Does digitization hurt book readership and sales? Evidence from the Google books project

Abhishek Nagaraj
UC Berkeley-Haas
nagaraj@berkeley.edu

Imke Reimers
Northeastern University
i.reimers@northeastern.edu

June 26, 2018

Preliminary work in progress

Abstract

Despite the promise of digitization to deliver a centralized, digital repository of all books ever published, copyright challenges have prevented the realization of this vision. Copyright holders are concerned that digitization might cannibalize readership and sales for printed material, although this claim lacks empirical evidence. We shed light on this topic using newly-collected data on the digitization of almost 90,000 books from Harvard's Widener Library by the Google Books project between 2005 and 2009. We exploit the quasi-random timing of the digitization process across books to estimate the causal impact of digitization on readership and sales. We find that while digitization does reduce local use within Harvard by 38%, contrary to some predictions, it increases sales of digitized titles by about 36%. This overall sales effect can be explained by increased discovery for less popular titles; sales for a small set of popular titles decrease following digitization, but sales for less popular titles increase considerably. Overall, our evidence suggests that rather than losing revenue from digital distribution, copyright holders might benefit from digitization through increased discovery for less popular works.

1 Introduction

Digitization and the advent of the internet have dramatically transformed the creation and distribution of information goods such as books, movies and music (Greenstein et al., 2013; Waldfogel, 2017). Not only has digitization facilitated the creation of new products, but it has also significantly expanded access to the catalog of existing works. Much like a modern-day Library of Alexandria, there is the real possibility that the internet could serve as a repository of all knowledge in digital form (Samuelson, 2011). This idea is not just a pipe dream. Efforts led by for-profit organizations such as the Google Books project as well as non-profit groups like the Hathi Trust and the internet archive, have spent tens of millions of dollars digitizing the world's books and by last count, over 25 million books had already been digitized through these efforts (Somers, 2017).

Despite the technological progress and the financial investment, the vision of a "digital Library of Alexandria" has not come to life. There exists no single, digital repository where the sum of human knowledge can be accessed freely and at low cost. This result can be largely attributed to legal considerations, especially copyright challenges from traditional publishers and authors.[1] This challenge has been led by authors and traditional publishers, who are significantly concerned about the possibility that digitized versions would serve as substitutes for already-published material, thus hurting an industry that made over $40 billion in revenue in 2008.[2] While this concern has been salient in legal and policy discussions, it is backed by little empirical evidence. Therefore, while the broader issue of the role of copyright in allowing (or preventing) mass digitization projects like Google Books to continue involves a number of different considerations, an important part of the argument – and of the policy decisions – hinges on quantifying the impact of digitization on the use and sales of works distributed through traditional channels.

Quantifying the relationship between digitization and the diffusion of works through traditional channels is challenging. Theories can be marshalled to support claims for both a negative and a positive impact of digitization on printed works. On one hand, offering an already available book

[1] Despite court decisions in favor of Google's digitization project, the company "all but shut down its scanning operation. See Somers (2017).

[2] See Michael Healy, Book Industry Study Group, Books and e-Books: Some Industry Numbers, at the D is for Digitize Conference at the NY Law School 2009, http://www.nyls.edu/innovation-center-for-law-and-technology/iilp-archive/iilp-conferences/d_is_for_digitize/.

for free in digital format can displace demand through other channels, even if the free version is of lower overall quality. On the other hand, many works have become obscure over time, and an easily accessible free version can increase awareness and discovery of the work, thus increasing demand.[3] In addition, many titles have become entirely unavailable as their copyright holder cannot be found. These "orphan works" are in essence lost to consumers because nobody can obtain the license to publish them. Digitization projects such as Google Books might bring such books back to life, and increase the possibility that they will be republished and distributed in print.

Given that both sides of the argument are likely to operate in parallel, the question of whether the digitization through Google Books hurts works published through traditional channels is an empirical one. However, credible empirical evidence on this topic is hard to obtain due to a few inter-related challenges. First, most collections are either available in digitized form or not, and it is difficult to obtain meaningful variation in the digitization status of otherwise similar works. Second, even when the digitization status varies between works, this variation is driven by a careful selection process. In other words, digitized works are usually more popular or useful than those not digitized, making it difficult to compare across these two groups. And finally, even when relatively comparable works with varying levels of digitization can be identified, it is difficult to measure how the digitization of works online might affect their physical counterparts offline.

In this paper, we highlight a natural experiment that provides an ideal setting to tackle all three of these empirical challenges and provide causal estimates of the impact of digitization on the use and sales of physical books. Specifically, we focus on the Google Books Digitization project which was launched in 2004 with a vision to digitize all works ever created. At its inception, the project was launched in partnership with a handful of selected libraries. Among Google's first library partners was Harvard University's Widener library, which provided public domain books and texts to be digitized by Google, a process that started in 2005 and continued till at least 2009. We were able to obtain proprietary data on this four-year long digitization process for almost 90,000 books, including the dates on which every individual book was digitized. These data help us identify variation in digitization status within a large and important collection. Further, while not explicitly random,

[3] In line with both arguments, a 2012 survey of users of a Norwegian digitization effort found that 20% of respondents purchased a book after first finding it on the depository, whereas 18% reported that they did not. See Jøsevold (2016).

the timing of the digitization of books was not selective and did not prioritize the most interesting books for early digitization. Rather, the digitization process proceeded on a "shelf by shelf" basis and was driven by convenience.

The quasi-random nature of the timing of the digitization process helps us estimate the causal impact of digitization on printed works. We collect data on three sets of outcomes for printed works: their local use, nationwide sales, and the variety in which the works are available in print. Specifically for about 90,000 titles with at least one library loan in our data, we are able to obtain data on their library loans within Harvard, the weekly sales of all related editions from the Nielsen BookScan database, and publications of new editions from the Bowker BooksInPrint database. Having performed this match, we compare our three outcomes for books which were made available online to titles which were not (yet) digitized in a difference-in-differences setting controlling for non-parametric book and calendar year fixed effects.

We find that the impact of digitization on use is negative when considering internal readership at Harvard, but is positive when considering sales. Specifically, digitization lowers the number of checkouts within Harvard by about 38%, but increases sales by 36%. These results support a theory where the added benefits from digitization are low within Harvard (where the library website and librarians already facilitated discovery), but are significant for general consumers (where such services were very limited before). This argument is supported by our finding that the positive effect of digitization on sales is largely driven by less popular books, for whom increased discovery through Google Books is most relevant. In contrast, sales for popular books – where discovery is less relevant – decrease following digitization. Further, we find that digitization leads to an uptick of new in-print editions through other publishers, likely making these books more easily available to consumers, but this "bringing books back to life" effect explains only a part of the increased overall sales due to digitization.

In sum, contrary to copyright holders' concern that Google Books digitization will have a large, negative and widespread effect on physical editions, we find that this substitution effect is limited to local, library use and for a small set of very popular titles. On average, however, the facilitated discovery and improved availability of content through all channels lead to an increase in sales.

While our results are largely based on the digitization of out-of-copyright works, they provide much-needed empirical evidence in the debate around the digitization and copyright protection for works whose copyright has not yet expired as well. We find an increase in sales of books when they are digitized and made available in their entirety. When a work is made available only in snippets, substitution to the Google Books version likely is even less of a concern.

This paper adds to a growing literature on the impact of copyright on the availability, price, and re-use of existing works. In the context of books, Heald (2007) documents a large effect of the 1998 Copyright Term Extension Act on the availability of books at Amazon, Li et al. (2018) find that an increase in copyright protection for works under the U.K. Copyright Act of 1814 substantially increased prices, and Reimers (2018) finds that such changes have significant welfare impacts in the publishing market. In addition, it appears that making works freely accessible significantly increases potential follow-on innovation. A copyright is narrow enough that an innovation would be covered by a new copyright. Yet, such an innovation may not happen without the ability to reuse the original work. Accordingly, Nagaraj (2018) shows that copyright hurts the reuse of information from the Baseball Digest, and Watson (2017) suggests that this fact may have large welfare impacts.

While this emerging literature has shown how lifting a copyright can be quite beneficial for general consumers (through reduced prices, increased access and follow-on innovation), we know less about how digitization might hurt or help publishers and copyright holders. Our results suggest that the negative effects of digitization on copyright holders might be more limited than previously thought (and even positive in many cases). This finding should help pave the way for future copyright legislation enabling digitization and lessen concerns on the side of publishers and authors about reduced sales through complementary channels.

More broadly, this paper adds to the literature on the impact of digitization and aggregation on traditional markets. A large literature addresses the impact of (illegal) file sharing – often through free copies – in the music and movie industries, as summarized by Smith and Telang (2012) and Peitz and Waelbroeck (2006). The impact of digitization in the book industry is less well understood, although a few papers examine the impact of copyright protection on legal book sales (Reimers, 2016; Hardy et al., 2014). Other papers have examined the impact of news aggregators

on competition in and consumption of news on the internet (e.g. Jeon and Nasr, 2016; Athey et al., 2017). To our knowledge, this is the first paper to study the impact of digitization and its accompanying information aggregation on the sales and use of works distributed through traditional channels.

Our paper proceeds as follows. In Section 2 we discuss the Google Books project and our data and research design, Section 3 describes our quantitative results and Section 4 concludes.

2 Data and Research Design

2.1 The Google Books Project: A Brief Background

The Google Books project (originally known as the Google Print Library Project)[4] was announced by Google in December, 2004. The aim of the project was to make offline information contained in printed works available and searchable online. At the project's inception, Google partnered with the libraries of Harvard University, Stanford University, the University of Michigan and the University of Oxford as well as the New York Public library to digitally scan books from their collections. Once these works were scanned, they were to be made digitally available on the Google Books website for the general public to read.

Soon after its launch, the Google Books project was met with staunch opposition from a group of authors and publishers, including the Authors Guild and the Association of American Publishers, who filed class action suits against Google for copyright violation.[5] The lawsuits were centered on the idea that Google's digitization effort caused material harm to the authors and the publishers of the printed works, violating their exclusive rights to profit from their works, and were illegal under the terms of copyright law.[6] Google Books' major defense was centered on the idea of fair use. The argument here was that Google Books' digitization efforts were a substantially transformative effort that increased discovery of printed works and "increase[d] the visibility of in and out of print books,

[4] https://googleblog.blogspot.com/2004/12/all-booked-up.html

[5] See Samuelson (2009), and https://googleblog.blogspot.com/2008/10/new-chapter-for-google-book-search.html.

[6] See https://tinyurl.com/y7hsxalw.

and generate[d] book sales."[7] The suits were eventually settled (publishers) or rejected (authors), but the process took quite long, lasting over a decade before an appeal by the Authors Guild was rejected in the Second Circuit.

2.2 Google Books and Harvard Libraries' Natural Experiment

While the copyright cases described above were quite contentious and involved a number of parties and issues, our focus in this paper is much narrower. We focus our study on digitization of works from Harvard's libraries through the Google Books project. Given the potential for significant legal challenges to Google's efforts, Harvard Libraries' participation in the Google Books project was limited to works which were already in the public domain and for whom the copyright was deemed to have expired. This included public domain works from Harvard's largest and most prestigious Widener Library that houses a total of over 3.5 million books in its collections. Specifically, works published in the United States before the year 1923, and those published internationally before 1909 were provided to Google for scanning. We focus on the real effects of the digitization of these works in order to shed light on the broader implications of the digitization of printed works on readership and sales.

The digitization proceeded as follows. The Google Books project had set up a scanning facility in the Greater Boston area to process the books from the Harvard libraries. For the purposes of the scanning effort, Google Books was assigned a special library patron code, and books would be "loaned" to Google under this special code to be taken to the scanning facility. Once the book had been scanned, it would be returned to the library and also made available on the Google Books website after a short delay.

Our natural experiment relies on the fact that the scale of Google's scanning project at Harvard implied that the total duration of the project was over five years (from 2005 to 2009), after which it was shut down. Further, the order in which books were scanned was driven by convenience, rather than an explicit selection mechanism. Specifically, the books were scanned on a shelf-by-shelf and wing-by-wing basis until all out-of-copyright books in the relevant sections were processed. It is

[7] http://googlepress.blogspot.com/2004/12/google-checks-out-library-books.html

this quasi-random variation in the timing of the scanning project that we exploit to estimate the impact of digitization on eventual readership and sales. Figure 1 shows the number of books that were digitized in each year between 2005 and 2009.

2.3 Data

We obtained proprietary data from the Harvard Libraries with an entire record of their holdings that were scanned, as well as all works published between 1923 and 1943 that were not scanned. We also obtained separate information on all library checkouts between 2003 and 2011. These data contain information on the specific patron code checking out the work (for example, faculty, student or visitor), including whether a book was being checked out by Google Books. This allows us to estimate the time when a book was digitized and when it was made available online.

In addition to internal data from Harvard libraries on book digitization and loans, we also collected data from two other sources. We obtained access to NPD (formerly Nielsen) BookScan, which provides weekly sales information for printed books. These data are collected through tracking book sales using scanner data from a large panel of retail booksellers including major bookstore chains, discount retailers such as Costco and major online retailers like Amazon. They claim to track about 85 percent of total retail sales.[8] Our data from Harvard do not contain global unique identifiers such as ISBN numbers, so we manually searched NPD BookScan for the book titles to find suitable matches, aggregating sales of all editions for each title. Given the tedious data collection process, we searched for sales data for all English-language books in the Harvard collection before 1943 with at least 5 loans between 2003 and 2011, for a total of about 5600 titles.

In addition to sales data, we also collected data on the number of editions of a work that is available in print from the Bowker Books-in-Print database. This database tracks all editions of a particular work that is available in print. We matched titles in the Harvard database to this database, and were able to find matches for almost 25,000 unique titles.

Combined, the Harvard libraries data on book digitization and loans, the Nielsen BookScan data on book sales and the Bowker Books in Print database on editions allow us to characterize the

[8] See (Sorensen and Rasmussen, 2004) and https://tinyurl.com/y94qpsqt, accessed June 26, 2018.

impact of the digitization on reuse and sales. We organize these data into a balanced panel at the book-year level between 2003 and 2011. These data contain loans information for about 88,000 books (those with at least one loan), including 50,263 books that were not digitized and 37,743 that were. We also have 2412 books with at least one sale, and 24,667 books with at least one edition in the Bowker Books-in-Print database. These data are summarized in Table 1. The average book has about 0.25 loans per year, sells about 1640 copies and has 1.08 editions, although the median value for all three of these outcomes is zero.

2.4 Preliminary Evidence

Our research design relies on the randomness of the timing of digitization, including whether a book is digitized at all. That is, we assume that – on average – digitized books would have followed a similar path as books that were not digitized, were it not for their digitization. We implicitly test this conjecture in Figure 2, which plots the average annual loans (left panel) and sales figures (right panel) for digitized and non-digitized books.[9] Before the digitization period, loans and sales moved in similar directions for digitized and non-digitized books. However, the trends changed after Google began to digitize works. Loans through the Harvard Libraries decreased for all works, but they fell more for works that were digitized. In contrast, sales of digitized works increased toward the end of the digitization period, compared to books that were not digitized.

While informative, the trends in Figure 2 do not account for the exact timing of digitization and the nature of each book. To identify the true effect of the digitization through Google Books, as well as the mechanism of this effect, we employ a more formal estimation strategy as explained in the next section.

3 Results

We follow the use of (and demand for) titles that were scanned and made available on Google Books, and we compare the evolution of these measures with that of titles which were not (yet)

[9] We control for book fixed effects by calendar year for illustrative purposes.

digitized in a difference-in-differences setting. Formally, we estimate

$$Y_{it} = \alpha \times PostScan_{it} + \gamma_i + \mu_t + \varepsilon_{it}, \tag{1}$$

where $PostScan_{it}$ is an indicator that is 1 if book i has been made available on Google Books before year t, and γ_i and μ_t are book and year fixed effects, respectively. The dependent variable, Y_{it}, denotes book- and year-specific measures of demand (loans and sales). To account for the discrete nature and low average values of the dependent variables, we assume that the error term ε_{it} follows a Poisson distribution, and we therefore estimate the model in a maximum likelihood estimation.

3.1 Loans and Sales

We first estimate the impact of digitization on demand through traditional channels: library check-outs (through Harvard's Widener library), and sales of physical copies. Table 2 displays the results from this specification without year fixed effects (columns 1 and 3) and with year fixed effects (columns 2 and 4). Columns 1 and 2 show that digitization through Google Books significantly decreases the number of loans through libraries. In column 1 (not including year fixed effects), we find that making the book available on Google Books decreases Harvard library loans by about 48 ($= e^{-0.668} - 1$) percent. The impact is slightly smaller – a decrease of 38 percent – when including year fixed effects, suggesting that loans decreased over time for all books.

Unlike loans, the number of *sales* does not seem negatively affected by digitization through Google Books. Rather, sales through traditional channels increased after digitization. When including both book and year fixed effects (column 4), we estimate an increase in sales of 0.36 ($= e^{0.309} - 1$) percent per year due to digitization. The coefficient is statistically significant at the 10% level.

3.2 Heterogeneous Effects

The above results suggest that Google Books did not displace sales and revenues from publishers – a concern central to the argument for shutting down Google's digitization project. The differences in impacts across channels may be due to differences in existing institutions for discovery. Libraries

– especially university libraries – often have institutions in place which facilitate the discovery of content, either electronically or through trained personnel. Consequently, Google Books predominantly displaces use in libraries. Traditional book stores do not have these institutions in place, and the digitization through Google Books may serve as a complement by taking on the role of a discovery tool.

We further investigate this mechanism by examining whether books of different popularities are impacted differently by the Google Books project. Did previously obscure titles receive attention from potential buyers? Or were sales of the better-known books displaced by the new consumption channel? We repeat the analyses from Table 2 with an additional interaction term of the Post-Scanned variable with an indicator that is 1 if the book was checked out at Harvard's libraries more than five times in the three years before digitization, i.e. popular.[10]

Table 3 shows the results from these specifications. There are significant differences in the impacts of digitization for both loans and sales, with more popular books being affected more negatively. As suggested in Table 2, both popular and less popular books see a significant decrease in loans after digitization, compared to books which were not digitized or digitized later. However, the impact is significantly larger for popular works, which experience a decrease of about 81 ($= e^{-0.419-1.243} - 1$) percent, compared to a decrease of 34 percent for less popular books, according to the point estimates in column 2.

The relative impacts on sales are similar, although no book types seem hurt by digitization. The least popular books experience a 42 percent increase in sales, whereas the estimated impact on sales of popular books is negative but not statistically significantly different from zero.

3.3 Editions and Mechanism

The above analyses suggest that the extent of facilitated discovery impacts the effects of digitization. Likewise, it is possible that making works available freely improves availability through other channels, allowing consumers to buy editions that previously did not exist. We examine this pos-

[10]This definition identifies 1232 books as popular, accounting for 1.4% of books in the loans data, and 15% of books in the sales data.

sibility here. We first estimate the impact of digitization through Google Books on the number of *new* editions that enter the market, and we then examine whether the changes in use and demand can be attributed to improved availability.

Table 4 shows the results of these specifications. The first two columns show the first stage: the impact of digitization on the number of newly available editions. We find that titles become available in many more editions after their digitization, with an average increase of 71% (see column 2). Many of these editions move the title back into print at all, thus potentially increasing sales from zero to a positive number.

Columns 3 and 4 of Table 4 estimate the impact of digitization on loans and sales, respectively, controlling for the number of newly introduced editions of the work. Estimates from both regressions show that the impact in the previous tables is not driven solely by changes in access. Use of the printed work in the library system decreases robustly after digitization, and some of that decrease may even be due to competition through additional editions in other channels. Sales of printed works continue to increase significantly after digitization, although a part of this increase can be attributed to the added access to editions.

Overall, our results do not support the publishers' concern that making books available for free through Google Books limits their ability to profit through other channels. Although use through the library system (which typically does not generate revenue for the publisher anyway) decreases, sales through traditional channels do not. Instead, at least for the least popular titles, digitization is able to spur previously untapped demand, thus providing additional revenue flows for the publishers.

4 Discussion

Some copyright holders fear that digital availability will cannibalize the use of printed works and cause financial harm. This concern has largely blocked projects that have tried to create a centralized and digital repository containing digital versions of all books ever published. In this paper, we provide empirical evidence on the relationship between digitization and the use and sales of physical works in the context of the digitization of books from Harvard's Widener Library by Google Books.

Specifically, we uncover three sets of findings. First, digitization lowers the local use of books through physical checkouts at Harvard. Second, digitization increases sales of books to general consumers and increases their availability in print. And finally, the positive effects of digitization on sales are largely driven by the increased discovery of less popular works, while more popular titles do not benefit from digitization in terms of sales.

Our results have important implications for ongoing legal and policy debates on the design of copyright law for the digital age. First, our evidence contradicts the popular notion that digitization necessarily harms copyright holders in terms of the use of printed works. While our data do support this conclusion when discovery of new works is less important (within the Harvard system and – to a lesser degree – for the sales of popular titles), in a large majority of cases, digitization increases sales of works. Combined with evidence from other studies in this literature that point to the positive benefits of digitization to consumers, our results suggest that existing copyright holders should be more supportive of the digitization of their catalog, especially for less popular and out-of-print works.

Second, while our evidence comes from the digitization of public domain books (published before 1923), it also speaks to debates about the digitization of newer, in-copyright works. Our evidence comes from providing the full text of public domain books in digital form, whereas for in-copyright works the debate is about providing "snippets" of relevant text. This difference means that our positive result on physical sales, where the discovery effect is muted due to the cannibalization effect from the provision of full text, could be even stronger for in-copyright works. Our results strongly suggest that copyright holders and policy-makers should encourage the digitization and discovery of less popular in-copyright works, at least through the provision of limited amounts of text. Finally our estimates help strengthen the value proposition of mass-digitization projects and help support their proponents such as Google Books, the Hathi Trust or the Internet Archive. While previous negotiations between these parties and copyright holders have tried to weigh the benefits to society as compared to the harm to copyright holders, our estimates suggest that this tradeoff might be relevant only when there is little potential for additional discovery through digitization, as in the case of extremely popular titles. For a large majority of works, mass digitization projects seem to create value both for copyright holders and customers.

While we advance the broader debate on the impact of digitization in the market for books, it is important to acknowledge the limitations of our study. First, we focus on the digitization of a sample of books from a single, albeit important, library's collections and our evidence is restricted to books in the public domain. It is possible that these effects could be different for a more general sample of contemporary books. Further, the overall welfare effect depends on how digitization changes the dynamic incentives of authors and publishers to produce and finance new work. Our estimates do not measure the elasticity of this important margin. Our work also suggests multiple avenues for future work. First, we provide the first set of estimates and a research framework to evaluate the impact of digitization of books on physical sales. Future work should evaluate the robustness of our estimates in different contexts. Further, our research design looked at the impact of digitization through providing the full text of books. Future work should look at the impact of providing "snippets" and evaluate the optimal length of such limited access, balancing the positive effects from discovery and low access costs with the harmful effects to publishers from the availability of full text.

In sum, while the debate on the mass digitization of published work is complicated and involves a number of different questions, our study clarifies the important role of digitization in enabling discovery and helping copyright holders increase the sales of physical editions of their works. Our evidence therefore provides an argument for increasing access to published knowledge for all through mass digitization.

References

Athey, S., M. M. Mobius, and J. Pál (2017). The impact of aggregators on internet news consumption.

Greenstein, S., J. Lerner, and S. Stern (2013). Digitization, innovation, and copyright: What is the agenda? *Strategic Organization 11*(1), 110–121.

Hardy, W., M. Krawczyk, and J. Tyrowicz (2014). Internet piracy and book sales: a field experiment.

Heald, P. J. (2007). Property rights and the efficient exploitation of copyrighted works: an empirical

analysis of public domain and copyrighted fiction best sellers. *UGA Legal Studies Research Paper* (07-003).

Jeon, D.-S. and N. Nasr (2016). News aggregators and competition among newspapers on the internet. *American Economic Journal: Microeconomics 8*(4), 91–114.

Jøsevold, R. (2016). A national library for the 21st century–knowledge and cultural heritage online. *Alexandria 26*(1), 5–14.

Li, X., M. MacGarvie, and P. Moser (2018). Dead poets' propertyhow does copyright influence price? *The RAND Journal of Economics 49*(1), 181–205.

Nagaraj, A. (2018). Does copyright affect reuse? evidence from the google books digitization project. *Management Science*.

Peitz, M. and P. Waelbroeck (2006). Piracy of digital products: A critical review of the theoretical literature. *Information Economics and Policy 18*(4), 449–476.

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

Reimers, I. (2018). Copyright and generic entry in book publishing.

Samuelson, P. (2009). Legally speaking: The dead souls of the google book search settlement. *Communications of the ACM 52*, 28.

Samuelson, P. (2011). The google book settlement as copyright reform. *Wis. L. Rev.*, 479.

Smith, M. D. and R. Telang (2012). Assessing the academic literature regarding the impact of media piracy on sales.

Somers, J. (2017). Torching the modern-day library of alexandria. *https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/*.

Sorensen, A. T. and S. J. Rasmussen (2004). Is any publicity good publicity? a note on the impact of book reviews. *NBER Working paper, Stanford University*.

Waldfogel, J. (2017). How digitization has created a golden age of music, movies, books, and television. *Journal of Economic Perspectives 31*(3), 195–214.

Watson, J. (2017). What is the value of re-use? complementarities in popular music.

5 Tables and Figures

Table 1. **Summary**

	N	Mean	SD	Median	Min	Max
Scanned	792054	0.43	.4949148	0	0	1
Year Scanned	339453	2006.98	1.194472	2007	2005	2009
Loans	792054	0.25	.894747	0	0	189
Sales	52038	785.95	8283.127	0	0	626610
Editions	262107	1.08	4.968459	0	0	542

Note: Observations for books identified by a Bib-Doc-Id. Scanned: 0/1 for books that have been scanned in the time period 2003 to 2011. Year Scanned is the year the book was scanned. Loans is the number of times a book has been loaned. Sales is the number of book copies sold. Editions is total number of different editions of a book available for sale.

Table 2. **Estimates for the Impact on Loans and Sales**

	(1) Loans	(2) Loans	(3) Sales	(4) Sales
Post-Scanned	-0.668***	-0.484***	0.211	0.309*
	(0.0120)	(0.0149)	(0.164)	(0.167)
Book FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes
N	792054	792054	21708	21708

Notes: This table presents estimates from a poisson model. Loans represents the total number of times a book has been loaned in a given year. Sales is the number of sold copies of that title in a year. Post-Scanned: 0 - for years before a book has been digitized and the year it is digitized, 1 - years after it is digitized. Book fixed effects are included in all models, year fixed effects in columns 2 and 4. Standard errors in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table 3. **Estimates for the Impact on Loans and Sales by Popularity**

	(1) Loans	(2) Loans	(3) Sales	(4) Sales
Post-Scanned	-0.603***	-0.419***	0.249	0.347**
	(0.0118)	(0.0149)	(0.172)	(0.175)
Post Scanned x Popular	-1.243***	-1.243***	-0.414**	-0.417**
	(0.0833)	(0.0833)	(0.174)	(0.174)
Book FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes
N	792054	792054	21708	21708

Notes: This table presents estimates from a poisson model. Loans represents the total number of times a book has been loaned in a given year. Sales is the number of sold copies of that title in a year. Post-Scanned: 0 - for years before a book has been digitized and the year it is digitized, 1 - years after it is digitized. Popular: 0 - for all books that were loaned at most 5 times before 2006, 1 - for all books that were loaned more than 5 times before 2006. Book fixed effects are included in all models, year fixed effects in columns 2 and 4. Standard errors in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

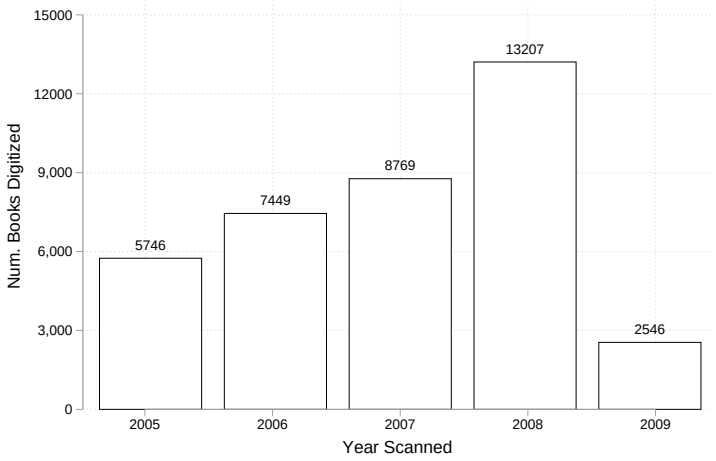
Table 4. **The Impact on New Editions, and their Impact on Loans and Sales**

	(1) Editions	(2) Editions	(3) Loans	(4) Sales
Post-Scanned	2.334***	0.538***	-0.388***	0.276**
	(0.0287)	(0.0371)	(0.0207)	(0.138)
Editions			-0.00409***	0.00847***
			(0.00151)	(0.00304)
Book FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes
N	222003	222003	262107	15408

Notes: This table presents estimates from a poisson model. Loans represents the total number of times a book has been loaned in a given year. Sales is the number of sold copies of that title in a year. Post-Scanned: 0 - for years before a book has been digitized and the year it is digitized, 1 - years after it is digitized. Popular: 0 - for all books that were loaned at most 5 times before 2006, 1 - for all books that were loaned more than 5 times before 2006. Book fixed effects are included in all models, year fixed effects in columns 2, 3, and 4. Standard errors in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Figure 1. **Timing of Book Digitization for Digitized Books**

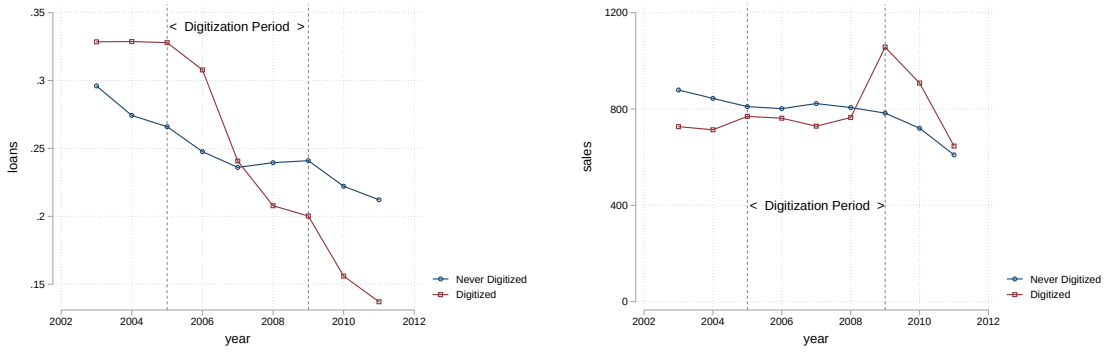
Panel A: Loans



Note: This histogram illustrates the distribution of the year in which books were digitized by the Google Books program within the Harvard library system. In total 37,743 books were digitized. 50,263 books in our data (published between 1923-43 with at least one loan) were not digitized and are not shown in this histogram.

Figure 2. **Trends in Loans and Sales by Year for Digitized and non-Digitized Books**

Panel A: Loans **Panel B: Sales**



Note: This figure illustrates annual loans and sales figures for digitized and non-digitized books. We calculate the (residualized) average level of loans (panel A) and sales (panel B) after controlling for book fixed effects by calendar year.