Bakos, Brynjolfsson, and Lichtman (1999).

[2] Throughout this paper, *pre-Napster* refers to June 1997 through May 1999, while *post-Napster* or *Napster period* refer to June 1999 through June 2001, the period in which Napster was operating.

[3] Napster was the first widely used software designed to facilitate the exchange of files between Internet users, and remained the dominant file sharing service until early 2001.

Figure 1: Total Real Value of Record Shipments in the U.S.[a]



[a] Refer to the RIAA's Yearend Statistics. Total sales include CDs, cassettes, LPs, and music videos.

consumers may have responded differently to Napster. Because this heterogeneity is likely highly correlated with age, I focus on demographic groups defined by age in particular.

To date, at least four econometric studies have attempted to measure the effect of file sharing on record sales. Using product-level data, Oberholzer and Strumpf (2007) and Blackburn (2004) seek to measure how many units a given hit album would sell if it were downloaded less. The product-level analyses, however, are limited in that they cannot distinguish between the sales decline among file sharing users and the decline among non-users. For example, it is possible that numerous songs were downloaded mostly by those who never purchased music, while most music albums were purchased by consumers who never downloaded music.[4] As noted by Rob and Waldfogel (2006), ignoring this problem can yield a spurious relationship between file sharing and music sales. Consequently, one needs micro-level data. Two other studies make use of micro-level data. These studies, nevertheless, use only cross-sectional data, which restricts their analyses on

[4] In this case, we could observe the increase in downloading of some songs as well as the change in sales of the same songs. Using only album-level information, we would therefore find correlations between downloading and record sales. In this example, however, downloading did not cause the change in sales of these songs because most music albums were purchased by those who never downloaded music. Without observing who downloaded music and how much their music expenditures changed, it is therefore difficult to infer a causal relationship between downloading and record sales.

changes in record sales. Besides, Zentner (2006) observes only binary variables on downloading and music purchase, which prevents him from measuring the magnitudes of the effect on record sales. Rob and Waldfogel (2006) collect their own survey data on albums either purchased or downloaded by college students, but their sample is not representative and is limited to a small number of college students.

The key to estimating the effect of file sharing on record sales is constructing the counterfactual estimates of record sales in the absence of file sharing. As described above, it is difficult to construct such counterfactuals from a comparison of sales of different albums. Likewise, a cross-sectional comparison of different individuals is unlikely to allow one to estimate the counterfactuals, unless it is plausible to assume that on average music expenditures of file sharing users would be the same as those of non-users in the absence of file sharing.[5] Ideally, one would use panel data of individual music expenditures in which file sharing were randomly assigned to similar individuals. Using such data, one could construct the counterfactuals by applying a conventional difference-in-differences (DD) approach. Unfortunately, such ideal data have never been available.

Despite a lack of ideal panel data, however, this paper shows that one can construct the counterfactuals by exploiting information in publicly available micro data, using an exogenous event, and improving upon existing econometric methods. Specifically, I use detailed household-level, repeated cross-sectional data from the *Consumer Expenditure Survey* (CEX). The CEX contains information on recorded music expenditures and Internet access fees. In contrast to data used in prior studies, the CEX consists of random samples of households designed to be representative of the U.S. population. Using the weights provided by the CEX, I can further measure the effect of file sharing on total record sales.

The empirical analysis treats the emergence of Napster as a technological event that only Internet users experienced. Changes in music expenditures after the introduction of Napster are attributed to a time effect and the effect of the presence of Napster. Internet non-users are subject

[5] This assumption is unlikely to hold because file sharing is not a random event, and individual-specific heterogeneity is likely to determine both file sharing and music expenditures.

only to the time effect. Applying the conventional DD to this context, however, entails two potential problems. The first is the likely relationship between music expenditure and the propensity to adopt the Internet. Internet users tend to spend more money on recorded music than non-users. More importantly, post-Napster Internet users might spend less on recorded music than pre-Napster Internet users simply because more consumers with low willingness to pay for music might have adopted the Internet over time. This compositional change in Internet users can lead to a serious bias in the conventional DD approach. The second concern is that the effect of the presence of Napster can be confounded with other online activities that originated around the same time. For example, Internet users during the Napster period could have reduced music expenditures because they started to use online secondary markets for used CDs. I account for each problem using two distinct methodologies that enhance the DD approach.

To address the first concern, I exploit the detailed demographic information in the CEX and use a nonparametric DD matching (DDM) method developed by Heckman, Ichimura, and Todd (1997, 1998). This method assumes random selection conditional on various observables that determine Internet adoption. I extend the standard DDM method by matching each post-Napster Internet user with pre-Napster users, as well as with pre- and post-Napster Internet non-users, based on two propensity scores: the probability of having Internet access given observables for the pre-Napster period; and this same probability for the post-Napster period. To the extent that consumers with similar propensity scores are equally likely to adopt the Internet, the DDM approach accounts for compositional changes in Internet users and isolates the effect of the presence of Napster.

I find that the average U.S. household with an Internet connection during the Napster period would have spent \$1.45 (7.6%) more per quarter on recorded music in the absence of Napster. Aggregating this number to the population of U.S. households explains approximately 40% of the total record sales decline during that period. About half of this decline is driven by households with children aged 6-17, which is estimated precisely. The other half of this decline is explained by households aged 15-34, but the DDM estimate for this group is not precisely estimated.

These results, nevertheless, could be confounded with other new online activities during the Napster period. Hence, I consider an additional approach that complements the DDM by decomposing the effect of the presence of Napster into a music downloading effect and the effects due to other new online activities. Because the CEX does not contain any information on music downloading, I use a complementary data set[6] with detailed demographic variables and information on downloading activitiy. The second data set, however, does not include music expenditures, and thus, it must be combined with the CEX. I first maintain the DDM approach by estimating non-parametric bounds, but find that they are not very informative. Therefore, I consider a two-sample instrumental variable (2SIV) approach which exploits linearity in the DD regressions to allow for data combination and provide more informative results. Because the DD requires more restrictive assumptions than the DDM, I examine the validity of these assumptions for different demographic groups and do not find any statistically significant evidence against these assumptions for households aged 15-34 and those with children aged 6-17. Accordingly, the 2SIV results for these two groups can be interpreted as decomposing the effect of the presence of Napster.

I find that actual downloading was a major factor in the decline in recorded music expenditure for households with children aged 6-17, while the effect due to other new online activities is statistically indistinguishable from zero for this demographic group. This suggests that the DDM estimate for this group almost exclusively represents the effect of file sharing. In contrast, the effect of actual downloading for all other households is insignificant, implying that we can rule out a potential negative effect of file sharing particularly for those aged 15-34. Recall that households with children aged 6-17 account for about half of the effect of the presence of Napster on total record sales. Therefore, the DDM results and the data combination results together imply that file sharing can explain approximately 20% of the total sales decline during the Napster period, and this decline is driven almost entirely by downloading activities of those with children aged 6-17.

These findings contribute in part to the large literature on copyright protection in a digital era.[7]

[6] The data set is an annual household-level survey on Internet usage collected by the UCLA Center for Communication Policy. See Section 4.2 for more details.

[7] See, e.g., Posner (2005) and Varian (2005) for recent discussion on the issue. See also Landes and Posner (2003)

Digital technologies have dramatically reduced the cost of copying, thereby increasing consumers' access to copyrighted materials. Under this new environment, however, conventional copyright protection might fail to secure revenues of copyright holders, hence reducing financial incentives to create new works. In accessing the social efficiency of more restrictive copyright protection in a digital era, one necessary (but not sufficient) piece of information is the extent to which digital technologies have reduced revenues of copyright holders. My findings provide this information in the case of the recording industry.

The rest of the paper proceeds as follows. Section 2 provides background on Napster and record sales. Section 3 explains potential problems with the DD approach and develops empirical strategies to account for them. Section 4 describes the data and examines basic patterns in the data. Sections 5 discusses econometric framework, and Section 6 presents estimation results. Section 7 concludes.

2 Background on Napster and Record Sales

Systematic file sharing began with Napster. After its introduction in June 1999,[8] Napster quickly became popular among Internet users. The number of users grew extraordinarily, and numerous music files were exchanged via Napster.[9] Though other minor file sharing programs appeared during the Napster period, Napster was undoubtedly the dominant file sharing service until early 2001.[10] For this reason, as well as the lack of further information, I do not distinguish file sharing via Napster from file sharing via other programs during the Napster period.

Napster allowed its users to share a variety of individual songs, thereby providing access to unbundled songs, as opposed to bundled albums, in addition to providing the songs for free. This is

for more extensive discussion and references.

[8] According to Shawn Fanning (2000), he and his uncle incorporated Napster in May 1999 ("Napster" was his nickname). He released a beta version during the summer, and it spread quickly to Internet users. The exact date on which Napster was introduced, however, is unclear. I use June 1999 because Napster was certainly not available before then. Moreover, each observation in the CEX covers expenditures for three months, so that the earliest observations during the Napster period includes expenditures for the period from June 1999 through August 1999, during which Napster became available to Internet users.

[9] According to *Newsbytes*, July 20, 2000, citing Napster's report on its membership, the number of Napster users grew from 1 million in November 1999 to 10 million in late April 2000, to 15 million in mid June, and to 20 million in mid July 2000. Fanning (2000) reports that it was over 32 million in early October 2000. Romer (2002) cites Webnoize, a web consulting firm, estimating that Napster members exchanged 2.8 billion files in February 2001.

[10] See Chapter 3 in Fisher (2004) for more information on various file sharing programs.

likely to have changed consumers' expenditures on recorded music, but the direction and magnitude of this change is unclear and is likely to differ across consumers. Zero prices could lead consumers to substitute CD purchases with music downloading, hence reducing their music expenditures. On the other hand, inexpensive access to individual songs provide consumers with better information on a variety of music, which could increase consumers' music expenditure. Moreover, considerable heterogeneity in music preferences and in the costs associated with downloading implies that consumers responded differently to the presence of Napster.

Nevertheless, this event coincided with the start of the ongoing slump in recorded music sales. According to the Recording Industry Association of America (RIAA), total real value of shipments in the United States had reached its peak of \$14,270 million in 1999. After Napster appeared, the total real value of record sales decreased by 5% in 2000, 6.7% in 2001, 9.6% in 2002, and 8.1% in 2003 (see Figure 1). Accordingly, the recording industry concluded that this decline was largely a result of file sharing. Subsequent legal action by the recording industry based on these grounds succeeded in closing Napster in 2001.

This coincidence, however, does not substantiate the negative impact of Napster and file sharing on record sales, nor does it provide any magnitude of the impact. Despite the apparent negative correlation between file sharing and record sales, it is unclear to what extent the sales decline is attributable to file sharing. There are a variety of factors, other than file sharing, that can also account for the recent slump in record sales. First, some entertainment goods might be substitutes or complements for recorded music, and changes in relative prices for these goods could lead to the decline in recorded music demand. Second, the rapid penetration of the Internet might lead Internet users to spend more time on the Internet and less time on listening to music, which could decrease demand for recorded music.[11] To evaluate the effect of file sharing on record sales, one needs to construct counterfactual record sales in the absence of file sharing, accounting for potentially confounding factors. The next section develops my empirical strategies to do so.

[11] See Hong (2006) for further investigation on these possibilities.

3 Empirical Strategy

The goal of the empirical analysis is to quantify the extent to which file sharing was responsible for changes in recorded music expenditures. To do so, I consider post-Napster Internet users and compare their music expenditures with those of pre-Napster Internet users, as well as those of pre- and post-Napster Internet non-users. The underlying assumption is that the presence of Napster is an exogenous technological event to which only Internet users were exposed. Changes in music expenditures after the introduction of Napster are attributed to a time effect and the effect of the presence of Napster. Non-users were subject only to the time effect. To the extent that (i) Internet non-user groups are comparable to user groups and (ii) there are no compositional changes between these groups, this difference-in-differences (DD) approach can account for the time effect and general effects from using the Internet, hence identifying the effect of the presence of Napster on recorded music expenditure. Furthermore, if (iii) the presence of Napster was the only event unique to post-Napster Internet users, then estimating the effect of the presence of Napster can isolate the effect of file sharing on recorded music expenditure during the Napster period.

These three conditions are unlikely to hold, however. As for (i) and (ii), a major concern is the relationship between music expenditure and the propensity to adopt the Internet, which I ascertain through examining the CEX data in Section 4. Internet users tend to have different demographic and consumption patterns from non-users. More importantly, pre-Napster Internet users differ from post-Napster users. Early Internet adopters are likely to be richer, younger, and more highly-educated than later adopters or non-adopters. This difference is important because later adopters may include more consumers with a lower willingness to pay for recorded music, so that a negative effect computed from the DD approach may simply reflect this compositional change due to Internet diffusion.

To see this, consider the following example, where there are two types of consumers. One is young and spends $30 on CDs in each period. The other is old and does not buy recorded music at all. There are two periods, *pre-Napster* and *post-Napster*. Young consumers have high propensity to adopt the Internet. Accordingly, 8 out of 10 had an Internet connection in both periods. In contrast, old consumers have low propensity to adopt the Internet, and so only 2 out of 10 had Internet access in the pre-Napster period. However, the Internet became more accessible over time, so that 7 out of 10 old consumers had Internet access in the post-Napster period. This example is illustrated below.

An Example of Compositional Changes in Internet Users[a]

	Type	pre-Napster		post-Napster	
		Y	N	Y	N
Internet user	young	\$30	8	\$30	8
	old	\$0	2	\$0	7
Non-user	young	\$30	2	\$30	2
	old	\$0	8	\$0	3

[a] Y denotes music expenditure, and N denotes the number of consumers.

In this example, no one changed her music expenditure. That is, an Internet user in the post-Napster period would continue to spend the same amount on recorded music as before even in the absence of Napster. However, if one uses the conventional DD approach which assumes the composition of Internet users remained the same in both pre- and post-Napster period, then the estimated effect is -\$14.[12] The example illustrates that the naive DD approach uses incorrect weights to compute the counterfactual. To fix the problem, one should use weights for the two ages of users in terms of post-Napster Internet users. Noting that these weights are $\frac{8}{15}$ and $\frac{7}{15}$, respectively for each age, one can correctly infer that an average Internet user in the post-Napster period would have spent \$16 even in the absence of Napster, and therefore, the effect of the presence of Napster is zero.[13] Note that this is equivalent to separating two types, and computing the DD estimates for each age, and then calculating a weighted mean of these estimates. Conditional on types, Internet adoption is randomly assigned, so that the likely relationship between music expenditure and Internet adoption disappears.

[12] $(\$30 \times \frac{8}{15} + \$0 \times \frac{7}{15}) - \left[(\$30 \times \frac{8}{10} + \$0 \times \frac{2}{10}) + (\$30 \times \frac{2}{5} + \$0 \times \frac{3}{5}) - (\$30 \times \frac{2}{10} + \$0 \times \frac{8}{10})\right] = \$16$ - \$30.

[13] $(\$30 \times \frac{8}{15} + \$0 \times \frac{7}{15}) - \left[(\$30 \times \frac{8}{15} + \$0 \times \frac{7}{15}) + (\$30 \times \frac{8}{15} + \$0 \times \frac{7}{15}) - (\$30 \times \frac{8}{15} + \$0 \times \frac{7}{15})\right] = \$16 - \$16$.

Therefore, to account for the compositional changes in Internet users, one needs to identify different types of consumers, where types are associated with the propensity to adopt the Internet. For this identification, I assume that these types are determined by observed characteristics. Detailed demographic information in the CEX motivates this assumption. In particular, I assume that probabilities of having an Internet connection for both pre- and post-Napster periods can summarize all the necessary information, so that conditioning on these two probabilities are sufficient for identification.[14]

With regard to the condition (iii), that is, the presence of Napster was the only event unique to post-Napster Internet users, the concern is that other new online activities, aside from file sharing, may also have led Internet users to change their music expenditure. This is plausible because of several innovations on the Internet during the Napster period. For example, online secondary markets, particularly for used recordings, were developed during the Napster period. In January 2000, Half.com launched a person-to-person marketplace where consumers could buy and sell various items including used CDs. This company became popular and was acquired by eBay in July 2000. Many other companies began offering similar services as well.[15] Facing lowered prices from these online secondary markets, Internet users could reduce their music expenditure.

These additional technologies can confound the relationship between file sharing and music expenditure. To account for this, one needs to further compare Internet users who downloaded music files with those who did not. This requires observing actual downloading activity.[16] Because the CEX does not contain information on music downloading, I address this concern by using an

[14] In the example, each type can be characterized by probabilities of having Internet access in each period. Young consumers have 0.8 probability for both periods, while old consumers have 0.2 probability in the pre-Napster period and 0.7 probability for the post-Napster period. Instead of conditioning on type, one can condition on these probabilities. This is the basic idea behind the DD matching using two propensity scores described in Section 5.

[15] As for Half.com, refer to *Business Wire*, January 19, 2000, and *PR Newswire*, October 30, 2000. Djangos.com, another online secondary market for used recordings, launched its service in May 2000 (refer to *Business Wire*, September 25, 2000). Tower Record started to sell used CDs at its online store in September 1999 (refer to *PR Newswire*, September 23, 1999), and Amazon.com began offering its customers the option to buy either new or used CDs from its music store in late 2000 (refer to *Billboard*, November 18, 2000).

[16] Consumers can also download music files from legitimate online music stores, but such an activity is *not* file sharing. Legal music downloading, however, did not take off until the success of Apple's iTunes Music Stores which had been launched on April 28, 2003. Hence, music downloading during the Napster period was predominantly done by file sharing (see http://en.wikipedia.org/wiki/Music_downloading and http://en.wikipedia.org/wiki/Itunes, accessed July 25, 2006). For this reason, I use file sharing and music downloading interchangeably.

additional data set described in the next section. I combine this data set with the CEX data based on common demographic variables. Section 5.5 presents the formal approach for data combination.

4 Data and Descriptive Statistics

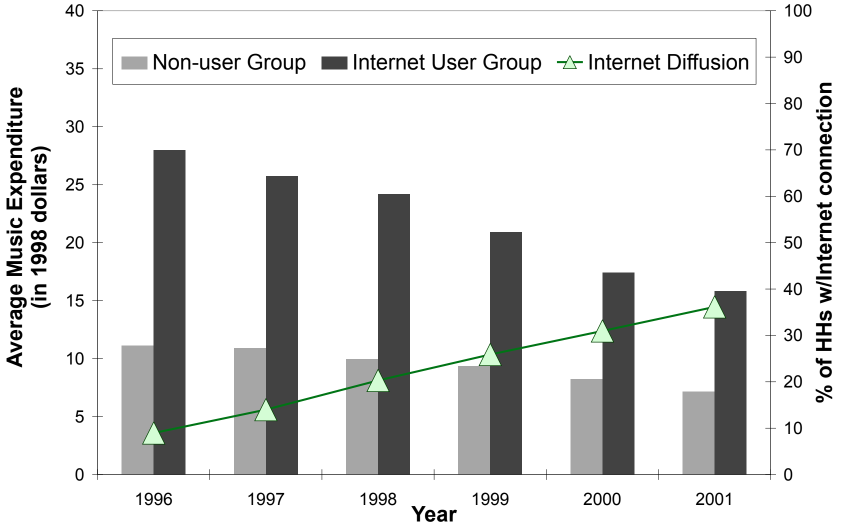
4.1 The Consumer Expenditure Survey (CEX)

The primary source of data is the 1996-2002 Interview surveys of the *Consumer Expenditure Survey* by the U.S. Bureau of Labor Statistics. The CEX is publicly available and consists of random samples of households designed to be representative of the total U.S. population. It is a repeated cross-section with a rotating panel structure. Because of several problems, however, I do not exploit this limited panel structure (see Appendix A for these problems and more details on the CEX). The unit of my analysis is quarterly expenditures of U.S. households.

The CEX is useful for my purposes because it contains a rich set of economic, demographic, and sociological variables as well as detailed data on various expenditures including recorded music and Internet service fees. Recorded music expenditures are defined to be the sum of expenses on CDs, tapes, and LPs purchased other than through a mail-order club as well as from a mail-order club. For Internet access, I use two pieces of information. The first is computer information service expenditures which consist mainly of Internet service fees. The second is whether the household is living in a college dormitory. The CEX identifies students living in college dormitories as separate households from their parents. It is highly likely that most college dormitories already had broadband connections in the late 1990s. Consequently, I define an Internet user group as households that either spent positive amounts on computer information service, or were living in a college dormitory. Appendix A.3 provides more detailed discussion on recorded music expenditures and Internet access in the CEX.

Figure 2 shows the percentage of the Internet user group and changes in average music expenditures for the Internet user group and the non-user group. Note that years in the figure and henceforth refer to the period from June of the year through May of the next year in order to separate pre-Napster and post-Napster periods conveniently. The decline in recorded music expen-

Figure 2: Internet Diffusion and Average Quarterly Music Expenditure in the CEX



ditures for the user group is accompanied by the diffusion of the Internet. This decline could result from the emergence of Napster. However, it is also possible that the use of the Internet in general substituted for music listening. Moreover, the decline can simply result from the compositional change owing to the diffusion of the Internet. Figure 3 documents similar patterns for different demographic groups.[17]

Descriptive statistics organized by Internet adoption and year are presented in Table 1. The table shows substantial differences between the Internet user group and the non-user group. Internet users are younger, richer, more educated, and likely to live in urban or more populated areas. This demographic pattern is also observed in other U.S. data (see, e.g., U.S. Census Bureau 2001; NTIA 2002). Moreover, Internet users tend to spend more money on recorded music and entertainment goods. In particular, about 80% of Internet non-users did not purchase recorded music, and most did not spend on entertainment, either. Hence, the non-user group does not appear to be comparable to the user group.

Table 1 also documents significant compositional changes between the two groups over time.

[17] Age refers to the age of the household head. All age groups exclude households with children aged 6-17, which consist of a separate group.

Table 1: Descriptive Statistics[a] for Internet User and Non-user Groups

Year[b]	1997		1998		1999		2000	
	Internet User (1)	Non-user (2)	Internet User (3)	Non-user (4)	Internet User (5)	Non-user (6)	Internet User (7)	Non-user (8)
Average Expenditure								
Recorded Music	$25.73	$10.90	$24.18	$9.97	$20.92	$9.37	$17.42	$8.22
Entertainment	$195.03	$96.71	$193.38	$84.92	$182.42	$80.19	$164.88	$71.44
Zero Expenditure								
Recorded Music	.56	.79	.60	.80	.64	.81	.68	.83
Entertainment	.08	.32	.09	.35	.14	.39	.17	.44
Demographics								
Age	40.2	49.0	42.3	49.0	44.1	49.4	44.3	49.9
Income	$52,887	$30,459	$51,995	$28,169	$49,970	$26,649	$47,510	$26,336
High School Grad.	.18	.31	.17	.32	.21	.32	.22	.33
Some College	.37	.28	.35	.27	.34	.27	.36	.27
College Grad.	.43	.21	.45	.21	.42	.20	.37	.20
Manager	.16	.08	.16	.08	.14	.08	.14	.07
Professional	.23	.11	.22	.10	.21	.10	.19	.10
Living in a Dorm	.12	0	.08	0	.05	0	.05	0
Urban	.93	.87	.93	.86	.91	.87	.89	.86
Inside a MSA	.84	.78	.83	.78	.83	.78	.81	.78
Pop. Size > 4 million[c]	.34	.26	.30	.26	.31	.25	.28	.25
Appliance Ownership								
Computer[d]	.79	.27	.81	.28	.80	.28	.81	.32
Sound System	.81	.57	.79	.58	.78	.56	.76	.56
VCR	.83	.72	.86	.74	.86	.72	.85	.72
Total Households[e]								
(in million)	15	91	22	86	28	80	34	76
Observations[f]	3,163	19,052	5,624	21,550	8,191	22,810	9,606	20,919

[a] Statistics are weighted using the weights provided by the CEX.

[b] Years refer to the period from June of the year to May of the next year.

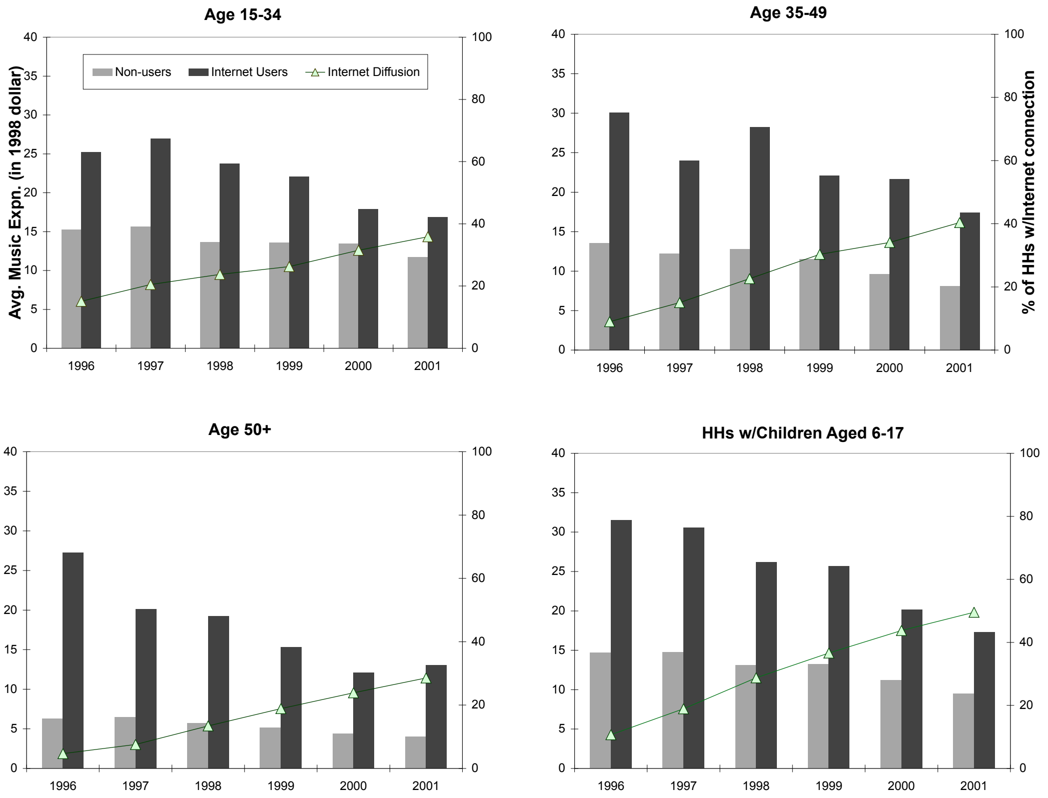
[c] Population size of the primary sampling unit to which the household belongs.

[d] Households can have Internet access without having a computer if they live in a dormitory or use other devices such as WebTV.

[e] The number is the sum of the CEX weights. Total of Internet users and non-users is equal to the number of U.S. households.

[f] Beginning with the first quarter of 1999, the sample size in the CEX has increased by approximately 50 percent.

Figure 3: Internet Diffusion and Average Music Expenditure in the CEX by Age Family Group



For many technologies, it is common that early adopters tend to represent the small fraction of the population that is technologically savvy, whereas later adopters have more diverse demographic and economic characteristics. Similar patterns are also observed in the diffusion of the Internet. Internet users in 1997, for example, are likely to be younger, richer, and more educated than those in 2000. Furthermore, later adopters, who used to be in the non-user group, say in 1997, are likely to be included in the Internet user group, say in 2000.[18]

4.2 The UCLA Internet Survey (UCLAIS)

Because the CEX does not contain information on music downloading, I use an additional data set with this information. Specifically, I use annual household-level surveys on Internet usage collected by the Center for Communication Policy at University of California, Los Angeles (henceforth, *UCLA*

[18] See the Web Appendix for further evidence on compositional changes in the CEX.

Internet Survey, or UCLAIS) for 2000-2002.[19] The surveys were conducted with approximately 2,000 households per year that were selected randomly from the U.S. population (see Appendix B for more details on the UCLAIS).

The UCLAIS contains detailed information on household-level demographics and individual-level Internet usage including, hours spent on music downloading in a typical week. However, it does not include any data on music expenditure. This motivates combining the CEX and the UCLAIS based on demographic data. Though the CEX contains more demographic variables than the UCLAIS, both data sets include a number of common variables. Appendix B compares the CEX and the UCLAIS. It suggests that the two data sets seem to be comparable overall, although there are some differences.

5 Econometric Framework

5.1 The Main Parameters of Interest

To formally define the main parameter of interest and be clear about underlying assumptions required for the identification of the main parameter, consider the following switching regression model with four regimes. I consider four regimes because the CEX is a repeated cross-section. For the i-th household in the CEX sample, define random variables Y_{dti} representing what would be equilibrium recorded music expenditure had the household been in a regime characterized by d and t. The subscript d takes two values, where "1" denotes an Internet access and "0" otherwise. The subscript t is either b or a, indicating periods *b*efore and *a*fter the introduction of Napster. The observed music expenditure Y_i for the i-th observation is then given by

$$Y_i = D_i T_i Y_{1ai} + D_i(1 - T_i)Y_{1bi} + (1 - D_i)T_i Y_{0ai} + (1 - D_i)(1 - T_i)Y_{0bi}, \tag{1}$$

[19] There were several consumer-level surveys by a few private research firms on music downloading and CD purchases. However, their representativeness seemed to be unclear, their demographic information tended to be limited, and more importantly, they were proprietary. In contrast, the UCLAIS was publicly available when I obtained it in January 2004. To acquire the UCLAIS dataset at that time, one had only to agree to use it for research purpose and not to share it with others. The UCLAIS collected random samples, designed to be representative of the U.S. population and contains rich demographic information. See the UCLA Internet Report (2000, 2001, 2003) for more detail on survey methodology.

where D_i is a dummy variable for an Internet access, and $T_i = 1$ if the i-th household is observed at period a and $T_i = 0$ if observed at period b. For an observation with $D_i = 1$ and $T_i = 1$ (i.e. post-Napster Internet user), only Y_{1ai} is observed. Similarly, when $D_i = 1$ and $T_i = 0$ (i.e. pre-Napster Internet user), only Y_{1bi} is observed, and vice versa for the other two cases.

Rearranging the terms in (1) yields

$$Y_i = D_i T_i \theta_i + D_i \gamma_i + T_i \delta_i + Y_{0bi}, \tag{2}$$

where $\gamma_i = Y_{1bi} - Y_{0bi}$, $\delta_i = Y_{0ai} - Y_{0bi}$, and $\theta_i = Y_{1ai} - (Y_{1bi} + \delta_i)$. These three coefficients reflect changes in equilibrium music expenditure attributable to different effects. Specifically, γ_i denotes the general effect of having an Internet access before the presence of Napster, and δ_i is the time effect that a household i would experience without an Internet access. Assuming that the presence of Napster was the only event unique to post-Napster Internet users, $Y_{1bi} + \delta_i$ represents counterfactual music expenditure that Internet users in period a would have incurred if Napster had not been introduced. Hence, θ_i reflects the effect of the presence of Napster for the i-th household.

Given this framework, the main parameter of interest, M, is then defined as

$$\begin{aligned} M &\equiv E(\theta_i | D_i = 1, T_i = 1) \\ &= E(Y_{1ai} | D_i = 1, T_i = 1) - E(Y_{1bi} + \delta_i | D_i = 1, T_i = 1). \end{aligned} \tag{3}$$

Note that I consider the average effect as the main parameter of interest, partly because the mean effect of "treatment on the treated" is one of common parameters to be estimated (see, e.g., Heckman, Lalonde, and Smith 1999), but mostly because one can infer changes in total record sales attributable to the presence of Napster, directly from the mean effect M in (3).

5.2 The DD Matching Estimator

In general, M cannot be identified without any assumption because Y_{1bi}, Y_{0ai}, and Y_{0bi} are not observed for post-Napster Internet users. However, if we had panel data of individual music expenditures and the adoption of the Internet were randomly assigned, M could be identified because

$E(Y_{1bi}|D_i = 1, T_i = 1) = E(Y_{1bi}|D_i = 1, T_i = 0)$ and $E(\delta_i|D_i = 1) = E(\delta_j|D_j = 0)$. Note that I use the subscript j to indicate that the conditional mean is estimated by using different samples from the "treatment group" whose sample is denoted by i.

Unfortunately, however, the CEX is a repeated cross-section, and the adoption of the Internet is unlikely to be random. To address these problems, I thus impose the following assumptions.[20]

(A-1) $E(Y_{1bi}|D_i = 1, T_i = 1; P_b, P_a) = E(Y_{1bj}|D_j = 1, T_j = 0; P_b, P_a)$,

(A-2) $E(\delta_i|D_i = 1, T_i = 1; P_b, P_a) = E(Y_{0aj}|D_j = 0, T_j = 1; P_b, P_a) - E(Y_{0bj}|D_j = 0, T_j = 0; P_b, P_a)$,

where P_b is the propensity to adopt the Internet for the period *b*efore the introduction of Napster, and P_a is the same propensity for the period *a*fter. Specifically, $P_b \equiv \Pr(D = 1|T = 0, X)$ and $P_a \equiv \Pr(D = 1|T = 1, X)$, where X is a vector of observed characteristics. The assumption (A-1) means that conditional on these propensities, the mean of unobserved Y_{1bi} for post-Napster Internet users can be computed from observed Y_{1bj} for pre-Napster Internet users who have the same propensities. The assumption (A-2) implies that conditional on these propensities, the adoption of the Internet is likely to be random, so that the conditional mean of unobserved time effect δ_i for post-Napster Internet users can be computed from observed time effect for Internet non-users.[21]

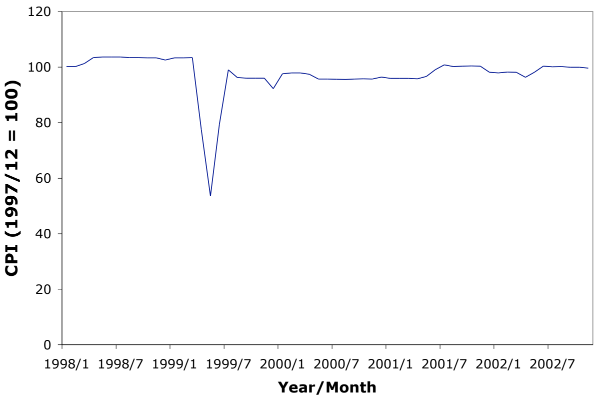
To motivate these assumptions, suppose that there are three types of consumers: early Internet adopters, late Internet adopters, and non-adopters. *Early adopters* are consumers with high income, low learning costs, and high preference for new technology. *Non-adopters* are consumers with low income, high learning costs, and low preference for new technology. Lastly, *late adopters* are in-between. The adoption of the Internet is determined by both consumer heterogeneity and the supply of Internet services. In essence, consumers adopt the Internet if the net benefits of using the Internet exceed the thresholds which vary across different types of consumers over time.

In the pre-Napster period, the Internet was still developing, and hence, was not easy to use. Furthermore, Internet access was expensive. Accordingly, only early adopters had high propensities

[20] Section 5.3 derives (A-1) and (A-2) from rather weaker and more common assumptions based on X.

[21] It is also assumed that conditional on these propensities, the time effect for Internet non-users can be computed from differencing the average music expenditures for post-Napaster non-users and those for pre-Napster non-users.

Figure 4: Consumer Price Index of Internet Services



to adopt the Internet. Note that not all early adopters adopted the Internet in the pre-Napster period. Similarly, it is possible that some late or non-adopters did adopt the Internet in this period. This is likely to have resulted from learning and network effects, which are unobserved in the data. Some early adopters may not have lived close to other early adopters, whereas some late adopters could reside in places surrounded by many early adopters who already adopted the Internet. In the post-Napster period, however, the Internet became better and more accessible to consumers. It is also plausible that Internet service providers lowered Internet access fees.[22] Moreover, learning and network effects imply that more consumers are likely to have Internet access over time. As a result, both early adopters and late adopters have high probabilities of adopting the Internet, while non-adopters still have low probabilities in the post-Napster period.

These three types of consumers can be then characterized by using two propensities P_b and P_a, in that consumers with high P_b and high P_a are early adopters, those with low P_b and high P_a are late adopters, and those with low P_b and low P_a are non-adopters. In this respect, (A-1) and (A-2) indicate that conditional on early adopters, for example, post-Napster Internet users can be matched with pre-Napster Internet users and pre- and post-Napster Internet non-users who

[22] This is the actual case around May through July in 1999 according to the Consumer Price Index (CPI) for the Internet services. A significant drop in the CPI occurred during this period because of special rebates in monthly service charges for some Internet service providers. See Figure 4.

are also early adopters with similar probabilities of adopting the Internet in both periods b and a. Accordingly, the mean of unobserved $Y_{1bi}+\delta_i$ for post-Napster Internet users can be obtained from observed music expenditures of Internet non-users and pre-Napster users with similar P_b and P_a.

To see specifically how (A-1) and (A-2) lead to the identification of M, rewrite (3) as

$$\begin{aligned} M &= \sum_{P_{bi},P_{ai}} E(Y_{1ai}-Y_{1bi}-\delta_i|D_i=1,T_i=1;P_{bi},P_{ai}) \times \Pr(P_{bi},P_{ai}|D_i=1,T_i=1) \\ &= \sum_{P_{bi},P_{ai}} [E(Y_{1ai}|D_i=1,T_i=1;P_{bi},P_{ai}) - E(Y_{1bj}|D_j=1,T_j=0;P_{bi},P_{ai}) \\ &\quad -E(Y_{0aj}|D_j=0,T_j=1;P_{bi},P_{ai}) + E(Y_{0bj}|D_j=0,T_j=0;P_{bi},P_{ai})] \times \Pr(P_{bi},P_{ai}|D_i=1,T_i=1), \end{aligned}$$

where the first equality uses the law of iterated expectation and the second equality follows from (A-1) and (A-2). Though I describe only three types for expositional simplicity, there can be actually a continuum of consumers who can be characterized by P_b and P_a. Consequently,

$$\begin{aligned} M &= \int_{P_{bi},P_{ai}} [E(Y_{1ai}|D_i=1,T_i=1;P_{bi},P_{ai}) - E(Y_{1bj}|D_j=1,T_j=0;P_{bi},P_{ai}) \\ &\quad -E(Y_{0aj}|D_j=0,T_j=1;P_{bi},P_{ai}) + E(Y_{0bj}|D_j=0,T_j=0;P_{bi},P_{ai})]\, dF(P_{bi},P_{ai}|D_i=1,T_i=1). \end{aligned}$$

This equation suggests the following estimator,

$$\begin{aligned} \widehat{M} &= \sum_{i\in I_{1a}} \Big[Y_{1ai} - \widehat{E}(Y_{1bj}|D_j=1,T_j=0;P_{bi},P_{ai}) \\ &\qquad -\widehat{E}(Y_{0aj}|D_j=0,T_j=1;P_{bi},P_{ai}) + \widehat{E}(Y_{0bj}|D_j=0,T_j=0;P_{bi},P_{ai})\Big] \times w_i, \end{aligned} \tag{4}$$

where I_{1a} denotes the post-Napster Internet user group, and $\widehat{E}(Y_{1bj}|D_j = 1, T_j = 0; P_{bi}, P_{ai})$, $\widehat{E}(Y_{0aj}|D_j=0,T_j=1;P_{bi},P_{ai})$, and $\widehat{E}(Y_{0bj}|D_j=0,T_j=0;P_{bi},P_{ai})$ are the conditional expectation estimators for each group conditional on P_{bi} and P_{ai} of i observation in I_{1a}, and the weight for i is given by $w_i = \frac{(\text{CEX weight})_i}{\sum_{k\in I_{1a}}(\text{CEX weight})_k}$. See Appendix C for more details on the estimator.

The proposed estimator in (4) is a modified version of the nonparametric DD matching (DDM) method developed in Heckman, Ichimura, and Todd (1997, 1998). To account for compositional changes and the repeated cross-sectional feature of the CEX, I modify the standard DDM method by nonparametrically matching each observation in the post-Napster Internet user group with observations in the pre-Napster Internet user group and the Internet non-user group based on P_{bi} and P_{ai} of each i in I_{1a}.

5.3 Discussion on Identifying Assumptions for the DDM

The DDM method assumes random selection conditional on observables that determine the adoption of the Internet. Similarly to Heckman, Ichimura, and Todd (1997, 1998) and Heckman and Smith (1999), I formally assume that $E(Y_{1bi}+\delta_i|D_i=1,T_i=1;X_i)-E(Y_{1bj}|D_j=1,T_j=0;X_i)=E(Y_{0aj}|D_j=0,T_j=1;X_i)-E(Y_{0bj}|D_j=0,T_j=0;X_i)$, which is implied by

(A-1)′ $E(Y_{1bi}|D_i=1,T_i=1;X_i)=E(Y_{1bj}|D_j=1,T_j=0;X_i)$,

(A-2)′ $E(\delta_i|D_i=1,T_i=1;X_i)=E(Y_{0aj}|D_j=0,T_j=1;X_i)-E(Y_{0bj}|D_j=0,T_j=0;X_i)$,

This assumption is a variant of "selection on observables" (Heckman and Robb 1985). It says that in the absence of Napster, changes in average music expenditure of Internet users with observed characteristics X_i are equal to those of Internet non-users with the same X_i. This assumption is made in order to exploit the detailed demographic information in the CEX. The conditional expectation in (A-1)′ and (A-2)′, however, is difficult to estimate nonparametrically because of the high-dimensionality in X_i. Accordingly, it is desirable to reduce the dimension of X_i while keeping sufficient information to maintain random selection. Rosenbaum and Rubin (1983) provides an useful result in this regard. They establish that if Y_i is independent of D_i conditional on X_i, and $0<\Pr(D_i=1|X_i)<1$, then random selection is ensured by conditioning only on the propensity score, $\Pr(D_i=1|X_i)$. This implies that matching can be based on a one-dimensional propensity score, instead of the high-dimensional observables X_i.

In the repeated cross-section context with compositional changes, however, matching should be based on at least P_b and P_a to separate different types of consumers such as early adopters, late adopter, and non-adopters.[23] Moreover, using these two propensity scores has economic meaning as discussed in the previous section. In Appendix C.1, I show that the assumption of (A-1)′ and (A-2)′ imply (A-1) and (A-2) under independence between X and T. As a result, conditioning only on

[23] If one uses the propensity score based on just one period, say a, then early adopters and late adopters would have similar $\Pr(D=1|X,T=1)$. One could use the propensity score based on the entire periods of b and a. Two observations may have the same $\Pr(D=1|X)=0.6$, for instance. However, one observation may have $\Pr(D=1|X,T=t)=0.6$, $\forall t=0,1$, whereas the other can have $\Pr(D=1|X,T=0)=0.45$ and $\Pr(D=1|X,T=1)=0.75$. Consequently, it is difficult to distinguish different types of consumers by using only one propensity score.

two propensity scores is sufficient for identification of M.[24] In Appendix C.3, I also plot estimated densities of P_b and P_a and further provide an empirical justification for using two propensity scores.

Selection on observables indicates that conditional on a vector of observed variables, Internet adoption is independent of recorded music expenditures. See Appendix C.5 for demographic variables that I use to account for the selection problem. These include age, race, education, appliance ownership (e.g. video tape or disc players, sound component system, etc.), occupation, family composition, work, housing, income, and geographic information. Upon conditioning on rich demographic variables in the CEX, most determinants of Internet adoption discussed in the previous section, except Napster, are unlikely to be correlated with recorded music expenditures.[25] Testing this assumption, nonetheless, is not feasible. To justify the use of the DDM method, I thus examine a case in which the assumption is least likely to be satisfied and consider its implication.

For consumers who adopted the Internet during the pre-Napster period, the introduction of Napster can be treated as exogenous technological event. The diffusion of the Internet, however, did not stop after the emergence of Napster. Though many later Internet adopters started to use the Internet without any consideration of Napster, some consumers might have adopted the Internet just to download free music. They are likely to have high preference for free digital music and thus have substantially reduced music expenditure after using Napster. For this reason, the post-Napster Internet user group might include more consumers with high preference for free digital music. Because this preference is unobserved, selection on observables is not satisfied. In other words, even upon conditioning on various observables, Internet adoption during the Napster period might be correlated with zero or low recorded music expenditure. Accordingly, the DDM method may contain a possible negative bias due to this “selection on unobservables”. Unfortunately, without observing preference for free digital music, I cannot account for this potential negative

[24] Note that independence between X and T is innocuous in my application because I consider two years before and after the introduction of Napster. These four years seem to be too short to lead to significant changes in observed demographics of the U.S. households.

[25] One may be concerned that Internet users might be more likely to copy CDs or use pirated CDs. However, note that these activities have existed even before the introduction of Napster. Though there may be some difference in these activities between Internet users and non-users, this problem is not likely to be serious because I compare not only Internet users with non-users, but also pre-Napster Internet users with post-Napster Internet users.

bias. Nevertheless, the DDM method is still of value because it provides meaningful upper bounds for the negative effect of the presence of Napster on recorded music expenditure.

5.4 Comparison of the DDM Method and the DD Regressions

The DDM approach estimates the effect of the presence of Napster on equilibrium music expenditures, while flexibly accounting for selective differences between Internet users and non-users over time. As discussed in Section 3, however, the estimated effect could be confounded with other new online activities during the Napster period. To complement the DDM, I thus consider additional approach to decompose the effect of the presence of Napster into the effect of actual downloading and the effect of other new online activities. Specifically, I combine the CEX and the UCLAIS because music expenditure and downloading are not observed together. In this additional approach, one should still account for compositional changes. For this reason, I first consider nonparametric bounds that use the DDM estimates. Using the Fréchet bound, this approach combines two data and provides bounds on the effect of music downloading, without imposing arbitrary parametric assumptions. This approach, however, uses only partial information contained in the UCLAIS, and the estimated bounds are not sufficiently informative.[26]

Consequently, I use a two-sample instrumental variable (2SIV) approach[27] which exploits more information from the data and uses linearity in the DD regressions to allow for data combination. The 2SIV is straightforward to implement in my application and provides point estimates of the effect of actual music downloading. Nevertheless, it is based on the DD regressions, which raises a question about whether the 2SIV would account for the selective differences between Internet users and non-users. For this reason, I examine below what other assumptions are required for the DD regressions to identify M. The next section then develops the 2SIV based on the DD regressions.

The conventional DD regression basically attempts to estimate $M \equiv E(\theta_i|D_i = 1, T_i = 1)$ directly from (2). Because Y_{1bi}, Y_{0ai}, and Y_{0bi} are not observed for post-Napster Internet users,

[26] See the Web Appendix for details on the bound approach and the results.

[27] See Angrist and Krueger (1992), Arellano and Meghir (1992), and Moffitt and Ridder (2003) for further discussion on the 2SIV.

however, the identification of M in the DD regression also requires (A-1)$'$ and (A-2)$'$. The DD regression, however, requires further assumptions because of two additional problems. First, Y_{0bi} in equation (2) is observed only for households without an Internet access in period b, and second, all three coefficients in (2) have the subscript i, indicating heterogeneous effects. To account for these additional problems, the DD regression further imposes parametric assumptions on conditional expectation of Y_{0bi} and restricts heterogeneous effects. Specifically, it assumes

(A-3) $E(Y_{0bi}|D_i, T_i, X_i) = \alpha + X_i\beta,$

(A-4) $E(\theta_i|D_i = 1, T_i = 1, X_i) = E(\theta_i|D_i = 1, T_i = 1) \equiv \theta,$

$E(\gamma_i|D_i = 1, T_i = 0, X_i) = E(\gamma_i|D_i = 1, T_i = 0) \equiv \gamma,$

and $E(\delta_i|D_i = 0, T_i = 1, X_i) = E(\delta_i|D_i = 0, T_i = 1) \equiv \delta, \quad \forall i.$

Note that (A-3) implicitly supposes no selective differences in Y_{0bi} between Internet users and non-users conditional on X_i. Given these assumptions, the DD regression is written as

$$Y_i = D_iT_i\theta + D_i\gamma + T_i\delta + \alpha + X_i\beta + \nu_i. \tag{5}$$

It is straightforward to show $E(\nu_i|D_i, T_i, X_i) = 0$ under (A-1)$'$, (A-2)$'$, (A-3), and (A-4). See the Web Appendix for more details.

5.5 Two Sample Instrumental Variable

To decompose the effect of the presence of Napster based on the DD regressions, consider (1). For each household with $D_i = 1$ and $T_i = 1$, let Y^1_{1ai} be what would be recorded music expenditure had the household downloaded music, and similarly Y^0_{1ai} had the household not downloaded. Hence, $Y_{1ai} = DM_iY^1_{1ai} + (1 - DM_i)Y^0_{1ai}$, where $DM_i = 1$ for households who download music and $DM_i = 0$ otherwise. Inserting this into equation (1), one can extend equation (2) as follows.

$$Y_i = D_iT_i\theta_{0i} + D_iT_iDM_i\theta_{1i} + D_i\gamma_i + T_i\delta_i + Y_{0bi},$$

where $\theta_{0i} = Y^0_{1ai} - (Y_{1bi} + \delta_i)$, representing the effect of other new online activities during the Napster period, and $\theta_{1i} = Y^1_{1ai} - Y^0_{1ai} - (Y_{1bi} + \delta_i)$, reflecting the effect of music downloading.

Using (A-1)$'$, (A-2)$'$, (A-3), and (A-4), one can rewrite the previous equation as

$$Y_i \quad = \quad D_i T_i \theta_0 + D_i T_i DM_i \theta_1 + D_i \gamma + T_i \delta + \alpha + X_i \beta + \varepsilon_i, \tag{6}$$

where $\theta_0 = E(\theta_{0i}|D_i = 1, T_i = 1)$, $\theta_1 = E(\theta_{1i}|D_i = 1, T_i = 1)$, and $\varepsilon_i = D_i T_i\{\theta_{0i} + DM_i\theta_{1i} - E(\theta_{0i}|D_i = 1, T_i = 1, X_i) - DM_i E(\theta_{1i}|D_i = 1, T_i = 1, X_i)\} + D_i\{\gamma_i - \gamma\} + T_i\{\delta_i - \delta\} + \{Y_{0bi} - E(Y_{0bi}|X_i)\}$. Accordingly, this equation decomposes M into θ_0 and θ_1.

In equation (6), two issues need to be addressed in order to identify θ_0 and θ_1. First, the error term ε_i can contain selection bias. Therefore, I maintain the assumptions (A-1)$'$, (A-2)$'$, (A-3), and (A-4). Similarly to the DD regressions, it can be shown that these assumptions imply $E(\epsilon_i|D_i, T_i, X_i) = 0$. These assumptions, however, may not hold in the data. To check their validity in the CEX samples, the Web Appendix proposes a simple test and reports the results. From this test, I find statistically significant evidence against these assumptions for households aged 35-49 and those over 50, whereas I do not find such evidence in the case of households aged 15-34 and those with children aged 6-17. This suggests that only for the latter two demographic groups, the 2SIV is likely to be a valid approach to decompose the effect of the presence of Napster.

The second problem for the identification is that DM_i is not observed in the CEX. Hence, I use the UCLAIS and impute the probability of downloading. To explain the idea, let $Z_i = (Z_{1i}, Z_{2i}, D_i, T_i)$ denote a vector of common variables in both the CEX and the UCLAIS, where Z_{1i} is a vector of common variables included in X_i, while Z_{2i} is excluded from X_i. Because the CEX contains more demographic variables than the UCLAIS, I also define X_{2i} to be a vector of covariates only in the CEX, so that $X_i = (Z_{1i}, X_{2i})$. Taking a conditional expectation of (6) yields

$$E(Y_i|Z_i, X_{2i}) = D_i T_i \theta_0 + D_i T_i E(DM_i|Z_i, X_{2i})\theta_1 + \alpha + D_i\gamma + T_i\delta + X_i\beta + E(\varepsilon_i|Z_i, X_{2i}). \tag{7}$$

Though $E(DM_i|Z_i, X_{2i})$ cannot be estimated from the CEX, it can be imputed using the UCLAIS. Assume that X_{2i} does not determine music downloading. This conditional expectation is then equal to $\Pr[DM_i = 1|Z_i]$. In the application, I first estimate $\Pr[DM_i = 1|Z_i]$, applying a Probit to the UCLAIS. Using the estimates, I next compute the predicted value of $\Pr[DM_i = 1|Z_i]$ for each observation in the CEX. Estimation of equation (7) is then straightforward. In this estimation, exclusion restrictions (i.e. Z_{2i}) are not critical for the identification, since the assumptions (A-1)$'$, (A-2)$'$, (A-3), and (A-4) are sufficient for $E(\epsilon_i|D_i, T_i, X_i) = 0$.[28]

6 Estimation Results

6.1 The Nonparametric DD Matching Estimates

Table 2 reports the estimates from the DDM method.[29] Local linear matching[30] is used to construct the counterfactual, $E(Y_{1bi} + \delta_i|D_i = 1, T_i = 1)$. The DDM estimate quantifies the main parameter of interest defined in Section 5.1. The estimate is estimated precisely. The results indicate that the average Internet user during the Napster period would have spent \$1.45 more per quarter on recorded music in the absence of Napster. Note that the estimated magnitude is smaller than those from the conventional DD methods.[31] This implies the importance of accounting for composition bias. In other words, the DD methods are less likely to account for a large number of late adopters with a low willingness to pay for recorded music. These late adopters might have downloaded music files but would not have purchased CDs even in the absence of Napster.

For further implications of the DDM result, consider the following back-of-the-envelope calculation of the impact of Napster on total record sales during the Napster period. The average percentage of Internet users in the CEX is 26% during the first year of the Napster period and

[28] Nevertheless, the estimation implicitly uses both a functional form for $\Pr[DM_i = 1|Z_i]$ (i.e. Probit) and exclusion restrictions, which further ensures the identification of the parameters. Specifically, I use Internet connection speed as Z_{2i}. It might be plausible that having high speed Internet connection is correlated with actual downloading but it may not be related to recorded music expenditure. The UCLAIS provides information on the Internet speed. However, this information is limited in the CEX. As a result, in the CEX I define households as having high speed Internet access if they are living in college dormitory or they spend more than a certain cutoff amount on computer information service. For this cutoff, I use information contained in the *Current Population Survey, August 2000: Internet and Computer Use Supplement* collected by the U.S. Bureau of Census. For each demographic group, I compute the average monthly Internet service fees paid by households with high speed Internet access.

[29] Two propensity scores are estimated by using Probit estimation. The Probit results are reported in Appendix C.5. The numbers of samples for the pre- and post-Napster Internet user (or non-user) groups are 7,806 and 17,797 (or 38,318 and 43,729). 0.5% of samples in the post-Napster user group are trimmed because there is no match in other groups within the neighborhood defined by the fixed bandwidth.

[30] As Heckman, et al. (1997) discuss, local linear estimators avoid the boundary bias problem associated with kernel estimators and have several nice properties that make them superior to the standard kernel regression estimators. Because a substantial fraction of observations in the CEX are at a boundary close to zero, I thus use local linear matching instead of kernel matching. See Appendix C.2 for more detail on local linear matching.

[31] For the purpose of comparison, the Web Appendix presents the results from the conventional DD approach.

Table 2: The DDM Estimate for the Main Parameter of Interest[a]

	DDM	$E(Y_{1ai} \mid D_i = 1, T_i = 1)$	$E(Y_{1bi} + \delta_i \mid D_i = 1, T_i = 1)$
Estimate	-1.446 (0.624)	19.048 (0.268)	20.494 (0.565)

[a] The results are estimated from local linear regression matching. A fixed bandwidth of .07 and biweight kernel, described in Appendix C, are used. Bootstrapped standard errors are reported in parentheses. They are based on 100 replications with 60% sampling, also described in Appendix C.

31% during the second year. There were approximately 100 million households in the U.S. during this period. Noting that the estimate is quarterly expenditure in terms of 1998 dollars, total record sales decline in the period attributable to Napster is then given by $-\$329.69$ million $=$ 100 million $\times (0.26 \times 4 + 0.31 \times 4) \times (-\$1.446)$. According to the CEX, the decrease in total record sales[32] in the Napster period compared to the pre-Napster period is \$832.24 million. This suggests that 39.6% of sales decline could be attributable to the presence of Napster.

6.2 The DDM Estimates for Age and Family Groups

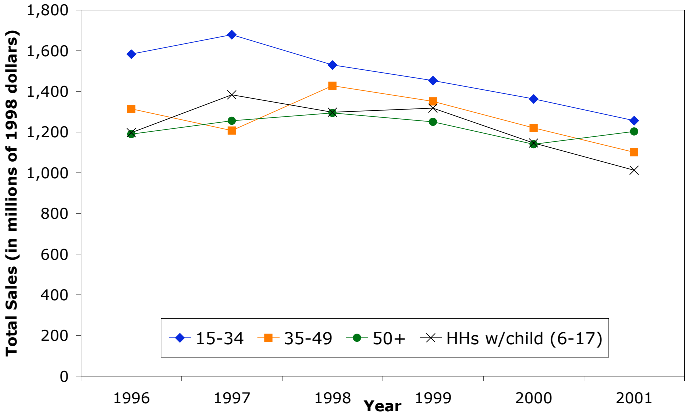
To investigate which demographic groups are responsible for the sales decline attributable to file sharing, I consider different age and family groups. Specifically, I examine four mutually exclusive groups: households with children aged 6-17,[33] households with heads aged 15-34, households with heads aged 35-49, and households with heads aged over 50. I choose these four groups for four distinct reasons. First, the number of observations in each group is similar, except those aged over 50.[34] Second, total record sales for each group are comparable though the 15-34 age group spent

[32] Total record sales are the weighted sum of recorded music expenditure using the CEX weights. A weight assigned to each sample household in the CEX can be used to estimate the number of U.S. households with the same demographic composition as that household, so that the weighted sum of expenditures can reflect sales among total U.S. households.

[33] Young consumers aged 6-17 are presumed to be heavy music buyers. In the CEX, however, their expenditures are included in their parents' expenditure unless they are financially independent. To account for music expenditures for those young consumers, I consider a separate group for households with children. The other groups do not have children aged 6-17. However, one may worry about a potential measurement error in music expenditures for those with children aged 6-17, because parents might not know about their children's music expenditures. Nonetheless, there is no reason to believe that such a measurement error is correlated with Internet access or file sharing. Therefore, even if the CEX has the measurement error for this demographic group, the DD approach can account for this problem.

[34] The group aged 15-34 contains 21% of samples; those aged 35-49 include 20%; those with children aged 6-17 contain 18%; those aged over 50 include 41% of samples. In particular, the percentage of each group in Internet users during the Napster period is comparable.

Figure 5: Total Music Expenditure by Age Group



slightly more than other groups (see Figure 5). Third, each group seems to have experienced a different level of compositional changes (see the Web Appendix for more discussion). Fourth, the DDM estimates for different groups may imply the effect of different intensity in music downloading on music expenditure.

Panel A of Table 3 presents the DDM estimates of the main parameter of interest for these four groups. The DDM estimate for those with children aged 6-17 is -$3.26, and its standard error is 0.75. The estimate for those aged 15-34 is -$2.99 but is not estimated precisely. For those aged 35-49 and those aged over 50, the estimated effect is statistically insignificant. For comparison, panel B of Table 3 reports the conventional DD estimates of the same parameters. In particular, the DD estimates for those aged over 50 show a significant negative impact of Napster on recorded music expenditure. The comparison of the DDM and DD estimates for this group implies that the DD overestimates the negative impact. On the other hand, the group with children aged 6-17 does not appear to have experienced significant compositional changes (see the Web Appendix). This seems to explain the slight difference between the DDM and DD estimates for this group.

The DDM estimate is estimated precisely only for those with children aged 6-17. Nonetheless, the estimated magnitudes can be used for similar back-of-the-envelope calculations as in the previous section. The CEX reports that the percentages of four groups in the post-Napster Internet

Table 3: The Estimates for Age and Family Groups[a]

	Age 15-34	Age 35-49	Age 50+	HHs w/children Aged 6-17
	(1)	(2)	(3)	(4)
	A. The DDM Estimates[b]			
θ	-2.992 (3.012)	-0.453 (0.988)	-0.408 (0.765)	-3.256 (0.748)
	B. The Conventional DD Estimates[c]			
θ	-3.432 (1.284)	-2.097 (1.342)	-3.449 (0.774)	-3.258 (1.203)
	C. Results from the 2SIV[d]			
θ_1	-2.719 (2.079)	13.054 (12.295)	2.334 (4.586)	-22.510 (6.889)
θ_0	-2.427 (0.949)	-4.160 (1.955)	-3.762 (0.798)	-0.120 (1.092)
	D. The Mean of $\widehat{\Pr}(DM=1\|Z)$ for post-Napster Internet users[e]			
π	0.346	0.158	0.133	0.140

[a] Standard errors are in parentheses. The dependent variable is music expenditure in 1998 dollar.

[b] The DDM results are estimated from local linear regression matching. The propensity scores for both pre- and post-Napster periods are estimated separately for each group, excluding age and family composition variables from Probit estimation in Appendix C. A fixed bandwidth of .07 and biweight kernel are used. Bootstrapped standard errors are reported in parentheses. They are based on 100 replications with 100% sampling.

[c] The table reports D_iT_i interaction in the DD regression (5) that includes controls such as age, education, income, appliance, occupation, family composition, and region.

[d] Bootstrapped standard errors are in parentheses. They are based on 500 replications with 80% sampling, described in Appendix D. The table reports coefficient estimates for D_iT_i (i.e. θ_0) and $D_iT_iDM_i$ (i.e. θ_1) in the regression (7) that includes age, income, education, appliance ownership, occupation, family composition, and region. The regression excludes a dummy for high speed Internet access, which is defined to be 1 if living in college dormitory, or computer information service $\geq$ \$96.58 for Age 15-34, \$92.76 for Age 35-49, \$85.53 for Age 50+, and \$87.11 for Family w/children aged 6-17. For each group, I compute the cutoffs using the CPS. These cutoffs are 3×average monthly Internet service fees that households with high speed Internet access paid.

[e] The reported π is the average imputed probabilities of DM for post-Napster Internet users in the CEX.

user group are 21% for those aged 15-34, 23% for those aged 35-49, 31% for those aged over 50, and 24% for those with child aged 6-17. Using these percentages and the DDM estimates, together with the same information in the previous section, I find that the DDM estimate for households with children aged 6-17 is translated into -\$196 million, which accounts for about 20% of total record sales decline during the Napster period.[35]

For households aged 35-49 and those over 50, the DDM estimates imply -\$26 million and -\$31 million, respectively. The standard errors for these estimates, however, suggest that these are statistically indistinguishable from zero. As for those aged 15-34, the DDM estimate is translated

[35] The 95% confidence interval is given by (-\$286 million, -\$106 million).

into -\$159 million, which approximately explains another 20% of total sales decline during the Napster period. However, we cannot be confident about this result because of the high standard error for this demographic group. Moreover, other new online activities may confound this result.

6.3 Combining Results from the DDM and the 2SIV

To obtain the 2SIV estimates, I first apply a Probit to the 2000-2001 UCLAIS data[36] and estimate the probability of downloading conditional on common variables that include age, education, income, computer ownership, family composition, working status, Internet access, and high speed Internet access. Appendix D reports results from this Probit estimation. Note that I compute separate Probit estimates for each demographic group. Using these estimates, I next impute the probabilities of DM for all observations in the CEX. I then estimate equation (7) for each demographic group. Panel C of Table 3 presents the coefficient estimates of θ_0 and θ_1 from the 2SIV method. The 2SIV approach decomposes θ in panel B into θ_1, the effect of actual downloading, and θ_0, the effect of other online activities during the Napster period. This decomposition can be interpreted as $\theta = \theta_0 + \theta_1 \pi$, where π, reported in Panel D, is the mean of the imputed probabilities of downloading for post-Napster Internet users in the CEX.

To connect the 2SIV results with the DDM findings, consider Table 3 again. For households with children aged 6-17, the DDM estimate indicates a precisely estimated substantial negative effect of the presence of Napster. Note that for these households, I do not find statistically significant evidence against underlying assumptions in the DD regressions (see the Web Appendix). Moreover, the magnitude of the estimated θ from the DD regression is almost identical to that from the DDM. For this reason, the 2SIV results for this demographic group can be interpreted as decomposing the DDM estimate of θ. The 2SIV results report that the effect of actual downloading is considerably negative and statistically significant, while the effect of other new online activities is statistically indistinguishable from zero. Consequently, these results suggest that the DDM estimate for this

[36] I use the 2000-2001 UCLAIS because each survey was conducted around June of each year, so that these surveys cover the Napster period.

demographic group is likely to represent the effect of file sharing.[37]

The DDM estimate for households aged 15-34 reports a fairly substantial negative effect of the presence of Napster. However, it is not estimated precisely. Furthermore, the 2SIV results for these households do not indicate significant negative effect of downloading. Recall also that I do not find statistically significant evidence against the DD assumptions for this demographic group. As a consequence, these results suggest that we can more confidently rule out the significant negative effect of file sharing on recorded music expenditure for this demographic group.

In contrast to the estimates for the preceding two demographic groups, the estimates for those aged 35-49 and over 50 do not seem to be comparable between the DDM results and the 2SIV. Note that I find statistically significant evidence against the DD assumptions for these households (see the Web Appendix). Moreover, the DDM estimates for these groups are much different from the estimated θ from the DD regression. Therefore, the 2SIV results do not appear to be connected to the DDM results. Nevertheless, the 2SIV estimates suggest a positive effect of downloading and a negative effect of other new online activities. More importantly, the magnitude of the DDM estimates for these households is relatively small and statistically insignificant. Accordingly, these results are unlikely to be consistent with the significant negative effect of file sharing.

7 Conclusion

To what extent is file sharing culpable for the recent slump in record sales, and which demographic group is primarily responsible for the sales decline due to file sharing? I study changes in household-level recorded music expenditure between the periods before and after the introduction of Napster, accounting for the likely relationship between music expenditure and the propensity to adopt the Internet, as well as potentially confounding factors. I find that file sharing can account for approximately 20% of the sales decline in recorded music during the Napster period. This negative effect of file sharing on recorded music expenditure is concentrated in a particular demographic group:

[37] Note that the DD estimate for this group is decomposed as $-3.258 \approx -0.120 + (-22.510) \times 0.140$. As a result, the back-of-the-envelope calculation using the 2SIV results for this group is almost identical to that in Section 6.2.

households with children aged 6-17.

These findings have further implications for the effectiveness of the recording industry's recent attempts to prosecute copyright violators as well as some current copyright-related legislation. The recording industry's lawsuits against file sharing services, and individual users in particular, could stop some users from downloading music. However, according to my findings, this is unlikely to lead many Internet users to spend more on recorded music. Moreover, intensive users with technological savvy, such as teenagers, could continue to use more advanced file sharing technology. Regarding current copyright-related legislation, my findings question the strength of the alleged negative relationship between file sharing and record sales that is presented as evidence to support legislation pending in the U.S. Congress (see, e.g., Section 2 of Congressional Findings in the "Artists' Rights and Theft Prevention Act of 2004" (S. 1932); Findings in the "Piracy Deterrence and Education Act of 2004" (H.R. 4077)).

The approaches in this paper are not limited to the analysis of the recording industry. The emergence of new digital technologies and their impact on the sales of copyrighted goods is a more general issue. The evaluation of these impacts, nonetheless, is complicated by problems similar to those addressed in this paper. First, the diffusion of these technologies entails compositional changes in their users. Second, as concomitant technologies are introduced, it becomes more difficult to discern the effect of the original technologies. The approaches taken in this paper account for these problems and can be extended to other applications involving the effect of digital technology on the sales of copyrighted goods such as movies, computer software, or video games.

References

Angrist, Joshua D. and Krueger, Alan B. "The Effect of Age at School Entry on educational Attainment: An application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association*, June 1992, *87*, pp. 328-36.

Arellano, Manuel and Meghir, Costas. "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets." *Review of Economic Studies*, July 1992, *59*(3), pp. 537-57.

Bakos, Yannis; Brynjolfsson, Erik and Lichtman, Douglas. "Shared Information Goods." *Journal of Law and Economics*, April 1999, *42*, pp. 117-55.

Battistin, Erich. "Errors in Survey Reports of Consumption Expenditures." Manuscript. The Institute for Fiscal Studies, May 2003, IFS working paper 03/07.

Besen, Stanley M. and Kirby, Sheila N. "Private Copying, Appropriability, and Optimal Copyright Royalties." *Journal of Law and Economics*, October 1989, *32*(2), pp. 255-80.

Blackburn, David. "On-line Piracy and Recorded Music Sales." Manuscript. Harvard University, October 2004.

Fanning, Shawn. "Testimony of Shawn Fanning, Founder, Napster, Inc. Before the Senate Judiciary Committee." *United States Senate Committee on the Judiciary Online Library*, October 9, 2000, available at
http://judiciary.senate.gov/testimony.cfm?id=199&wit_id=273

Fisher III, William W. *Promises to Keep: Technology, Law, and the Future of Entertainment.* Stanford: Stanford University Press, 2004.

Heckman, James J.; Ichimura, Hidehiko and Todd, Petra E. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies*, October 1997, *64*(4), pp. 605-54.

_________ "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies*, April 1998, *65*(2), pp. 261-94.

Heckman, James J.; Lalonde, Robert J. and Smith, Jeffrey A. "The Economics and Econometrics of Active Labor Market Programs." In *The Handbook of Labor Economics, Volume 3A*, edited by Orley C. Ashenfelter and David Card. Elsevier, 1999, pp. 1865-2097.

Heckman, James J. and Robb, Richard. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, edited by James Heckman and Burton Singer. Cambridge: Cambridge University Press, 1985, pp. 156-245.

Heckman, James J. and Smith, Jeffrey A. "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies." *Economic Journal*, July 1999, *109*(457), pp. 313-48.

Hong, Seung-Hyun "The Diffusion of the Internet and Changes in the Household-Level Demand for Entertainment." Manuscript. University of Illinois, October 2006.

Landes, William M. and Posner, Richard A. *The Economic Structure of Intellectual Property Law* Cambridge, Mass.: Harvard University Press. 2003.

Moffitt, Robert and Ridder, Geert "The Econometrics of Data Combination." Manuscript. November 2003. Forthcoming in *The Handbook of Econometrics, Volume 6*, edited by James Heckman and Edward Leamer.

National Telecommunications and Information Administration (NTIA). "A Nation Online: How Americans are Expanding Their Use of the Internet." U.S. Department of Commerce, February 2002.

Oberholzer, Felix and Strumpf, Koleman. "The Effect of File Sharing on Record Sales: An Empirical Analysis." Forthcoming in *Journal of Political Economy*, 2007.

Posner, Richard A. "Intellectual Property: The Law and Economics Approach." *Journal of Economic Perspective*, Spring 2005, *19*(2), pp. 57-73.

Rob, Rafael and Waldfogel, Joel. "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students." *Journal of Law and Economics*, April 2006, *49*(1), pp. 29-62.

Romer, Paul. "When Should We Use Intellectual Property Rights?" *American Economic Review*, May 2002, *92*(2), pp. 213-16.

Rosenbaum, Paul and Rubin, Donald "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, April 1983, *70*(1), pp. 41-55.

Sherman, Cary. "Testimony of Cary Sherman, President and General Counsel, Recording Industry Association of America, Before the Senate Judiciary Committee." *United States Senate Committee on the Judiciary Online Library*, September 9, 2003,
available at http://judiciary.senate.gov/testimony.cfm?id=902&wit_id=2562

Takeyama, Lisa N. "The Welfare Implications of Unauthorized Reproduction of Intellectual Property in the Presence of Demand Network Externalities." *Journal of Industrial Economics*, June 1994, *42*(2), pp. 155-66.

The UCLA Center for Communication Policy. "The UCLA Internet Report: Surveying the Digital Future." November 2000, November 2001, and February 2003.

U.S. Census Bureau. "Home Computers and Internet Use in the United States: August 2000." U.S. Department of Commerce, September 2001.

Varian, Hal R. "Copying and Copyright." *Journal of Economic Perspective*, Spring 2005, *19*(2), pp. 121-138.

Zentner, Alejandro. "Measuring the Effect of File sharing on Music Purchases." *Journal of Law and Economics*, April 2006, *49*(1), pp. 63-90.

Appendix A Details on the Consumer Expenditure Survey

A.1: Basic Description

The main source of data in this paper is the CEX. The CEX consists of an Interview panel survey and a Diary survey. These surveys are conducted at the level of a consumer unit (CU) which is essentially a household. Both surveys contain the similar information on demographics and expenditures, but each survey uses its own questionnaire and independent sample. The Interview survey has more detailed information on expenditures relevant to the analysis in this paper than the Diary survey. Moreover, it includes data on household ownership of appliances such as computers and sound component systems, whereas the Diary does not. As a result, I use the Interview survey.

A.2: Rotating Panel Structure of the CEX and Its Problems

The Interview survey interviews each household five times for every three months: the first for general demographics and the second through the fifth for household expenditures. In the Interview survey, new panels are added every month of the year, so that twenty percent of the sample that finished its final interview in the previous quarter are replaced by panels that are initiated newly in each quarter. Because of this rolling panel feature in the CEX, I organize data based on the first month of households' quarterly expenditure.

However, I do not exploit this limited panel structure for several reasons. First and most importantly, the period of the panel survey lasts only one year for a household. Note that music expenditures normally peak around December due to the Christmas season, which explains a substantial part of the changes in music expenditures within a household. This seasonality, therefore, is likely to bias within-sample estimators. Second, more than half of households in the sample skipped a few interviews. Third, the CEX provides different weights for different interviews of the same household. Fourth, a rotating structure in the CEX limits the number of complete panels, say, in 1999 because most of samples in 1999 also performed Interviews either in 1998 or 2000.

A.3: Recorded Music Expenditures and Internet Access in the CEX

Recorded music expenditures include expenses on two types of items: first, CDs, audio tapes, or records purchased other than through a mail-order club; second, CDs, tapes, videos, or records purchased from a mail-order club. I define the sum of both expenses as recorded music expenditures. Note, however, that the weighted sum of recorded music expenditures from the CEX is approximately 40% of the value of total record sales reported by the RIAA. The CEX tends to underestimate the total value of expenditures compared to the national accounts such as the *Personal Consumption Expenditure*, an aggregate time-series for U.S. consumer expenditures estimated by the Bureau of Economic Analysis. This has been noted by Battistin (2003) and references therein. One possibility of this underestimation is a recall problem. That is, survey respondents often forgot their purchases. The other is that the CEX surveys only households, so that expenditures from institutions including the government, businesses, libraries, and radio stations are not included in the CEX. Note that this problem of underestimated expenditures is not critical to my analysis, because this measurement error is unlikely to be correlated with Internet access or file sharing, and therefore, the DD approach can account for this problem. Moreover, the trend of record sales in the CEX are closely related to that from the RIAA.

An Internet user group is defined to be households that either spent positive amount on computer information service, or were living in a college dormitory. Computer information service in the CEX includes: DSL or ISDN; Internet access and data services; and Internet connection, other data service not reported. The CEX started to collect this expense from the first quarter of 1996. Consequently, I use the Interview survey from 1996 to 2002. From 2001 on, the CEX began to collect more detailed data on Internet services such as expenses on DSL or ISDN, but this information is not used in this paper. Defining Internet adoption by using Internet service fees, however, appears to underestimate the population rate of Internet adoption. To check this, I examine the *Current Population Survey supplements for Internet and computer use* (CPS) by the U.S. Bureau of Census. According to the CPS, 37.1% of households (not individuals) had home Internet access by “dial-up” telephone service, and 4.5% of households had access by high speed Internet access service in August 2000. However, if I define Internet access by positive amounts on Internet service fees, then the CPS reveals that 34.5% of households had Internet access. In contrast, the CEX reports that approximately 31% of households had Internet access around August 2000. Because I do not have further information on Internet access in the CEX, I cannot account for this problem.

Appendix B Details on the UCLAIS

The surveys for the UCLAIS were conducted during mid- to late-spring in 2000, May through July 2001, and April through June 2002. They were conducted under the name of the UCLA Internet Project and partially sponsored by the National Science Foundation. The UCLA Center for Communication Policy moved to the University of Southern California in August 2004. It is now called the USC Annenberg School Center for the Digital Future. In the UClAIS, households were randomly selected based on telephone numbers. One individual from the household was then randomly selected and interviewed by telephone. See the UCLA Internet Report (2000, 2001, 2003) for more detail on survey methodology and patterns in the data.

Table B.1 compares the UCLAIS with the CEX. It reports the mean of each variable, with standard deviations in parenthesis. The mean without a standard deviation is the proportion of the households in the sample. To compare the two data sets, I use the CEX for the period from June 1999 through May 2002, because the surveys for the UCLAIS were conducted in the middle of each year (2000, 2001, and 2002). Overall, the two data sets seem to be comparable, although there are some differences due to the following reasons. First, the UCLAIS interviewed an individual who was randomly selected from a household, whereas the CEX interviewed the head of the household who tends to be an adult. This explains the relatively lower age of reference persons in the UCLAIS than in the CEX. Note, however, that demographic information in the UCLAIS is not individual-level, but household-level. In addition, the UCLAIS appears to sample more households that are richer and more highly educated. As a result, the households in the UCLAIS are more likely to have computers and an Internet connection than those in the CEX. Second, the average income differs

Table B.1: Comparison of the CEX and the UCLAIS

Variable[a]	CEX		UCLAIS			CEX	UCLAIS
age	48.37	(17.52)	43.81	(18.37)	hw.child.in.sch	0.14	0.15
male	0.53		0.42		hw.child.af.sch	0.08	0.07
hsgrad	0.29		0.24		sp.child.bf.sch	0.03	0.03
lesscol	0.29		0.23		sp.child.in.sch	0.03	0.05
colgrad	0.26		0.37		headwrk[b]	0.71	0.62
income	$35,535	(42,832)	$47,446	(38,317)	retired	0.18	0.18
fam.size	2.58	(1.52)	2.72	(1.52)	col.student	0.01	0.03
persot64	0.32	(0.62)	0.28	(0.59)	computer	0.48	0.67
perslt18	0.64	(1.08)	0.80	(1.17)	internet.home	0.32	0.59
single	0.28		0.24		high.internet[c]	0.05	0.11
hw.young	0.03		0.04		dm[d]	n.a.	0.14
hw.old	0.17		0.17				
hw.child.bf.sch	0.05		0.06		observations	92,304	6,511

[a] See the Web Appendix for definition of variables.

[b] In the UCLAIS, headwrk = 1 if the reference person of a household is employed.

[c] For the CEX, I set high.internet equal to 1 if the household in the CEX spent more than $90 on computer information service, or was living in college dormitory.

[d] For the UCLAIS, dm is equal to 1 if the reference person spent positive hours on downloading music.

significantly between the CEX and the UCLAIS. In the CEX, nevertheless, the average income for households only with positive income are $42,312 (43,532). In the UCLAIS, income is reported as a dummy variable for different intervals. I use the lower bound of the interval to impute income for a household. Note also that the average income in the UCLAIS is computed excluding households that did not report income, which leaves 5,520 observations.

Appendix C Details on the Nonparametric DD Matching

C.1: Two Propensity Scores as Sufficient Statistics for Identification[38]

Lemma 1 *Consider the following two sets of assumptions:*

(A-1)′ $E(Y_{1bi}|D_i = 1, T_i = 1; X) = E(Y_{1bj}|D_j = 1, T_j = 0; X)$,

(A-2)′ $E(\delta_i|D_i = 1, T_i = 1; X) = E(Y_{0aj}|D_j = 0, T_j = 1; X) - E(Y_{0bj}|D_j = 0, T_j = 0; X)$,

(A-1) $E(Y_{1bi}|D_i = 1, T_i = 1; P_b, P_a) = E(Y_{1bj}|D_j = 1, T_j = 0; P_b, P_a)$,

(A-2) $E(\delta_i|D_i = 1, T_i = 1; P_b, P_a) = E(Y_{0aj}|D_j = 0, T_j = 1; P_b, P_a) - E(Y_{0bj}|D_j = 0, T_j = 0; P_b, P_a)$,

where $P_b = \Pr(D = 1|T = 0, X)$ *and* $P_a = \Pr(D = 1|T = 1, X)$. *Suppose that* $X \perp\!\!\!\perp T$ *holds, where* $\perp\!\!\!\perp$ *denotes independence. Then, (A-1)′ implies (A-1), and (A-2)′ implies (A-2).*

[38] I am grateful to Ed Vytlacil for helpful insights on the discussion in this appendix.

PROOF: Note first that $f(X|D=1,T=0;P_b,P_a) = f(X|T=0;P_b,P_a)$, since

$$\begin{aligned}
f(X|D=1,T=0;P_b,P_a) &= \frac{f(X,D=1|T=0;P_b,P_a)}{\Pr(D=1|T=0;P_b,P_a)} \\
&= \frac{\Pr(D=1|T=0;P_b,P_a,X)f(X|T=0;P_b,P_a)}{\Pr(D=1|T=0;P_b,P_a)} \\
&= \frac{P_b f(X|T=0;P_b,P_a)}{P_b} = f(X|T=0;P_b,P_a),
\end{aligned}$$

where $f(\cdot)$ denotes a density function. Similarly, $f(X|D=1,T=1;P_b,P_a) = f(X|T=1;P_b,P_a)$. Because $X \perp\!\!\!\perp T$, it follows that $f(X|T=0;P_b,P_a) = f(X|T=1;P_b,P_a)$. Hence,

$$f(X|P_b,P_a) = f(X|T=t;P_b,P_a) = f(X|D,T=t;P_b,P_a), \qquad \forall t=0,1.$$

Therefore,

$$\begin{aligned}
& E(Y_{1bi}|D_i=1,T_i=1;P_b,P_a) - E(Y_{1bj}|D_j=1,T_j=0;P_b,P_a) \\
= & \int E(Y_{1bi}|D_i=1,T_i=1;P_b,P_a,X)dF(X|D_i=1,T_i=1;P_b,P_a) \\
& -\int E(Y_{1bj}|D_j=1,T_j=0;P_b,P_a,X)dF(X|D_j=1,T_j=0;P_b,P_a) \\
= & \int \left[E(Y_{1bi}|D_i=1,T_i=1;X) - E(Y_{1bj}|D_j=1,T_j=0;X)\right] dF(X|P_b,P_a) = 0,
\end{aligned}$$

where the second equality follows from (A-1)$'$. Similarly,

$$\begin{aligned}
& E(\delta_i|D_i=1,T_i=1;P_b,P_a) - E(Y_{0aj}|D_j=0,T_j=1;P_b,P_a) + E(Y_{0bj}|D_j=0,T_j=0;P_b,P_a) \\
= & \int E(\delta_i|D_i=1,T_i=1;P_b,P_a,X)dF(X|D_i=1,T_i=1;P_b,P_a) \\
& -\int E(Y_{0aj}|D_j=0,T_j=1;P_b,P_a,X)dF(X|D_j=0,T_j=1;P_b,P_a) \\
& +\int E(Y_{0bj}|D_j=0,T_j=0;P_b,P_a,X)dF(X|D_j=0,T_j=0;P_b,P_a) \\
= & \int [E(\delta_i|D_i=1,T_i=1;X) - E(Y_{0aj}|D_j=0,T_j=1;X) \\
& +E(Y_{0bj}|D_j=0,T_j=0;X)]\, dF(X|P_b,P_a) = 0,
\end{aligned}$$

where the second equality follows from (A-2)$'$. ■

C.2: The DDM Estimator

The proposed DDM estimator is defined as

$$\begin{aligned}\widehat{M} &= \sum_{i\in I_{1a}} \Big[Y_{1ai} - \widehat{E}(Y_{1bj}|D_j = 1, T_j = 0; P_{bi}, P_{ai}) \\ &\qquad -\widehat{E}(Y_{0aj}|D_j = 0, T_j = 1; P_{bi}, P_{ai}) + \widehat{E}(Y_{0bj}|D_j = 0, T_j = 0; P_{bi}, P_{ai})\Big] \times w_i,\end{aligned}$$

where I_{1a} denotes the post-Napster Internet user group, and w_i is the CEX weight for each $i \in I_{1a}$ given by $\frac{(\text{CEX weight})_i}{\sum_{k\in I_{1a}} (\text{CEX weight})_k}$. $\widehat{E}(Y_{1bj}|D_j = 1, T_j = 0; P_{bi}, P_{ai})$, $\widehat{E}(Y_{0aj}|D_j = 0, T_j = 1; P_{bi}, P_{ai})$, and $\widehat{E}(Y_{0bj}|D_j = 0, T_j = 0; P_{bi}, P_{ai})$ are the conditional expectation estimators for each group conditional on P_{bi} and P_{ai} of each $i \in I_{1a}$. They are estimated by local linear regressions below.

$$\widehat{E}(Y_{dtj}|D_j, T_j; P_{bi}, P_{ai}) = e_1'(X_{P_i}'W_{P_i}X_{P_i})^{-1}X_{P_i}'W_{P_i}Y,$$

$$e_1 = (\begin{array}{ccc} 1 & 0 & 0 \end{array})', \quad Y = (Y_{dt1}, \cdots, Y_{dtJ})', \quad X_{P_i} = \begin{pmatrix} 1 & \cdots & 1 \\ (P_{b1} - P_{bi}) & \cdots & (P_{bJ} - P_{bi}) \\ (P_{a1} - P_{ai}) & \cdots & (P_{aJ} - P_{ai}) \end{pmatrix},$$

$$W_{P_i} = \text{diag}(G_{1i}, \cdots, G_{Ji}), \quad G_{ji} = K(\frac{P_{bj} - P_{bi}}{h}) \times K(\frac{P_{aj} - P_{ai}}{h}),$$

where h is a fixed bandwidth, and $K(\cdot)$ is a biweight kernel function as

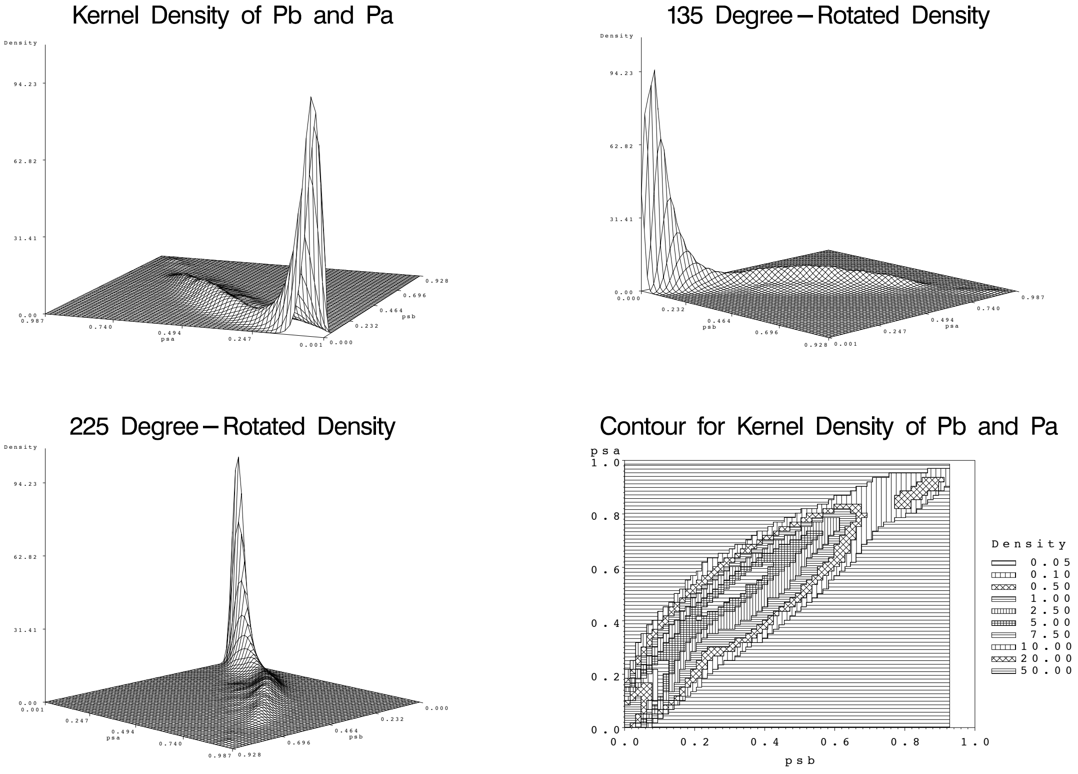
$$\begin{aligned} K(s) &= 15/16(s^2 - 1)^2 \qquad |s| < 1; \\ &= 0 \qquad\qquad\qquad\quad |s| \geq 1. \end{aligned}$$

I use 0.07 for the fixed bandwidth throughout the estimations. Results are comparable for other fixed bandwidths within ±0.02 of 0.07.

C.3: An Empirical Justification for the Use of Two Propensity Scores

To check possible sensitivity of the DDM method, I further experimented with other local averaging method such as kernel matching and did not find much difference in results. However, using a one-dimensional propensity score yields estimates that are larger than those using two-dimensional

Figure C.1: Estimated Density of P_b and P_a for All Samples in the CEX



propensity score in an absolute magnitude, suggesting that the one-dimensional propensity score is not sufficient to remove negative bias due to compositional changes. Using two propensity scores is further rationalized by density estimation of these propensity scores for all observations in the CEX. If one propensity score is sufficient, then most observations should be centered at the 45 degree line, where x-axis is P_b and y-axis is P_a. However, the plots in Figure C.1 show that most observations are located above the 45 degree line, indicating that consumers with the same characteristics are more likely to have an Internet connection in the post-Napster period than in the pre-Napster period. Note that the relationship between P_b and P_a is close to $P_a = g(P_b) + \epsilon$, where $g(\cdot)$ is a nonlinear function and ϵ is an error term. One could estimate this relationship and use one propensity score. However, I use both P_b and P_a, because using both does not complicate the estimation procedure.

C.4: Details on the Computation of the Standard Errors

I use a symmetric and compact kernel function in order to rely on the asymptotic distribution derived in Heckman, Ichimura, and Todd (1998). Nevertheless, it is difficult to derive a specific expression of asymptotic variance for the estimator as above. As a result, I use a bootstrap method to compute the standard errors of the DDM estimator. Heckman, et al. (1998) prove that a class of matching estimators, including the local linear matching estimator in this paper, is asymptotically normally distributed under regularity conditions, which assures that the bootstrap will lead to valid standard errors. There are two remarks on the actual computation of bootstrapped standard errors. First, I recompute propensity scores for each bootstrapped sample in order to account for the estimation error from the first stage Probit. Note that, in the DDM estimation, I use predicted propensity scores, instead of true propensity scores. Second, I resample bootstrap samples based on the CEX weights, in that the selection of observations is performed with probability proportional to the weights. Note that the CEX weights are the inverse of the probabilities of selection of the household in the original CEX samples. Therefore, bootstrap resampling based on the CEX weights ensures that oversampled observations in the original CEX samples (those with lower CEX weights) are relatively undersampled, and those undersampled in the original samples (those with higher CEX weights) are proportionally oversampled. By doing so, all observations are equally likely to be selected in bootstrap resamples. As a result, each bootstrap estimation of the DDM estimate is computed without using the CEX weight.

C.5: Probit Estimation Results for the Propensity Scores

	Before		After			Before		After	
Variable	Est.	S.E.	Est.	S.E.	Variable	Est.	S.E.	Est.	S.E.
intercept	-0.485	(0.157)	-0.864	(0.116)	hw.young	-0.028	(0.047)	-0.111	(0.038)
age	-0.014	(0.004)	0.025	(0.003)	hw.child.bf.sch	0.011	(0.055)	0.061	(0.042)
$(\text{age})^2$	0.000	(0.000)	0.000	(0.000)	hw.child.in.sch	0.162	(0.049)	0.127	(0.036)
white	-0.104	(0.094)	0.186	(0.067)	hw.child.af.sch	0.118	(0.045)	0.096	(0.032)
black	-0.388	(0.099)	-0.052	(0.070)	sp.child.bf.sch	-0.323	(0.077)	-0.078	(0.050)
male	0.086	(0.020)	0.095	(0.014)	sp.child.in.sch	0.029	(0.060)	0.045	(0.044)
hsgrad	0.379	(0.041)	0.317	(0.025)	retired	0.028	(0.049)	0.045	(0.033)
lesscol	0.612	(0.040)	0.545	(0.025)	headwrk	0.192	(0.061)	0.114	(0.045)
colgrad	0.705	(0.042)	0.616	(0.026)	spouwrk	-0.097	(0.046)	-0.058	(0.034)
tv	0.009	(0.008)	0.015	(0.006)	incwk1	-0.005	(0.001)	-0.002	(0.001)
computer	1.113	(0.021)	1.160	(0.016)	inchr1	-0.001	(0.001)	-0.002	(0.001)
soundcp	0.022	(0.023)	-0.028	(0.017)	incwk2	0.001	(0.001)	0.000	(0.001)
vcr	-0.344	(0.028)	-0.376	(0.021)	inchr2	0.000	(0.001)	0.002	(0.001)
vehq	0.021	(0.006)	0.031	(0.005)	owner	-0.958	(0.043)	-1.228	(0.050)
manager	0.183	(0.032)	0.129	(0.025)	renter	-1.225	(0.041)	-1.377	(0.049)
teacher	-0.021	(0.047)	-0.068	(0.036)	fincbtax	0.047	(0.004)	0.049	(0.003)
prof	0.186	(0.031)	0.108	(0.024)	$(\text{fincbtax})^2$	-0.001	(0.000)	-0.001	(0.000)
admin	0.108	(0.036)	0.090	(0.027)	ne	0.067	(0.026)	0.079	(0.020)
tech	0.210	(0.042)	0.112	(0.033)	mw	-0.020	(0.023)	0.026	(0.017)
sales	0.044	(0.036)	0.108	(0.027)	west	0.065	(0.022)	0.041	(0.017)
service	0.048	(0.037)	0.003	(0.026)	urban	0.243	(0.044)	0.097	(0.034)
fam.size	-0.016	(0.019)	-0.044	(0.012)	msa	-0.094	(0.055)	-0.088	(0.040)
no.ch.le11	-0.086	(0.024)	-0.016	(0.016)	ps4mil	0.129	(0.048)	0.103	(0.033)
no.ch.1217	-0.034	(0.022)	0.009	(0.015)	ps1mil	0.132	(0.048)	0.098	(0.033)
persot64	-0.087	(0.029)	0.000	(0.019)	ps330k	0.183	(0.050)	0.093	(0.035)
hw	0.037	(0.050)	-0.016	(0.036)	ps125k	-0.001	(0.050)	0.011	(0.035)
single	-0.041	(0.043)	-0.164	(0.029)	observations[a]	46,124		61,526	

[a] Beginning with the first quarter of 1999, the sample size in the CEX has increased by approximately 50 percent.

Appendix D Details on the Two-Sample IV

D.1: Details on the Computation of the Standard Errors

Both Angrist and Krueger (1992) and Arellano and Meghir (1992) derive asymptotic normal distributions for two-sample IV estimators under regularity conditions, which assures that a bootstrap method would lead to valid standard errors. One could make use of their results and derive a correct expression of asymptotic variance that accounts for the estimation errors in the first stage Probit. Nevertheless, I use the bootstrap method because of its computational convenience. Note that I resample both the CEX and the UCLAIS for each replication. There are similar remarks as in the computation of the DDM standard errors. First, I perform the Probit estimation of music downloading for each resampled UCLAIS data. In doing so, I account for the first stage estimation errors. Second, I resample CEX samples based on the CEX weights, and then compute each bootstrap estimation of the 2SIV estimates without using the CEX weights. However, I do not use any weight for the UCLAIS because weights are not provided for the UCLAIS data.

D.2: Probit Results for Music Downloading from the UCLAIS

	Age 15-34		Age 35-49		Age 50+		HHs w/children aged 6-17	
Variable	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
intercept	1.831	(1.444)	-0.795	(6.623)	-1.383	(3.917)	1.001	(0.749)
age	-0.208	(0.119)	-0.063	(0.317)	-0.045	(0.123)	-0.130	(0.034)
$(\text{age})^2$	0.003	(0.002)	0.001	(0.004)	0.000	(0.001)	0.001	(0.000)
male	0.328	(0.098)	0.127	(0.137)	0.330	(0.148)	0.027	(0.130)
hsgrad	-0.010	(0.195)	0.251	(0.407)	0.049	(0.398)	0.208	(0.249)
lesscol	0.084	(0.195)	0.349	(0.406)	0.008	(0.402)	0.122	(0.253)
colgrad	-0.022	(0.221)	0.247	(0.397)	-0.096	(0.402)	0.098	(0.263)
col.student	0.169	(0.150)						
fam.size	-0.004	(0.046)	0.016	(0.071)	0.013	(0.126)	-0.040	(0.104)
single	-0.074	(0.156)	0.208	(0.205)	0.119	(0.244)		
no.ch.le11							0.061	(0.124)
no.ch.1217							-0.061	(0.155)
employed	-0.153	(0.125)	-0.229	(0.210)	0.354	(0.184)	0.169	(0.159)
computer	0.155	(0.051)	0.127	(0.079)	0.102	(0.085)	0.052	(0.066)
internet	0.895	(0.138)	1.289	(0.250)	1.427	(0.273)	0.994	(0.189)
high.internet	0.489	(0.138)	-0.181	(0.198)	0.379	(0.213)	0.348	(0.175)
income	-0.030	(0.037)	-0.022	(0.056)	-0.074	(0.068)	-0.093	(0.048)
$(\text{income})^2$	0.000	(0.002)	-0.001	(0.003)	0.000	(0.004)	0.003	(0.003)
observations		1,312		787		1,500		912

Additional Web Appendix for Reviewers

(These are not intended for publication but will be available on the Web)

E. Illustrations of Compositional Changes in the CEX samples

The compositional change discussed in Section 4.1 is further illustrated in Figure E.1. The figure plots the percentages of households in the CEX by Internet adoption and whether they spent any money on recorded music. For example, area (3) denotes households with an Internet connection who spent nothing on recordings. More households adopted the Internet over time, so that area (2) and (3) become larger. However, the percentage of households who spent nothing on music (area (3) and (4)) has increased little, and more non-music buyers adopted the Internet over time. As a result, the post-Napster Internet user group includes more households with low reservation prices for recorded music than the pre-Napster user group does. This shows one reason why the decrease in the average music expenditure for the Internet user group may have nothing to do with Napster.

Figure E.1: % of Households in the CEX by Internet Adoption and Music Expenditure

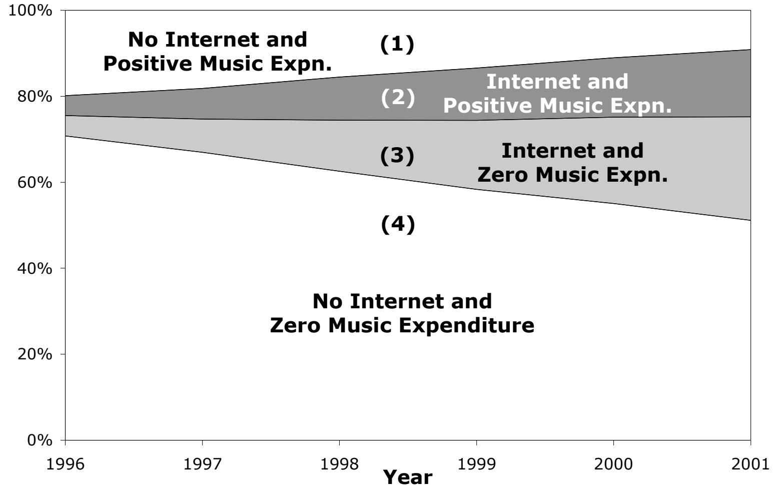
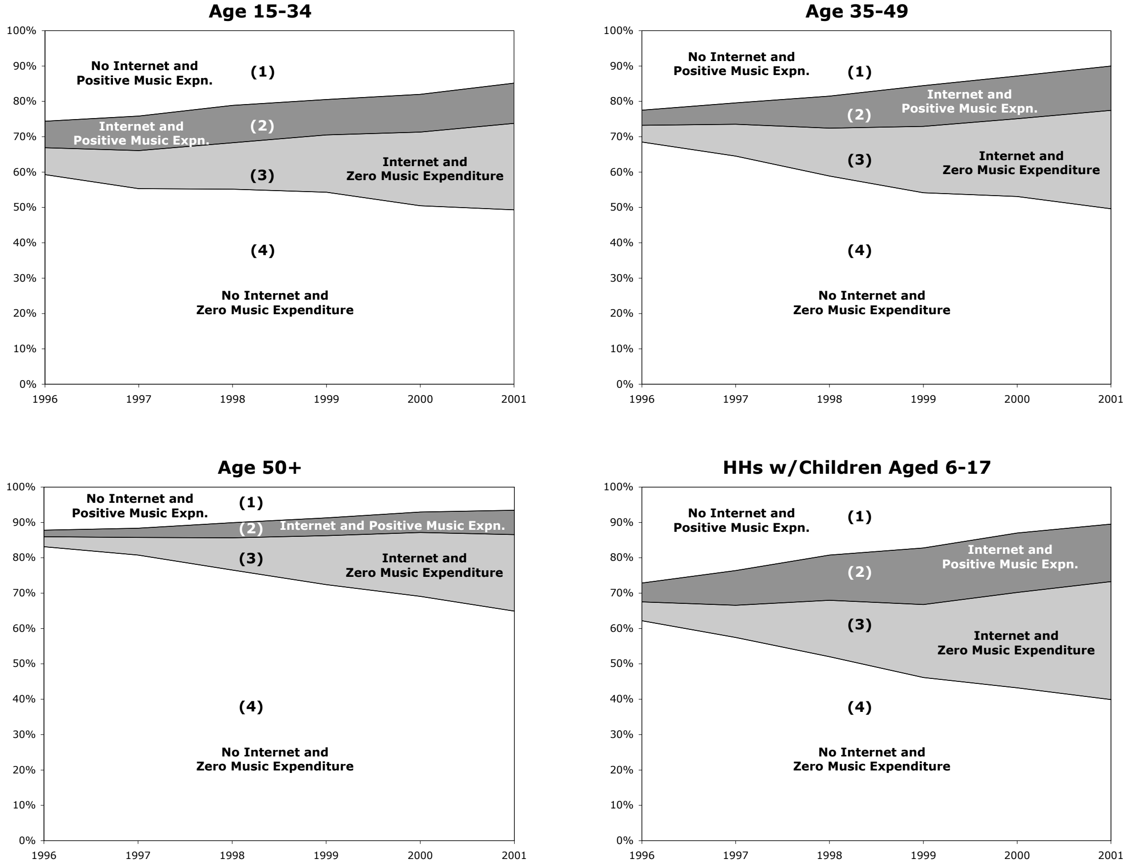


Figure E.2 plots similar percentages for different demographic groups. For those over 50 in particular, compositional changes seem to be substantial because the decline in the percentage of music non-buyers without Internet access (area (4)) is more severe than that of music buyers without Internet access (area (1)). Note that the percentage of music buyers has declined little (area (1) and (2)), while more music non-buyers seem to have adopted the Internet (area (3)). In contrast, such compositional changes appear to be less significant for households aged 15-34 because the decline in area (4) seems to be comparable to the decline in area (1). This also applies to those with children aged 6-17. Though a number of music non-buyers have adopted the Internet (from (4) to (3)), a similar number of music buyers have adopted the Internet as well (from (1) to (2)). Note also that the percentage of music buyers (area (1) and (2)) declined after 1999, suggesting that Napster may be responsible for decreases in music expenditure for this demographic group.

Figure E.2: % of HHs in the CEX by Internet and Music Expenditure by Age Family Groups



F. The Difference-in-Differences Estimates

For the purpose of comparison, I report the results from the conventional DD method. Table F.1 reports baseline DD estimates. They are ordinary least squares[39] estimates, where *Napster* is a dummy for the Napster period, and *Internet* indicates the adoption of the Internet. The coefficient for Internet×Napster implies the effect of the presence of Napster. The coefficient estimates in both columns 1 and 2 indicate a significant negative impact of Napster on recording expenditure. As explained in Section 5, however, these DD estimates are likely to be negatively biased because of compositional changes. An ad hoc approach to account for the problem is to estimate propensity score weighted regressions using the probability of Internet adoption given demographic variables. The results of these regressions are reported in columns 3 and 4 in Table F.1. The estimated coefficients are negative, but their magnitudes are smaller than those of the simple DD. This suggests that proper weighting might help reducing the negative bias.

[39] Though I call 'ordinary least squares' or 'not weighted' regressions, I actually estimate weighted least squares regressions using the CEX weight assigned to each observation. I use these terms in order to distinguish weighted regressions using only the CEX weights from weighted regressions using both propensity scores and the CEX weights.

Table F.1: Baseline and Alternative DD Estimates[a]

	Not Weighted		P_b-Weighted[b]	
	No Control	Control[c]	No Control	Control
	(1)	(2)	(3)	(4)
Internet×Napster	-4.589 (0.552)	-3.307 (0.540)	-2.892 (0.592)	-2.177 (0.581)
Napster	-1.749 (0.253)	-1.366 (0.248)	-3.252 (0.382)	-3.180 (0.376)
Internet	14.781 (0.432)	6.475 (0.444)	11.543 (0.440)	6.266 (0.441)

[a] Standard errors in parentheses. The dependent variable is quarterly recorded muisc expenditure in 1998 dollar. Samples in the pre- and post-Napster periods are used. The number of observations is 107,650. Note that observations whose quarterly expenditures occurred in April-June or May-July 1999 are excluded, since these periods include both pre- and post-Napster periods.

[b] Weighted least squares using a propensity score of Internet adoption for the pre-Napster period.

[c] Controls include age, education, income, appliance, occupation, family composition, and region.

G. Nonparametric Bounds

G.1: Empirical Framework

I estimate nonparametric bounds of the effect of actual downloading on recorded music expenditure, following the data combination approach developed in Cross and Manski (2002). The UCLAIS provides a marginal distribution of music downloading conditional on common demographic variables, whereas the CEX provides a marginal distribution of music expenditures conditional on the same variables. As Moffitt and Ridder (2003) point out, all that can be learned from two marginal distributions without additional assumptions is the Fréchet bound on the joint distribution, from which bounds on the conditional expectation of music expenditures given music downloading can be estimated. This idea of data combination is further developed in Cross and Manski (2002).

To explain the idea, I decompose the effect of the presence of Napster on recorded music expenditure, using the law of iterated expectation as follows.

$$\begin{aligned} M &= E_Z\left[E(Y_{1ai} - Y_{1bi} - \delta_i | D_i = 1, T_i = 1, Z_i)\right] \\ &= E_Z\left[\sum_{j=0,1} E(Y_{1ai} - Y_{1bi} - \delta_i | D_i = 1, T_i = 1, Z_i, DM_i = j) \times \Pr(DM_i = j | D_i = 1, T_i = 1, Z_i)\right], \end{aligned}$$

where Z denotes common variables, DM indicates a dummy variable for downloading music, and D is a dummy for Internet adoption. Because $Y_{1bi} + \delta_i$, the music expenditure in the absence of Napster, would be the same for either DM equal to 0 or 1, the equation above is rewritten as

$$M = E_Z\left[\sum_{j=0,1} \{E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = j) - E(Y_{1bi} + \delta_i | D_i = 1, T_i = 1, Z_i)\} \times \Pr(DM_i = j | D_i = 1, Z_i)\right]. \quad (8)$$

Note that $E(Y_{1bi} + \delta_i | D_i = 1, T_i = 1, Z_i)$ is the counterfactual estimated by the DDM method for a given Z. In practice, this term cannot be computed if Z is a continuous vector, since there are few observations for a fixed Z. For this and other reasons below, I consider a discrete case for Z in the actual estimation of the bounds. That is, Z denotes different age and family groups.

The first term in equation (8) requires data combination because Y_{1ai} and DM_i are not observed together in the same data. Using the results in Cross and Manski (2002), I obtain a bound of $E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = j)$ for fixed z as follows.

$$\left[E\left(Y_{1ai} \mid D_i = 1, T_i = 1, Z_i = z, Y_{1ai} < F^{-1}(\pi_{z1})\right), \quad E\left(Y_{1ai} \mid D_i = 1, T_i = 1, Z_i = z, Y_{1ai} \geq F^{-1}(\pi_{z0})\right)\right], \tag{9}$$

where $\pi_{zj} = \Pr(DM_i = j|D_i = 1, T_i = 1, Z_i = z)$, and $F(\cdot)$ is a cumulative distribution function of Y_{1ai}. $F^{-1}(\pi_{z1})$ is thus the π_{z1}-th quantile of the distribution of Y_{1ai}. The expectation is taken with respect to L, the right-truncated (or U, the left-truncated) distribution for the lower (or upper) bound. They are defined by

$$L = \begin{cases} \frac{\Pr[Y_{1ai} \leq y|D_i=1,T_i=1,Z_i=z]}{\pi_{z1}} & \text{if} \quad y < F^{-1}(\pi_{z1}) \\ 1 & \text{if} \quad y \geq F^{-1}(\pi_{z1}) \end{cases}, \quad U = \begin{cases} 0 & \text{if} \quad y < F^{-1}(\pi_{z0}) \\ \frac{\Pr[Y_{1ai} \leq y|D_i=1,T_i=1,Z_i=z]-(1-\pi_{z1})}{\pi_{z1}} & \text{if} \quad y \geq F^{-1}(\pi_{z0}) \end{cases}.$$

The basic intuition for the bound is as follows. Consider a specific demographic group $Z = z$. Suppose that 20% of Internet users in this group actually downloaded music, which is computed from the UCLAIS. Next, order Internet users for this group in the CEX by their recorded music expenditure, that is, from households who spent zero on recordings to those whose music expenditure is the largest. The lower bound of $E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = 1)$ is then attained when only the lowest 20% of Internet users for this group in the CEX downloaded music. Likewise, the upper bound is attained when only the highest 20% downloaded music. Similarly, one can compute bounds on $E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = 0)$ as well.

I compute the bound in (9) for each demographic group. Specifically, I first use the UCLAIS and estimate $\widehat{\pi_{z1}}$. I then estimate the $\widehat{\pi_{z1}}$-th quantile and the $\widehat{\pi_{z0}}$-th quantile of Y_a^1 for the group in the CEX. Because each group contains a considerable number of observations, the conditional expectation given in (9) can be computed by taking a mean for observations with recorded music expenditure below (or above) the quantile. If Z is a continuous vector, however, it is not feasible to compute the conditional expectation. This problem still remains if Z is a discrete vector but defined finely. Consequently, I consider discrete cases of Z such as age and family groups.

Given these bounds, one can decompose the effect of the presence of Napster into the effect of actual downloading on music expenditure, denoted by θ_1, and the effect of other new online activities during the Napster period, denoted by θ_0. Specifically, the bounds on θ_0 are obtained by subtracting $E(Y_{1bi}+\delta_i|D_i = 1, T_i = 1, Z_i = z)$ from the bounds on $E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = 0)$. Similar bounds for $DM = 1$, however, entail not only θ_1 but also θ_0, since θ_0 is the common effect for all Internet users during the Napster period, whether or not they downloaded music. Therefore, bounds on θ_1 can be obtained by subtracting bounds on $E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = 0)$ from bounds on $E(Y_{1ai}|D_i = 1, T_i = 1, Z_i, DM_i = 1)$.

G.2: Estimation Results

I use the 2000-2001 UCLAIS because each survey was conducted around June of the year, so that both surveys cover the Napster period. Rows 1 and 2 of Table G.1 report the empirical probabilities of music downloading DM for Internet users in each age and family group. Using these probabilities, I then estimate bounds on average recorded music expenditure for Internet users during the Napster period given DM for each group. These bounds are reported in rows 3 and 4. As mentioned before, the bounds tend to be too wide when the probability of DM is low. This explains the wide bounds for average music expenditure given $DM = 1$. This further explains relatively tight bounds for

Table G.1: Nonparametric Bounds Estimates on θ_1 and ${\theta_0}^a$: [*lower bound, upper bound*]

	Age 15-34	Age 35-49	Age 50+	HHs w/children Aged 6-17
	(1)	(2)	(3)	(4)
$\Pr(DM = 1 \vert D = 1, Z = z)^b$	0.409	0.161	0.093	0.173
$\Pr(DM = 0 \vert D = 1, Z = z)$	0.591	0.839	0.907	0.827
Bounds on $E(Y_{1ai} \vert D_i = 1, T_i = 1, Z_i, DM_i = 1)^c$	[0.000, 19.806]	[0.000, 103.009]	[0.000, 103.884]	[0.000, 96.103]
Bounds on $E(Y_{1ai} \vert D_i = 1, T_i = 1, Z_i, DM_i = 0)$	[0.000, 19.806]	[6.488, 21.876]	[4.762, 13.468]	[7.933, 22.625]
$E(Y_{1bi} + \delta_i \vert D_i = 1, T_i = 1, Z_i = z)$ from the DDM[d]	22.715	22.355	13.843	25.974
Bounds on ${\theta_1}^e$	[-19.806, 19.806]	[-21.876, 96.521]	[-13.468, 99.122]	[-22.624, 88.170]
Bounds on ${\theta_0}^f$	[-22.715, -2.909]	[-15.867, -0.479]	[-9.081, -0.375]	[-18.042, -3.350]

[a] The number of observations for each group in the UCLAIS (or the CEX) is 1,312 (or 22,943) for Age 15-34, 787 (or 21,614) for Age 35-49, 1,500 (or 44,007) for Age 50+, and 912 (19,086) for Family w/Child 6-17.

[b] The probability of DM computed from the UCLAIS. DM denotes a dummy variable for downloading music, D is a dummy for Internet adoption, and Z indicates age and family groups.

[c] The bounds defined in equation (5) are estimated using the CEX data and the probabilities of downloading from the UCLAIS.

[d] Counterfactual average recorded music expenditure of Internet users for the age group in the absence of Napster.

[e] The effect of actual music downloading.

[f] The effect of other new online activities during the Napster period.

those aged 15-34 and for $DM = 0$, since the probabilities of DM is relatively high. For those aged 15-34 in particular, estimated bounds for $DM = 1$ and $DM = 0$ are identical, but this is simply because the probabilities of DM are similar.

Given these bounds, I then decompose the effect of the presence of Napster into the effect of actual downloading on music expenditure, denoted by θ_1, and the effect of other new online activities during the Napster period, denoted by θ_0. Bounds on θ_0 and θ_1 are presented respectively in rows 6 and 7 of Table G.1. I do not compute standard errors for bounds, so that it is unclear how precisely these bounds are estimated. Nonetheless, it is evident that the estimated bounds are too wide to provide any useful information. However, bounds on θ_0, which are relatively tight, imply that the effect of other new online activities could be significantly negative.

H. Testing Assumptions in the DD Regressions

H.1: Additional Assumptions in the DD Regressions

The identification of the main parameter of interest in the DD regressions requires not only (A-1)$'$ and (A-2)$'$ in Section 5.3, but also parametric assumptions on conditional expectation of Y_{0bi} and constant effects. To see this, rewrite equation (2) in Section 5.1 as

$$\begin{aligned} Y_i = D_iT_i\{E(\theta_i|D_i=1,T_i=1)+\eta_\theta(X_i)\} + D_i\{E(\gamma_i|D_i=1,T_i=0)+\eta_\gamma(X_i)\} \\ + T_i\{E(\delta_i|D_i=0,T_i=1)+\eta_\delta(X_i)\} + E(Y_{0bi}|X_i)+\nu_i, \qquad (10) \end{aligned}$$

$$\begin{aligned} \text{where} \quad \nu_i &\equiv D_iT_i\{\theta_i - E(\theta_i|D_i=1,T_i=1,X_i)\} + D_i\{\gamma_i - E(\gamma_i|D_i=1,T_i=0,X_i)\} \\ &\quad + T_i\{\delta_i - E(\delta_i|D_i=0,T_i=1,X_i)\} + \{Y_{0bi} - E(Y_{0bi}|X_i)\}, \quad \text{and} \\ \eta_\theta(X_i) &\equiv E(\theta_i|D_i=1,T_i=1,X_i) - E(\theta_i|D_i=1,T_i=1), \\ \eta_\gamma(X_i) &\equiv E(\gamma_i|D_i=1,T_i=0,X_i) - E(\gamma_i|D_i=1,T_i=0), \\ \eta_\delta(X_i) &\equiv E(\delta_i|D_i=0,T_i=1,X_i) - E(\delta_i|D_i=0,T_i=1). \end{aligned}$$

It can be shown that if the assumptions (A-1)$'$ and (A-2)$'$ hold and $E(Y_{0bi}|D_i,T_i,X_i) = E(Y_{0bi}|X_i)$, then $E(\nu_i|D_i,T_i,X_i) = 0$.[40] Therefore, an additional identifying assumption for the DD regressions is that conditional on X_i, Y_{0bi} is mean independent of Internet access over time. In particular, the standard DD regressions assume (A-3): $E(Y_{0bi}|X_i) = \alpha + X_i\beta$.

The assumptions (A-1)$'$, (A-2)$'$, and (A-3), nevertheless, are not sufficient to identify $E(\theta_i|D_i = 1, T_i = 1)$. Notice that the coefficients on D_i, T_i, and D_iT_i in equation (9) are conditional on X_i. To see the problem clearly, rewrite equation (9) as

$$Y_i = Z_i'(\mu + \eta_i) + \nu_i,$$

where Z_i is a vector defined as $Z_i \equiv [D_iT_i, D_i, T_i, X_i]$, μ is a vector that includes the main parameter of interest and is defined as $\mu \equiv [E(\theta_i|D_i = 1, T_i = 1), E(\gamma_i|D_i = 1, T_i = 0), E(\delta_i|D_i = 0, T_i = 1), \beta']$, and η_i is a vector defined as $\eta_i \equiv [\eta_\theta(X_i), \eta_\gamma(X_i), \eta_\delta(X_i), 0']$. Ordinary least squares then

[40] After some algebra, one can show that (A-1)$'$ and (A-2)$'$ imply $E(\nu_i|D_i = 1, T_i = 1, X_i) = E(Y_{0bi}|D_i = 1, T_i = 0)+E(Y_{0bi}|D_i = 0, T_i = 1, X_i)-E(Y_{0bi}|D_i = 0, T_i = 0, X_i)-E(Y_{0bi}|X_i)$. Therefore, if $E(Y_{0bi}|D_i, T_i, X_i) = E(Y_{0bi}|X_i)$, then $E(\nu_i|D_i = 1, T_i = 1, X_i) = 0$. It is straightforward to show that this assumption also implies $E(\nu_i|D_i, T_i, X_i) = 0$ for $(D_i, T_i) = (1, 0)$, $(0, 1)$, or $(0, 0)$.

yield the following estimator for μ.

$$\begin{aligned}\widehat{\mu} &= \left[\sum_{j=1}^{N} Z_j Z_j'\right]^{-1} \sum_{i=1}^{N} Z_i Y_i = \left[\sum_{j=1}^{N} Z_j Z_j'\right]^{-1} \sum_{i=1}^{N} \left\{Z_i Z_i'(\mu + \eta_i) + Z_i \nu_i\right\} \\ &= \mu + \left[\sum_{j=1}^{N} Z_j Z_j'\right]^{-1} \sum_{i=1}^{N} Z_i Z_i' \eta_i + \left[\sum_{j=1}^{N} Z_j Z_j'\right]^{-1} \sum_{i=1}^{N} Z_i \nu_i. \qquad (11)\end{aligned}$$

In order to estimate μ consistently, the second and third terms in equation (11) should converge to zero as the sample size grows. The third term reflects selection bias, but it converges to zero because the preceding assumptions of selection on observables imply $E(\nu_i|Z_i) = 0$. However, the second term is unlikely to converge to zero because $E(Z_i Z_i' \eta_i)$ is not equal to zero in general. One could assume constant effects by setting $\eta_\theta(X_i) = \eta_\gamma(X_i) = \eta_\delta(X_i) = 0$, so that the second term is set to be zero. This is equivalent to the assumption (A-4) in Section 5.4. Nonetheless, it might be too strong to assume only constant effects. For example, a young music buyer is unlikely to have experienced the same effect of Internet or Napster on her music expenditure as an old music non-buyer. As a consequence, identification of the main parameter of interest requires restricting heterogeneous effects, either by assuming constant effects for different age groups, or by imposing parametric assumptions for heterogeneous effects. Possible parametric assumptions include $\eta_\delta(X_i) = X_i \beta_\delta$, and $\eta_\theta(X_i) = \widehat{\Pr}(D_i = 1|X_i)$.

One might argue that the DD regressions also identify $E(\theta_i|D_i = 1, T_i = 1)$ without restricting heterogeneous effects, and that the difference between the DDM and the DD regressions is essentially the weighting scheme used to pool $E(\theta_i|D_i = 1, T_i = 1, X_i)$. This appears to be true because the first element of $\mu + \eta_i$ in the preceding equation is $E(\theta_i|D_i = 1, T_i = 1, X_i)$ and the DD regressions pool these heterogeneous effects by using $\left[\sum_{j=1}^{N} Z_j Z_j'\right]^{-1} Z_i Z_i'$ as weights, so that the first element of $\widehat{\mu}$, which is the regression coefficient for $D_i T_i$, appears to be a weighted mean of $E(\theta_i|D_i = 1, T_i = 1, X_i)$.

The distinction between the DDM and the DD regressions, however, is not simply different weighting scheme. Note that $\mu + \eta_i$ includes not only $E(\theta_i|D_i = 1, T_i = 1, X_i)$ but also $E(\gamma_i|D_i = 1, T_i = 0, X_i)$ and $E(\delta_i|D_i = 0, T_i = 1, X_i)$. As a result, the first element of $\left[\sum_{j=1}^{N} Z_j Z_j'\right]^{-1} Z_i Z_i'(\mu + \eta_i)$ is a combination of not only heterogeneous effects of the presence of Napster, but also diverse effects of the Internet in general as well as heterogeneous time effects. Without restricting heterogeneous effects, therefore, the DD regressions are unlikely to isolate the effect of the presence of Napster from other confounding effects, and so the main parameter of interest is not identified.

H.2: Testing Assumptions in the DD Regressions

The DD regression assumes (A-1)$'$, (A-2)$'$, (A-3), and (A-4). The assumptions (A-1)$'$, (A-2)$'$, and (A-3) are variants of selection on observables and account for selection bias represented by the third term in equation (11). The assumption (A-4) restricts heterogeneous effects and thus assumes away a bias denoted by the second term in (11).

How likely are these assumptions to hold in the CEX data? To answer this question, one needs to test the validity of these conditions in the CEX samples. Though it is difficult to test them directly, an indirect way of testing is possible. The idea is based on equation (11). Note that if the preceding assumptions hold, the second and third terms in equation (11) converge to zero.

Table H.1: “Pre-treatment” Test for Age Groups: A Test of Assumptions in the DD Regressions]“Pre-treatment” Test for Age Groups: A Test of Assumptions in the DD Regressions[a]

	Age 15-34		Age 35-49		Age 50+		HHs w/children Aged 6-17	
	Control	Control w/ Year×X_i	Control	Control w/ Year×X_i	Control	Control w/ Year×X_i	Control	Control w/ Year×X_i
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Not Weighted								
Internet×1997	0.75 (1.98)	0.76 (2.15)	-5.17 (2.41)	-6.04 (2.59)	-6.97 (1.59)	-8.11 (1.75)	-0.79 (2.15)	0.59 (2.33)
Internet×1998	0.25 (1.93)	0.57 (2.11)	-1.93 (2.27)	-2.31 (2.45)	-5.67 (1.46)	-5.34 (1.61)	-3.29 (2.05)	-1.50 (2.24)
Internet×1999	-0.99 (1.92)	2.19 (2.08)	-5.68 (2.20)	-4.13 (2.36)	-8.35 (1.41)	-7.69 (1.56)	-3.92 (2.03)	-0.88 (2.21)
Internet×2000	-4.22 (1.88)	-2.55 (2.03)	-3.69 (2.19)	-3.10 (2.35)	-10.56 (1.38)	-8.72 (1.52)	-6.52 (2.01)	-3.87 (2.17)
Internet×2001	-4.16 (1.85)	-1.17 (2.00)	-6.65 (2.16)	-4.39 (2.31)	-9.13 (1.37)	-7.35 (1.51)	-7.38 (2.00)	-3.81 (2.16)
B. P_b-Weighted[b]								
Internet×1997	-0.08 (2.14)	-0.22 (2.32)	-2.33 (2.40)	-5.31 (2.51)	-5.22 (1.53)	-5.69 (1.64)	0.54 (2.16)	0.57 (2.29)
Internet×1998	0.89 (2.01)	1.58 (2.20)	-0.99 (2.30)	-2.75 (2.39)	-1.82 (1.44)	-4.34 (1.53)	-1.22 (2.11)	-2.09 (2.24)
Internet×1999	-1.08 (1.96)	2.13 (2.12)	-0.54 (2.27)	-0.59 (2.35)	-3.94 (1.42)	-5.63 (1.52)	-1.89 (2.09)	-1.23 (2.22)
Internet×2000	-4.31 (1.94)	-2.27 (2.09)	-0.79 (2.28)	-2.28 (2.36)	-8.05 (1.39)	-8.54 (1.47)	-5.17 (2.09)	-4.81 (2.20)
Internet×2001	-4.57 (1.91)	-1.23 (2.07)	-3.49 (2.24)	-2.90 (2.31)	-5.91 (1.38)	-6.39 (1.46)	-5.99 (2.07)	-4.46 (2.16)

[a] Standard errors are reported in parentheses. The dependent variable is recorded music expenditure. All samples in the periods from year 1996 to year 2001 are used. Years refer to June of the year to May of the next year. The table reports Internet×year interactions in regressions that include year and Internet dummies, with 1996 as the base period. The regressions with controls include covariates such as age, income, education, appliance ownership, family composition, and region. The regressions with controls w/year×X_i include the same covariates as well as year×covariates.

[b] Weighted least squares using a propensity score of Internet adoption for the pre-Napster period.

Moreover, if I compare samples in two periods before the introduction of Napster, say, year 1997 and 1998, the effect of the presence of Napster should be zero in principle. Accordingly, under the given assumptions, a regression of the previous equation using pre-samples should yield a zero coefficient for D_iT_i, where $T_i = 1$ if i is observed in year 1998 and $T_i = 0$ if observed in 1997. Similar test can be performed for different periods before the introduction of Napster. Specifically, I compare 1996 (the base period) with 1997 as well as with 1998. If (A-1)$'$, (A-2)$'$, (A-3), and (A-4) hold, the coefficient estimates for Internet×1997 and Internet×1998 should be statistically indistinguishable from zero. Even though it is possible that these coefficient estimates can be close to zero while some assumptions do not hold, this test is useful because it provides necessary evidence for the validity of the DD assumptions.

Table H.1 presents the results from the proposed tests for different age groups. It reports the coefficient estimates of Internet×year in regressions that include year and Internet dummies with 1996 as the base period. Controls include age, income, education, appliance ownership, family composition, and region. The regressions with controls w/year$\times X_i$ include the same covariates as well as year×covariates. Note that adding $T_i \times X_i$ is equivalent to assuming $\eta_\delta(X_i) = X_i\beta_\delta$, which is an ad hoc approach to allow for heterogeneous time effects. Panel A shows the results from 'not weighted' regressions in that they are not weighted by propensity scores but weighted by the CEX weights. Panel B reports the results from 'weighted' least squares weighted by both propensity scores and the CEX weights.

For households with heads aged 15-34 and those with children aged 6-17, I do not find statistically significant evidence against underlying assumptions in the DD regressions. For the other two demographic groups, however, I find statistically significant evidence against underlying assumptions in that at least one of coefficients for Internet×1997 and Internet×1998 is significantly different from zero in most specifications. Although both coefficients are statistically insignificant in the P_b-weighted regression for households aged 35-49, this does not provide evidence for the validity of underlying assumptions in this demographic group, since if all three assumptions in the DD regressions hold, both coefficients should be zero regardless of different weighting schemes.

The results in Table H.1 suggest that the DD regressions are likely to identify the main parameter of interest for households aged 15-34 and those with children aged 6-17, whereas they are unlikely to do so for the other two demographic groups. For the former two groups, therefore, it is valid to decompose the effect of the presence of Napster based on the DD regressions. In other words, the 2SIV results can be interpreted as decomposing the DDM estimates of the main parameter of interest. For the latter two groups, however, the 2SIV is not likely to be a valid approach for the decomposition, so that it would be difficult to connect the 2SIV results with the DDM results. To use the 2SIV for households aged 35-49 and over 50, I need to relax the assumptions (A-3) and (A-4) in the DD regressions. Possible approaches include assuming parametric functional forms for $\eta_\theta(X_i)$, $\eta_\gamma(X_i)$, and $\eta_\delta(X_i)$, as well as using nonparametric estimation of conditional mean function of Y_{0bi}, instead of assuming $E(Y_{0bi}|D_i, T_i, X_i) = \alpha + X_i\beta$. Nevertheless, I do not further investigate these modifications of the 2SIV for households aged 35-49 and over 50, mainly because the DDM estimates for these demographic groups are small and statistically insignificant.

I. 2SIV Results for Alternative Specifications

Table I.1 presents the coefficient estimates of θ_0 and θ_1 from the 2SIV method for alternative specifications. I consider four specifications, each of which corresponds to each DD regression examined in the Web Appendix H.

Table I.1: 2SIV Estimates for Age Groups with Alternative Specification[a]

		Age 15-34	Age 35-49	Age 50+	HHs w/children Aged 6-17
		(1)	(2)	(3)	(4)
			Not weighted		
Control	θ_1[b]	-2.72 (2.08)	13.05 (12.29)	2.33 (4.59)	-22.51 (6.89)
	θ_0[c]	-2.43 (0.95)	-4.16 (1.95)	-3.76 (0.80)	-0.12 (1.09)
Control w/ $T_i \times X_i$	θ_1	-0.04 (2.54)	7.57 (13.86)	-4.18 (4.98)	-22.50 (6.97)
	θ_0	-0.63 (1.07)	-1.62 (2.28)	-1.13 (0.86)	1.66 (1.18)
			P_b-weighted[d]		
Control	θ_1	-4.16 (2.73)	11.37 (13.89)	8.53 (8.44)	-25.48 (10.37)
	θ_0	-0.06 (1.36)	-1.30 (2.49)	-4.18 (1.40)	0.18 (1.52)
Control w/ $T_i \times X_i$	θ_1	1.51 (2.97)	-16.31 (16.91)	-10.36 (8.91)	-27.94 (11.86)
	θ_0	0.56 (1.41)	4.81 (2.89)	-0.79 (1.42)	1.95 (1.76)

[a] Bootstrapped standard errors are in parentheses. They are based on 500 replications with 80% sampling, described in Appendix D. The dependent variable is recorded music expenditure in 1998 dollar. The CEX samples in the pre- and post-Napster periods are used. The table reports coefficient estimates for D_iT_i (i.e. θ_0) and $D_iT_iDM_i$ (i.e. θ_1) in regressions that include age, income, education, appliance ownership, occupation, family composition, and region. DM is the imputed variable from the UCLAIS. It is equal to 1 if the households spent positive hours on music downloading. T_i is a dummy for the post-Napster period, and D_i indicates the adoption of the Internet. All regressions exclude a dummy for high speed Internet access, which is defined to be 1 if living in college dormitory, or computer information service $\geq$ \$96.58 for Age 15-34, \$92.76 for Age 35-49, \$85.53 for Age 50+, and \$87.11 for Family w/children aged 6-17. For each group, I compute the cutoffs using the CPS. These cutoffs are 3×average monthly Internet service fees that households with high speed Internet access paid.

[b] θ_1 is the effect of actual downloading.

[c] θ_0 is the effect of other new online activities during the Napster period.

[d] Weighted least squares using a propensity score of Internet adoption for the pre-Napster period. The propensity score is estimated separately for each group.

For most demographic groups, the coefficient θ_1 is not estimated precisely. For households with children aged 6-17, however, θ_1 is precisely estimated and is approximately -\$23. Though its magnitude changes slightly depending on specifications, the estimated effect of downloading for this demographic group remains to be significantly negative. For this same group, the magnitude of estimated θ_0 is small and statisticallhy insignificant. These results suggest that actual downloading was a significant factor in the decline in music expenditures for these households. By constrast, the estimated θ_1 for households aged 15-34 is fairly small and statistically insignificant, indicating that the effect of music downloading may not be a significant factor in decreases in recorded music expenditure. As for households aged 35-49 and over 50, the estimates of θ_1 seem to be sensitive to changes in specifications, but they are statistically insignificant and often positive, implying that downloading is unlikely to be a significant factor in the decline in music expenditures.

J. Definition of Variables

AGE	age of reference person in the household
WHITE	1 if the reference person is white
BLACK	1 if black
MALE	1 if male
HSGRAD	1 if highest education is high school graduate
LESSCOL	1 if some college, less than college
COLGRAD	1 if college graduate
MANAGER	1 if the job best fits the category of administrator, manager
TEACHER	1 if the job best fits the category of teacher
PROF	1 if the job best fits the category of professional
ADMIN	1 if the job best fits the category of administrative support, including clerical
SALES	1 if the job best fits the category of sales, retail
TECH	1 if the job best fits the category of technician
SERVICE	1 if the job best fits the category of service
FAM.SIZE	the number of members in the household
PERSOT64	the number of persons older than 64
NO.CH.LE11	the number of children younger than 11
NO.CH.1217	the number of children ages between 12 and 17
HW	1 if CU is family with husband and wife only
SINGLE	1 if CU is single
HW.YOUNG	1 if family only with husband and wife, and AGE under 40
HW.OLD	1 if family only with husband and wife, and AGE over 45
HW.CHILD.BF.SCH	1 if husband and wife with children before school
HW.CHILD.IN.SCH	1 if husband and wife with children in school
HW.CHILD.AF.SCH	1 if husband and wife with children after school
SP.CHILD.BF.SCH	1 if CU is single parent with children before school
SP.CHILD.IN.SCH	1 if single parent with children in school
COL.STUDENT	1 if reference person is attending college
RETIRED	1 if CU is retired
HEADWRK	1 if the head of the household is working
SPOUWRK	1 if the spouse of the household is working
EMPLOYED	1 if either the head or the spouse is working
INCWK1	the number of weeks in a year that head worked
INCWK2	the number of weeks in a year that spouse worked
INCHR1	the number of hours in a week that head worked
INCHR2	the number of hours in a week that spouse worked
FINCBTAX	Real final income before tax (in $10,000)
OWNER	1 if CU owns house
RENTER	1 if CU rents house
COLDORMI	1 if CU is living in college dormitory
NE	1 if household resides in Northwest Census region
MW	1 if household resides in Midwest Census region
WEST	1 if household resides in Census Western region
URBAN	1 if household resides in urban area
MSA	1 if household resides in Metropolitan Statistical Area
PS4MIL	1 if household resides in area with population size over 4 million
PS1MIL	1 if population size between 1.2 million and 4 million
PS330K	1 if population size between 330 thousand and 1.2 million
PS125K	1 if population size between 125 thousand and 330 thousand
INTNET	1 if expense on computer information service is positive
TV	Number of televisions in the household
COMPUTER	Number of computers
SOUNDCP	Number of sound components
VCR	Number of VCR
VEHQ	Number of vehicles