9-26-2019 4:00 PM

Essays on the Economics of Digital Piracy

Zhuang Liu
The University of Western Ontario

Supervisor
Salvador Navarro
The University of Western Ontario
David Rivers
The University of Western Ontario

Graduate Program in Economics
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

My thesis consists of three chapters relating to the economics of digital piracy. Digital piracy is a debated topic catching tremendous academic and public attention. My studies contribute to the understanding of the impact of digital piracy on legitimate sales revenue with a focus on the motion picture industry.

In my first chapter, I examine the effects of screener piracy on the movie box office. Screeners are movie copies sent to critics and industry professionals for evaluation purposes. Sometimes, screeners are leaked and made available to download on the Internet. This chapter 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 compare the box office performance of leaked and non-leaked movies both cross-sectionally and employing a difference-in-difference strategy to identify the causal effect of piracy on movie box office. I find evidence that screener piracy reduces the box office revenue of the leaked movie in subsequent weeks. However, the negative impact depends on the timing of leaks. Damages to the total box office for movies with late leaks are much smaller compared to pre-release screener leaks.

One reason behind the lack of consensus regarding the impact of piracy in the empirical literature is the dearth of data on actual piracy activity. Data limitation restricts most empirical researchers to reduced-form and quasi-experimental methods, where estimates on industry loss may be biased due to extrapolation of the local average treatment effects to the population. To deal with the data challenge, in the next two chapters I have collected a dataset of weekly illegal downloads on BitTorrent in the United States from 40,267 torrent files for 255 movies released between March 27th and December 27th, 2015. Being able to directly observe piracy activities, my studies avoid many issues such as measurement error or sample selections which exist in studies using proxies or conjoint surveys.

In my second chapter, I utilize this novel dataset of illegal movie downloads to estimate a random-coefficient demand model of movie consumption. Using counterfactual experiments, I quantify the effects of piracy on movie box office revenue. The model distinguishes two channels of piracy substitution effects: (1) intra-movie substitution effects, which are the focus of most previous literature; and (2) inter-movie substitution effects, which represent the spillover of one title's piracy to the other titles' revenue. While the latter channel has been rarely explored in the literature due to restrictions imposed by data and methods, the novel dataset and demand model allow me to quantify its relative impact. I find that the revenue loss due to piracy's inter-movie substitution is on average 4 times as large as the intra-movie substitution effects. Omitting inter-movie substitution severely underestimates the actual damage of piracy. Other counterfactuals show that piracy reduces the total box office revenue of the motion picture industry by $93 million in total, 1.09% of the current box office.

In my third chapter, I extend the result of my second chapter by considering heterogeneity in the effects of piracy. Based on the data and methodology employed in the second chapter, I utilize information on video quality of illegal files to separate piracy by video quality. Augmented by data on home-video sales, I quantify the heterogeneous effects of piracy by distributional channels and video quality. In addition, a heatedly debated topic on digital piracy is about the potential positive effects of piracy due to channels like word-of-mouth. I incorporate the word-of-mouth effects of piracy in my model and quantify the contribution of piracy's word-of-mouth on movie sales revenue. I find that piracy's effects are large on the home-video market and small on the theatrical market. For high-quality piracy, substitution effects dominate word-of-mouth effects. For low-quality piracy, substitution effects are dominated by word-of-mouth effects which make low-quality piracy a potential promotional tool for studios.

Keywords: Industrial Organization, Digital Piracy, Intellectual Property Rights, Word-of-mouth, File-sharing, Motion Picture Industry

Summary for Lay Audience

Nowadays, technology like file-sharing makes it very easy to get illegal copyrighted content on the Internet. The digital piracy problem has attracted a great deal of public attention. There are concerns that digital piracy might kill the creative content industry. My thesis examines the impact of digital piracy on legitimate sales revenue using data on the motion picture industry.

I find that the effects of piracy are more complicated than we used to believe. First, I find that screener piracy accidents reduced leaked movies' box-office. The effects are more pronounced for pre-release leaks. Second, using actual downloads data on movies, I find that piracy of one movie also negatively affects the revenues of other movies in theaters at the same time. Third, I find that the negative effects of piracy depend on a lot of factors. Physical home-video market revenues are heavily affected by piracy, but the negative effects on theatrical market revenue are very small. High-quality piracy with high resolution is mainly responsible for the lost sales, while low-quality piracy can be used as a promotional tool if we consider the word-of-mouth produced by pirates. In the end, the results suggest that the effects of piracy are very different. The conclusion we draw regarding piracy in one industry may not be easily extended to the others. Public policy on digital piracy should be cautious about over-generalizing evidence from other industries.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the very important developments on the Internet is the emergence of peer-to-peer (P2P) file-sharing. In less than 20 years, P2P file-sharing has experienced dramatic growth and has become one of the most common activities on the Internet. The most widely used file-sharing protocol, BitTorrent, now has more than 170 million active users worldwide. It is claimed that BitTorrent moves as much as 40% of the world's Internet traffic on a daily basis.[1] The wide use of file-sharing has provided Internet users free and easy access to unauthorized copies of digital content such as movies and music, resulting in a surge in digital piracy.[2]

These facts have raised concerns among both policy makers and academic researchers about the economic effects of file-sharing on relevant industries. However, there is yet no consensus on the impact of file-sharing. Many people, especially copyright holders in the movie and music industries, treat file-sharing as the major reason for declining sales. Several widely quoted industry investigations have indicated evidence of huge economic losses, for example, software piracy cost the U.S. economy about 63.4 billion dollars in 2011 (Business Software Alliance's [BSA's] 2011 Piracy Study). Digital piracy causes 58 billion dollars in actual U.S. economic losses and 373,000 lost jobs (IPI 2005 study)[3]. However, the reliability of some of these estimates is under criticism

[1] BitTorrent Inc: http://www.bittorrent.com/company/about

[2] http://arstechnica.com/tech-policy/2015/08/riaa-says-bittorrent-software-accounts-for-75-of-piracy-demands-action/

[3] http://www.prnewswire.com/news-releases/58-billion-in-economic-damage-and-373000-jobs-lost-in-us-due-to-copyright-piracy-58354582.html

for the unrealistic assumptions made in these studies.[4] This relatively "naive" methodology inevitably inflates the estimated loss. A reliable quantification of the damage caused by piracy is therefore valuable for understanding the potential impact of piracy, and for resolving the current debate on intellectual property.

My thesis consists of three chapters relating to the economics of digital piracy. Digital piracy is a debated topic attracting tremendous academic and public attention. My studies contribute to the understanding of the impact of digital piracy on legitimate sales revenue with a focus on the motion picture industry.

In my first chapter, I examine the effects of screener piracy on the movie box office. Screeners are movie copies sent to critics and industry professionals for evaluation purposes. Sometimes, screeners are leaked and made available for download on the Internet. This chapter 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 compare the box office performance of leaked and non-leaked movies both cross-sectionally and employ a difference-in-difference strategy to identify the causal effect of piracy on movie box office revenue. I find evidence that screener piracy reduces the box office revenue of the leaked movie in subsequent weeks. However, the negative impact depends on the timing of leaks. Damages to the total box office revenue for movies with late leaks are much smaller compared to pre-release screener leaks.

One reason behind the lack of consensus regarding the impact of piracy in the empirical literature is the dearth of data on actual piracy activity. The data limitation restricts most empirical researchers to reduced-form and quasi-experimental methods, where estimates on industry loss may be biased due to extrapolation of the local average treatment effects to the population. To deal with the data challenge, in the next two chapters I have collected a dataset of weekly illegal downloads on BitTorrent in the United States from 40,267 torrent files for 255 movies released between March 27 and December 27, 2015. Being able to directly observe piracy activities, my studies avoid many issues

[4] For instance BSA admits that they assume that every download counts as one lost sale in their study

such as measurement error or sample selection that exist in studies using proxies or conjoint surveys.

In my second chapter, I utilize this novel dataset of illegal movie downloads to estimate a random-coefficient demand model of movie consumption. Using counterfactual experiments, I quantify the effects of piracy on movie box office revenue. The model distinguishes two channels of piracy substitution effects: (1) intra-movie substitution effects, which are the exclusive focus of the previous literature, and (2) inter-movie substitution effects, which represent the spillover of one title's piracy to the other titles' revenue. The latter channel has rarely been explored in the literature because of restrictions imposed by data and methods, but the novel dataset and demand model used in my study allowed me to quantify its relative impact. I found that the revenue loss due to piracy's inter-movie substitution is on average four times as large as the intra-movie substitution effects. Omitting inter-movie substitution would severely underestimate the actual damage caused by piracy. Other counterfactuals show that piracy reduces the total revenue of the motion picture industry from box office sales by $ 93 million in total, 1.09 % of the current box office revenue.

In my third chapter, I extend the result of my second chapter by considering heterogeneity in the effects of piracy. Based on the data and methodology employed in the second chapter, I utilize information on the video quality of illegal files to separate piracy by video quality. Augmented by data on home-video sales, I quantify the heterogeneous effects of piracy by distributional channels and video quality. In addition, I consider the potential positive effects of piracy due to channels like word-of-mouth (WOM), apossibility that has been the subject of much debate. I incorporate the WOM effects of piracy in my model and quantify the contribution of piracy's WOM effects on movie sales revenue. I find that piracy's effects are large on the home-video market and small on the theatrical market. For high-quality piracy, substitution effects dominate WOM effects. For low-quality piracy, substitution effects are dominated by WOM effects, which makes low-quality piracy a potential promotional tool for studios.

Chapter 2

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

2.1 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 because of screener piracy. Screeners are movie copies that are sent to movie critics and reviewers for awards consideration. Screener piracy involves 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 is mostly from boot-leg recordings in theatre.

Hollywood studios have taken serious precautionary efforts to prevent screener leaks. Firstly, in addition to the industry code of conduct and reputation which indirectly discipline the reviewers, screeners are also watermarked with the name of reviewing person or parties, so that the leaking person or parties can be traced and are held accountable. Secondly, most screeners DVD are protected by encryption which prohibits illegal copying. However, all of these safeguards don't guarantee perfect protections. Pirates sometimes managed to obtain screeners from "internal contacts" in the reviewing parties, and successfully decrypted the digital protection.

Despite their substantial 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, a first-generation file-sharing tool, music record sales have declined dramatically, naturally suggesting that file sharing or piracy activities are the primary suspect for such a dramatic decline. On the other hand, however, other confounding factors, such as the change in digital distribution channels and the emergence of other means of digital entertainment, which happened roughly at the same time as the surge of digital piracy and file sharing, make the issue more complicated.

One of the challenges in identifying the effect of digital piracy on the sales revenue of digital content is the endogeneity of piracy downloads, as both illegal downloads and sales are strongly correlated with unobservable movie characteristics. Better movies naturally attract more pirates and have higher piracy downloads. The selection problem would result in substantial positive bias, masking the true effect of piracy in an OLS regression.

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

The study in this chapter finds several interesting results. First, difference-in-difference estimates indicate that screener piracy leaks reduce the subsequent box office revenue of affected movies by 29.8% on average. Second, the negative effect depends on the timing of leaks. Cross-sectional analysis shows that on average movies with pre-release leaks under-perform in box office revenue by 3.2 million dollars. However, damages to the total box office revenue for movies with late leaks are more moderate and statistically insignificant compared to pre-release screener leaks. Generally damage from piracy decreases in leak time.

The rest of the chapter is organized as follows. Section 1.2 describes the relevant literature. Section 1.3 discusses the data and background. Section 1.4 presents preliminary pre-analysis checks. Section 1.5 discusses results of the cross-sectional analysis. Section 1.6 describes the alternative difference-in-difference analysis and results. Section 1.7 concludes.

2.2 Literature Review

This chapter adds to a large body of literature that focuses on the effect of piracy/file-sharing on sales of digital products. Overcoming the potential endogeneity problem of piracy is a challenging task in the empirical literature on piracy/file-sharing. Better movies naturally attract more pirates and are associated with more piracy activities. This selection problem contributes to substantial positive bias that could mask the true effect in an OLS regression. One of the earliest studies was 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 could not explain the observed reduction in sales.

Using panel data on 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 reduced an individual's probability of music purchase 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) studied the same topic using survey micro data of 412 U.S. college students and their album purchase information. After instrumenting for downloads using access to broadband connection, they found that each download reduced purchases by about 0.2 in their sample. Their welfare analysis showed that file-sharing significantly increased consumer welfare and the reduction of deadweight loss due to file-sharing doubled the loss of producer profits.

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

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

Ma et al. (2014) studied pre-release movie piracy and found that pre-release piracy caused a 19.1% decrease in revenue compared with piracy that occurred post-release. In a related paper, Lu et al. (2019) studied the substitution and WOM effects of piracy. They find that substitution effects dominate WOM effects for pre-release piracy and WOM effects dominate substitution effects for post-release piracy. While Ma et al. (2014) and Lu et al. (2019) do not emphasize the quality differentiation of movie piracy, I focus on the impact of high-quality piracy resulting from screener leaks. In addition, piracy availability is correlated with many endogenous factors, such as the international release gap; focusing on screener leaks provides cleaner identification because screener release is determined by the timing of film festivals and is less dependent on movie quality.

2.3 Data

2.3.1 Data Collection

Two datasets were employed in this paper. First, data on both weekly and total movie box office for 9,799 movies released from 2003 to 2016 were collected from the box office reporting website *BoxofficeMojo.com*. For each movie I also collected associated characteristics including genre, Motion Picture Association of America (MPAA) rating, studio, movie runtime, and budget from the *International Movie Database (IMDB)*. Movie rating data were obtained from *Rottentomatoes*, which took integer values from 1 to 100. I also collected information about film award wins and nominations for each movie from IMDB, particularly awards and nominations with respect to the Academy Awards (Oscars). Some time-varying variables, including opening screens, and total number of screens, were also collected from *BoxofficeMojo.com*.

The second dataset consisted of the dates of screener leaks for each Oscar nominee. The data were collected by scraping the piracy search engine *thepiratebay.org*. I choose *thepiratebay.org* because it was one of the most popular torrent sites in 2017[1], the website has a long history going back to 2003, which allowed me to trace historic leaks of earlier movies.[2] To collect the data on screener leaks, for each Oscar nominees between 2003 and 2016, I searched on *thepiratebay.org* with keywords including combinations of the movie name with one of three identifying keywords: **"screnner", "dvdscr", "scr"**.[3] I used an automated script to extract search results for each query. *thepiratebay.org* provides upload date for each torrent file in the search result, which allowed me to track the earliest date of piracy. For each movie I took the earliest upload date in the search result as the leak date of its screener.[4]

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

[2] 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.

[3] These keywords are the most common screener format indicators in a torrent file name.

[4] Baio (2016) also collect statistics on leaks date for Osacar nominees, I find the leak date collected from *thepiratebay.org* to be different from Baio (2016) for a significant fraction of movies. For movies with different leak dates in the two data, I take the earlier one. In addition, 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 (2016), I first used screener receipt dates reported by movie critic Ken Rudolph from his personal website (http://kenru.net/movies/) as the screener release date. The website records the date every movie screener was received by the critic from 2001 to 2017. I also used the earliest reviews posted

2.3.2 Descriptive Statistics

Table 2.1 shows how the number of leaks changed 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 having screener leaks before or during the first week of release. For those leaked movies, the average time between the U.S. release date and the average leak date was above 50 days. After 2009, the pre-release leaks happened less frequently, averaging 1 movie every year during the period 2010 to 2015. The average leak-release gap, however, has shrunk from about 50 days to 38 days. The incidence of leaks showed large variations over the years, which demonstrates the underlying randomness governing the success of obtaining screeners by pirates. In 2016, the number exploded because of the number of leaks by the group HIVE-CM8 mentioned earlier [5]. A large number of movies was leaked during their theatrical runs in December, six movies were leaked before their scheduled release dates and two movies during 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 2.1. Figure 2.1 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 potential links between screener piracy activities and theatrical market performances.

I first compared total box office revenue of Oscar nominees with the intensity of leaks for each year, using the measures mentioned in Table 2.1. Figure 2.1 shows the evolution of yearly aggregate box office revenue of all Oscar nominees and two measures of intensity of screener leaks: the number of movies leaked before or during the first week of release, and the average num-

in IMDB as an alternative check—for movies with different release dates in the two sets of data, I took the earlier one as the release date.

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

Figure 2.1: Pattern of Correlation between Box office and Leaks over Time

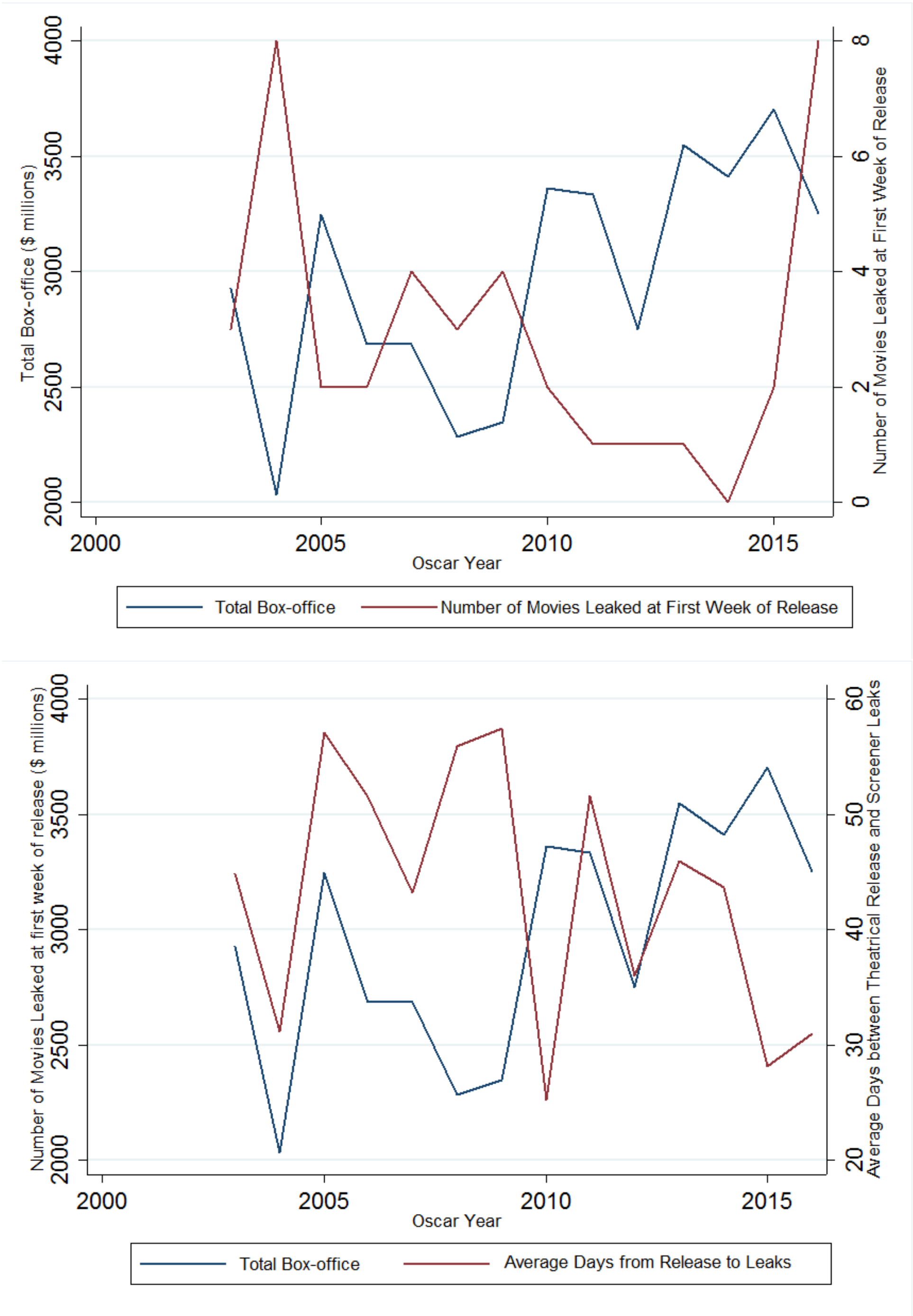


Table 2.1: Time Trend of Screener Leaks

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

ber of days between U.S. release and leaks. The time series patterns show that there is significant correlation between box office performance and leaks: box office revenue tends to be lower if there are fewer leaks and box office higher if on average leaks happen later. The correlation suggests potential links between screener piracy activities and theatrical market performance.

Table 2.2 reports summary statistics for four sets of samples. As a reference group, column (1) shows descriptive statistics for all movies released in the United States theatrical market during 2003 to 2016. There are 9,799 movies in the whole sample. On average one movie yields box office revenue of about 16.9 million dollars, with a huge dispersion represented by the standard deviation of 47.88 million. Similar dispersions exist among other variables such as the number of screens, first week box office. Budget data are only available for around 2,000 movies. Of those available, the average budget is 56.39 million dollars, with standard deviation of 82.75 million. About 4.55% of all movies are nominated for the Academy Award.

In this chapter, I focus on the sample of Academy Award (Oscar) nominees[6]. The subsample of Oscar nominees represents 4.5% of the total sam-

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

ple. As there is huge heterogeneity in movies, restricting the sample to Oscar nominees made the sample more comparable in terms of quality and market reception, reducing the potential endogeneity due to unobserved quality. The summary statistics are presented in column (2) of the table. Clearly Oscar nominees have more box office appeal because of the better quality, with average total box office revenue increase from 16.9 million dollars to 93.2 million dollars. These movies also invest more in budget (83.45 million dollars compared with 56.39 million dollars) and are assigned to more opening screens by theatres (2,019 compared with 718). If we look at the distribution of genres, drama and science fiction movies are overrepresented in the sample of Oscar nominees—their shares rise from 22.7% and 2.1% to 41.0% and 5.8%, respectively. Leaks happen in about 60% of the Oscar nominees, and about 11% of leaks happen prior to theatrical releases.

Column (3) reports summary statistics for the leaked Oscar movies and column (4) reports statistics for the non-leaked movies. To compare the observable movie characteristics for leaked and non-leaked movies, I conducted a balance test, discussed in the next section.

2.4 Pre-analysis Checks

Before proceeding to the main empirical analysis, I first conducted several tests on the exogeneity of leaks. A natural way to analyse my research question is to directly compare box office revenue between leaked and non-leaked movies—the underlying identification assumption is that leaks are independent of other movie-specific, unobservable confounding factors that also relate to box office revenue. The result of a cross-sectional comparison of box-office performance will be biased by any movie-specific unobservable factors affecting both demand and leaks. For example, it might be the case that leaks are affected by unobservable movie demand shocks. Movie pirates target the most popular movies, hence movies with higher box office appeal are more likely to suffer from leaks[7]. It is therefore important to check the validity of my identification assumption. Although there is no conclusive test of the identification

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

Table 2.2: Summary Statistics

	(1)		(2)		(3)		(4)	
	All Movies Sample		Oscar Nominees		Oscar Nominees with Leaks		Oscar Nominees without Leaks	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
Key Variables								
Number of Screens at Peak	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	

assumption, I can exploit the richness of my data to present some evidence supporting the exogeneity of screener leaks incidence.

2.4.1 Balance Test

To address these potential concerns, I did a balance test for observable characteristics between the leaked and unleaked samples. For each characteristic, I ran a t-test for equality of group means on major observable characteristics. The t-statistics are reported in Table 2.3. We would expect small t-statistics across most observables if the two samples are relatively homogeneous. Based on the result, for the most important variables, the mean of rating between two samples were not significantly different. For example, two samples had very similar ratings, with 79.42 for leaked movies and 78.31 for unleaked movies. On average, leaked movies and unleaked movies have similar number of screens at peak time (2,014 versus 2,028). As the total number of screens is strategically adjusted according to change in demand, this indicates that ex post market reception is similar for leaked and unleaked movies. In terms of scheduled screens during the first week, however, I did observe a significant difference: unleaked movies have more screens than leaked ones. The difference resulted from a higher number of leaked movies choosing limited release initially[8], which indicates that unleaked movies are expected to perform better in market ex ante. The results also show that for genres such as action, comedy, horror, and romance there is no significant difference between the two groups. However, I did observe that genres such as drama and crime are over-represented in the leaked group so it is important to control for them in the cross-section regression.

I then switched to comparisons between movies with pre-release leaks and other movies. There are 30 movies with pre-release leaks, making the sample size of the treatment group to be much smaller than the control group. The differences in most covariates are more pronounced—movies with pre-release leaks tend to have smaller release, with fewer opening screens scheduled and fewer screens at peak. The difference in genre distribution is less significant. Similar to the previous result, drama movies are over-represented in the treat-

[8] Limited-released movies usually are first released in theatres in major metropolitan areas such as New York and Los Angeles. After gauging their market appeal, some movies will then be released nationwide.

Table 2.3: Test of Baseline Balance

	Treatment: Screener Leaks				Treatment: Screener Leaks before Theatrical Release			
	Treatment	Control	Control-Treatment	(t-statistics)	Treatment	Control	Control-Treatment	(t-statistics)
t-tests Between Group Means								
Key Covariates								
Rating	79.42	78.31	-1.12	(-0.63)	81.37	78.81	-2.56	(-0.75)
Number of Screens at Peak	2014.28	2028.34	14.06	(0.10)	1090.37	2086.93	996.56***	(3.76)
Scheduled screens for first week	1309.15	1703.31	394.16*	(2.44)	345.37	1546.55	1201.19***	(3.84)
Budget	77.79	93.46	15.67	(1.49)	46.16	85.83	39.68	(1.88)
Genres								
Action	0.06	0.11	0.05	(1.81)	0.03	0.08	0.05	(0.95)
Drama	0.46	0.32	-0.14***	(-2.97)	0.60	0.40	-0.20*	(-2.19)
Animation	0.09	0.19	0.10***	(3.16)	0.03	0.14	0.11	(1.68)
Comedy	0.14	0.09	-0.46	(-1.46)	0.10	0.13	0.03	(0.40)
Crime	0.05	0.01	-0.04**	(-2.69)	0.03	0.03	0.00	(0.01)
Horror/Thriller	0.09	0.08	-0.01	(-0.19)	0.17	0.08	-0.08	(-1.59)
Romance	0.02	0.03	0.01	(0.37)	0.00	0.03	0.03	(0.90)
Science Fiction	0.04	0.09	0.05*	(2.33)	0.03	0.06	0.03	(0.60)
Documentary	0.01	0.02	0.01	(0.92)	0.00	0.01	0.01	(0.60)
Observations	268	178			30	416		
Joint Test of orthogonality								
F-statistics:			1.15				1.72	

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.

ment group.

I also conducted a joint test of orthogonality to check if differences in key covariates significantly influenced the assignment of treatment. I ran a linear regression of the dummy variable of leaks as the dependent variable and all the covariates excluding market outcome variables as regressors. The result of the F-statistics is shown in Table 2.3. I can not reject the hypothesis that all coefficients are 0.

2.5 Cross-sectional Evidence

Perhaps the most straightforward approach to this paper's question regarding the impact of piracy on legitimate movie revenue is to directly compare total sales of leaked and unleaked movies. I therefore start with a cross-sectional comparison using total box office revenue. I also exploit the timing of leaks by comparing movies with early leaks and those with late leaks.

2.5.1 Leaked vs. Non-leaked

First, I conduct a simple comparison of total box office revenue of leaked and unleaked movies. One challenge is the comparability between leaked and unleaked groups. Although the balance test described in the previous section suggested 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. I use the following specification:

$$Y_i = \alpha + \beta Leak_i + X_i'\gamma + \sum_{k=2003}^{2016} \lambda_k \mathbb{1}(RY_i = k) + \sum_{j=1}^{11} \tau_j \mathbb{1}(RM_i = j) + \epsilon_i \quad (2.1)$$

On the left-hand side, $Leak_{it}$ is a dummy variable equal to 1 if the movie's screener is leaked during its run in theatres. X_{it} is a set of controls. As movie fixed effects are unavailable, I add a rich set of controls, including the total number of screens, the length of theatrical run, Rotten Tomatoes Rating, and dummies for genre, MPAA rating, and major studios. RY_i denotes the release year of movie i, $\mathbb{1}(RY_i = k)$ is a dummy variable equal to 1 if movie i was released in year k. k ranges from 2003 to 2016. Similarly, RM_i denotes the release month of movie i, $\mathbb{1}(RM_i = j)$ is a dummy variable equal to 1 if movie i was released in month j ranging from 1 to 11.

For the dependent variable Y_i, I used two sets of outcome variables: total movie box office revenue and surprise sale. Following Moretti (2011), I used the number of opening theatres as a proxy for expected demand and defined 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 dummies for month and year). Specifically, for each movie i released at specific time t, I regressed its opening screens $OpenScreen_i$ and release month and year dummies on its box office R_i.

$$R_i = \alpha + \beta OpenScreen_i + \sum_{k=2003}^{2016} \lambda_k \mathbb{1}(RY_i = k) + \sum_{j=1}^{11} \tau_j \mathbb{1}(RM_i = j) + \epsilon_i \quad (2.2)$$

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

$$Surprise_i = R_i - \hat{R}_i \quad (2.3)$$

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 unleaked movies as the control group and test if the outcome variables is significantly different between these two groups. If movie piracy indeed severely cannibalizes sales, I would expect the average sale of leaked movies to be much lower than that of the unleaked movies, given the assumption that screener leaks do not depend on unobservable heterogeneity that drives sales. In addition to the baseline specifications, I also try using another definition of treatment: I define treatment as “Leaked prior to Release”, which represent a more intense treatment.

The result of the cross-sectional regression is reported in Table 2.4. Column (1) and (2) show the result using total box office revenue as the dependent variable, and column (3) and (4) show the result using surprise sale as the dependent variable. Beginning with the baseline comparison between leaked and unleaked movies in column (1) and (3), I did not find strong evidence that leaked movies perform worse than unleaked movies. The coefficient has a negative sign; the magnitude indicates that on average leaked movies did worse by \$0.36 million for total box office revenue and \$0.08 million for surprise sale. The coefficients are of little statistical significance, however. The results are not surprising as for a large fraction of leaked movies, leaks happen relatively late in their theatrical runs. Because movie revenues are concentrated in the beginning of the theatrical run, even if screener piracy causes severe harm to subsequent box office revenue, the total impact can still be moderate if the screener is leaked very late.

For further analysis, I switch to pre-release leaks as the treatment group.

Table 2.4: Cross-sectional Results

	(1)	(2)	(3)	(4)
Dependent Variables:	Total Box office	Total Box office	Surprise Sale	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

Note: Standard errors in parentheses

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

The results for total box office revenue and surprise sale are reported in column (2) and (4), respectively. Once I restrict 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 average, movies with screeners leaked prior to release did worse by \$ 3.202 million for total box office revenue and \$ 2.99 million for surprise sale. In general the results reveal that the impact of leaks is negative but lacks statistical significance. It also confirms that the impact is more pronounced if leaks happen earlier, suggesting that the intensity margin of leaks might be important as well.

2.5.2 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 during theatrical runs as the intensity of treatment. If leaks do harm sales, I would observe that after controlling for quality, movies leaked earlier on average yield lower box office revenue on average than movies leaked late.

For implementation, I constructed a variable LW_i that denoted the number of weeks between movie i's U.S. release date and its screener leak date.

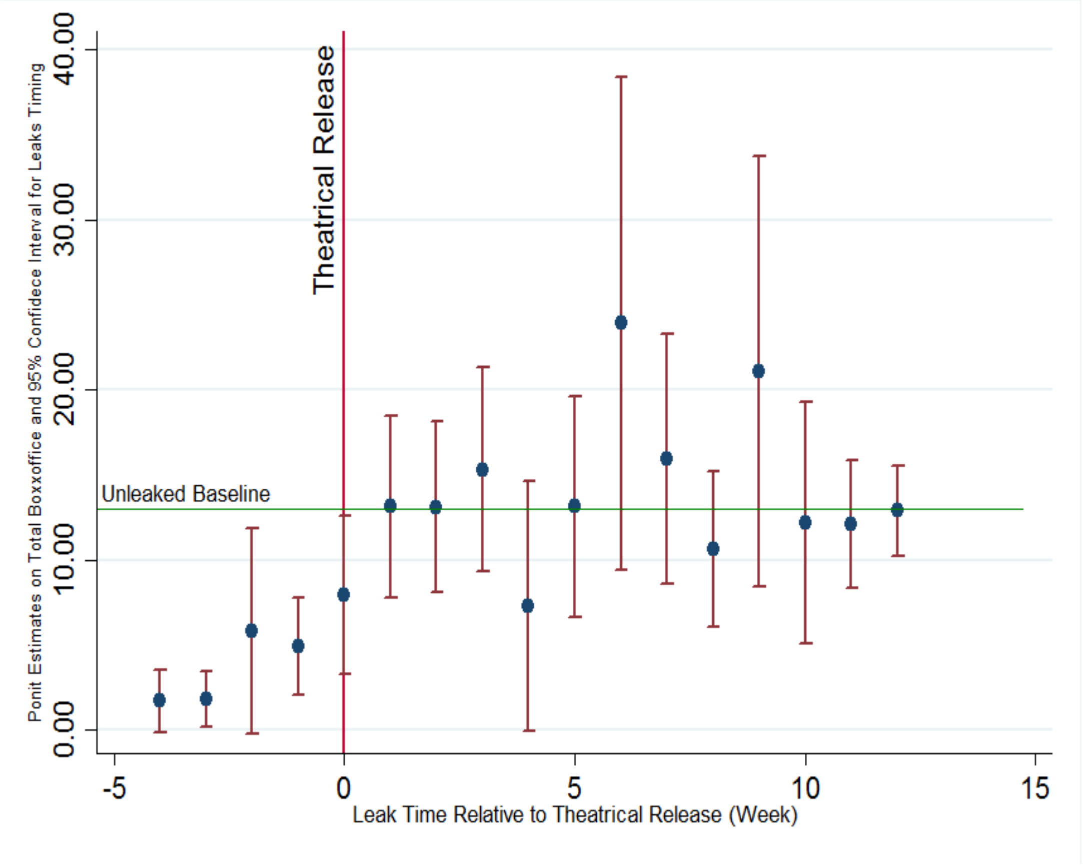
Noting that the marginal effect might vary depending on different treatment levels, I dropped the uniform marginal effect assumption and transformed the treatment variable into a series of dummies by different values of LW_i .

$$Y_i = \sum_{j=-4}^{15} \beta_j \mathbb{1}(LW_i = m) + X_i'\gamma + \sum_{k=2003}^{2016} \lambda_k \mathbb{1}(RY_i = k) + \sum_{j=1}^{11} \tau_j \mathbb{1}(RM_i = j) + \epsilon_i \quad (2.4)$$

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

Figure 2.2 reports the point estimates and 95% confidence interval on β_j the effects of leak timing on movie outcomes. Because I did not include an intercept in the specification, the point estimate corresponding to each leak week dummy can be treated as the average box office revenue in that leak week category after controlling for release time and year. There was a significant upward

Figure 2.2: Cross-Sectional Comparison: Effects of Leak Timing on Total Box-office



trend in the early stage, especially before theatrical release. This was consistent with the belief that early leaks before theatrical release attract a significant number of consumers with high willingness-to-pay, especially when there is a lack of legal channels for viewing the movie , and therefore the harm from pre-release piracy should be the highest. For subsequent leaks after the theatrical release, I find no compelling evidence of harm to box office revenue, as most point estimates are statistically indistinguishable from the unleaked baseline. This suggests that the harm to box office performance is significantly moderated by the late occurrence of leaks. Because box office revenue for a movie is not uniformly distributed over time in that a large fraction of box office revenue comes from the first few weeks in the theatrical window, the harm from leaks is expected to decrease in leak occurrence time.

2.6 Difference-in-Difference Analysis

2.6.1 Empirical Specifications

Complementary to the cross-sectional analysis, I adopt a difference-in-difference (DD) strategy using weekly box office data. 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 is not correlated with screener leaks.

This study's main identification strategy involved implementing a DD strategy using panels of weekly box office data. Screener leaks incidence generates 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 the same calendar time and release time.[9]

The DD specification take the following forms:

$$ln(Sale_{it}) = \alpha Leak_{it} + X_{it}\beta + \sum_{j} \pi_j \mathbb{1}\{\tau_{it} = j\} + \xi_i + \lambda_t + \varepsilon_{it} \tag{2.5}$$

On the left-hand side, the dependent variable is $ln(Sale_{it})$, the log of weekly box office revenue for movie i at time t. On the right-hand side of the regression I included movie fixed effects ξ_i, and calender time fixed effects λ_t. In addition, because box office revenue has a natural exponential declining pattern, I need to control for weeks after release. As such decline 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 variable, which equals 1 if movie

[9] 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 revenue 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 revenue and downloads both increase when summer holiday begins).

i's screener has already leaked at time t. ε_{it} is the idiosyncratic error term. I clustered the standard error at the movie level, allowing for serial correlation of error within movies.

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

2.6.2 Results

The results of the panel regressions using weekly data are reported in Table 2.5. As a starting point I first present the simple OLS results at first two columns. Column (1) shows estimates for the simplest OLS specification without 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 in 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 decreases the weekly box office revenue by 12.6%[10] . In comparison to the previous OLS estimates, it suggests that adding controls for quality help reduce a significant amount of upward bias.

Now we turn to the DD estimates. In columns (3) and (4), I further add movie fixed effects, calendar week fixed effects, and week of release fixed effects to control for additional sources of endogeneity, for reasons discussed in the previous section. After differencing out the time-invariant movie-specific component, the 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 compar-

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

ison highlights again the importance of movie heterogeneity and time-varying demand effects in influencing the estimate. The result in column (3) indicates that the occurrence of screener piracy lowers weekly box office by 29.8%.

To test the identification assumption on the 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 estimated the following specification:

$$ln(Sale_{it}) = \sum_{k=-2}^{2} \alpha_k \mathbb{1}\left\{t = T_i + k\right\} + X_{it}\beta + \sum_{j} \pi_j \mathbb{1}\left\{\tau_{it} = j\right\} + \xi_i + \lambda_t + \varepsilon_{it} \quad (2.6)$$

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

I report the results of the 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. All treatment effects after the leak are of the same expected signs and are mostly statistically significant. I also observe that the magnitude of the treatment effect at the leak ($t = T_i$) becomes smaller after the inclusion of leads and lags, and higher for treatment effects of 2 weeks lags ($t = T_i$+2). This observation suggests that treatment effects of leaks to movie box office revenue accumulate and grow over subsequent weeks, potentially due to the fact that more consumers have discovered the leak over time.

2.6.3 Heterogeneous Effects by Rating

After the baseline specification, another natural question to ask is whether there exists any effect heterogeneity, especially how the harm caused by screener leaks differs by movie rating. Were good movies hurt more than bad movies or the opposite? To examine the effect heterogeneity for rating, I added interaction terms between leaks and rating. If heterogeneity is significant, we would expect the coefficient to be statistically significant. As the result in column

Table 2.5: 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, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(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, a 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 at the box office in subsequent weeks after leaks, while a hypothetical movie rated as low as 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 highlights the positive role of piracy from various channels including, for example, sampling effect, network effect, and WOM effect (Liebowitz (1985), Peitz and Waelbroeck (2006), Belleflamme and Peitz (2014), Peukert et al. (2017),Lu et al. (2019)). It suggests that higher-rated movies better capitalize on the sampling feedback or positive WOM from piracy. The increase in sales neutralizes piracy's negative cannibalization, yielding a lower estimate of negative effect as shown in the result.

2.7 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 study I used screener leaks as a natural experiment for identification. Screeners are copies of movies sent to critics and industry professionals for evaluation purposes. Sometimes screeners get leaked accidentally and made available for download on the Internet. These incidences of leaks 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 to 2016, I estimated the impact of movie piracy on box office revenue exploiting the plausibly exogenous 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 the majority of screener piracy occurs at a late stage of theatre runs, piracy therefore only affecting demand at the residuals. Second, evidence suggests that the effects of leaks "spillover" to other, unleaked, movies. I found significant negative indirect displacement effects among other movies: one additional contemporaneous leak led to a 3% decrease of box office revenue of other unleaked movies. The magnitude of the indirect displacement effect is stronger for movies that are closer substitutes. The presence of indirect displacement implies negative externalities of piracy. The total cost of an additional piracy leak will be higher for the total industry revenue than for a particular leaked movie.

The result suggests that it is important to deter emergence of high-quality piracy during the early stage of release because of their potentially detrimental effects on sales. It should also be noted that piracy is of differentiated quality—the effects could be different by quality. As screener leaks are high-quality piracy compared 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.

Chapter 3

Decomposing the Intra-movie and Inter-movie Substitution Effects of Movie Piracy

3.1 Introduction

The empirical literature on digital piracy has focused on the magnitude of the substitution effects of piracy on sales. One question that remained unanswered is the source of substitution. Take the motion pictures industry as an example: substitution on sales can come from either the piracy of the same movie title or from piracy of different movie titles. A large body of the literature on movie piracy has focused exclusively on the direct intra-movie substitution effect of piracy, attempting to measure the response of one movie's sales revenue to its own piracy. How piracy of one movie affects other movies' revenue is mostly ignored in the literature.

Several features of the movie market suggest that the substitution pattern of piracy might be quite complicated. First, the primary sources of high-quality piracy usually come from ripping screener or physical home-video copies like DVD or Blu-ray, and most downloads appear at the later stage of the theatrical run, which accounts for a relatively small fraction of movie revenue. The lack of interaction in time weakens the intra-movie or direct substitution effect. Second, fierce competition exists between similar movies. This competition can arise due to constraint in consumers' time, attention, and budget. There is a great deal of evidence suggesting that similar movies with close release windows significantly cannibalize each other's sales (Einav, 2010, Dhar and Weinberg, 2016). Such competition could also exist across channels. In reality, piracy of a product may not only displace the revenue of its own legitimate sales but also have a "spillover" effect on the revenue of the products that are close substitutes.

Such spillover or inter-movie substitution effects of piracy are important to our understanding and evaluation of piracy's impact on the motion picture industry. Why do we care about the inter-movie substitution effect of piracy? The presence of "spillovers" implies that the damage caused by a movie's piracy with regard to total industry revenue will be higher than the damage to its own revenue. Research only measuring the direct intra-movie substitution effect will potentially underestimate the total harm of piracy to the industry. Despite their importance, these spillover effects have not yet been systematically quantified in the literature.

The goal of this chapter is to answer these questions of interests. First, this chapter quantifies the effect of piracy on movie box office revenue by estimating a random coefficient demand model of movies using a novel dataset I collected on actual downloads on BitTorrent. Second, using the estimated model, I decompose the inter-movie and intra-movie substitution effect by counterfactual experiment.

The main contributions of this chapter are as follows. First, for movies released during a 40-week sampling period in 2015, I collect 40,267 relevant movie torrent files via major torrent search engines. Using the collected torrents, I construct a dataset of weekly movie downloads for movies in my sampling period. Due to the lack of data on actual downloads, previous research on file sharing has mainly explored various proxies and quasi-experiments on piracy to study the effects of file sharing. Data limitations may hamper the identification of the true effects of file sharing because of issues related to measurement error and data representativeness. This chapter avoids these concerns using direct information on downloads of pirated products.

Second, this chapter contributes to the empirical literature on movie piracy

(Rob and Waldfogel, 2007; Smith and Telang, 2010; Danaher and Smith, 2014; Danaher and Waldfogel, 2012; Peukert et al., 2017; Lu et al., 2019) by first applying aggregate illegal download data on BitTorrent to estimate a random-coefficient logit demand model for movie piracy. The use of a random-coefficient demand models brings several benefits. (1) linear reduced-form estimations employed in previous literature focus on exclusively intra-movie effect of piracy (how movie A's piracy affect its own revenue), while ignoring the inter-movie substitution (how movie A's piracy compete and cannibalize other movie's revenue).[1] One advantage of logit demand models is that they naturally incorporate these competition effects between movies and therefore can be used to study the inter-movie substitution effects of piracy. (2) As the reliability of estimates of industry loss relies heavily on accurately identifying the true substitution patterns, random-coefficient logit demand models that allow for flexible substitution patterns will generate a more accurate estimate on the effects of piracy compared with the multinomial logit models. The rich set of variation in the choice set due to movie theatrical release and exit along with leaks of piracy provides a sufficient source of variation for identification. (3) I can use counterfactual experiments to test the efficacy of various anti-piracy policies. With data on aggregate downloads, I can also quantify industry loss due to piracy, which is difficult to obtain from reduced-form studies. The industry-wide loss estimate based on the flexible demand model will be useful to test and improve estimates from the previous widely cited industry studies. (4) Estimation of a demand model allows the calculation of consumer welfare; therefore, I can assess the welfare effect of file sharing and digital piracy.

This chapter's findings are as follows. First, using the estimated demand model, I decompose the intra-movie and inter-movie substitution effects of piracy. I find that the revenue loss due to piracy's inter-movie substitution is on average 4 times as large as the intra-movie substitution. This is quite surprising as it implies that considering only intra-movie substitution will seriously underestimate piracy's damage. Second, my damage estimates show that piracy reduces the total box-office revenue of the motion pictures industry by $ 93 million in total, 1.09% of the current box-office. The estimates are smaller than widely cited industry estimates constantly referenced in policy making.

[1] It's very difficult to incorporate competition effects using a linear reduced-form regression framework as it requires including a full sets of each movies' sales and piracy on the right-hand side of the equation.

A "naive" methodology, which assumes full sale displacement of piracy, will inflate the true cost of piracy by 13.6 times.

The rest of this chapter is organized as follows. Section 3.2 provides a review of relevant literature. Section 3.3 provides background information on the motion picture industry and file-sharing. Section 3.4 describes the data and section 3.5 shows some preliminary evidence. Section 3.6 presents the model. The estimation procedure and results are presented in Section 3.7. Section 3.8 gives the results of counterfactual experiments, and Section 3.9 concludes the chapter.

3.2 Literature Review

This paper adds to several strands of literature. First, this paper is related to the empirical literature on digital piracy and file sharing. Identifying the effects of digital piracy on the sales of digital products is an empirically challenging question because of issues related to data limitations and the endogeneity of downloads. The displacement effects of file sharing on sales have been widely studied in the literature. Generally researchers agree that piracy has a nontrivial displacement effect on sales as the majority of papers find significant negative effects on sales of both music (Liebowitz, 2004; Zentner, 2006; Rob and Waldfogel, 2004; Hong, 2013) and movies (Danaher and Waldfogel, 2012; Rob and Waldfogel, 2007; De Vany and Walls, 2007; Bai and Waldfogel, 2012; Ma et al., 2014; Lu et al., 2019), but the estimated magnitude of the displacement effects differ substantially across papers. There are also a number of papers finding moderate and insignificant negative effects, or even positive effects (Oberholzer-Gee and Strumpf, 2007; Hammond, 2014; Aguiar and Martens, 2016).

One reason behind the disparity of these empirical results is data limitations. Due to the difficulty of observing actual downloads, researchers have came up with different ways to overcome this empirical issue. Judging by their methodologies, most research on file sharing can be classified into three categories. First, many researchers employ various proxies, such as geographic variation in the Internet penetration rate, broadband connection rate (Liebowitz, 2004, 2006; Zentner, 2006). Second, some papers take advantage of quasi-experiments, such as development of file sharing technology, the sudden close of file sharing websites, or variation in international movie release windows (Hong, 2013; Danaher and Waldfogel, 2012; Danaher and Smith, 2014; Peukert et al., 2017). Third, the many use survey data collected from groups of consumers (Rob and Waldfogel, 2004, 2007; Bai and Waldfogel, 2012; Leung, 2015).

Each of these research methods has merits, but in absence of data on actual file sharing activities, questions may arise, such as to what degree these proxies and quasi-experiments can capture the true variation of file sharing activities, and to what extent the consumers sampled in the survey are representative of the true population. Having data on actual downloads can be a good complement to these studies. A few studies on music piracy have utilized actual file sharing download data (Oberholzer-Gee and Strumpf, 2007; Hammond, 2014; Aguiar and Martens, 2016). The data used in these studies include data on Napster, data from private BitTorrent trackers, and clickstream data. Most of them find no significant effect or a very moderate negative effect. The file sharing data used in above-mentioned papers is exclusively about music piracy. In comparison, I used file-sharing data on movies from a more recent period in 2015 in this paper. There are substantial differences between music piracy and movie piracy: pirated MP3 music are generally of the same quality as legal purchase, but there are significant differences in quality among movie piracy. The sequential introduction of movies into different channels of sales (box office and DVD) and their effects on the availability of piracy make the issue more complicated than music piracy. In addition, the landscape of file sharing has changed dramatically, using data from more recent period is especially more relevant in the context. Lastly, instead of using data from one tracker, I attempt to estimate the aggregate download using data obtained from a more comprehensive list of 84 popular public BitTorrent trackers.

Methodologically this paper is closely related to Leung (2013), who also structurally estimated a random-coefficient logit model to study software piracy using a conjoint survey of 281 college students. My papers are different in several aspects. Leung (2015) studies the software industry, and this paper focuses on the motion picture industry. While Leung (2013) focused on substitution patterns under a single product setting, this paper instead focuses on estimating the total cost of piracy at the industry level, taking into considerations the substitution between different movie titles.

In addition to the empirical literature on file sharing, this paper is also related to the growing literature on the motion picture industry. Researchers have studied different aspects of the motion picture industry, for example: movie word-of-mouth (Chintagunta et al., 2010; Moul, 2007; Moretti, 2011; Gilchrist and Sands, 2016b), seasonality in the motion picture industry (Einav, 2007), uniform pricing practices (Orbach and Einav, 2007), estimation of price elasticity (Davis, 2002; De Roos and McKenzie, 2014), vertical integration (Gil, 2008, Gil, 2010), effects of uncertainty in the movie industry (De Vany and Walls, 1999; Elberse and Eliashberg, 2003), spatial competition in movie theatres (Davis, 2006), and strategic entry and exit decisions of studios and theatres (Einav, 2010; Takahashi, 2015; Dalton and Leung, 2017). This paper adds to the literature on the effect of file sharing on the motion picture industry.

The third strand of related literature is the broad literature on intellectual property, especially copyright. The emergence of file sharing may require governments to adjust the existing strength of copyright protection accordingly. However, there is no consensus on the optimal degree of intellectual property protection. On the one side, as Boldrin and Levine (2002) point out, strong property rights not only include the right to own and sell ideas but also the right to regulate their use after sale, which will create a socially inefficient intellectual monopoly. On the other side, Klein et al. (2002) argue that file-sharing restricts the ability of copyright holders to exercise price discrimination and effectively control price, so file sharing services are likely to reduce the value of copyrighted work. They argue that the use of strong property rights to restrict piracy should be implemented even if there is a substantial cost of restricting the consumer's "fair use." Empirical evidence on the effects of file sharing will provide useful insight on the debate on optimal copyright protection.

3.3 Background: File sharing and BitTorrent

Peer-to-peer (P2P) file sharing is a decentralized file-transfer technology. In traditional downloading methods, files are downloaded from centralized servers which store the source file. Because of the limited bandwidth, download speed deteriorates as the number of clients requesting services from the server increases. For P2P file sharing, clients download the file from other clients who are also downloading the file or those who have already downloaded the file. P2P file sharing efficiently utilizes the upload bandwidth of clients to facilitate downloading, therefore it successfully overcomes the bandwidth bottleneck of centralized servers and significantly increases download speed. Due to these advantages, P2P file sharing has quickly gained popularity among Internet users.

The history of file sharing dates back to 1999. An American computer programmer named Shawn Fanning developed a peer-to-peer file sharing platform called Napster. Napster was used to share music files among users and it quickly became popular among Internet users. At its peak in 2001, Napster had about 80 million registered users all over the world. In July 2001, Napster was involved in a series of copyright lawsuits and was forced to shut down by US court. After the shutdown of Napster, subsequent file sharing services have been developed including Gnutella, Freenet, Kazaa, FastTrack, E-Mule, and so on. Among those followers, BitTorrent has become the dominant file sharing service, accounting for on average 40% of Internet upstream traffic according to broadband management company Sandvine.[2] Most files transfered in BitTorrent are media files like movies, TV shows and music, and most of these files are pirated. According to research conducted by RIAA, Bit Torrent may account for about 70% of piracy activities around the world.

Due to the dominance of BitTorrent over other file sharing platforms, I focus on BitTorrent in the study of file sharing in this paper. As a dominant protocol for file sharing and on-line piracy, BitTorrent is generally representative of the population of file-sharers and pirates. Although BitTorrent is not the only P2P file sharing service, behavior of file-sharers are not systematically different across different platforms (Oberholzer-Gee and Strumpf, 2007). Second, even if there are difference across platform, BitTorrent is so dominating nowadays that the share of other substitutes is negligible. According to the research of Ipoque[3], by 2011 the traffic of the second most popular file sharing tool E-Donkey was only 2.6%. [4]

[2] TorrentFreak: https://torrentfreak.com/bittorrent-still-dominates-global-internet-traffic-101026/

[3] TorrentFreak, https://torrentfreak.com/p2p-traffic-still-booming-071128/

[4] In later years a significant fraction of piracy activities have shifted from BitTorrent to the use of on-line illegal streaming service. Many popular illegal movie streaming websites such as *Popcorntime* is powered by BitTorrent

3.4 Data

3.4.1 Data Description

In this paper, I combine data from several data sources. I assemble a dataset including weekly information on movie-level box office sales, downloads, and sales on DVD in the United States. I collected data covering a 40-week period from March 27 to December 27 in 2015. The data contain weekly downloads on BitTorrent in the United States collected from 40,267 torrent files for 255 movies released between March 27 and December 27, 2015. The data were collected using computer science techniques following several studies on BitTorrent (Erman, 2005; Layton and Watters, 2010). The details of the data collection method are presented in Appendix B.3.

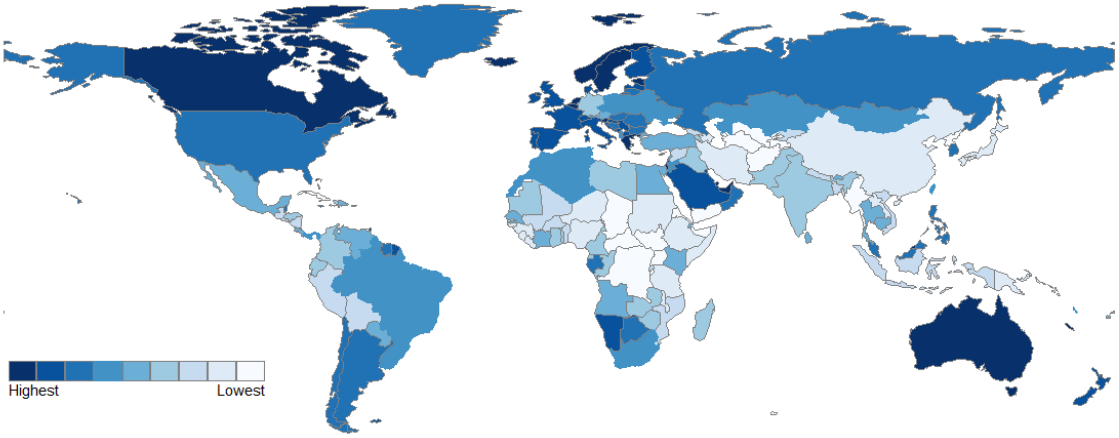
The box office data were collected from box office reporting service website *Boxofficemojo.com*. I collected information on weekly box office and movie characteristics for all movies showing during the sampling period. I included characteristics such as movie ratings, sequel, genres, MPAA rating (G/PG/PG-13/R), and weeks after release, which are commonly used in studies on the motion picture industry. Movie rating data were collected from Internet Movie Database(imdb.com). Because the uniform pricing practice in movie theaters and movie price data are hard to collect, only country-level average admission price in the United States was obtained[5]. The box office revenues of some independent movies are extremely small so that their market shares are indistinguishable from zero. Inclusion of these "zero" market share movies would introduce numerical problems to the estimation procedure so I dropped all observations with market share smaller than 0.01% in my sample.

I matched the box office data with the collected file-sharing data. Table 3.1 provides statistics about the top downloaded movies and top selling movies. The top downloaded movies are generally blockbuster movies with big budgets and massive advertising campaigns. Most best-seller movies also appear to be

and are therefore included in my data. Streaming through file-hosting service such as *Openload* is however not included in my data. As a consequence, using only data on BitTorrent underestimates piracy activities.

[5] Admission price variation is very small. Although prices are different across screen types(IMAX/3D/Ordinary), these variations are perfectly correlated with movie characteristics and therefore offer little identification power for the price elasticity. There is also price discrimination on different age group and selected days in a week (e.g. Cheap Tuesday), but actual data on admission by type and price are hard to obtain, so I did not attempt to estimate price elasticity in this study.

Figure 3.1: file sharing Activities in the World



Notes: Darker colors denote higher numbers of file-sharers adjusted by country population. Frequency of file sharing activities in each country is based on a sample of 1,698,846 movie downloaders' IP addresses that I collected from public BitTorrent trackers during a 5 day period. The geographic information of IP addresses are obtained using Maxmind's geoip database.

the most downloaded.

3.4.2 Descriptive Statistics

Figure 3.1 shows the intensity of file sharing activities worldwide. The intensity is measured by the number of file sharers I found in the sample period adjusted by country population. File sharing is indeed penetrating almost every place in the world. Out of 177 countries and regions in the study, file sharing activities are found in 170 countries. In terms of the total number of file sharers, the United States has the largest number of file sharers, comprising 13.7% of the total number. Other followers include Russia (6.3%) and France (5.4%). Not surprisingly, file sharing activities are positively correlated with the gross domestic product (GDP) per capita and population size,[6] but they are only mildly correlated with Internet speed.[7]

Now I describe the summary statistics. Because of the panel structure of my data, the basic observation is at the product-time level. For any movie there are potentially two channels for a consumer to get access (theater/piracy). A product is therefore defined as a movie and channel pair.

[6] The correlation coefficient between GDP and file sharing is 0.7649, and the correlation coefficient between population and file sharing is 0.3262.

[7] The correlation coefficient between Internet speed and file sharing is 0.082. Due to data limitations I am only able to collect the average Internet speed for 59 countries. Since most of the countries with low Internet speed are not presented in the data, this selection problem may explain the low correlation found between Internet speed and file sharing activities.

Table 3.1: Top-sellers and top downloaded movies

Top Selling Movies	
Title	Admission(million)
Jurassic World	184.09
Furious 7	167.97
Avengers: Age of Ultron	155.83
Minions	120.15
The Hobbit: The Battle of the Five Armies	106.22
Inside Out	84.63
The Hunger Games: Mockingjay Part 1	83.57
Interstellar	75.00
Big Hero 6	73.09
Mission: Impossible - Rogue Nation	72.94
Top Downloaded Movies	
Title	Download(million)
Furious 7	35.85
Interstellar	35.08
Fifty Shades of Grey	30.54
Kingsman: The Secret Service	27.18
Big Hero 6	23.41
The Hobbit: The Battle of the Five Armies	21.65
American Sniper	21.28
Avengers: Age of Ultron	18.84
Taken 3	18.57
Jupiter Ascending	16.02

Notes: Box office and download data are up to September 11th, 2015. Box office and downloads are all global numbers.

Table 3.2: Summary Statistics on Movie Characteristics

	Mean	Std.Dev	Min	Max
Pirated	.539	.498	0	1
Rating	6.426	1.554	4	9.3
Genre				
Action	.141	.348	0	1
Comedy	.179	.384	0	1
Drama	.219	.414	0	1
Horror	.086	.281	0	1
Cartoon	.066	.248	0	1
Foreign	.043	.204	0	1
Science Fiction	.061	.240	0	1
MPPA Rating				
PG	.160	.367	0	1
PG13	.347	.476	0	1
R	.363	.481	0	1
Market Share				
Sale(%)	.077	.405		
Illegal Downloads(%)	.010	.020		
Observations	6877			

Note: The summary statistics are at product-time level. Rating are of a scale of 0-10. Pirated is a dummy variable which equals 1 if the movie have pirated version available online. Sale and Downloads in movie characteristics section are measured in units. Action, Animation, Comedy, Drama, Horror, Science Fiction, PG, PG13, R are all genre and MPAA Rating dummy variables. In the market share section, the market share is an average of one movie's market shares in all weeks.

Table 4.4 provides sample descriptive statistics for the data at the products level. About 54% of the products are pirated. The bottom panel of Table 4.4 provides information on the average product market shares by channels. The average product market share of the ticket sales of a movie is about 0.077%, piracy downloads on average take up 0.01% of market share. Taken together, the illegal downloads account for less than 10% of all movie-watching activities. The three most common genres are drama (21.9%), comedy (17.9%), and action (14.1%). Ratings on IMDB are on a scale of 0 to 10, the distribution of rating has smaller dispersions, with an average of 6.43 and a standard deviation of 1.56.

3.5 Inter-Movie Substitution Effects of Piracy

As one kind of "experience" good, movies exhibit short product life cycles. Consumers have strong preferences for new movies, and most advertising budgets are spent in the first few weeks after the release date. Therefore, movie sales concentrate at the beginning of the release. The common showing period of a movie is about 6 to 10 weeks. For blockbuster movies, the box office revenue of the opening week usually accounts for about 20% of the total box office revenue. Figure 4.2 shows the pattern of the average weekly audience (in millions) compared to other channels, including DVD sales and downloads (in hundreds of thousands) by the number of weeks after the initial release. Weekly audience attendance in the theatre decays exponentially, quickly dropping to almost zero at around 10 weeks after the initial release date.

Downloads of pirated movies exhibit a very different pattern. They start low and peak after around the 10th to 20th week after release, the trend is less smooth partly because the new supply of better quality torrents in the later periods. Some important facts to highlight are that: (1) Qualitywise, early piracy is hardly comparable to the legitimate theatrical product[8]. (2) For many movies, a big part of their downloads happen at the end of the theatrical run, where availability of movie in theatres become limited. Figure 3.3 breaks total downloads into three part: downloads during wide theatrical availability period (when studio arrange more than 1000 screens nationwide), downloads during limited theatrical availability period (0-1000 screens) and downloads after theatrical exit. The figure shows top 50 movies in terms of aggregate downloads.[9] As the figure shows, despite some heterogeneity in the composition, most movies' downloads happen during the period with limited or no theatrical availability. For many downloaders of piracy, illegal versions are the only options available to them. Even if there were no piracy, the limited theatrical availability restrict

[8] During the first few weeks after release, most available pirated movies are the "CAM" version with very low quality copy made in a cinema using a camcorder or mobile phone by audience, they are hardly comparable with the quality of normal movies in theater. Around 5-10 weeks after release, many better quality "TC" version which are usually copy produced by transfering the movie from its analog reel to digital format. pirated movies come out and downloads start to increase. Download usually peaks at some time between 10-20 weeks after theater release when the "DVDRip/BlurayRip" version pirated movies become available due to the movie's DVD/Bluray release. At this moment, movies' theatrical windows have closed for a long time.

[9] To make it easy to compare, I restrict movies to have at least 40 weeks in theatre by the end of my sampling period.

copyright holder's ability to recover their profits. It mitigates the potential harm caused by intra-movie substitution.

Based on these facts, it is hard to believe that direct intra-movie substitution by piracy is the only important source of damages to studio revenue. If competition and substitution between titles are cross-channel and strong enough, the data pattern suggests the possibility of sizable spillover or inter-movie substitution effects, which could even dominate the direct intra-movie substitution effect. A simple decomposition might be useful to illustrate this point:

Let R be the industry revenue, which can be expressed as the sum of revenues of all individual movies R_j. The effect of movie j's piracy D_j on total industry revenue can be decomposed into two parts: one is the inter-movie effect $\frac{\partial R_j}{\partial D_j}$, which measures how movie j's piracy affects movie j's own revenue; the second is an inter-movie effect $\frac{\partial \sum_{j' \neq j} R_{j'}}{\partial D_j}$ which measures how movie j's piracy affects other movie's revenue.

In the previous literature, empirical research has exclusively focused on the direct effect. The magnitude of the inter-movie effect remains unclear as most research assumed that each movie can be treated as a segregated market, piracy's competition effects are assumed to exist only within-movie. This is understandable because most research used a linear reduced-form regression framework, making it very hard to incorporate competition effects as this requires including a full set of each movie's piracy on the right-hand side of the equation, which is computationally infeasible in reality.

The use of the logit demand model, in contrast, provides an opportunity to examine inter-movie effects of piracy, as competition between movies is naturally incorporated in the model. I conducted a formal decomposition exercise to quantify the magnitude of intra-movie and inter-movie effects of piracy, described in a later section.

3.6 Model

Models of movies demand with realistic substitution pattern and taking into account consumer heterogeneity are pivotal in examining the effect of file-sharing. In this section, I present a static random coefficient demand model of movies from both legal source and file-sharing based on Berry et al. (1995).

Figure 3.2: Average Weekly Audience and Downloads per Movie by Weeks after Release

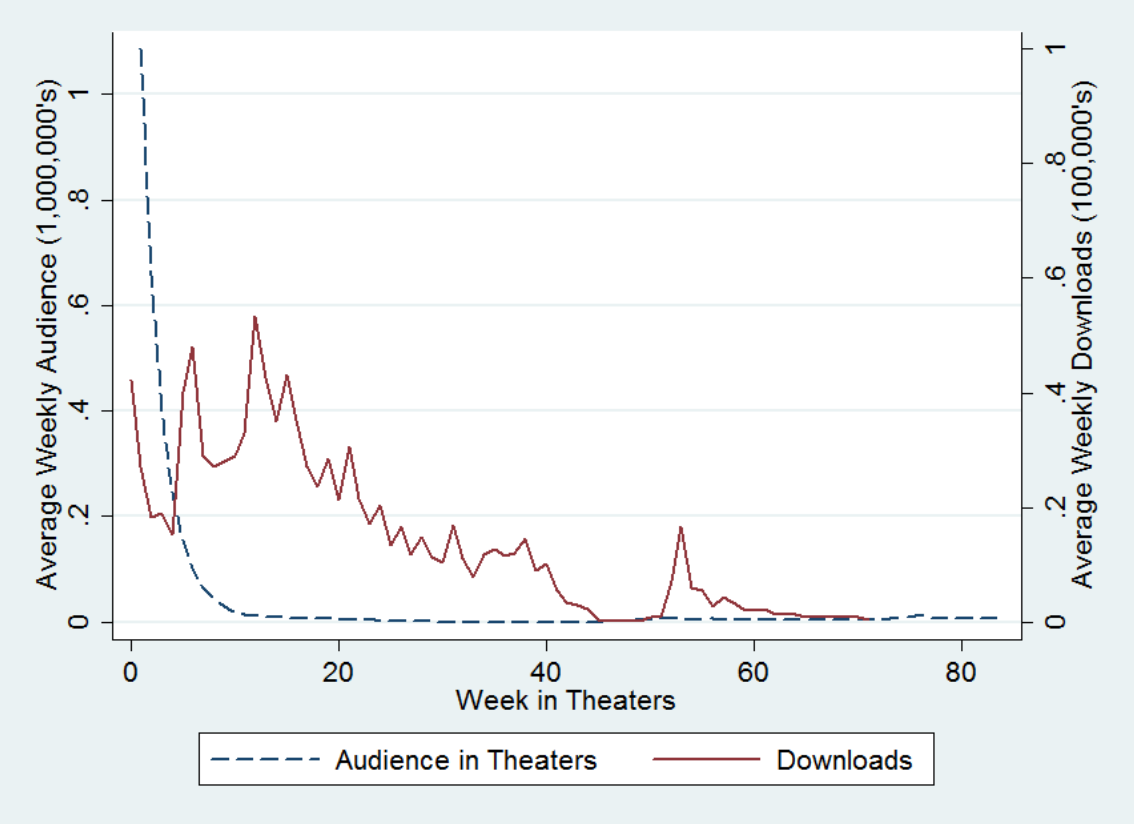
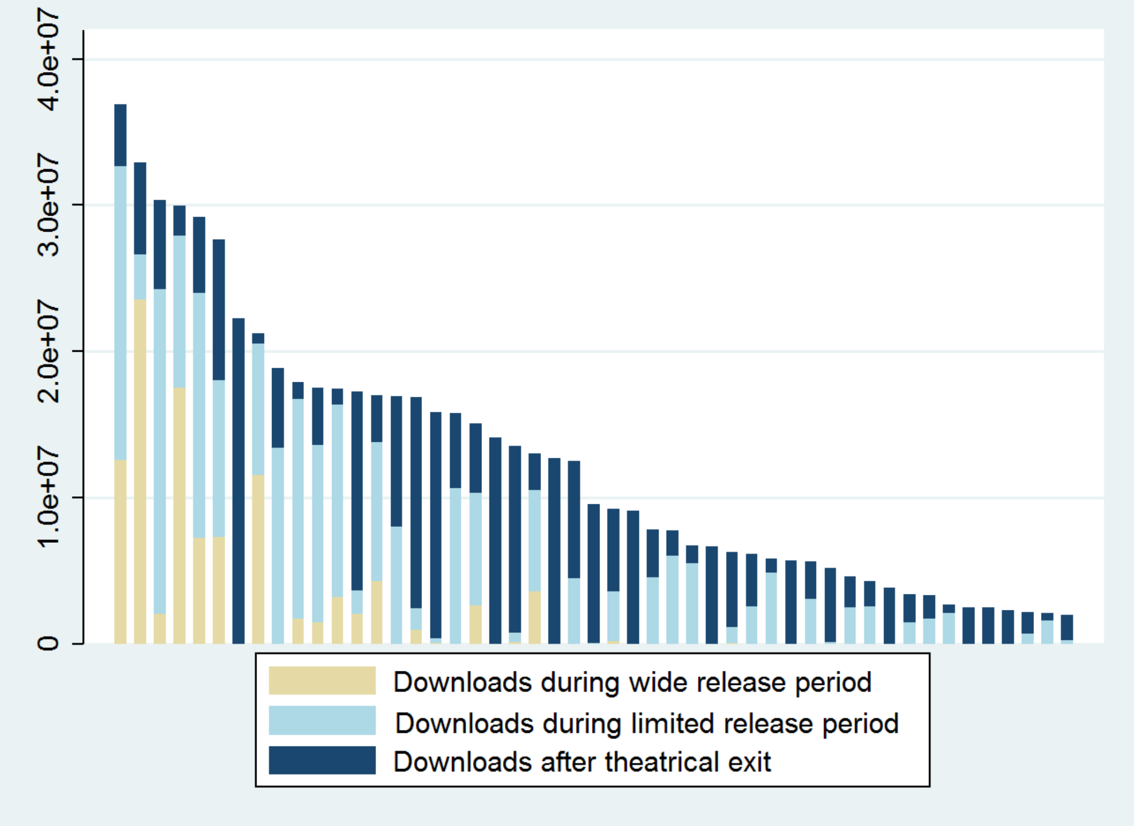


Figure 3.3: Downloads Composition by Theatrical Availability for Top 50 Downloaded Movies



Notes: This figure breaks total downloads into three part: downloads during wide release period (when studio arrange more than 1000 screens nationwide), downloads during limited release (0-1000 screens) and downloads after theatrical exit. The figure shows top 50 movies in terms of aggregate downloads, for comparison I restrict movies to have at least 40 weeks in theatre by the end of my sampling period.

It is well acknowledged that random coefficient models can generate better substitution pattern that can get rid of the unrealistic IIA assumption in multinomial logit demand models.

In the model, time is discrete and indexed by t, the decision period is one week in length. At each time period consumer face a set of products in the market. A product is defined as a movie that is currently showing in the cinemas or available to download on the Internet at a given period. Let m indexes movie titles and b denotes channel (i.e. $b = 1$ denotes illegal download and $b = 0$ denotes in legal purchase cinemas). A product is indexed by j, and defined as the pair consist of a movie title m and a channel b. The set of all available products is $\mathcal{J}$. For notational convenience, I first define a mapping $f : \mathcal{J} \mapsto \mathcal{M}$ and a mapping $g : \mathcal{J} \mapsto \mathcal{B}$ that map a given product $j \in \mathcal{J}$ to its movie title $m \in \mathcal{M}$ and its channel $b \in \mathcal{B}$, respectively.

Consumer i's utility from a product j which belongs to movie m at time t via channel b is:

$$u_{ijt} = X_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \tag{3.1}$$

where X_{jt} is a vector of observed movie characteristics including the following time-invariant characteristics average movie ratings in IMDB, genres, MPAA rating, last and most importantly a dummy variable *Pirated* which equals 1 if it's illegal piracy, it captures the utility difference between choosing illegal piracy and legal purchase.[10] Besides the above-mentioned variables, X_{jt} also includes time-varying characteristics weeks after release which are used to capture the decay pattern of sales and downloads. ξ_{jt} measures the time-varying product unobservable qualities. β_i is a vector of individual-specific taste parameters associated with these observed movie characteristics. Lastly, ε_{ijt} is the idiosyncratic consumer taste shock following Type-I Extreme Value distribution.

Consumer have heterogeneous taste over a series of characteristics. The heterogeneous taste takes a random-coefficient logit form which is a standard in literature (Berry et al., 1995; Nevo, 2001).The distribution of consumer tastes parameters for movie characteristics is modelled as multivariate normal: the taste of consumer i for characteristic k is denoted by $\beta_{ik} = \beta_k + v_{ik}\sigma_k$ where v_{ik}

[10] Notice that I did not include price coefficient in this specification, so one should treat the coefficient associated with *Pirated* as a combination a taste effect and a price effect.

is a mean zero taste shock for characteristic k. To write in matrix form:

$$\beta_i = \bar{\beta} + \Sigma v_i \tag{3.2}$$

I did not include random coefficient on all characteristics, let $\mathcal{K}$ denote the set of characteristics that have random coefficients and suppose I have K characteristics that have random-coefficients. $v_i = \{v_{i1}, v_{i2} ... v_{iK}\}$ is the vector form of unobservable consumer characteristics following a multivariate standard normal distribution. Σ is a scaling diagonal matrix.

I then rewrite the utility function into the following form:

$$u_{ijt} = \underbrace{\delta_{jt}}_{\text{Mean Utility}} + \underbrace{\sum_{k \in \mathcal{K}} X_{jt,k} v_{ik} \sigma_k + \varepsilon_{ijt}}_{\text{Error Component}} \tag{3.3}$$

δ_{jt} is the commonly called "mean utility" which captures the deterministic component of utility that is common to all consumers:

$$\delta_{jt} = X'_{jt}\bar{\beta} + \xi_{jt}$$

Now let us consider the error component of equation (3.3), which is the "random" or individual specific part of the utility. The second component $\sum_{k \in \mathcal{K}} X_{jt,k} v_{ik} \sigma_k$ introduce correlation for choice of same characteristics in X_{jt}. For instance, let characteristics $X_{jt,k}$ be the dummy variable corresponding to Action movies. If v_{ik} is high, which indicate consumer i have higher taste for Action movies, then consumer tastes for all alternative action movie will be high. This component is the main difference between random-coefficient logit demand and multinomial logit demand. If we remove this component, the model becomes standard multinomial logit.

Consumer i can also choose the outside option to neither buy nor pirate any movies. The introduction of outside option gives consumers flexibilities to turn to other non-movie activities, therefore rules out the unrealistic assumption that one download must transfer into one sale if file sharing is disabled. The utility of outside option is defined as:

$$u_{i0t} = \varepsilon_{0t} \tag{3.4}$$

Consumer i chooses one among all options to maximize his utility. Since the error term ε_{jt} follows extreme value distribution, consumer i's choice probability of movie j at time t can be written as:

$$Pr_{ijt} = \frac{exp(\delta_{jt} + \sum_{k\in\mathcal{K}} X_{jt,k}v_{ik}\sigma_k)}{1 + \sum\limits_{j'} exp(\delta_{j't} + \sum_{k\in\mathcal{K}} X_{j't,k}v_{ik}\sigma_k)} \tag{3.5}$$

And the market share of product j is then:

$$s_{jt} = \int Pr_{ijt} f(v_i; \Sigma) dv_i \tag{3.6}$$

Since I observe the same movie title across multiple weeks and channels, I can add movie title dummies to control for all time/channel-invariant characteristics of movies following Nevo (2001). One important benefit of including movie titles dummies and time fixed effects is that it helps improve fit of the model and serves to correct the potential bias caused by the correlation between observable movie characteristics and unobservable quality. Now market specific deviation from mean valuation $\Delta\xi_{jt} = \xi_{jt} - \xi_m$ with $f(j) = m$. $\Delta\xi_{jt}$ will serve as the new econometric error term, compared with the previous assumption, it is more plausible to assume that movie characteristics are predetermined and not responsive to shocks of unobservable $\Delta\xi_{jt}$.

Let I_j be a $M \times 1$ vector of movie title dummies. The vector of characteristics X_{jt} can be separated into two parts: one part $X_m^{(1)}$ is time/channel-invariant and movie title-specific characteristics,[11] and the rest part $X_{jt}^{(2)}$. When movie titles dummies are added, the demand model specified in equation (3.3) remains the same, but I do need to modify δ_{jt}:

$$\delta_{jt} = I_j'\theta + X_{jt}^{(2)'}\bar{\beta}^{(2)} + \Delta\xi_{jt}$$

Where θ is a $1 \times M$ vector of coefficients on movie title dummies representing mean quality corresponding to a movie title. Once movie title dummies have been introduced, the new vector of observable characteristics X_{jt} can not include time-invariant characteristics, because all variations in time-invariant variables are absorbed by these movie title dummies. Therefore, $X_m^{(1)}$ has to be

[11] I use $X_m^{(1)}$ as a short-hand notation for $X_j^{(1)}$ when $f(j) = m$

dropped from in the above equation. Let θ_m be the mth element of θ. For each movie title m I have:

$$\theta_m = X_m^{(1)'}\bar{\beta}^{(1)} + \xi_m \tag{3.7}$$

I allow for heterogeneous tastes for movie titles. There are good reasons to do so. Adding random coefficients on observables only allows consumers to have different tastes for observable characteristics. While this is perhaps enough to generate flexible substitution patterns for some markets, such as cereal, where products differ in relatively few observable dimensions, it becomes difficult in other settings, such as the movie market, where product differentiations take place in so many dimensions. For instance, consumers could be particularly fond of a particular story, a particular character, a particular actress, etc. This is difficult to capture with a limited number of observables. Therefore, allowing consumer preference heterogeneity on these unobservable movie characteristics will help to generate more flexible substitution patterns needed in the context of this paper.

The existence of multiple channels helps here, as I am able to observe multiple products associated with each movie title. Therefore, I can adopt an easy solution to incorporate heterogeneous taste on unobservable movie quality. Specifically, I add random coefficients to the movie title dummies. This helps generate correlation of preference within the same movie title. For example, it allows for an *Ironman* movie ticket to be a closer substitute to an *Ironman* pirated movie than a *Batman* movie. The implementation is straightforward. To begin, I modify equation (3.3) by adding an additional random component:

$$u_{ijt} = \delta_{jt} + \sum_{k\in\mathcal{K}} X_{jt,k}v_{ik}\sigma_k + \sum_{m=1}^{M} I_{j,m}w_{im}\sigma_w + \varepsilon_{ijt} \tag{3.8}$$

where $I_{j,m}$ denotes the mth element of I_j. Similarly to v_i , $w_i = \{w_{i1}, w_{i2}, ..., w_{iM}\}$ is another set of unobservable consumer characteristics representing taste for a given movie title, also following multivariate normal distribution: $w_i \sim N(0, \Omega_w)$, where Ω_w is the diagonal variance-covariance matrix. Because of the substantial number of movie titles, it would be computationally difficult to allow the standard deviations of random coefficients to be movie title-specific. For computational simplicity, I restrict the standard deviations of taste on movie titles to be the same for all titles. i.e. $\Omega_w = \sigma_w I$. Compared with the previous specification, this adds only one more parameter, σ_w, to estimate.

3.7 Estimation Procedure and Result

3.7.1 Estimation Procedure

Following the estimation procedure of Berry et al. (1995), I use GMM to estimate the model's parameters. The estimation procedure is a nested fixed point algorithm: in the inner loop I solve a contraction mapping to get the mean utility δ's from the market shares. In the outside loop the econometric error term $\Delta\xi$ is interacted with instruments to form the GMM objective function. The GMM estimator is obtained by minimizing the objective function using Nelder-Mead method:

$$(\hat{\Sigma}, \hat{\sigma_w}) = \underset{\Sigma,\sigma_w}{argmin} \Delta\xi(\Sigma, \sigma_w)' Z \Phi^{-1} Z' \Delta\xi(\Sigma, \sigma_w) \tag{3.9}$$

Where Φ^{-1} is the optimal GMM weighting matrix. My data consists of movie characteristics $\{X_{jt}\}$ and market shares $\{s_{jt}\}$. The parameters need to estimate includes $\{\bar{\beta}^{(2)}, \theta, \Sigma, \sigma_w\}$. Given the data and a guess of nonlinear parameters $\{\Sigma, \sigma_w\}$, I can solve the contraction mapping in the inner loop of the estimation algorithm:

$$\delta_{jt}^{n+1} = \delta_{jt}^{n} + ln(s_{jt}) - ln(S(X_{jt}, \delta_{jt}^{n}; \Sigma, \sigma_w)) \tag{3.10}$$

where $S(X_{jt}, \delta_{jt}^{n}; \Sigma, \sigma_w)$ is the simulated market share:

$$S(X_{jt}, \delta_{jt}^{n}; \Sigma, \sigma_w) = \frac{1}{n_{ind}} \sum_i Pr_{ijt}(X_{jt}, v_i, w_i; \beta, \Sigma, \sigma_w) \tag{3.11}$$

n_{ind} is the number of simulated individuals in the model, which is set to 500. Following Dube et al. (2012), I set the convergence tolerance of the contraction mapping to be 10^{-8} to avoid propagation of simulation error which affects parameter estimates.

Including movie title dummies requires modifications of the conventional estimation procedure. One can recover the mean taste for time-invariant characteristics by a two step procedure as implemented in Nevo (2001). First, after I solve for the mean utility δ's, I can regress them on $X_{jt}^{(2)}$ and movie title dummies I_j to obtain the estimates of $\{\bar{\beta}^{(2)}, \theta\}$. In addition, the estimated residuals from this regression correspond to the econometric error term $\Delta\xi$'s. I then apply GMM to the set of moment conditions in order to estimate $\{\Sigma, \sigma_w\}$:

$$E[Z\Delta\xi(\Sigma, \sigma_w)] = 0 \tag{3.12}$$

where Z is a set of instruments discussed in previous section. Once the nonlinear parameters $\{\Sigma, \sigma_w\}$, together with $\{\bar{\beta}^{(2)}, \theta\}$ is obtained, at the second step I can use equation (3.7) to recover mean taste parameters for the title-invariant characteristics via the linear regression depicted before.

Identification Distributions of random coefficients are identified using variations in choice sets and the corresponding change in market shares. For example, if three movies A, B, and C are offered, A and C have the same budget but very different ratings, while B and C have the same rating but very different budgets. Suppose we observe that movie C exits the market, then the magnitude of how consumers of C shift to movie A and movie B will help determine the distributions of the random coefficient on budget and rating, respectively. The features of the movie market provide good sources of variation in the choice set. In my 40-week period of data sample, I observe a large number of entries and exits of products coming from the theatrical release and the exit of the movie, release of the DVD, and leaks of the pirated version. This rich variation in the choice set provides a key source of identification of the random coefficients and associated substitution patterns.

Instruments For identification of the random coefficients, I maintain the assumption that own time-invariant product characteristics are uncorrelated with market specific deviation of mean valuation $\Delta\xi$. Given the assumptions, I choose a set of differentiation-instruments in line with Gandhi and Houde (2016) which approximate the optimal instruments of Chamberlain (1987). The instruments are:

- own product characteristics
- $\sum_{j'} ||X_{jt,k} - X_{j't,k}||^2$ for each characteristics k
- sum of number of rival product where difference between rival product characteristics and own product characteristics less than one standard deviation of product characteristics.
  $\sum_{j'} \mathbf{1}\{||X_{jt,k} - X_{j't,k}|| < sd(X^k)\}$ for each characteristics k

3.7.2 Estimation Results

This section reports the estimation results of my models. I report the results of demand estimations in Table 3.3. The mean coefficient on piracy, which represents the taste for piracy is -24.248, the standard deviation of the random coefficient is 12.391, which shows that people's preference for piracy is quite dispersed. The magnitude of the piracy random coefficient suggests that for preference for piracy there are many consumers at the extreme: on one hand, a significant fraction of consumers have extremely high preference for piracy, representing those "die hard" pirates; on the other hand, there are also consumers having extremely negative utility for consuming piracy, representing those people with no access to file-sharing technology. This has important implication for counterfactuals. Basically, the substitution pattern indicates that it will be very hard for consumers at those extremes to move across channels (illegal downloads/legal purchase). On the other hand, the standard deviation for movie fixed effects is significant but has relatively small magnitude. Therefore, the high dispersion restricts the effectiveness of piracy content removal.

For movie genres, the standard deviation of comedy, foreign and science fiction have relatively high magnitude compares with other genres like drama and cartoon. It suggest that consumers have very heterogeneous preference for comedy, foreign and science fiction movies. For MPAA ratings, the standard deviation terms are small in magnitude and statistically insignificant, indicating that consumer heterogeneity mainly exists on genres.

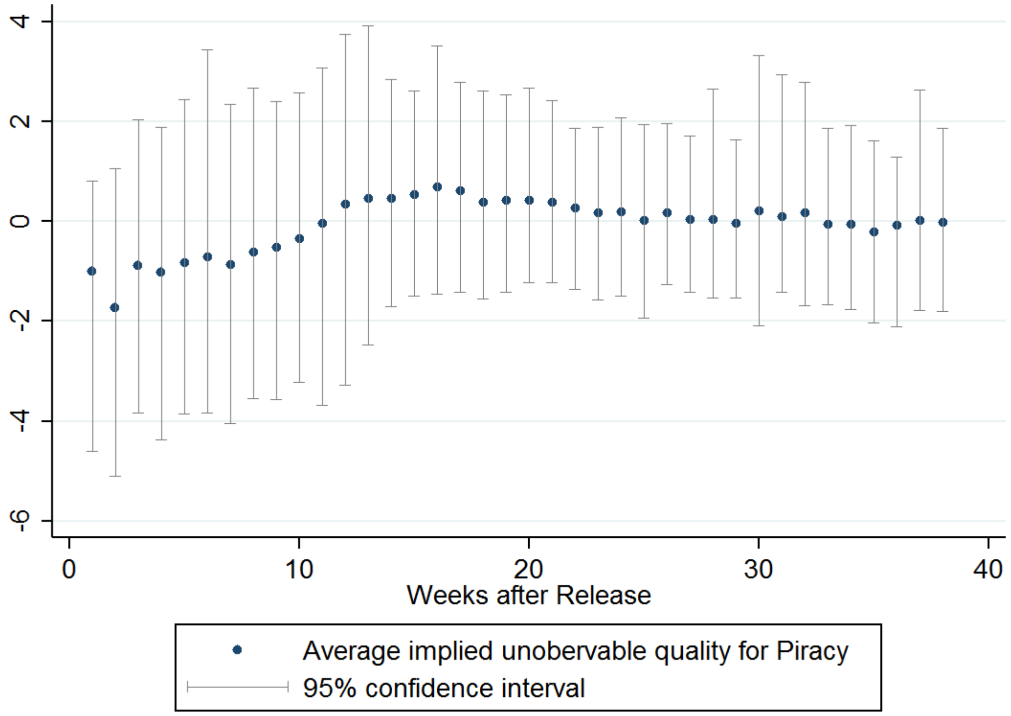
Now I turn to discuss how products progress after their releases. The decay pattern of movie consumption suggests that consumer prefer newly released

Table 3.3: Demand Estimation Results

Variable	Mean	Std Dev
Pirated	-24.248***	12.391***
	(2.712)	(1.320)
Weeks after Release	-1.182***	
	(0.114)	
Movie FE		0.939***
		(0.177)
Invariant Movie Charactersitics		
Rating	0.141***	
	(0.043)	
Sequel	0.103	2.774***
	(0.277)	(0.168)
Genres		
Action	0.794***	0.009
	(0.294)	(0.885)
Comedy	-10.546***	6.617***
	(0.260)	(0.075)
Drama	-0.744***	1.925***
	(0.221)	(0.126)
Science Fiction	-8.749***	7.289***
	(0.507)	(0.114)
Horror	0.733*	0.015
	(0.341)	(1.354)
Cartoon	-0.231	1.589***
	(0.414)	(0.100)
Foreign	-4.863***	3.206***
	(0.312)	(0.159)
MPAA Ratings		
PG	1.492***	0.003
	(0.297)	(0.689)
PG-13	1.898***	0.002
	(0.221)	(0.274)
R	1.155***	0.204
	(0.201)	(0.202)

Notes: Standard errors in parentheses. ***,**, and * denote statistical significance at 0.005, 0.01, and 0.05 levels respectively. Based on 5625 observations. Cast and Director are variables ranging from 0 to 100 measuring the strength of cast and director in terms of previous box-office performance. Vriable Age is a binary variable indicates whether or not individual is older than 40. Variable Income is the log of annual income and variable Internet Speed is the log of the speed of Internet. For full model, movie dummies and interaction terms of Pirated with contry dummies are included. Coefficients of time-invarying movie characterristics are obtained from regressing movie fixed effects on time-invarying movie characterristics.

Figure 3.4: Growth of Average Implied Unobservable Quality $\Delta\xi_{jt}$ for Pirated Products by Tenure



movies. In the estimation result, the negative coefficient on Weeks after Release confirms the fact that consumer utility for movie product diminishes over time. When tenure of movie increase by one week, consumer utility drops by 1.182. Another interesting question is how does piracy quality change over time? Due to a lack of actual quality data, I use piracy products' estimated $\Delta\xi_{jt}$ as measures for perceived quality after controlling for time trends. Figure 3.4 plots the growth if average implied quality for piracy products by tenure. There is a clear upward trend suggest that piracy quality improves continually during the first 15 weeks. The quality of piracy peaks around 16th week after theatrical release, roughly in line with the average DVD/Bluray release time for movies. The implied quality then stabilizes and slightly decreases possibly due to the drop in concurrent downloaders which limits download speed.

3.8 Counterfactual Experiments

The most important task in this study was to estimate the true cost of file-sharing on movie box office revenue and its welfare implications. In this section, I conduct several counterfactual experiments to estimate the true cost of file-sharing on box office revenue. First, I conduct a "No-Piracy" experiment that removes all pirated movie products in my models and compared the counterfactual box office revenue and consumer welfare with the benchmark. Second, I use partial removal experiments to decompose intra-movie and inter-movie substitution effects of piracy, I iteratively remove only pirated versions of each movie, while leaving other movies' pirated version untouched. Comparisons between recovered revenue for the removed movie and all other movies are then used to quantify the two effects.

3.8.1 Estimate Industry Loss

I remove all pirated movies in the model and recalculated counterfactual market shares using the estimated full model parameters in Table 3.3. Assuming price is the same after the no-piracy policy, I can then calculate counterfactual industry revenue as the product of market share times market size and price. Following Train (2003), consumer welfare at time t is calculated as the mar-

Table 3.4: Result of No-Piracy Counterfactual Experiment

	With Piracy	No Piracy	Percentage Change
Industry Revenue(billion)	8.527	8.620	1.09%
Consumer Welfare(billion)	10.849	8.614	-20.6%

ket size times the average of expected maximum value of indirect utility of simulated individuals:

$$CS_t = M\frac{1}{|\alpha|}\frac{1}{n_{ind}}\sum_i E[maxu_{ijt}] \tag{3.13}$$

where α is the mean price coefficient used to translate utility into terms of money value[12] and M denotes market size.

The result of the counterfactual experiment is shown in Table 3.4. The elimination of pirated movies on file-sharing will result in an increase of industry revenue of $93 million during a 40 week period. The number represents a 1.09% increase in total box office revenue.[13] Consumer welfare decreases by $ 2.235 billion when piracy is banned, mostly affecting consumers who have high preference for piracy. The number is much higher than the increase in motion picture industry revenue. There is a dead weight loss of $ 2.14 billion if we ban movie piracy. In general, the counterfactual result suggests that piracy indeed reduces firm revenue, but also increases consumer welfare which is higher than the initial loss. Hence, the policy that eradicating movie file-sharing may result in transfer of large reduction of consumer welfares into small increase in industry revenue, resulting in socially inefficient outcomes from just the social welfare point of view. However, if I decompose the welfare improvement by consumer preference for piracy (a_i), I find the piracy's welfare improvement disproportionately benefits those "pirates": consumers who have positive preference for piracy ($a_i > 0$). The decomposition shows that 98.3% of piracy's total welfare improvement is on consumers with positive preference on piracy. Meanwhile only 1.7% of welfare improvement belongs to consumers with neg-

[12] Because I did not attempt to estimate price elasticity in this paper, I parametrize the α as 0.16 according to Davis (2002).

[13] For comparison with widely cited industry studies: In 2005, the Motion Picture Association of America (MPAA) estimated that they were losing $3 billion in box office sales due to piracy according to De Vany and Walls (2007).

ative preference for piracy ($a_i < 0$). The removal of piracy will mainly affect pirates and have relatively small effects on consumers from legitimate channels.

If we use a "naive" way to estimate the revenue loss, assuming that one download equals one lost sale of paid movies, then the estimated revenue loss amounts to 1.265 billion dollars for the same time periods, which is 13.6 times the revenue loss calculated in the counterfactual experiment. Many widely cited industry studies have employed this "Naive" method in their estimation of piracy's cost. The result shows that using such methodology will substantially inflate the true loss from piracy.

I also calculate the average displacement rate of pirated movies on legitimate movie sales in theatres. On average one download displaces legitimate sales by 0.07 unit.

3.8.2 Decomposing Inter-Movie Substitution and Intra-Movie Substitution

In this subsection, I decompose the intra-movie and inter-movie substitution effect of piracy. Start again from the simple decomposition in previous section. An analog in counterfactual experiment for the total effects would be the sum of recovered revenue for all movie when movie j's piracy is removed. Intra-movie effects of movie j's piracy can be quantified using the recovered revenue for movie j itself, while inter-movie effects can be quantified using sum of recovered revenue from all the other movies.

$$\frac{\partial R}{\partial D_j} = \underbrace{\frac{\partial \sum_{j'} R_{j'}}{\partial D_j}}_{\text{total effect}} = \underbrace{\frac{\partial R_j}{\partial D_j}}_{\text{intra-movie effect}} + \underbrace{\frac{\partial \sum_{j' \neq j} R_{j'}}{\partial D_j}}_{\text{inter-moive effect}}$$

Therefore, the ideal counterfactual experiment would be to partially remove only piracy of each particular movie and compare the intra-movie recovered revenue with inter-movie recovered revenue.

Such counterfactual experiment can be also seen as approximation of private copyright protection measure. Copyright protections are not always initiated by the government or legislation, where policy are tend to affect the whole industry. In recent years private copyright protection initiated by firms targeting at individual copyrighted work becomes more and more prevalent. In motion picture industry, studios hires internet surveillance company to monitor and send DMCA notices to take down torrents files on file-sharing websites.

As Reimers (2016) pointed out, such private copyright protections are effective in the book publishing industry. How effective are those private copyright protection efforts targeted to remove piracy for individual movies? Will the downloader substitute its paid version, other pirated movies, or simply the outside options?

To implement this counterfactual experiment, for each movie I eliminate all its piracy across all time periods, but leave pirated versions of other movies untouched. I then calculate counter-factual market shares and counter-factual revenue increase for that movie.

Table 3.5 shows the comparison of the average movie's revenue increase between this partial removal counterfactual experiment and the full removal experiment. Not surprisingly, the average revenue increase dropped from 0.170 to 0.014 million dollars, only 8% of the average recovered revenue by eradicating all piracy. In this counterfactual, most downloaders will choose the other available pirated movie or other similar movies instead because in many cases the availability of the original movie in theatres is small. This comes as no surprise given the high dispersion in estimated preference for piracy. Removing other movies' piracy, however, indirectly helps this movie's revenue through positive spillovers from inter-movie substitution.

On average, I find sizable spillover effects resulting from inter-movie substitution. Removal of one movie can indirectly increase the revenue of other movies by 0.055 million dollars, roughly four times the recovered revenue from intra-movie substitution effects.

The existence and large magnitude of the inter-movie substitution effects indicate: (1) The successful battle against piracy requires all studios to take action against piracy, movie studios are interconnected as industry revenue will be severely affected if some studios fail to deter piracy of their movies. (2) Anti-piracy efforts can have large positive externalities to other firms, this public goods nature might disincentivize provision of anti-piracy efforts, especially for studios treating each other as competitors than cooperators. Government involvement may be needed for subsidy and coordination of the copyright protection efforts. (3) Research attempt to estimate the effects of piracy should

Table 3.5: Comparison of Revenue Increase from two Counter-Factual Experiment

($ millions)	Full Removal	Partial Removal
Average Recovered revenue	0.170	0.014

Notes: This Panel summarize the distribution of recovered revenue under full removal and partial removal. For partial removal exercises, I iteratively remove and only remove the piracy product for each movie and calculate the improved revenue.

Table 3.6: Decomposition: Intra-movie vs Inter-movie Effect

($ millions)	Intra-movie Effects	Inter-movie Effect
	(self)	(others)
Average Recovered revenue	0.014	0.055

Notes: This Panel reports the size of intra-movie effects and inter-movie effects. The "Inter-movie Effect" calculate the sum of improved revenue on all other movies and "Intra-movie Effects" measures improved revenue on the movie itself.

also consider the inter-movie substitution effects, especially for the motion pictures industry. Focusing exclusively on intra-movie effects will underestimate the actual damage from piracy.

3.9 Conclusion

This paper examines the effect of piracy on movie box-office revenue using a novel data of illegal movie download from BitTorrent networks. Motivated by the data pattern that piracy downloads happen relatively late in most movies' theatrical windows, I estimate a random-coefficient demand model of movie consumption to decompose the intra-movie and inter-movie substitution effects. I find that the revenue loss due to piracy's inter-movie substitution is on average four times as large as intra-movie substitution. This is quite surprising considering only direct or intra-movie substitution will seriously underestimate piracy's damage. In addition, my damage estimates show that piracy reduces total revenue of the motion picture industry from box office sales by $ 93 million in total, 1.09% of the current box office revenue. The estimates are smaller than widely cited industry estimates constantly referenced in policy making, the "naive" methodology that assumes full sale displacement will inflate the true cost 13.6 times. In addition, anti-piracy campaigns that remove piracy for individual movies have limited benefits to box office revenue because most downloaders just substitute into other pirated movies.

The findings of this paper serve to provide extra evidence to assist resolution of the current heated debate on controversial issues regarding intellectual property. For policy makers, the findings in this paper highlight the importance of considering outside option and substitution in evaluating the effect of file-sharing;, research omitting these factors will substantially overestimate the negative effects of file- sharing and should be treated with caution for policy making. For the industry, the finding in this paper can be used by motion picture studios to determine the optimal level of copyright protection given the high cost of supervision and litigation.

An interesting question I did not answer in this paper is how the supply of movies is affected by file-sharing since I take movie release as exogenous in my model. An interesting extension of this paper would be to model the movie release decision as an entry game given the estimated demand system. This would help to find the effect of file-sharing on producer incentives to supply new products, which is also an important question worth exploring in the future.

Chapter 4

Quantifying the Heterogeneous Effects of Piracy on the Demand for Movies

4.1 Introduction

Although two decades of research on file-sharing and piracy has not produced a consensus on the impact of piracy on legitimate sales, it does show us that the issue is more complex than we think. Traditionally researchers have focused attention on the cannibalization effect of piracy, yet studies on different markets over different time periods generate drastically different conclusions regarding the impact of piracy on sales. Such differences indicate that the issues of piracy are more subtle than we previously perceived.

First, there might be substantial heterogeneity in the effects of piracy. There are two aspects of heterogeneity. (1) The market for pirated goods is highly differentiated by circulation method (e.g., street piracy vs. online piracy) and quality (e.g., pirated movies from low-quality camera recordings vs. pirated movies ripped from Blu-ray). (2) Copyrighted goods such as movies are usually distributed via different channels, which have distinct product features, prices, and release times. As substitutability is driven by these characteristics, different types of piracy and copyrighted goods from different distribution channels may have heterogeneous substitution patterns. For example, pirated movies ripped from Blu-ray might be closer substitutes to home-video sales of DVDs/Blu-rays than a low-quality bootleg video from a theatre recording.

Second, besides the obvious cannibalization effect, it is possible that piracy can have positive effects on sales through various channels.[1] One notable channel for such positive effect is through word-of-mouth (WOM). Pirated consumers can spread the word and make recommendations to other consumers. It helps create a "buzz" for the product and eventually benefits its sales. Several papers have focused on the WOM aspect of piracy (Lu et al., 2019, Ma et al., 2014, Peukert et al., 2017) and have shown that there is significant WOM generated from piracy consumption. Given the existence of WOM, piracy consumption might induce positive spillovers onto legitimate sales. Therefore, it will be difficult to conclude the true effect of digital piracy without knowing (1) the heterogeneous effects of different pirated consumption on legitimate sales and (2) the magnitude of the positive effect of pirated consumption on sales.

The goal of this paper is to answer these questions that are at the centre of current debates. Using a novel dataset of movie illegal downloads on file-sharing network BitTorrent with information on torrent quality, I classify piracy into different types and estimate a rich demand model to quantify the heterogeneous effects of piracy on movie sales. I focus on two types of heterogeneity: (1) how the effects differ by video quality of pirated movies; (2) how the effects differ on two different channels of sales: box office and home-video (DVD/Blu-ray) sales.

Utilizing data on movie's weekly WOM measured by Google search volume, I also take into account the potential complementarity between piracy and sales through the spread of WOM. I decompose the effects of piracy on sales into a negative cannibalization effect and a positive WOM effect.

Quantifying the heterogeneous effects of piracy has important managerial implications. Understanding the heterogeneous effects of piracy helps firms to effectively allocate their efforts on protection for different channels of sales against different types of piracy according to the differences in their effects. Under some circumstances, when the positive effects outweigh the negative

[1] Peitz and Waelbroeck (2006) and Belleflamme and Peitz (2014) provide a more comprehensive survey of the literature on the positive effects of digital piracy

effects, firms can utilize piracy as a promotional tool under the right timing.

The findings of this paper are as follows. First, on-line movie piracy reduces the total revenue of the motion picture industry from the box office by \$231 million in total, or about 2.71% of the current box office during my 40 weeks sampling period in 2015. The estimates are smaller than many widely cited industry estimates often referenced in media. The “naïve” methodology of assuming full sale displacement will inflate the true cost by a factor of 5. However, responses differ substantially by channels. Unlike the box office, in the home-video market, DVD revenue would increase by 36% if there were no piracy. On average, one movie suffers a 40-week monetary loss of \$1.24 million because of file sharing in 2015. Second, different qualities of piracy play different roles. High quality is a closer substitute to sales, but the removal of high-quality piracy alone does not solve the problem, as consumers have a strong preference for piracy and will substitute to low-quality piracy. Third, the results of the welfare analysis show that file sharing increases consumer welfare by a total of \$7.05 billion. Lastly, I examine the magnitude of the word-of-mouth (WOM) effects of piracy on sales revenue and find that the WOM effects have a small and positive impact on the industry revenue.

The most important contribution of this paper to the literature is to quantify the heterogeneous effects of piracy. Most studies on movie piracy have focused exclusively on the total effect of piracy on only one particular distribution channel. Very few studies have assessed and compared the heterogeneous effects by different types of piracy and by different distribution channels. The heterogeneous results I obtain in this paper suggest that public policies towards piracy should not be over-generalized, as different types of piracy have distinct influence over different channels of sales. For example, the positive WOM brought by low-quality piracy make it useful as a promotional tools for studios.

A few papers have attempted to examine the positive role played by piracy. Notable channels include positive effects on demand for complementary goods (Papies and van Heerde, 2017; Leung, 2015), indirect appropriability (Liebowitz, 1985), sampling effects (Kretschmer and Peukert (2017)). The most similar papers in the context of movies are Lu et al. (2019) and Ma et al. (2014), which also examine the word-of-mouth effects of piracy on the box office. I complement these studies by directly modelling the consumer choice of piracy using a flexible random-coefficient demand model that attempts to capture and explain the rich heterogeneity of the effects of piracy on movie sales.

In addition to the empirical literature on file sharing, this paper is also related to the growing literature on the motion picture industry. Researchers have studied different aspects of the motion picture industry, for example: spatial competition of movie theatres (Davis, 2006), social spillover, and WOM (Chintagunta et al., 2010; Moul, 2007; Moretti, 2011; Gilchrist and Sands, 2016b), seasonality in the motion picture industry (Einav, 2007), uniform pricing practices (Orbach and Einav, 2007), movie price elasticity (Davis, 2002; De Roos and McKenzie, 2014), effects of uncertainty in the movie industry (De Vany and Walls, 1999; Elberse and Eliashberg, 2003), and strategic entry and exit decisions of studios and theatres (Einav, 2010; Takahashi, 2015; Dalton and Leung, 2017). This paper adds to the literature on the effect of file sharing on the motion picture industry.

This paper is also related to the marketing and economics literature studying the WOM effect and viral marketing (Dellarocas, 2003; Godes and Mayzlin, 2004; Chevalier and Mayzlin, 2006; Trusov et al., 2009; Zhu and Zhang, 2010). Many papers try to assess the role of WOM as either a predictor or influencer for sales. There are a number of papers focused on the motion picture industry (Liu, 2006; Duan et al., 2008; Chintagunta et al., 2010; Dhar and Weinberg, 2016). This paper adds to the literature by studying the WOM effect induced by movie piracy. I quantify the positive effects that movie piracy brings to the movie sales revenue through WOM.

The chapter is organized as follows. Section 4.2 describes the data and some preliminary evidence. Section 4.3 discusses the model. The estimation results are presented in Sections 4.4. Section 4.5 gives the results of the counterfactual experiments, and Section 4.6 concludes the paper.

4.2 Data

4.2.1 Data Description

In this paper, I combine data from several data sources. I assemble a dataset including weekly information on movie-level box office sales, downloads, and sales on DVD in United States. The movie piracy download data contain weekly downloads on BitTorrent in the United States. The details of the data and its collection have been discussed in Chapter 3 and Appendix B.3.

The box office and DVD sale data are collected from the box office reporting service websites *Boxofficemojo.com* and *The-numbers.com*. I collect information on weekly box office and movie characteristics for all movies showing during the sampling period. I also collect information on movie characteristics, such as movie ratings, sequels, genres, MPAA rating (G/PG/PG-13/R), and weeks after release, which are commonly used in studies of the motion picture industry. The movie critic rating data are collected from *Internet Movie Database (IMDB)*. Due to the uniform pricing practices in movie theatres and the fact that movie price data are difficult to collect, only a country-level average admission price is obtained.

Word-of-mouth(WOM) data are collected from Google Trends, The Google Trends search index measures the popularity of topics according to the number of search queries using Google. As same movie can relate to multiple similar search queries, to ensure better comparability and precision, I utilize Google's *Freebase* topic classification engines and extract the freebase topic identifier related to each movie.[2] Using the identifier, I collect weekly US Google trends index in sampling period for each movie.[3] The summary statistics are shown in the next section.

4.2.2 Preliminary Data Pattern

Heterogeneity in Movie Piracy Pirated movies come in different formats/versions. Table 4.1 shows descriptions of different formats and classifications based on video quality. There is substantial heterogeneity in quality across different formats, and the video quality will generally improve over time. After a movie has been released in the theatre, the earliest piracy is usually of CAM(Camcording) format, produced by camera recording in the theatre. The

[2] For example, keywords "Furious 7" and "Fast and Furious 7" are all related to the same movie. the freebase topic identifier automatically takes into consideration all search queries related to movie *Furious 7*.

[3] The raw Google trends data is normalized and takes integer value from 0 to 100, with 100 the highest search volume and 0 the lowest. I use movie Titanic as reference group, with the search volume of *Titanic* at first week of November 2015 at 97, all movie's WOM is then standardized with respect to *Titanic*. Because Google censors all search volume that are sufficiently low to 0, given that the scale of WOM is relatively large with average of 150 and max of 8400, in the end when take logarithm of WOM I take the $log(1 + x)$ transformation.

Table 4.1: Pirated Movie Format and Quality Classification

Format	Description	Timing	Quality
CAM	bootleg recording made in theaters	Early	Low
Telesync(TS)	improved bootleg recording of a film recorded in a movie theater, filmed using a professional camera on a tripod in the projection booth.	Early	Low
Telecine(TC)	Pirated movie copy captured from film print using a machine that transfers the movie from its analog reel to digital format	Early	High
DVDSCR	Pirated movies copied from movie screener for review purpose	Early/Late	High
Web-DL/WebRip	Pirated movies ripped from streaming service	Late	High
HDRip/HDTV	Pirated movies captured using analog capture card from HDTV	Late	High
DVDRip	Pirated movies ripped from retail DVD	Late	High
Bluray/BRRip	Pirated movies ripped from Blu-ray	Late	High

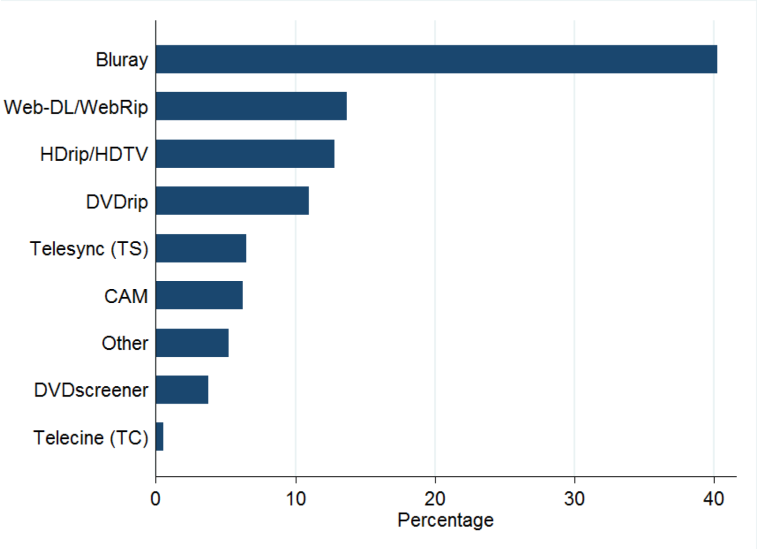
Notes: Descriptions are abridged from the Wikipedia page of each format. The classfication of quality is based on source, any source from bootleg recording is classified as Low quality.

subsequent release of the *Telesync* version significantly improves the quality of the CAM version, but it is still inferior compared to the quality of a DVD. Other piracy formats come with much better quality, almost comparable to the quality of a DVD, but take longer to release. Figure 4.1 shows the distribution of the illegal torrent file formats.

Based on the uploaded date of illegal torrent files, I obtain the leak date by illegal versions for each movie. Table 4.2 shows some summary statistics on the timing of piracy leaks. On average, the first leak (usually the CAM version) appears 4.7 weeks after the initial theatrical release. About 7.4 weeks after the initial release comes the first high-quality piracy release (Telecine/Screener/Webrip). About 17 weeks after the initial release, distributors will release DVD and Blu-ray versions of the movie into the home-video market. The DVD and Blu-ray versions of piracy (DVDrip/BRrip) happen almost instantly after the home-video retail release. Piracy appears in only a fraction of movies. My sample period covers the full theatrical windows of 694 movies, and for 78.4% of movie titles, no available piracy of any kind is observed during theatrical runs.

Are the effects of piracy on sales differ between high-quality and low quality piracy? I present some preliminary evidence to verify the conjecture on the differential effects between high-quality and low-quality piracy on sales.

Figure 4.1: Torrents Files Format Distribution

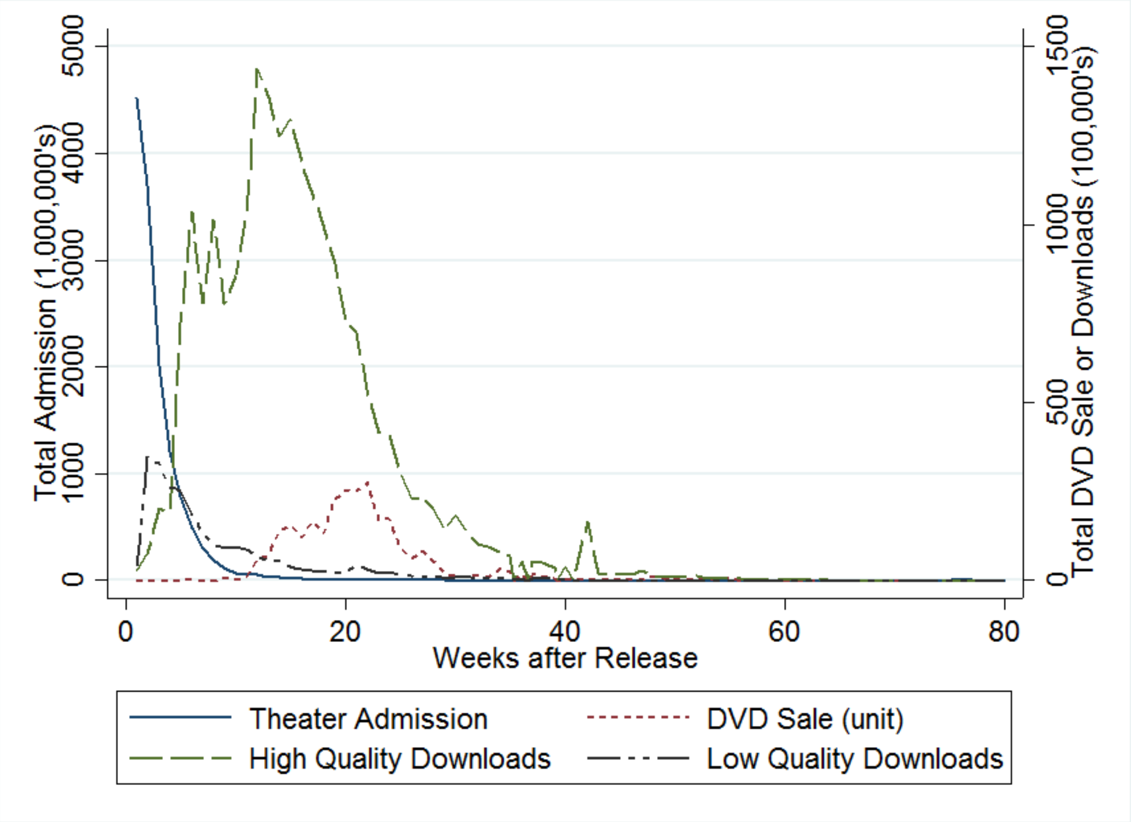


Notes: Figure 2 shows the distribution of torrent files formats. Most common formats are included in my dataset. Category Other includes other uncommon formats, and torrents that shows no information of format. For those with unknown format, I impute their quality use their appearance time, if it come at time with more high quality torrents of the same movie, it is then classified as high quality torrent.

Table 4.2: Summary: Timing of Piracy Release by Format, Quality

	Mean	Std Dev	Min	Max
Timing				
Weeks from Theatrical Release to Any Piracy Leaks	4.733	4.734	-1	23
Weeks from Theatrical Release to First High Quality Piracy Leaks *(Telecine/DVDscreener)*	7.423	5.081	-1	23
Weeks from Theatrical Release to DVD/Blu-ray Piracy Release *(DVDrip/BRRip)*	17.161	5.820	6	52
Fraction of Movies with No Leaks during Theatrical Run	78.4 %			
Fraction of Movies with No High Quality Leaks during Theatrical Run	79.5 %			

Figure 4.2: Theatre Admission, DVD Unit Sale and Downloads by Weeks after Release



A shown in Table 4.2, for most movies, the availability of piracy resources during the theatrical run differs substantially across movies and time. I can compare changes in weekly box office sales of leaked movies before and after the time low-quality and high-quality piracy become available, against a baseline of changes for movies with no piracy availability at the same period.

The specification is as follows:

$$ln(Boxoffice_{it}) = \beta_H HQLeak_{it} + \beta_L LQLeak_{it} + \sum_j \mathbb{1}\{\tau_{it} = j\} + X'_{it}\gamma + \alpha_i + \lambda_t + \eta_{it} \tag{4.1}$$

Here, the dependent variable $ln(Boxoffice_{it})$ represents the natural log of the weekly box office sales for movie i at time t. I include movie fixed effects α_i and calendar week fixed effects λ_t. To capture the decaying pattern of ticket sales, I also include a series of release week dummies $\sum_j \mathbb{1}\{\tau_{it} = j\}$, where τ_{it} is the count of the weeks after theatrical release with $j = 1, 2, ..., J$. Moreover, $HQLeak_{it}$ is an indicator variable that equals 1 if high-quality piracy for movie i has already leaked at time t. Similarly, $LQLeak_{it}$ is an indicator variable equal to 1 if low-quality piracy for movie i has already leaked at time t. In the end, X_{it} is the control including the log number of screens in week t for movie i.

The specification resembles a difference-in-difference (DD) specification. The first difference is taken using the movie fixed effects, which controls for time-invariant movie heterogeneity between the treated and control data. The second difference is taken using the time fixed effects, which controls for the general movie-invariant time trends before and after treatment (leaks). The coefficients β_H and β_L measure the percentage changes of the box office of movies after the emergence of high-quality piracy (β_H) and low-quality piracy (β_L), respectively, against the baseline change of movies without presence of piracy. I interpret negative coefficients of β_L and β_H as evidence of harm against sales.[4]

[4] Although not pursuing a causal interpretation of my estimates, the estimates still give me some good ideas on the correlation of piracy since the specification has already taken care of major sources of bias. A number of unobservable factors are accounted for in the specification: (1) movie-specific time-invariant unobservable heterogeneity (e.g., better movies attract more piracy and have more downloads), (2) general decreasing trend of the box office over its release time (e.g., natural decaying patterns will not be falsely attributed to the effect of piracy), and (3) time-variant but movie-invariant factors (e.g., box office and downloads both rise when the summer holiday begins).

Table 4.3 shows the results of these regressions. Column (1) reports estimates without the calendar week fixed effects, while column (2) includes full sets of fixed effects. In addition, β_H is negative and statistically significant in both specifications. Estimates with full set of fixed effects in column (2) suggest that, as high-quality piracy becomes available, the box office drops by 19%. However, for low-quality leaks, the estimate is no longer statistically significant and has a much smaller magnitude. These result are suggestive that high-quality piracy is harmful to sales, but the harm from low-quality piracy is indefinite and at least much more moderate.

Low-quality and high-quality piracy seem to play different roles here. Low-quality piracy tends to be less substitutable with sales compared to high-quality piracy. In addition, since low-quality piracy appears earlier than high-quality piracy, it is more likely to benefit legitimate sales through WOM. These differences motivate me to model the types of piracy to highlight the different roles played by these two types of piracy.

4.2.3 Summary Statistics

Before describing the model, I first describe the final data used for estimation of the model. The basic observation is at product-time level. Because for one movie there are potentially four channels that a consumer can get access(theater/DVD/low-quality piracy/high-quality piracy), a product is therefore defined as a movie and channel pair.[5]

Table 4.4 provides sample descriptive statistics for the data at products level. About 55% of the products are pirated movies, and about 7.8% are DVDs. 82.5% of all products are of high video quality, which indicates 17.5% of piracy product are low quality piracy. Bottom panel of Table 4.4 provides information on the average product market shares by channels. The average product market share of the ticket sales of a movie is about 0.077%, DVD sales on average take up 0.039% of market share, while the average product market share of the high video quality downloads of a movie is about 0.01%, low quality piracy products have the smallest average market share of 0.003%. Taken

[5] The box office of some independent movies is extremely small so that their market shares are indistinguishable from zero. Inclusion of these "zero" market share movies creates numerical problems to the estimation procedure, so I drop all observations with a market share smaller than 0.0001% in my sample.

Table 4.3: Preliminary Regression Estimates of impact of piracy by quality

	(1)	(2)
Dependent Variable:	$ln(Boxoffice_{it})$	$ln(Boxoffice_{it})$
After Piracy Leaks		
High Quality	-0.1441^{**}	-0.1904^{***}
	(0.0498)	(0.0507)
Low Quality	-0.0374	-0.0568
	(0.0584)	(0.0580)
Controls		
$ln(Screens_{it})$	0.8841^{***}	0.8769^{***}
	(0.0077)	(0.0077)
Movie FE	✓	✓
Week After Release Dummies	✓	✓
Calendar Week FE		✓
Observations	6805	6805
Adjusted R^2	0.8586	0.8637

Standard errors in parentheses,observations are movie-week level

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

together, the illegal downloads account for less than 10% of all movie-watching activities.

In terms of genre, the three most common genres are drama (19.1%), comedy (17.7%), and action (16.6%). Around 24% of products are sequel of previous movies. The average Word-of-mouth(WOM) index for a product is around 150, the distribution is skewed substantially to the right by the top big budget "hit" movies. Rating on IMDB are on a scale of 0 to 10, the distribution of rating has smaller dispersions with an average of 6.43 and a standard deviation of 1.56.[6]

4.3 Model

Models of movie demand with a realistic substitution pattern, considering consumer heterogeneity, are pivotal in examining the effects of digital piracy. In this section, I present a static random-coefficient demand model of movies from both legal and piracy sources based on the work by Berry et al. (1995). A random-coefficient demand model introduces heterogeneity in consumer preference. By allowing for rich dimensions of heterogeneity in consumer preference, the model will allow for more flexible substitution patterns. This is crucial for precisely measuring the effects of digital piracy on sales.

Basic Setup In the model, time is discrete and indexed by t, and the decision period is one week in length. Each time period, consumer observes a number of products in the market for movies in the US. A product is defined as a movie that is currently showing in the cinemas, available on DVD/Blu-ray, or available to download on the Internet at a given period. Let m denote the movie title and $\mathcal{M}$ be the set of all available movie titles. Let b denote the channel through which consumers watch a movie. For a given movie title $m \in \mathcal{M}$, there are up to four channels to watch movie m depending on availability: by purchasing a ticket at the theatre ($b = 0$), purchasing a DVD ($b = 1$), downloading a low-quality illegal copy online ($b = 2$), or downloading a high-quality illegal copy

[6] As for the sensitivity of using IMDB ratings as measure of movie quality, I collect popular metrics from other movie review websites including *RottenTomatoes* and *Metacritic*. The correlation coefficients between these measures are relatively high. For wide release movies, the correlation coefficient between IMDB, *RottenTomatoes* and *Metacritic* are 0.8021 and 0.7903, respectively.

Table 4.4: Summary Statistics on Product Characteristics

	Mean	Std.Dev	Min	Max
Pirated	.553	.497	0	1
DVD	.078	.268	0	1
Home	.631	.482	0	1
High Quality	.825	.379	0	1
Word-of-mouth	150.382	327.698	1	8400
Rating	6.434	1.556	4	9.3
Sequel	.243	.429	0	1
Genres				
Action	.166	.372	0	1
Comedy	.177	.382	0	1
Drama	.191	.393	0	1
Science Fiction	.077	.268	0	1
Horror	.081	.273	0	1
Cartoon	.075	.264	0	1
Foreign	.035	.186	0	1
MPAA Rating				
PG	.168	.374	0	1
PG-13	.368	.482	0	1
R	.354	.478	0	1
Market Share				
Ticket Sale(%)	.077	.405		
DVD Sale(%)	.039	.405		
High Quality Downloads(%)	.010	.019		
Low Quality Downloads(%)	.003	.011		
Observations	8617			

Note: Pirated is a dummy variable which equals 1 if the movie has a pirated version available online. Rating are on a scale of 0-10. Sale and Downloads in movie characteristics section are measured in units. Action, Animation, Comedy, Drama, Horror, Science Fiction, PG, PG13, R are all genre and MPAA Rating dummy variables. In the market share section, the market share is the average market shares for all observations.

online ($b = 3$). Let $\mathcal{B}$ be the set of channels: $\mathcal{B} = \{0, 1, 2, 3\}$. A product is indexed by j, and defined as the combination of a movie title m and a channel b. The set of all available products is $\mathcal{J}$. For notational convenience, I first define a mapping $f : \mathcal{J} \mapsto \mathcal{M}$ and a mapping $g : \mathcal{J} \mapsto \mathcal{B}$ that map a given product $j \in \mathcal{J}$ to its movie title $m \in \mathcal{M}$ and its channel $b \in \mathcal{B}$, respectively. In each market, there are a number of consumers indexed by i. The market size is set to the total population of the United States and is constant in the sampling period.[7]

The utility of consumer i from product j of movie title m at time t is as follows:

$$u_{ijt} = W'_{jt}\eta_i + \alpha P_{jt} + a_i Pirated_j + \psi_i HQ_j + \gamma_i H_j + \phi ln(WOM_{mt-1}) + \tau_t + \xi_{jt} + \varepsilon_{ijt} \tag{4.2}$$

where W_{jt} is a vector of observed movie characteristics, including movie title-invariant characteristics, such as movie ratings in IMDb, genres, MPAA rating, sequel, as well as weeks after release. η_i is a vector of individual-specific taste parameters associated with these observed movie characteristics.

P_{jt} denotes the price of the product, with price of the piracy set to 0. i.e., $P_{jt} = 0$ if $g(j) \in \{2, 3\}$. The DVD prices are calculated by dividing the total weekly sales revenue by the total weekly units sold. Because of the uniform pricing practice in movie theatres and the difficulty to collect ticket price data, I use the US country-level average admission price. Because of the lack of price variation on the movie ticket, it is difficult to estimate the price elasticity α in this paper. I parametrize the price coefficient according to the existing literature on the movie industry. Davis (2002) uses a randomized control price experiment in US movie theatres to allow the price coefficient to be precisely estimated. Therefore, I parametrize my price coefficient according to the estimated price coefficient in the last column of Table 5 in Davis (2002).[8] The corresponding value of α is -0.16. The own price elasticity implied by the imputed price coefficient is $-\alpha P_{jt}(1 - s_{jt}) = 1.7515$ using the average product market shares of theatre ticket and average admission price.

[7] The constant market size assumption is not unresonable given that my data spans only 40 weeks

[8] In the estimated equation in Davis (2002), the dependent variable is $\delta_j = ln(s_j) - ln(s_0)$ and explanatory variables include mainly days of week dummies, linear and quadratic terms of "week at theater", and interactions of price with theater.

$Pirated_j$ is an indicator variable which equals 1 if $g(j) \in \{2, 3\}$, i.e. product j is from an illegal source(download). Therefore, a_i denotes the individual-specific difference in the mean valuation of legal movies and pirated movies. Here I model a_i with the following structure:

$$a_i = a_{0i} + r_{genre} \tag{4.3}$$

Where a_{0i} is the individual-specific constant term, and r_{genre} captures the genre-specific difference in taste for piracy. The estimates of r_{genre} indicates the relative amenability of each movie genres to piracy.

HQ_j is a dummy variable for "high quality", which equals 1 if $g(j) \in \{0, 1, 3\}$ (i.e., consumers choose to watch a movie through one of three high-quality channels: theatre, DVD, or high-quality download). The corresponding ψ_i measures individual-specific tastes for high quality. H_j is a dummy variable for the product to have the "watched at home" attribute. It equals 1 if $g(j) \in \{1, 2, 3\}$ (i.e., consumers choose to watch through DVD or either low-quality or high-quality download). γ_i measures individual-specific taste for home watching experience. Because the seasonality in movie demand (Einav, 2007), I include the calendar week fixed effects τ_t to control for any general time-specific demand shocks. ε_{jt} is the unobservable product characteristics and ε_{ijt} denotes an idiosyncratic shock following a Type I extreme value distribution.

Word-of-mouth and Complementarity In the setting of the BLP model, pirated movies and paid movies are by construction substitutes only. However, complementarity might exist between piracy and the box office. A number of recent researchers have found evidence of such complementarity through various channels. This could come from several different sources, such as the sampling effect (Peitz and Waelbroeck, 2006; Kretschmer and Peukert, 2017), network effect (Peitz and Waelbroeck, 2006; Belleflamme and Peitz, 2014), observation learning (Newberry, 2016), or backward spillover on product discovery (Hendricks and Sorensen, 2009). One of the most important channels is from spreading of WOM. The importance of WOM in influencing consumer decisions has been extensively studied in economics and marketing literatures (Chevalier and Mayzlin, 2006; Liu, 2006; Gayer and Shy, 2003). It is particularly important in the context of the movie market. Moretti (2011) found that peer effects and social learning from WOM are important in movie consumption. Gayer and Shy (2003) showed evidence supporting the existence of network externalities in movie consumption.

Here, I model WOM as a bridge through which the complementarity between piracy consumption and sales is established. According to the previous literature (Liu, 2006), two aspects of WOM have been highlighted as the most important: volume and valence. *volume* measures the quantity of WOM generated, while *valence* focuses on the quality aspects that capture the positiveness of WOM. In my model, I measure the valence with the movie rating in IMDB. I assume that piracy does not influence the rating, and I put more emphasis on the ability of piracy to influence the volume of WOM. Specifically, to quantify the volume of WOM for a particular movie, I use the Google Trends search volume as a proxy. The Google Trends search index measures the popularity of topics according to the number of Google search queries.[9] The Google Trends index will proxy the WOM, especially online WOM, since WOM activities are usually associated with searches on Google. In the utility function in equation (4.2), I use the log of the Google Trends index to measure WOM_{mt}. Because word-of-mouth WOM_{mt} is affected by both illegal downloads and legal sales, it then allows piracy to have a spillover effect on current demand. The parameter ϕ measures the magnitude of WOM in influencing demand.

Evolution of word-of-mouth I assume that both concurrent piracy and legitimate consumption of a specific movie can help increase the current period volume of WOM through communication and influence with uninformed consumers. Evolution of WOM_{mt} is assumed to follow an AR(1) process. It is a function of movie title m's previous word-of-mouth WOM_{mt-1} and total views across all channels at time t. WOM_{mt} evolves as below:

$$WOM_{mt} = \rho WOM_{mt-1} + \kappa TotalViews_{mt} + \pi_t + \omega_m + \epsilon_{mt} \tag{4.4}$$

with $TotalViews_{mt}$ defined as the sum of movie m's views from all channels at time t. π_t is time fixed effects, ω_m is movie title fixed effects which control for unobservable time-invariant factors such as total advertisement and movie

[9] It should be noted that weekly Google trends index is a flow value.

quality.

Preference Heterogeneity Similar to the model setup in Chapter 3, consumers have heterogeneous tastes over a series of characteristics. The heterogeneous taste takes a random-coefficient logit form which is a standard in literature (Berry et al., 1995; Nevo, 2001).

For notational simplicity I collapse all product characteristics into a vector X_{jt}. $X'_{jt} = \left\{W'_{jt}, P_j, Pirated_j, HQ_j, H_j, WOM_{mt-1}\right\}$ and the same for their corresponding taste parameters $\beta_i = \{\eta_i, \alpha_i, \delta_i, \gamma_i, \phi, a\}$. Now let the dimension of new X_{jt} be $K \times 1$ with K being the total number of characteristics. The demand model described in equation (4.2) can be rewritten as:

$$u_{ijt} = X'_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \tag{4.5}$$

Now let us turn to consumer tastes. Consumers have heterogeneous taste in the model. The distribution of consumer tastes parameters for movie characteristics is modelled as multivariate normal: the taste of consumer i for characteristic k is denoted by $\beta_{ik} = \beta_k + v_{ik}\sigma_k$ where v_{ik} is a mean zero taste shock for characteristic k. To write in matrix form:

$$\beta_i = \bar{\beta} + \Sigma v_i \tag{4.6}$$

I did not include random coefficient on all characteristics, let $\mathcal{K}$ denote the set of characteristics that have random coefficients and suppose I have K characteristics that have random-coefficients. $v_i = \{v_{i1}, v_{i2}...v_{iK}\}$ is the vector form of unobservable consumer characteristics following a multivariate standard normal distribution. Σ is a scaling diagonal matrix.

To highlight consumer heterogeneity, equation (4.5) can be rewritten in the following form:

$$u_{ijt} = \underbrace{\delta_{jt}}_{\text{Mean Utility}} + \underbrace{\sum_{k\in\mathcal{K}} X_{jt,k} v_{ik}\sigma_k + \varepsilon_{ijt}}_{\text{Error Component}} \tag{4.7}$$

δ_{jt} is defined as "mean utility" similar to Chapter 3, and the remaining component of equation (4.7) represents the "random" or individual specific part of the utility.

$$\delta_{jt} = X'_{jt}\bar{\beta} + \xi_{jt}$$

Consumers can also choose the outside option to neither buy nor download any movies. The utility of outside option is defined as:

$$u_{i0t} = \varepsilon_{0t} \tag{4.8}$$

Consumers choose one among all options to maximize his utility. Since the error term ε_{jt} follows extreme value distribution, consumer i's choice probability of movie j at time t can be written as:

$$Pr_{ijt} = \frac{exp(\delta_{jt} + \sum_{k\in\mathcal{K}} X_{jt,k}v_{ik}\sigma_k)}{1 + \sum\limits_{j'} exp(\delta_{j't} + \sum_{k\in\mathcal{K}} X_{j't,k}v_{ik}\sigma_k)} \tag{4.9}$$

And the market share of product j is obtained by integrating individual choice probability over the distributions of consumer characteristics:

$$s_{jt} = \int Pr_{ijt} f(v_i; \Sigma) dv_i \tag{4.10}$$

Adding Movie Title Dummies I add movie title dummies to control for all time/channel-invariant characteristics of movies following Nevo (2001). As described in Chapter 3, one important benefit of including movie titles and time fixed effects is the improvement of fit of the model and help in correcting the potential bias caused by the correlation between observable movie characteristics and unobservable quality. I extract market specific deviation from mean valuation $\Delta\xi_{jt} = \xi_{jt} - \xi_m$ with $f(j) = m$. $\Delta\xi_{jt}$ will serve as the new econometric error term.

Let I_j be a $M \times 1$ vector of movie title dummies. The vector of characteristics X_{jt} can be separated into two parts: one part $X_m^{(1)}$ is time/channel-invariant and movie title-specific characteristics,[10] and the rest part $X_{jt}^{(2)}$. When movie titles dummies are added, δ_{jt} is modified accordingly:

$$\delta_{jt} = I'_j\theta + X_{jt}^{(2)'}\bar{\beta}^{(2)} + \Delta\xi_{jt}$$

[10] I use $X_m^{(1)}$ as a short-hand notation for $X_j^{(1)}$ when $f(j) = m$

Where θ is a $1 \times M$ vector of coefficients on movie title dummies representing mean quality corresponding to a movie title. Once movie title dummies have been introduced, the new vector of observable characteristics X_{jt} can not include time-invariant characteristics, because all variations in time-invariant variables are absorbed by these movie title dummies. Therefore, $X_m^{(1)}$ has to be dropped from in the above equation. Let θ_m be the mth element of θ. For each movie title m I have:

$$\theta_m = X_m^{(1)\prime}\bar{\beta}^{(1)} + \xi_m \tag{4.11}$$

Allow Heterogeneous Taste on Movie Titles In addition to the above-mentioned specifications as benchmark, I also experiment with other specifications. One extension is allowing for heterogeneous tastes for movie titles as in Chapter 3.

$$u_{ijt} = \delta_{jt} + \sum_{k\in\mathcal{K}} X_{jt,k}v_{ik}\sigma_k + \sum_{m=1}^{M} I_{j,m}w_{im}\sigma_w + \varepsilon_{ijt} \tag{4.12}$$

where $I_{j,m}$ denotes the mth element of I_j. Similarly to v_i , $w_i = \{w_{i1}, w_{i2}, ..., w_{iM}\}$ is another set of unobservable consumer characteristics representing taste for a given movie title, also following multivariate normal distribution: $w_i \sim N(0, \Omega_w)$, where Ω_w is the diagonal variance-covariance matrix. Because of the substantial number of movie titles, it would be computationally difficult to allow the standard deviations of random coefficients to be movie title-specific. For computational simplicity, I restrict the standard deviations of taste on movie titles to be the same for all titles. i.e. $\Omega_w = \sigma_w I$. Compared with the previous specification, this adds only one more parameter, σ_w, to estimate.

4.4 Estimation Results

The estimation procedure of the model is similar to that of Chapter 3. I use GMM method to estimate the model. Differentiation instruments following Gandhi and Houde (2016) is used to identify the random coefficients, the details of which have been described in Chapter 3. The econometric error term $\Delta\xi$ is interacted with instruments to form the GMM objective function. Model parameters are obtained by minimizing the GMM objective function. Lastly, the

AR(1) process of WOM is estimated outside of the main estimation procedure via OLS using movie-week level observations.

This section reports the estimation results of my models. I report the results of demand estimations of two specifications, including a random coefficient logit model without heterogeneous taste on title-specific unobservable quality, and a full random coefficient logit model including heterogeneous tastes for movie titles. The results are shown in Table 4.5. I also show the results on the law of motion for WOM, which is estimated outside of the main model.

I estimate two versions of the random-coefficient logit model. The first version adds random coefficients on observables including pirated, high quality, home, movie genres, MPAA rating, which is shown in column (1). The second version, which is the full version, allows consumers to have heterogeneous taste on movie titles by including additional movie random coefficients on movie title dummies. In terms of random coefficients, the estimates show that consumers have heterogeneity in genre characteristics, such as action, science fiction, and cartoons. In terms of MPAA ratings, not much heterogeneity seems to exist, as most standard deviation terms are not small.

Now, I turn to variables that are most important in determining substitution between piracy and sales. The standard deviation of the random coefficient in *pirated* is significant with a magnitude of 27.107, indicating evidence of substantial consumer heterogeneity in taste for piracy. The distribution of consumer taste in piracy is shown in Figure 4.3. Roughly 6.5% of all consumers have positive preference for piracy. Consumer heterogeneity in *home* is also large, with a standard deviation of 19.771. The standard deviation of the random coefficient on movie title dummies is also significant, but with a smaller magnitude of 5.321.

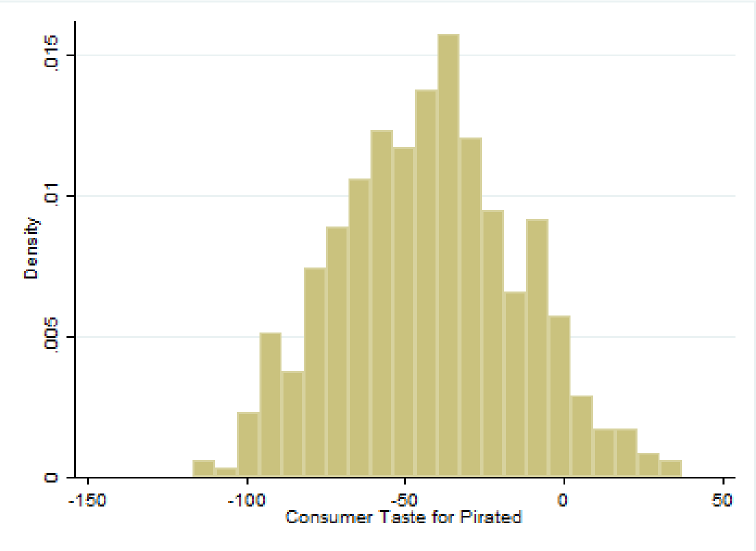
For movie genres, when I include the random coefficient on the movie title, the standard deviation of genres, such as science fiction, decreases, indicating that some persistence of choice in genres is actually because of consumer persistence in movie titles. As discussed in the previous section, the above-mentioned parameters play a crucial role in determining the substitution patterns. The standard deviation for a movie title is large relative to the standard deviation for piracy, we should expect consumers to have strong persistence with movie titles. If the movie *m* piracy version is eradicated, more consumers would be diverted to other channels of watching movies *m* (DVD and theatre),

Table 4.5: Demand Estimation Results

	(1)		(2)	
	Random Coefficient Logit		Random Coefficient Logit	
	No Title RC		Full Model	
	Mean β	Std Dev σ	Mean β	Std Dev σ
Pirated	-33.672***	33.111***	-41.430***	27.107***
	(4.421)	(4.056)	(7.578)	(4.341)
High Quality	12.099***	10.118***	1.510	5.095***
	(0.864)	(0.509)	(1.736)	(0.622)
Home	-44.080***	27.757***	-30.254***	19.771***
	(6.719)	(2.444)	(4.438)	(1.805)
Weeks after Release	-0.021		-0.079***	
	(0.020)		(0.019)	
Word-of-mouth	0.799***		0.931***	
	(0.091)		(0.092)	
Movie Title				5.321***
				(0.513)
Invariant Movie Charactersitics				
Rating	-0.036		0.605*	
	(0.052)		(0.285)	
Sequel	0.474	1.021	-3.768***	8.425***
	(0.309)	(1.311)	(0.469)	(0.368)
Genres				
Action	-2.529***	5.595***	-6.525***	6.907***
	(0.350)	(0.381)	(0.507)	(0.234)
Comedy	-0.247	1.300*	0.389	1.196*
	(0.314)	(0.605)	(0.457)	(0.515)
Drama	-1.589***	1.947***	-1.710***	3.600***
	(0.266)	(0.430)	(0.388)	(0.623)
Science Fiction	-40.485***	32.444***	-11.063***	9.204***
	(0.578)	(0.994)	(0.879)	(0.289)
Horror	-0.571	0.658	-1.181*	2.459***
	(0.403)	(0.927)	(0.581)	(0.529)
Cartoon	-1.889***	3.988***	-3.602***	4.696***
	(0.499)	(0.681)	(0.731)	(0.388)
Foreign	-0.626	0.000	-0.953	0.000
	(0.391)	(6.654)	(0.557)	(2.548)
MPAA Ratings				
PG	-0.461	0.000	0.149	0.618
	(0.349)	(0.649)	(0.526)	(0.679)
PG-13	-0.292	0.949	0.775*	1.913***
	(0.267)	(0.582)	(0.386)	(0.381)
R	-0.643**	0.000	-0.048	0.000
	(0.251)	(0.985)	(0.351)	(0.916)
Constant	-38.528***		-29.102***	
	(0.383)		(0.365)	
Movie Title Dummies	✓		✓	
Time Fixed Effect	✓		✓	
Observations	8617		8617	

Notes: Standard errors in parentheses. ***,**, and * denote statistical significance at 0.005, 0.01, and 0.05 levels respectively. Based on 8617 observations anf 40 week period in United States. For full model, movie title dummies, time fixed effects and interaction terms of Pirated with genres are included. Coefficients of time-invarying movie characterristics are obtained from regressing movie fixed effects on time-invarying movie characterristics.

Figure 4.3: Frequency Distribution of Consumer Taste for Pirated



Note: Frequency distribution of consumer taste for Pirated. About 6.5% of individuals' tastes on Pirated are positive.

while larger standard deviations for piracy will drive more consumers to other pirated movies of different titles. In my result, I find the second scenario to be most likely true in reality. However, strong persistence of preference also exists in *Home*, which to some extent helps, as more consumers may substitute DVDs. For the WOM effect, the estimated coefficients on volume of WOM is 0.931 in the current full model, which means for a movie product with mean WOM index of 150, an 10% increase in WOM leads to increase in utility by 0.038. The WOM coefficient is significant across all specifications, indicating that controlling for observable variables, there is strong evidence that consumer demand is influenced by WOM.

Law of Motion for WOM Using observations at the movie-week level, I estimate the law of motion to describe the evolution of WOM. Table 4.6 shows the result for the law of motion of WOM. All coefficients are precisely estimated. The coefficient on the lagged volume WOM_{mt-1} (ρ) is 0.423, suggesting that WOM is persistent over time, but also decays at a rate of 0.425. The estimated coefficient on the total views (κ) is 0.132, indicating that an increase in 1,000 views in contemporaneous total views increase the WOM index by 0.132. For a movie with average WOM of 150, the index would double with an increase of around 1.1 millions views.

Table 4.6: Law of Motion of Word-of-mouth

WOM_{mt-1}	0.4253^{***}
(ρ)	(0.0072)
$TotalViews_{mt-1}$(in thousand)	0.1323^{***}
(κ)	(0.0025)
Calendar Week FE	✓
Movie FE	✓
Observations	7983
Adjusted R^2	0.7825

Note: Observations are at movie-time level. WOM_{mt} is the volume of word-of-mouth collected from Google Trends. $TotalViews_{mt-1}$ represent the sum of $LegalViews_{mt-1}$ and $IllegalViews_{mt-1}$. and Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.5 Counterfactual Experiments

In this section, I simulate several counterfactual experiments to estimate the true cost of file sharing on box office revenue. First, to examine the heterogeneous effects of piracy on different channel of sales, I conduct a "no-piracy" experiment that eradicates all pirated movie products in my models and compares the counterfactual box office and DVD revenue and consumer welfare with the benchmark. Second, to examine the heterogeneous effect by piracy quality, I remove high-quality and low-quality separately to examine the heterogeneous effects on industry revenue. Lastly, I shut down the WOM effect channel to measure the magnitude of WOM effect of piracy on movie sales.

4.5.1 Full Removal: Heterogenous Effects by Distribution Channels

I now turn to my primary research question: What is the cost of digital piracy for the motion picture industry? How is the effect different by distribution channels. To quantify the effect of piracy on the industry revenue, I simulate a counterfactual experiment where I remove all pirated movies and recalculate counterfactual market shares using the estimated full model parameters in column (2) of Table 4.5. This experiment can be treated as an anti-piracy campaign

Table 4.7: Counterfactual Experiment: Removing All Piracy

(in $ billion)	Current	Remove All Piracy	Change	(%)
Box office Revenue	8.5278	8.7592	+0.2314	2.71%
DVD Revenue	1.4660	1.9934	+0.5274	35.98%
Consumer Welfare	30.1059	23.0533	-7.0506	-23.41%

Notes: This Panel report the result of industry revenue and consumer welfare after fully removal of all piracy products of all 40 weeks periods in United States. All first time period observations' last week views are not altered and I updated last week views for other periods considering the removal of piracy.

at the legislative level that bans piracy nationwide.

Updating Word-of-mouth Since the removal of piracy changes the illegal views of all movies with piracy availability, legal views will also change, as some pirated viewers switch to paid channels. The subsequent volume of the WOM needs to be adjusted accordingly. The updating procedure is very straightforward, I calculate counterfactual market shares sequentially, after counterfactual market shares in period t are calculated. The volume of WOM for movie m in period $t + 1$ can be calculated using its legal and illegal views in period t and its volume in period t. I then use volume in period $t+1$ to calculate the counterfactual market shares in $t + 1$.

Welfare Analysis Assuming price is the same after the no-piracy policy, I can then calculate counterfactual industry revenue as the product of market share times market size and price. Following Train (2009), consumer welfare at time t is calculated as the market size times the average of expected maximum value of indirect utility of simulated individuals:

$$CS_t = M\frac{1}{|\alpha|}\frac{1}{n_{ind}}\sum_{i}^{n_{ind}} E[maxu_{ijt}] \tag{4.13}$$

where a is the mean price coefficient used to translate utility into terms of money value, M denotes market size.

Result The result of the full removal counterfactual experiment is shown in Table 4.7. As the result shows, the elimination of piracy has different results on revenues from different channels.

The home-video market is severely affected by piracy, as home-video sales go up substantially by 35.98% after the removal of piracy. The total monetary loss on the home-video sales sums up to 527 million dollars during the 40-week period in the United States. The elimination of pirated movies results in an increase in box office revenue of $231 million dollars, less than half of the number in the home-video market. The figure represents a 2.71 % increase in current total box office revenue, only a mild impact compared with the home-video market.[11]

4.5.2 Partial Removal: Heterogenous Effects by Video Quality

Another interesting exercise is to examine the heterogeneous effect by piracy video quality. As shown in the previous preliminary regressions, there is potentially large degree of heterogeneity in quality across the two types of pirated products. In terms of cannibalization, the two types of piracy might play different roles: high-quality piracy is a closer substitute to legitimate sales, but low-quality piracy appears at an earlier time and interacts with ticket sales for more time than high-quality piracy. Because of this timing, low-quality piracy might be able to boost the box office more than high-quality piracy in terms of the WOM effect.

To quantify and compare the effects on revenue, I run two sets of counterfactuals. Specifically, I remove only high-quality piracy and low-quality piracy in two separate counterfactuals and compare the response of the revenue to the full eradication benchmark in the last section. I also run additional counterfactuals that remove high/low-quality piracy but do not update the change in

[11] The counterfactual result on box office revenue is different from the result discussed in Chapter 3, where removal of all piracy improves box office revenue by 1.09%. This discrepancy is mainly because the inclusion of new data source results in different estimated substitution patterns. Two additional sources of variations might help explain the observed discrepancy in couterfactuals. First, separating high-quality and low-quality piracy provides more variations in the choice set, allowing distribution of consumer preference to be estimated more precisely. Second, the inclusion of home-video sales provides an additional source of variations. A comparison of the results indicates that these additional variations result in a larger estimate on the standard deviation of random coefficient regarding movie titles in the results of Chapter 4. It generates more intra-movie substitution and contributes to the higher recovered revenue in Chapter 4's counterfactual experiment. To verify, I estimate a similar version of the model in Chapter 4, excluding home-videos sales data. I find that the counterfactual results with only quality data are very similar to the result in Table 4.7. So it is likely that additional variations on video quality mainly contribute to the observed difference.

WOM as a benchmark on pure cannibalization effects. A comparison between pure cannibalization results and the full results will allow me to see the effects of WOM. I also calculate the diversion ratio among channels. The diversion ratio is calculated by examining how the share of the removed channels (high quality/low quality) is allocated to different channels.

Table 4.8 presents the results on the industry revenue. There are several interesting findings. Removing high-quality piracy results in a much higher revenue increase than removing the low-quality piracy. Box office and DVD revenue increase by 1.29% and 5.64% after removal of high quality piracy, but the number is much smaller than the recovered revenue in the full removal benchmark. As shown in Table 4.9, the reason is that 56.6% of the previous consumers choosing high-quality piracy now switch to low-quality piracy, where, in the full removal, the low-quality option is not available, and many choose a paid option instead. As low-quality piracy is removed, the affected consumers almost exclusively switch to either high-quality piracy (54.4%) or an outside option (43.8%), so the actual gain of revenue is very limited for both box office and DVD.

When I allow WOM to update, another interesting result arises. While the removal of low-quality piracy causes DVD sales to increase, the box office revenue drops by 0.19%, compared to the previous result of a 0.03% growth. The difference highlights the positive role of early low-quality piracy in spreading of WOM, because of both its positive role in spreading WOM and the fact that they are relatively imperfect substitutes for any paid channels. The positive effect of WOM actually outweighs the limited negative cannibalization effects. This indicates that low-quality piracy can be used as a promotional tools for the studios. It's possibly profitable to orchestrate piracy deliberately in order to reap the free WOM generated by piracy[12].

[12] It is still worth to note that, for more than half of all pirated movies, low-quality piracy acts as the second-best option against the high-quality version. Despite the fact that removal of low-quality piracy brings no benefit, it is still worth noting that the efficacy of removal of high-quality piracy will be severely affected if low-quality piracy is not removed at the same time. As shown in Table 4.9, half of the high-quality users choose low-quality piracy if it is available, but when low quality is removed altogether, half will eventually switch to paid channels.

Table 4.8: Partial Removal by Video Quality

		Box office Revenue	DVD Revenue
	Update WOM	(in $ billions)	(in $ billions)
Current	-	8.5278	1.4660
		(-)	(-)
Remove All Piracy	No	8.7269	1.9731
		(2.34%)	(34.59%)
	Yes	8.7592	1.9935
		(2.71%)	(35.98%)
Remove High-Quality Piracy	No	8.7221	1.5616
		(2.28%)	(6.52%)
	Yes	8.6376	1.5486
		(1.29%)	(5.64%)
Remove Low Quality Piracy	No	8.5307	1.4879
		(0.03%)	(1.49%)
	Yes	8.5107	1.4744
		-(0.19%)	(0.57%)

Notes: This Table results of counterfactual experiment where I remove all piracy, remove only high quality (HQ) piracy or remove only Low quality piracy. For row 2,4,6 I shut down the word-of-mouth updating process, so the results reflect pure substitution effects. In row 3,5,7, I sequentially update word-of-mouth and new market shares are calculated using updated word-of-mouth, the difference in result between these two sets of experiments can be treated as the impact from word-of-mouth

Table 4.9: Diversion of Piracy Consumers by Destinations

Destination	No change in word-of-mouth			Update word-of-mouth		
(divertion ratio)	Remove All	Remove HQ	Remove LQ	Remove All	Remove HQ	Remove LQ
Box office	15.8%	10.0%	1.8%	18.4%	17.7%	-10.6%
DVD sale	25.8%	5.4%	3.1%	27.4%	4.8%	11.9%
Low-Quality Piracy	-	56.6%	-	-	56.5%	-
High-Quality Piracy	-	-	51.4%	-	-	51.3%
Outside Option	58.4%	28.0%	43.8%	54.3%	21.0%	47.4%

Notes: This Table shows consumer diversion ratio in the piracy removal experiment. The number are percentage of consumers that are diverted to certain option when their first choice are eliminated.

Table 4.10: Comparison of Counter-factual Revenue: With WOM vs No WOM

(in $ billions)	Box office revenue	DVD Revenue	Consumer surplus
No WOM	8.5081	1.4367	30.0517
With WOM	8.5278	1.4660	30.1059
Contribution of WOM Effects from Piracy	0.0197	0.0490	0.0543
(percentage)	(0.23%)	(2.00%)	(0.18%)

4.5.3 Positive Word-of-mouth Effect from Piracy

In the last counterfactual experiment, I quantify the magnitude of the WOM effect from pirated consumption. In the model, demand is influenced by the WOM. By generating more WOM, higher previous market share in a pirated movie can therefore benefit the demand for paid movies in the next period. Based on the estimates, the WOM effect is statistically significant. To directly assess the magnitude of WOM effects, in this counterfactual experiment, I shut down the WOM effect of piracy. Specifically I assume piracy no longer affect evolution of WOM, this is done by cutting piracy views from total views, re-calculate WOM and compare the counterfactual revenue with the benchmark to quantify the magnitude of spillover effect on the industry revenue.

The results are shown in Table 4.10. The contribution from the spillover effect on the industry revenue is relatively moderate. It increases the total box office revenue by 19.7 million dollars for 40 weeks in the US, representing 0.23% of the total box office revenue. The small magnitude in benefits to the box office may be attributed to the fast decay of movie attendance in theatres, as most downloads take place late in a movie's life cycle in theatres. Spillover effects happen too late to affect sales, as movie availability in theatres drops quickly. In terms of DVDs, the number increases compared to the box office, amounting to roughly 2% of the total DVD revenue.

4.6 Conclusion

This paper examines the heterogeneous effect of file sharing on movie box office and home-video sales revenue. To allow for flexible substitution patterns, I estimate a random-coefficient demand model of movies, allowing demand to be influenced by spillover from pirated consumption. Using a representative sample of download data from BitTorrent networks, I have several findings. First, file sharing reduces the total revenue of the motion picture industry from the box office by $ 231 million in total or 2.71% of the current box office in the US for my sample of 40 weeks in 2015. Unlike the box office, in the home-video market, DVD revenue increase by a surprising 36% when all piracy is removed. Second, different qualities of piracy play different roles. The positive WOM effects from low quality piracy outweigh its negative cannibalization effects, while for high quality piracy, its cannibalization effect dominates. Interestingly, removal of high-quality piracy alone does not solve the problem, as consumers that have a strong persistence for piracy will switch to low-quality piracy instead. In the end, I examine the magnitude of the WOM effect of piracy on box office revenue. I find that the WOM effect contributes to the box office revenue by a total of 68.7 million dollars for a 40-week period.

The findings of this paper highlight the potential danger in over-generalization in policy making regrading piracy. As the cannibalization effect of piracy varies greatly by video quality and distribution channels, anti-piracy resource and efforts should be directed towards the most harmful piracy type and the most vulnerable markets. For industry, these results also have important managerial implications. The proper estimate of the cannibalization and WOM effects of piracy helps managers in the motion picture industry better assess potential threats from piracy and design better strategies to harness the WOM effects of piracy.

Bibliography

Luis Aguiar and Bertin Martens. Digital music consumption on the internet: evidence from clickstream data. *Information Economics and Policy*, 34:27–43, 2016.

Andrew Ainslie, Xavier Drèze, and Fred Zufryden. Modeling movie life cycles and market share. *Marketing Science*, 24(3):508–517, 2005.

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.

Jie Bai and Joel Waldfogel. Movie piracy and sales displacement in two samples of chinese consumers. *Information Economics and Policy*, 24(3-4):187–196, 2012.

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

Paul Belleflamme and Martin Peitz. *Industrial organization: markets and strategies*. Cambridge University Press, 2010.

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

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

David Blackburn. Does file sharing affect record sales. *PhD dissertation. Harvard University*, 2004.

Michele Boldrin and David Levine. The case against intellectual property. *American Economic Review*, pages 209–212, 2002.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

Pradeep K Chintagunta. Heterogeneous logit model implications for brand positioning. *Journal of Marketing Research*, pages 304–311, 1994.

Pradeep K Chintagunta, Dipak C Jain, and Naufel J Vilcassim. Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research*, pages 417–428, 1991.

Pradeep K Chintagunta, Shyam Gopinath, and Sriram Venkataraman. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29(5):944–957, 2010.

Bram Cohen. The bittorrent protocol specification. 2015.

John T Dalton and Tin Cheuk Leung. Strategic decision-making in hollywood release gaps. *Journal of International Economics*, 105:10–21, 2017.

Brett Danaher and Michael D Smith. Gone in 60 seconds: The impact of the megaupload shutdown on movie sales. *International Journal of Industrial Organization*, 33:1–8, 2014.

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

Brett Danaher, Samita Dhanasobhon, Michael D Smith, and Rahul Telang. Converting pirates without cannibalizing purchasers: The impact of digital distribution on physical sales and internet piracy. *Marketing Science*, 29 (6):1138–1151, 2010.

Peter Davis. Estimating multi-way error components models with unbalanced data structures. *Journal of Econometrics*, 106(1):67–95, 2002.

Peter Davis. Spatial competition in retail markets: movie theaters. *RAND Journal of Economics*, pages 964–982, 2006.

Nicolas De Roos and Jordi McKenzie. Cheap tuesdays and the demand for cinema. *International Journal of Industrial Organization*, 33:93–109, 2014.

Arthur De Vany and W David Walls. Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of Cultural Economics*, 23(4):285–318, 1999.

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.

Chrysanthos Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424, 2003.

Tirtha Dhar and Charles B Weinberg. Measurement of interactions in nonlinear marketing models: The effect of critics' ratings and consumer sentiment on movie demand. *International Journal of research in Marketing*, 33 (2):392–408, 2016.

Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, 45 (4):1007–1016, 2008.

Jean-Pierre Dube, Jeremy T Fox, and Che-Lin Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267, 2012.

Liran Einav. Seasonality in the us motion picture industry. *RAND Journal of Economics*, pages 127–145, 2007.

Liran Einav. Not all rivals look alike: Estimating an equilibrium model of the release date timing game. *Economic Inquiry*, 48(2):369–390, 2010.

Anita Elberse and Jehoshua Eliashberg. Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Science*, 22(3):329–354, 2003.

Jehoshua Eliashberg and Steven M Shugan. Film critics: Influencers or predictors? *The Journal of Marketing*, pages 68–78, 1997.

David Erman. *Bittorrent traffic measurements and models*. PhD thesis, Blekinge Institute of Technology, 2005.

Amit Gandhi and Jean-François Houde. Measuring substitution patterns in differentiated product industries. 2016.

Amit Gayer and Oz Shy. Internet and peer-to-peer distributions in markets for digital products. *Economics Letters*, 81(2):197–203, 2003.

Ricard Gil. Revenue sharing distortions and vertical integration in the movie industry. *The Journal of Law, Economics, & Organization*, 25(2):579–610, 2008.

Ricard Gil. An empirical investigation of the paramount antitrust case. *Applied Economics*, 42(2):171–183, 2010.

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

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

David Godes and Dina Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4):545–560, 2004.

Robert G Hammond. Profit leak? pre-release file sharing and the music industry. *Southern Economic Journal*, 81(2):387–408, 2014.

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

Ken Hendricks and Alan Sorensen. Information and the skewness of music sales. *Journal of Political Economy*, 117(2):324–369, 2009.

Seung-Hyun Hong. Measuring the effect of napster on recorded music sales: Difference-in-differences estimates under compositional changes. *Journal of Applied Econometrics*, 28(2):297–324, 2013.

Benjamin Klein, Andres V Lerner, and Kevin M Murphy. The economics of copyright" fair use" in a networked world. *American Economic Review*, pages 205–208, 2002.

Tobias Kretschmer and Christian Peukert. Video killed the radio star? online music videos and recorded music sales. 2017.

Robert Layton and Paul Watters. Investigation into the extent of infringing content on bittorrent networks. *Internet Commerce Security Laboratory*, pages 8–10, 2010.

LEK. The cost of movie piracy. 2005.

Tin Cheuk Leung. What is the true loss due to piracy? evidence from microsoft office in hong kong. *Review of Economics and Statistics*, 95(3):1018–1029, 2013.

Tin Cheuk Leung. Music piracy: Bad for record sales but good for the ipod? *Information Economics and Policy*, 31:1 – 12, 2015.

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

Stan J Liebowitz. Pitfalls in measuring the impact of file-sharing on the sound recording market. *CESifo Economic Studies*, 51(2-3):435–473, 2005.

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

Yong Liu. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3):74–89, 2006.

Shijie Lu, Xin Shane Wang, and Neil Thomas Bendle. Does piracy create online word-of-mouth? an empirical analysis in movie industry. *Forthcoming, Management Science (2019)*, 2019.

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.

Jordi McKenzie. Revealed word-of-mouth demand and adaptive supply: Survival of motion pictures at the australian box office. *Journal of Cultural Economics*, 33(4):279–299, 2009.

Jordi McKenzie. The economics of movies: A literature survey. *Journal of Economic Surveys*, 26(1):42–70, 2012.

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

Charles C Moul. Measuring word of mouth's impact on theatrical movie admissions. *Journal of Economics and Management Strategy*, 16(4):859–892, 2007.

Aviv Nevo. Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The RAND Journal of Economics*, pages 395–421, 2000a.

Aviv Nevo. A practitioner guide to estimation of random-coefficients logit models of demand. *Journal of Economics and Management Strategy*, 9(4): 513–548, 2000b.

Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

Peter W Newberry. An empirical study of observational learning. *The RAND Journal of Economics*, 47(2):394–432, 2016.

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.

Motion Picture Association of America. Theatrical market statistics 2014. 2014.

Barak Y Orbach and Liran Einav. Uniform prices for differentiated goods: The case of the movie-theater industry. *International Review of Law and Economics*, 27(2):129–153, 2007.

Dominik Papies and Harald J van Heerde. The dynamic interplay between recorded music and live concerts: The role of piracy, unbundling, and artist characteristics. *Journal of Marketing*, 81(4):67–87, 2017.

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

Kathleen Reavis Conner and Richard P Rumelt. Software piracy: an analysis of protection strategies. *Management Science*, 37(2):125–139, 1991.

Imke Reimers. Can private copyright protection be effective? evidence from book publishing. *The Journal of Law and Economics*, 59(2):411–440, 2016.

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.

Joshua Slive and Dan Bernhardt. Pirated for profit. *Canadian Journal of Economics*, pages 886–899, 1998.

Michael D Smith and Rahul Telang. Competing with free: the impact of movie broadcasts on dvd sales and internet piracy. *MIS Quarterly*, 33(2):321–338, 2009.

Michael D Smith and Rahul Telang. Piracy or promotion? the impact of broadband internet penetration on dvd sales. *Information Economics and Policy*, 22(4):289–298, 2010.

Olaf van der Spek. Udp tracker protocol for bittorrent. 2015.

Koleman Strumpf. Using markets to measure the impact of file sharing on movie revenues. 2014.

Yuya Takahashi. Estimating a war of attrition: The case of the us movie theater industry. *American Economic Review*, 105(7):2204–41, 2015.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Michael Trusov, Randolph E Bucklin, and Koen Pauwels. Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of Marketing*, 73(5):90–102, 2009.

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

Feng Zhu and Xiaoquan Zhang. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2):133–148, 2010.

Appendix A

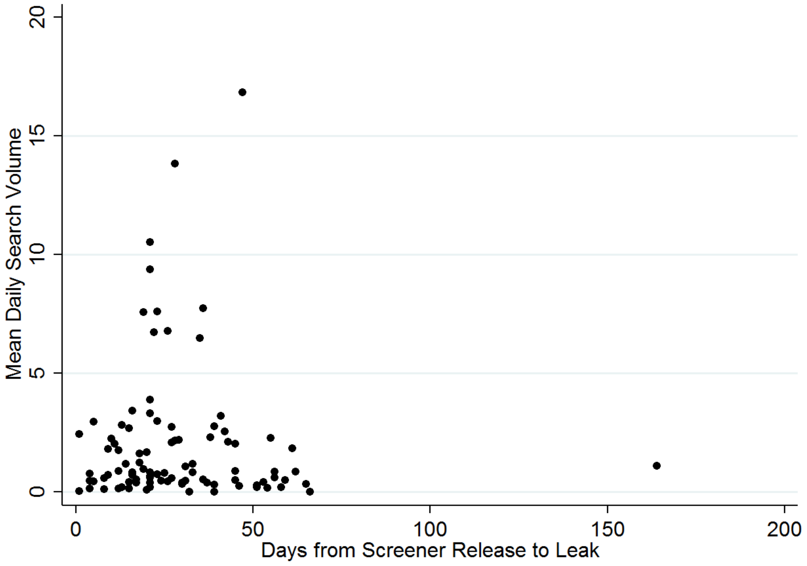
Chapter 2 Appendices

A.1 Additional Test: Are Popular Movies Leaked Earlier?

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 WOM, 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.

To be specific, I collect daily search volume of each leaked movie from 2010 to 2016, from 30 days prior to 30 days after the leak 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 leak for each movie. Figure A.1 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 leak. If movie pirates are influenced by demand factors, then we should expect the sign on release-to-leak time to be significantly negative, as more popular titles will have earlier leaks. The result from Table A.1 column (1) shows that the coefficient of interest is close to zero and not statistically signif-

Figure A.1: 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 A.1: 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

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

icant. In addition, I regress mean daily search volume on time between U.S. release and screener leaks; the results are shown in column (2). This regression specification addresses the possibility that better or more popular titles are well protected so earlier leaks is harder a nd they enjoy longer "screener piracy-free" windows as a result. The results in column (2) do not 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.

Appendix B

Chapter 3 Appendices

B.1 Preliminary Evidence

In this section I utilize the screener leaks incidence to provide some preliminary evidence on the potential inter-movie substitution effect of piracy. I construct a variable $NumLeaks_{it}$ which denotes the number of leaks incidence in specific weeks. For example, if three movies are leaked at time t, then the $NumLeaks_{it}$ is 3 for all observations during that week. To gauge the inter-movie substitution effect, I restrict the sample to unleaked movies. The baseline specification is a straightforward one:

$$ln(BoxOffice_{it}) = \alpha NumLeak_{it} + X_{it}\beta + \lambda_t + \xi_i + \epsilon_{it} \tag{B.1}$$

where the dependent variable is the log of weekly box office revenue, 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 denotes the movie fixed effects. ϵ_{it} is the idiosyncratic error term.

Is the magnitude of inter-movie substitution 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 the treatment to number of leaks that are of the same genre and MPAA rating as movie i, as movies in the same genres and MPAA ratings are generally considered as closer substitutes. Increased magnitude of the estimated coefficient would suggest that inter-movie effects are stronger for unleaked movies that are closer substitutes to the original leaked movie.

Table B.1: 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	✓	✓	✓
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. Based on 78730 observations. 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, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The regression result of the baseline model is shown in Table B.1. From the baseline result in column (1), an additional contemporaneous leaks will reduce the same-week box office revenue of other unleaked movies by about 3%. The results suggest that the inter-movie substitution effects are 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.

Columns (2) and (3) report results using leaks in the same genres and MPPA ratings, as movies falling under the same genres or MPAA ratings can be treated as closer substitutes. I find that an additional leak will reduce the box office revenue of unleaked movies in the same genre by 5%. In addition, an additional leak will reduce the box office revenue of unleaked movies with the same MPAA rating by 3.9%. The higher magnitude of the estimated coefficient confirms that inter-movie effects are stronger for closer substitutes.

B.2 Substitution Patterns

Table B.2 shows how consumers substitute into other products when their initial choice was eliminated for a selected number of movies in the United States at one particular time period. An examination of Table B.2 reveals basic patterns for intra-movie and inter-movie substitution effects. For example, after the elimination of pirated versions of the movie *Minions*, 26 % of downloaders of Minions chose to go to watch Minions in theatres. In addition, 4.5 % of downloaders chose to watch another cartoon, Inside Out. In general, except for blockbuster movies such as *Avengers: Age of Ultron* or *Jurassic World* that have few concurrent competitors and lots of concurrent downloads during release, most movies can only reclaim a small fraction of recovered revenue given full removal of piracy due to the timing of downloads and cross-substitution into other piracy or paid movies.

B.3 Collection of File sharing Data

This section provides a description of the procedures of downloading on BitTorrent and my data collection methodology.

It is very easy for BitTorrent users to download movie files online, they only need to find the .torrent file associated to the requesting file, the .torrent file is a descriptor meta-file containing important information to facilitate file transfer. Each .torrent file is indexed by a unique 40-bit identifier called torrent info-hash. The torrent file usually can be obtained from popular torrent search engines such as Piratebay.com, Torrentz.com, and so on. Upon receiving the .torrent file, the BitTorrent client software installed on a user's computer will help download the file automatically. The information on the .torrent file will guide the client to contact BitTorrent trackers and get a list of clients (so called "peers") who are also downloading the same file. The role of trackers is essentially directing the traffic in the Bit Torrent network. The tracker server does not keep the file content itself, instead it keeps tracks of who are downloading the file and informs the client who they should contact for file transfer. The tracker server keeps the current number of downloads for each registered torrent file and these numbers can be scraped by sending an HTTP or UDP request given

Table B.2: Substitution Patterns upon Eradication of Particular Movie's Piracy for US in week 15

Percent(%)		outide option	Paid Movies						Pirated Movies			
			Home	Inside Out	Jurassic World	Terminator	Minions	Spy	Home	Jurassic World	Minions	Spy
outside option		-	-	-	-	-	-	-	-	-	-	-
Paid Movies	Home	33.4069	-100	9.1391	0.5217	0.1418	54.009	0.0058	0.2265	0.0291	0.093	0.0011
	Inside Out	35.2809	0.1467	-100	0.5684	0.1522	60.7835	0.0061	0.2481	0.0309	0.1019	0.0011
	Jurassic World	77.0618	0.0085	0.5938	-100	8.0485	3.5092	0.0043	0.001	0.8649	0.0004	0.0005
	Terminator	80.4048	0.003	0.2091	9.9687	-100	1.2357	0.0495	0.0006	0.1373	0.0003	0.0006
	Minions	55.5423	0.4934	34.5919	1.1557	0.279	-100	0.0102	0.7127	0.0493	0.2927	0.0013
	Spy	30.3505	0.0005	0.0355	0.0255	0.2027	0.2096	-100	0.0001	0.001	0	9.4217
Pirated Movies	Home	59.8264	0.067	4.6948	0.0191	0.009	27.745	0.0002	-100	0.5691	3.639	0.0012
	Jurassic World	90.7006	0.0017	0.1201	3.2379	0.3738	0.7099	0.0006	0.1426	-100	0.0586	0.0103
	Minions	56.3862	0.065	4.5548	0.019	0.0089	26.9177	0.0002	8.2414	0.5626	-100	0.0012
	Spy	17.5318	0.0002	0.0114	0.0053	0.0044	0.0676	15.822	0.0006	0.0241	0.0002	-100

the info-hash of torrent file[1].

To obtain the estimates of weekly downloads on BitTorrent, I first collected the torrent files of each movie by web crawling the popular BitTorrent search engines. Every week the web crawler sent search queries about each movie on major BitTorrent search engines (Torrentz, KickassTorrents, isoHunt, The Pirate Bay, Extratorrent) and extracted the identifier (info-hash) of relevant movie torrent files from the torrent information page.

To ensure the extracted torrent files were truly relevant, I added several restrictions in the search queries:

- The file size has to be bigger than 200 MB.
- The file format has to be a video format such as mp4,avi,wmv,mkv,rmvb,etc.
- The file age cannot be older than the earliest release date of the movie[2].
- I filter out several keywords such as: trailer, featurette, soundtrack, OST, xxx, etc.

After obtaining a collection of info-hashes (torrent identifiers) for each movie, I collected a list of all working public BitTorrent trackers. There are currently 84 trackers in the list.

According to BitTorrent protocols, BitTorrent trackers will respond to HTTP or UDP GET requests with information including number of downloads, current number of seeders, number of leechers, and list of peers. The procedures of obtaining downloads for a movie go as follows:

- For each movie (e.g. Furious 7), searches the name plus filter in torrent search engine as shown in Figure 1.
- The webcrawler will collect the infohashes for all search results shown in Figure 2.

[1] Though trackers coordinate most of the downloads on BitTorrent, it is not the only way to download files on BitTorrent, downloading can happen in a decentralized way using DHT (Distributed Hash Table) without trackers ., I did not count download incidence in DHT for this study, because monitoring the DHT traffic is difficult. I am working on estimating the scale of downloads in DHT for possible correction on the download estimates.

[2] One exception is DVDSCR format, as screener piracy prior release has been found very frequently.

- Specifically, for each torrent file in search results, for example: "Fast.and.Furious.7.HDRip.XviD.AC3-EVO", the crawler will get access to the Torrent information page and record the infohash as shown in Figure 3:
  35a89cb57246dbdfdbf581403c33010d177a30dd

- The computer program then transforms the infohash into codes that can be understood by trackers (Bencode):

  ```
  5%A8%9C%B5rF%DB%DF%DB%F5%81%40%3C3%01%0D%17z0%DD
  ```

- For each tracker in the tracker list (e.g. http://www.todotorrents.com:2710/announce), the program sends a HTTP GET request[3]:

  ```
  GET http:///www.todotorrents.com:2710/scrape?info_hash=5%A8
     %9C%B5rF%DB%DF%DB%F5%81%40%3C3%01%0D%17z0%DD
  ```

- The tracker response contains information about the current number of seeders (complete), leechers (incomplete) and the number of completed downloads (downloaded) for the file:

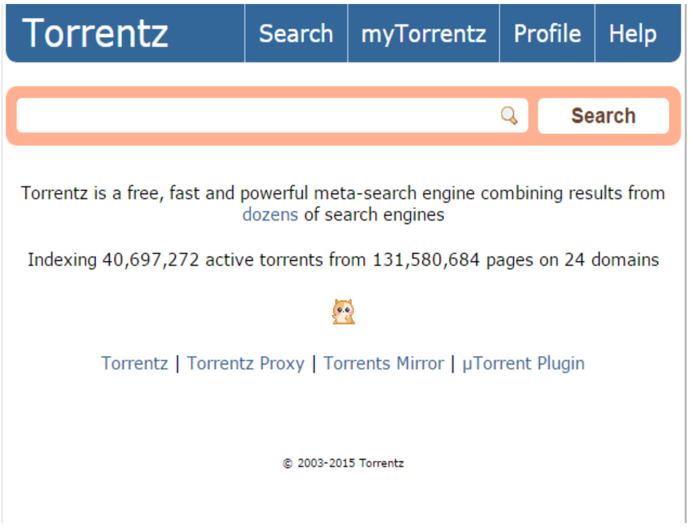
  ```
  {'files': {'5\xa8\x9c\xb5rF\xdb\xdf\xdb\xf5\x81@<3\x01\r\
     x17z0\xdd': {'downloaded': 659, 'complete': 3, '
     incomplete': 4}}}
  ```

  From the response, 'downloaded' indicates stock value of completed downloads, 'complete' refers to number of seeders, 'incomplete' is the number of leechers. The current number of downloads registered in this tracker for this torrent is 659.

- The program records this number and repeats previous steps for all trackers and all torrents.

I aggregated the number of downloads of each torrent file to get the current stock value of download count for each movie. Weekly flow value of downloads is obtained by taking the difference of download counts of consecutive

[3] The UDP request is similar so I omit the description of UDP.



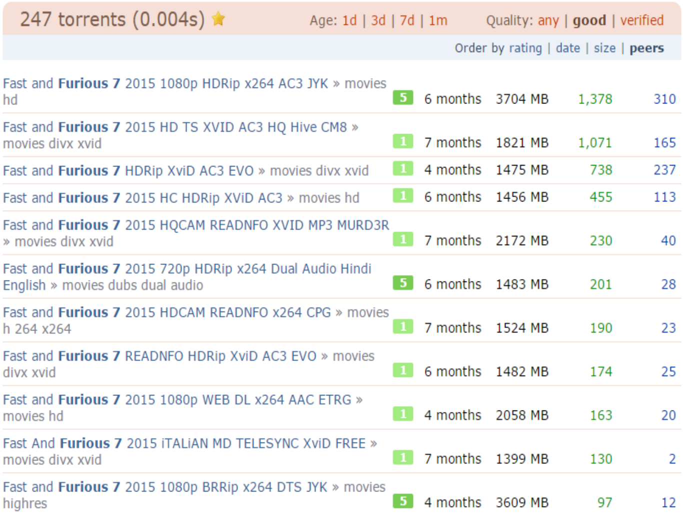
Torrentz is a free, fast and powerful meta-search engine combining results from dozens of search engines

Indexing 40,697,272 active torrents from 131,580,684 pages on 24 domains

Torrentz | Torrentz Proxy | Torrents Mirror | µTorrent Plugin

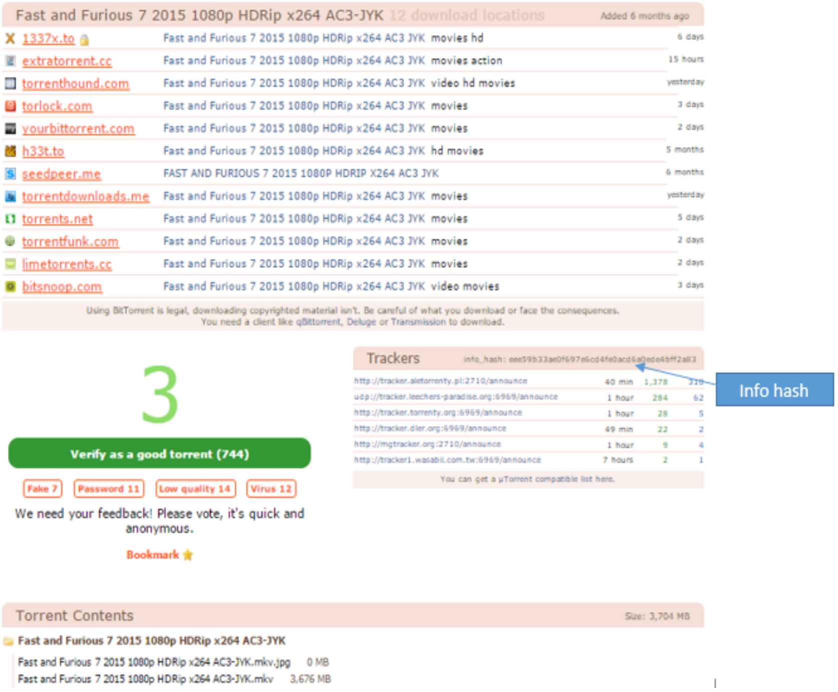
© 2003-2015 Torrentz

Figure B.1: Home Page of a Torrent Search Engine


247 torrents (0.004s) ★
Age: 1d | 3d | 7d | 1m
Quality: any | good | verified
Order by rating | date | size | peers

Fast and **Furious 7** 2015 1080p HDRip x264 AC3 JYK » movies hd	5	6 months	3704 MB	1,378	310
Fast and **Furious 7** 2015 HD TS XVID AC3 HQ Hive CM8 » movies divx xvid	1	7 months	1821 MB	1,071	165
Fast and **Furious 7** HDRip XviD AC3 EVO » movies divx xvid	1	4 months	1475 MB	738	237
Fast and **Furious 7** 2015 HC HDRip XViD AC3 » movies hd	1	6 months	1456 MB	455	113
Fast and **Furious 7** 2015 HQCAM READNFO XVID MP3 MURD3R » movies divx xvid	1	7 months	2172 MB	230	40
Fast and **Furious 7** 2015 720p HDRip x264 Dual Audio Hindi English » movies dubs dual audio	5	6 months	1483 MB	201	28
Fast and **Furious 7** 2015 HDCAM READNFO x264 CPG » movies h 264 x264	1	7 months	1524 MB	190	23
Fast and **Furious 7** READNFO HDRip XviD AC3 EVO » movies divx xvid	1	6 months	1482 MB	174	25
Fast and **Furious 7** 2015 1080p WEB DL x264 AAC ETRG » movies hd	1	4 months	2058 MB	163	20
Fast And **Furious 7** 2015 iTALiAN MD TELESYNC XviD FREE » movies divx xvid	1	7 months	1399 MB	130	2
Fast and **Furious 7** 2015 1080p BRRip x264 DTS JYK » movies highres	5	4 months	3609 MB	97	12

Figure B.2: Search Result


Fast and Furious 7 2015 1080p HDRip x264 AC3-JYK 12 download locations Added 6 months ago

1337x.to	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies hd	6 days
extratorrent.cc	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies action	15 hours
torrenthound.com	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK video hd movies	yesterday
torlock.com	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies	3 days
yourbittorrent.com	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies	2 days
h33t.to	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK hd movies	5 months
seedpeer.me	FAST AND FURIOUS 7 2015 1080P HDRIP X264 AC3 JYK	6 months
torrentdownloads.me	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies	yesterday
torrents.net	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies	5 days
torrentfunk.com	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies	2 days
limetorrents.cc	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK movies	2 days
bitsnoop.com	Fast and Furious 7 2015 1080p HDRip x264 AC3 JYK video movies	3 days

Using BitTorrent is legal, downloading copyrighted material isn't. Be careful of what you download or face the consequences.
You need a client like qBittorrent, Deluge or Transmission to download.

We need your feedback! Please vote, it's quick and anonymous.

Trackers info_hash: eee59b33ae0f697e6cd4fe0acd6a0ede4bff2a83

http://tracker.aletorrenty.pl:2710/announce	40 min	1,378	310
udp://tracker.leechers-paradise.org:6969/announce	1 hour	284	62
http://tracker.torrenty.org:6969/announce	1 hour	28	5
http://tracker.dler.org:6969/announce	49 min	22	2
http://mgtracker.org:2710/announce	1 hour	9	4
http://tracker1.wasabii.com.tw:6969/announce	7 hours	2	1

You can get a µTorrent compatible list here.

Torrent Contents Size: 3,704 MB

Fast and Furious 7 2015 1080p HDRip x264 AC3-JYK

Fast and Furious 7 2015 1080p HDRip x264 AC3-JYK.mkv.jpg	0 MB
Fast and Furious 7 2015 1080p HDRip x264 AC3-JYK.mkv	3,676 MB

Figure B.3: Torrent Information Page

weeks. This number can be treated as the total global downloads because the trackers' responses to SCRAPE requests contain no geographical information. Additional HTTP and UDP 'announce' requests were sent on a weekly basis to trackers to get a snapshot list of IP address of users currently downloading the files. I then used the IP address to identify the source country of downloaders and the share of downloads from each country. Country-specific weekly downloads were estimated using this geographic share information.

B.4 Reliability of the Download Estimates

Given the difficulty in estimating traffic on BitTorrent, concerns might be raised regarding the precision of the data collected in this paper, as indeed certain types of BitTorrent activities are omitted in my data collection procedures. For example, the data collection process was unable to track download activity happening through the trackerless protocol (DHT) and private trackers. It would be ideal to compare my data with more reliable statistics from sources such as Internet surveillance companies to further assess the quality of my data. While the data on downloading via BitTorrent for movies are scarce, I found yearly download statistics for a limited number of movies in 2015 estimated by the professional piracy tracking company Excipio. Table 12 shows the comparison of the download estimates in this paper and Excipio's estimates.

As the table shows, indeed there are some differences between the two columns. Generally my data tend to underestimate the download compared with Excipio's, my average is 28,155,435 compared with their average: 33,221,557. The correlation coefficient is 0.88. The high correlation suggested that variation in my data well matched the variation in the file-sharing network.

Table B.3: Comparison between Download Estimates from Excipio and this paper

Movie Title	Excipio's Estimates	Estimates in this paper
Interstellar(2014)	46,762,310	37,615,912
Furious 7(2015)	44,794,877	37,961,921
Avengers: Age of Ultron (2015)	41,594,159	36,418,665
Mad Max: Fury Road (2015)	36,443,244	29,645,492
Terminator: Genisys (2015)	31,001,480	30,399,370
San Andreas (2015)	25,883,469	20,376,013
The Minions (2015)	23,495,140	22,071,636
Inside Out (2015)	22,734,070	22,135,244
Jurassic World (2015)	36,881,763	27,094,954
American Sniper (2014)	33,953,737	24,423,823
Fifty Shades of Grey (2015)	32,126,827	34,442,676
The Hobbit: Battle Of The Five Armys (2014)	31,574,872	24,179,608
Mean	33,211,557	28,155,435
Correlation Coefficient: 0.88		

Notes: All download estimates number are up to Dec 31, 2015.

B.5 List of Trackers

```
udp://open.demonii.com:1337/announce
udp://9.rarbg.com:2710/announce
udp://tracker.leechers-paradise.org:6969/announce
udp://glotorrents.pw:6969/announce
http://bttracker.crunchbanglinux.org:6969/announce
http://i.bandito.org/announce
udp://www.eddie4.nl:6969/announce
udp://coppersurfer.tk:6969/announce
udp://shadowshq.eddie4.nl:6969/announce
http://tracker.dutchtracking.nl/announce
http://tracker.flashtorrents.org:6969/announce
udp://tracker.internetwarriors.net:1337/announce
http://www.todotorrents.com:2710/announce
http://pow7.com/announce
udp://inferno.demonoid.ph:3389/announce
http://torrent.gresille.org/announce
udp://tracker4.piratux.com:6969/announce
http://opensharing.org:2710/announce
http://anisaishuu.de:2710/announce
http://tracker.tvunderground.org.ru:3218/announce
http://tracker2.wasabii.com.tw:6969/announce
```

```
udp://mgtracker.org:2710/announce
udp://shadowshq.yi.org:6969/announce
http://bt.careland.com.cn:6969/announce
http://teentorrent.com:7070/announce
http://tracker.dler.org:6969/announce
http://bigfoot1942.sektori.org:6969/announce
udp://sugoi.pomf.se:80/announce
http://tracker.blazing.de:6969/announce
udp://exodus.desync.com:6969/announce
udp://open.nyaatorrents.info:6544/announce
http://tracker.tricitytorrents.com:2710/announce
udp://tracker.blackunicorn.xyz:6969/announce
http://tracker.ex.ua/announce
udp://bt.rutor.org:2710/announce
http://announce.torrentsmd.com:6969/announce
http://tracker.aletorrenty.pl:2710/announce
http://210.244.71.11:6969/announce
udp://tracker.torrenty.org:6969/announce
http://pubt.net:2710/announce
http://tracker.best-torrents.net:6969/announce
http://tracker.files.fm:6969/announce
http://retracker.uln-ix.ru/announce
http://bulkpeers.com:2710/announce
http://tracker3.infohash.org/announce
http://bt.mp4ba.com:2710/announce
udp://tracker.opentrackr.org:1337/announce
udp://p4p.arenabg.ch:1337/announce
http://retracker.telecom.kz/announce
http://tracker.mg64.net:6881/announce
http://tracker.trackerfix.com/announce
udp://zer0day.ch:1337/announce
udp://tracker.piratepublic.com:1337/announce
udp://tracker.sktorrent.net:6969/announce
http://xbtrutor.com:2710/announce
http://85.17.19.180/announce
http://tracker.bittorrent.am/announce
http://siambit.org/announce.php
http://retracker.krs-ix.ru/announce
http://tracker.baravik.org:6970/announce
http://tracker.tntvillage.scambioetico.org:2710/announce
http://tracker.mininova.org/announce
http://tracker.frozen-layer.com:6969/announce
```

```
http://www.mvgroup.org:2710/announce
http://bt.edwardk.info:6969/announce
http://share.camoe.cn:8080/announce
http://tracker.otaku-irc.fr/bt/announce.php
http://tracker.anirena.com:81/announce
http://tracker.dm258.cn:7070/announce
http://tracker.minglong.org:8080/announce
http://www.smartorrent.com:2710/announce
http://tracker.zaerc.com/announce.php
http://www.spanishtracker.com:2710/announce
http://www.todotorrents.com:2710/announce
http://www.tribalmixes.com/announce.php
http://funfile.org:2710/announce
http://mixfiend.com/announce.php
http://firesharing.altervista.org/announce.php
http://tracker.desitorrents.com:6969/announce
http://fafs.fansubanime.net/announce.php
http://all4nothin.net/announce.php
http://www.crnaberza.com/announce.php
http://www.gameupdates.org/announce.php
```

Zhuang Liu

Research Interests

Industrial Organization

Intellectual Property

Economics of Digitization

Applied Econometrics

Education

Ph.D. Economics, University of Western Ontario, Expected in 2019.

Dissertation Title: Essays in Industrial Organization.

Comittee: Prof. Salvador Navarro (Co-chair), Prof. David Rivers (Co-chair), Prof. Tai-Yeong Chung, Prof. Roy Allen

M.A. Economics, Simon Fraser University, 2013.

B.A. Economics and Applied Psychology, Lanzhou University, 2012.

Working Papers

Quantifying the Heterogeneous Effects of Piracy on the Demand for Movies, 2019.

Will the Leak Sink the Ship? Screener Leaks and the Impact of Movie Piracy, 2018

Works in Progress

Government Subsidy and Productivity Dynamics in Chinese Steel Industry (with Salvador Navarro and Jin Zhou), 2018

Effects of Digital Piracy on Sales, Entry and Welfare in Motion Picture Industry. 2018

Honors and Awards

Best PhD Student Paper Award, 1st Doctoral Workshop on the Economics of Digitization, 2017

Graduate Fellowship, University of Western Ontario, 2013-2017

Graduate Fellowship, Simon Fraser University, 2012-2013

First Class Scholarship, Lanzhou University, 2011

Conference Presentations

"Quantifying the Heterogeneous Effects of Piracy on the Demand for Movies"

16th International Industrial Organization Conferences (Rising Star Session), Indianapolis, 2018

2017 China Meeting of the Econometric Society, Wuhan, China, 2017

2017 Asian Meeting of the Econometric Society, Hong Kong, China, 2017

1st Doctoral Workshop on the Economics of Digitization, Munich, Germany, 2017

Tenth IDEI-TSE-IAST Conference on The Economics of Intellectual Property, Software and the Internet, Toulouse, France, 2017.

Western Economics 50th Anniversary Conference (Poster Session), London, Canada, 2016.

Canadian Economic Association 50th Annual Meetings, Ottawa, Canada, 2016.

UWO labor lunch seminar, London, Canada, 2015.

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

Canadian Economic Association 52th Annual Meetings, Montreal, Canada, 2018.

Research Experience

Research Assistant for Prof Salvador Navarro, 2017-2018

Research Assistant for Prof Mark Desjardine, 2016

Teaching Experience

Instructor, King's University College, University of Western Ontario

International Finance (Evaluations: 6.1/7.0), Winter 2017

Instructor, Department of Economics, University of Western Ontario

Intermediate Microeconomics II (Evaluations: 6.1/7.0), Summer 2017

Teaching Assistant, Department of Economics, University of Western Ontario

Intermediate Microeconomics II, Public Finance, Econometrics II, Mathematical Economics, Intermediate Microeconomics, Mathematical Economics(Graduate), Principles of Macroeconomics, Principles of Microeconomics

Teaching Assistant, Department of Economics, Simon Fraser University

International Finance, Principles of Macroeconomics, Principles of Macroeconomics

Skills

Stata, Python, Fortran, SQL

SAS Certified Base Programmer