Quantifying the Heterogeneous Effects of Piracy on the Demand for Movies

Zhuang Liu*

University of Western Ontario

February 10, 2019

Latest Version: http://publish.uwo.ca/~zliu328/jobmarketpaper.pdf

Abstract

The debate on digital piracy has attracted significant public attention. An accurate estimate of the loss due to piracy relies crucially on correctly identifying the substitution between pirated and paid consumption. Using a novel dataset of weekly piracy downloads collected from the BitTorrent network, I estimate a random-coefficient logit demand model of movies to quantify the effect of movie piracy on movie revenue from two distribution channels: box office and DVD sales. Counterfactual results reveal that digital piracy has heterogeneous effects on different channels of sales. When all piracy is removed, US box office revenue will only increase by 2.71% ($ 231 million) while US DVD sales will increase by 36% ($ 527 million) over 40 weeks in 2015. In addition, the effects on sales differ substantially by quality of pirated videos. I find strong evidence that one movie's piracy have negative indirect effects on other movie's revenue. Decomposition exercises show that the magnitude of piracy's indirect effects is much larger than the direct effects on its own revenue. Lastly, I allow piracy to have potentially positive effects on sales through word of mouth (WOM). The positive WOM effects from all pirated consumption have relatively moderate contributions to the industry revenue amounted to $ 68.7 million over 40 weeks in 2015.

*This paper was previously titled "A Structural Model of Movie Piracy with Word-of-mouth", "Estimating the Effects of fil-sharing on Movie Box office". I thank Salvador Navarro, David Rivers, Tai-Yeong Chung, Tim Conley, Paul Belleflamme, Marc Rysman, Mark Roberts, Tin Cheuk Leung, Giulia Pavan, Scott Orr, Christian Peukert, Jonathan Williams, Roy Allen and participants of the 16th IIOC, 2017 Asian Meeting of Econometric Society, 1st Doctoral Workshop on the Economics of Digitization, Tenth IDEI-TSE-IAST Conference on The Economics of Intellectual Property, Software and the Internet, 50th Annual Conference of the Canadian Economics Association, Western Economics 50th Anniversary Conference and UWO labor lunch seminar for their help and useful comments. All errors are mine.

1 Introduction

One of the most 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 with 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 policymakers and academic researchers about the potential harm of digital piracy on the relevant industries. However, there is no consensus yet on this question. On the one side, many people, especially copyright holders in the movie and music industries, view piracy as the major cause for declining sales. Several widely quoted industry investigations have indicated evidence of huge economic loss. For example, according to a 2011 piracy study by Business Software Alliance (BSA), software piracy costs the economy about 63.4 billion dollars in 2011.[3] Another study by the Institute for Policy Innovation (IPI) claims that digital piracy causes 58 billion dollars in actual US economic losses and 373,000 lost jobs.[4] However, the reliability of some of these estimates is under criticism for the unrealistic assumptions made in these studies.[5] To accurately estimate the true cost of piracy, it is crucial to have a thorough investigation of the substitution patterns between pirated goods, paid goods, and outside options.

In addition, there might be substantial heterogeneity in the effects of piracy, as 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). When substitutability

[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] https://phys.org/news/2012-05-software-piracy-billion.html

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

[5] For instance BSA admits that they assume that every download counts as one lost sale in their study. This relatively "naïve" methodology will inevitably inflate the estimated loss due to piracy.

is driven by product characteristics, different types of piracy will be likely to target different market segments and result in heterogeneous substitutability for paid goods from different sources. 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.

Besides the obvious cannibalization effect, it is possible that piracy can have positive effects on sales through various channels.[6] Several papers have focused on the word-of-mouth (WOM) aspect of piracy (Peukert et al., 2016; Lee, 2016) 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 true substitutability between legitimate and pirated consumption 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. To be specific, this paper complements the empirical literature on digital piracy and file sharing by estimating a random-coefficient demand model of movies using a novel dataset of movie downloads I collected on BitTorrent. Using the estimated model, I conduct a set of counterfactual experiments to quantify the heterogeneous effects of movie piracy on different channels of sales.

The main contributions of this paper are as follows. First, for movie released during a 40 weeks 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 paper avoids these concerns using direct information on downloads of pirated products.

Second, to the best of my knowledge, this paper is the first attempt to apply

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

aggregate download data on BitTorrent to estimate a random-coefficient logit demand model for movie piracy. The use of random-coefficient demand model brings several benefits. (1) linear reduced-form estimations employed in previous literature focus on exclusively **direct effect** of piracy (How one movie A's piracy affect its own revenue), while ignores the **indirect effects** (How movie A's piracy compete and cannibalize other movie's revenue) [7]. One advantage of logit demand models is that it naturally incorporate these indirect effects between movies. Using the estimated demand model, I find that the magnitude of indirect effects is on average 4 times as large as the direct effects. (2) As the reliability of estimates on industry loss relies heavily on identifying the true substitution patterns, a demand model that allows for flexible substitution patterns will generate a more accurate estimate on the effects of piracy. The rich set of variation in the choice set due to movie theatrical release and exit along with leaks of piracy provides sufficient source of variation for identification. By examining the implied substitution patterns, I can investigate more deeply into the heterogeneity of the effects of piracy, specifically how the piracy of one movie displaces sales and the factors affecting the displacement rate. (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.

Third, the detailed data on movie piracy allows me to study piracy along several dimensions. I classify piracy into different types and estimate a rich demand model to quantify the heterogeneous effects of piracy. I focus on three 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) sales; and (3) I decompose the effects of piracy on sales into a negative cannibalization effect and a positive WOM effect. Quantifying the spillover effect from piracy has important

[7] It's very hard 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.

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 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 relatively 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-weeks 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. Fourth, anti-piracy campaigns that remove piracy for an individual movie have limited benefits to its sales revenue, because most downloaders will substitute to other pirated movies. Fifth, I find strong evidence that one movie's piracy have negative spillover effects on other movie's revenue, and decomposition exercises show that the magnitude is on average 4 times as large as the direct effects on its own revenue. 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 paper is organized as follows. Section 2 provides an overview of the relevant literature. Section 3 provides background information on piracy and file sharing. Section 4 describes the data and some preliminary evidence. Section 5 discusses the model. The estimation procedure and results are presented in Sections 6 and 7. Section 8 gives the results of the counterfactual experiments, and Section 9 concludes the paper.

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), 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, this paper also decomposes the total effects of piracy into a pure substitution effect and a positive spillover effect which is not considered by Leung (2013).

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)). Most similar papers in the context of movie is 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 attempt 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, 2016), 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.

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.

Lastly, 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.

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.[8] 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 popula-

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

tion 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[9], by 2011 the traffic of the second most popular file sharing tool E-Donkey was only 2.6%. [10] 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 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. In Appendix B and C, I discuss reliability of download estimates and possible adjustment to include illegal streaming.

4 Data

4.1 Data Description

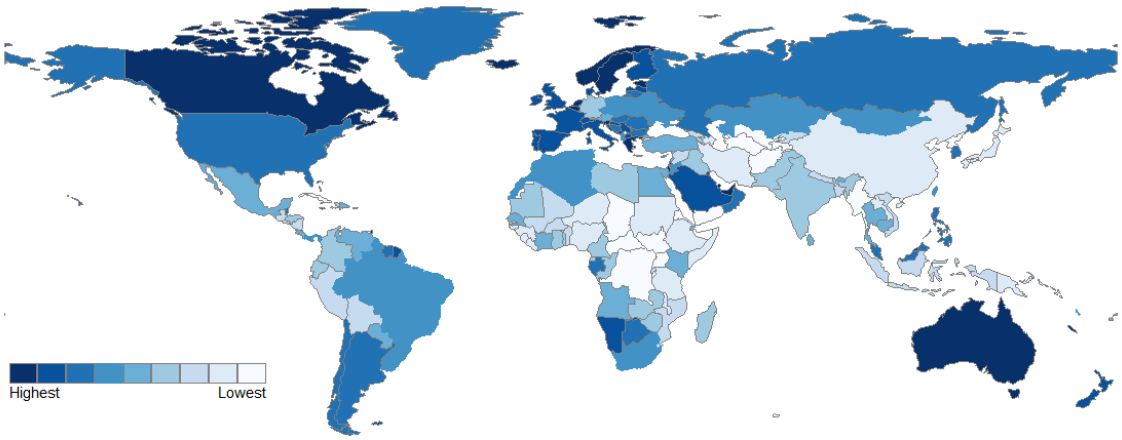
In this paper, I combine data from several data sources. I assemble a dataset including weekly information on movie-level box office sales, downloads, and sales on DVD in United States. I collected data covering a 40-week period from March 27th to December 27th in 2015. The data contain weekly downloads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

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,

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

[10] ISPreview, http://www.ispreview.co.uk/story/2011/05/18/bittorrent-p2p-filesharing-dominates-eu-broadband-isp-internet-traffic.html

Figure 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.

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.[11] Using the identifier, I collect weekly US Google trends index in sampling period for each movie.[12] The summary statistics are shown in the next section.

4.2 Preliminary Data Pattern

Figure 4 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,[13] but they are only mildly correlated with Internet speed.[14]

I match the box office data with the collected file sharing data. Table 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 advertisement campaigns. Most best-seller movies also appear to be the most downloaded.

Heterogeneity in Movie Piracy Pirated movies come in different formats/versions. Table 2 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 subsequent release of the *Telesync* version significantly improves the quality of the CAM version, but it is still inferior

[11] 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*.

[12] 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.

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

[14] 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 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 2: 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

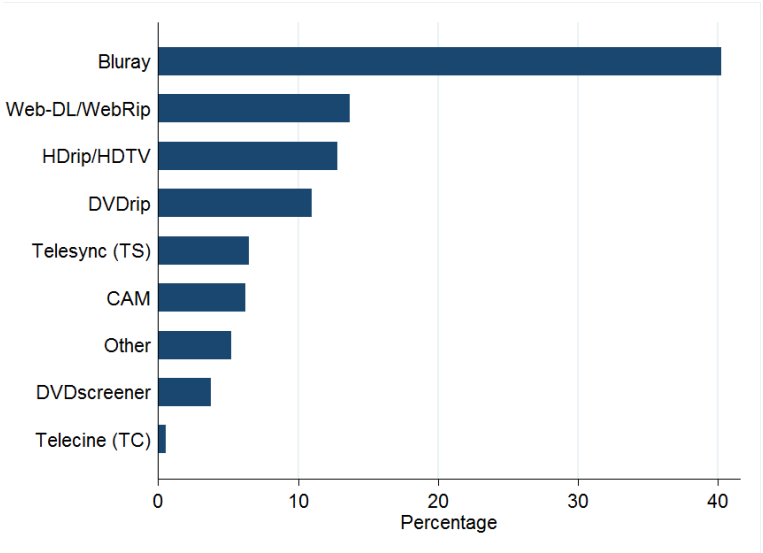
Notes: Descriptions are abridged from the Wikipedia page of each format. The classfication of quality is based on source, any source from bootleg recording is classified as Low quality.

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 2 shows the distribution of the torrent file formats.

Table 3 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.

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 few weeks around the release date. Therefore, movie sales concentrate at the beginning of the release. The common showing period of a movie is about 6-10 weeks. For blockbuster movies, the box office revenue of the opening week usually accounts for around 20% of the total box office revenue. Figure 3 shows the

Figure 2: Torrents Files Format Distribution

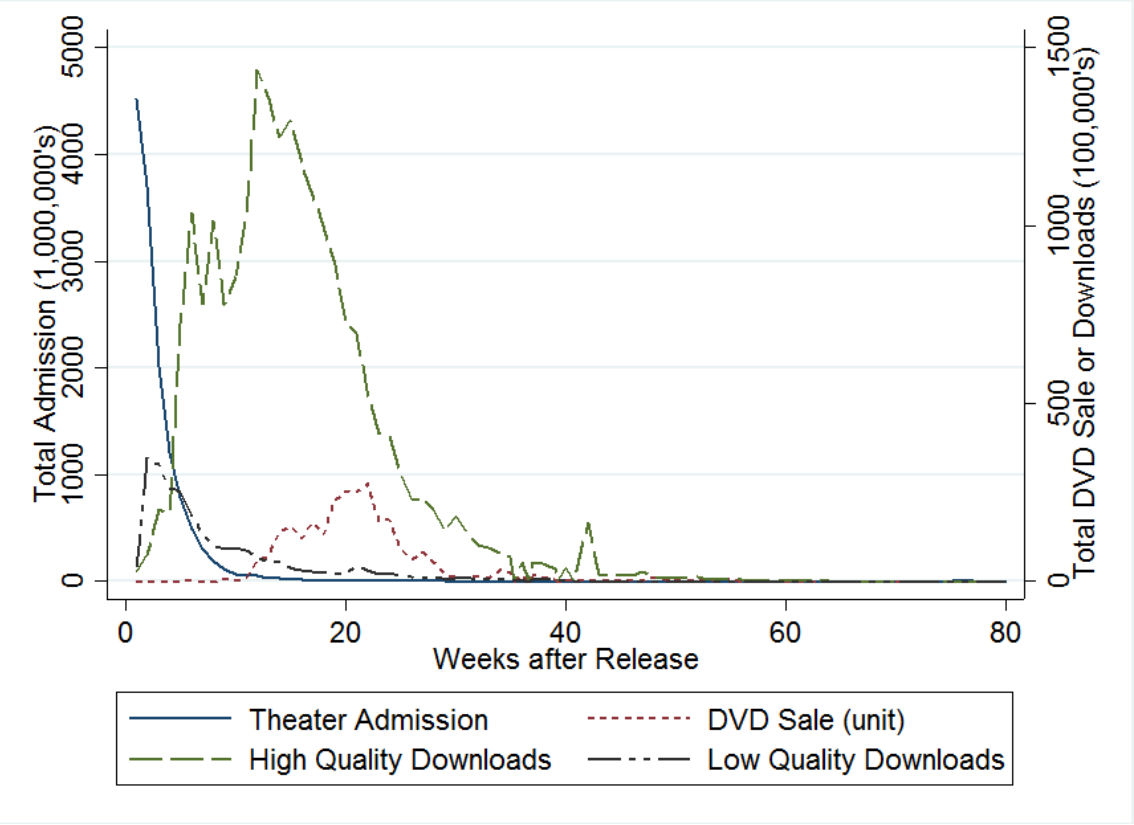


Notes: Figure 2 shows the distribution of torrent files formats. Most common formats are included in my dataset. Category Other includes other uncommon formats, and torrents that shows no information of format. For those with unknown format, I impute their quality use their appearance time, if it come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

Table 3: 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 3: Theatre Admission, DVD Unit Sale and Downloads by Weeks after Release



pattern of the average weekly audience (in millions) compared to other channels, including DVD sales and downloads (in hundred 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. After the close of the theatrical window, studios release the DVD into the market, and sales peak at around 20 weeks after the initial theatrical release date and gradually decrease after that. In terms of piracy, low-quality downloads quickly become available, but downloads are much lower in scale compared to high-quality piracy. Low-quality piracy is quickly substituted by the high-quality version, high-quality downloads have a larger scale than the low-quality counterpart and DVDs. It exhibits a more persistent pattern, partly because of the continuous supply of better-quality torrent files in the later period.

An important aspect to highlight is that, for most movies, the majority of downloads, especially those of high-quality format, happen after the end of the theatrical run. Thus, the download and box office do not have much overlap in time. Further-

more, the quality of downloaded movies is hardly comparable to the movie quality in the theatre. Judging by these facts, one conjecture is that the movie's own downloads might not displace its box office sales by much. However, the fact that a movie's own downloads do not overlap with its box office sales does not mean file sharing is not hurting studio revenue. Though the effect on the box office sales might be low, two other potential effects exist. One is to displace its own sales revenue on DVD/Blu-ray, and another is to displace the similar movies that release later. A simple decomposition might be useful to illustrate this point:

$$\underbrace{\frac{\partial R}{\partial D_j} = \frac{\partial \sum_{j'} R_{j'}}{\partial D_j}}_{\text{Total Effect}} = \underbrace{\frac{\partial R_j}{\partial D_j}}_{\text{Direct Effect}} + \underbrace{\frac{\partial \sum_{j' \neq j} R_{j'}}{\partial D_j}}_{\text{Indirect Effect}}$$

Let R be the industry revenue and it can be expressed as the sum of all individual movies R_j. The effect of movie j's piracy D_j on total industry revenue can be decomposed into two part, one is the direct effect $\frac{\partial R_j}{\partial D_j}$, which measures how movie j's piracy affects movie j's own revenue. Plus a indirect 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 previous literature, empirical research has exclusively focused on the direct effect. The magnitude of indirect effect remains unclear as most research assumed that each movie can be treated as segregated market, competition between movies is assumed to be zero. This is understandable since most research uses a linear reduced-form regression framework, it's very hard to incorporate competition effects as it requires including a full sets of each movies' piracy on the right-hand side of the equation, which are computationally infeasible in reality.

The use of logit demand model, on the other hand, provides an opportunity to examine indirect effects of piracy, as competition between movies are naturally incorporated in the model. I will conduct a formal decomposition exercise to quantify the magnitude of direct and indirect effects of piracy in Section 9.

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.

A shown in Table 3, 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:

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

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.[15]

Table 4 shows the results of these regressions. Column (1) reports estimates without the calendar week fixed effects, while column (2) includes full sets of fixed

[15] 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: Preliminary Regression Estimates of impact of piracy by quality

Dependent Variable:	(1) $ln(Boxoffice_{it})$	(2) $ln(Boxoffice_{it})$
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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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.

4.3 Summary Statistics

Before describing the model, I first describe the final data used for estimation of the model.[16] 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.

Table 5 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 5 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 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.

5 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 de-

[16]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 5: 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.

mand 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, I observe 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 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.[17]

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

$$u_{ijt} = W'_{jt}\eta_i + \alpha P_{jt} + a_i Pirated_j + \psi_i HQ_j + \gamma_i H_j + \phi ln(WOM_{mt-1}) + \tau_t + \xi_{jt} + \varepsilon_{ijt} \quad (2)$$

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

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

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).[18] 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.

$Pirated_j$ 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:

$$a_i = a_{0i} + r_{genre} \tag{3}$$

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

[18] 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.

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.[19] 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 (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

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

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:

$$WOM_{mt} = \rho WOM_{mt-1} + \kappa TotalViews_{mt} + \pi_t + \omega_m + \epsilon_{mt} \tag{4}$$

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 quality.

Preference Heterogeneity 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).

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 (2) can be rewritten as:

$$u_{ijt} = X'_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \tag{5}$$

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:

$$\beta_i = \bar{\beta} + \Sigma v_i \tag{6}$$

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 (5) can be rewritten in the following form:

$$u_{ijt} = \underbrace{\delta_{jt}}_{\text{Mean Utility}} + \underbrace{\sum_{k\in\mathcal{K}} X_{jt,k} v_{ik}\sigma_k + \varepsilon_{ijt}}_{\text{Error Component}} \tag{7}$$

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 (7), 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 Action movies dummies. 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 download 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:

$$u_{i0t} = \varepsilon_{0t} \tag{8}$$

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:

$$Pr_{ijt} = \frac{exp(\delta_{jt} + \sum_{k \in \mathcal{K}} X_{jt,k} v_{ik} \sigma_k)}{1 + \sum_{j'} exp(\delta_{j't} + \sum_{k \in \mathcal{K}} X_{j't,k} v_{ik} \sigma_k)} \tag{9}$$

And the market share of product j is then:

$$s_{jt} = \int Pr_{ijt} f(v_i; \Sigma) dv_i \tag{10}$$

Adding Movie Title Dummies 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,[20] and the rest part $X_{jt}^{(2)}$. When movie titles dummies are added, the demand model specified in equation (7) 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 dropped from in

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

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

$$\theta_m = X_m^{(1)'}\bar{\beta}^{(1)} + \xi_m \tag{11}$$

Allow Heterogeneous Taste on Movie Titles In addition to the above-mentioned specifications as benchmark, I also experiment with other specifications. One possible extension is allowing 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 I need in the setting 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 will help 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 (7) by adding an additional random component:

$$u_{ijt} = \delta_{jt} + \sum_{k\in\mathcal{K}} X_{jt,k}v_{ik}\sigma_k + \sum_{m=1}^{M} I_{j,m}w_{im}\sigma_w + \varepsilon_{ijt} \tag{12}$$

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.

The magnitude of σ_w also serve as a good test of the degree of competition between movies. As σ_w play a crucial role in determining the substitution patterns. Holding other model parameters equal, if σ_w becomes larger, consumer preference will have stronger correlation with movie titles, and we have weaker competition between movies. If movie m's piracy version is removed, more consumers would be diverted to other channels of the same movie m (DVD, theater). In contrast, larger value of σ_P will drive more consumers to other pirated movies of different titles, it increase competition between movies because movies especially like pirated movies become more substitutable.

6 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:

$$(\hat{\Sigma}, \hat{\sigma_w}) = \underset{\Sigma, \sigma_w}{argmin} \Delta\xi(\Sigma, \sigma_w)' Z \Phi^{-1} Z' \Delta\xi(\Sigma, \sigma_w) \tag{13}$$

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:

$$\delta_{jt}^{n+1} = \delta_{jt}^{n} + ln(s_{jt}) - ln(S(X_{jt}, \delta_{jt}^{n}; \Sigma, \sigma_w)) \tag{14}$$

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

$$S(X_{jt}, \delta_{jt}^n; \Sigma, \sigma_w) = \frac{1}{n_{ind}} \sum_i Pr_{ijt}(X_{jt}, v_i, w_i; \beta, \Sigma, \sigma_w) \tag{15}$$

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 $\{\Sigma, \sigma_w\}$:

$$E[Z\Delta\xi(\Sigma, \sigma_w)] = 0 \tag{16}$$

where Z is a set of instruments discussed in previous section. Once the non-linear parameters $\{\Sigma, \sigma_w\}$, together with $\{\bar{\beta}^{(2)}, \theta\}$ is obtained, at the second step I can use equation (11) to recover mean taste parameters for the title-invariant characteristics via the linear regression depicted before.

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

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

7 Results

This section reports the estimation results of my models. I report the results of demand estimations of two specifications, including a mixed-logit model without heterogeneous taste on title-specific unobservable quality, and a full mixed-logit model including heterogeneous tastes for movie titles. The results are shown in Table 6. 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

Table 6: Demand Estimation Results

	(1) Random Coefficient Logit No Title RC		(2) Random Coefficient Logit 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.

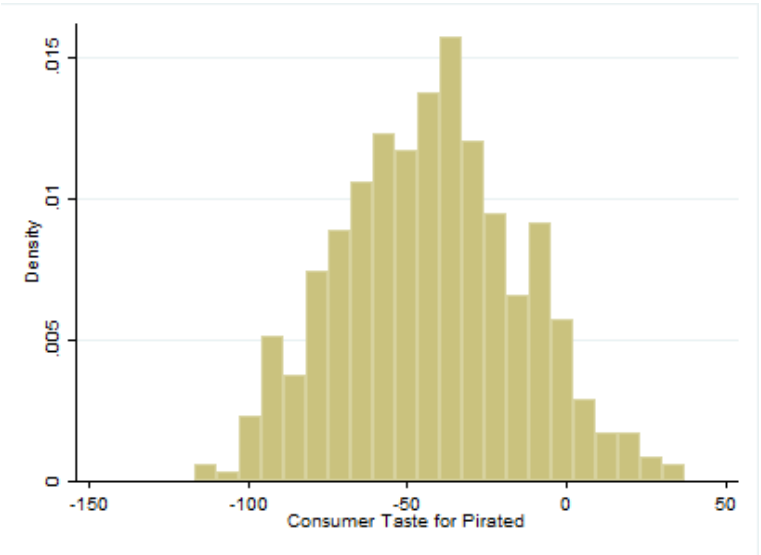
ratings, not much heterogeneity seems to exist, as most standard deviation terms are not small.

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. 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), 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 7 shows the result for the law of motion of WOM. All coefficients are precisely estimated. The coefficient

Figure 4: 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.

Table 7: 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, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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.

8 Counterfactual Experiments

In this section, I simulate several counterfactual experiments to estimate the true cost of file sharing on box office revenue. First, 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, I remove high-quality and low-quality separately to examine the heterogeneous effects on industry revenue. Third, I consider a firm-level anti-piracy experiment for each movie by removing just pirated versions of individual movie, while leaving other movies' pirated versions untouched. Lastly, I shut down the WOM effect channel to measure the magnitude of WOM effect of piracy on movie sales.

8.1 Remove All Piracy

I now turn to my primary research question: What is the cost of digital piracy for the motion picture industry? To quantify the total 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 6. This experiment can be treated as an anti-piracy campaign 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

Table 8: 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.

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:

$$CS_t = M\frac{1}{|a|}\frac{1}{n_{ind}}\sum_{i}^{n_{ind}} E[maxu_{ijt}] \tag{17}$$

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 8. The elimination of pirated movies on file sharing will result in a increase of box office revenue of $231 millions during 40 week period in US. The number represent a 2.71 % increase in current total box office revenue. Judging by the magnitude of this estimate for 40 weeks, the number for the whole year of 2015 is lower than most widely cited estimates.[21] On the other hand, DVD revenue benefit substantially from

[21]For example, in 2005, the Motion Picture Association of America (MPAA) and LEK Consulting estimated that total cost of US piracy to all studios to be around $ 2.7 billion, with $ 918 million due to Internet piracy. It is hard to make direct comparison between my estimates and these industry estimates directly. Because first it's difficult obtain an annual number for my estimates since I can not directly extrapolate the annual revenue loss from 40-weeks result. Second, the year of studies are different from each other. Readers should be cautious about comparing these estimates directly.

the removal of piracy: total sales go up by 35.98%.

If I use a "naive" way to estimate the revenue loss, assuming that one download equals one lost sale of movie ticket, then the estimated revenue loss amount to $1.25 billion for the same time periods, which is almost 5 times of the revenue loss in box office calculated in counter-factual experiment. Many widely cited industry studies have employed this "naive" methods in their estimation on piracy's cost. The result shows that using such methodology will substantially inflate the true loss of piracy. I also calculate the average displacement rate of pirated movies on both ticket and DVD sales. On average one download displaces ticket sales by 0.158 units. Since DVDs are estimated to be closer substitutes to pirated products, their displacement rate is higher, at about 0.258 units.

Consumer welfare decreases by $ 7.05 billion when we ban piracy, which is almost 4 times higher than the increase in motion picture industry revenue. There is a dead weight loss of $6.29 billions if we ban movie piracy. In general, the counterfactual result suggests several things: First, piracy does lower firm revenue, but also increases consumer welfare which is higher than the initial loss. So the piracy acts like a redistributive tool that use a small fraction of firm revenue to "subsidize" consumers who have low willingness-to-pay for movies and generates additional consumer welfare that is much higher than the lost revenue. Policies that eradicating movie piracy may result in the transfer of a large reduction of consumer welfare into small increase in industry revenue, resulting in socially inefficient outcomes from just the social welfare's point of view (holding movie provision fixed). One caveat is that the welfare analysis results are based on static models. It only answers the question regarding static efficiency but cannot tell much about dynamic efficiency. In the long run, when one considers reduced entry because of piracy, dynamic efficiency analysis could have different welfare implications. It would be interesting for future researchers to incorporate movie entry and evaluate the dynamic efficiency of piracy.

To assess the heterogeneity in responses to removal of piracy, I calculate the displacement rate and recovered revenue for each movie. Table 11 shows some descriptive statistics on the distribution of recovered revenue. There is substantial heterogeneity in terms of movies recovered revenue from piracy eradication because of the difference in position in characteristics space and level of competition faced. I

calculate each movie's revenue gain from the no-piracy counter-factual experiment. On average each movie's revenue increases by $1.24 million, the distribution is quite dispersed with a standard deviation of $6.11 million.

8.2 Partial Removal: High Quality vs. Low 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 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 9 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 10, 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

Table 9: Counterfactual Experiment

	Update WOM	Box office Revenue (in $ billions)	DVD Revenue (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

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.

However, 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 10, 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 10: Diversion of Piracy Consumers by Destinations

Destination (divertion ratio)	No change in word-of-mouth			Update word-of-mouth		
	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 11: Full Removal vs Partial Removal

($ millions)	Mean	Std Dev	Min	Max
Full removal	1.24	6.11	0.00	99.08
Partial removal	0.03	0.27	0.00	13.00

Decomposition: Direct vs Indirect Effect

($ millions)	Direct Effect (self)	Indirect Effect (others)
Average Recovered revenue	0.03	0.13

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. It is reported also as the "Direct Effect". I also calculate the sum of improved revenue on all other movies and it is reported under"Indirect Effects".

8.3 Partial Removal by Movie

The previous conterfactual experiment resembles the copyright protection at the public and legislative level, where policy tends to affect the whole industry. However, copyright protections are not always initiated by government or legislation. In recent years, private copyright protection initiated by firms targeting individual copyrighted work has become increasingly prevalent. As Reimers (2016) pointed out, such private copyright protections are effective in the book publishing industry. In the motion picture industry, studios also hire Internet surveillance companies to monitor and send DMCA (Digital Millennium Copyright Act) notices to take down torrent files on file sharing websites.

How effective are those private copyright protection efforts targeted to remove piracy for individual movies? Will downloaders substitute the paid version, other pirated movies, or simply outside options? To answer the question, I conduct a movie-level piracy removal counterfactual experiment. In this experiment, for each movie, I simulate a firm-level private copyright protection campaign, which eliminates

all pirated versions across all periods but leaves pirated versions of other movies untouched. I then calculate counterfactual market shares and counterfactual revenue increase for that movie.

Table 11 shows the comparison of recovered revenue per movie between this partial-eradication counterfactual experiment and the full eradication experiment. Not surprisingly, the average recovered revenue dropped to 0.03 million, less than 10% of the average recovered revenue by eradicating all piracy. In this counterfactual, most downloaders will choose the other available pirated movies or other similar movies instead because, in many cases, the availability of the movie in the theatres is low.

Table 12 shows the diversion ratio for a selected number of products in the US at one particular period. The diversion ratio shows us how consumers substitute into other products when their initial choice is eliminated. I find that the substitution between low-quality piracy for both tickets and DVDs are generally very limited. The second-best option for consumers choosing low-quality piracy is usually another low-quality movie. For example, when the low-quality pirated movie *Mission Impossible: Rogue Nation* is removed, almost 0% switch to buying the ticket; instead, some choose to download other low-quality pirated moves, such as *Avengers: Age of Ultron* (4.3%), but most of them simply choose the outside option (71.6%).

Not surprisingly, high-quality piracy has a higher substitutability for sales, especially DVD sales. For instance, 34% of high-quality piracy consumers of animation film *Home* are diverted to buying DVDs if the high-quality piracy is removed, compared to a much lower 0.9% that switch to buying tickets.[22]

Interestingly, Table 12 reveals that the substitution patterns is not restricted to substitution within the movie title. There are also notable indirect substitution effects of piracy across movie titles with similar characteristics. For example, after the elimination of high-quality piracy of the science fiction movie *Ex Machine*, 4.1% of affected consumers switched to the DVD of another science fiction movie *Jupiter Ascending*. The substitution patterns across movie titles implies significant positive externalities of these private anti-piracy campaigns. When studios fight piracy, they

[22]One thing to note is the availability of ticket sales, as many movie tickets are at the end of the theatrical run. The low number of screens indicates limited accessibility for consumers, to a large extent explaining why there are few diversions to ticket sales.

also benefit competitors by bringing a windfall on the revenue of other movies.

As Table 12 demonstrates, externalities from private copyright protection to the other movies are large in magnitude. On average, other movies gain 0.13 million dollars in total, much larger than the gain from the protected movie. So the magnitude of indirect effects is 4 times as high as the direct effects. The result to some extent indicates that the biggest threat to movie box office revenue is not the piracy of its the movie, but rather the movies whose downloads overlap the release window. For private copyright protection to secure the box office revenue, other studios' copyright protection efforts are equally important, so coordination and cooperation in copyright protection efforts may be beneficial to studios.

8.4 How Big is the Word-of-mouth Spillover Effect?

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, recalculate 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 13. 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.

Table 12: Substitution Patterns upon Removal of Particular Movie's Piracy

				Diversion Ratio (In Percent)																	
					Avengers			Ex Machina			Home				Inside Out		Jupiter Ascending		Mission: Impossible	
Movie title	Genre	Type	ID	outside option	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Avengers:	Action	Low Quality Piracy	1	73.5	-100.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.4	0.0	1.2	0.0	0.0	0.0	0.6
Age of Ultron		High Quality Piracy	2	0.1	0.1	-100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Ticket	3	27.8	0.0	0.0	-100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	2.0	0.0
Ex Machina	Sci-Fi	DVD	4	60.9	0.0	0.0	0.0	-100.0	4.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	1.7	4.1	0.0	0.0
		High Quality Piracy	5	57.9	0.0	0.0	0.0	3.5	-100.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	1.7	3.9	0.0	0.0
		Ticket	6	99.2	0.0	0.0	0.0	0.0	0.0	-100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Home	Cartoon	Low Quality Piracy	7	71.4	4.3	0.0	0.0	0.0	0.0	0.0	-100.0	0.0	0.0	0.3	0.0	1.2	0.0	0.0	0.0	0.6
		Ticket	8	71.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-100.0	14.4	3.1	0.1	0.0	0.0	0.0	0.1	0.0
		DVD	9	55.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	-100.0	12.3	0.0	0.0	0.0	0.0	0.0	0.0
		High Quality Piracy	10	45.1	0.2	0.0	0.0	0.3	0.3	0.0	0.0	0.9	34.1	-100.0	0.0	0.0	0.1	0.3	0.0	0.0
Inside Out	Cartoon	Ticket	11	83.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-100.0	0.0	0.0	0.0	0.3	0.0
		Low Quality Piracy	12	72.1	4.4	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.3	0.0	-100.0	0.0	0.0	0.0	0.6
Jupiter Ascending	Sci-Fi	High Quality Piracy	13	54.4	0.0	0.0	0.0	3.3	3.6	0.0	0.0	0.0	0.0	0.7	0.0	0.0	-100.0	3.6	0.0	0.0
		DVD	14	61.1	0.0	0.0	0.0	3.7	4.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	1.7	-100.0	0.0	0.0
Mission: Impossible	Action	Ticket	15	64.5	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-100.0	0.0
Rogue Nation		Low Quality Piracy	16	71.6	4.3	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.3	0.0	1.2	0.0	0.0	0.0	-100.0

Notes: This table reports the substitution patterns between different products in US in week 20. The table show in each row a scenario when one particular movie is removed in the choice set, which product does affected consumer substitute into. The number indicates fraction of affected consumer that choose the product in the column.

Table 13: 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%)

9 Conclusion

This paper examines the effect of file sharing on movie box office 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. The estimates are considerably smaller than the widely cited industry estimates that are often referenced in the press. Using the naïve methodology that assumes full sales displacement will inflate the true cost by 5 times. On average, each movie suffers a monetary loss of 0.5 million because of file sharing. However, the responses differ substantially by channel. 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. High quality is a closer substitute for sales, but 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. Third, the results of welfare analysis show that file sharing increases consumer welfare by a total of $ 7.05 billion; therefore, banning file sharing services will result in a dead-weight loss of $ 6.29 billion. In addition, removing piracy for individual movies have limited benefits to box office revenue because most downloaders just substitute other pirated movies. 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 serve to provide extra evidence to assist the resolution

of the current heated debate on controversial issues regarding intellectual property. For policymakers, the findings in this paper highlight the importance of considering outside options and substitutions in evaluating the effects 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 industry, these results also have important managerial implications. The proper estimate of the cannibalization and WOM effects of file sharing will help managers in the motion picture industry better assess their movie's vulnerability to piracy and better determine the optimal level of copyright protection given the supervision and litigation costs.

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.

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.

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 diss. 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 non-linear 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.

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

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.

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, Jörg 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 Collection of file sharing Data

This section provides a description of the procedures of downloading on BitTorent 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 an unique 40 bits 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 getting the .torrent file, the BitTorrent Client software installed on user's computer will help download the file automatically. The information on .torrent file will guide 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, tracker server don't keep the file content itself, instead it keeps tracks of who are downloading the file and tell a client who they should contact for file transfer. Tracker server keep the current number of downloads for each registered torrent file and these number can be scraped by sending an HTTP or UDP request given the info-hash of torrent file[23].

Now I describe my data collection methods. To obtain the estimates of weekly download on BitTorrent, I first collect the torrent files of each movie by webcrawling the popular BitTorrent search engines. Every week the web crawler will send search queries about each movie on major Bit Torrent search engines (Torrentz, KickassTorrents, isoHunt, The Pirate Bay, Extratorrent) and extract the identifier (infohash) of relevant movie torrent files from the torrent information page.

To ensure the extracted torrent file are truly relevant, I add 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.

[23]Though trackers coordinate most of the downloads on BitTorrent, it is not the only way to download file on BitTorrent, downloading can happen in a decentralized way using DHT without trackers, I did not currently count download incidence right now in DHT because monitoring the DHT traffic is difficult. I am working on a estimating of the scale of downloads in DHT for possible correction on the download estimates

- The file age can not be older than the earliest release date of the movie[24].
- I filter out several keywords such as: trailer, featurette, soundtrack, OST, xxx, etc.

After obtaining a collection of infohashes (torrent identifiers) for each movie, I collect a list of all working public Bit Torrent trackers. There are currently 84 trackers in the list.

According to Bit Torrent protocols, BitTorrent trackers will respond to HTTP or UDP GET request 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.
- 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[25]:

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

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

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

- 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. 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 will aggregate the number of downloads of each torrent file to get the current stock value of download count for each movie. Weekly flow value of download is obtained by taking difference of download count of consecutive 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' request is sent on weekly basis to trackers to get a snapshot list of IP address of users currently downloading the files. I then use the IP address to identify the source country of downloaders and the share of downloads from each country. Country-specific weekly downloads is estimated using this geographic share information.

Appendix B Reliability of the Download Estimates

Given the difficulty in estimating traffic on BitTorrent, concerns might be raised regarding the precision of the collected data in this paper, as indeed certain types of BitTorrent activities are omitted in my data collection procedures. For example, the data collection process is unable to track the 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

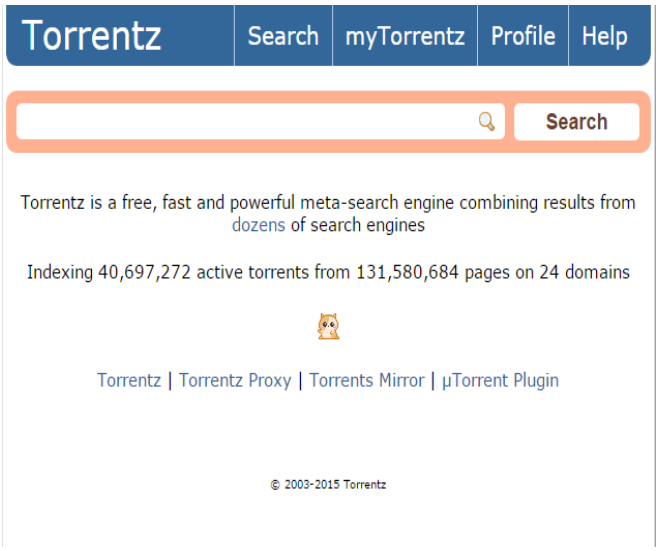


Figure 5: Home Page of a Torrent Search Engine

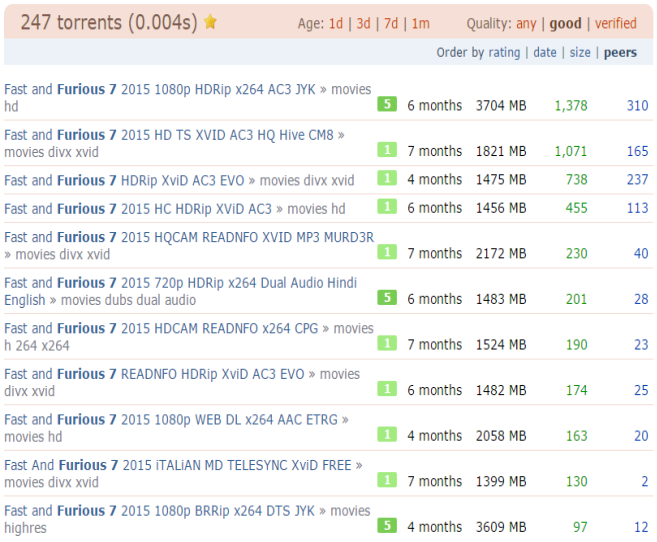


Figure 6: Search Result

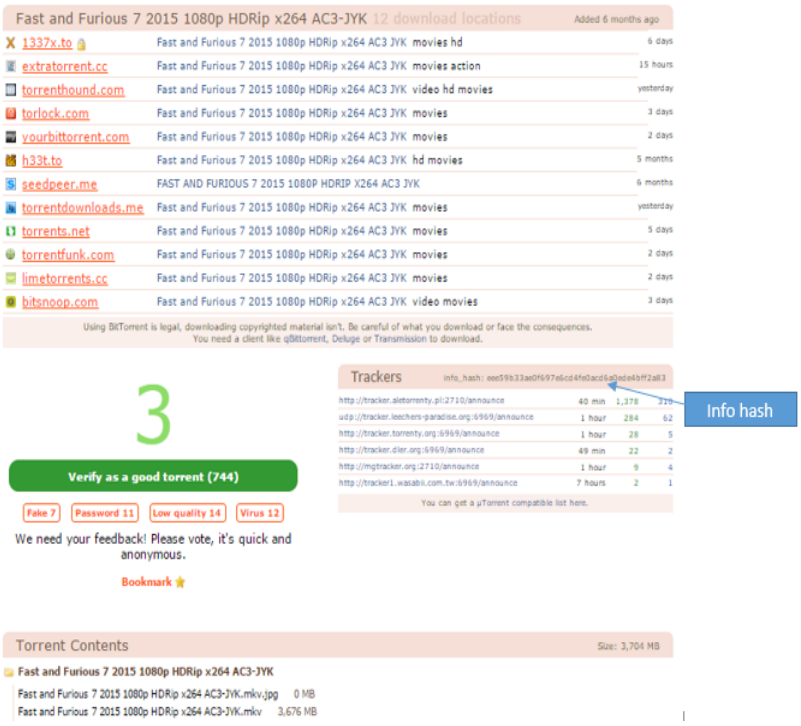


Figure 7: Torrent Information Page

Table 14: 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.

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.

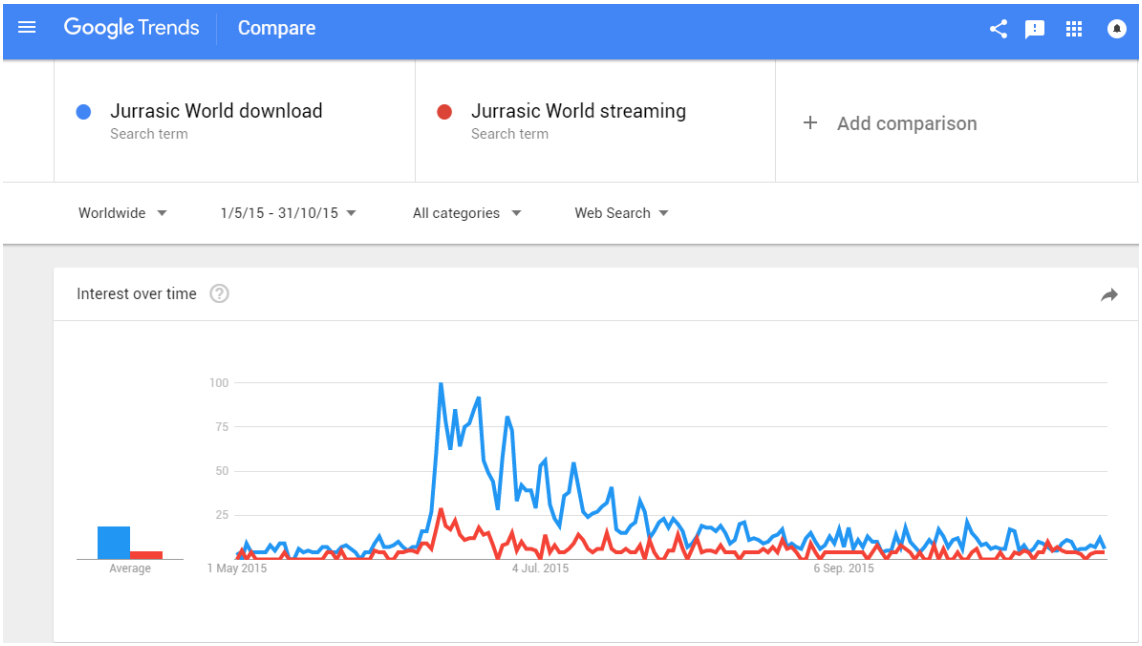
As the Table shows, indeed there are some difference between the two columns, generally my data tend to underestimate the download compared to Excipio's, my average is 28,155,435 compared with their average: 33,221,557. The correlation coefficient is 0.88. The high correlation suggest that variation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

Appendix C Illegal Streaming

With the emergence of Pirated streaming website like Popcorntime, Putlocker and Movie4k, many file sharing users have switched from downloading to streaming. In 2015, streaming has already taken up a significant proportion of total piracy activity. In order to taken into account the increasing popularity of illegal streaming service, the volume of illegal streaming need to be estimated. Unfortunately it is technically very difficult to monitor the movie streaming traffic.

To overcome the difficulties in direct estimation of streaming traffic, I choose to leverage search traffic data for streaming and downloading in Google Trends as

Figure 8: Screenshot of Google Trends

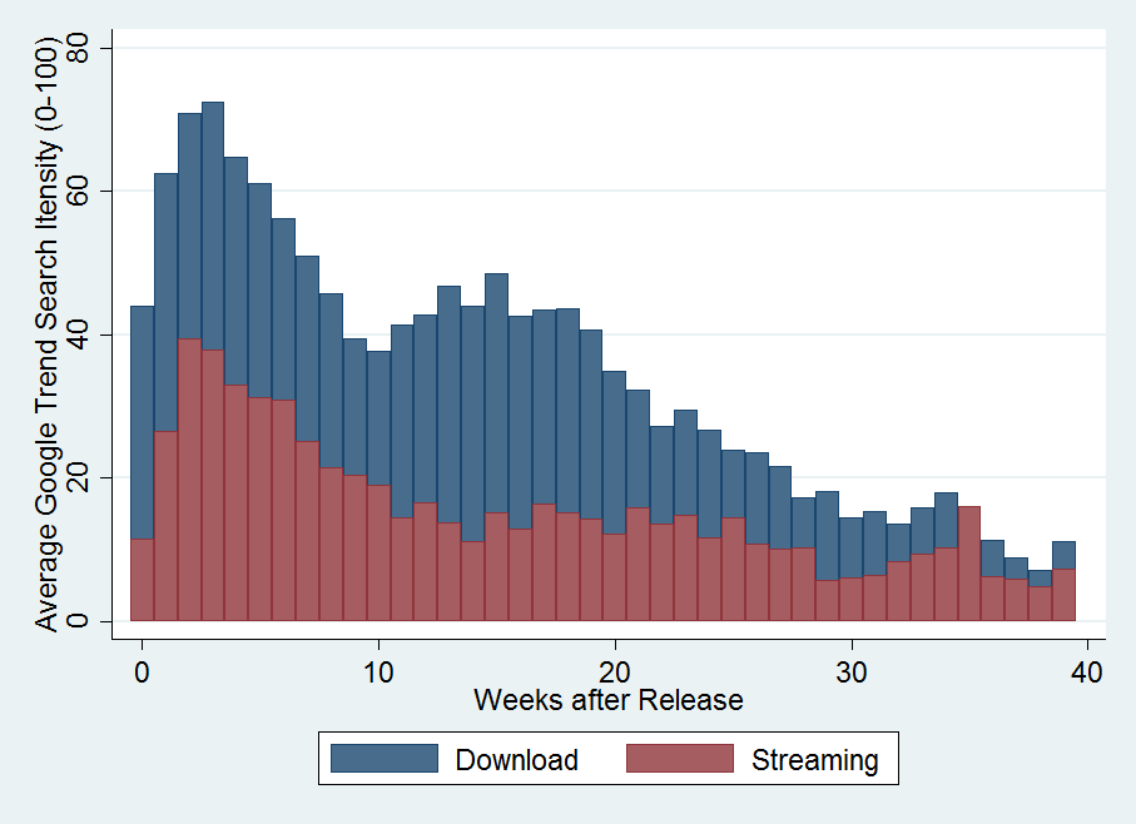


proxies for actual downloading/streaming activities. Given my estimated downloads on BitTorrent I can come out with an estimates to the volume of streaming activities using ratio of Google Trends search traffic index between streaming and downloading.

The procedure is as follows. First, I keep track of a list of most common search queries about streaming/downloading that appear in the Top related search queries list related to movie category. Second, I divide those queries into Streaming-related and Downloading-related and retrieve values of their weekly search traffic index for each movie in my sample. Third, I calculate the ratio between aggregate download-related and streaming-related traffic for each movie. The ratio for each movie is then used to adjust total piracy views estimates. I can then re-estimate the model using the new piracy views estimates. Alternatively, I can also scale up piracy by a multiplier and re-estimate the model.

[IMCOMPLETE]

Figure 9: Download traffic vs Streaming traffic by weeks after release in theater



Appendix D 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