Will the Leak Sink the Ship? Screener Leaks and the Impact of Movie Piracy

Zhuang Liu

December 10, 2018

Latest version: `http://publish.uwo.ca/~zliu328/screener.pdf`

Abstract

Screeners are movie copies sent to critics and industry professionals for evaluation purposes. Sometimes screeners are leaked accidentally and made available to download on the Internet. This paper exploits the plausibly exogenous variation of file sharing/piracy activities caused by screener leaks of Oscar nominated movies to estimate the impact of movie piracy on box office revenue. Using information on leak dates collected from *thepiratebay.org*, I employ a difference-in-difference strategy to identify the causal effect of piracy on movie box office. The paper finds two interesting results. First, screener piracy caused by leaks reduces the box office revenue of the leaked movie in subsequent weeks by 29.8% on average. However, the negative impact on total box office is to a large extent moderated by the late occurrence of leaks. Second, there are significant negative indirect effects on other movies: An additional contemporaneous leak lead to a 3% decrease of box office revenue of other unleaked movies.

1 Introduction

The Christmas season of 2016 was not as sweet as it used to be for Hollywood movie-makers. On 20 December 2016, Hollywood was shocked to find that high-quality pirated versions of two blockbuster movies—*The Hateful Eight* and *The Revenant*—had been leaked to the BitTorrent network, joining an unprecedented long list of leaks that week, including *Creed, Legend, In the Heart of the Sea, Joy, Steve Jobs, Concussion and Spotlight*. Many rate this unprecedented week of leaks as the worst incidence of piracy leaks in Hollywood history.

The end of December is a special time for movie pirates due to 'screener piracy'. Usually, copies of pirated movie are ripped from . Screeners are movie copies that are sent to movie critics and reviewers for awards consideration. Screener piracy are pirated videos ripped from screeners and are of similar quality to videos ripped from media such as DVDs (digital video discs) or online streaming websites. Screener piracy is a concern to studios mainly because some appear relatively early in the theatrical run and have better video quality than the other early low-quality CAM (camcorder) piracy, which are mostly from boot-leg recordings in theatre.

Hollywood studios have taken serious precautionary efforts to prevent screener leaks. Despite their substantial precautionary efforts, bad luck does still strike. Every year, a few "unlucky" DVD screeners are leaked and then flood the Internet. While these leaks might be mishaps to the movie industry, the potential randomness of these events provide a great opportunity to investigate the role of piracy on movie sales. To date, whether and to what degree piracy hurts revenue is still a hotly debated empirical question. On the one hand, since the invention of Napster, music record sales have declined dramatically, naturally suggesting that file sharing or piracy activities are the primary suspect for such a dramatic decline. However, on the other hand, other confounding factors, such as the change in digital distribution channels and the emergence of other means of digital entertainment, which happen roughly at the same time with the surge of digital piracy and file sharing, make the issue more complicated.

This paper explores the impact of movie piracy on box office revenue using the exogenous variation in piracy activities created by the screener leak shocks. The identification builds on the orthogonality between being leaked and the movie's unobservable quality. Using a dataset of leak dates for Hollywood movies released between 2003 and 2016 obtained on *thepiratebay.org*, and a dataset including each movie's weekly and total box office from *Boxofficemojo*, I explore the relationship between screener piracy and industry box office revenue.

This paper finds two interesting results. First, pre-release screener piracy caused by the leaks reduces the box office revenue of affected movies 29.8% on average. Second, the results reveal a significant indirect effect of piracy on different movie titles. An additional contemporaneous screener leak negatively affects box office revenue of other unleaked movies by a baseline of 3%, which indicates the effect of piracy also spillovers to the legitimate sale of other movies. It suggests that total cost of an additional piracy leak will be higher to the industry as a whole than its cost for the particular leaked movie.

The rest of the paper is organized as follows. Section 2 describes the relevant literature. Section 3 discusses the data and background in this paper. Section 4 present the main empirical strategy. Section 5 discusses the estimation result. Section 6 shows alternative analysis and robustness check. Section 7 concludes the paper.

2 Literature Review

This paper adds to a large body of literature that focuses on the effect of piracy/file-sharing on sale of digital product. Overcoming the potential endogeneity problem of piracy is a challenging task in the empirical literature piracy/file-sharing. As better movies naturally attracts pirates and are associated with more piracy activities. The selection problem would contribute to substantial positive bias which could mask the true effect in an OLS regression. Earliest study including Liebowitz (2004) who assessed various possible explanations for the recent decline in music sales and found that MP3 downloads do harm music sales because alternative reasons cannot explain the observed reduction in sales.

Using a panel data of aggregate music sales by country and individual-level cross-section data, Zentner (2006) studied the effect of music downloads on music purchases.

Using the number of broadband Internet users as measures of file sharing activities, and using degrees of Internet sophistication and Internet speed as instruments, he found that file sharing reduces an individual's probability of purchase music by an average of 30%. Based on his estimates, he concluded that without file sharing, music sales in 2002 would have increased by 7.8%.

Rob and Waldfogel (2004) study the same topic using survey micro data of 412 US college students and their album purchase information, after instrumenting for downloads using access to broadband connection, they find that each downloads reduce purchases by about 0.2 in their sample. Their welfare analysis shows that file-sharing significantly increase consumer welfare and the reduction of deadweight loss due to file-sharing doubles the loss of producer profits.

In another paper, Danaher and Waldfogel (2012) uses the international release gap and find that longer release windows are associated with decreased box office returns, even after controlling for film and country fixed effects. Also the effect is much stronger after adoption of BitTorrent and in those heavily-pirated genres.

However, another study by Oberholzer-Gee and Strumpf (2007) have found opposite results, they collected data on weekly album sales and weekly downloads of album on Napster, a first generation file sharing tool. Using international school holidays as instruments, their results have shown that the effect of file sharing on album sales is statistically indistinguishable from 0.

This paper mostly relates to Ma et al. (2014) who studied pre-release movie piracy and found that pre-release piracy causes a 19.1% decrease in revenue compared with piracy that occurs post-release. My paper and theirs differ in several dimensions. While Ma et al. (2014) stressed the importance of timing and focus on the piracy occurring prior to theatrical release, I emphasize the importance of piracy quality and focus on the impact of high-quality piracy resulting from screener leaks. As pre-release piracy might be caused by endogenous factors, such as the international release gap, focusing on screener leaks provides cleaner identifications as I show later in the paper.

3 Data

3.1 Data Collection

The data employed in this paper mainly consists of two part. First is data on both weekly and total movie box office for 9799 movies released from 2003 to 2016. The box office data is collected from the box office reporting website *BoxofficeMojo.com*, for each movie I also collect associated movie characteristics including movie genres, MPAA ratings, studios, movie runtime, budget[1] from *International Movie Database (IMDB).* Move rating data is obtained from *Rottentomatoes*, which take integer value from 1 to 100. I also collected information about film award wins and nominations for each movie from IMDB, particularly the awards and nominations with respect to the Academy Awards (Oscar). Some time-varying variables including opening screens, total number of screens are also collected from *BoxofficeMojo.com.*

The second part of the data consists of the dates of screener leaks for each Oscar nominee. The data is collected by scraping piracy search engine *thepiratebay.org.* I choose *thepiratebay.org* because it is one of the most popular torrent sites in 2017[2], the website has a long history since 2003, which allows me to trace back historic leaks of previous movies.[3] To collect the screener leaks data, for each Oscar nominees between 2003 to 2016, I searched on *thepiratebay.org* with keywords including combinations of movie name with one of three keywords to identify screeners: **"screnner", "dvdscr", "scr"**.[4] I use a automated script to extract search result of each query. *thepiratebay.org* provides upload date for each torrent file in search result, which allow me to track the earliest date of piracy. For each movie I take the earliest upload date in the search result as the leak date of its screener.[5]

[1] I can only obtain budget information for about one thirds of all movies, the movie missing budget informations are small/micro budget movies, so I calculate the median of all existing budgets, and construct the budget variables in our analysis as a discrete variable. It equals 1 if a movie's budget is higher than the calculated median, and equals 0 if the budget is smaller than the median or missing.

[2] https://torrentfreak.com/top-10-most-popular-torrent-sites-of-2017-170107/

[3] Although its service did get interrupted several times in history because of legal issues, the database is not affected so previous information before interruptions is still available.

[4] These keywords are most common screener format indicators in torrent file name.

[5] Baio (Accessed in December 2016) also collect statistics on leaks date for Osacar nominees, I find

Data of movie theatrical release dates in US are collected from *BoxofficeMojo.com*. Public data on screener release date (the date when screeners are sent to critics) is difficult to find. Following Baio (Accessed in December 2016), I first use screener receipt dates reported by movie critic Ken Rudolph from his personal website[6] as the screener release date. The website record the critic's date of every movie screener received from 2001 to 2017. I also use earliest reviews posted in IMDB as an alternative check, for the movies with different release date in the two data, I take the earlier one.[7]

3.2 Descriptive Statistics

I then proceed to discuss the definition of important variables in the analysis. Variable $Leak_{it}$ is constructed as a dummy variable which equals 0 if at week t movie i's screener is not leaked yet. It equals 1 if week t past the leak date for movie i. I also construct a cross sectional version of the treatment variable which equals 1 if the movie has experience leaks on Internet and 0 otherwise.

Table 1 shows the how the number of leak incidence change over time. From the table, screener leaks in BitTorrent exhibit significant time-series variations. A higher level of leaks is observed during the period from 2003 to 2009, with on average 3.7 movies get screener leaks before or during first week of release. for those leaked movies, the average time between the US release date and leak date is about 50 days on average. After 2009, the pre-release leaks happens less frequently, averaging 1 movies every year during the period of 2010-2015. The average leak-release gap, however, has shrunk from about 50 days to 38 days. In 2016, the number exploded because of the massive leak incidence by the group HIVE-CM8 which are mentioned in the begining of the paper[8]. A large number of movies were leaked during their theatrical runs in December, 6 were

the leak date collected from *thepiratebay.org* to be different from Baio (Accessed in December 2016) for a significant fraction of movies. For movies with different leak dates in the two data, I take the earlier one.

[6]http://kenru.net/movies/

[7]Measurement error is likely to exist on screener release dates, but screener release data is only used for testing exogeneity of leaks. Because these data are not used in the main empirical analysis, main results in this paper is not affected.

[8]Source: https://www.theverge.com/2015/12/24/10663146/hollywood-s-christmas-is-being-ruined-by-unprecedented-leaks

Table 1: Time trend of Screener Leaks

Year	Average Time between US release date and Screener leak date	Number of movies leaked on or before first week in theatre	Number of movies leaked before release date
2003	44.88	3	2
2004	31.125	8	6
2005	57.1538	2	2
2006	51.5417	2	2
2007	43.2333	4	2
2008	55.9412	3	2
2009	57.5	4	1
2010	25.2222	2	2
2011	51.6667	1	1
2012	36	1	1
2013	45.9412	1	1
2014	43.7143	0	0
2015	28.1538	2	2
2016	30.9444	8	6

leaked before their scheduled release dates and 2 at the first week after release.

I first compare total box office of Oscar nominees over the year with the intensity of leaks every year, using the measures mentioned in Table 1. Figure 1 plot the evolution of yearly aggregate box office of all Oscar nominees and two measures of intensity of screener leaks: number of movies leaked before or during first week of release, and the average days between US theatrical release to leaks. The time series patterns show that there is significant correlation between box-office performance and leaks: box-office tends to be lower if there are fewer leaks, and box-office are higher if on average leaks happen at later time. The correlation suggests that part of the variations in box office could be potentially attributed to screener piracy activities.

Table 2 reports summary statistics for four sets of samples. As a reference group, column (1) shows descriptive statistics for all movies released in United States theatrical market during 2003-2016. there are 9799 movies in the whole sample. On average one movie yields box office revenue of around 16.9 million dollars, with a huge dispersion represented by the standard deviation of 47.88 million. Similar dispersions exist among

Figure 1: Pattern of Correlation between Box office and Leaks over Time

other variables such as Total number of screens, first week box office. Budget data are only available for around 2000 movies, of those available, budget is averaged at 56.39 million with standard deviation of 82.75 million. Of the whole sample, about 4.55% are nominated for the Academy Award.

This paper will focus on the sample of Academy Award (Oscar) nominees[9]. The subsample of Oscar nominees represent 4.5% of the total sample. As heterogeneity in movies are huge, restricting the sample to Oscar nominees will make the sample more comparable in term of quality and market reception, reducing the potential endogeneity due to unobserved quality. All Oscar nominees have their screeners sent out prior to the award ceremony and becomes potential target of the piracy leaks. The summary statistics are presented in column (2). Clearly Oscar nominees are of more box office appeal because of the better quality, with average total box office revenue increase from 16.9 to 93.2 million dollars, these movies also invest more in budget (83.45 million compared with 56.39 million), and are assigned to more opening screens by theatres (2019 compared with 718). If we look at the distribution of genres, drama and science fiction movies are over-represented in the sample of Oscar nominees, their shares rise from 22 % and 2.1% to 41% and 5.8% respectively. Leaks happen in about 60% of the Oscar nominees. And around 11% of leaks happens prior to theatrical releases.

Column (3) reports summary statistics for the leaked Oscar movies, column (4) report statistics of the unleaked movies. Comparison between these two columns shows that observable differences for most movie characteristics between the leaked and unleaked movies is quite small.

On average, leaked movies and unleaked movies have similar number of screens (2014 vs 2028). As total number of screens are strategically adjusted according to change in demand, then it indicate that ex post market reception are similar between leaked and unleaked movies. But in terms of opening screens, I do observe some difference: unleaked movies have more screens than those leaked ones, the difference are resulted from a higher number of leaked movies choosing limited release initially[10], which indicates that ex ante

[9]Here I include Oscar nominees for all award categories except short film (live action).

[10]Limited released movie usually first release in theatres in major metropolitan areas like New York

unleaked movies are expected to perform better in market. The distribution of movie genres are slightly different between leaked and unleaked Oscar nominees. Genres like drama and crime are more prevalent in leaked movies, while number for animation and science fiction are less than those in the unleaked sample. Difference in the other genres are not statistically significant. As for the market outcome variables, unleaked movies have higher total box office on average. To better gauge the difference between two samples, I conduct a balance test and results are shown in the next section.

4 Testing the Exogeneity of Leaks

Before proceeding to the main empirical analysis, I first conduct several tests on the exogeneity of leaks. As a natural way of analyzing my research question is to directly compare box office between leaked and unleaked movies, the underlying identification assumption is that leaks are independent of other movie-specific unobservable confounding factors that also relate to box office revenue. Result of cross-sectional comparison of box office will be biased by any movie-specific unobservables affecting both demand and leaks. For example, It may be the case that leaks are affected box office demand, movie pirates are targeting at the most popular movies, so movies with higher box office appeal are more likely to suffer from leaks[11]. It is therefore important to check the validity of my identification assumption. Although there is no conclusive test of the identification assumption, I can exploit the richness of my data to present some evidence supporting the exogeneity of screener leaks incidence.

4.1 Balance Test

To address these potential concerns, first I did a balance test for observable characteristics between the leaked and unleaked samples. For each characteristics, I run a t-test for

and Los Angeles. After gauging their market appeals, some movies will then proceed to nationwide release.

[11] As discussed in the Empirical Strategy section, the DD specification is immune to this case as any time-invariant difference would be taken care of using movie fixed effect. However the this case would bias any cross-sectional result as it is impossible to include movie fixed effect.

Table 2: Summary Statistics

	(1) All Movies Sample		(2) Oscar Nominees		(3) Oscar Nominees with Leaks		(4) Oscar Nominees without Leaks	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
Key Variables								
Total Number of Screens	718.5	1209.9	2019.9	1421.1	2014.3	1288.4	2028.3	1604.2
Opening Screens	670.4	1209.1	1465.4	1675.9	1309.1	1576.12	1703.3	1796.1
Budget	56.3	82.7	83.4	89.4	77.7	95.7	93.4	76.3
Screener is Leaked	0.027	0.163	0.600	0.490	-	-	-	-
Pre release leaks	0.003	0.057	0.067	0.251	0.112	0.316	-	-
Oscar nomination	0.045	0.208	-	-	-	-	-	-
Genre								
Action	0.060	0.238	0.078	0.269	0.059	0.237	0.106	0.309
Drama	0.227	0.419	0.410	0.492	0.466	0.499	0.325	0.470
Animation	0.031	0.175	0.134	0.341	0.093	0.291	0.196	0.398
Comedy	0.193	0.395	0.123	0.329	0.141	0.349	0.095	0.294
Crime	0.026	0.160	0.033	0.180	0.052	0.222	0.005	0.074
Horror/Thriller	0.118	0.323	0.087	0.282	0.089	0.286	0.084	0.278
Romance	0.042	0.201	0.024	0.155	0.022	0.148	0.028	0.165
Science Fiction	0.021	0.143	0.058	0.234	0.037	0.189	0.089	0.286
Documentary	0.152	0.359	0.011	0.105	0.007	0.086	0.016	0.129
Market Outcome								
Total Box office (Millions)	16.9	47.8	93.2	117.6	89.4	105.2	98.9	134.3
Box office in opening week (Millions)	5.2	15.0	20.8	34.1	16.2	26.4	27.7	42.5
Observations	9799		446		268		178	

Table 3: Test of Baseline Balance

	Oscar Movies with Leaks (A)	Oscar Movies no Leaks (B)	Leaked - Unleaked (A-B)	(t-statistics)
t-tests Between Group Means				
Key Covariates				
Rating	79.42	78.31	-1.12	(-0.63)
Total Number of Screens	2014.28	2028.34	14.06	(0.10)
Scheduled screens for first week	1309.15	1703.31	394.16*	(2.44)
Budget	77.79	93.46	15.67	(1.49)
Genres				
Action	0.06	0.11	0.05	(1.81)
Drama	0.46	0.32	-0.14***	(-2.97)
Animation	0.09	0.19	0.10***	(3.16)
Comedy	0.14	0.09	-0.46	(-1.46)
Crime	0.05	0.01	-0.04**	(-2.69)
Horror/Thriller	0.09	0.08	-0.01	(-0.19)
Romance	0.02	0.03	0.01	(0.37)
Science Fiction	0.04	0.09	0.05*	(2.33)
Documentary	0.01	0.02	0.01	(0.92)
Observations	268	178		
Joint Test of orthogonality		F-statistics:	1.15	

Note: This table reports (1) the differences in mean variables and the corresponding t-statistics between the leaked and unleaked movies. (2) Joint test of orthogonality is done by F-test on a linear regression of leak indicator variable as dependent variable and all the covariates excluding market outcome variables as regressors.

equality of means for major observable characteristics and the t-statistics are reported in the last column of Table 3. We would expect small t-statistics across most observables if two samples are relatively homogeneous. Based on the result, for the most important variables, the mean of rating between two samples are not significantly different. For example, two samples are almost identical in ratings, with 79.428 for leaked movies and 78.31 for unleaked movies. When it turns to total number of screens which usually indicate responses of market performance, the difference are also not significant. These indicate evidence that leaks might not correlate with demand factors such as quality and box office attractiveness of the movie. The result also shows that for total box office and genres like Action, Comedy, Horror, Romance there are no significant difference between two groups. However, I do observe that genres like Drama and Crime are over-represented in the leaked group so it is important to control for them in our cross-section regression.

I also conduct a joint test of orthogonality. To test whether there are any factors driving the selection of leaks. I run a linear regression of dummy variable of leaks as dependent variable and all the covariates excluding market outcome variables as regressors. Result of the F-statistics can not reject the hypothesis that all coefficient are 0.

According to result of balance test, When compare box office cross-sectionally, I include controls for selected covariates that are significantly different between two groups: mainly genres and screens. I also control for a number of other factors: (1) *Year Dummies*. pirates activities vary significantly over years. For example, after the outburst of leaks in 2016, one of piracy group responsible for leaks released apology on-line and promised to delay release of leaked screener in subsequent years[12]. Indeed we observe less piracy in year of 2017. (2) *Calender Week Dummies*. i will add Calender Week Dummies to control for timing of release. There exists strong seasonality of piracy activities, leak incidences happen more frequently mostly during December and January and die down during months like April and June. As seasonality also exist in theatrical market[13], it is therefore necessary to employ calendar week fixed effects as control so that seasonality doesn't confound the result.

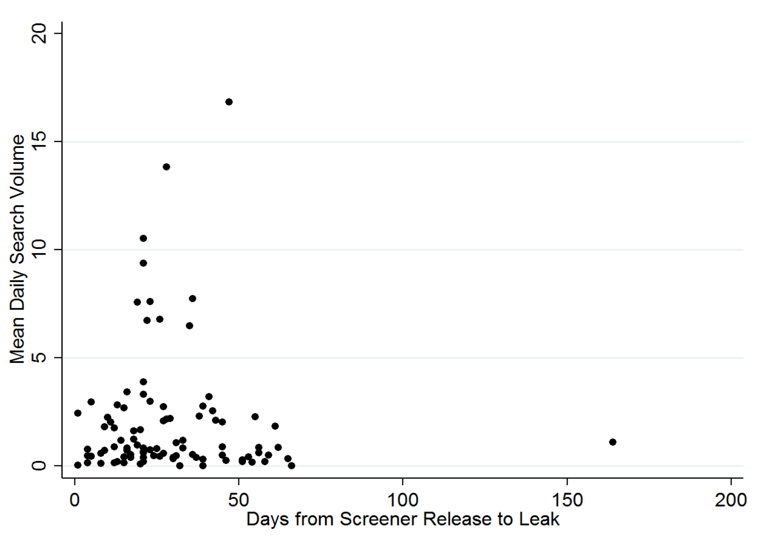
4.2 Are timing of leaks affected by demand?

Even if leaks are not connected to time-invariant movie quality, still it is possible that leaks are driven by unobservable time-varying demand shocks such as sudden increase in consumer interest or word-of-mouth, which will again introduce endogeneity to my main empirical analysis. To test whether the decision to pirate is the outcome of demand shocks, here I take advantage of Google Trends' search volume data as a measure of consumer interests or popularity, which is informative on unobservable demand shocks a movie is facing.

[12] https://variety.com/2016/digital/news/hateful-eight-piracy-group-apology-1201670927/

[13] The demand and supply seasonality in motion picture industry has been discussed in Einav (2007). For example, blockbuster movies are mostly scheduled around summer season and Christmas holidays. Screener leaks also have specific timings in proximity to the date of several major film awards, therefore it is expected that systematic quality difference driven by supply seasonality will be correlated with screener leaks.

Figure 2: Mean Daily Search Volume by Time to Leaks



Notes: Search Volume data are collected from Google Trends. *Titanic* is chosen as the reference title, all titles' search volume is normalized relative to those of *Titanic*. Plotted points represent mean daily search volume at time of screener leak against the length of time from the release of screeners to reviewers to the leak.

Table 4: Search Volume and Early Screener Leaks

	(1)	(2)
screener release to screener leaks	-0.0031	
	(0.0141)	
US release to screener leak		-0.0085
		(0.0096)
Observations	98	98
Adjusted R^2	-0.0102	-0.0023

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To be specific, I collect daily search volume of each leaked movie from 2010-2016 from 30 days prior to 30 days after the leaked date. First I calculated the mean daily search volume around this 2 month window, along with the time interval between each movie's screener release date and screener leak date, a measure of how fast the screener leak happens. I then plot the title's mean daily search volume over the time from release to leaks for each movie. Figure 2 indicates that the mean search volume of those leaked earlier is not higher than those leaked later. To further substantiate our observation from the figure, I regress mean daily search volume on the time from screener release to leaks. If movie pirates are influenced by demand factors, then we should expect the sign on release to leaks time to be significantly negative, as more popular titles will have earlier leaks. The result from Table 4 column 1 shows that the coefficient of interest is close to zero and not statistically significant. In addition, in column 2 I regress mean daily search volume on time between US release to screener leaks. This regression specification addresses the possibility that better or more popular titles are well protected so earlier release is harder and they enjoy longer "screener piracy-free" window as a result. Result in column 2 doesn't support this hypothesis as again the coefficient of interest is close to zero and statistically insignificant. In general the results suggest that these leaks are not driven by demand.

5 Cross-sectional Evidence

Perhaps the most straightfoward approach to this paper's question is to directly compare total sales of leaked and unleaked movies. Therefore I start with a cross-sectional comparison using total box office revenue. I will also exploit the leak timing by comparing movies with early leaks and those with late leaks.

5.1 Leaked vs. Unleaked

First, I conduct a simple comparison of total box office of leaked and unleaked movies. One challenge is the comparability between leaked and unleaked group. Although balance test in previous section suggests that observable differences between two samples are mostly insignificant, it is still necessary to control for observable characteristics especially those that are significantly different. Additionally, sample heterogeneity is reduced by limiting the analysis to a more homogeneous sample of only Oscar nominated movies. Specifically I use the following specification:

$$Y_i = \alpha + \beta Leak_i + X_i' \gamma + \sum_{k=2003}^{2016} \lambda_k \mathbb{1}(ReleaseYear_i = k) + \sum_{j=1}^{11} \tau_j \mathbb{1}(ReleaseMonth_i = j) + \epsilon_{it} \tag{1}$$

On the left hand side, $Leak_{it}$ is a dummy variable equals 1 if the movie's screener is leaked during its run in theatre. X_{it} is a set of controls. As movie fixed effects are unavailable, I add a rich set of controls including total number of screens, length of theatrical release, Rotten Tomatoes Rating, genre dummies, MPAA rating dummies, dummies for major studios, $ReleaseWeek_i$ is the calendar week at which movie was released, and $ReleaseYear_i$ is the year of release.

As for the dependent variable Y_i, I use two set of outcome variables: total movie box office and surprise sale used in the previous literature (Moretti (2011), Gilchrist and Sands (2016)).

Following Moretti (2011), I uses the number of opening theaters as proxy for expected demand and define the "**Suprise sale**" of a movie as the residual demand that is not predicted by the number of opening theaters and seasonality component (controlled for by month and years dummies). Specifically, for each movie I release at specific time t, I regress its opening screens $OpenScreen_i$ and release month and year dummies on its box office R_i.

$$R_i = \alpha + \beta OpenScreen_i + \sum_{k=2003}^{2016} \lambda_k \mathbb{1}(ReleaseYear_i = k) + \sum_{j=1}^{11} \tau_j \mathbb{1}(ReleaseMonth_i = j) + \epsilon_{it} \tag{2}$$

Let the model predicted box office be $\hat{R}_i$ and define the Surprise sale as the residual term:

$$Surprise_{it} = R_i - \hat{R}_i \tag{3}$$

The constructed surprise sale essentially measures the ability of a movie to outperform its market expectation. Under this specification, I treat the leaked movies as the treatment group and the unleaked movies as control group and test if the outcome variables is significantly different between these two groups. If movie piracy indeed severely cannibalize sales, I would expect the average sale of leaked movies to be much lower than the unleaked movies, given the assumption that screener leaks does not depends on unobservable heterogeneity that drive sales. In addition to the baseline specifications, I also try using other definition of treatment: I define treatment as "Leaked prior to Release", which represent a more intense treatment.

I report the result of cross-sectional regression in Table 5. Column (1) and (2) show result using Total Box office as dependent variable, column (3) and (4) show result using Surprise Sale as dependent variable. Start with the baseline comparison between leaked and unleaked in column (1) and (3), I did not find strong evidence that leaked movie performs worse than unleaked. The coefficient is with a negative sign, the magnitude indicate that on average leaked movies did worse by \$ 0.36 million for total box office and \$ 0.08 million for surprise sale. But the coefficient are of little statistical significance. The result are not surprising as for a large fraction of leaked movies, leaks happen at relatively late time in their theatrical runs. Due to the fact that movie revenue are concentrated in the beginning of theatrical run, even if screener piracy cause severe harm on subsequent box office, the total impact can still be moderate if screener is leaked at a very late time.

For further analysis, I switch to pre-release leaks as treatment and the result for total box office and surprise sale are reported in column (2) and (4), respectively. Once I restrict the attention to a more intense treatment, I find that the negative impact on box office becomes much higher in magnitude and statistically significant ($p < 0.1$). On

Table 5: Cross-sectional Results

Dependent Variables:	(1) Total Box office	(2) Total Box office	(3) Surprise Sale	(4) Surprise Sale
Leaked	-0.364		-0.081	
	(1.500)		(1.506)	
Leaked prior to Release		-3.202*		-2.993*
		(1.702)		(1.640)
Total Number of Screens	0.008***	0.008***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)
Length of Release (Days)	0.062***	0.060***	0.072***	0.070***
	(0.017)	(0.016)	(0.017)	(0.016)
Rotten Tomatoes Rating	0.072**	0.073***	0.086***	0.087***
	(0.028)	(0.028)	(0.028)	(0.028)
Additional Controls				
Genres Dummies	✓	✓	✓	✓
MPAA Rating Dummies	✓	✓	✓	✓
Major Studios Dummies	✓	✓	✓	✓
Calender Weak of Release Dummies	✓	✓	✓	✓
Year of Release Dummies	✓	✓	✓	✓
Observations	411	411	409	409
Adjusted R^2	0.660	0.663	0.412	0.415

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

average, movies with screener leaked prior to release did worse by \$ 3.202 million for total box office and \$ 2.99 million for surprise sale. In general the result reveals that the impact of leaks is negative but lacks of statistical significance. It also confirms that the impact is more pronounced if leaks happen earlier, suggesting that intensity margin of leaks might be important as well.

5.2 Early Leaks vs. Late leaks

Inspired by the previous result, the second empirical test explores the impact of leaks on the intensive margin. I utilize variations of the timing of leaks in theatrical run as the intensity of treatment. If leaks do harm sales, I would observe that after controlling for quality, movies leak earlier on average yield lower box office than movies which leak late.

For implementation, I construct a variable $LeakWeek_i$ which denote the number of weeks between movie i's US release date and its screener leak date.

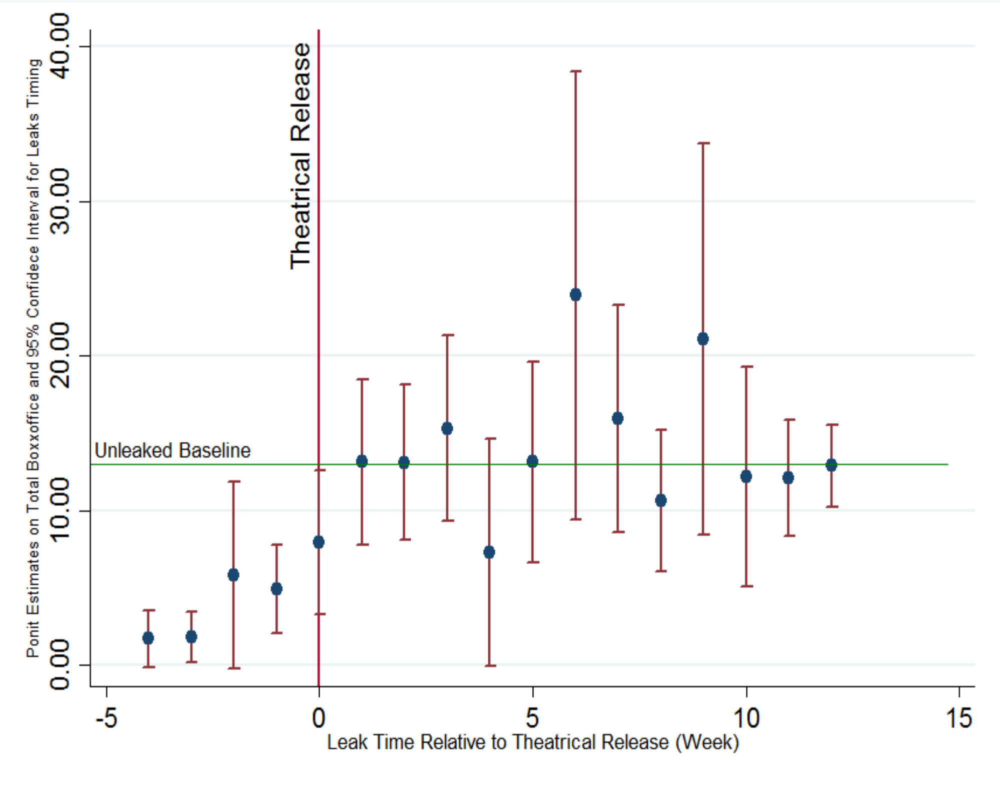
Notice that the marginal effect might vary depending on different treatment levels, I drop the uniform marginal effect assumption and transform the treatment variable into a series of dummies by different values of $LeakWeek_i$.

$$Y_i = \sum_{j=-4}^{15} \beta_j \mathbb{1}(LeakWeek_i = m) + X_i'\gamma + \sum_{k=2003}^{2016} \lambda_k \mathbb{1}(ReleaseYear_i = k) + \sum_{j=1}^{11} \tau_j \mathbb{1}(ReleaseMonth_i = j) + \epsilon_{it} \tag{4}$$

Here we $\mathbb{1}(LeakWeek_i = m)$ is a set of dummies equal 1 if The movie get leaked in the mth week after release. The estimated coefficients of β_j measures the week-specific effect of leak on movie outcome variable Y_i.

Figure 3 reports the point estimates and 95% confidence interval on β_j the effects of leak timing on movie outcomes. Because I did not include a intercept in the specification, the point estimate corresponding to each leak week dummy can be treated as the average box office in that leak week category after controlling for release time and year. There exists a significant upward trend in the early stage especially before theatrical release. These evidence are consistent with the belief that early leaks before theatrical release

Figure 3: Cross-Sectional Comparison: Effects of Leak Timing on Total Box-office



attract significant amount of consumer with high willingness-to-pay especially when there is a lack of legal channels, therefore the harm from pre-release piracy should be the highest. For subsequent leaks after the theatrical release, I did not find compelling evidence of harm to box office, as most point estimates are statistically indistinguishable from the unleaked baseline. This suggests that the harms on box office are significantly moderated by the late occurrence of leaks. Because box office for a movie is not uniformly distributed over time, a much larger fraction of box office is from first few weeks in the theatrical window, harm of leaks to box office is expected to be decreasing in the time of occurrence.

6 Empirical Strategy

I adopt a difference-in-difference (DD) strategy in the main analysis of this paper. Unlike the cross-section analysis, which hinges on the strong assumption that leaks are orthogonal to unobservable heterogeneity, the difference-in-difference strategy relies on a much weaker identification assumption of parallel pre-leak time trend between leaked and unleaked movies, which also implies that the time-varying movie-specific unobservable heterogeneity affecting sales are not correlated with screener leaks.

This paper's main identification strategy involves implementing a DD strategy using panels of weekly box office data. Screener leaks incidences generate substantial variations of piracy activity (downloads), both across movies and across time. The DD strategy will explore these variations and estimate the impact of piracy by comparing the change in ticket sales before and after screener leaks for leaked movies against a baseline of changes in ticket sales of those unleaked movies at same calendar time and release time.

The DD specification will take care of a majority part of concerns for selection as the following unobservable factors are accounted for in the specification: (1) movie-specific time invariant unobservable heterogeneity (e.g., better movies attract more pirates and have more downloads); (2) general decreasing trend of box office over its release time (e.g., natural decaying patterns will not be falsely attributed to the effect of piracy); (3) calendar time-variant but movie-invariant factors (e.g., box office and download both rise when summer holiday begins).

The DD specification take the following forms:

$$ln(Sale_{it}) = \alpha Leak_{it} + X_{it}\beta + \sum_{j} \mathbb{1}\left\{\tau_{it} = j\right\} + \xi_i + \lambda_t + \varepsilon_{it} \tag{5}$$

On the left hand side, the dependent variable is $ln(Sale_{it})$, the log of weekly box office for movie i at time t. On the right hand side of the regression I include movie fixed effect ξ_i, calender time fixed effects λ_t. In addition, because box office has a natural exponential declining pattern, I would need to control for weeks after release. As such declining is in a non-linear fashion I include a series of release week dummies

$\sum_j \mathbb{1}\{\tau_{it} = j\}$ where τ_{it} is the count of weeks after theatrical release with $j = 1, 2, ..., J$. It works non-parametrically to control the pattern. $Leak_{it}$ is an indicator variables which equals 1 if movie i's screener piracy has already leaked at time t. ε_{it} is the idiosyncratic error term. I cluster the standard error at movie level, allowing for serial correlation of error within movies.

The first difference is taken using the movie fixed effects which take care of time-invariant movie heterogeneity. The second difference is taken using time fixed effect which take care of the general time trends that are movie-invariant. The coefficient of interest is β will be the percentage changes of box office of movies after emergence of screener piracy (β) against the baseline change of movies without presence of piracy. If the estimated coefficient of β comes significant negative, it would be evidence indicating that piracy significantly cannibalize sales.

7 Results

7.1 OLS

Result of panel regressions using weekly data are reported in Table 6. As a starting point I first present the simple OLS result at first two columns. In column (1), I estimate the simplest OLS specification without including any fixed effects. Not surprisingly, the estimates are small in magnitude (-0.028) and insignificant as most of the upward bias due to endogeneity of movie quality has not been corrected yet. As a first attempt to correct for endogeneity, I proceed by including rating into the OLS regression as a control for quality. As the result shows in column (2), once rating is added, the coefficient on leaks becomes -0.135, which implies that occurrence of screener piracy decrease the weekly box office revenue by 12.6%[14] . In comparison to the previous OLS estimates, it suggests that adding controls for quality help reduce a significant amount of upward bias.

[14]Interpretation of coefficient is calculated as $(exp(-0.135) - 1) * 100$

7.2 Difference-in-Difference

Now we turn to the Difference-in-Difference (DD) estimate. In column (3) and (4), I further add movie fixed effects, calendar week fixed effects and week of release fixed effects to control for additional source of endogeneity for reasons discussed in previous section. After differencing out time-invariant movie-specific component, general seasonality component and decaying pattern using these three sets of fixed effects, the coefficient on Leaks becomes -0.355 and is much larger in magnitude compared with the OLS estimates. The comparison highlights again the importance of movie heterogeneity and time-varying demand effects in influencing the estimate. Result in (3) indicates that occurrence of screener piracy lower weekly box office by 29.8%.

To test the identification assumption on parallel pre-treatment trend, I conduct a falsification test as in Autor (2003) by including leads and lags of treatment in the DD specification.

Let T_i be time of leak for movie i, I estimate the following specification:

$$ln(Sale_{it}) = \sum_{k=-2}^{2} \alpha_k D_{it}(t = T_i + k) + X_{it}\beta + \sum_j \mathbb{1}\left\{\tau_{it} = j\right\} + \xi_i + \lambda_t + \varepsilon_{it} \qquad (6)$$

Here I include 2 leads and 2 lags of treatment effects, α_k is the coefficient on the kth lags or leads. The falsification test essentially test the hypothesis that all coefficients of leads are 0: $\alpha_k = 0$ for $k < 0$. I employ the same set of data and controls in the test and in the previous DD specifications.

I report the result of falsification test in column (5). The coefficients of all pre-treatment variables are statistically indistinguishable from 0, consistent with the assumption of parallel time trend. For all treatment effects after the leak, they are of same expected signs and are mostly statistically significant. I also observe the magnitude of treatment effect at the leak ($t = T_i$) become smaller after the inclusion of leads and lags, and becomes higher for treatment effects of 2 weeks lags ($t = T_i$+2). This observation suggests that treatment effects of leaks to movie box office accumulate and grow over subsequent weeks, potentially due to the fact that more consumers have discovered the leak over time.

7.3 Effect Heterogeneity by Rating

After the baseline specification, another natural question to ask is whether there exists any effect heterogeneity, especially how the harm due to screener leaks differs by movie rating. Did good movies hurt more than bad movies or the opposite? To examine the effect heterogeneity for rating, I add interaction terms between leaks and rating. If heterogeneity is significant, we will expect the coefficient to be statistically significant. As the result in column (4) shows, the coefficient of the interaction term is positive, although not very precisely estimated (significant at 10% level). The estimate indicates sizable effect heterogeneity for movie rating. For instance, movie rated at 90 out of 100 will suffer from a drop of $-(exp(-0.557 + 80 \times 0.003) - 1) \times 100 = 24.9$ percent of sales for box office at subsequent weeks after leaks, while movie rated lower at 20 out of 100 will suffer from a drop of $-(exp(-0.557 + 20 \times 0.003) - 1) \times 100 = 39.1$ percent of sales.

The result is consistent with the theoretical and empirical literature that highlight the positive role of piracy from various channels including for example sampling effect, network effect and word-of-mouth effect(Liebowitz (1985), Peitz and Waelbroeck (2006), Belleflamme and Peitz (2014), Peukert et al. (2017),). It suggest that higher rating movies better capitalize the sampling feedback or positive word-of-mouth from piracy. The increase in sales neutralize piracy's negative cannibalization, yielding a lower estimates of negative effect as shown in the result.

7.4 Indirect Displacement: Effects on Unleaked

A large body of the literature has focused on the "direct effects" of piracy: namely the response of sales revenue to its own piracy downloads. The most common analytical framework is based a single good setting, with emphasis on the direct displacement effect between one product's piracy and its own sales. The competition between different products and complex interactions between piracy across substitutable products are

Table 6: OLS and difference in difference estimates of impact of leaks on weekly box office

	(1) OLS	(2) OLS	(3) DD	(4) DD	(5) Falsification Test
Leak	-0.028	-0.135***	-0.355***	-0.557***	
	(0.019)	(0.019)	(0.023)	(0.119)	
Leak × Rating				0.003*	
				(0.001)	
Rating		0.006***			
		(0.000)			
Log number of screens	1.065***	1.084***	0.974***	0.969***	0.976***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.006)
Falsification Test					
Pre-Treatment (T-2)					0.016
					(0.052)
Pre-Treatment (T-1)					-0.008
					(0.034)
Treatment (T)					-0.086*
					(0.042)
Post-Treatment (T+1)					-0.032
					(0.039)
Post-Treatment (T+2)					-0.115**
					(0.043)
Movie FE			✓	✓	✓
Calendar week FE			✓	✓	✓
Weeks since release dummies			✓	✓	✓
Observations	79454	73096	79454	73096	42320
Adjusted R^2	0.867	0.875	0.865	0.871	0.905

Notes: The table provides the OLS and difference-in-difference estimates on the the effects of piracy on box office. The dependent variable is the log of weekly box office of movie i. Column (1) reports the baseline estimates using OLS, column (2) adds rating as control, column (3) reports the baseline difference-in-difference estimates. column (4) interacts Leaks with rating. Results of falsification test is reported in column (5). All specifications include movie FE and calendar week FE. Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

largely ignored in the literature.

In this section I intend to utilize the screener leaks incidence to investigate into the potential "indirect displacement effect" to the sale of other available movie substitutes. under the context of this paper, I use "indirect effect" to refer to the displacement effect of screener leak on sales of other unleaked movie titles. I construct a variable $NumLeaks_{it}$ which denotes the number of leaks incidence happened in the specific weeks. For example, if three movie's screener are leaked at t, then the $NumLeaks_{it}$ is then 3 for all movie. The baseline specification is a straightforward one:

$$ln(Sale_{it}) = \alpha NumLeak_{it} + X_{it}\beta + \lambda_t + \xi_i + \epsilon_{it} \tag{7}$$

Where the dependent variable is the log of weekly box office, one the right hand side, X_{it} is a vector of time-varying control variables including *number of screens, weeks since release*. λ_t is the time (week) fixed effects. ξ_i denote the movie fixed effects. ϵ_{it} is the idiosyncratic error term.

Does the magnitude of indirect effects affected by the degree of substitutability between affected movie and leaked movie? To answer the question, I explore different measures of $NumLeaks_{it}$ in addition to the baseline model. I narrow down the treatment to number of leaks that are of same genre and MPAA ratings as movie i, as movies in same genres and MPAA ratings are generally considered as closer substitutes. If increased magnitude of the estimated coefficient is observed, it will suggest that indirect effects get stronger for unleaked movies that are closer substitutes to the original leaked movie.

Regression result of the baseline model is shown in Table 7. From the baseline result in column 1, An additional contemporaneous leaks will reduce the same week box office of other unleaked movies by around 3%. The results suggest that the indirect displacement effects is significant. Although the magnitude is relatively small compared to previous DD estimates, the total cost in revenue might be higher given the large number of movies screening together in theatres. The presence of indirect displacement implies negative externalities of piracy. Magnitude of the negative externalities will be stronger for markets with higher level of competition and products that are closer substitutes. Total cost of an additional piracy leak will be higher to industry than its cost for the particular leaked movie. Research addressing only direct displacement will potentially underestimate the total harm of piracy to the industry, especially under the context of movie consumption.

Column (2) and (3) report results using leaks in same genres and MPPA ratings. I find that an additional leak will reduce box office of unleaked movies in same genre by 5%. In addition, an additional leak will reduce box office of unleaked movies in same MPAA rating by 3.9%. The higher magnitude of the estimated coefficient indicates that indirect displacement indeed gets stronger for closer substitutes.

It should be noted that the presence of externalities on other movies also implies contaminations as the treatment of "leaks" are also affecting the control group of unleaked movies. The existence of externalities will have downward bias of the DD estimates, underestimating the negative direct displacement effect of leak on subsequent movie's box office itself.

8 Conclusion

There has been a long debate on the causal effect of piracy on revenues of digital content industries. The empirical difficulty for identification arises from the confounding effect of unobservable demand factors affecting sales and download simultaneously. In this paper I use screener leaks as an natural experiment for identification. Screeners are copies of movies sent to critics and industry professionals for evaluation purpose. Sometimes screeners get leaked accidentally and made available to download on the Internet. These leaks incidences provide a good opportunity for identifying the causal effects of piracy on movie sales.

Using box office and screener leaks data of Oscar nominees from 2003-2016, I estimate the impact of movie piracy on box-office revenue exploiting the plausibly exogenous

Table 7: Indirect Displacement Effects on Unleaked

	(1) All Leaks	(2) Same Genre Leaks	(3) Same MPAA Rating Leaks
Number of Leaks	-0.030***		
	(0.008)		
Leaks under same genres		-0.049*	
		(0.021)	
Leaks under same MPAA Rating			-0.038**
			(0.014)
Weeks After Release	-0.099***	-0.099***	-0.099***
	(0.005)	(0.005)	(0.005)
Number of Screens	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
Movie FE	✓	✓	✓
Calendar week FE	✓	✓	✓
Observations	78730	78730	78730
Adjusted R^2	0.609	0.609	0.609

Notes: The table provides estimates on the the effects of piracy on box office of other unleaked titles. The dependent variable is the log of weekly box office of movie i. Column (1) reports the estimates using total number of new weekly leaks as the treatment variable, column (2) re-defines the treatment variable as new weekly leaks under the same genre, column (3) re-defines the treatment variable to new weekly leaks under the same MPAA rating. Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

variation of piracy activities caused by screener leaks of Oscar nominees. Difference-in-difference estimates suggest that screener piracy caused by the leaks reduces leaked movies' box-office revenue in subsequent weeks by 29.8% on average. Although the magnitude is relatively large, total revenue loss due to screener piracy might be much lower as majority of screener piracy occurs at late stage of theatrical run, piracy therefore only affecting demand at the residuals. Second, evidence suggests that effects of leaks "spillover" to other unleaked movies. I find significant negative indirect displacement effects among other movies: one additional contemporaneous leak lead to a 3% decrease of box-office revenue of other unleaked movies. Magnitude of the indirect displacement effect is stronger for movies that are closer substitutes. The presence of indirect displacement implies negative externalities of piracy. Total cost of an additional piracy leak will be higher for the total industry revenue than its cost for the particular leaked movie.

The result suggests that it is important to deter emergence of high-quality piracy at early stage of release because of their potentially detrimental effects on sales. It should also be noted that piracy are of differentiated qualities, the effects could be different by quality. As screener leaks are high-quality piracy comparable to the quality of common home video media, the implication of this paper's finding should be limited to the scope of high-quality piracy only.

Bibliography

David H Autor. Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of labor economics*, 21(1):1–42, 2003.

Andy Baio. Pirating the oscars, `http://waxy.org/2016/01/pirating_the_oscars_2016`. Accessed in December 2016.

Paul Belleflamme and Martin Peitz. Digital piracy: an update. 2014.

Brett Danaher and Joel Waldfogel. Reel piracy: The effect of online film piracy on international box office sales. 2012.

Arthur S De Vany and W David Walls. Estimating the effects of movie piracy on box-office revenue. *Review of Industrial Organization*, 30(4):291–301, 2007.

Liran Einav. Seasonality in the us motion picture industry. *The Rand Journal of Economics*, 38(1):127–145, 2007.

Duncan Sheppard Gilchrist and Emily Glassberg Sands. Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy*, 124(5):1339–1382, 2016.

Ben B Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, pages 219–236, 2008.

Stan Liebowitz. Will mp3 downloads annihilate the record industry? the evidence so far. *Advances in the Study of Entrepreneurship, Innovation, and Economic Growth*, 15:229–260, 2004.

Stan Liebowitz. File sharing: creative destruction or just plain destruction? *Journal of Law and Economics*, 49(1):1, 2006.

Stan J Liebowitz. Copying and indirect appropriability: Photocopying of journals. *Journal of political economy*, 93(5):945–957, 1985.

Liye Ma, Alan L Montgomery, Param Vir Singh, and Michael D Smith. An empirical analysis of the impact of pre-release movie piracy on box office revenue. *Information Systems Research*, 25(3):590–603, 2014.

Enrico Moretti. Social learning and peer effects in consumption: Evidence from movie sales. *The Review of Economic Studies*, 78(1):356–393, 2011.

Felix Oberholzer-Gee and Koleman Strumpf. The effect of file sharing on record sales: An empirical analysis. *Journal of Political Economy*, 115(1):1–42, 2007.

Martin Peitz and Patrick Waelbroeck. Piracy of digital products: A critical review of the theoretical literature. *Information Economics and Policy*, 18(4):449–476, 2006.

Christian Peukert, Jörg Claussen, and Tobias Kretschmer. Piracy and box office movie revenues: Evidence from megaupload. *International Journal of Industrial Organization*, 52:188–215, 2017.

Rafael Rob and Joel Waldfogel. Piracy on the high c's: Music downloading, sales displacement, and social welfare in a sample of college students. Technical report, National Bureau of Economic Research, 2004.

Rafael Rob and Joel Waldfogel. Piracy on the silver screen. *The Journal of Industrial Economics*, 55(3):379–395, 2007.

Alejandro Zentner. Measuring the effect of file sharing on music purchases. *Journal of Law and Economics*, 49(1):63–90, 2006.