GRAPE

Group for Research in APplied Economics

GRAPE Working Paper #93

Internet piracy and book sales: a field experiment

Wojciech Hardy, Michał Krawczyk, Joanna Tyrowicz

FAME | GRAPE, 2023

Foundation of Admirers and Mavens of Economics
Group for Research in Applied Economics

Internet piracy and book sales: a field experiment

Wojciech Hardy
University of Warsaw

Michał Krawczyk
University of Warsaw, and JCR

Joanna Tyrowicz
University of Warsaw, University of Regensburg, IZA and FAME | GRAPE

Abstract

The widespread Internet ``piracy'' continues to fuel the debate about business models impervious to copyright infringement. We studied the displacement effects of ``piracy'' on sales in the book industry. We conducted a year-long large-scale field experiment: in the treatment group, we removed unauthorized copies appearing on the Internet and observed the sales data, whereas in the control group, we simply observed sales. We were able to substantially curb the unauthorized distribution, which resulted in a small, positive effect on sales. While using classical analysis we found it not to be significantly different from zero, a Bayesian approach using previous ``piracy'' studies to generate a prior led to the conclusion that protecting from piracy resulted in a significant sales boost of about 9 percent..

Keywords:

digital piracy, copyright infringement, sales displacement, books, field experiment

JEL Classification

C93, D12, K42, L82, O34

Corresponding author

Michał Krawczyk, mkrawczyk@wne.uw.edu.pl

Acknowledgements

The authors would like to thank Wlodzimierz Albin, Justyna Kossak and Magdalena Przek-Slesicka of Wolters Kluwer Polska, Krzysztof Gutowski and Sebastian Kawczynski 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 Sylwestrzak, Agnieszka Borowska and Lucas van der Velde for research assistance. The earlier versions of this paper received valuable comments from Raul Caruso, Andrea Galeotti, Lorenz Goette, Jakub Growiec, Keith Marzilli Ericson, Tomasz Michalak, Imke Reimers, Koleman Strumpf, Juuso Valimaki, as well as participants of seminars and conferences in Alicante, Lausanne, Warsaw, Aix-en-Provence (EALE 2014) and New York (EEA 2015). The remaining errors are ours. The authors gratefully acknowledge the support of the National Science Centre, grant 2011/01/D/HS4/03937. All opinions expressed are those of the authors and have not been endorsed by the NSC..

.

Published by: FAME | GRAPE
ISSN: 2544-2473

Foundation of Admirers and Mavens of Economics
Koszykowa 59/7
00-660 Warszawa
Poland

W | grape.org.pl
E | grape@grape.org.pl
TT | GRAPE_ORG
FB | GRAPE.ORG
PH | +48 799 012 202

1 Introduction and motivation

With digitisation, many businesses were disrupted by the emergence of zero-price alternatives in the form of online "piracy". Indeed, unauthorised file-sharing has often been blamed for the decline in legal sales of music and films throughout the 2000s (Kim et al., 2018). The rise of e-book technology and the proliferation of mobile devices introduced the threat of sales displacement to the book publishing industry as well (Wischenbart, 2014). Nielsen and Digimarc (2017) argued that the losses to the book industry reached approx. $315 million in the US alone.[1]

Theoretical literature, however, suggests that "piracy" could carry both positive and negative effects on legal consumption. Besides the potential displacement, "piracy" could facilitate word-of-mouth, exposure, promotion or exploration (see e.g. Bounie et al., 2007) or simply reflect additional consumption occurring side-by-side with the paid one (but not instead of it). Moreover, many researchers proposed alternative explanations for the decline in sales of cultural goods in the 2000s (e.g., Boldrin and Levine, 2008; Martikainen, 2014; Krueger, 2019).

The problem remains debated due to the difficulty in capturing the causal relationship. First, consumers who consume more content often do so through both authorised and unauthorised sources. Second, higher-quality and more popular content is typically consumed more through both sources. Third, any kind of "piracy" is difficult to measure as it requires either online user tracking or self-reporting of controversial behaviour. Past research applied various approaches including panel surveys, instrumental variables and quasi-experiments that often lacked a clear counterfactual and were non-replicable in other contexts.

We propose a new method of analysing the effects of piracy that does not suffer from reverse causality or omitted variables issues and can be replicated across different countries, industries and years. In our field experiment, we worked with publishers who provided recent or upcoming book titles for the study. We then combined the titles into pairs or small groups of comparable items. Within each of these pairs or groups, half were randomly assigned to a group protected from unauthorised sharing. Finally, we compared the sales of books from the protected and non-protected groups.

We find that our experimental treatment successfully lowered unauthorised availability of the treated book titles. As our sample is statistically underpowered, when assessing the impact on sales, we complement the classical approach with Bayesian statistics. We find evidence of an effect comparable to that identified in other studies of "piracy" effects.

We first discuss the relevant empirical work and how our study stands out in terms of its design. We then provide a thorough description of our approach at each stage of the experiment, including design, communication with the stakeholders, control of the implementation, evaluation and feedback. We summarise our main findings in Section 4. Finally, we provide a discussion of our results and the limitations of our study, suggesting how this framework could be re-applied to other industries.

[1] Notably though the estimate was based on a survey of e-book readers and questions such as "what would you do if this ["pirated"] version was not available?"

2 Sales displacement and online "piracy" in books: the framework

Typical market competition perspectives assume that both the product and its potential substitutes have a positive price and compete within symmetrical sets of rules. These assumptions are violated within the context of digitised cultural goods, whereas "pirated" copies have a price of zero, are distributed illegally, and are ethically controversial.

This asymmetry implies that "pirate" distribution occurs outside of the industry networks and is rarely measured directly. Consequently, while data on sales can be accessed at the very least by publishers and distributors, data on unauthorised distribution is difficult to collect. Even if such data can be approximated, it is difficult to establish the relationship between sales and file-sharing. This is because the general quality and popularity of cultural goods typically translates into both higher sales and more unauthorised downloads.

An ideal approach to solving this issue would be to manipulate the supply of unauthorised content. In a typical competition setting, this could be done by, e.g., exogenously changing the price of one good while holding that of the other fixed. However, in the context of "piracy," the closest alternative is to affect the availability of the "pirated" copies themselves.

In this vein, several studies looked at the impact of "pirate" websites shutdowns. Danaher and Smith (2014) and Peukert et al. (2017) studied the effects of the sudden shutdown of the prominent file-hosting website Megaupload. The studies reached different conclusions, with Danaher and Smith (2014) finding an increase in legal digital downloads and Peukert et al. (2017) finding no evidence of a change in box office revenue. One reason for this discrepancy may be that a "pirated" copy is a better substitute for a copy downloaded from a legitimate source than for a visit to the movie theater.[2]

A larger study by Danaher et al. (2019) considered three events of "pirate" website blocking in the UK. The three blocks covered one prominent website, 14 websites, and 52 websites respectively. Notably, the authors found that while the first block mainly redirected users to other unauthorised sources, the other two successfully curbed "piracy", increasing authorised consumption by approximately 7%-12%.

In the most related study, Reimers (2016) measured the effects of take-down notices as means of protecting books from "piracy". For identification, she exploited the variation in the timing of the takedown notices (TDN). Applying a difference-in-differences identification strategy, she found an increase of 10-20% in sales of protected e-books and no evidence of effects for physical books. Additionally, Reimers (2016) showed that the timings of protection measures for particular titles were not related to the title popularity (measured via Google Trends) and that thus the timings could be considered as randomised.

Finally, some insight into the impact of "piracy" on authorised sales may also be gained indirectly from studies looking at the cannibalization of print formats by e-books. Chen et al. (2019) studied the relationship between the e-book premiere date of titles sold in physical form by utilizing a **quasi-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 negatively impacted the e-book sales. The authors argued

[2]The two studies also differ methodologically and in the sample of countries covered.

that most consumers make a choice of the mode of consumption (physical or digital) first. Therefore, e-books are a weak substitute to physical copies (consistent with Lee et al., 2014, though these authors also note substantial heterogeneity of observed effects).

While these studies provide important information, they struggle with three main challenges. First, as researchers have no say in title protection choices, they need to prove that the assignment was randomised (or that the selection criteria can be determined) for the experimental approach to work. This additionally means that some factors such as title-specific publisher strategies (based on whether a title is protected or not) cannot be fully controlled for. In the context of website shutdowns, the lack of a clear counterfactual forces the authors to look for a substitute identification strategy. Second, the analyses cannot be reliably replicated for other countries, industries, or time periods. This is because they all rely on events and processes outside of the researchers' control.

In our study, we design and document an approach that could be replicated - even though cooperation with industry stakeholders would be required. Our method allows us to start the treatment execution with a reliable counterfactual set of titles. We also monitor the process of anti-piracy protection from the start. Moreover, the publishers in our study are unaware of which titles belong to the treated group and therefore cannot make strategic choices (e.g. advertisement expenditures) based on this information. As stated by Danaher et al. (2014), "controlled experiments are the 'gold standard' [because they do not] suffer from the endogeneity problem". These authors find it regrettable that controlled field experiments on "piracy" are virtually nonexistent. Our study fills this important gap.

3 Experimental design

The purpose of this experiment was to verify the hypothesis that unauthorised online sharing of books is detrimental to their sales in authorised channels. To do so, we partnered with an anti-piracy agency specialised in the detection and removal of unauthorised copies (via TDNs), as well as with several prominent Polish publishers who provided book titles for the study. We randomly assigned half of the titles for protection from unauthorised online distribution, while leaving the other (a priori, comparable) half without protection. We collected sales data (see Appendix F) and unauthorised distribution data for both groups. The experiment ran for a whole year, making our analysis robust to seasonal effects. We describe the core of the experimental design in three stages: recruitment of publishers, treatment assignment and treatment execution (and effectiveness). Detailed information for each step is available via online appendices.

3.1 Recruitment of publishers

With the help of the Polish Book Chamber and our e-mail campaign, 70 Polish book publishers were invited to participate in the experiment.[3] We were also supported by independent literary agents, who invited the publishers on our behalf. We subsequently called those book publishers to reiterate the invitation and give them a chance to ask questions about the experiment. At this stage of the

[3]Background information on the Polish book market and its recent changes is provided in Appendix E.

experiment, we informed the publishers about the objective of the experiment, its duration, and the intended treatment.

Publishers were asked to provide a list of up to approximately 50 books each and to commit to providing sales data on these books.[4] We were explicit both in writing and in verbal communication that the assignment to treatment and control groups was to be random and could not be influenced by the publishers. We openly explained that we want to mitigate the risk that the publishers selectively change pricing or other elements of marketing for the treated books.[5]

Eventually, 13 major publishers agreed to take part in the experiment. One of them failed to supply the data necessary for the start of the study. The titles from the remaining 12 publishers were used in the treatment assignment (see section 3.2). By the end of the study, three publishers dropped out: two failing to provide the sales data and one due to interference in the treatment procedure. We discuss the details of the recruitment procedure, the market coverage, and publisher dropouts in Appendix A.

The initially participating publishers provided between 5 and 53 book titles, which was typically only a fraction of their catalog. The invitation letter specifically asked that the books selected for the experiment be relatively new (or even forthcoming) and relatively popular. This sets our study apart from the study of Reimers (2016). To identify other potential undisclosed selection criteria, we compared the provided lists to the rest of the publishers' catalogs (see Appendix B). We found little evidence of selection on variables other than the release date.

The publishers provided detailed data on each book. Although we had 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 A2 reflects this updated list. Additionally, the information included the author's name, page count, price for different formats (paperback, hardcover, audiobook, e-book) and date of release. We have also acquired data on the number of previous editions (if any), the first print run, and past sales (if any). Publishers were also asked to provide quarterly or (preferably) monthly sales forecasts. After the first half-year and after the end of the study, publishers provided sales reports for each title. For a description of the provided book data, see Table A4.

3.2 Treatments and treatment assignment

The experimental treatment in this study consists of issuing automated takedown notices for the identified unauthorised copies of the protected titles in our sample. If they comply with requests to remove infringing content, platforms are generally not considered liable for hosted content. Automated removal of files following TDNs issued by copyright holders or their authorised representatives constitutes a popular low-cost compliance technology. It was also implemented by most file-hosting platforms in Poland.

In this experiment, we enlisted the help of Plagiat.pl, the largest Polish agency specialising in online protection of copyrighted content. Plagiat.pl routinely browses through internet sources in search of unauthorised content and issues TDNs whenever a copy is found. Of the finally participating

[4] We signed non-disclosure agreements with all the participating publishers.

[5] In any event, the publishers' control of the retail prices is limited in the Polish market: the bookstores typically sell at substantial and discretionary discounts with respect to the list price.

publishers, all reported that they had not applied any anti-piracy strategies prior to the experiment, although they believed that "piracy" hurt their sales.[6]

We provided Plagiat.pl with a complete list of books and the treatment assignment. Plagiat.pl did not inform publishers about the treatment assignment and did not provide publishers with any information during the experiment. Instead, Plagiat.pl performed its tasks and reported its findings for both title groups (on a monthly basis) to us. No steps were taken against the file uploaders (who remained anonymous).

The participating publishers varied in segments (see Table A2), pricing, formats, as well as business models (e.g. promoting domestic authors versus translating foreign bestsellers). Due to a high number of such factors, we decided to employ a matched-subject design rather than pure randomisation of the treatment (Appendix C). The matched-subject design enables the efficient use of within-treatment heterogeneity in estimating the treatment effects (see for example Meyer and Van Klaveren, 2013; Hainmueller et al., 2015), enhancing covariate balance between treatment groups and therefore the power of statistical tests.

The matching procedure was conducted for 12 publishers, but (as noted in section 3.1), only nine of them remained in the sample for the final analysis. In the remaining sections, we describe the data and conduct the analyses for these nine publishers. Importantly, only five titles from the retained publishers were *a priori* matched with the titles of dropped publishers, with one also having a direct match with another retained title (see Appendix C). This is because most of the titles in the sample had their best match within the same publisher.

3.3 Treatment execution

The treatment execution started on the same date for all the books released prior to the experiment and on the first day for those released during the experiment. The publication date has been used in matched-subject randomization, hence the books in treated and control groups are perfectly comparable along this dimension.

Plagiat.pl identified unauthorised copies on 53 different websites. For books in the Enforcement Treatment (ET, "treated" books), Plagiat.pl traced the unauthorised distribution and issued TDNs to the service providers. TDNs affect all copies of unauthorised files, regardless of how many users share it. These actions were repeated continuously and TDNs were reiterated every time given content reappeared without authorization. The agency also verified whether the TDNs were complied with and reported its findings. For books assigned to the Control Treatment (CT, "untreated" books), Plagiat.pl recorded the number of copies distributed without authorization and reported them to us without issuing TDNs. The findings were compiled in monthly reports, including the number of unauthorised copies of titles from both the CT and ET samples.

We verify the effectiveness of our treatment empirically through two manipulation checks. First, we compared the availability of the unauthorised copies of books belonging to CT and ET using the reports from Plagiat.pl. We find a significant drop in the number of copies in the ET relevant to those in CT. Second, to confirm whether this translated into lower chances of finding copies of the books,

[6]Notably, they presented no evidence for why this would be; still, they tended to be frustrated about the availability of their books on file-hosting websites.

we asked three research assistants to search the Internet for copies of randomly selected titles from both CT and ET. We found lower chances of finding the copies for the ET titles relative to the CT titles. For the copies found, the time for the search was longer and the search often ended in less known and popular (and thus riskier) sources. We describe both approaches and analysis in more detail in Appendix G. In the Appendix, we also discuss the relationship between the number of available copies and the number of downloads based on supplementary data provided to us by the leading file-hosting platform Chomikuj.pl for another time period. We find a positive link between the number of available copies and the number of downloads. We also notice that the relationship gains strength for higher numbers of copies available – i.e. that reducing the number to only a few copies might be enough to curb downloading as well.

4 Results

We analyse the treatment effects from three angles. First, we report basic sample statistics. Because our experiment is a randomised trial, we can interpret these results as primary evidence. However, if treatment effects are heterogeneous, means/medians may be uninformative. We thus, secondly include regression analyses. Finally, we take a Bayesian perspective to consider our results in the light of other studies on the effects of "piracy", taking the results of the most notable ones as *a priori* information for the analysis.

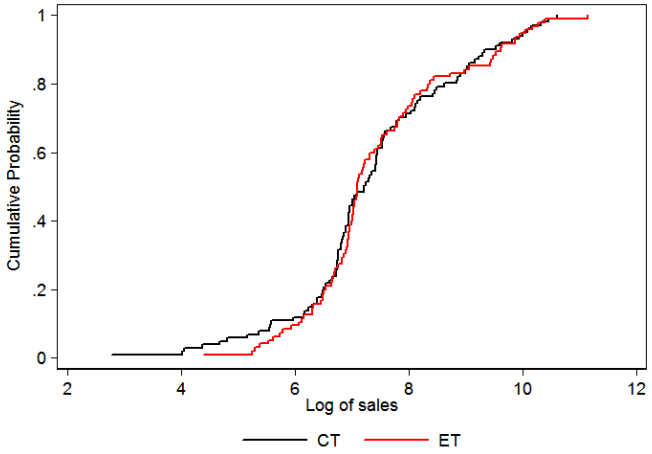
4.1 Statistical tests of the means

The comparison of mean sales between the two treatments reveals that on average, sales were approximately 5% higher in ET. Yet, the difference is not statistically significant. Figure 1 reports cumulative distribution functions for the (log of) sales in ET and in CT.[7] The observed differences between CT and ET remain statistically insignificant.

Statistical tests cannot reject the null hypothesis that there is no difference in sales across treatment and control groups. Table 1 reports the results of Mann-Whitney's non-parametric rank test on overall sales of the analysed books. Additionally, we perform the Wilcoxon test on matched pairs based on our design. The maximised/minimised 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. Regardless of specification, the data fail to reject the null hypothesis of no difference in sales between the ET and CT groups. As a last check, we test for differences in variance in both groups – it is plausible that the treatment affected various genres or titles in different ways. As the treatment could not have a bearing on the titles with no unauthorised copies ever uploaded, we repeated all the tests restricted to the sample that saw at least one copy of the "pirated" version available at some point. Again, we find no statistically significant difference.

[7] We take the natural logarithm of copies sold. Since for a fraction of our titles, during the experiment window, sales were smaller than returns, we added the first print run to the reported sales for all the titles. Note that the titles were matched on the first print run prior to the treatment assignment, hence the first print run is orthogonal to the treatment and does not affect the estimations.

Figure 1: Cumulative distribution functions of sales (in logarithms)



Note: sales (topped with first print run figures to avoid negative values, see footnote 7), in the sample of 148 titles that were ever observed to be subject to unauthorised file-sharing.

Table 1: The difference in sales between CT and ET - test statistics

Difference in sales	Mann-Whitney test	Wilcoxon test		Levene test
		maximized variant	minimized variant	of equal variance
total 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
if unauthorized copy identified	$z = -0.710$	$z = -0.995$	$z = -0.139$	$W = 2.128$
(*p-value*)	(0.4779)	(0.32)	(0.89)	(0.147)
n	CT:82, ET:72	56 pairs	56 pairs	CT:82, ET:72

Notes: To improve the robustness of our results, we use two extreme approaches. First, we take the max of the annual sales of the treated books and the min of the untreated ones 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 the max of the untreated within each group ('minimized'), which makes it easiest to observe a negative treatment effect. The 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.

4.2 Regression analysis

The lack of evidence for a significant effect may be driven by insufficient statistical power. Given the dispersion in sales across book titles as observed in our data, with a treatment effect of approximately 10%, the statistical power of 80% would require a sample of approximately 900 rather than the current 239 titles. Though the issue is partially offset by matched-subject randomization, it is possible our sample size still falls short of the required number.

To mitigate this problem, we re-estimate the treatment effect in a regression model. In these specifications, we proxy for the intensity of digital "piracy" using information about the length of unauthorised availability and electronic (authorised) distribution. We report the results in Table 2: the (intensity-adjusted) treatment effect is estimated at roughly 10%, though it remains statistically non-significant. We study separately the titles that were ever distributed in unauthorised channels and the titles for whom the TDNs had a permanent effect (i.e. reduced the number of unauthorised copies to 0 with no reappearance). Note that this estimator is less underpowered than the tests on means, medians and variances in Table 1. The estimated effects remain statistically non-significant even though their sign and size seem in line with prior literature on displacement effects of "piracy".

Table 2: Estimating the treatment effect - OLS specification

	Whole sample					Only titles ever available in unauthorized channels					CT and ET with no copies reappearing				
Treatment	0.109	0.064	0.140	0.076	0.201	0.160	-0.004	0.344	0.027	1.029	-0.095	-0.116	0.337	0.095	-0.368
	(0.203)	(0.209)	(0.256)	(0.237)	(0.387)	(0.250)	(0.261)	(0.564)	(0.530)	(0.741)	(0.375)	(0.406)	(0.442)	(0.406)	(0.775)
log(number of copies)			0.269*	0.189	0.242*			0.251*	0.107	0.226			0.338	0.359*	0.338
			(0.144)	(0.136)	(0.144)			(0.141)	(0.131)	(0.140)			(0.211)	(0.209)	(0.212)
Ebook exists (EE)		1.852***	1.696***	1.305***	1.701***		1.809***	1.623***	1.223***	1.628***		1.556***	1.366***	1.028***	1.366***
		(0.252)	(0.248)	(0.253)	(0.246)		(0.290)	(0.290)	(0.291)	(0.286)		(0.354	(0.349)	(0.357)	(0.351)
Treatment * EE		-0.082	-0.144	-0.066	-0.272		0.074	0.057	-0.027	-0.212		-0.033	0.156	0.577	0.474
		(0.353)	(0.349)	(0.317)	(0.391)		(0.412)	(0.408)	(0.369)	(0.480)		(0.741)	(0.721)	(0.668)	(0.843)
Time available (TA)			-0.029	-0.039	-0.020			0.012	-0.024	0.021			-0.035	-0.078	-0.035
			(0.054)	(0.051)	(0.054)			(0.075)	(0.070)	(0.075)			(0.08)	(0.077)	(0.081)
Treatment * TA			0.076	0.044	0.042			0.049	0.037	-0.032					
			(0.048)	(0.047)	(0.056)			(0.076)	(0.074)	(0.093)					
Segment	No	No	No	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Treatment * segment	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	No	Yes
Constant	7.150***	6.538***	6.230***	7.016***	6.227***	7.361***	6.703***	5.912***	7.326***	5.898***	7.057***	6.615***	6.162***	7.219***	6.162***
	(0.143)	(0.145)	(0.194)	(0.298)	(0.193)	(0.173)	(0.175)	(0.501)	(0.558)	(0.495)	(0.177)	(0.189)	(0.272)	(0.481)	(0.274)
R^2	0.00	0.32	0.37	0.51	0.41	0.00	0.36	0.42	0.58	0.46	0	0.2	0.26	0.46	0.29
N	228	228	228	228	228	148	148	148	148	148	104	104	104	104	104

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$, standard errors in parentheses. Estimated with OLS.
The explained variable is the natural logarithm of aggregate sales topped with the first print run (see footnote 7). The nature of the sales data coupled with missing first print data lowers the sample to 228 titles (see Appendix F for details). The number of copies is defined as monthly average over the 12 months observed (+1 so that the logarithm is defined also for titles that were never shared).

4.3 Regression analysis with priors

The regression analyses make the effect size more in line with other literature on "piracy" effects. However, the estimates of treatment effects remain statistically insignificant. With no other source of information, we cannot reject the null hypothesis for of no detrimental effects of "piracy" on sales. To further improve the information value of our study, we leverage information from other studies on the effects of "piracy" on sales. While these studies differ in the specifics of industry and measurement, they make it possible for us to deploy a Bayesian approach, using the findings from other studies as a *prior*.

Non-Bayesian (i.e. frequentist) methods typically yield fixed point estimates and derive their conclusions purely from the observed data. Bayesian analysis combines prior beliefs (expressed as probability distributions) with observed data to provide a posterior distribution that encapsulates a range of plausible parameter values. This method allows us to account for prior evidence on "piracy" effects when estimating the effects of the analysed treatment. An *informative prior*, in this context, can be based on estimates from prior empirical literature, including their variance. The resulting *posteriori* outcomes combine inference from the data and the prior findings (for more information on the implementation of Bayesian regression and choosing priors see e.g. Korner-Nievergelt et al., 2015; Muth et al., 2018).

At first glance, the study of Reimers (2016) might be considered a suitable source of priors. Indeed, Reimers (2016) studied (non-experimentally) the effects of of a similar TDN strategy on sales. However, the books in her sample were quite different. First, their electronic versions represented a much larger part of sales. Second, the books were much older: in our sample, all the books were published within the last 15 years, whereas in Reimer's data only "a few". Other studies suggest this could play a big role. For one, older books could be more susceptible to the promotional effects of "piracy". Tanaka (2019) finds that a similar protection strategy conducted for manga comic books in Japan boosted the sales of ongoing series but harmed those of the (typically older) finished ones. In a related setting, Nagaraj and Reimers (2021) found that digitization of older books – and particularly less popular ones – increased their print sales, while Chen et al. (2019) found a non-significant negative effect for recent books and Sharma et al. (2019) observed a significant negative one. As a result, the study of Reimers (2016) cannot inform our beliefs about our findings.

By contrast, the study of Danaher et al. (2019) - despite focusing on movies - bears several important similarities to our context. First, the study analyses a similar time period of 2012-2014. Second, it focuses on ongoing consumption, dominated by current titles rather than older ones. Third, it analyses the effects of website blocking, which amounts to a considerable reduction in availability of unauthorised copies (i.e. treatment similar to ours). Fourth, it was conducted in the UK, a European country and (back then), an EU member. Arguably, legal and cultural differences between the UK and Poland are smaller than those between Poland and Japan (Tanaka, 2019) or the USA (Reimers, 2016).

Danaher et al. (2019) consider several specifications, concluding that the effects fall within the range of 7-12%. We base our *prior* on these conclusions, assuming a normal distribution with the mean at the middle of these values (i.e. 9.5%) and a standard deviation placing 95% of the distribution within this range (i.e. 1.25%). We also verify the possibility of a less informative *prior* by

instead assuming that 68% of the distribution falls within that range (i.e a std. dev. of 2.5%). For the estimation, we mimic our prior OLS approach from Table 2), using the Bayesian generalised linear models implemented in the *rstanarm* package for R – estimated using the MCMC method and assuming a Gaussian distribution. For the specifications with control variables, we use the approach implemented in the *rstanarm* package aimed at providing *weakly informative priors* by default. The *priors* for the other coefficients thus assume a normal distribution centered on 0, with the variance rescaled statistic based on the standard deviations of the explained variable and the respective covariate (see: Goodrich et al., 2023). The informative *prior* used for the coefficient on the Treatment variable remains non-rescaled.

Owing to the additional information from the *prior*, the *posterior* uncertainty intervals show a positive effect of the protection. In the strongly informative *prior* scenario, the effect is close to that of Danaher et al. (2019), falling within the range of 7.4% and 11.7%. When assuming a less informative *prior*, the interval widens to 5.6%-13.5%, thereby still excluding a null relationship. Across the different specifications of the model included in Table 2 (i.e. when looking only at the titles ever available in unauthorised channels or ones with no reappearing copies; and adding control variables) the results for the posterior intervals remain stable with the lower and upper bounds shifting by at most 1 pp. While these results support our main hypothesis, they also show that the posterior intervals are largely driven by the information from the *prior* rather than the more volatile information from our data. Assuming a standard deviation of 5% in the *prior* (i.e. lowering the share of values in the range provided by Danaher et al. (2019) to as few as 38%) yields a consistent lower bound of 1% for the posterior intervals across the specifications, i.e. only just above the null. Unfortunately, as few other studies considered the complete removal of pirate availability, we cannot test our data against other *priors*.

5 Conclusions

We proposed an experimental framework to test if the availability of unauthorised copies has an effect on sales. We worked with prominent Polish book publishers who provided titles for the study. We then split the sample into two comparable groups and protected one from unauthorised online sharing for a full year with the help of an agency specialised in the enforcement of takedown notices. Finally, we analysed sales data for the two groups, which were provided by the publishers. While ultimately our sample proved small in terms of statistical power, we found our results do not go against those from related studies, including the estimates that Reimers (2016) reported for printed books. To verify this, we supplemented traditional econometric approaches with a Bayesian perspective, where information from a prior study with a similar design was combined with the information derived from new data. This analysis is more supportive of a positive effect of our treatment than a negative one, although the low number of observations in our data remains an issue.

Our main contribution, however, comes from the applied method. Our design allowed us to greatly reduce causality issues while providing us with a clear counterfactual and full control of the treatment. This is in contrast to prior literature on “piracy”, where even the most robust studies rely on one-time events that were often specific to particular countries, and which were outside of the control of the

researchers. While the mentioned studies provided important and definitive information about the effects of these events, these approaches cannot be replicated in other contexts.

Our proposed framework relies on the cooperation of several stakeholders, including publishing companies, anti-piracy agencies and researchers. It was made possible largely due to the complementary incentives of the participants. On the one hand, the publishers benefited from having some of their titles protected from “piracy” (which they suspected of hurting their sales). On the other hand, the agency had the opportunity to present its efficiency in removing unauthorised copies and the potential effectiveness of this approach to the publishers. This collaboration was easier because we – the researchers involved – handled almost all of the design and data analysis. At the end of the experiment, we also provided the participants with individualised reports on the outcomes.

Throughout the article, we document our approach and the design of our experiment. We hope this study can be used as a reference for future studies in the contexts of other industries and countries.

References

Boldrin, M., Levine, D.K., 2008. Against intellectual monopoly. volume 8. Cambridge University Press Cambridge.

Bounie, D., Bourreau, M., Waelbroeck, P., 2007. Pirates or explorers? analysis of music consumption in french graduate schools 50.

Chen, H., Hu, Y.J., Smith, M.D., 2019. The Impact of E-book Distribution on Print Sales: Analysis of a Natural Experiment. Management Science 65, 19–31. URL: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2017.2940, doi:10.1287/mnsc.2017.2940. publisher: INFORMS.

Danaher, B., Hersh, J.S., Smith, M.D., Telang, R., 2019. The effect of piracy website blocking on consumer behavior. Available at SSRN 2612063 .

Danaher, B., Smith, M., Telang, R., 2014. Piracy and copyright enforcement mechanisms, in: Lermer, J., Stern, S. (Eds.), Innovation Policy and the Economy, Volume 14. University of Chicago Press, Chicago, Illinois.

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

Goodrich, B., Gabry, J., Ali, I., Brilleman, S., 2023. rstanarm: Bayesian applied regression modeling via Stan. URL: https://mc-stan.org/rstanarm/. r package version 2.26.1.

Hainmueller, J., Hiscox, M.J., Sequeira, S., 2015. Consumer demand for fair trade: Evidence from a multistore field experiment. Review of Economics and Statistics 97, 242–256.

Kim, A., Lahiri, A., Dey, D., 2018. The'invisible hand'of piracy: An economic analysis of the information-goods supply chain. MIS Quarterly 42.

Korner-Nievergelt, F., Roth, T., Felten, S.v., Guelat, J., Almasi, B., Korner-Nievergelt, P., 2015. Chapter 15 - Prior Influence and Parameter Estimability in: Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN. Academic Press, Boston. URL: https://www.sciencedirect.com/science/article/pii/B9780128013700000150, doi:10.1016/B978-0-12-801370-0.00015-0.

Krueger, A.B., 2019. Rockonomics: A Backstage Tour of What the Music Industry Can Teach Us About Economics and Life. Broadway Business.

Lee, K., Han, K., Lee, E., Lee, B., 2014. How consumers' content preference affects cannibalization: an empirical analysis on e-book market, in: 35th International Conference on Information Systems, pp. 1–15.

Martikainen, E., 2014. Does file-sharing reduce dvd sales? NETNOMICS: Economic Research and Electronic Networking 15, 9–31.

Meyer, E., Van Klaveren, C., 2013. The effectiveness of extended day programs: Evidence from a randomized field experiment in the netherlands. Economics of Education Review 36, 1–11.

Muth, C., Oravecz, Z., Gabry, J., 2018. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. The Quantitative Methods for Psychology 14, 99–119. URL: http://www.tqmp.org/RegularArticles/vol14-2/p099, doi:10.20982/tqmp.14.2.p099.

Nagaraj, A., Reimers, I., 2021. Digitization and the Demand for Physical Works: Evidence from the Google Books Project. URL: https://papers.ssrn.com/abstract=3339524, doi:10.2139/ssrn.3339524.

Nielsen, Digimarc, 2017. Inside the Mind of a Book Pirate. Technical Report. Ontario Media Development Corporation.

Peukert, C., Claussen, J., Kretschmer, T., 2017. Piracy and box office movie revenues: Evidence from Megaupload. International Journal of Industrial Organization 52, 188–215.

Reimers, I., 2016. Can private copyright protection be effective? evidence from book publishing. The Journal of Law and Economics 59, 411–440.

Sharma, S., Telang, R., Zentner, A., 2019. The Impact of Digitization on Print Book Sales: Analysis using Genre Exposure Heterogeneity. URL: https://papers.ssrn.com/abstract=3579521, doi:10.2139/ssrn.3579521.

Tanaka, T., 2019. The Effects of Internet Book Piracy: Case of Comics. Technical Report. Institute for Economics Studies, Keio University.

Wischenbart, R., 2014. Global Trends in Publishing 2014. Annual report. Frankfurter Buchmesse Business Club. URL: www.book-fair.com/businessclub.