Internet “piracy” and book sales. A field experiment - Online Appendix

Wojciech Hardy* Michał Krawczyk† Joanna Tyrowicz‡

May 2023

*University of Warsaw
†University of Warsaw, corresponding author, email: mkrawczyk@wne.uw.edu.pl
‡FAME|GRAPE, University of Warsaw, and IZA

A Recruiting publishing houses

Using the advise from the Polish Book Chamber, we composed a list of 70 leading publishing houses in Poland. We contacted them by email, and explained the nature of the experiment. We emphasized that they could have the unauthorized copies of some of their titles removed from online distribution, free of charge. We explained that the procedure would be applied to a random selection of titles only. Finally, we explained that we would require information about sales of all the titles participating in the experiment. In addition to the official letter sent out by the Polish Book Chamber, an independent literary agent contacted *circa* 20 publishers. After the email campaign has ended, we contacted publishing houses by telephone, inquiring if they would be willing to participate in the experiment.

The publishing houses were at liberty to participate in the experiment and if they chose to participate, they were at liberty to provide the list of book titles from their current and future offer. In the phone interviews, most publishing houses showed little interest in participating. Most often, this was due to difficulties in reaching the person in charge. Managing directors were often too busy to give the project full consideration. We also experienced general skepticism towards experimental methods. In four cases, the good will of the managers lost the battle with urgent matters and the managers eventually ceased communicating. Some publishers, especially the smaller ones, said they lacked the workforce necessary to prepare the required data. Despite our readiness to sign a non-disclosure agreement, in some cases, the legal adviser of the publishing house would block the partnership due to data sensitivity concerns. The biggest concern voiced in all phone calls was the fear that their sales data would be disclosed to a third party, e.g. other publishing houses. Indeed, in Poland only few publishing houses acquire market information, whereas market data on book sales and market shares come from the sellers rather than from the publishing houses and are not comprehensive (i.e. for selected segments of the market, rather than the whole market). Only few publishers related to the crux of the intended manipulation, namely to the “piracy” itself. Of those few cases, one publisher informed us that they were already participating in an anti-“piracy” project, but did not specify its nature. In the remaining cases, the publishing houses representatives maintained either of the two extreme views: that their business did not suffer from “piracy” or that no research is necessary, because “piracy” obviously damages sales. In both cases, they insisted that engagement in the project was not likely to benefit them. To the extent that such insights were correct, the treatment effect in our sample may be stronger than would have been in the general population of publishers.

At the start, thirteen publishing houses decided to participate in the project. However, the WAB publisher did not provide the list of titles for the study, despite having signed the required agreements. Thus, titles from twelve publishers were used for the treatment assignment procedures and took part in the experiment. Over the course of the project, we detected an anomaly in the unauthorized availability of untreated titles of one of the publishers (PWN), forcing us to conclude anti-piracy efforts going beyond the scope of the study and to remove the publisher from the study. Finally, two publishers (Buchmann and Galaktyka) did not provide the required sales data at the end of the study period. The former engaged in a merger with another publishing house and hence lacked the administrative capacity to provide the data. The latter underwent a change in management and the new CEO did not approve disclosing the sales data to us, despite the previously signed agreement.

The participating publishing houses are established firms, functioning in the market for many years. Book sales reports for 2012 and 2013 (the years relevant for the experiment) reveal that academic, professional and technical books constituted the largest segment of the market with a share of 36.5%, followed by school books (31.5%) and general fiction (13.05%). The participating publishing houses constitute a fair share of the relatively dispersed market, especially within their respective segments (see Table A1).

Table A1: Publishing houses participating in the project

Publisher	Main segment	Ranking	Market share Total	Market share Segment	Main title/Notes
Wolters Kluwer Polska	Prof.	3	6.9%	18.8%	no single major title/series
Lexis Nexis	Prof.	9	1.8%	4.8%	legal commentaries
GW Foksal [dropped]	Fict.	10	2.0%	15%	former WAB, Buchmann and Wilga
Proszynski	Fict.	12	1.2%	8.8%	Interview with a Polish mafia boss, "Revival" by Stephen King
Sonia Draga	Fict.	15	0.8%	6.3%	"50 shades of Gray" by E.L. James
Czarne	Fict.	33	0.3%	2.2%	no single major title/series
Insignis	Fict.	35	0.2%	1.3%	"God never blinks" by Regina Brett
WKiL	Prof.	37	0.1%	0.3%	no single major title/series
Jaguar	Fict.	38	0.1%	0.9%	fantasy series of John Flanagan
Cambridge University Press	n.a.				CUP English textbooks
Dropped out – did not provide data					
Buchmann	Fict.	26	0.5%	4.0%	
Galaktyka	n.a.				series of Jeremy Clarkson books
WAB	Fict.	21	0.7%	5.4%	
Dropped out – engaged in own TDNs					
PWN	Prof.	4	4.7%	13.0%	no single major title/series

Note: data on market shares and segments comes from several reports, a financial statement and websites compiling information on the Polish publishing market - Biblioteka Analiz (2013b, 2014, 2016); Rynek-Ksiazki.pl (2012); wirtualnywydawca.pl (2015); Jaguar (2018). Note that PWN, Buchmann, WAB and Galaktyka are the four publishers removed from study. We report their market shares to highlight that our efforts were at least partially successful in reaching to major market participants. The data for Buchmann and WAB is from 2012; the two publishers later merged with a third and formed GW Foksal - presented with data for 2013. The sales figures for Lexis Nexis were taken from 2011 and the data for Jaguar was taken from 2016/17. All market shares calculated using 2012-13 totals. Rankings are an approximation based on all of the available data and the most comprehensive available ranking (from one of the links). Prof. stands for academic, professional and technical books; Fict. stands for general fiction. Both reflect the key segments of the participating publishers.

B Selection of titles

The selection of books participating in the experiment covers a wide selection of genres and audiences. In addition to general fiction, our experiment covered non-fiction, academic books, professional books, legal books, self-help, foreign language textbooks, etc. Table A2 summarizes the coverage.

One potential concern related to the book titles proposed by the publishers participating in the experiment concerns the segments of the book market. Relatively large share of our books are professional books (law and business/economics, science and research) and educational books (academic books and foreign languages books). Note that in Poland, neither tertiary education institutions nor language schools equip their students with the textbooks. In fact, the acquisition of books is entirely individual decision, but the institutions select which books will be followed in the curriculum. The professional and educational books are thus arguably less substitutable by a different book from authorized channel. This implies that the TDNs should have a stronger effect on sales, because substitutes are not available.

Table A2: Number of books per segment and publisher

	CUP	Cza	Ins	Jag	LN	Pro	SD	WKP	WKiL	All
General fiction	-	12	-	-	-	8	22	-	-	42
Fantasy and Sci-Fi	-	-	2	13	-	-	-	-	-	15
Non-fiction	-	40	2	-	-	2	2	-	-	46
Foreign languages	18	-	-	-	-	-	-	-	-	18
Academic books	-	-	-	-	18	-	-	14	-	32
Science & Research	-	-	-	-	5	-	-	1	-	6
Business & Economics	-	-	-	-	-	-	-	9	-	9
Law	-	-	-	-	16	-	-	21	-	37
Professional & Technical	-	-	-	-	-	-	-	-	22	22
Self-Help	-	-	2	-	-	-	-	2	6	10
Other	-	1	-	-	-	-	-	1	-	2
All	18	53	6	13	39	10	24	48	28	239

Note: Non-fiction books include biographies, memoirs, essays, etc. The publisher abbreviations are: CUP - Cambridge University Press; Cza - Czarne; Ins - Insignis; Jag - Jaguar; LN - Lexis Nexis; Pro - Prószyński; SD - Sonia Draga; WKP - Wolters Kluwer Polska; WKiL - Wydawnictwa Komunikacji i Łączności.

To infer if the book titles submitted by the publishers for the experiment differed substantially from their overall catalog at the time, we used a web scraping tool, collecting information on the books released up to 2013 by each of the publishers in our sample. This was only performed a few years after the experiment, so not all of the book titles participating in the experiment were still available in the catalogs. Moreover, two of the participating publishers merged in the meantime, which made it a challenge to identify their separate catalogs. To make the data complete, we added manually the book titles committed to the experiment if they were missing from the current catalog. Overall, while 239 book titles participated in the experiment, we identify the contemporaneous catalog of 1156 titles. Recall that until the end of the experiment, the publishers did not know which books were in the control group and which were in the treated group.

Having gathered this additional data, we were able to determine statistically if publishers selected books purposefully. To this end, we ran a logistic regression of book characteristics on the likelihood

that a given title was selected by the publisher to participate in the experiment. We report the results in Table A3.

Table A3: Odds ratio from a logit regression on selection for the study

Odds ratios	All publishing houses (1)	Without professional books (2)
Format (square cm)	1.003	1.003
	(0.003)	(0.002)
Hardcover	1.415	1.524
	(0.81)	(0.98)
Price per page	21.91	13.25
	(37.19)	(15.58
Page count	2.38**	1.88*
	(0.76)	(0.43)
Year of first print	1.26	1.22
	(0.22)	(0.20)
E-book exists	1.58	1.94
	(1.30)	(1.50)
Audiobook exists	1.95	2.36
	(0.90)	(1.13)
No of previous works by author	0.995	0.999
	(0.02)	(0.03)
Observations	1,269	1,216
Pseudo R^2	0.13	0.11

Notes: Robust std. errors clustered at publisher level in parentheses. Cambridge University Press dropped from the sample as the catalogue presented problems for coding (books had multiple versions difficult to automatically process, e.g. "Teacher's book", "Student's book", "Student's book with key", "Teacher's book with DVD", etc.).

The publishers indeed followed the principle of choosing newer titles. However, it does not seem that "piracy" threat was a consideration, as titles with official e-book versions were not more likely to be chosen for the study (as discussed in the literature review, e-book versions – at least of older titles – were previously shown to be more likely to be affected by piracy - Reimers (2016)). The publishers were, by contrast, more likely to suggest titles with an audiobook version available (in total 18 book titles out of 239 in the experiment were distributed also as audiobooks). Similarly, more expensive book titles were more often included. However, the publishers did not seem to bet on specific authors or cost of production.

In terms of market representation of the titles in the sample, we note some divergence. Most of the sales in our sample (averaged per book across all months) occurred within the Fantasy and Science Fiction sample (30.4%), followed by non-fiction (biographies, memoirs, essays, etc. - 29.7%) and general fiction (12.2%). The book sales reports for 2013 instead reveal that academic, professional and technical books comprised 36.5% of the market (11.9% in our sample), followed by school books with 31.5% (1.7% comprising English textbooks for various levels in our sample) and general fiction with 13.2% (12.2% in our sample).

We argue, however, that accurate representation in terms of segment shares might not be desirable, and some discrepancies can be controlled for. For the former, while school books constitute a large share of the market, they may be inadequate for the studies of "piracy" effects as the segment targets a very specific group (non-working school youth), represents books that are often mandated by schools (constraining choices), are often purchased secondhand (from older school youth), and

often have to be in print format for classes. For the latter, this is because the sales shares in our sample are largely driven by heterogeneous release dates of the analysed titles. In the non-professional segments it is typical that much of the sales occurs directly after the release. In our sample, two books in the fantasy/sci-fi and two in the non-fiction segments were released within a month since the experiment start. As such, we deal with these discrepancies by controlling for, e.g., release dates.

C Matched-subject randomization procedure

Matched-subject randomization requires data of high quality already 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 most objective characteristics. First, we included sales forecasts as reported by the publishers. A possible bias in these forecasts is not a problem, non-trivial *correlation* with sales allows to match books that will sell similarly. 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). We used available quarterly data as bounds and used within-segment seasonal variation from the available monthly data in the interpolation, to account for seasonal effects. Specifically, denoting the monthly reference series by $I_1, \ldots, I_{12}$ and the quarterly series to be distributed by $Q_1, \ldots, Q_4$, the (proportional) Denton method seeks to minimize the sum of squared differences between subsequent ratios of the resulting monthly series $X_1, \ldots, X_{12}$ and the reference series:

$$\min_{X_1,\ldots X_{12}} \sum_{t=2}^{t=12} \left(\frac{X_t}{I_t} - \frac{X_{t-1}}{I_{t-1}} \right)^2$$

s.t.

$$X_1 + X_2 + X_3 + X_4 = Q_1, \ldots, X_9 + X_{10} + X_{11} + X_{12} = Q_4.$$

Second, we included data on the basic book characteristics, such as the type of the book, the 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 A4 reports in detail the variables available prior to the experiment and used for matching.

Table A4: Book characteristics

Variable	Median	Mean	Std. Dev.	Matching
Publication date (for the current edition)	27.04.2012	-	580 days	Yes
Which edition[a]	1	2	3	Yes
Previous edition publication date (if applies)	28.04.2010	-	742 days	No
E-book available		23%		No
E-book release date (if applies)	25.09.2012	-	75 days	No
Page count[b]	352	415	304	Yes
Versions available (hardcover, e-book, etc.)	-	-	-	Yes
Price	39.99 PLN	50 PLN	21 PLN	No
Price per page	0.14 PLN	0.17 PLN	0.10 PLN	Yes
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[c]	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

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

[b] The average length of the book in the 2013 catalog was 263 pages with a standard deviation of 230 pages.

[c] 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.

For the matching procedure, the Mahalanobis distances were computed between each two observations within the data set. 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 flows, written for R by Mark Fredrickson and Ben Hansen. For additional information, see Hansen and Klopfer (2006) or the *optmatch* package for R-CRAN.

As a result, 125 matched pairs were created, 22 groups of three and one group of five. 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 randomized manner, so 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.

Table A5: Publishers and titles by treatment groups

publisher	treated	control	dropped
Buchmann	0	0	51
Cambridge University Press	9	9	0
Czarne	27	26	0
Galaktyka	0	0	5
Insignis	3	3	0
Jaguar	6	7	0
Lexis Nexis	21	18	6
Proszynski	5	5	0
PWN	0	0	20
Sonia Draga	11	13	1
WAB	0	0	0
Wolters Kluwer Polska	24	24	0
WKiL	13	15	0
Total	119	120	83

By the end of the experiment period, some publishers and titles had to be dropped from the sample (see Table A5). One of the titles from Sonia Draga was an outlier in terms of sales and was therefore dropped from the matching procedure and further analyses. WAB did not provide a title list for the study despite an initially signed agreement. PWN was dropped due to detected own anti-piracy activities. Buchmann and Galaktyka did not provide sales data at the end of the study. Six titles of Lexis Nexis that were supposed to premiere during the course of the experiment were dropped from the publishing schedule. To ensure these changes did not affect our initial randomisation, we have compared all our variables used for matching by the treatment group, using t tests and Wilcoxon Rank Sum tests. We found no support for dismissing the null hypothesis of no difference between groups for any of the variables at any conventional significance level.

D Reporting of the treatment execution

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, during the duration of the experiment – i.e. between the 24th of October 2012 and the 23rd of November 2013 – Plagiat.pl issued in total 11,952 notices to take down unauthorized files of 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 the actual date identification (the reports only contained the concluding date of the period of searching). The 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 were committed in this group. We matched the data from the second report to those from the first to infer 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).

Table A6: 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

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 size files before sending the take-down notice. 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 take-down notice and 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.

Table A7: Effectiveness of reducing the availability unauthorized copies

Number of unauthorized copies	Coefficient	Marginal effect
Treatment	-6.09	-4.04
	(2.65)	[-7.08 -0.57]
Month of experiment	2.35	1.56
	(0.13)	[1.31 1.64]
Month of experiment * treatment	-3.20	-2.12
	(0.18)	[-2.27 -1.75]
No. of unauthorized copies identified prior to the experiment	0.045	0.03
	(0.004)	[0.02 0.03]
E-book exists	-4.90	-3.25
	(2.42)	[-6.05 -0.10]
Constant	6.81	6.81
	(4.77)	[-2.54 16.17]
Number of observations	2514	
Number of titles	228	

Notes: Panel tobit regressions on the monthly Plagiat.pl data about the number of copies in unauthorized online distribution. Segments and publisher dummies included, not reported, available upon request. The average marginal effect is calculated at month=0 (first month of the reports), therefore describing the situation prior to the treatment. Standard errors in round parentheses, confidence intervals in square parentheses. The sample was reduced to the 228 titiles included in regressions from Table 2.

E The Polish book market

Our experiment is implemented in a highly suitable environment. Generally speaking, the Polish book market is a fairly typical mid-sized European market, which raises hopes about potential external validity of our findings. This market is about as large as those of Belgium or The Netherlands (Simon and De Prato, 2012; International Publishers Association, 2015), with relatively high market concentration, a bulk of sales attributable to a handful of authors, a significant share of translated work. Appendices A and B provide more information on how the publishers and titles in our sample related to the whole book market.

According to yearly reports published by the National Library of Poland, about 40% of adult population reads books (at least one book a year). Combined with a relatively high number of new titles per one million inhabitants – about 750 – mean print run has decreased from more than 15 thousand in the early nineties to just about three thousand copies at the time of the experiment (Biblioteka Analiz, 2013a), considerably lowering publishers' profit margins. Under the circumstances, the case of Poland constituted a useful playground for our field experiment: linguistically constrained readership, growing use of digital books and a medium size market allowed our experimental design to be relevant, targeted and effective. Considering the setting, we believe our design could be replicated across other countries – particularly ones where English is not the first language.

While 10 years have passed since the conduction of the experiment, as of this writing, we believe the book market did not undergo major shifts in that period (IKP, 2015; IKP, 2020). The share of large publishers decreased by 4.9 pp. to 70.1%, with medium-sized publishers increasing by 3.9 pp. to 26.4%, and small-sized to 3%. Book industry income has been further declining (from 640 million EUR in 2013 to 532 million in 2019[1]), but the segment composition remained largely the same, with

[1]We're deliberately comparing the numbers with 2019 as the first two years of the COVID-19 pandemic might have

a slight increase of the share of children's books (from 6% in 2013 to 16% in 2019) and a decline in academic/professional books (from 36.5% to 28.3%). The other segments recorded changes lower than 6 pp. Notably though, 2019 was a slightly odd year with fiction receiving a boost from the Nobel Prize of Olga Tokarczuk, and a global hit TV series based on the Polish fantasy series The Witcher. The distribution market also did not change much, with most of the change between 2014 and 2019 attributable to a decline of the bookshop share by 11 pp. to 24% and the growth of the internet share by 9 pp. to 44%. Contrarily to expectations, the number of bookshops stayed mostly the same, with 1,974 bookshops in 2013 and 1,914 in 2019 (having grown since the low point of 1,820 in 2017). Finally, the share of e-books and audiobooks in the total market experienced a modest growth from 3% in 2013 to 11% in 2019. This is despite a reduction in the tax for e-books from 23% to 5% in 2019, which was largely absorbed by the publishers (IKP, 2020). Meanwhile, the main file-hosting website - Chomikuj.pl - remains active.

severely affected the market, and data for 2022 is not yet available.

F Sales data

The publishers do not have actual sales data in Poland. Few of the biggest publishing houses purchase bar scanner data, but even this source of data is incomplete, as many bookstores operate without bar scanners. To collect the sales data, the publishers utilize the netting of the relationship with retail bookstores and intermediaries. It is customary in this industry that, after publishing a book, a fairly large number of copies are shipped to the intermediaries and to the bookstores. Both, the intermediaries and the bookstores, keep the stock of books for the period they deem 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 as well as the books returned by them – not on contemporaneous bookstore sales. The clearing of the transactions typically occurs at a quarterly or semi-annual basis. We aggregate this data to annual sales per title.

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 A8, we solve this problem by adding the print run to the aggregate sales data reported by the publishers.

We have also asked the publishers about e-book sales of their titles. However, the sale numbers proved very low with an average of 94 electronic copies sold throughout the whole experiment period, and the e-books were sold only for app. 32% of the titles in our sample. This is perhaps not surprising, given that the value of the e-book market in Poland comprised approximately 3% of the whole book market in 2013. As such, we have not followed on the e-book numbers further and did not include them in our analysis.

Table A8: Sales: descriptive statistics

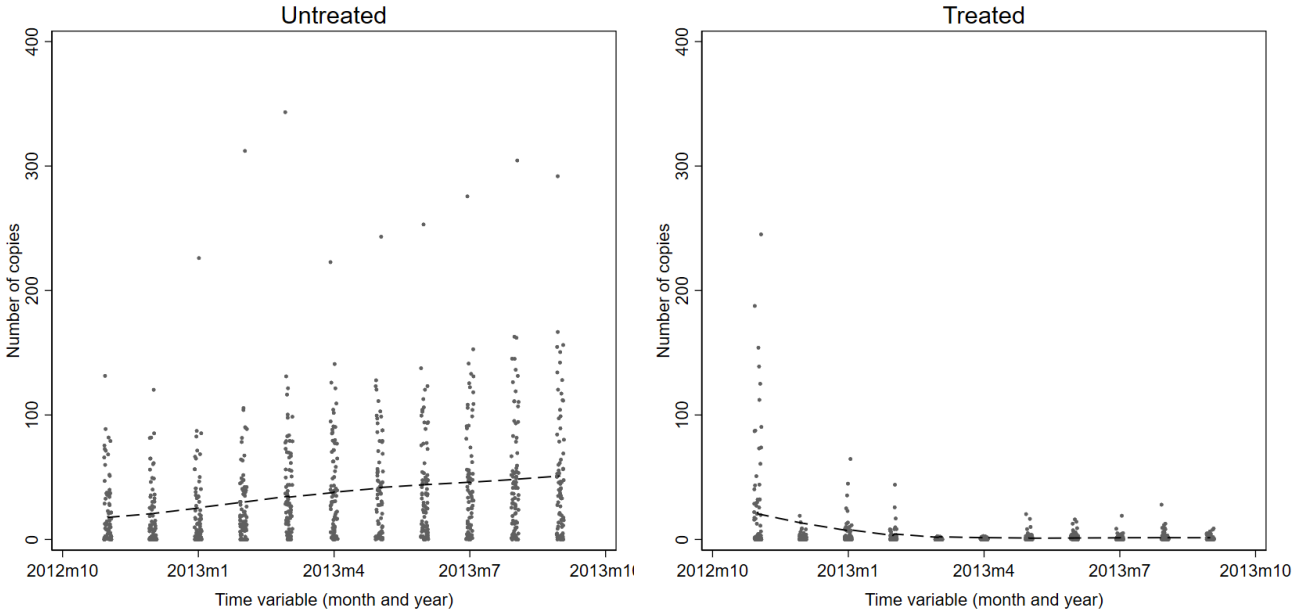
	No of titles	Mean	Std. dev.	Median	Min	Max
	Original sales data					
Total	239	1 358	3 423	309	-5 893	34 577
CT	120	1 244	2 895	256	-5 893	18 534
ET	119	1 473	3 892	343	-5 037	34 577
	Sales data corrected with first print run					
Total	228	4 103	8 052	1 152	4	69 577
CT	115	4 008	7 220	1 112	4	40 534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

Notes: 11 book titles had missing first run data and negative annual 'sales', which effectively reduced the sample to 228 titles.

G Treatment effectiveness

Manipulation check 1. Data from Plagiat.pl Figure A1 shows the average number of copies (per book) available monthly, in the control and treatment groups respectively. The significant difference between the two groups is plainly visible. With the initial level of sharing being identical by virtue of the random treatment assignment, the number of shared books grows steadily over time in the control treatment, while it declines sharply for treated titles (even if they sometimes resurface for short periods). We estimate formally the decline in availability of unauthorized copies and report these estimates in Table A7. Across the twelve months of the experiment, at least three times *less* copies were available in the treated group (relative to the first measurement, before the experiment), whereas two times *more* copies were available in the control group.

Figure A1: Number of unauthorised copies of books over time by treatment group



Note: Each dot represents the number of unauthorised copies available of a given book in a given month (average from all days in the month). The lines represent locally weighted regression lines.

Manipulation check 2. RA's In a second manipulation check, three research assistants, who reported being familiar with acquiring uploaded files, searched for a specified set of books on the Internet.[2] The lists comprised twenty randomly chosen pairs of titles from our initial sample. The assistants did not know about the experiment design, nor treatment assignment. Two of the assistants (B and C), received the same list of titles, so that we would have some sense of individual differences between users, while Assistant A received a different one so that more books could be covered. Their task was to search for each title for up to five minutes. The assistants could use services that requested a payment for transfer with up to 5GB of transfer; apart from that, they were allowed to spend up to PLN2 per title found on other platforms in case downloading required paying a small fee.

[2]Note that under Polish legislation, downloading is legal, but uploading violates copyright.

If they found the book during that time, or reached 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 successfully acquire the listed book. The results of this manipulation check are reported in Table A9.

This manipulation check reveals that treated books were substantially more difficult to acquire: about 57% of titles found in CT, versus only 32% in ET, confirmed by a formal test of equality of proportions; $z = 2.78$. The lower availability of the treated group was also manifested in somewhat longer search times, but this difference has not proven statistically significant (on average about 63 seconds for CT and 94 seconds for ET, $z = -1.13$). The two assistants with the same list (assistants B and C) found almost the same set of titles. Assistant B found three titles that Assistant C did not and Assistant C found four titles that Assistant B did not. They both found the same set of 16 titles, although assistant B reported longer search times for the found copies (on average twice as long, i.e. around 113 seconds per copy found). Except for this speed difference, the outcome of the searches seems consistent and both of the assistants mostly found books from the CT group. The source of the majority of successfully downloaded files was one of the most popular file-hosting platforms (see Table A10), which reinforces the findings of the first manipulation check.

Table A9: 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%
CT: avg. time until found	60 sec.	108 sec.	27.5 sec.	63 sec.
Enforcement Treatment	4	7	8	6.33
Enforcement Treatment (%)	20%	35%	40%	31.67%
ET: avg. time until found	60 sec.	120 sec.	85 sec.	94 sec.

Notes: All of the title lists included 20 pairs of books (10 in the CT and 10 in the ET group). Assistants B and C had the same list of titles. Note that under Polish legislation, download is legal, it is uploading that violates copyright.

The research assistants were requested to spend no more than five minutes looking for any source. 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. For the treated titles, the assistants often had to consult specialized sources, which could be discouraging for downloaders who are not familiar with them. For example, to download a file through the P2P network, one typically needs to search a website that hosts *torrent* or similar files files (or magnet links) and to have appropriate software installed for the download of the actual content. Since Google policy change in 2013, a general search in this engine would only return *torrent* files if one uses this keyword in the search. Although the marginal cost of both actions is negligible, one has to actually possess this knowledge prior to successfully downloading any content distributed without authorization. Meanwhile, with the file-hosting platforms, no prior actions are needed and no keywords are necessary.

We carried out the following exercise to inspect the importance of P2P networks as compared to file-hosting platforms. We took a random sample of 50 titles represented in our study and then we asked a research assistant experienced in finding unauthorized versions of various media and unaware of the hypotheses, design, or findings of the main study, to find the complete text of each of these books on 1) P2P networks and 2) one the most popular file-hosting platforms. The assistant in

this experiment was not one of the three assistants involved in the second manipulation check. The assistant was asked to spend up to 5 minutes for each of the 100 searches and to spend no more than 2 PLN (.50 EUR) on any of them. In practice, these constraints made very little difference. The research assistant was able to find 14 books (28%) on one the most popular file-hosting platforms and just one (4%) on peer-to-peer networks. We take it as additional evidence of relative importance of file-hosting platforms as compared to P2P networks in the market for unauthorized versions of Polish-language books.[3]

In our attempts to verify the effectiveness of treatment in curbing "piracy", the student RAs were indeed able to find some of the books in unauthorized distribution. However, the source of the majority of successfully downloaded files was one of the most popular file-hosting platforms, see Table A10. This particular file-hosting platform was very responsive to TDNs in general and removed the content literally within hours of receiving the notice. The fact that our student RAs were able to find content there is merely a consequence of the fact that the manipulation check was performed *after* the end of the twelve month experiment. Note, that the student RAs recorded the *first* place where they were able to find the book (as most users would) and did not invest further time in finding additional copies. Plausibly, some of the books found on Chomikuj.pl could be available on other websites as well, but accessing content in those sources is more time consuming and requires greater expertise. It can also be riskier in terms of malicious or misleading content.

Table A10: Manipulation check – sources where titles were found

Source	Control Treatment	Experimental Treatment	Total	Type
5fantastic.pl	2	2	4	file-hosting
chomikuj.pl	28	6	34	file-hosting
download.freebiz.pl	2	0	2	file-hosting
forumwpia.org	0	1	1	file-hosting
freedisc.pl	0	3	3	file-hosting
sendspace.com	0	2	2	file-hosting
share.pdfonline.com	0	1	1	file-hosting
torrenty.org	0	1	1	P2P
ulozto.net	0	1	1	file-hosting
uploaded.net	0	1	1	file-hosting
vgi***.com	1	0	1	personal website
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. The source represents the first location where the file was found.

Correlation between number of copies and downloads at Chomikuj.pl It would be desirable to look at the number of actual downloads, rather than just the number of copies available. Unfortunately, this data is not available. However, these values are highly correlated, because an almost constant share of the downloaders leave a copy in their account. We verify empirically this claim using a separate data set obtained from the operator of the file hosting platform which has proven to be the most popular in the search queries implemented by Plagiat.pl – Chomikuj.pl.

[3]This test was performed about 15 months after concluding the experiment, in response to comments collected during seminars and conferences, hence the effectiveness of our treatment enforcement cannot be meaningfully inferred from this exercise.

Chomikuj.pl agreed to provide data with the number of downloads and the number of copies, per file, within 31 days since it was first uploaded to their platform. The data is narrowed to relevant file types (pdf, epub, mobi, doc, txt, rtf), covering all the files that were uploaded within this specified time period (Chomikuj.pl agreed to provide us with data from 2017, but did not disclose the exact dates). We were only advised to collect this data after the end of our experiment, it was impossible to acquire data contemporaneous to the time of our experiment, but there is no indication that these general patterns in the data have changed.[4]

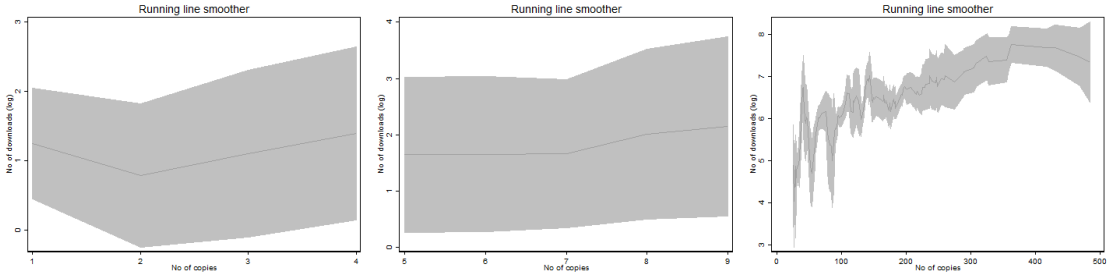
The correlation between the number of copies and downloads (logs thereof) is high and statistically significant (app. .6 with a p-value of 0.00). In Figure A2 we show a lowess approximation of the correlation, split by intervals for the number of copies (we also add the logically necessary zero downloads for zero copies to create the figures). Interestingly, with few copies (left and central panel) the correlation is virtually zero. It is for higher number of copies (right panel) that the correlation increases, to become flat again for very high number of copies in our sample.

These results demonstrate as well that there is no big jump at the first copy. Quite the contrary, for a low number of copies on average the number of downloads is low: files that have few copies tend to have only few downloads. Although there appears to be some heterogeneity, the upper bound of the 95% confidence interval plotted in figures below reveals literally less than 10 downloads for unpopular items. It is for the items with a (relatively) large number of copies that the correlation with the number of downloads increases. Hence, we infer cautiously that even if not literally all copies were effectively removed, the probable number of downloads was substantially reduced.

Ideally, we would acquire data on downloads from Chomikuj.pl at the time of the experiment. However, at that point, there was a growing conflict between the Polish Book Chamber and the file-hosting service, making negotiations problematic. In 2012, a group of publishers decided to sue Chomikuj.pl for pirate practices. Chomikuj.pl denied involvement, citing they've been consistently removing any infringing files when asked to, claiming they have removed as many as 13.5 million files in just the two prior years. In turn, they counter-sued, alleging the publishers had no rights to call their service "piracy". In 2014, it was ruled that calling it a 'pirate service' was not a violation. It took 11 years for the larger suit to reach the conclusion that Chomikuj.pl did indeed break the law by not doing enough to prevent piracy at its website. However, since the ownership of the service changed in that time, not much has resulted from it. Still, some time after the first suit was resolved, Chomikuj.pl agreed to provide us with the data mentioned here.

[4]This data refers to file formats associated with books (*.pdf, *.doc, *.epub, etc.), hence our conclusions should not be extended to music files or films and tv series.

Figure A2: Nonparametric estimation of the correlation between number of copies available in unauthorized distribution and number of downloads, split by the number of copies.



H Descriptive statistics

Table A11: Sales and copy availability data summary

	Total sample		Ever shared		Circulated prior to the experiment			
					All		Ever shared	
	CT	ET	CT	ET	CT	ET	CT	ET
N**	120	119	82	72	84	78	61	53
Average sales	1244	1473	1312	1979	1278	1445	1406	1588
Average sales*	3769	3929	4191	5498	3474	3950	3923	4768
Median sales	256	343	495	550	317	331	450	343
Median sales*	1052	1124	1277	1238	1104	1146	1104	1203
P5 sales	-119	-166	-90	-28	-27	-165	-26	-42
P5 sales*	10	0	0	198	16	54	16	191
P95 sales	6088	6080	5820	8628	4854	4259	4854	4259
P95 sales*	21317	21080	21957	26379	18385	23195	24058	26379
During the experiment								
Mean no. of months a copy was found	6.2	2.4	9.1	3.9	6.6	2.4	9.2	3.5
Median no. of months a copy was found	8	1	10	3	9	1	11	3
Average no. of copies identified	267	27	391	44	314	24	437	35
Median no. of copies identified	91.5	2	311	19.5	121	3	325	16
Min. no. of copies identified	0	0	2	1	0	0	2	1
Max. no. of copies identified	2724	436	2724	436	2724	436	2724	436
P5 no. of copies identified	0	0	7	1	0	0	7	1
P95 no. of copies identified	931	148	982	176	982	120	1141	126
Prior to the experiment								
Average no. of copies shared	101	90	146	145	144	137	196	196
Median no. of copies shared	3	3	8	17	8	15.5	43	45
Min. no. of copies shared	0	0	0	0	1	1	1	1
Max. no. of copies shared	3490	1444	3490	1444	3490	1444	3490	1444
P5 no. of copies shared	0	0	0	0	1	1	2	1
P95 no. of copies shared	576	821	583	1056	583	1056	593	1073

Notes: * Sales corrected with the initial print run, see footnote 7.
** 11 titles were published during the experiment. For 57 titles we have data on pre-treatment sales data (28 CT and 29 ET). Copies identified is equivalent to the number of copies removed in the case of the ET.

References

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

Biblioteka Analiz, 2013a. Badanie rynku ksiazki w Polsce. Biblioteka Analiz.

Biblioteka Analiz, 2013b. Polski rynek ksiazki 2013. URL: https://www.instytutksiazki.pl/rynek-ksiazki,7,raporty,18,polski-rynek-ksiazki-20

Biblioteka Analiz, 2014. Polski rynek ksiazki 2014. URL: https://www.instytutksiazki.pl/rynek-ksiazki,7,raporty,18,polski-rynek-ksiazki-20

Biblioteka Analiz, 2016. Polski rynek ksiazki 2016. URL: https://rynek-ksiazki.pl/wp-content/uploads/2017/08/RPK-2016.T-1_342_elektroniczny

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

International Publishers Association, 2015. Annual report.

Jaguar, 2018. Jaguar annual financial statement. URL: https://aleo.com/pl/firma/w003-sp-z-oo-warszawa/dane-finansowe.

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

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

Rubin, D.B., 1980. Bias reduction using Mahalanobis-metric matching. Biometrics 36.

Rynek-Ksiazki.pl, 2012. Lexisnexis polska 2011 - podsumowanie roku. URL: https://rynek-ksiazki.pl/aktualnosci/lexisnexis-polska-2011/.

Simon, J.P., De Prato, G., 2012. Statistical, ecosystems and competitiveness analysis of the media and content industries.

wirtualnywydawca.pl, 2015. Ranking wydawnictw 2014. URL: https://wirtualnywydawca.pl/2015/04/ranking-proba/.

Zhao, Z., 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. Review of Economics and Statistics 86.