UNIVERSITY OF WARSAW
Faculty of Economic Sciences

WORKING PAPERS
No. 23/2014 (140)

WOJCIECH HARDY
MICHAŁ KRAWCZYK
JOANNA TYROWICZ

INTERNET PIRACY AND BOOK SALES: A FIELD EXPERIMENT

WARSAW 2014

UNIVERSITY OF WARSAW
Faculty of Economic Sciences

Internet piracy and book sales: a field experiment

WOJCIECH HARDY
Faculty of Economic Sciences,
University of Warsaw
e-mail: whardy@wne.uw.edu.pl

MICHAŁ KRAWCZYK
Faculty of Economic Sciences,
University of Warsaw
e-mail: mkrawczyk@wne.uw.edu.pl

JOANNA TYROWICZ
Faculty of Economic Sciences,
University of Warsaw
e-mail: j.tyrowicz@uw.edu.pl

Abstract

We report the results of an experimental study analyzing the effects of Internet piracy on book sales. We conducted a year-long controlled large-scale field experiment with pre-treatment pair matching. Half of the book titles received experimental treatment, in which a specialized agency would immediately remove any unauthorized copy appearing on the Internet. For the other half we merely registered such occurrences, but no countermeasures were taken. For all the titles we obtained print and e-book sales statistics from the publishers. We find that removal of unauthorized copies was an effective method of curbing piracy, but this had no bearing on legal sales.

Keywords:
digital piracy, e-books, field experiment

JEL:
C93, D12, K42, L82, O34

Acknowledgments:
Authors would like to thank: Włodzimierz Albin, Justyna Kossak and Magdalena Przek-Ślesicka of Wolters Kluwer Polska, Krzysztof Gutowski and Sebastian Kawczyński of Plagiat.pl, Kamila Kanafa of Macademia Literary Agency as well as representatives of publishers participating in the study for help in acquiring the data. We are also grateful to Marta Dyrkacz, Agnieszka Borowska and Lucas van der Velde for research assistance. The remaining errors are ours. Authors gratefully acknowledge the support of the National Science Centre, grant DEC-2011/01/D/HS4/03937. All opinions expressed are those of the authors and have not been endorsed by the NSC.

1 Introduction and motivation

The problem of copyright infringement became prevalent in the book trade in the beginnings of the 17th century. The reprinting of previously published works – especially musical scores – by other publishers, was a major concern for original authors and heralded the invention of first copyright laws, see Kawohl and Kretschmer (2003). The 20th century saw the increased popularity of what is now often referred to as “piracy”.[1] With the advent of new technologies, the culture industry had to face unauthorized distribution of their products: music, movies, software and books.[2] The market changed again as Internet was invented and broadband connection spread worldwide. With the rapid evolution of technology over the subsequent years, culture industry faced new forms of ‘piracy’ and file-sharing. Notably, the book industry is now witnessing the rise of the e-book format and much like music industry before, book publishers begin to face the consequences of the fact that sharing a digital copy is becoming increasingly easy.

Most of the studies find that the scope of unauthorised distribution is large and expanding, but the effects on authorised sales are not always detrimental while the estimates themselves are subject to a lot of controversy. The impact of unauthorised distribution on film, music and software industries has been subject to several studies over the past two decades – e.g. Givon et al. (1995); Fine (2000); Zentner (2006); De Vany and Walls (2007); Smith and Telang (2009) just to name a few. Generally speaking the results of these studies tend to be mixed if not equivocal when it comes to estimating the effects of increasing “online piracy” on sales. This is not surprising given how complex and multi-faceted social phenomenon “digital piracy” is. For instance, the effects on box office revenues and the DVD sales seem to differ – see Rob and Waldfogel (2007); Zentner (2011) – which suggests that visiting a cinema and watching a movie at home are not very good substitutes. Furthermore, not all industries are affected in the same way and distribution channels differ, see Gunter (2009). The confusion stems also from objective constraints on data quality. Researchers are often forced to rely on more or less adequate proxies rather than actual variables of interest, because the latter is relatively scarce, especially for the indicators of “piracy”. In addition, the link between “piracy” and sales is clearly endogenous, which necessitates the use of – possibly weak – instruments. As the applied econometric methods evolve, so do controversies around them – e.g. the study by Oberholzer-Gee and Strumpf (2007) and its critique by Liebowitz (2008, 2010). These challenges may add noise to the obtained results. Many problems (such as omitted variables or reverse causality) could be overcome in an experimental framework. There are some *quasi*-experimental studies, which explore natural experiments of legislative or institutional changes – e.g. Danaher et al. (2010, 2012) – but the experimental evidence remains scarce.

This paper contributes to the literature by providing experimental evidence on the effect that digital “piracy” has on the book sales. We design a field experiment, observing the authorized sales of selected books over a period of 12 months. We randomly assign books into a treatment group, where unauthorized copies were removed from the distribution channels – and a control group where the number of unauthorised copies was recorded but no actions were taken. We carefully control for the actual availability of unauthorised copies throughout the entire period of the experiment and include in our study a variety of book genres, such as academic, language, self-help, fiction, fantasy, etc. The sample of titles

[1] This term is often used with reference to both downloading and sharing content online, despite stark differences in legal constraints. Typically sharing is prosecuted in many legal systems, whereas downloading is not. In the reminder of this paper we use interchangeably the terms file-sharing and digital piracy to refer to a more restrictive definition of “piracy” i.e. unauthorized sharing of cultural content.

[2] Although software is typically not subject to copyright protection (*sui generis* restrictions apply), this industry is typically considered a part of creative industries and thus we will be referring to this type of cultural products as well.

we use in this study contains cases for both best-sellers and niche items. We also control for the time since first edition and a variety of measures for the book price components (hard cover, number of pages, etc.). Our manipulation checks indicate that removing unauthorized copies is an effective way to reduce online availability. Yet, our experimental treatment, which consisted of removing these unauthorized files shared on public websites, resulted in only negligible increase in sales, regardless of the book's genre and popularity. These results are in line with some other studies which argue that decreasing sales are not a consequence of the increasing prevalence of "online piracy", but other factors – e.g. Oberholzer-Gee and Strumpf (2007); Martikainen (2011).

This paper is structured as follows. First, we review carefully the relevant literature, in order to nest our experimental design in the earlier contributions. Second, we discuss the experiment, the sample, the treatment as well as the rationale for the pre-treatment matching, which was pursued in this study. In section 4 we discuss at length the characteristics of our dataset, whereas in section 5 we outlay the results of this experiment, various robustness checks and the conclusions emerging from our study. We conclude in section 6 with the most important policy recommendations and limitations of our study.

2 Literature review

The path of technological progress made the book industry suffer the problems of other creative industries later. Indeed, unlike music, films and software, books have been, until recently, quite hard to digitalise. A significant change has been introduced by digital scanners, which made it much easier and cheaper to have the entire printed book uploaded on the Internet. Additionally, until user-friendly reader devices were developed, a digitalised copy was a poor substitute of the printed book. With the introduction of readers and e-books, demand for digital copies soared. The new format also allowed to produce unauthorised copies faster and without losing the original quality - i.e. directly from the legal sources. Therefore, what has been on the agenda of music and film producers since the 1980s (and the introduction of walkman and video cassette recorder), became a major headache for the book industry only in late 2000s.

The novelty of the problem is associated with a scarcity of literature. We are not aware of a serious empirical study devoted explicitly to book piracy. There have been some attempts to assess how sales are affected by publishers' or authors' *voluntary* sharing of their books online. Hilton and Wiley (2011) followed this approach for eight religion books from a single publisher. Books that were downloadable free of charge had slightly higher sales. Similar insights come from Hilton and Wiley (2010), who looked into data on sales of 41 books that were at some moment made available online for free by the publisher. Once again, they found a positive impact of downloads on the print sales, although the effect was not significant when this new option was only short-termed (e.g. one-week-only offer) or involved additional trade-offs (e.g. signing up for a newsletter).

Most evidence is even more anecdotal. Flint (2002) described the 'Free Library' project whose main purpose was "to provide a literary equivalent of a test drive to the readers" and states that despite free online availability, books continued to sell well, whereas in some cases online distribution contributed to the 'sales recovery'. Many other publishers and authors argued in favour of free online distribution through essays, articles or blog entries, see Doctorow (2009); O'Reilly (2002); Coelho (2008). Admittedly, many could be found who consider any kind of book sharing harmful to their sales, see Pogue (2008); Ingram (2012).

Such studies have well known limitations for causal interpretation: a low number of observations, often in a single market segment, sometimes even a single title. Admittedly the results seem consistent across different studies in showing that voluntary dissemination free of charge or at PWYW price, gives consumers the privilege of dipping the toe. The actual causal effect on sales, however, cannot be determined.

The fact that unauthorized copies can typically be downloaded cheaply or at no cost at all makes "piracy" resemble free distribution by the IPR owner. Even, however, if the latter indeed increases overall sales of a given title, it would be naive to imply that the same must hold for "piracy". These effects may be different for a number of reasons. First, while downloading an unauthorised copy, a "pirate" is not exposed to the advertisement or other encouragements to actually buy the downloaded product, as analysed in the case of (Hilton and Wiley, 2011). Second, while one may argue that a publisher sharing his own books for free evokes a form of gratitude among his readers (i.e. gains good reputation), the same cannot be said about a publisher whose works are shared against her will. Third, information on free legal offers might spread faster than in the case of unauthorised channels of distribution, while the former remains unfettered by the subjects' moral judgement. Finally, unauthorised versions of a book might conjure less positive impressions (i.e. crude scans, low resolution pictures, etc.) than a professionally-prepared official version of a book, and thus induce a less positive effect on the reader's willingness to pay.

A similar point has been emphasised by O'Leary (2009) who focused on differences between piracy and free online distribution. He compared the sales of eight books across two periods. In the first one, books were available online for free. When this promotion period has ended, he noted the appearance of first illicit versions on P2P file sharing networks. He compared the four weeks after the end of the promotion period, to the four weeks of free online promotion period. He found little relationship between sales and the file sharing. Sample size was clearly too small to provide general conclusions but the study shows an interesting complementarity between free online authorised distribution and unauthorised file sharing. Yet O'Leary (2009) does not report dates, so it is also possible that his results are driven by the seasonal effects, e.g. higher sales during Christmas time.

An electronic monochrome scan is clearly not a perfect substitute for a hard cover colour print. Yet, a *.epub or a *.mobi file downloaded from an unauthorized source are a perfect substitute for the exact same file to be purchased from the publisher or an online bookstore. Widespread broadband Internet access and the increasing popularity of digital book formats make it even more important to study the causal effect of the "piracy" on book sales. Like in the case of other cultural goods, the two main channels i.e. "dipping the toe" and substitution of sales, both seem plausible and work in the opposite directions.

Although there have been many attempts at measuring "piracy" effect on music and movie sales, experimental methodology has only very rarely been adopted. What comes closest are quasi-experiments, such as those based on changes of the law – e.g. Danaher et al. (2012) study the effect of HADOPI law implementation in France – or sudden supply changes in the market – for example Danaher et al. (2010) analyzed the changes in piracy levels after NBC removed its content from iTunes, while Peukert et al. (2013); Danaher and Smith (2014) focused on the impact of Megaupload shutdown. We are not aware of any studies that would establish a 'clean' counterfactual for causal relationship analysis.

Despite scarce empirical literature on the impact of piracy on book sales, it is quite clear what features an experimental design aimed at investigating this effect should show. First, for the experiment to be meaningful, a diversified and large sample of books and authors is needed, because results seem to be to a large extent driven by readers' characteristics. Second, randomisation needs to account for book specificity, because both sales and downloads are highly diversified across segments of the literature and across authors. It is also true that different genres of books attract different kinds of readers who might also differ in their attitude toward buying and downloading. Third, seasonal effects, like Christmas time or the starting of a school term need to be controlled for.

We address these three issues by including titles from several publishers and controlling for book genres, as well as other characteristics. We also observe a whole year of shares and sales and use a control group, which makes the conclusions robust to factors like seasonal effects. On the one hand, the market we are analysing is arguably much smaller than those being subject to most previous studies of music, games and films. On the other hand, it makes it feasible to actively manipulate availability of unauthorised copies, enabling effective implementation of experimental treatments. To our best knowledge, this is the first application of a large-scale natural field experiment to measure the impact of piracy on sales.

Some additional insight relevant to our topic can be drawn from studies on e-book demand, and its relation to piracy and other book forms. Zegners (2014) showed that providing free samples of e-books is often associated with higher prices and a better match between the reader and the title. Although his results were based on titles published only in the electronic format, it is clear that the availability of a free sample may have a strong influence on the demand for the so-called "experience goods". At the same time we could expect e-book availability to decrease some of the physical sales for at least two reasons: first, that some people would rather buy a cheaper digital version than a physical one, and second, that e-book existence usually implies higher-quality pirated versions as well. Hu and Smith (2013) studied the relationship between the e-book premiere date of titles sold in physical form, by utilizing a natural experiment where the e-book availability of some titles was delayed by two months. They found that the delay caused an increase in physical sales, although it was small and statistically insignificant. At the same time the delay heavily and negatively impacted the e-book sales. The authors argued that most consumers make a choice about the mode of consumption (physical or digital) first and therefore that e-books are a weak substitute to physical copies. In the context of piracy this could imply that the widespread sharing of scans and e-books has more of an effect on the still developing digital format, and less on the physical one.

3 Experimental design

The purpose of this experiment was to test whether unauthorised online supply of books is detrimental to their sales. In order to investigate the causal effect we conducted an experiment where books from various segments of the market were matched in pairs of similar titles, from which one was randomly assigned to a treatment group while the other to a control group. We subsequently intensively combated unauthorised online distribution in the treatment group, leaving the control group unprotected.

Book sales are well known to exhibit Christmas and holiday seasonal effects (buying books as gifts or because there is more time for leisure). Also, some of the participating titles are academic or school books, so it seems natural that they would be in higher demand as the new educational cycle starts. To address this problem the experiment ran for twelve months starting October 1, 2012, thus encompassing a full year of observations on file-sharing traffic and sales.

We describe the experimental design in three stages: recruitment of publishers, treatment assignment and treatment execution.

3.1 Recruitment of publishers

With the help of the Polish Book Chamber and the e-mail campaign more than 70 Polish book publishers were invited to participate in the experiment[3]. At the stage of invitation the publishers were informed about the objective of the experiment, its duration and the

[3] In addition to the official letter sent out by the Polish Book Chamber, an independent literary agent contacted ca. 20 publishers.

intended treatment. It was also explicit in the invitation that the assignment to treatment and control groups was to be random and could not be influenced by the publishers. Of the initial group, 11 major publishers eventually took part in the experiment.

Publishers were asked to provide a list of up to ca. 50 books each, the actual numbers ranging from eight to sixty-one. The invitation letter specifically asked that the books selected for the experiment are relatively new (or even forthcoming) and relatively popular titles. This was to ensure that most of our sample consists of books potentially susceptible to the problem of sales loss and not ones that were already out of sale, or selling poorly, for which we would not be able to observe any effect. While this requirement is meant to reduce the odds that the final results are driven by a few specific cases, it is not likely to limit the validity of our findings. First, research shows that best-selling books are most frequently pirated as well. Second, the publishers did not know at the moment of designing the book list, which of the books would be treated and which would be in the control group.

The participating publishers covered various segments and represented different business models (e.g. promoting domestic authors versus translating foreign bestsellers). As such, their sales' reaction to our treatment could differ. Table 1 presents the numbers of books from each of the segments included in the study. Although we have initially provided the publishers with a list of segments they were to choose from, the list has evolved in accordance with their advice and their own segmentation. Table 1 reflects this updated list.

Table 1: Number of books per segment.

Segment	Number of books
General fiction	43
Fantasy and Science Fiction	15
Non-fiction (biographies, memoirs, essays, etc.)	46
Foreign languages	18
Academic books	32
Science & Research	9
Business & Economics	9
Law	40
Professional & Technical	22
Self-Help	10
Other	2

The publishers were also asked to provide us with detailed data on each book. This information included the author's name, page count, price for different formats (hardcover, audiobook, e-book) and date of release and was used for the matching procedures as explained in the next section. In addition to the basic characteristics we have also acquired data on the numbers of previous editions (if any), first print run, past sales (if any). Publishers were also asked to provide quarterly or (preferably) monthly sales forecasts and to indicate whether they expected the unauthorised distribution (if any) to affect these figures positively or negatively. The fact that they uniformly anticipated the latter effect speaks to the importance of the study. Finally, after the first half-year and after the end of the study, publishers provided monthly or quarterly sales reports for each title. For description of the provided book data, see Table 2 in section 3.2.

3.2 Treatments and treatment assignment

The experimental treatment in this study consists of issuing notice to take down to all identified locations where unauthorized distribution took place. Enforcement of copyright is not automatic in most countries. In most cases, copyright owners either develop own algorithms of buy a service from an agency specializing at identifying the unauthorized copies. Most file-sharing sites develop algorithms which permit application-to-application communication for automatic take down of the unauthorized content.

In this experiment we cooperated with one such specialized agency – Plagiat.pl. Plagiat.pl was provided with the complete list of books, each assigned to one of the two treatments. With respect to the titles in the Enforcement Treatment (ET, "treated" books), Plagiat.pl would trace the unauthorized distribution and urge the appropriate service providers to remove the hosted content. These actions were a continuous effort. For the titles assigned to the Control Treatment (CT, "untreated" books), the number of copies distributed without authorization was recorded and reported to the research team but no efforts were made to enforce copyright protection.

Given the diversity of books and publishers participating in this study, we chose the treatment assignment strategy to maximize *ex ante* the efficiency of the treatment effect estimation. Thus, instead of basic randomization we have applied a matched-subject design. This procedure is somewhat unusual in experimental economics, but permits efficient use of within-treatment-group heterogeneity in estimating the treatment effects. More specifically, if any allocation was equally likely, diversity of popularity between segments of the books would inflate the standard errors and reduce the efficiency of the estimates. Yet, if treatment is assigned randomly among fairly similar books, the standard errors will stem only from unexplained, random component and not the diversity of the analyzed sample.

The advantages of applying matched-subject design in the process of randomisation are numerous. First, such a design allows for a better control of the differences between the ET and the CT, assuring that similar titles appear in both groups. Second, statistical properties of such experiments are far more satisfactory, as they enhance the covariate balance (and therefore the power of statistical tests, e.g. of treatment effects), and the efficiency of procedures such as covariance adjustment, see Greevy et al. (2004). In fact, Greevy et al. (2004) show that blind randomization (as opposed to matched-pair design) results in what they call an "occasional disaster", i.e. a greatly lowered efficiency of the statistical tests due to an uncontrolled and *unlucky* random allocation between groups.

Matched-subject randomization requires high quality of data prior to the experiment. This data was collected for each book before the study commenced. Yet, it was not obvious *ex ante* which characteristics may moderate the treatment effect. To limit the scope of interference, we relied on the few most objective variables. First, we included sales forecasts as reported by the publishers. Whether or not sales forecasts are a good prediction of the sales is irrelevant as long as books are paired within similar "classes" of sales. Some of the publishers were only able to deliver quarterly forecasts for the sales of their titles. Such data were interpolated into monthly series using the Denton method (Baum and Hristakeva, 2014). The quarterly data were used as a constraint for the sum of monthly data, while the monthly data from the same segments but other publishers, who provided monthly data, were used as a reference for the seasonal effects and trends. Second, we included data on the basic book characteristics, such as the type of the book, date of publication and the number of editions. Third, we included characteristics which may (or may not) explain the price of a book, i.e. hard/soft cover, number of pages, etc. Table 2 reports in detail the variables available prior to the experiment and used for matching.

For the matching procedure, the Mahalanobis distances were computed between each two observations within the dataset. We use this measure as it is fairly robust to sample size and is often reported to perform well in comparison with other methods, cfr. Rubin (1979, 1980); Zhao (2004). For the matching itself we used an algorithm based on network

Table 2: Book characteristics

Variable	Median	Mean	Std. Dev.	Matching
Publication date (for the current edition)	27.04.2012	-	580 days	Yes
Which edition[1]	1	2	3	Yes
Previous edition publication date (if applies)	28.04.2010	-	742 days	No
E-book release (if applies)	25.09.2012	-	75 days	No
Page count	352	415	304	Yes
Versions available (hardcover, e-book, etc.)	-	-	-	Yes
Price [2]	39.99 PLN	50 PLN	21 PLN	No
First print run (no of copies)	800	3 257	10 476	No
Sales before the experiment	0	1 632	8 215	No
Sales forecasts for experimental period	1900	13 639	47 883	Yes
No of unauthorized copies before the experiment[3]	3	94	303	Yes

Notes: Matching column denotes if a variable was included in the pair matching prior to randomization. For prices, 1 PLN $\sim$ 0.3 USD

[1] Some publishers do not make a clear distinction between editions and print runs.

[2] Price per page was used for matching

[3] Number of files shared was identified immediately before the experiment commenced (October 2012) by Plagiat.pl. Files smaller than 1MB were not reported. The actual variable used in the matching procedure was a $ln(x + 1)$. Some titles debuted during the experiment, which drives the mean and median downwards.

flows, written for R by Mark Fredrickson and Ben Hansen.[4]

As a result, 94 matched pairs were created, 13 groups of three and one group of five, see Table A.1 in Appendix A. In every case the books within the same group tended to be similar on the dimensions taken into account. Within each of the matched groups, books have been assigned to either the treatment or control group in a randomised manner, such that there was always one treated and one untreated book in each pair, one or two of either type in each group of three and two or three of either type in the group of five. Another constraint was that the numbers in the odd groups would 'balance' each other out, i.e. the total numbers of the treated and untreated titles were also equal or different by one.

3.3 Treatment execution

During the course of the experiment, Plagiat.pl performed regular Internet searches for unauthorized copies of all books participating in the experiment. For the treatment group (ET) the notice to take down was issued to the service provider immediately upon finding. In fact, almost each day consecutive titles would be sought. Once a month a report was drafted by Plagiat.pl with the list of notices issued and the number of copies from ET still available at the end of the reporting period. For the control group (CT) the number of copies at the end of the reporting period was reported.

It is important to mention that during the course of the study the publishers were not informed about the treatment assignment of their titles. This was done so as to reduce the risk of the publishers selectively changing their behavior in pricing or other strategies. Of course, it cannot be excluded that some of them could infer it from the availability of unauthorized copies if they decided to conduct a search on their own. Yet, as we report, for all publishers and in each segment there were some book titles that were never shared online in both treatment and in control groups. The titles which we never identified to be online would blur any inference by the publishers.

Although Plagiat.pl was crawling the web regularly, one of the potential "black spots"

[4] For additional information see Hansen and Klopfer (2006) or the *optmatch* package for R-CRAN.

could be peer-to-peer networks, such as Torrent. These kinds of networks tend to be extremely difficult to monitor and are impossible to issue a notice to take down (there is no entity to which the notice could be issued). Thus, one might argue that identifying and tracking down unauthorized P2P file-sharing is fruitless. However, P2P networks are not very popular in Poland; in particular they tend to offer limited supply of Polish-language content, which makes many downloaders turn to domestic websites. In fact, according to the reports on the largest continuing commercial study on Internet use in Poland – PBI/Gemius Megapanel – the site that Poles would most often turn to for files was the file-sharing service Chomikuj.pl. Chomikuj.pl is a *pay-for-transfer* platform, where people upload files for free but need to pay small amounts to download content. Alexa web analytics place Chomikuj.pl as the 17th most popular website in Poland in 2013. The second general file-sharing platform is the (in)famous Pirate Bay at rank 66.[5].

This is of importance, because Chomikuj.pl maintains a policy of swift response to the take down notices. They develop an application-to-application algorithm which permits copyright owners upload directly the bulk lists of copyrighted content. In addition, if the same file is shared by other users than reported by the copyright owner, these files too are removed after the notice to take down. Thus, relying on Plagiat.pl in tracking down unauthorized file-sharing was indeed likely to reduce the amount of book piracy significantly enough for the effects to be measurable. The actual effectiveness of these activities is tentatively measured and reported in section 5.1.

4 Data

In this study we collected the data on sales and on the file sharing in the experimental context of removing. In this section we review the data collected in the experimental study. We first discuss the nature of treatment and the reporting on the file-sharing in the context of treated and untreated sample. In section 4.2 we discuss the descriptive statistics of the sales data.

4.1 Treatment data

Data on file-sharing come from Plagiat.pl who monitored the availability of books participating in the experiment and sent the notice to take down in the case of the treated books. The reports from Plagiat.pl arrived monthly, but the actual dates were subject to the timing of web crawling performed by Plagiat.pl and thus differ between the months. Plagiat.pl constantly updated the mechanics of the web crawling algorithms in order to improve the ability to scan the new file-sharing sites as well as to strengthen the identification of a particular cultural content (for example new types of file names, formats, etc.). The full list of report dates, and the covered periods is provided in Table 3.

The reports cover the whole year of experiment duration, from the end of October 2012, to the end of October 2013. The reports on the experimental treatment (ET) arrived in two forms. In the first report we received information on all notices to take down the unauthorized copies over the periods described in column (1). For example, between 24th of October 2012 and 23rd of November 2013, Plagiat.pl issued in total 11,952 notices to take down the unauthorized files with books from our experimental group.

In the second report, as covered in column (2), Plagiat.pl reported the statistics of all files available online for each book in the experimental group without actual date identification (the reports contained only the concluding date of the period of searching). The

[5] Source: http://www.alexa.com/siteinfo/chomikuj.pl, http://www.alexa.com/siteinfo/thepiratebay.sx; accessed 16-12-2013.

Table 3: The reports on treatment execution

No.	ET - with notices (From-To) (1)	ET - descriptive (2)	CT - descriptive (3)
1	24 Oct 2012 – 23 Nov 2012	23 Nov	26 Nov 2012
2	-	-	2 Jan 2013
3	4 Nov 2012 – 17 Jan 2013	17 Jan 2013	16 Jan 2013
4	5 Jan 2013 – 18 Feb 2013	18 Feb 2013	22 Feb 2013
5	2 Feb 2013 – 18 Mar 2013	11 Mar 2013	18 Mar 2013
6	2 Mar 2013 – 15 Apr 2013	9 Apr 2013	15 Apr 2013
7	3 Apr 2013 – 15 May 2013	14 May 2013	17 May 2013
8	7 May 2013 – 18 Jun2013	13 Jun 2013	19 Jun 2013
9	3 Jun 2013 – 9 Jul 2013	12 Jul 2013	11 Jul 2013
10	2 Jul 2013 – 19 Aug 2013	14 Aug 2013	19 Aug 2013
11	6 Aug 2013 – 9 Sep 2013	9 Sep 2013	13 Sep 2013
12	3 Sep 2013 – 28 Oct 2013	-	30 Oct 2013

web crawling algorithm utilized to produce the reports from column (2) sometimes resulted in items that were not listed in reports from column (1). The differences, however, were negligible with the second report simply containing more clutter – as the ET reports concluded with taking action against the uploaded files, Plagiat.pl put extra effort into making sure that no mistakes are committed in this group. We matched the data from the second report to those from the first to infer some additional knowledge on the found ET group files (e.g. their size). A report analogous to that from column (2) was compiled by Plagiat.pl for the control group, column (3).

Summarizing, the reports on file-sharing permit to identify per period the number of files available in both ET and CT groups. In addition, they permit to identify if treatment was executed, i.e. if a notice to take down was issued, as Plagiat.pl reported dates of checking whether the files were actually removed. If no date was given (presumably because the file hadn't been removed before the date of the check) the reports for the subsequent months allowed to identify if the notice was effective, i.e. if the book was permanently removed from a particular site at which it was shared without authorization.

There are two additional remarks to be made. In the ET notice reports Plagiat.pl filtered away some of the smaller files. This is based on the premise that the smaller files may not contain an actual book, but a promotional fragment. Plagiat.pl inspected each case of small file before sending the notice to take down. For example, Plagiat.pl was able to identify the cases where a complete book was cut into smaller files to avoid being identified. Case-by-case inspections reveal such situations to Plagiat.pl. If a file was not identified as copyrighted content, Plagiat.pl would not issue the notice to take down and it would not include it in the monthly reports. As a general rule, Plagiat.pl stated that most of the files under 1MB are actually promotional fragments.

Second, despite notices being issued on a regular basis, sometimes our treatment was not fully effective. One such case includes an individual who re-uploads a copy after it was taken down. Such activities are not very frequent – according to our data this happened on approximately 4% of occasions. However, since the web crawling algorithm worked through monthly periods, it may be that the reemerged files were available for some short periods before being taken down again in the subsequent month. Another such case is possible when a book is dissected into smaller particles (e.g. chapters), whereas the notice is effective only for some of these files. It remains debatable if a fraction of a digital copy of the book constitutes a viable substitute for the whole book. These two cases were, however, rare. The parts of the book were removed at the latest within the next month of web crawling. The time needed for the algorithm to find the new copies was short, the improvements were implemented on a daily basis.

For the following analysis it seems natural to focus on an actual *copy* of the book (rather than a file) as the unit of analysis. As our data reported findings of each separate file, we have defined a *copy* as a collection of one or more files in a particular format (e.g. pdf or mp3), uploaded by a particular user and referring to the same book title. This means that a hundred image files (e.g. scanned pages) of one book uploaded by one user accounted for one copy, as well as one (large) pdf file of this book uploaded by the same or another user could constitute another copy. Moreover, as the dates of the reports were irregular, we made some assumptions to be able to analyze the data aggregated to monthly intervals. First, we assumed that a copy would remain available until (and if) we received information that it was removed. This we infer if the subsequent report made no mention of that copy. Files once uploaded by the users are not removed unless there is a specific reason to do so. One such reason is notice to take down. The other is replacing a lower quality file with a higher quality one. While the former constitutes treatment, the latter does not. Thus, we identify a copy as available if the same URL address (i.e. the same user) has a file with the same book (even if a different file size or not exactly the same size name). With the transformed dataset we are able to track the numbers of uploaded copies across time, and – consequently – the number of unique copies that appeared on the Internet in each of the months of the experiment.

4.2 Sales data

The data on sales comes from the publishers reports. It is customary in this industry that after publishing a book a fairly large number of copies is shipped to the intermediaries and to the bookstores. Both the intermediaries and the bookstores keep the stock of the books for the period they find adequate and subsequently the unsold items are returned to the publishers. Sales reports of the publishers comprise the data on books sent to the intermediaries and bookstores and the books returned by them – not on actual bookstore sales. Clearing of the transactions typically occurs at a quarterly or semi-annual basis.

These features of the book industry imply that typically monthly sales data comprise a lot of zeros (most of the months), some large positive numbers (when the edition is published) and some large negative numbers (in the transaction cleaning periods). This large variance is not exactly informative of monthly sales, because it does not reflect whatever happens at bookstores in particular months. We thus aggregate the sales data to annual sales.

Table 4: Sales data - descriptive statistics

	No of observations	Mean	Std. dev.	Median	Min	Max
	Original sales data					
Total	239	1 358	3 423	309	-5 893	34 577
CT	120	1 244	2 895	256	-5 893	18 534
ET	119	1 473	3 892	343	-5 037	34 577
	Sales data corrected with first print run					
Total	228	4 103	8 052	1 152	4	69 577
CT	115	4 008	7 220	1 112	4	40 534
ET	113	4 200	8 850	1 181	54	69 577

Notes: First run data were missing for 11 book titles, effectivelu reducing the sample to 228 titles.

A small number of the aggregate annual sales data turn out to be negative. This is possible if a book was published (and sent to intermediaries/bookstores) prior to the beginning of the experiment, but the returns occurred within our observation window.

Thus, the negative sales data are not actually negative sales, but rather returns higher than the contemporaneous shipping. As depicted by Table 4, we solve this problem by adding the print run to the aggregate sales data reported by the publishers.

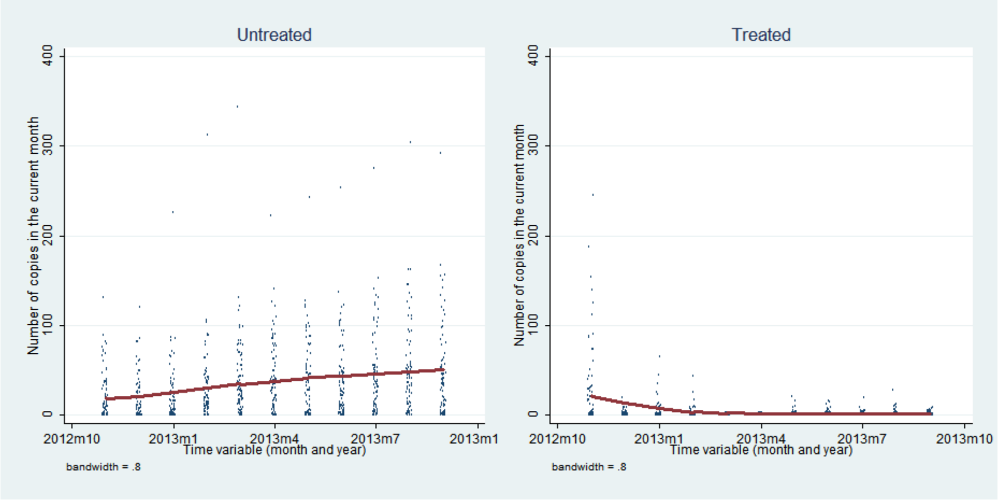
5 Results

This section reports the results of the experiment. We first show that the treatment applied was effective, i.e. that the books from the treated group were much less frequently available online than the books from the control group. This is extensively discussed in section 5.1. In section 5.2 we move to analyzing the treatment effects. We do that in two ways. First, we compute simple sample statistics. Because our experiment is a randomized trial, we can interpret these results as primary evidence. However, if treatment effects are heterogeneous, means/medians may be uninformative. We thus additionally include treatment regression analysis in section 5.3.

5.1 The effectiveness of treatment

Our manipulation check strategy inolved a two-tier test. First, we compared the availability of the illegal copies of books belonging to CT and ET using the reports from Plagiat.pl. If our treatment manipulation was at all meaningful, we would expect that fewer copies of treated books could be found, compared to the control. However, even if this test was passed, we could not be sure that these enforcement efforts were actually making it more difficult for the readers to find an unathorised copy of a book on the Internet. For example they could still be easily downloadable from websites missed by Plagiat.pl. We have therefore applied a second-level test, asking research assistants to find copies of selected books from our list on the Internet, again comparing the results for the two treatments.

Figure 1: The effectiveness of treatment



Note: Numbers of copies (per book) - raw values (dots) and a locally weighted regression lines, zero observations excluded.

At the first level, Figure 1 shows the average numbers of copies (per book) available monthly, respectively in the control and treatment groups. By eyeballing, there is a significant difference between the two groups. While the shared numbers of the control books steadily grow over time, we witness a sharp decline in the numbers of treated titles (even if they sometimes resurface for short periods). Table 5 reports estimates of a panel regression model, explaining the effect of the treatment on the number of copies and the probability of at least one copy being available at a given time. The results indicate a significant role of the treatment in the observed availability.

Table 5: Effectiveness of reducing the available unauthorized copies

Number of unauthorized copies	Coefficient	P-value
Treatment	-6.09	0.021
Month of experiment	2.35	0.000
Month of experiment * treatment	-3.20	0.000
No of unauthorized copies identified prior to the experiment	0.45	0.000
E-book exists	-4.90	0.043
Constant	6.81	0.153
Number of observations	2514	
Number of titles	228	

Notes: Panel tobit regressions on the monthly data. Number of copies provided by Plagiat.pl. Segments and publisher dummies included, not reported, available upon request.

At the second level, three research assistants, who reported being proficient or at least familiar with acquiring uploaded files, searched for a specified set of books on the Internet. The lists comprised of twenty randomly chosen pairs of titles from our initial sample. The assistants did not know which of the books have been treated and, as a matter of fact, that there was any treatment at all. Two of them received the same list of titles, so that we would have some sense of individual differences between users, while the third received a different one so that more books could be covered. Their task was to search for each title for up to five minutes. Should they find the book during that time, or they reach their time limit, they were asked to move to the next title. For each book, we asked them to record the number of failed attempts (i.e. the number of times a downloaded file proved fake or not working) and the time it took them to get it. We also prepared a Chomikuj.pl account with sufficient transfer (5GB for the whole search) and specified a money budget (2 PLN) per title in case the assistants found some other platform providing the title for a small fee. While the Chomikuj.pl account proved useful, no other paid services were utilized.

Table 6: Numbers of books for which an unauthorized version could be found

	Assistant A	Assistant B	Assistant C	On average
Control Treatment	10	12	12	11.33
Control Treatment (%)	50%	60%	60%	56.67%
Enforcement Treatment	4	7	8	6.33
Enforcement Treatment (%)	20%	35%	40%	31.67%

Notes: All of the title lists included 20 pairs of books (20 in the CT and 20 in the ET group). Assistants B and C had the same list of titles.

The results of this manipulation check are reported in Table 6. Although we have not been able to completely eliminate the availability of the treated titles, they were visibly harder to acquire (about 57% of titles found in CT, versus only 32% in ET), which was confirmed by a formal test of equality of proportions (z=2.776). The lower availability of the treated group was manifested also in longer times of searching (on average about 63 seconds for CT and 94 seconds for ET), although formal testing did not find this difference significant (z=-1.127), possibly due to the sample of found titles being small. The source of most successfully downloaded files was Chomikuj.pl (see Table A.2 in Appendix A), which reinforces the findings of the first manipulation check (i.e. that we successfully affected the upload traffic even if we look only at Chomikuj.pl). For the treated titles, the assistants also often had to turn to websites other than Chomikuj.pl, which could be discouraging for downloaders who are not familiar with them.

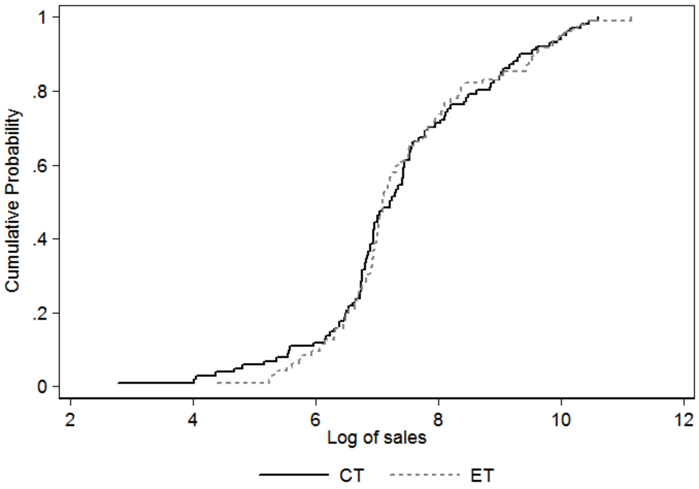
Most of the found titles were the same for the two assistants with the same list (assistants B and C) . Assistant B found 2 titles that Assistant C did not and Assistant C found 3 titles that Assistant B did not. They both found the same set of 16 titles, although assistant B reported longer times of searching of the found copies (on average twice as long, i.e. around 88 seconds per copy found). Still, despite this difference in quickness, the outcome of the searches seems consistent and both of the assistants mostly found titles from the CT group.

The research assistants were requested to look for any source for five minutes. It is plausible, however, that many Internet users limit themselves only to their favorite websites and/or stop if the search does not provide results immediately. We therefore conclude that our treatment successfully reduced the download traffic of the treated titles, perhaps even more than we observed.

5.2 Treatment effects

The comparison of mean sales between the two treatments reveals that on average sales were approximately 5% higher in ET. Yet, the difference does not seem to be actually economically relevant or statistically significant. Figure 2 reports cumulative distribution functions for the (log of) sales in ET and in CT. While there are some discrepancies, tracing them to the rigid enforcement of the copyright woould be questionable.

Figure 2: Cumulative distribution functions of sales (in logarithms)



Note: sales data topped with first run data to avoid negative values, in total 148 titles were ever subject to unauthorized file-sharing.

Indeed, proper statistical tests cannot reject the null hypothesis that in fact there is no difference in sales across the two treatment groups. Table 7 reports the results of Mann-Whitney's non-parametric rank test on overall sales of the analyzed titles. Additionally, we perform Wilcoxon test on matched-pairs (the matching based on our initial design). The maximized/minimized columns refer to the choice of titles for the testing. As described in section 3.2 we had several groups of more than two titles. To run the test, just two had to be selected and compared. To maximize robustness of our results, we use two extreme approaches. First, we take the *max* of the annual sales of the treated books and *min* of the untreated in each group. This approach, which we call 'maximized' is the one that makes it most likely that the test indicates a positive treatment effect. We also take the *min* of the treated and *max* of the untreated within each group ('minimized'), which makes it easiest to observe negative treatment effect. Yet, in either case the data fail to reject the null hypothesis that within a matched group of treated and untreated books there is no difference in sales over the observed period.

Table 7: The difference in sales between CT and ET - test statistics

Difference in sales	Mann-Whitney test	Wilcoxon test		Levene test
		maximized	minimized	of equal variance
whole sample	$z = -0.076$	$z = -1.731$	$z = -0.848$	$W = 0.132$
(*p-value*)	(0.9396)	(0.08)	(0.40)	(0.72)
n	CT:120, ET:119	82 pairs	82 pairs	CT:120, ET:119
unauthorized copy exists	$z = -0.710$	$z = -0.995$	$z = -0.139$	$W = 2.128$
(*p-value*)	(0.4779)	(0.32)	(0.89)	(0.147)
n	CT:82, ET:72	56 pairs	56 pairs	CT:82, ET:72

Notes: Difference in sales reported as $z - statistic$ for Mann-Whitney and Wilcoxon matched-pair test, and as $W - statistic$ for Levene's test. $p - values$ in parentheses.

As a last check we test for differences in variation in both groups – it is plausible that treatment affected various genres or titles in different ways. For example, Blackburn (2006) finds that piracy affects music sales differently, depending on the popularity of the artists. If that indeed was true for books, the positive and negative effects could cancel out. We thus test explicitly if variance of sales is different in CT and ET. Data fail to reject the null hypothesis of equal variances.

Obviously, our manipulation could have no bearing on the titles, for which no single unauthorized copy was ever uploaded. As a robustness check we have thus repeated all the tests restricted to the sample that saw at least one pirated version available at some point. The results do not change.

While treatment assignment maximized the comparability of the two groups, it is possible that our sample is too small to account for all the heterogeneity between the titles. Therefore, as a robustness check, we have conducted a treatment regression model to double-check our matching strategy and to account for any unresolved problems.

5.3 Regression analysis

The comparison of distributions, as included in section 5.2 cannot capture the possible nonlinearities associated with the process of unauthorized file-sharing. For example, it may hold that a larger number of unauthorized copies or the sites where these copies are shared makes it effectively easier for the potential customers to find a book online and thus abstain from purchasing. In addition, some of the books could exist also as e-books, which makes the substitution between the authorized and unauthorized versions closer to each other. Finally, it may be *when* the unauthorized copy is available rather than *if* it is available, that affects the sales. For example, a textbook is sought after in the beginning of the academic cycle or in the end, but its availability in the middle of the semester may have absolutely no effects on sales. We put all of these contentions to an empirical test in order to eliminate the channels possibly blurring the treatment effect.

Table 8 presents the results of these tests. The columns (1)-(4) present the results on the whole sample, while columns (5)-(7) present analogous regressions for sample where at least one unauthorized copy for each title was identified during the experiment. This restriction helps to identify books for which piracy could have any direct effects on sales.

Table 8: Treatment effect estimation

	Whole sample				Only titles uploaded (ever)		
Log of aggregate sales	OLS	Treatment regressions					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.06	4.99	4.23	5.88	1.95	0.095	2.34
	(0.30)	(1.07)	(1.08)	(1.08)	(1.14)	(0.07)	(1.32)
* time available			0.10	0.03		0.05	-0.034
			(1.56)	(0.28)		(0.64)	(0.37)
* e-book exists	-0.08	-0.35	-0.15	-0.26	-0.33	0.06	-0.21
	(0.23)	(0.67)	(0.37)	(0.42)	(0.71)	(0.15)	(0.46)
* segment	No	No	No	Yes	No	No	Yes
E-book exists	1.85***	1.73***	1.58***	1.55***	1.77***	1.63***	1.59***
	(7.35)	(4.77)	(4.53)	(3.62)	(6.25)	(5.72)	(5.61)
No of copies			0.32	0.30		0.25*	0.21
			(1.63)	(1.24)		(1.84)	(1.55)
Time available			-0.05	-0.04		0.01	0.03
			(0.66)	(0.46)		(0.14)	(0.37)
No of observations	228	228	228	228	148	148	148
R^2	0.32						

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$, t-statistics in parentheses.
The first step for treatment regressions comprises price per page, page count, previous sales, initial report, date of publication, dummies for segments, i.e. variables used for matching. All of them prove insignificant. The explained variable is log of aggregate sales topped with first print run. No of copies is defined as a monthly average of the 12 months observed ($log(x + 1)$ to avoid dropping unshared files in specifications (3) and (4).

We find no treatment effect. Treatment is also insignificant in interactions with segments and how long the unauthorized copies were available. The same is true if we account for the existence of an e-book version for each title.

Finally, we seek to identify the treatment effect via a quantile regression. The rationale behind this approach relies on the understanding that for unpopular items, the availability of unauthorized copies can have different effect on sales than for the titles who sell in tens of thousands of copies. For example, *niche* titles can see their demand totally destroyed by piracy, whereas for a bestseller file-sharing may have only marginal importance. The opposite may hold as well – bestsellers could have a larger number of customers diverted by file-sharing than *niche* titles. We test these contentions by means of quantile regression with the same controls as in Table 8. The results are reported in Table 9. Along the entire sales distribution there is no indication that the treatment effect could be at work. Although the sign changes from positive to negative as sales increase, the estimates are insignificant.

The estimate for the interaction of the treatment with availability of unauthorized copies show that the results from Table 8 were no mistake. For the popular books there is no link between the availability of unauthorized copies and sales. The only exception is the 75th percentile, but given the insignificance for the higher quantiles this suggests rather a blip in the data than a strong pattern. Traditional sales seem higher if an e-book exists, but this is probably associated with the fact that publishers decide opt for this form only in the case of popular titles. Thus, this coefficient does not have a causal interpretation.

6 Conclusions and discussion

The phenomenon of intensifying piracy – unauthorized file-sharing – has been often considered responsible for the drop in sales of music albums, cinema tickets and, recently,

Table 9: Quantile estimates of the treatment effects

	Q10	Q25	Q50	Q75	Q90
Treatment	1.35	1.36	0.82	-0.18	-0.16
	(1.22)	(1.42)	(0.87)	(-0.33)	(-0.20)
Treatment * time available	-0.14	-0.08	0.05	0.16	0.07
	(-0.90)	(-0.57)	(0.45)	(1.65)	(0.63)
Treatment * e-book exists	-0.21	0.07	-0.34	-0.19	0.44
	(-0.27)	(0.12)	(-0.57)	(-0.28)	(0.59)
E-book exists	1.63**	1.18***	1.68***	1.65***	1.60***
	(2.38)	(2.70)	(4.08)	(3.84)	(2.87)
Avg. number of copies	-0.02	0.31	0.26	0.47*	0.22
	(-0.07)	(1.47)	(1.16)	(1.67)	(0.68)
Time available	0.08	0.06	0.05	-0.10	0.05
	(0.52)	(0.53)	(0.42)	(-0.95)	(0.33)
No of observations	148	148	148	148	148

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$, t-statistics in parentheses. Model estimated as a quantile regression (not to be confused with quantile treatment effects estimation).

also books. The threat from piracy is emphasized by the publishers and other stakeholders, such as agencies of collective management of copyright or cultural goods chambers. In this study we have developed an experimental framework to test if the availability of unauthorized copies has any effect on sales. We have conducted a large field experiment where treatment comprised fairly effective removal of the unauthorized copies of books, for which we observed the traditional sales. While most of the publishers suspected a negative impact of piracy on legal sales, we find no evidence of a significant shift in sales due to pirate copies being available on the internet.

The failure to reject the null hypothesis may stem from several reasons. First, one of the 'pro-piracy' arguments often raised in the debate is that piracy can be used as means to sample products, before buying them. It is possible that this effect balances out the negative effect coming from people downloading instead of buying.

Second, the result is likely to be associated with the fact that digital files are in fact poor substitutes to paper versions of books. They are indeed of a different form and require specific devices to be read. We cannot state if our results would hold for sales of e-books – the Polish e-book market is as of yet too small for it to be studied in a similar manner.

Third, our sample was not representative of the market in terms of genres; it cannot be excluded that piracy does affect sales for segments, for which we only had very few titles. However, the fact that we do not observe segment-specific treatment effects suggests that genre of the book and its target group may not be strong modulator of the piracy effect. Furthermore, it should be noted that the publishers had an incentive to select the titles that would be affected negatively, because they would then benefit from free protection of, on average, half of them *and* they could use study results in their lobbying for stronger anti-piracy protection. To the extent that they were actually able to pick those, we would, if anything, predict that piracy could have a relatively *positive* impact on the non-selected titles.

Fourth, it is possible that we are not able to observe any effect on sales, because other substitutes for our titles were also available through unauthorized channels. Indeed, if a consumer could easily substitute one title with another, he would not switch to legal channels after not having found his book of choice on the internet, but instead would just download some other title. It would seem, however, that this is of more relevance for some genres than others; thus, again, the fact that we do not see differentiated treatment effects across segments, provides some limited evidence against this concern. Unfortunately it seems impossible to put the impact of preventing availability of *all unauthorized books* to a proper experimental test.

Our work contributes to the current research done on piracy in two main ways. First, to the best of the authors' knowledge it is the first study on piracy's effect on sales in the book industry, with a large sample of books and a long period of analysis. We therefore provide the first major insight into piracy's impact on this industry. Second, we have applied an innovative method that allowed to eliminate the causality issues – a common problem in studies on piracy. Similar design could be applied to other industries as well although controlling online availability may be very difficult in other markets.

Recent years have seen quite a few negative findings on piracy effect in various industries. For example, Koh et al. (2013) suggested that the threat to music industry may have already passed, with the industry now successfully competing with the free, unauthorized alternative. Similarly, Martikainen (2011) found no impact of file-sharing on DVD sales in 2009 in the USA. Lilewise, Adermon and Liang (2014) observed no impact of downloading on sales of DVDs and cinema tickets. We believe our results represent an important addition to this body of evidence suggesting that the fear of piracy may have been excessive.

References

Adermon, A., Liang, C.-Y., Sep. 2014. Piracy and music sales: The effects of an anti-piracy law. Journal of Economic Behavior & Organization 105, 90–106.

Baum, C. F., Hristakeva, S., Jul. 2014. DENTON: Stata module to interpolate a flow or stock series from low-frequency totals via proportional denton method.

Blackburn, D., 2006. The heterogeneous effects of copying: The case of recorded music. Job Market Paper (Harvard Ph. D. Programme).

Coelho, P., 2008. Pirate coelho. http://paulocoelhoblog.com/2008/02/03/pirate-coelho/.

Danaher, B., Dhanasobhon, S., Smith, M. D., Telang, R., Oct. 2010. Converting pirates without cannibalizing purchasers: The impact of digital distribution on physical sales and internet piracy. Marketing Science 29 (6), 1138–1151.

Danaher, B., Smith, M., Telang, R., Chen, S., 2012. The effect of graduated response anti-piracy laws on music sales: Evidence from an event study in france. SSRN Scholarly Paper ID 1989240.

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

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

Doctorow, C., Aug. 2009. Why free ebooks should be part of the plot for writers. http://www.guardian.co.uk/technology/2009/aug/18/free-ebooks-cory-doctorow.

Fine, M., 2000. Soundscan study on napster use and loss of sales. Tech. rep., RIAA.

Flint, E., 2002. Prime palaver #6. http://www.baen.com/library/palaver6.htm.

Givon, M., Mahajan, V., Muller, E., 1995. Software piracy: Estimation of lost sales and the impact on software diffusion. The Journal of Marketing 59 (1), 29–37.
URL `http://www.johnstockmyer.com/enmu/JMPiracyArticle.pdf`

Greevy, R., Lu, B., Silber, J. H., Rosenbaum, P., 2004. Optimal multivariate matching before randomization. Biostatistics 5 (2), 263–275.

Gunter, W. D., 2009. Internet scallywags: A comparative analysis of multiple forms and measurements of digital piracy. W. Criminology Rev. 10, 15.
URL `http://wcr.sonoma.edu/v10n1/Gunter.pdf`

Hansen, B. B., Klopfer, S. O., 2006. Optimal full matching and related designs via network flows. Journal of Computational and Graphical Statistics 15 (3), 609–627.

Hilton, J., Wiley, D., 2010. Free: Why authors are giving books away on the internet. TechTrends 54 (2), 43–49.
URL `http://www.springerlink.com/content/6t6388667v317711/abstract/`

Hilton, J., Wiley, D., 2011. Free e-Books and print sales. Journal of Electronic Publishing 14 (1).
URL `http://hdl.handle.net/2027/spo.3336451.0014.109`

Hu, Y. J., Smith, M. D., 2013. The impact of ebook distribution on print sales: Analysis of a natural experiment. SSRN Scholarly Paper ID 1966115, Social Science Research Network.

Ingram, M., Aug. 2012. The e-book lending wars: When authors attack. http://gigaom.com/2012/08/11/the-e-book-lending-wars-when-authors-attack/.

Kawohl, F., Kretschmer, M., 2003. Abstraction and registration: Conceptual innovations and supply effects in prussian and british copyright (1820-55). Intellectual Property Quarterly 2 (2).

Koh, B., Murthi, B. P. S., Raghunathan, S., 2013. Shifting demand: Online music piracy, physical music sales, and digital music sales (shifting demand: Piracy, physical & digital music sales).

Liebowitz, S. J., 2008. Sequel to liebowitz's comment on the oberholzer-gee and strumpf paper on filesharing. SSRN Scholarly Paper ID 1155764.

Liebowitz, S. J., 2010. The oberholzer-Gee/Strumpf file-sharing instrument fails the laugh test. SSRN Scholarly Paper ID 1598037.

Martikainen, E., Jan. 2011. Does file-sharing reduce DVD sales? SSRN Scholarly Paper ID 1742443, Social Science Research Network.

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

O'Leary, B., May 2009. Impact of P2P and Free Distribution on Book Sales. O'Reilly Media, Inc.
URL `http://www.google.pl/books?id=A1mgZxf3EE0C`

O'Reilly, T., 2002. Piracy is progressive taxation, and other thoughts on the evolution of online distribution. OpenP2P. com.

Peukert, C., Claussen, J., Kretschmer, T., Aug. 2013. Piracy and movie revenues: Evidence from megaupload: A tale of the long tail? SSRN Scholarly Paper ID 2176246, Social Science Research Network.

Pogue, D., 2008. Can e-publishing overcome copyright concerns? http://pogue.blogs.nytimes.com/2008/05/22/can-e-publishing-overcome-copyright-concerns/.

Rob, R., Waldfogel, J., 2007. Piracy on the silver screen. The Journal of Industrial Economics 55 (3), 379–395.
URL `http://www.jstor.org/stable/4622391`

Rubin, D. B., 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association 74 (366).

Rubin, D. B., 1980. Bias reduction using mahalanobis-metric matching. Biometrics 36 (2).

Smith, M., Telang, R., 2009. Competing with free: The impact of movie broadcasts on DVD sales and internet piracy. MIS Quarterly 33 (2), 321–338.

Zegners, D., 2014. Voluntary disclosure of product information: The case of e-book samples. SSRN Scholarly Paper ID 2380216, Social Science Research Network, Rochester, NY.

Zentner, A., 2006. Measuring the effect of online music piracy on music sales. Journal of Law & Economics.

Zentner, A., 2011. Measuring the impact of file sharing on the movie industry: An empirical analysis using a panel of countries. SSRN scholarly paper, Social Science Research Network.

Zhao, Z., 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and monte carlo evidence. Review of Economics and Statistics 86 (1).

A Data appendix

Table A.1: Publishers and titles by treatment groups

Publisher	CT	ET	Total
Cambridge University Press	9	9	18
Czarne	26	27	53
Insignis	3	3	6
Jaguar	7	6	13
Lexis Nexis	19	21	40
Proszynski	5	5	10
Sonia Draga	13	12	25
Wolters Kluwer Polska	24	24	48
Wydawnictwa Komunikacji i Łacznosci	15	13	28
Total	121	120	241

Table A.2: Locations of the unauthorized copies

Source	CT	ET	Total
5fantastic.pl	2	2	4
chomikuj.pl	28	6	34
download.freebiz.pl	2	0	2
forumwpia.org	0	1	1
freedisc.pl	0	3	3
sendspace.com	0	2	2
share.pdfonline.com	0	1	1
torrenty.org	0	1	1
ulozto.net	0	1	1
uploaded.net	0	1	1
vgi***.com (personal website)	1	0	1
Total	33	18	51

Note: Results from a manipulation check which consisted of seeking unauthorized copies for a random subsample from the titles participating in the experiment.

Faculty of Economic Sciences
University of Warsaw
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl