Nagaraj and Reimers (2021): Replication and Extension

David Scolari, John Bowman, Blake Lin

May 2022

Abstract

1 Introduction

Motivated by this omission, we examine whether it is possible for free digital distribution of books to increase rather than depress physical sales, especially when accompanied by a full-text search technology. We begin our analysis by developing a simple theoretical framework that incorporates the discovery mechanism when considering the role of free digital distribution in shaping demand for physical books. The framework clarifies that while the net effect of book digitization is ambiguous, sales could increase if the discovery channel can compensate for the cannibalization of physical sales via the digital channel. Investigating the effects of such search-enabled digital provision is important, given the large size of the publishing industry, and because discovery through digital distribution might be- come increasingly relevant in other settings as well. While the finding that free digital distribution tends to cannibalize (or not increase) physical sales seems relatively well established, less attention has been paid to the possibility that free digital distribution can also enhance search and discovery, thereby stimulating offline demand. Such an effect could be particularly salient in the market for books, where digital distribution can be accompanied by technologies that allow consumers to search through the full-text and discover (and buy) new content that would be otherwise hard to find (Ellison and Ellison, 2018).

The heart of our study empirically analyzes the effects of a prominent, search-enabled, free digital distribution program: the Google Books digitization project. Launched in 2005, Google Books is one of the landmark projects of the digital age, with commentators likening it to a "modern-day Library of Alexandria" (Somers, 2017). Google Books did not just scan a book's textual material but also made it searchable via optical character recognition (OCR) technology through its "Google Book Search" feature (referenced by Cambridge University Press in the epigraph). Further, a large portion of the Google Books corpus included less well-known and older books (including public domain content) that are of significant consumer interest but have become forgotten over time.

We tackle the empirical challenges through a unique natural experiment leveraging a research partnership with Harvard's Widener Library, which provided books to seed the Google Books program. The digitization effort at Harvard only included out of copyright works, which – unlike in-copyright works – were made available to consumers in their entirety. This allows us to fairly assess the tradeoff between cannibalization (by a close substitute) and discovery (through search

technology). Owing to the size of the collection, book digitization (and subsequent distribution) at Widener took over five years, providing signifi- cant variation in the timing of book digitization.

First, we collect data on the shelf- level location of books within the Harvard system between 2003 and 2011 along with information on their loan activity. Since most books are never loaned, our analyses focus on 88,006 books (out of over 500,000) that had at least one loan in the sample period (and are robust to using a smaller sample of books with at least one loan before the start of digitization). Second, for a subset of 9,204 books (in English with at least four total loans), we obtain weekly US sales data on all related physical editions from the NPD (formerly Nielsen) BookScan database. The sales data must be manually collected and matched, which restricts the size of this sample. Finally, we are interested in the effect of digital distribution on physical supply through the release of new editions. Accordingly, we also collect data from the Bowker Books-In-Print database on book editions and prices, differentiating between established publishers and independents. We use these combined data and the natural experiment we outlined to examine the effects of free digital distribution on the demand and supply of physical editions. Our panel data structure allows for a difference-in-differences design that can incorporate time and, notably, book fixed effects, increasing confidence in the research design.

Next, we examine the importance of the discovery and substitution channels by studying two paral- lel settings where one of these two effects is muted. First, we investigate the effects of digitization on loans within the Harvard system. In this setting, the discovery effect is muted since Harvard students and professors already had access to alternate discovery mechanisms through library services. In this setting, digitization reduces rather than increases demand, measured as loans within Harvard.

Answering our research questions is important, not only because it helps make theoretical progress on the literature on cross-channel substitution, but also because of its industry and policy relevance. The net revenue of the book publishing industry in the US in 2019 was more than thrice as large as the music rev- enues ($25.9B as compared to $7.3B; AAP 2020 and RIAA 2020, respectively). Further, the advent of the internet, platforms like Amazon.com and mass digitization projects like Google Books have presented numerous questions about balancing physical sales, digital distribution (via e-books) and mass digital distribution (via projects like Google Books) for firms in the publishing industry. In fact, legal debates have analyzed the specific case we focus on (the Google Books project) and debated whether it increases or de- creases sales (Samuelson, 2009). Our research provides quasi-experimental evidence to policy-makers and legal scholars, who have largely relied on anecdotal and theoretical data up to this point.

1.1 How Nagaraj and Reimers think digitization is working wrt demand (and supply) for books

How might digital distribution increase the sales of physical books? For end customers, the question of whether the digitization of books increases or reduces demand for physical works depends on two counter- acting forces. The first is the substitution effect of digital distribution as a competitor for existing, physical products as studied in the literature. Some consumers who would otherwise consume physical copies will switch to digital versions, driving the substitution effect. This is likely to happen when a consumer's search costs are low, and when she has a taste for digital consumption, and it may be less relevant when books are

However, Google Books might stimulate a discovery effect due to increased awareness and search-ability (see appendix Figure D.1.) Specifically, some consumers may start consuming the physical

version for the first time, after digitization lowers search costs for books. This is likely to happen if they were made aware of a book through Google Books' search engine and prefer to purchase physical copies rather than read online. This second mass of consumers will drive the discovery effect. The net effect of digitization is ambiguous and depends on the magnitude of these two margins.

The tradeoff between substitution and discovery further differs for different margins of books and con- sumers. Notably, for popular books, already well-known to consumers (e.g. The Wealth of Nations), the substitution effect is likely to dominate. On the other hand, obscure books are likely to benefit from discov- ery, and unlikely to face the costs of substitution. The effect of Google Books on demand should therefore be more positive for less popular books. In addition, if consumers discover a particular author through a dig- itized copy, they might also seek out other books by the same author, even if these have not been digitized. Therefore, digitization might lead to an increase in physical sales for non-digitized works of a digitized author as well.

Further, when the discovery channel is muted, the positive effects on demand should reduce or disappear altogether. For instance, for consumers within Harvard, who already benefit from access to search technol- ogy (through Harvard's librarians and internal catalog system) the substitution effect is likely to dominate the discovery effect. Therefore, when considering loans within Harvard, the effect of digital distribution is likely much smaller, and even negative. On the flip side, when a digital platform provides access only to the search function, but not the entire text of the book (as is common with "snippet view"), we expect the positive effect of demand to remain strong. Our empirical analysis sheds light on these predictions as well.

To summarize, our theoretical predictions are threefold. First, digital distribution allows consumers to search for topics they are interested in and discover works previously unknown to them. Second, if the digital medium offers a poor substitute for a physical book, demand for physical works is more likely to increase. Finally, digital provision will also allow publishers (especially small and independent ones) to identify new material and introduce new editions. Our theoretical arguments speak to past work on the impact of digital distribution on demand and supply for physical products. On the demand side, our arguments about the role of digital distribution in enhanc- ing consumer discovery add to work on the effects of digital technology on the diversity of consumption patterns (Brynjolfsson et al., 2006; Kumar et al., 2014; Holtz et al., 2020). This work shows that digital distribution channels tend to change consumption patterns by helping consumers discover more novel and niche products. Related work has looked at the complementary effects of news content sampled via social networks, which drives traffic to news websites by helping consumers discover specific news articles (Chiou and Tucker, 2017; Sismeiro and Mahmood, 2018). However, this literature focuses on how digital channels change consumption patterns and has not explored how online access coupled with access to digital search and discovery tools affects demand for and supply of the same product in physical form. Our theoretical and empirical analyses extend this work in this direction.

On the supply side, past work has looked at other channels through which access to existing work improves the supply of new content. Most notably, the literature in copyright and digitization shows that free access to past work can often stimulate follow-on production of knowledge (Watson, 2017; Reimers, 2019; Heald, 2007). For example, the (copyright-related) digitization of magazine content improved the quality of content on Wikipedia (Nagaraj, 2018). Further, existing literature also suggests that digital access can be particularly helpful for smaller players and stimulate entry (Nagaraj, 2020; Zhang, 2018). However, whether or not digital distribution itself can improve the supply of physical products and whether it benefits smaller or larger players on this margin remains unexamined.

2 Literature Review

The motivation for this project comes from literature which examines the effect of free or low cost digital distribution of information goods on the demand on their physical counterparts. (Smith and Zentners 2016 review(1).

The is a breadth of resources which investigate the impact of illegal downloads. The bulk of this research focus on the impact of file sharing on music sales (2), while others look at the impact on the film industry(3). Recently, researchers has incorporated legal forms of cheap digital media distribution such as online streaming and found that it also shifted demand of more expensive substitutes(4). In these industries, the literature overall tends not to find that digital distribution shifts supply for the physical counterpart, which is perhaps due to digital distribution doe not explicitly improve the information environment.

There are several key differences to low cost digital distribution for print media. Digitized print media that is scanned in by the entire text can be searched by the entire text, which enables a more complete match between the content and the users query. This match quality can increase demand because of the higher match quality.(5) Specifically, readers are able to find books that they would not want to read or not even know existed. Consumers who want a more user friendly format besides the low quality reader buy the physical book, entering the market.

Another method of acquisition is checking the book out form the library. Books that may have been left on the shelf are now being checked out and read because people were made aware of their existence. However scanning books could also have the opposite effect on the demand for them. A book that's easily available on google books could save a potential consumers a trip the book store or library, or the wait for the online order to deliver. Depending on what type of book the consumer seeks, it could even be more advantageous to have it in a digital format. Large textbooks for example offer a much more utility if they book itself can be queried for specific information, not to mention forgone cost of the physical version.

Since the theoretical argument remains ambiguous, an empirical study into the causal effect is warranted. When designing a causal research question, model specification is key. The most common specification is Two Way Fixed Effects, which compares outcomes for observations with the treatment to observations without the treatment. It is a simple model design, which is flexible to various controls, but has a critical flaw. In circumstances when there are multiple treatment groups, Two Way Fixed Effects is specified in a way which causes observations which have been treated to be compared to newly treated observations. This contaminates the leading and lagging indicators when other assumptions are imposed like parallel trends and limited anticipation of the treatment.(6) The remedy for this issue is found in Callaway Sant'anna (7) specification, which bins the treated observations into time treated cohorts. This separates out the treatment effects and prevents overlapping. It also enables us to check the plausibility of prior trends and see how the treatment effect changes over time.

2.1 MINE THE PAPERS MENTIONED HERE FOR SOURCES THAT SAY: the extant literature does not tend to find that digital distribution stimulates demand for physical products

Our work is closely related to the literature that has looked at the impact of free or low-cost digital distribution of information goods on demand for physical alternatives. This work has largely studied the effects of piracy on the markets for movies and music (See Smith and Zentner (2016) for a

review). A number of papers have looked at the effect of illegal online distribution in the form of piracy on sales of legal music and movies. The majority of this work looking at the impact of free distribution via file sharing on music sales finds a negative effect (Bounie et al. (2006); Rob and Waldfogel (2007); Zentner (2005, 2006); Rob and Waldfogel (2006) see Danaher et al. (2014) and Oberholzer-Gee and Strumpf (2010) for a review), although some work on the movie industry (Bai and Waldfogel, 2012; Danaher et al., 2010) suggests a less pronounced effect. More recently, scholars have also looked at legal forms of cheap, digital distribution such as online streaming and found that it too tends to depress sales in other channels (Yu et al., 2018; Aguiar and Waldfogel, 2018). In these industries, the extant literature does not tend to find that digital distribution stimulates demand for physical products, perhaps because digital distribution does not explicitly improve the information environment.

2.2 lit directly related to e-books' effect on physical copies

Scholars have studied aspects of the market for books such as price dispersion (Ellison and Ellison, 2018), comparing product variety in online and offline formats (Brynjolfsson et al., 2003), substitution be- tween used and new books (Ghose et al., 2006) or platforms for books access (Baye et al., 2015), but the impact of digital distribution on physical sales and especially the role of full-text search are less well under- stood. The few studies that do look at the impact of digital distribution on physical sales in the publishingindustry study contexts where digital distribution is provided without the added benefit of full-text search. For example, Chen et al. (2019) consider the impact of e-book distribution on physical book sales, finding no effect. Similarly, Forman et al. (2009) find that digital and physical distribution channels for books are largely substitutes when considering Amazon.com. A key question that motivates our paper remains unan- swered: What is the effect of digital distribution – combined with full-text search technology – on physical demand in the market for books?

2.3 causal sources

calloway santanna 2021 goodman bacon 2021: Difference-in-differences with variation in treatment timing maybe scunning's mixtape, shout out our boy

3 Data

The raw data we use to replicate Nagaraj and Reimers 2021 is the loan activity for books within the Harvard library system between 2003 and 2011. After processing, 88,006 of the total books were loaned during that time period. Each of these 88,006 books is a panel unit for which we observe the number of loans between 2003 and 2011.

Since most books are never loaned, our analyses focus on 88,006 books (out of over Our panel data structure allows for a difference-in-differences design that can incorporate time and, notably, book fixed effects, increasing confidence in the research design.

The goal of my data cleaning approach is to make it a balanced panel of book level loan counts between 2003 and 2011, such that there are a total of 88,006 observations (one per book) in each calendar year. The main challenge that I found in doing this is that there is not a loan even for every unique book in each year. So to obtain the desired balanced panel, I need to implement "empty" loan events into each year to account for books that are not loaned in said year.

The main pipeline that I use for this task involve Stata's "collapse" and "reshape commands". This is perhaps not the best way to go about this, mainly because, as I will describe in detail, I am interested in the byproduct results of these commands and not the primary results. However, I was unable to find a more direct method of creating this result, and since this method does produce the desired panel, I proceed with it. In a research setting, I would consult colleagues and spend more time looking for a more direct methods of making this panel.

Collapsing the data by book will result in a column of summary statistics (the primary result) and a column of 88,006 unique book ids. The latter column is the vector we are after, and to it, I add a column for each year, which after reshaping, expand the unique cross-sectional book units into a panel that spans 9 years, 2003-2011. Each book is reflected once in each year. There are 792,054 observations in this matrix.

After creating this "blank" panel, I merge the raw data back onto it. The result will be the same as the original raw data with the addition of "empty" loan evens so that every unique book is considered in each year, even if not loaned. I then collapse again to sum loan events by book and year, making sure to preserve location and scanned variables. The result of this collapse is nearly the finished product, however, location and scanned variables do not carry over to the books that have 0 loans for a given year. I have a loop method that is described in detail in the comments of my code, however, I do want to mention that this loop relies on the variables yearscanned and location being invariant within cross-sectional units. This is true by construction for yearscanned—a book cannot be scanned again after it is already in the Google Books system. However, it seems possible for a physical book to be moved from Harvard Library location to another. I do some data exploration to convince myself that location does not vary within books. Given more time, I would prefer to double check this and perhaps check this with colleagues.

4 Methodology

5 Results

Table 1: Table 5 Replication

	(1) log-OLS	(2) log-OLS	(3) LPM	(4) LPM
Post-Scanned	-0.0511***	-0.0518***	-0.0613***	-0.0627***
	(0.00152)	(0.00146)	(0.00170)	(0.00163)
Book FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes
Year-Location FE	Yes	No	Yes	No
N	792054	792054	792054	792054

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Conclusion

Table 2: Additional Results

	(1) log-OLS	(2) log-OLS	(3) log-OLS	(4) log-OLS
Post-Scanned	-0.0518***	-0.0254***	-0.0179***	-0.0254***
	(0.00146)	(0.000862)	(0.000861)	(0.000862)
Faculty × Scanned		0.0760***		
		(0.00145)		
Doctorate × Scanned			0.0475***	
			(0.00167)	
Masters × Scanned			0.0688***	
			(0.00195)	
Undergrad × Scanned			0.0737***	
			(0.00150)	
In-Building × Scanned				0.0760***
				(0.00145)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	1584108	3168216	1584108

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Bacon Decompositison

2x2 Type	Avg. Estimate	Weight
All Borrowers		
Earlier vs Later Treated	-0.195	0.341
Later vs Earlier Treated	-0.184	0.250
Treated vs Untreated	-0.590	0.409
Doctoral Students		
Earlier vs Later Treated	-0.084	0.341
Later vs Earlier Treated	-0.468	0.250
Treated vs Untreated	-0.693	0.409
Faculty		
Earlier vs Later Treated	-0.213	0.341
Later vs Earlier Treated	-0.184	0.250
Treated vs Untreated	-0.485	0.409
In-Building		
Earlier vs Later Treated	-0.213	0.341
Later vs Earlier Treated	-0.184	0.250
Treated vs Untreated	-0.485	0.409
Masters Students		
Earlier vs Later Treated	-0.179	0.341
Later vs Earlier Treated	0.047	0.250
Treated vs Untreated	-0.371	0.409
Undergraduate Students		
Earlier vs Later Treated	-0.048	0.341
Later vs Earlier Treated	0.026	0.250
Treated vs Untreated	-0.242	0.409

Table 4: TWFE vs. Group CS Estimators

Borrower Group	TWFE	Calloway Sant'Anna	Significance	Sign Change
All Bowrrowers	-0.0518	-0.0388	More	No
	(0.00728)	(0.000342)		
Doctoral Students	-0.107	-0.025	More	No
	(0.0169)	(0.00305)		
Faculty	-0.0515	-0.0267	More	No
	(0.0126)	(0.000869)		
In-Building	-0.0515	-0.0267	More	No
	(0.0126)	(0.00196)		
Masters Students	-0.066	-0.0244	More	No
	(0.0149)	(0.00394)		
Undergraduate Students	-0.05	-0.0209	More	No
	(0.0116)	(0.0039)		

Standard errors in parenthesis