Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia

http://orcid.org/0000-0001-9049-0522Abhishek Nagaraj

To cite this article:
http://orcid.org/0000-0001-9049-0522Abhishek Nagaraj (2017) Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia. Management Science

Published online in Articles in Advance 26 Jul 2017

. https://doi.org/10.1287/mnsc.2017.2767

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–17
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia

Abhishek Nagaraj[a]

[a] Haas School of Business, University of California, Berkeley, Berkeley, California 94720
Contact: nagaraj@berkeley.edu, http://orcid.org/0000-0001-9049-0522 (AN)

Received: September 3, 2014
Revised: May 8, 2016; October 7, 2016
Accepted: November 10, 2016
Published Online in Articles in Advance: July 26, 2017

https://doi.org/10.1287/mnsc.2017.2767

Abstract. While digitization has greatly increased the reuse of knowledge, this study shows how these benefits might be mitigated by copyright restrictions. I use the digitization of in-copyright and out-of-copyright issues of *Baseball Digest* magazine by Google Books to measure the impact of copyright on knowledge reuse in Wikipedia. I exploit a feature of the 1909 Copyright Act whereby material published before 1964 has lapsed into the public domain, allowing for the causal estimation of the impact of copyright across this sharp cutoff. I find that, while digitization encourages knowledge reuse, copyright restrictions reduce citations to copyrighted issues of *Baseball Digest* by up to 135% and affect readership by reducing traffic to affected pages by 20%. These impacts are highly uneven: copyright hurts the reuse of images rather than text and affects Wikipedia pages for less-popular players greater than more-popular ones.

History: Accepted by Lee Fleming, entrepreneurship and innovation.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2017.2767.

Keywords: intellectual property • copyright • digitization • economics of innovation • Wikipedia

1. Introduction

The digital representation of information has impacted a wide range of economic activities (Greenstein et al. 2013). Digitization has reduced the cost of storage, computation, and transmission of information, and it has enabled massive changes in the ways that creative producers build on and reuse existing information (Goldfarb et al. 2014). For example, the digital availability of music has allowed artists to sample and remix previously hard-to-find musical pieces, enabling a new wave of creativity in the music industry. (Waldfogel 2014). Similarly, the availability of new digital maps of the earth's surface have led to a wave of discoveries in the gold exploration industry (Nagaraj 2017).

The digitization process is subject to constraint, however; notably, it is governed by intellectual property (IP) and copyright laws that were originally conceptualized for more traditional forms of content. Therefore, the question of whether and how copyright should be modified for the digital age has become a prominent topic of discussion in policy and legal circles (Merges et al. 2013). Some firms have argued for strengthening copyright protection given the difficulties in enforcing copyright on digital information (Anderson 2007), while others have argued that the current copyright regime severely undermines reuse and, therefore, limits the economic potential of digitization (Samuelson 1999, Lessig 2004). Despite the economic significance of these debates (e.g., see Supreme Court case *Authors Guild v. Google, Inc.*, 13-4829 (2d. cir. 2015)), there is little empirical evidence about whether and how copyright influences the diffusion and reuse of digital information. A recent essay describing gaps in the literature summarizes this problem quite succinctly: "[We understand little about] what would be the economic effects of various alternative copyright arrangements and proposals for its redesign" (Greenstein et al. 2010, p. 3).

Theoretically, it is difficult to predict how copyright would affect the possible gains from the digitization process. According to IP prospect theory (Kitch 1977), broad and strong IP protection is needed to spur reuse by facilitating the licensing and maintenance of information (Mazzoleni and Nelson 1998, Gallini and Scotchmer 2002) in the face of digital piracy (Rob and Waldfogel 2007), while transaction cost theories predict that when digitization reduces costs of access, copyright could significantly hinder reuse (Lessig 2005, Benkler 2006, Zittrain 2009, Lemley 2004). While diametrically opposed, the two theories are clear in their respective predictions: according to one theory, copyright lubricates the market for ideas (Gans and Stern 2003) and needs to be strengthened in the face of digitization, while according to the other, copyright impedes the free diffusion of digital information and needs to be reined in.

In this paper, I make empirical progress on the evaluation of the impact of copyright on the reuse of digital information by exploiting a natural experiment that occurred during a marquee project in the history of the Internet: the digitization of about 30 million

works by Google Books. In December 2008, Google Books digitized all existing issues (every issue published between 1940 and 2008) of *Baseball Digest*, a prominent baseball magazine, and made them available online to readers for free. The digitization of *Baseball Digest* is relevant for this study because, due to an accidental failure to renew copyrights, issues of the magazine published before 1964 lapsed into the public domain. Consequently, pre-1964 *Baseball Digest* issues can be freely reused, while those published after 1964 are copyrighted and their reuse without permission is legally prohibited. However, both pre-1964 and post-1964 issues of the magazine were digitized by Google Books and could be accessed and read online by anyone. By exploiting this idiosyncratic variation in copyright, this study sheds light on the broader question of how IP law affects the potential benefits of digitization. In particular, I focus on the effect of copyright on the reuse of magazine material on Wikipedia, a natural venue in which to investigate this question. Not only is Wikipedia the fifth most-visited website on the Internet (receiving about 7.5 billion page views every month)[1] as well as a common source of information about the history of baseball,[2] it also stores and provides open access to all past versions of every page, allowing the analyst to track how information on Wikipedia changed in response to the Google Books digitization event and to copyright.

Specifically, I track Wikipedia citations of all issues of *Baseball Digest* published between 1944 and 1984, both before and after the Google Books project, on Wikipedia (between 2004 and 2012). The unit of analysis is a "publication-year," all issues of the magazine published in a given year between 1944 and 1984. Using these data, I estimate whether out-of-copyright magazine issues (publication-years 1944–1963) were disproportionately more likely to be cited on Wikipedia compared to in-copyright issues (publication-years 1964–1984) after the Google Books digitization event in a difference-in-difference framework. This specification helps to isolate the causal impact of copyright on the reuse of *Baseball Digest* as a source of information after the digitization project. Figure A.2 in the online appendix provides a schematic illustrating the research design.

The key finding from the ensuing empirical analysis is that the digitization project greatly encouraged the reuse of *Baseball Digest* information on Wikipedia, but information from copyrighted issues was significantly less likely to be reused. Specifically, the econometric estimates indicate that after digitization, citations to out-of-copyright publication-years increased by about 135% compared to citations to in-copyright publication-years, even after controlling for publication-year and calendar-year fixed effects. Of course, if Wikipedia editors are able to create high-quality pages for players in the post-1964 era using information from alternate sources and attract similar levels of readership, then the welfare consequences of copyright on Wikipedia are likely not severe. My analysis shows, however, that pages affected by copyright had, on average, 20% lower gains in Internet traffic, suggesting a significant loss to Wikipedia from the inability to reuse copyrighted information. Finally, I explore the idea that the impact of copyright is more concentrated for certain types of information that have higher transaction costs of reuse. This impact is particularly large for images, for instance, which cannot be paraphrased the same way that textual information can and for information about less well-known players. This hypothesis too finds support.

This research contributes to the literature on digitization (Goldfarb et al. 2015, Miller and Tucker 2011, Waldfogel 2014) by evaluating the impact of IP law on the economic consequences of digitization. I show, for the first time, how copyright law could severely curtail potential benefits of digitization in online contexts. I also add to research on the role of digitization in influencing the differences in outcomes between more and less established players in a market (Qian 2014, Mortimer et al. 2012, Zhang 2016, Nagaraj 2017, Brynjolfsson et al. 2006). Finally, I contribute to the nascent empirical literature on copyright, including some work in the legal domain, that estimates the impact of copyright in the publishing context (Heald 2008, 2009a; Buccafusco and Heald 2012) and the impact of copyright on prices (Li et al. 2017, Reimers 2017, Mortimer 2007). This paper also speaks to the growing literature on copyright enforcement and piracy (Aguiar and Waldfogel 2014, Luo and Mortimer 2016, Danaher et al. 2010).

The rest of the paper is organized as follows. Section 2 describes the empirical setting including the *Baseball Digest* experiment and data collection. Section 3 analyzes the overall impact of the *Baseball Digest* copyright experiment, while Section 4 discusses distributional impacts. Finally, Section 5 concludes.

2. Empirical Context and Data

2.1. Empirical Context

2.1.1. The Digitization of *Baseball Digest*. To analyze the role of copyright in the age of digitization, I turned to Google Books, a Google initiative that has as its objective the digitization of all books ever published. It currently offers a catalog of about 30 million works (Wu 2015). The well-known copyright law academic Pamela Samuelson has called the project "one of the most significant developments in the history of books, as well perhaps in the history of copyright" (Samuelson 2009, p. 1308). To understand how copyright law influences the impact of the Google Books project, I focus on the

December 2008 announcement by Google Books that, with consent from the publisher, it had made available every page from all past issues of *Baseball Digest* (Foulser 2008). The digitization did not proceed gradually over time—all issues published before December 2008 were simultaneously accessible on the Google Books website as of December 9, 2008.

2.1.2. The 1964 Copyright Experiment.

Baseball Digest is the focus of this study because it is one of the few titles that I am aware of that offers variation in both copyright and digitization status. While it is generally assumed, both in practice and in the literature, that a work is either completely in copyright or in the public domain, magazines and periodicals can offer variation within a single publication because their copyright term is defined not in terms of the author's life term but by the publication date. Specifically, prior to 1964, periodicals were subject to the *copyright renewal* requirement. Under the 1909 Copyright Act, two copyright terms were provided: a 28-year initial term and a 28-year renewal term (Landes and Posner 2003). However, the renewal term was not automatically granted (Kupferman 1944), and if the renewal application was not filed on time, the work entered the public domain after the first 28 year term expired. For issues published after 1964, the 1909 Copyright Act no longer applies and these works are automatically granted a second 28-year copyright term. This stipulation meant that a magazine issue published in December 1963 would relinquish its copyright in 28 years, i.e., in 1991, if the copyright was not renewed. However, for an issue published in January 1964, copyright would last for 56 years, i.e., till 2020, because renewals were not necessary. In this way, as of 2012 (when the data collection stopped), issues of a single publication could have widely varying copyright protection, despite being printed only a few months apart from each other. Because this requirement was not well known, it "tripped up" many small publishers and a "failure to renew caused many works originally published from 1923 through 1963 to enter the public domain" (Public Domain Sherpa 2014).

I leverage the University of Pennsylvania library's "First Copyright Renewals for Periodicals," which conducted a thorough review of the Catalog of Copyright Entries for each periodical printed before 1964, to clarify the copyright status of *Baseball Digest*. According to this source,[3] no issues of *Baseball Digest* published before 1964 were ever renewed and have thus entered the public domain, while issues published in or after 1964 will remain under copyright. Thus the publication date of the periodical (before or after 1964) determines whether a particular issue was under copyright and the date of access (before or after December 2008) determines whether it was digitized.

Focusing on *Baseball Digest* is useful for three reasons. One, it is one of only a very small number of publications that have been digitized by Google Books in their entirety. Two, of these, *Baseball Digest* is the only one that I'm aware of that also has a sharp variation in its copyright status. Three, baseball is a good choice of topic because it has a thriving editor community on Wikipedia. Further, the experiment is likely to be economically meaningful given the widespread interest in both the game of baseball and in *Baseball Digest*. Over 45% of all Americans identify as baseball fans, and revenues from the sport of baseball in 2010 were estimated to be approximately 7 billion USD.

2.2. Data

To understand the impact of *Baseball Digest*'s copyright status on reuse, I look at citations on Wikipedia. There are many reasons why Wikipedia is a natural venue for such analysis. First, Wikipedia is the preeminent source of information on the Internet. A total of 56% of typical Google noun searches point to a Wikipedia page as their first result, and 99% point to a Wikipedia entry on the first page (Silverwood Cope 2012). Second, Wikipedia is built explicitly on the "No Original Research" rule, which requires editors to cite a secondary source for information, making citations to magazines like *Baseball Digest* typical on the site. Third, *Baseball Digest* frequently runs profiles of baseball players and teams, including detailed articles, interviews, and player images. Such biographical information forms the foundation of any encyclopedia (Greenstein and Zhu 2017) and is, therefore, particularly likely to be reused on Wikipedia. Finally, each revision of a Wikipedia page is archived and publicly accessible. This allowed me to collect repeated panel data on Wikipedia pages both before and after a digital version of *Baseball Digest* was made available.

2.2.1. Sample A: Publication-Year Level.

To construct the sample for the main specification (Sample A), I followed a four-step process. First, I searched the entire Wikipedia repository for pages that contained mentions of the word `baseball digest` and variants thereof. Second, for pages that contained references to *Baseball Digest*, I accessed and downloaded every past version of the page between 2004 and 2012 as it appeared on December 1 of that year.[4] Third, I wrote python scripts to detect citations to *Baseball Digest* magazine on each page snapshot, collecting all citations by publication-year for every year between 1944 and 1948.[5] Fourth, I organized the data such that it provides total citations for each publication-year of *Baseball Digest* for each calendar-year between 2004 and 2012. For example, for a given publication-year, say 1963, Sample A tracks citations on Wikipedia as of 2004, 2005, and so on, up to 2012. Summary statistics are available in Table 1. This shows that, on average,

Table 1. Summary Statistics

	Mean	SD	Median	Min	Max
(1) Sample A: Unit of observation—Publication-year ($N = 360$)					
Publication-Year	1,964.50	11.56	1,964.50	1,945	1,984
Wikipedia-Year	2,008.00	2.59	2,008.00	2,004	2,012
1(*Out-of-Copy*)	0.47	0.50	0.00	0	1
1(*Wikipedia-Year* > 2008)	0.44	0.50	0.00	0	1
Total Citations	4.19	7.75	0.00	0	51
Image Citations	1.19	4.03	0.00	0	30
Text Citations	3.00	4.75	0.00	0	21
(2) Sample B: Unit of observation—Player-page ($N = 4{,}869$)					
Player Debut-Year	1,966.12	10.19	1,966.00	1,944	1,984
Wikipedia-Year	2,008.00	2.58	2,008.00	2,004	2,012
1(*Out-of-Copy*)	0.38	0.49	0.00	0	1
1(*Wikipedia-Year* > 2008)	0.44	0.50	0.00	0	1
Total Citations	0.17	0.86	0.00	0	10
Total Images	0.66	1.30	0.00	0	18
Total Text	1.18	1.41	0.78	0	16
Average Traffic	101.49	224.94	34.73	0	3,395
Quality Percentile	2.62	1.29	3.00	1	4

Notes. This table presents summary statistics for the two main data samples used in this study. Both samples track citations to *Baseball Digest* on Wikipedia between 2004 and 2012. In Sample A, the unit of observation is a publication-year of *Baseball Digest*—i.e., all years between 1944 to 1984. For each of the 40 publication-years, I track total citations in every Wikipedia-year between 2004 and 2012, for a sample size of 360 observations (40 issue-years times 9 calendar years). For Sample B, the unit of observation is an individual Wikipedia player-page for 541 notable baseball players. On each player-page, citations to *Baseball Digest* are tracked (irrespective of the year of publication) between 2004 and 2012, for a total of 4,869 observations (541 pages times 9 calendar years). 1(*Out-Of-Copy*) is defined as all publication-years (Sample A) or debut-years (Sample B) before 1964. Traffic data are only available for years 2007–2013.

a publication-year of *Baseball Digest* receives 4.16 citations, of which about three are for text and the rest for images.

2.2.2. Sample B: Player-Page Level. Sample A is intended to provide an accurate assessment of the reuse of *Baseball Digest* information on Wikipedia as a function of copyright, but it is inadequate for evaluating the impact of copyright on the quality of Wikipedia pages. Thus, I build Sample B at the player-page level, which uses data on the amount of content on a Wikipedia page (i.e., number of words of text and quantity of images), as well as proxies for quality, such as player-page–level traffic. A player-page-level analysis also helps to directly characterize the differential impact of copyright on different types of content. In particular, by estimating whether Wikipedia pages for well-known players are more affected by copyright than pages for less well-known ones, the heterogeneous effects of copyright on different types of topics can be better understood.

To build this sample, I compiled a list of 541 players who have been nominated for election to the Baseball Hall of Fame and who made their debut appearances between 1944 and 1984. This screening process captures players who have been judged by others as significant to the game. The data set also provides biographical details of the players, including date of debut, and performance details, including experience, length of career, and number of appearances in all-star games. I then created a *quality* metric for each player based on their percentile rank in the number of all-star appearances they made. Next, I downloaded archival versions of each player's page as it appeared on December 1 for every year between 2004 and 2012. The full details of the data set construction process are described in Online Appendix A.3. The result is a data set that counts the number of citations to *Baseball Digest* on each player's Wikipedia page as well as the number of images and the number of words of text.

3. Empirical Results

3.1. Descriptive Analysis

To estimate the impact of the 1964 copyright experiment on the reuse of the digitized *Baseball Digest* magazine content, it is helpful to begin by exploring some simple descriptive statistics from the data.

3.1.1. Cross-Sectional Comparison of Out-of-Copyright and In-Copyright Groups. Table 2 provides simple descriptives comparing the likelihood of reuse of information in out-of-copyright and in-copyright issues of *Baseball Digest*. Specifically, panel (1) compares citation and reuse outcomes for in-copyright and out-of-copyright publication-years of *Baseball Digest* using Sample A, whereas panel (2) compares Wikipedia player-pages for in-copyright and out-of-copyright players using Sample B.

As these data make evident, there were important differences in reuse outcomes for in-copyright and out-of-copyright material as of 2012. Specifically, panel (1) of Table 2 reports that total citations to in-copyright publication-years was about 10.33, while out-of-copyright publication-years received almost double the number, for a total of about 20.95 citations.

Table 2. Cross-Sectional Comparison of Reuse Outcomes

	(1) Out-of-copy $\bar{y}$	(2) In-copy $\bar{y}$	(3) Diff.	(4) p-val.
(1) Sample A: *Baseball Digest* publication-years				
Total Citations	20.95	10.33	10.61	0.00
Image Citations	10	0.0952	9.905	0.00
Text Citations	10.95	10.24	0.709	0.65
(2) Sample B: Wikipedia player-pages				
Total Citations	0.602	0.334	0.268	0.03
Total Images	1.786	0.916	0.870	0.00
Total Text	2.128	1.645	0.483	0.00
Average Traffic	158.9	111.5	47.43	0.03

Notes. This table compares outcomes for out-of-copyright and in-copyright groups using cross-sectional data from 2012 Wikipedia data. For panel (1), $N = 40$; for panel (2), $N = 541$. In panel (1), column (1) includes publication-years 1944–1963, whereas column (2) includes publication-years 1964–1984. In panel (2), column (1) includes all out-of-copy player-pages (debut before 1964) and column (2) includes all in-copy player-pages (debut after 1964). The p-value reported in column (4) is from a t-test for a difference in mean outcomes across columns (1) and (2).

However, these differences seem to be derived largely by differences in image citations (i.e., citations for the reuse of images) as compared to differences in text citations. Similarly, panel (2) of Table 2 finds that out-of-copyright players receive almost double the number of citations to *Baseball Digest* (0.60 as compared to 0.33 per page on average) in 2012, have on average about 1.78 images as compared to 0.92 for in-copyright player-pages, and attract about 47 more visitors per month on average.

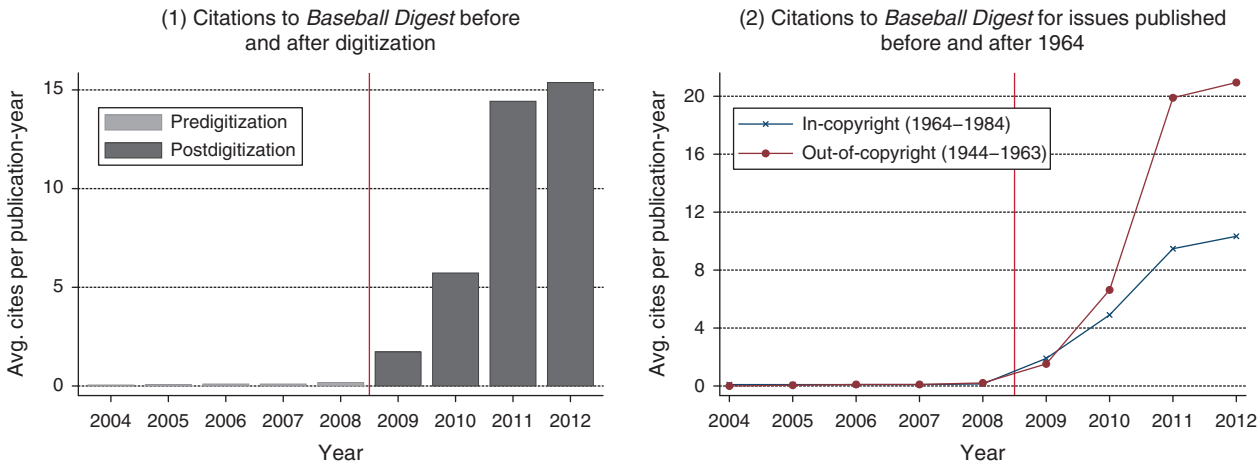
The large cross-sectional differences are a first striking piece of evidence that suggest a large impact of the 1964 copyright experiment on reuse outcomes on Wikipedia.

3.1.2. Time-Series Comparison of Citation Trends Between 2004 and 2012.

Having explored cross-sectional patterns, I now explore temporal trends in the data. First, using Sample A, I explore whether the digitization event had any impact on the reuse of material from *Baseball Digest*. Panel (1) of Figure 1 plots the average number of citations per *Baseball Digest* publication-year for the years 2004 and 2012. As these data indicate, the average *Baseball Digest* publication-year received a very small number of citations (about 0.125 cites) before the Google Books project made this material more accessible. However, after the Google Books digitization project, citations to *Baseball Digest* issues increased dramatically, to an average of 15.4 citations per publication-year by 2012. This dramatic increase in citations, which started in 2009, suggests the potential for large benefits to Wikipedia from the digital availability of the magazine on Google Books.

While this graph suggests a strong and positive effect of the digitization program on reuse, it does not provide a numerical estimate of its impact. In the online appendix, I include an analysis that provides a causal estimate of the impact of digitization on reuse. For this exercise, I collected data analogous to Sample B for a comparable set of basketball player-pages on Wikipedia. Using this sample, I am able to compare reuse outcomes for baseball player-pages compared to basketball player-pages while controlling for average time trends on page quality across Wikipedia in a difference-in-difference framework. As Table A.1 in the online appendix shows, the regression estimates indicate that the digitization program increased citations to *Baseball Digest* by about 0.34, an almost 300% increase as compared to the average level of 0.12. The impacts seem to translate across both image and text citations.

Figure 1. (Color online) Citations to *Baseball Digest* on Wikipedia (Sample A)



Notes. Panel (1) presents average citations per publication-year of *Baseball Digest* on Wikipedia before and after the Google Books digitization event in 2008. Panel (2) divides the data into two categories, in-copyright and out-of-copyright, and presents average citations per publication-year for each year of the period 2008–2012.

Having confirmed that the digitization of *Baseball Digest* appears to benefit Wikipedia as a source of information, panel (2) of Table 1 plots the average number of citations to *Baseball Digest* separately for out-of-copyright and in-copyright publication-years between 2004 and 2012. As this figure makes clear, the gains in citations were heavily concentrated among issues published before 1964 and thus out of copyright. In fact, my data suggest that while in 2004, both in-copyright and out-of-copyright issues of *Baseball Digest* averaged about 0.05 citations, in 2012, this number increased to about 21.1 citations for out-of-copyright issues compared to only 10.3 for in-copyright issues.

3.2. Estimating the Effects of Copyright on Reuse

A number of different theories could explain both the cross-sectional and temporal trends in the data; therefore, it is difficult to conclude from the descriptive analysis alone that the difference in copyright status of pre- and post-1964 issues of *Baseball Digest* is the primary driver of the empirical patterns. For example, if players who played before 1964 ultimately became more well known, then cross-sectional differences in the amount of content on their pages could be explained by the difference in reader interest rather than the copyright experiment. The cross-sectional evidence from Sample A suggesting that pre-1964 issues have higher citations to *Baseball Digest* as compared to post-1964 issues is perhaps more convincing. However, even in this instance, it is possible that pre-1964 issues were of higher quality (perhaps because certain well-known writers contributed articles) or because pre-1964 issues contained more material of interest to the general reader. Given these difficulties in interpreting the cross-sectional data, I use a regression framework in this section to directly test the central hypothesis of this paper, that difference in copyright status of pre- and post-1964 issues is the primary driver of the difference in the levels of reuse of magazine material on Wikipedia.

Specifically, using both Sample A and Sample B, I estimate versions of

$$Cites_{it} = \alpha + \beta_1 \times Post_t \times Out\text{-}of\text{-}Copy_i + \gamma_i + \delta_t + \varepsilon_{it},$$

where γ_i and δ_t represent unit-of-observation and time fixed effects, respectively, for unit i and Wikipedia-year t, and indicator variable $Post_t$ equals one if the observation is from any Wikipedia-year after 2008. The coefficient β_1 on the variable of interest $Post_t \times Out\text{-}of\text{-}Copy_i$ estimates the differential impact on the out-of-copyright group as compared to the in-copyright group after the *Baseball Digest* digitization event. The main outcome variable, $Cites_{it}$, measures the total number of citations to a given publication-year i in Wikipedia-year t (Sample A), or total citations to any issue of *Baseball Digest* magazine on a player-page i in Wikipedia-year t (Sample B).

In Sample A, the unit-of-observation fixed effect (FE) controls for publication-year fixed effects, which flexibly controls for time-invariant differences between publication-years 1944–1984. Further, $Out\text{-}of\text{-}Copy_i$ is an indicator variable that equals one for all publication-years before 1964 and zero otherwise. In Sample B, the unit-of-observation fixed effect controls for player-page fixed effects, which flexibly controls for inherent differences in player quality for each of the approximately 500 players in the sample. Further, $Out\text{-}of\text{-}Copy_i$ is an indicator variable that equals one if a player makes his debut before 1964. Table 3 presents estimates from ordinary least-squares (OLS) models for Samples A and B. For both samples, the first two models are estimated using OLS while the third model is estimated using a Log-OLS specification, where the dependent variable is logged.[6] The first OLS model is estimated without publication-year (Sample A) or player-page

Table 3. Impact of 1964 Copyright Experiment on Total Citations

	Sample A			Sample B		
	Cites	Cites	Log-cites	Cites	Cites	Log-cites
Out-of-Copy × *Post*	5.595	5.605	0.322	0.202	0.188	0.0690
	(1.785)***	(1.774)***	(0.158)**	(0.0588)***	(0.107)*	(0.0360)*
Unit of obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Page-age FE	—	—	—	Yes	Yes	Yes
Adj. R^2	0.646	0.709	0.903	0.0592	0.0851	0.116
N	360	360	360	4,869	4,869	4,869

Notes. This regression estimates the impact of the 1964 copyright exception on citations to *Baseball Digest* before and after digitization in a difference-in-difference framework using OLS. The estimates presented use both Sample A and Sample B. *Post* refers to all Wikipedia-years after 2008, and *Out-of-Copy* refers either to *publication-year* < 1964 (Sample A) or *debut-year* < 1964 (Sample B). Clustered standard errors are shown in parentheses.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

(Sample B) fixed effects. All models include Wikipedia-year fixed effects.

Consistent with the descriptive analysis, even after flexibly controlling for differences between units of observation and secular time trends, the estimates suggest that out-of-copyright *Baseball Digest* issues are cited at a significantly higher rate compared to in-copyright issues across the different specifications. The estimate in the second column from the regression using Sample A, which includes both publication-year and Wikipedia-year fixed effects, suggests that after the digitization event, *Baseball Digest* issues published before 1964 received about 5.6 more citations as compared to issues published in or after 1964. This is an increase of almost 135% above the average citation level of 4.2. These estimates are somewhat muted, although they remain large and significant in the Log specification in the third column, which suggests that citations to out-of-copyright issues increase by about 31.6% after the digitization event. The difference between the Log and OLS estimates is important to note,[7] and the Log estimates should be preferred for a more conservative estimate of the impact of copyright. Sample B provides additional evidence for the negative effect of copyright law on preventing the diffusion of material to Wikipedia.[8] In addition to the player-page and year fixed effects, these estimates also include fixed effects for page-age—i.e., the number of years since the page was created in the focal year. This additional fixed effect helps control for the concern that pages for more-popular players might have been created earlier, and that these pages had more time to accumulate citations. After adding this additional control, the estimates from Sample B suggest that after the Google Books digitization event, pages for players who made their debut before 1964 are more likely to cite material from *Baseball Digest* compared to in-copyright player-pages. The OLS estimate from Sample B, Column (2) suggests about 0.2 additional citations compared to an average of 0.12, an increase of about 160%. Similarly, the estimates from the log specification are also positive and significant, but are considerably smaller, suggesting an increase in citations of about 6.9%. This analysis reveals two findings: citations to out-of-copyright issues increased significantly after the digitization project, and this impact was mostly concentrated among players who made their debut before 1964, thus the players who were most likely to be affected by the 1964 copyright experiment.

Taken together, estimates from Samples A and B are able to robustly confirm the main hypothesis of this study: the copyright restrictions on digitized, post-1964 *Baseball Digest* material significantly reduced the likelihood of their reuse on Wikipedia.

3.3. Checking for Pretrends and Other Robustness Checks

The evidence thus far for the causal role that copyright status plays in determining the level of reuse of *Baseball Digest* content is strong, but alternative explanations still exist. The analyses below use alternative specifications to rule out these competing explanations, thus improving our confidence in copyright status as the true causal variable.

3.3.1. Time-Varying Estimates. The regression analysis reported above accounted for time-invariant differences by including controls for Wikipedia-year-, publication-year-, and player-page-level fixed effects. However, it remains possible that time-varying differences between in-copyright and out-of-copyright groups might be doing the heavy lifting. For example, if older issues of *Baseball Digest* are coming back into circulation, or if pre-1964 baseball players are coming back into fashion right before the digitization event, then the regression specification is likely to mistake a positive coefficient on β_1 for a causal effect of the copyright exception on reuse.

A standard way to investigate this concern in the difference-in-difference literature (Bertrand et al. 2004) is to explore the difference between the treatment and control group separately for each year before and after the causal event. The main identifying assumption for the specification (i.e., similar time-trends) implies that before the digitization event, the difference in citations between the out-of-copyright and in-copyright groups is constant and is not trending upward. If citations for the out-of-copyright group are increasing relative to the in-copyright group even before 2009, then the validity of main difference-in-difference specification becomes uncertain.

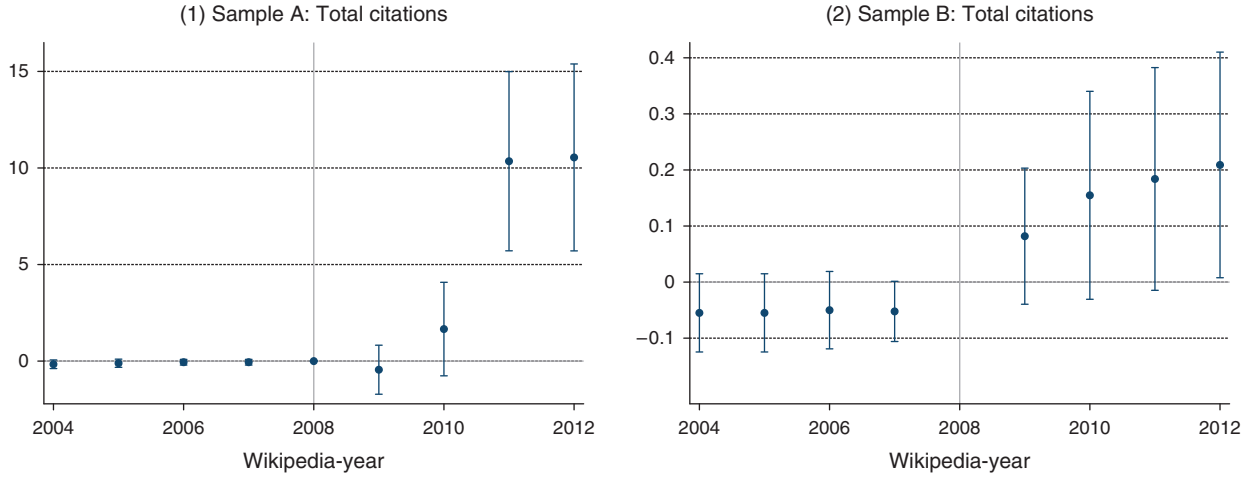
Accordingly, Figure 2 presents graphical versions of the following event study specification separately for Samples A and B:

$$Cites_{it} = \alpha + \gamma_i + \delta_t + \Sigma_t \cdot \beta_t \cdot Out\text{-}of\text{-}Copy_i \times 1(t) + \varepsilon_{it}$$

for unit i in Wikipedia-year t.

However, the time-varying coefficients in Figure 2 reveal no discernible evidence for an increase in citations for out-of-copyright groups compared to in-copyright groups before 2009. Panel (1), which estimates this specification with Sample A, finds virtually zero difference in the level of citations between pre- and post-1964 issues before 2009; a positive and significant difference emerges only after the digitization event in late 2008. Panel (2) paints a similar picture, although the differences by year are estimated less precisely. In particular, there seems to be a negative but insignificant difference between out-of-copyright and in-copyright player-page citations to *Baseball Digest* before 2009. However, this difference is relatively flat

Figure 2. (Color online) Time-Varying Estimates of the Impact of Copyright on Citations to *Baseball Digest*



Notes. This figure plots coefficients and 95% confidence intervals from the event study specifications described in Section 3.3.1. The reference year is 2008, the year of the digitization event. The coefficients are estimates from OLS models, standard errors are clustered, and the dependent variable in both panels is the total number of citations in a calendar year.

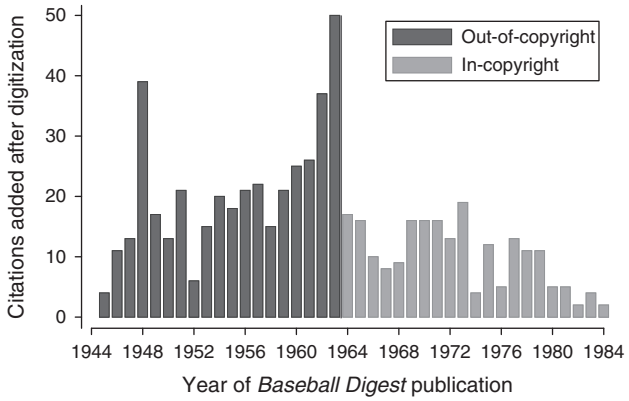
and does not seem to be changing before 2009. After 2009, we see a large positive effect, even though none of the coefficients are individually significant. Table A.2 in the online appendix presents these results in tabular format. The evidence from this analysis, especially from the precisely estimated coefficients in Sample A, significantly reduces the concern that citations to out-of-copyright and in-copyright groups were evolving at a different rate before the digitization event.

3.3.2. Exploiting Discontinuity Around the 1964 Copyright Cutoff. In addition to exploring pretrends directly, it is possible to examine the robustness of the main result using another feature of the setting that has so far been underexploited. Specifically, the setting allows for examination of the impact of the IP law using a strategy that exploits the sharp distinction in copyright status between issues published only a few years on either side of the 1964 cutoff. In principle, if the main effects are driven by the fact that older issues benefit more from the digitization project, then we should see a gradual decline in reuse between issues published before and after 1964. However, if the copyright law is affecting reuse directly, then we should see a discontinuous change in levels of reuse around publication-year 1964. This exercise is a robust check of the main specification because it provides a nonparametric method to examine the impact of copyright on reuse that does not rely on the "pretrends" assumption inherent to the difference-in-difference specification.

Accordingly, Figure 3 plots the net increase in citations that each publication-year from 1944 to 1984 experienced between 2008 and 2012. As the graph indicates, the number of new citations does not display a steady time trend that decreases from 1944 to 1984. Instead, issues published right before 1964 have a significantly higher increase in the number of new citations compared to issues published right after. For example, issues published in 1963 gained about 51 citations, while issues published in 1964 gained only about 17 citations. This sharp discontinuous difference in the likelihood of new citations for out-of-copyright publication-years increases confidence in the main hypothesis: the disproportionately large increase in the reuse of pre-1964 issues of *Baseball Digest* was caused by the difference in copyright status rather than by confounding factors such as different pretrends in citation patterns.

3.3.3. Additional Falsification Checks. The online appendix presents additional robustness checks that help build confidence in the main results. First, I conduct a falsification analysis,[9] in which I restrict the sample

Figure 3. Citations to *Baseball Digest* Published Before and After the 1964 Copyright Cutoff (Sample A)



Note. This figure plots the increase in the number of citations to *Baseball Digest* publication-years between 2008 and 2012.

to the "pre" period only (2004–2009) and assume that the treatment year is 2007 rather than 2009. If out-of-copyright groups are experiencing an increasing rate of citations compared to in-copyright groups, then we would expect the coefficient on β_1 in this regression to also be positive and significant. However, the estimates from this falsification check (see Table A.3 in the online appendix) are close to zero and not significant when both unit-of-observation fixed effects and time fixed effects are included.

The time-varying estimates, the discontinuity plots, and the falsification analysis help to address the concern that differing pretrends between in-copyright and out-of-copyright groups might be driving the main results. In addition to the pretrends, it is also important to address the pattern of the time trend in the estimates after the digitization event. Specifically, Figure 2 suggests that while the digitization happened in late 2008, the positive impact of out-of-copyright status seems to become apparent around 2011. Both qualitative evidence and additional robustness checks help to confirm that this increase around 2011 is not due to an unrelated external event that might have influenced reuse.

Specifically, my qualitative evidence suggests that in 2011, certain Wikipedia "power" editors became aware of the digitized *Baseball Digest* issues (through novice users) and were heavily involved in reusing material from the magazine to improve Wikipedia. This pattern, where certain novice users make contributions that help to attract the attention of core users, is quite common on Wikipedia (Gorbatai 2014). For example, a "power" Wikipedian belonging to the WikiProject:Baseball community I interviewed told me the following:

> I found out that the *Baseball Digest* issues from before 1964 fell into the public domain (PD) as the copyright expired (around 2010). As a result, any images in those issues are free to use. Originally found that out when I saw a Brooks Robinson free pic used from *Baseball Digest* and knew there would be other images out there.
>
> (Interview, December 2011)

This quote helps explains why the positive effects of the program could be concentrated in certain calendar years rather than being evenly distributed. In addition, for robustness, I include two specifications in the online appendix where I shorten the time-frame of the analysis and reestimate the main specifications. Specifically, panel (1) of Table A.4 in the online appendix reports the estimates of the specification using a sample from 2005–2011, whereas panel (2) uses a sample from 2006–2010. When I shorten the scope of the analysis to these years, the coefficient on β_1 remains positive, although in panel (2), the main estimate from the Sample A specification becomes imprecise given the shorter time span.

Taken together, the falsification and robustness checks help to build confidence that the main estimates are not driven by different alternative explanations.

3.4. The Effect of Copyright on Traffic

Thus far, I have established that citations to out-of-copyright issues increased at a significantly higher rate than citations to in-copyright issues of *Baseball Digest*. While this is important evidence of the impact of copyright law on the diffusion of digitized material, I now turn to Wikipedia traffic information to evaluate the welfare impact of copyright on Wikipedia. Specifically, if Wikipedia contributors are able to supplement copyrighted information not available from *Baseball Digest* with information from other sources, then we might find that the suppression of citations to *Baseball Digest* due to copyright does not in fact translate into lower-quality Wikipedia pages. However, if the lack of citations to in-copyrighted issues of *Baseball Digest* does cause lower traffic to Wikipedia pages, this will suggest that the impact of copyright on welfare could be significant. One anecdotal mechanism through which lower quality of Wikipedia pages seems to lead to lower traffic is a lower ranking of such pages on Google search results or their lower likelihood of being linked to or shared online.

Panel (2) of Table 2 presents a cross-sectional comparison of Wikipedia traffic information, indicating that, on average, *out-of-copy* player-pages have about one and a half times more traffic than *in-copy* player-pages. While this difference is striking, it could simply be driven by differences in player popularity over time, with the pre-1964 players being significantly more well known than their post-1964 counterparts. In this section, I utilize traffic data[10] from Sample B in a regression framework to shed light on the impact of the 1964 copyright experiment on traffic. The specification follows the form as the equation in Section 3.2, with the main outcome variable being $Traffic_{it}$ for player-page i in year t and with player-page and year fixed effects to account for systematic differences between players and traffic trends over time.

Table 4 reports estimates from such an analysis. Models (1) and (2) include year fixed effects, while model (3) includes separate year-trends for each of the four player-quality quartiles. Models (2) and (3) also include additional player-page-level fixed effects. The estimates in column (2) indicate that, on average, out-of-copyright player pages receive a boost of about 20 hits per month after controlling for player and year fixed effects. Against a mean of about 101 page views per month, this represents an increase of about 20%. The coefficient reduces slightly when *quality* × *year* fixed effects are included. Despite strong results, the limited traffic data from the predigitization period makes it difficult to validate the parallel trends assumption and the estimates should be interpreted with caution. Given this issue, in the online appendix, I examine the robustness of these estimates in log models using a cross-sectional specification. Columns (3)

Table 4. Impact of 1964 Copyright Experiment on Wikipedia Traffic (Sample B)

	Traffic		
	(1)	(2)	(3)
Out-of-Copy × *Post*	43.22	20.42	16.54
	(12.09)***	(9.883)**	(10.13)+
Player-page FE	No	Yes	Yes
Time FE	Year FE	Year FE	Quality × Year FE
Adj. R^2	0.0137	0.0810	0.0899
N	3,246	3,246	3,246

Notes. This regression estimates the impact of the 1964 copyright exception on traffic to Wikipedia player-pages before and after digitization in a difference-in-difference framework using OLS. The estimates presented use data from Sample B. *Post* refers to all Wikipedia-years after 2008, and *Out-of-Copy* refers to *debut-year* < 1964. In column (3), *Quality* × *Year FE* controls for separate time-trends by each of the four quartiles of player quality. Standard errors clustered at player-level are shown in parentheses.

+$p < 0.15$; **$p < 0.05$; ***$p < 0.01$.

and (4) of Table A.7 in the online appendix estimate the impact of copyright on traffic to be about 88.8%, or twice as large as the OLS estimates. However, a conservative estimate of the impact of out-of-copyright content on traffic to affected pages would be to boost page views on the order of 20%, a significant difference.

Overall, my estimates suggest a significant positive impact of the digitization program for traffic to out-of-copyright player-pages, implying that Wikipedia editors are unable to substitute copyrighted content with information from other sources. The negative impact of copyright on reuse, therefore, has real effects on Wikipedia readers. In other words, pages affected by copyright are unable to fully capture and deliver value to end users, and ultimately, copyright seems to harm not only the diffusion of material from *Baseball Digest* but also traffic to affected pages on Wikipedia. This impact is important when considering the welfare impact of the 1964 copyright experiment on the social value of Wikipedia.

4. Differential Effects of Copyright on Reuse

4.1. Theoretical Framework

Existing scholarship indicates that digitization affects markets by reducing the cost of access to information (Goldfarb et al. 2014, Bakos 1997). The theory is that the digital availability of information makes it much easier to locate relevant knowledge and then build on it (Shapiro and Varian 1998, Chiou and Tucker 2017). Therefore, when information from printed material is digitized, we should expect its reuse to increase. The empirical evidence presented so far is consistent with this hypothesis, as shown in panel (1) of Figure 1. Meanwhile, a robust literature in IP has argued that IP might introduce transaction costs that mute the benefits from the reduction in the cost of access (Waldfogel 2012, Williams 2013, Murray and Stern 2007). Together, this research suggests that, while digitization might encourage access and reuse, transaction costs imposed by copyright might mitigate potential gains from digitization (Gans 2015). These predictions are consistent with panel (2) of Figure 1 and the regression evidence presented in Section 3.

In this section, I use the logic of transaction and access costs to sketch a brief theoretical framework through which the heterogeneous impact of copyright can be understood. Specifically, I focus on differences in outcomes for different types of information (notably text versus images) and different player-pages on Wikipedia. I then proceed to empirically testing these hypotheses. A formalization of the ideas in this section is presented as a simple theoretical model in Online Appendix A.2.

4.1.1. The Differential Impact of Copyright: Images vs. Text.

Consider the impact of copyright on the reuse of images and textual material. Digitization lowers access costs for both types of media, but the transaction costs that are required to prevent copyright infringement are significantly different for images as opposed to text. For work to be reused without copyright infringement, some evidence of "transformative reuse" is often necessary. While copyright on text prohibits only verbatim reuse of large sections of text, the same rule does not (and cannot) apply to image. In other words, paraphrasing of text is possible and constitutes "fair use," but for images, the standard for "fair use" is much higher (Leval 1990).

Consequently, to prevent infringement, the reuse of textual information with sufficient modifications is typically straightforward and possible to do at low cost. However, for images to be legally reused, large modifications need to be made to satisfy the "transformative reuse" criterion, making the process more complicated and significantly costly. In practice, even when significant changes are made, copyright infringement is possible (see *Cariou v. Prince*, 714 F.3d 694 (2d. cir. 2013) for an example); therefore, end-users often avoid the reuse of copyrighted images without explicit permission from the creators.

I argue that differences in practical and legal standards for reuse impose different transaction costs on textual and visual content. For images, transaction costs of reuse are likely to be high, while for textual material, they are likely to be low. Therefore, while digitization lowers access costs for both types of content, for images, copyright limits some of these gains by imposing greater transaction costs compared to text. It follows naturally from this argument that the reuse of information from out-of-copyright status is likely to be higher for images than for text. This is the first hypothesis that I explore in this section.

4.1.2. The Differential Impact of Copyright by Player Quality. Second, I argue that copyright will also have distributional effects across different types of topics. Specifically, I propose that the effects of copyright on the level of information on Wikipedia pages is more pronounced for low-quality player-pages than for high-quality player-pages. (Note that I am referring to the quality of the players, not the quality of the page.)

To understand how copyright affects player-pages, I argue that the optimal level of knowledge on a Wikipedia page depends both on the value of new information to that page and the cost of adding new information. A large literature in media economics finds that the provision of news about events is directly proportional to commercial interest (Prat and Strömberg 2013, Strömberg 2007). In line with this research, we can expect the value of information about players to be directly proportional to player quality and the costs of sourcing information to be inversely related. Higher-quality players attract more interest from end users and therefore the value of information for these players is greater. However, there are a number of alternate sources of information for higher-quality players, which makes it cheaper to source information about them. The assumption for lower-quality players is the inverse. According to this logic, there are greater incentives to add information for high-quality players and this information is easier to obtain, even before the digitization of *Baseball Digest* by Google Books.

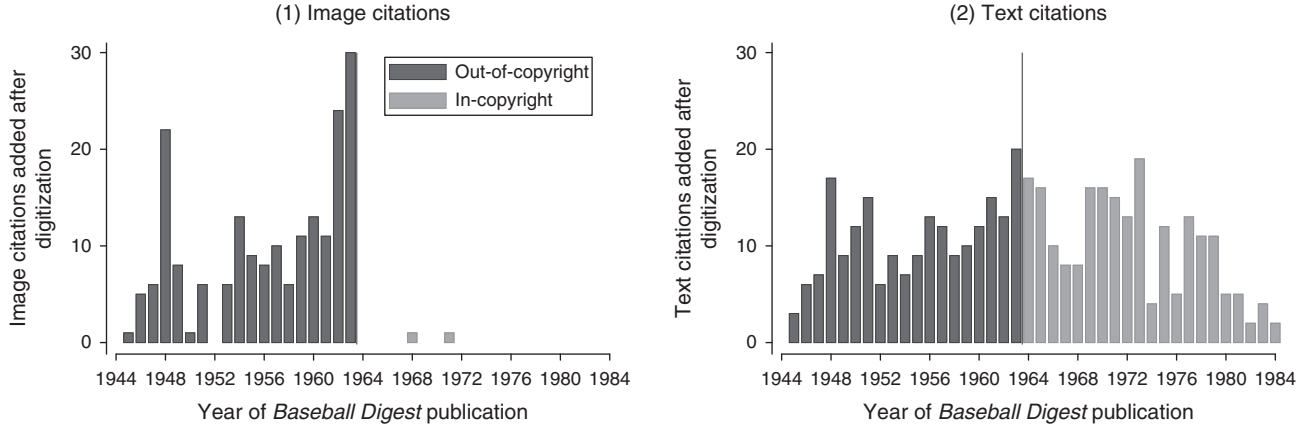
Now, consider the reduction in the cost of access to information due to digitization (Goldfarb et al. 2014) and the increased transaction costs due to IP (Williams 2013). For players of the highest quality, information was readily available even before digitization, and the marginal utility of new information is low. For obscure players, even if information exists on *Baseball Digest*, reuse is unlikely given the low value of adding information and the fact that these players are rarely featured on Wikipedia to begin with. In this framework, the value of out-of-copyright information is highest for players of middling quality: players who are good enough to merit encyclopedic inclusion, but for whom information was relatively difficult to source before digitization. Therefore, I hypothesize that out-of-copyright digital information is most valuable for player-pages at the middle tier of the quality distribution, rather than for players at the very top. Given that the lowest-quality players are unlikely to even be covered on Wikipedia, within the sample of about 500 players who have been nominated for the Hall of Fame, the bottom two quartiles are likely to represent this "middle tier."

This prediction is consistent with, and contributes to, a number of different papers that investigate the implications of reduced costs of access to information due to digitization. For example, the reduced cost of music due to digital distribution most benefited less well-known musicians (Mortimer et al. 2012), the digital availability of retail items benefits the sales of products in the "long tail" (Brynjolfsson et al. 2011), and reductions in the cost of communication due to BITNET "democratized" innovation by benefiting collaboration among medium-ranked universities (Agrawal and Goldfarb 2008). Similarly, the innovation literature has found that the reduced cost of access to scientific material through the establishment of scientific institutions benefits countries in the developing world the most, where such access is harder to obtain (Furman and Stern 2011).

4.2. Comparing the Reuse of Images vs. Text

To test the hypothesis that copyright status is likely to have a bigger impact on the reuse of images than of text, I first recreate the simple descriptive analysis from Figure 3 separately for image and text citations using data from Sample A. This chart is displayed as Figure 4.

Figure 4. Impact of Copyright on Image and Text Citations (Sample A)



Notes. This figure plots the growth in citations to *Baseball Digest* publication-years in 2012 compared to 2008. Panel (1) plots the growth in image citations. Panel (2) plots the growth in text citations.

As before, for each publication-year of *Baseball Digest*, I measure the total number of new (text or image) citations that were made between the 2008 and 2012 page versions on Wikipedia.

As is evident from this chart, the patterns for text and image citations differ dramatically across the 1964 copyright cutoff. For images, there are hardly any citations at all from issues published in or after 1964. In other words, the likelihood that an image will be reused from a post-1964 issue of *Baseball Digest* is very close to zero, even after digitization. However, the pattern for text citations is quite different. In this case, there are a significant number of citations both before and after the 1964 cutoff; copyright status seems to have little impact. As hypothesized, the descriptive analysis suggests that copyright law seems to influence the reuse of digitized material mostly by preventing the reuse of images rather than text.

The basic intuition of the descriptive analysis can also be tested in a regression framework. I estimate difference-in-difference specifications similar to the baseline specification used for Table 3. The main outcome variables are citations to images and text. The estimates from this analysis are presented in Table 5. The estimates uphold the prediction that the 1964 copyright cutoff has a more significant impact on the reuse of images than of text, as out-of-copyright publication-years see on average 5.4 more image citations than in-copyright publication-years; meanwhile, there is no significant effect on text citations. This conclusion is justified even when Log-OLS models are considered.[11]

Finally, similar to Figure 2, I test the validity of these estimates by plotting time-varying coefficients separately for image and text citations. These are represented in Figure 5. As panel (1) indicates, the difference in the reuse of images for in-copyright and out-of-copyright issues is close to zero before the digitization event. However, after 2008, there was an immediate increase in the difference, and by 2012, out-of-copyright publication-years had significantly higher levels of image reuse. However, as panel (2) indicates, this pattern does not hold for text citations. The difference in text citations before digitization is close to zero and constant, and this pattern does not change significantly after 2008. In-copyright and out-of-copyright text citations track each other pretty closely, suggesting that copyright has very little impact on preventing the reuse of digitized textual material.

4.3. Comparing Differential Effects Across Players

I now turn to a test of the differential impact of copyright law on pages for players of different quality by leveraging the data in Sample B.

To examine the heterogeneous impact of the copyright experiment by player quality, I estimate the following specification:

$$Y_{it} = \alpha + \beta_1 \times Post_t \times Out\text{-}of\text{-}Copy_i + \sum_{m=2}^{4} \hat{\beta}_m \times Post_t \times 1(Quality_i = m) + \sum_{n=2}^{4} \hat{\beta}_n \times Post_t \times Out\text{-}of\text{-}Copy_i \times 1(Quality_i = n) + \gamma_i + \delta_t + \varepsilon_{it},$$

where γ_i and δ_t indicate player-page and time fixed effects, respectively. The key indicator variable, $Quality_i$, is a categorical variable that indicates the "quality" percentile rank of a player as a number between 1 and 4 (top 25th percentile, 25–50th percentile, 50–75th percentile, and bottom 25th percentile). The key coefficients of interest, $\hat{\beta}_n$, estimate the difference between the $Out\text{-}of\text{-}Copy_i \times Post_t$ coefficient and $Out\text{-}of\text{-}Copy_i \times Post_t \times 1(Quality = n)$ coefficient for each quality percentile n. In other words, $\hat{\beta}_n$ provides estimates of the differences in the impact of copyright on reuse for players of different quality levels.

Regression estimates from the specification listed above are presented in Table A.8 in the online appendix. However, to interpret the results, I focus on Figure 6, which plots these coefficients separately for each quality percentile.[12] Panel (1) validates the hypothesis that the impact of copyright on the reuse of images is larger

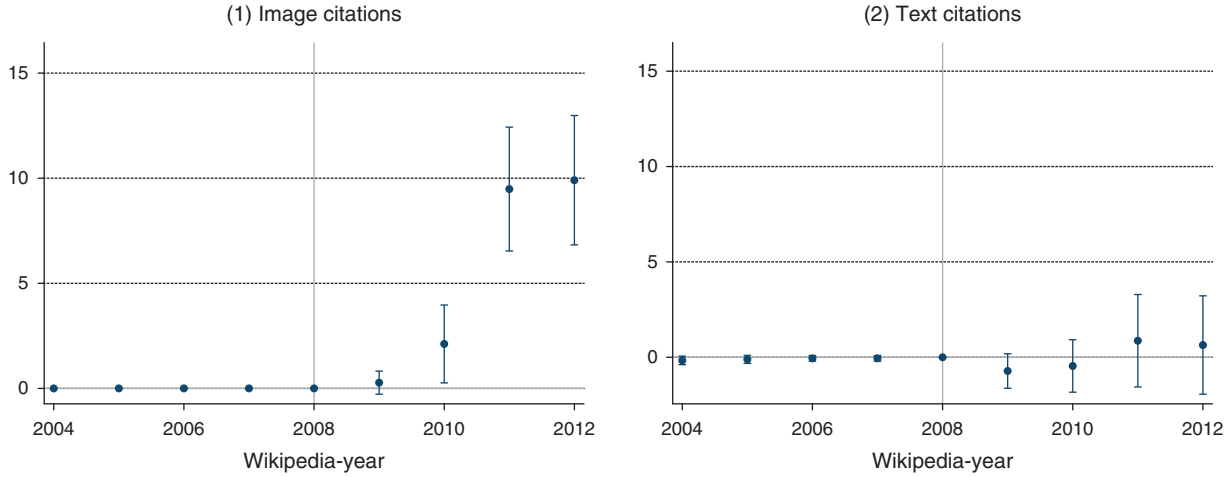
Table 5. Differential Impact of 1964 Copyright Experiment on Image vs. Text Citations (Sample A)

	Images			Text		
	OLS	OLS	Log-OLS	OLS	OLS	Log-OLS
Out-of-Copy × *Post*	5.444	5.444	1.173	0.151	0.161	−0.0203
	(1.094)***	(1.094)***	(0.151)***	(1.004)	(0.997)	(0.145)
Publication-year FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.413	0.416	0.574	0.738	0.811	0.911
N	360	360	360	360	360	360

Notes. This regression estimates the impact of the 1964 copyright exception on the reuse of images and text from *Baseball Digest* before and after digitization in a difference-in-difference framework. The estimates presented use data from Sample A. *Post* refers to all Wikipedia-years after 2008, and *Out-of-Copy* refers to *publication-year* < 1964. Clustered standard errors are shown in parentheses.

***$p < 0.01$.

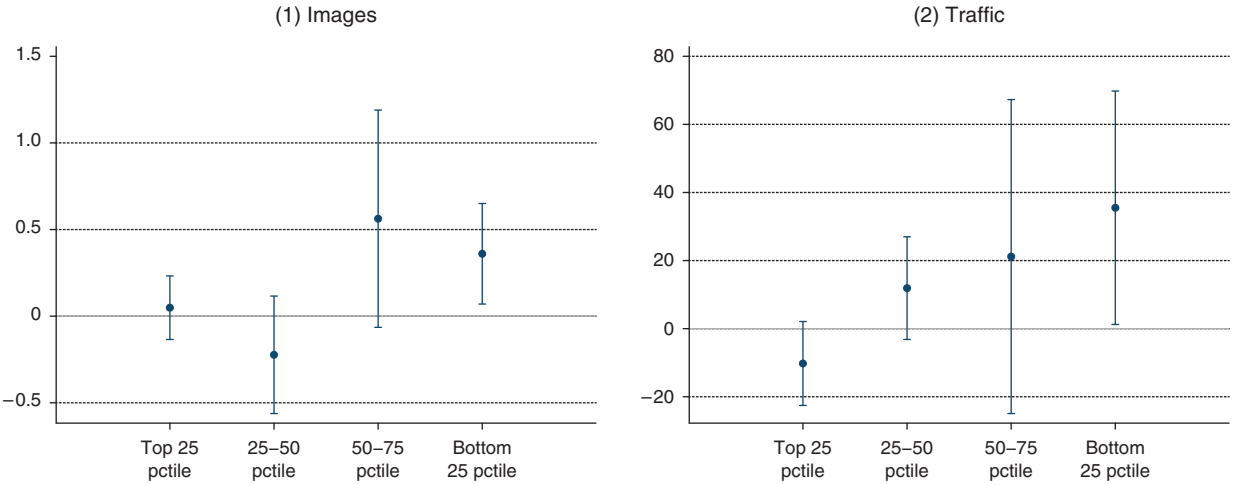
Figure 5. (Color online) Time-Varying Estimates for Image and Text Citations (Sample A)



Notes. This figure plots coefficients and 95% confidence intervals from the event study specifications described in Section 3.3.1 separately for image citations (panel (1)) and text citations (panel (2)). The reference year is 2008, the year of the digitization event. Coefficients are estimated from OLS models, standard errors are clustered, and the dependent variable in both panels is the total number of citations in a calendar year.

for players of lower quality than for players of higher quality. The estimate on the $\hat{\beta}_{n=3}$ coefficient is 0.71 and $\hat{\beta}_{n=4}$ is 0.47, although the first of these two estimates is marginally insignificant. In contrast, the coefficients on $\hat{\beta}_{n=1}$ and $\hat{\beta}_{n=2}$ are both statistically indistinguishable and close to zero. Panel (2), which estimates the impact of copyright on traffic to affected pages, shows a similar pattern, indicating that the impact of copyright on increasing traffic is most relevant for players in the lower tier of player quality. To confirm that these findings are robust, I replicate this analysis using an alternate measure of player quality by ranking players according to the number of years they played Major League Baseball (better players typically have longer careers), as shown in Figure A.1 in the online appendix. This analysis provides further support to the findings from Figure 6. However, it seems that the traffic results are supported only for players in the bottom quartile of the quality distribution. Finally, it is important to state that the estimates are not able to establish conclusively that the effects for players in the second and third quartile of quality are statistically different than the effects for players in the bottom quartile of player quality. More specifically, Table A.8 in the online appendix suggests that these differences are likely to be more robust for the results on page traffic than for image reuse. Despite this

Figure 6. (Color online) Heterogeneous Impact of Copyright on Wikipedia Pages by Player Quality (Sample B)



Notes. This plot documents the differential impact of the *Baseball Digest* copyright cutoff on baseball player-pages of different *quality* as described in Section 4.3. For this analysis, players are split into four different levels of quality based on their percentile rank within the sample of baseball players, and the main difference-in-difference estimates are calculated separately for each of the four quality percentiles. Panel (1) plots these estimates for image citations, whereas panel (2) plots estimates for traffic.

caveat, this analysis suggests that an important channel through which the digitization of *Baseball Digest* proved useful to Wikipedia was through the unlocking of unique material about famous but not superstar players on Wikipedia.

Sections 4.2 and 4.3 provide strong evidence for the proposition that copyright restrictions have important distributional implications that are especially relevant for images as compared to text, and they are particularly harmful for less popular topics for which alternate information is hard to find.

5. Discussion

> "Copyright is out of control. How – even if it's out of control, how does it stifle invention? [. . .] Anybody can make a movie, and the fact that that movie has a copyright, how does that hurt the Internet, for God's sake?"
>
> —Jack Valenti
> (USC Annenberg Norman Lear Center 2001)

This paper builds an empirical framework to answer Valenti's question, suggesting a mechanism through which copyright might affect the benefits of digitization: by prohibiting the reuse of digitized material, particularly within open, community-based innovation projects like Wikipedia. The paper makes three major findings. First, the digitization of *Baseball Digest* by Google Books had a positive impact on the reuse of material within Wikipedia, but these gains were much larger for out-of-copyright issues printed before 1964 than for in-copyright issues printed in or after 1964. Second, restricted reuse due to copyright had real effects on Wikipedia: affected pages experienced about a 20% drop in traffic. Finally, the impact of copyright on reuse was uneven—it affected the reuse of images while textual material was unaffected, and out-of-copyright material was most helpful for developing less well-known players' Wikipedia pages compared to more well-known ones.

5.1. External Validity

One concern with the results could be the lack of external validity. Specifically, one might be concerned that *Baseball Digest* specifically and Wikipedia generally represents an idiosyncratic setting in which to analyze the impact of copyright on reuse. To alleviate these concerns, I show that Wikipedia's practices when it comes to copyright protection appear to be shared by major commercial online firms. I also provide anecdotal evidence suggesting that reuse of other books and periodicals on Wikipedia is contingent on copyright status. Finally, I compare my results to similar findings in the literature.

One concern over external validity might stem from Wikipedia's status as a nonprofit: because it makes no money, it might have less to gain from reusing copyrighted work than commercial firms do, as they might find a monetary benefit to flouting copyright law. This would cause my estimates to be biased upward. This is a valid concern. However, Wikipedia's approach to copyright protection and copyright licensing appear to follow standard practices shared by commercial online firms. A number of other major commercial digital platforms, where one might expect reuse of digitized information to occur for profit, also have extensive programs for copyright enforcement. These include YouTube (Seidenberg 2009), Amazon, all major mobile application stores, and even Google's search engine.[13] For instance, Apple's App Store rejected about 1,000 applications in August 2009 because they used copyrighted images and books in their applications (see Ritchie 2009). Further examples can be found in the online appendix. Furthermore, similar to for-profit entities, Wikipedia appears to go to the trouble to properly license copyrighted content. For example, online volunteers are known to negotiate for licenses by leveraging Wikipedia's General Counsel, which acts similarly to a company's legal counsel.[14]

It is also worth noting that *Baseball Digest* is not the only instance in which copyright status affected the reuse of content on Wikipedia. Preliminary research that I have done indicates that a large portion of the anatomical images on Wikipedia are sourced from a 1918 edition of *Gray's Anatomy*,[15] rather than from a modern version, presumably because of copyright restrictions. Similarly, *Time* magazine images from before 1964 seem to also have lapsed into the public domain due to copyright nonrenewal; therefore, a large number of images from *Time* magazine from before 1964 find reuse on Wikipedia.[16]

Finally, the results I find here are compatible with the emerging empirical literature on the effects of copyright, which suggests that copyright has a negative effect on access, a precondition for any reuse to occur. Extant work (Heald 2008, 2009b, Buccafusco and Heald 2012) has shown that works produced before 1923, which are generally in the public domain, are much more accessible today, both in print and online, than works produced after 1923. A more recent study in the economics literature (Reimers 2017) analyzes the market for books in a similar time period and finds that copyright extensions decrease welfare from fiction bestsellers by decreasing variety, thereby causing a decrease in consumer surplus that outweighs the increase in profits.

In light of the anecdotes presented here and the recent empirical literature, it does seem plausible that the impact of copyright on Wikipedia that is measured in this paper could generalize to a number of other settings where the reuse of digital information is important. Finally, even if external validity is a concern, Wikipedia's prominence as a source of information

means that the potential gains from the reuse of out-of-copyright material remain significant for the digital economy.

5.2. Contributions and Managerial Implications

This paper makes a number of contributions to the empirical literature on digitization, a primary goal of which is to analyze the economic consequences of digital information. Previous scholarship on important economic activities such as consumer search and pricing has generally found that digitization reduces the cost of accessing information, which can often have beneficial implications for consumer welfare. This paper adds to this literature by considering the impact of IP restrictions, arguing that in settings where the ability of intermediaries to reuse information is important, copyright law might have important implications for the economic effects of digitization.

Furthermore, some recent work has suggested that the digitization process could influence the distribution of economic outcomes disproportionately in favor of smaller market participants. For example, it has been found that file sharing increases live performance revenues for less well-known artists, perhaps through increased awareness, but performance revenues for large, well-known artists are unaffected (Mortimer et al. 2012). Similarly, counterfeiting has been shown to have a larger advertisement effect for brands that were less well known at the time of infringement (Qian 2014).[17] My findings follow a similar intuition: when access costs are reduced through digitization and public domain status, less well-known topics benefit disproportionately. However, the presence of copyright could prevent this process.

My results are also related to scholarship on the empirical effects of IP on the diffusion of knowledge (Murray and Stern 2007, Murray et al. 2009, Williams 2013, Furman and Stern 2011, Sampat and Williams 2015, Galasso and Schankerman 2015). This study provides direct evidence that the costs of access (i.e., digitization) seem to matter for the impact of IP on reuse. From a policy point of view, this paper is able to address questions that are likely to be important going forward, such as: (a) How does the impact of copyright change when works are digitized and access costs are low? (b) Does copyright need to be modified for the digital age?

Finally, this study also has implications for managers in knowledge-intensive sectors of the economy. For those in charge of IP and digitization strategy, this study suggests that copyright can be an effective tool to manage digital assets. How effective copyright can be seems to depend on access and the medium in which information is expressed. This is a useful counter to the concern that piracy is so rampant on the Internet that additional tools (such as digital rights management (DRM)) are necessary to supplement toothless copyright law (Zhang 2016). Second, for managers who are interested in using user communities like Wikipedia to generate innovation (Boudreau et al. 2011, Franzoni and Sauermann 2014), this study suggests that the provision of external, uncopyrighted but digitized material can be extremely beneficial. From a policy point of view, measures that affect the availability and legal status of sources can either boost or retard innovative activity within online communities.

5.3. Limitations and Welfare Calculation

This paper does not evaluate the overall welfare consequences of the impact of copyright on digital information, but it does help to make progress in that direction. In a static setting, where new digital information does not build on preexisting work, stronger copyright law should incentivize the production of digital information (Watt and Towse 2006). However, in a more dynamic setting, where the production of new knowledge depends on preexisting information (e.g., the presence of Google Books helps the production of new knowledge on Wikipedia), whether stronger copyright will boost knowledge production is unclear (Scotchmer 1991). If transaction costs imposed by copyright prevent the reuse of existing work, then optimal copyright policy should provide for weaker IP than it would in the absence of such transaction costs. In this context, it becomes critical to obtain credible empirical measurements of the cost of copyright on preventing reuse of digital information. Without such measurements, we do not know, "whether copyright protection would need to be strengthened or weakened" in the digital age (Waldfogel 2012, p. 340). This paper helps to fill this gap.

Despite this contribution, there are other aspects of the welfare calculation that this paper does not address. In particular, if copyrights allow the publishers of *Baseball Digest* to profit from archival material and help them generate new combinations of preexisting work, then a weakening of copyright will hurt overall knowledge creation as well. In such a case, overall welfare gains from the removal of copyright protection for archival *Baseball Digest* issues could be small, especially if licensing archival content is a major source of revenue that is hurt by lost copyright protection.[18] However, I am not able to directly estimate the extent to which lost copyright would depress incentives for the production of new knowledge by the publisher.

Finally, notwithstanding this limitation, this study is especially useful in cases where issues of copyright policy arise for works already created. In these cases, the argument for extending copyright relies on the assumption that copyright on existing works furthers the diffusion of information. Such an argument was a feature of the "Mickey Mouse" law of 1998.[19] Even if

it is indeed possible that copyright provides an incentive for the creation of new material, the estimates in this paper show that there are attendant welfare losses from retroactive extensions of copyright.

Acknowledgments

The author thanks Pierre Azoulay, Scott Stern, Catherine Tucker, and Heidi Williams for their guidance throughout this project. The author is grateful to WikiProject Baseball contributors, including Delaywaves and Wizardman for useful feedback on this project. The author also thanks Erik Brynjolfsson, Daniel Fehder, Lee Fleming, Jeff Furman, Avi Goldfarb, Shane Greenstein, Paul Heald, Joshua Krieger, Matt Marx, Aruna Ranganathan, Fabian Waldinger, Joel Waldfogel, Pai-Ling Yin, three anonymous referees, and an anonymous associate editor at *Management Science*, as well as participants of the Economic Sociology Working Group, the Munich Summer Institute, the Northwestern Searle Roundtable on the Law and Economics of Digital Markets, and the Massachusetts Institute of Technology Economics Third Year Lunch. All errors remain the author's own.

Endnotes

[1] https://reportcard.wmflabs.org (accessed July 1, 2017).

[2] The Wikipedia page on "baseball" receives about 100,000 page views per month (see https://tools.wmflabs.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&start=2016-07-01&end=2016-07-31&pages=Baseball, accessed July 1, 2017).

[3] http://onlinebooks.library.upenn.edu/cce/firstperiod.html (accessed July 1, 2017).

[4] I chose December 1 because this is a few days before the digitization event happened on December 9, 2008.

[5] I choose these years because they form a window of 20 years before and after the copyright cutoff year of 1964. In theory it would be possible to perform the analysis in this paper with data extending beyond 1984.

[6] Specifically, the dependent variable is $\text{Log}(Cites_{it}+1)$.

[7] One reason for the lower Log-OLS estimates compared to the OLS estimates could be the large number of zeros in the outcome variable, making the $\text{Log}(Cites_{it}+1)$ variable quite consequential.

[8] While the main outcome variable in this case does not count citations to in-copyright and out-of-copyright issues separately, it is very likely that out-of-copyright player-pages make citations to out-of-copyright issues and vice versa.

[9] I would like to thank a referee for this idea.

[10] Traffic information is calculated as a monthly average for years 2007–2012 (data are not available before this period) and is recorded at the player-page level.

[11] I repeat this analysis using Sample B in Table A.6 of the online appendix. I find positive and large estimates of the impact of out-of-copyright status on the reuse for images, while for text, the estimates are positive in OLS models and negative in Log models, and their magnitude is economically insignificant. See notes in Table A.6 for more discussion.

[12] More precisely, I plot the coefficient on $Out\text{-}of\text{-}Copy \times Post_t$ for quality = 1 and add this estimate to the coefficients for other quality levels to compute marginal effects.

[13] https://www.google.com/transparencyreport/removals/copyright/explore/ (accessed July 1, 2017).

[14] See https://wikimediafoundation.org/wiki/User:GeoffBrigham_%28WMF%29 (accessed July 1, 2017).

[15] See https://commons.wikimedia.org/wiki/Category:Gray's_Anatomy_plates (accessed July 1, 2017) for a listing of these images.

[16] For example, https://commons.wikimedia.org/wiki/File:Shidehara_Kijuro_on_TIME_magazine_cover.jpg (accessed July 1, 2017).

[17] I thank a reviewer for pointing me toward this research.

[18] I did make a number of reasonable attempts to contact the publishers of *Baseball Digest* to investigate the possibility of licensing content for reuse, but my requests were met with no response. This suggests that, in this case at least, producer surplus from licensing archival material is fairly low.

[19] Sonny Bono Copyright Term Extension Act, Pub. L. No. 105-298, 112 Stat. 2827 (1998).

References

Agrawal A, Goldfarb A (2008) Restructuring research: Communication costs and the democratization of university innovation. *Amer. Econom. Rev.* 98(4):1578–1590.

Aguiar L, Waldfogel J (2014) Digitization, copyright, and the welfare effects of music trade. December 3, https://ssrn.com/abstract=2603238.

Anderson N (2007) New copyright alliance hopes to strengthen copyright law. *Ars Technica* (May 18), https://arstechnica.com/tech-policy/2007/05/new-copyright-alliance-hopes-to-strengthen-copyright-law/.

Bakos JY (1997) Reducing buyer search costs: Implications for electronic marketplaces. *Management Sci.* 43(12):1676–1692.

Benkler Y (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale University Press, New Haven, CT).

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *Quart. J. Econom.* 119(1):249–275.

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Sci.* 57(5):843–863.

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye Pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Sci.* 57(8):1373–1386.

Brynjolfsson E, Hu YJ, Smith MD (2006) From niches to riches: The anatomy of the long tail. *Sloan Management Rev.* 47(4):67–71.

Buccafusco C, Heald PJ (2012) Do bad things happen when works enter the public domain?: Empirical tests of copyright term extension. *Berkeley Tech. Law J.* 28(1), http://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=1972&context=btlj.

Chiou L, Tucker C (2017) Digital content aggregation platforms: The case of the news media. *J. Econom. Management Strategy.* Forthcoming.

Danaher B, Dhanasobhon S, Smith MD, Telang R (2010) Converting pirates without cannibalizing purchasers: The impact of digital distribution on physical sales and Internet piracy. *Marketing Sci.* 29(6):1138–1151.

Foulser D (2008) Search and find magazines on Google Book Search. *Google Blog* (December 9), https://googleblog.blogspot.com/2008/12/search-and-find-magazines-on-google.html.

Franzoni C, Sauermann H (2014) Crowd science: The organization of scientific research in open collaborative projects. *Res. Policy* 43(1):1–20.

Furman J, Stern S (2011) Climbing atop the shoulders of giants: The impact of institutions on cumulative knowledge production. *Amer. Econom. Rev.* 101(5):1933–1963.

Galasso A, Schankerman M (2015) Patents and cumulative innovation: Causal evidence from the courts. *Quart. J. Econom.* 130(1):317–369.

Gallini N, Scotchmer S (2002) Intellectual property: When is it the best incentive system? *Innovation Policy and the Economy*, Vol. 2 (MIT Press, Cambridge, MA), 51–78.

Gans JS (2015) Remix rights and negotiations over the use of copy-protected works. *Internat. J. Indust. Organ.* 41(July):76–83.

Gans JS, Stern S (2003) The product market and the market for ideas: Commercialization strategies for technology entrepreneurs. *Res. Policy* 32(2):333–350.

Goldfarb A, Greenstein SM, Tucker CE (2014) Introduction. *Economic Analysis of the Digital Economy* (University of Chicago Press, Chicago), 1–17.

Goldfarb A, McDevitt RC, Samila S, Silverman B (2015) The effect of social interaction on economic transactions: Evidence from changes in two retail formats. *Management Sci.* 61(12):2963–2981.

Gorbatai A (2014) The paradox of novice contributions in collective production: Evidence from Wikipedia. February 10, https://ssrn.com/abstract=1949327.

Greenstein S, Zhu F (2017) Do experts or collective intelligence write with more bias? Evidence from Encyclopaedia Britannica and Wikipedia. *MIS Quart.* Forthcoming.

Greenstein S, Lerner J, Stern S (2010) The economics of digitization: An agenda for NSF. American Economic Association, Ten years and beyond: Economists answer NSF's call for long-term research agendas. https://ssrn.com/abstract=1889153.

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

Heald PJ (2008) Property rights and the efficient exploitation of copyrighted works: An empirical analysis of public domain and copyrighted fiction best sellers. *Minnesota Law Rev.* 92:1031–1063.

Heald PJ (2009a) Testing the over- and under-exploitation hypotheses: Bestselling musical compositions (1913–32) and their use in cinema (1968–2007). *Rev. Econom. Res. Copyright* 6:31–60.

Heald PJ (2009b) Does the song remain the same—An empirical study of bestselling musical compositions (1913–1932) and their use in cinema (1968–2007). *Case Western Reserve Law Rev.* 60:1.

Kitch EW (1977) Nature and function of the patent system. *J. Law Econom.* 20(2):265–290.

Kupferman TR (1944) Renewal of Copyright. Section 23 of the Copyright Act of 1909. *Columbia Law Rev.* 44(5):712–735.

Landes WM, Posner RA (2003) Indefinitely renewable copyright. *Univ. Chicago Law Rev.* 70(2):471–518.

Lemley MA (2004) Ex ante versus ex post justifications for intellectual property. *Univ. Chicago Law Rev.* 71:129–149.

Lessig L (2004) *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity* (Penguin Books, New York).

Lessig L (2005) *Free Culture: The Nature and Future of Creativity* (Penguin Books, New York).

Leval PN (1990) Toward a fair use standard. *Harvard Law Rev.* 103(5):1105–1136.

Li X, MacGarvie M, Moser P (2017) Dead poets' property—How does copyright influence price? *RAND J. Econom.* Forthcoming.

Luo H, Mortimer JH (2016) Copyright enforcement: Evidence from two field experiments. *J. Econom. Management Strategy* 26(2): 499–528.

Mazzoleni R, Nelson RR (1998) Economic theories about the benefits and costs of patents. *J. Econom. Issues* 32(4):1031–1052.

Merges RP, Menell PS, Lemley MA (2013) *Intellectual Property in the New Technological Age* (Aspen Publishers, New York).

Miller AR, Tucker CE (2011) Can health care information technology save babies? *J. Political Econom.* 119(2):289–324.

Mortimer JH (2007) Price discrimination, copyright law, and technological innovation: Evidence from the introduction of DVDs. *Quart. J. Econom.* 122(3):1307–1350.

Mortimer JH, Nosko C, Sorensen A (2012) Supply responses to digital distribution: Recorded music and live performances. *Inform. Econom. Policy* 24(1):3–14.

Murray F, Stern S (2007) Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *J. Econom. Behav. Organ.* 63(4): 648–687.

Murray F, Aghion P, Dewatripont M, Kolev J, Stern S (2009) Of mice and academics: Examining the effect of openness on innovation. Technical report, National Bureau of Economic Research, Cambridge, MA.

Nagaraj A (2017) The private impact of public maps—Landsat satellite imagery and gold exploration. Working paper, University of California, Berkeley, Berkeley.

Prat A, Strömberg D (2013) The political economy of mass media. Acemoglu D, Arellano M, Dekel E, eds. *Advances in Economics and Econometrics: Tenth World Congress, Volume II, Applied Economics* (Cambridge University Press, New York), 135–187.

Public Domain Sherpa (2014) Copyright renewal: When it had to happen, or else. Last accessed July 1, 2017, http://www.publicdomainsherpa.com/copyright-renewal.html.

Qian Y (2014) Counterfeiters: Foes or friends? How counterfeits affect sales by product quality tier. *Management Sci.* 60(10):2381–2400.

Reimers I (2017) Copyright and generic entry in book publishing. Working paper, Northeastern University, Boston.

Ritchie R (2009) App Store cracks down on copyright, ejects 900+ aggregator apps, rejects e-books. *iMore* (August 6), https://www.imore.com/app-store-cracks-copyright-ejects-900-aggregator-apps-rejects-ebooks.

Rob R, Waldfogel J (2007) Piracy on the silver screen. *J. Indust. Econom.* 55(3):379–395.

Sampat B, Williams H (2015) How do patents affect follow-on innovation? Evidence from the human genome. NBER Working Paper 21666, National Bureau of Economic Research, Cambridge, MA.

Samuelson P (1999) Intellectual property and the digital economy: Why the anti-circumvention regulations need to be revised. *Berkeley Tech. Law J.* 14(2):519–566.

Samuelson P (2009) Google book search and the future of books in cyberspace. *Minn. L. Rev.* 94:1308–1374.

Scotchmer S (1991) Standing on the shoulders of giants: Cumulative research and the patent law. *J. Econom. Perspect.* 5(1):29–41.

Seidenberg S (2009) Copyright in the age of YouTube. *Amer. Bar Assoc. J.* 95:46.

Shapiro C, Varian HR (1998) *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business Review Press, Brighton, MA).

Silverwood Cope S (2012) Wikipedia: Page one of Google UK for 99% of searches. *Pi Datametrics* (February 8), https://www.pi-datametrics.com/wikipedia-page-one-of-Google-uk-for-99-of-searches/.

Strömberg D (2007) Natural disasters, economic development, and humanitarian aid. *J. Econom. Perspect.* 21(3):199–222.

USC Annenberg Norman Lear Center (2001) A debate on "creativity, commerce & culture" with Larry Lessig and Jack Valenti, November 29, https://learcenter.org/pdf/LessigValenti.pdf.

Waldfogel J (2012) Copyright research in the digital age: Moving from piracy to the supply of new products. *Amer. Econom. Rev.* 102(3):337–342.

Waldfogel J (2014) Digitization and the quality of new media products: The case of music. Goldfarb A, Greenstein SM, Tucker CE, eds. *Economic Analysis of the Digital Economy* (University of Chicago Press, Chicago), 407–442.

Watt R, Towse R (2006) copyright protection standards and authors' time allocation. *Indust. Corporate Change* 15(6):995–1011.

Williams H (2013) Intellectual property rights and innovation: Evidence from the human genome. *J. Political Econom.* 121(1):1–27.

Wu T (2015) What ever happened to Google Books? *The New Yorker* (September 11), http://www.newyorker.com/business/currency/what-ever-happened-to-google-books.

Zhang L (2016) Intellectual property strategy and the long tail: Evidence from the recorded music industry. *Management Sci.*, ePub ahead of print November 11, https://doi.org/10.1287/mnsc.2016.2562.

Zittrain J (2009) *The Future of the Internet—And How to Stop It* (Yale University Press, New Haven, CT).

A Online Appendices: *Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia*

A.1 Appendix A1 : Robustness Checks

Table A.1. **Estimating the Causal Impact of Digitization**

	Digitization DD		
	Citations	Images	Text
baseball X post	0.340	0.459	0.391
	(0.0494)***	(0.0610)***	(0.0650)***
Player FE	Yes	Yes	Yes
Time FE	Year	Year	Year
adj. R^2	0.0687	0.172	0.399
N	13260	13260	13260
Clusters	1105	1105	1105

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: This table provides estimates that help to determine the causal impact of the Google Books digitization event on reuse. I supplement data in Sample B, with similar data from Wikipedia player-pages for a comparable set of 564 basketball players. The estimates are provided from a difference-in-difference specification where the treatment group is the set of baseball player-pages and the post-period are the years 2009-2012 after the digitization event. All estimates are from ordinary-least-squares (OLS) models.

Table A.2. **Robustness: Exploring pre-trends between in-copyright and out-of-copyright Issues**

	Sample A			Sample B		
	Citations	Images	Text	Citations	Images	Text
$Digitization_{-3}$	-0.048	-0.000	-0.048	-0.008	-0.194	-0.642
	(0.048)	(0.000)	(0.048)	(0.004)*	(0.024)***	(0.031)***
$Digitization_{-2}$	-0.048	-0.000	-0.048	-0.008	-0.075	-0.359
	(0.048)	(0.000)	(0.048)	(0.004)**	(0.027)***	(0.022)***
$Digitization_{-1}$	-0.048	-0.000	-0.048	-0.003	-0.033	-0.106
	(0.048)	(0.000)	(0.048)	(0.003)	(0.019)*	(0.009)***
$Digitization_{+1}$	1.762	0.095	1.667	0.046	0.060	0.088
	(0.418)***	(0.066)	(0.415)***	(0.015)***	(0.016)***	(0.007)***
$Digitization_{+2}$	4.762	0.095	4.667	0.112	0.140	0.220
	(0.654)***	(0.066)	(0.656)***	(0.027)***	(0.020)***	(0.012)***
$Digitization_{+3}$	9.333	0.095	9.238	0.128	0.220	0.397
	(1.066)***	(0.066)	(1.060)***	(0.028)***	(0.026)***	(0.017)***
$Digitization_{+4}$	10.190	0.095	10.095	0.132	0.259	0.472
	(1.185)***	(0.066)	(1.178)***	(0.028)***	(0.026)***	(0.020)***
$Digitization_{-3}$ x out-of-copy	-0.110	0.000	-0.110	-0.040	-0.093	-0.180
	(0.126)	(0.000)	(0.126)	(0.027)+	(0.064)+	(0.084)**
$Digitization_{-2}$ x out-of-copy	-0.058	0.000	-0.058	-0.037	-0.013	-0.127
	(0.087)	(0.000)	(0.087)	(0.027)	(0.062)	(0.067)*
$Digitization_{-1}$ x out-of-copy	-0.058	0.000	-0.058	-0.035	0.033	-0.114
	(0.087)	(0.000)	(0.087)	(0.021)*	(0.052)	(0.050)**
$Digitization_{+1}$ x out-of-copy	-0.446	0.273	-0.719	0.077	-0.038	0.050
	(0.753)	(0.327)	(0.537)	(0.045)*	(0.042)	(0.024)**
$Digitization_{+2}$ x out-of-copy	1.659	2.115	-0.456	0.160	-0.013	0.124
	(1.437)	(1.100)*	(0.816)	(0.068)**	(0.055)	(0.039)***
$Digitization_{+3}$ x out-of-copy	10.351	9.484	0.867	0.187	0.383	0.101
	(2.754)***	(1.748)***	(1.438)	(0.073)**	(0.083)***	(0.046)**
$Digitization_{+4}$ x out-of-copy	10.546	9.905	0.642	0.206	0.454	0.117
	(2.872)***	(1.826)***	(1.531)	(0.074)***	(0.086)***	(0.055)**
Player FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Year	Year	Year	Year	Year	Year
adj. R^2	0.757	0.573	0.810	0.043	0.134	0.365
N	360.000	360.000	360.000	9945.000	9945.000	9945.000

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Table A.3. **Falsification Check – Alternate Treatment Years**

	Sample A			Sample B		
	Cites	Cites	Log-Cites	Cites	Cites	Log-Cites
out-of-copy X post	-0.100	-0.0576	-0.0472	0.0605	0.0546	0.0163
	(0.307)	(0.269)	(0.0773)	(0.0319)*	(0.0392)	(0.0143)
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Page-Age FE	—	—	—	Yes	Yes	Yes
adj. R^2	0.245	0.315	0.538	0.0405	0.0502	0.0762
N	240	240	240	3246	3246	3246

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: This table presents a falsification check of the baseline specification. In this regression, the panel is restricted to years 2004 to 2009, and the treatment year is assumed to be 2007 rather than 2009. The $out-of-copy$ variable is defined as before, and unit-of-observation fixed effects and time fixed effects are included as indicated.

Table A.4. **Robustness Check : Adding Panel Restrictions**

(1) Wikipedia-Years 2005-2011

	Sample A			Sample B		
	Cites	Cites	Log-Cites	Cites	Cites	Log-Cites
out-of-copy X post	3.922	3.911	0.192	0.186	0.170	0.0605
	(1.458)**	(1.441)***	(0.153)	(0.0655)***	(0.101)*	(0.0342)*
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Page-Age FE	—	—	—	Yes	Yes	Yes
adj. R^2	0.626	0.682	0.884	0.0552	0.0781	0.106
N	280	280	280	3787	3787	3787

(2) Wikipedia-Years 2006-2010

	Sample A			Sample B		
	Cites	Cites	Log-Cites	Cites	Cites	Log-Cites
out-of-copy X post	0.674	0.645	-0.0595	0.165	0.144	0.0491
	(0.968)	(0.948)	(0.158)	(0.0744)**	(0.0914)	(0.0314)
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Page-Age FE	—	—	—	Yes	Yes	Yes
adj. R^2	0.480	0.572	0.821	0.0487	0.0678	0.0932
N	200	200	200	2705	2705	2705

*p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: This table presents robustness checks for the baseline specification to alternate panel restrictions. The specification is similar to the baseline specification and is estimated using OLS. However, instead of using the complete panel from 2004-2012, Panel (1) only includes data from years 2005-2011, and Panel (2) includes data from year 2006-2010. The $out-of-copy$ and *post* variables are defined as before, and unit-of-observation fixed effects and time fixed effects are included as indicated.

Table A.5. **Robustness to Sample Restrictions, Alternate Variables and Treatment Definition (Sample B)**

	(1)	(2)	(3)	(4)
Panel A: Citations				
out-of-copy X post	0.0359	0.0845	0.0434	0.0517
	(0.0429)	(0.0340)**	(0.0306)	(0.0253)**
Panel B : Images				
out-of-copy X post	0.570	0.717	0.203	0.00904
	(0.244)**	(0.166)***	(0.128)+	(0.0309)
Panel C : Text				
out-of-copy X post	0.238	0.509	0.261	1779.0
	(0.227)	(0.158)***	(0.121)**	(812.7)**
FE	Yes	Yes	Yes	Yes
Time FE	Year	Year	Year	Year
N	3438	4869	3663	4398
Adj R-square	0.421	0.406	0.417	0.360

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: This table evaluates the robustness of the impact of copyright on reuse result to different modeling and data assumptions. Column (1) drops all players who played both before and after the copyright-cutoff year of 1964 and estimates the model using data from players who either retired before 1964 and those who made their debut after 1964. Column (2) uses an alternate definition of $out-of-copy$ using the year of a player's first all star game instead of the debut year for classification. Column (3) drops very well-known players (those who have played 15 all star games or more) before estimating the model. Column (4) uses alternate dependent variables: Citations and Images are replaced by indicator variables if variable is greater that 0, and text is measured by the size of the page in kilobytes. All estimates are from ordinary-least-squares (OLS) models.

Table A.6. **Differential Impact of 1964 Copyright Experiment on Image vs. Text Citations (Sample B)**

	Images			Text		
	OLS	OLS	Log-OLS	OLS	OLS	Log-OLS
out-of-copy X post	0.477	0.337	0.109	0.466	0.314	-0.113
	(0.102)***	(0.0819)***	(0.0272)***	(0.120)***	(0.0877)***	(0.0490)**
Player-Page FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.0992	0.159	0.205	0.218	0.394	0.844
N	13260	13260	13260	13260	13260	13260

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Clustered standard errors shown in parentheses.

Note: This regression estimates the impact of the 1964 copyright exception on affecting the reuse of images and text from Baseball Digest before and after digitization in a differences-in-differences framework. The estimates presented use data from Sample B. *post* refers to all Wikipedia-years after 2008, and *out-of-copy* refers to $publication - year < 1964$. The estimates for images are large and significant relative to the mean. However, the estimates for text are less clear – being positive in OLS models and negative in LOG models. The magnitude of the estimates for Text reuse are also smaller for OLS estimates as compared the magnitude of estimates for Image reuse. I interpret this evidence as supporting the broad conclusion that out-of-copyright status is more beneficial for image reuse, as compared to text reuse.

Table A.7. **Impact of Copyright on Images and Traffic: Robustness with "Out-of-copyright" Exposure Index**

	(1) Diff. Img	(2) Log Diff. Img.	(3) Diff. Traf	(4) Log Diff. Traf
Out-of-copy Exposure	1.298	0.582	25.90	0.404
	(0.218)***	(0.0675)***	(11.35)**	(0.195)**
Constant	0.455	0.267	42.41	2.816
	(0.0435)***	(0.0220)***	(5.016)***	(0.0660)***
Observations	541	541	541	541
Adjusted R^2	0.130	0.146	0.006	0.008

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Robust standard errors shown in parentheses.

Note: This table provides a robustness check to log models for Table 4. Simple log versions of the models in Table 4 were tried, however a lack of sufficient "pre" data (before 2008) means that the main coefficients were imprecisely estimated, and were not significant at conventional levels. As an alternative, the following table estimates cross sectional regressions that utilize the variance in *copyright exposure* to estimate log models. For each player, *copyright exposure* is defined as amount of their career that they played in the out-of-copyright period, i.e. before 1964. For players who retired before 1964, this index is set to one, for players who made their debuts after 1964 this index is set to zero, while for other players it is calculated as $\frac{1964-DebutYear}{FinalYear-DebutYear}$. Because player debut and retirement years are unlikely to be related to the 1964 copyright cutoff date, this variation provides an additional source of quasi-random variation that can then be used in the cross-section to estimate the impact of copyright on internet traffic, and that helps alleviate the problem of missing traffic data for years before 2007. Columns (1) and (2) show the impact of the Copyright Exposure variable on the reuse of Images, while Columns (3) and (4) estimate the effect for traffic. Coefficients are roughly the same order of magnitude as with the difference-in-difference specifications.

This regression uses page-year level observations. Sample includes all baseball pages in 2012. The specification is $Y_i = \alpha + \beta \times out - of - copyindex + \varepsilon_i$. All estimates are from ordinary-least-squares (OLS) models, and columns (2) and (4) use $Log(1+Y)$ as the dependent variable.

Table A.8. **Heterogeneous Impacts of Copyright on Wikipedia Pages by Player Quality (Sample B)**

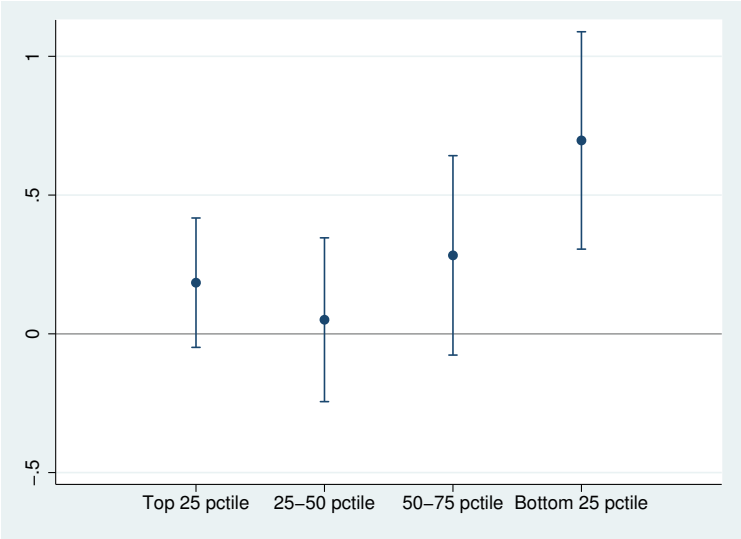
	Images	Traffic
post	0.851	49.39
	(0.0758)***	(5.556)***
out-of-copy X post	0.0489	-10.22
	(0.112)	(7.493)
post X quality=2	-0.00848	2.811
	(0.105)	(5.879)
post X quality=3	0.255	14.69
	(0.128)**	(8.212)*
post X quality=4	0.423	20.39
	(0.130)***	(15.81)
out-of-copy X post X quality=2	-0.272	22.13
	(0.234)	(11.84)*
out-of-copy X post X quality=3	0.514	31.40
	(0.397)	(29.01)
out-of-copy X post X quality=4	0.311	45.73
	(0.209)	(22.15)**
Unit of Obs. FE	Yes	Yes
Year FE	Yes	Yes
adj. R^2	0.213	0.0918
N	4869	3246

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Clustered standard errors shown in parentheses.

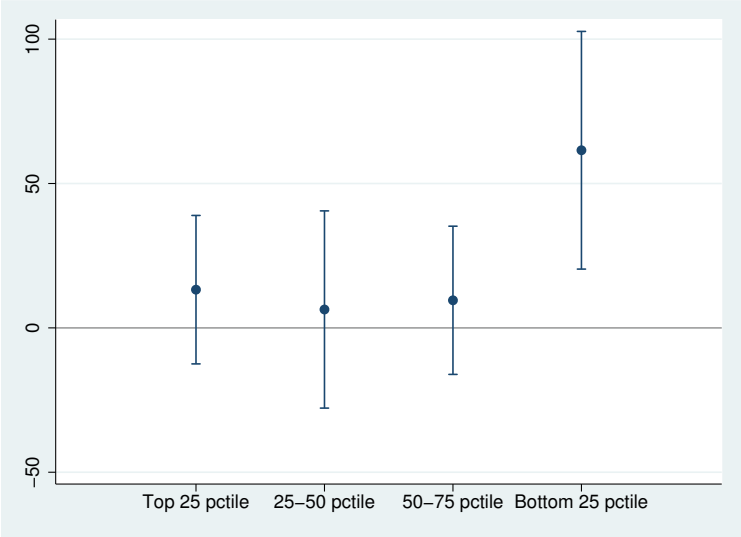
Note: This table presents estimates from the regression used to calculate the marginal effects presented in Figure 6. The estimates presented use data from Sample B. The $out-of-copy$ and *post* variables are defined as before, and unit-of-observation fixed effects and time fixed effects are included as indicated. Players are split into 4 different levels of quality based on their percentile rank within the sample of baseball players and the number of all-star games that they appeared in, and the main difference-in-difference estimates are calculated separately for each of the four quality percentiles. Column (1) plots these estimates for the reuse of images, while Column (2) plots estimates for traffic.

Figure A.1. **Heterogeneous Impacts of Copyright on Wikipedia Pages by Player Quality (Sample B)**

(1) Images



(2) Traffic

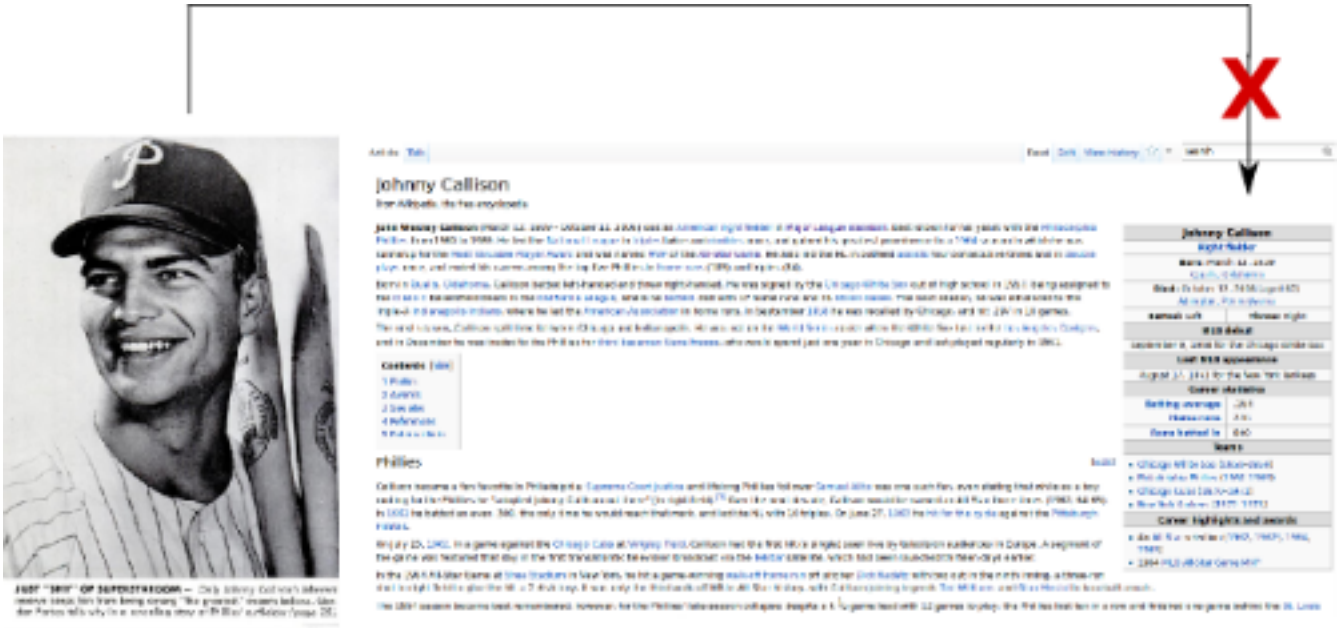


Note: This plot documents the differential impact of the Baseball Digest copyright cutoff on baseball player pages of different *quality* quartiles based on the number of games they have played in their career (as an additional robustness check). For this analysis, players are split into 4 different levels of quality based on their percentile rank within the sample of baseball players and the main difference-in-difference estimates are calculated separately for each of the four quality percentiles. Panel (1) plots these estimates for Image Citations, while Panel (2) plots estimates for Traffic.

Figure A.2. **An Illustration of How Copyright Might Affect the Reuse of Information**

(1) Felipe Alou's image in December 1963 (out-of-copyright) issue of Baseball Digest, reused on Wikipedia)

(2) Johnny Callison's image in January 1964 (in-copyright) issue of Baseball Digest, not reused on Wikipedia



A.2 Appendix A2 : Simple Theoretical Framework

This section builds a simple toy model to understand how copyright might affect the reuse of digitized information.

Setup

Consider a wikipedia page $W_{q,k}$ for an item of quality q and knowledge level k. The quality is a parameter that captures how inherently interesting a topic is, for example a famous, well-known baseball player will have higher q than a less well-known player. Knowledge level k captures how much information exists on a given page. Let $q \in \{0, \infty\}$ and $k \in \{1/4, \infty\}$.

Now define value, $V(W_{q,k}) = \sqrt{q} + \sqrt{k} - \frac{k}{4}$ to be the value that the Wikipedia community delivers from a page $W_{q,k}$. In a context like Wikipedia, V could be the traffic that a page receives for example. Note that while $\frac{dV}{dq} > 0$ and $\frac{dV}{dk} > 0$, $\frac{d^2V}{dq^2} < 0$ and $\frac{d^2V}{dk^2} < 0$. This simply implies diminishing but positive marginal returns from increased information and increased player quality to V.

Define $C(W_{q,k}) = \frac{k}{q}$ to be the cost of adding k units of information to a page with quality level q. Here, $\frac{dC}{dk} > 0$ implying higher costs of information acquistion for higher levels of knowledge, but $\frac{dC}{dq} < 0$, implying that it is easier to source information for higher quality topics, presumably because such information is more easily available.

Under this setup, the Wikipedia community solves the following, simple maximization problem to determine optimal levels of k, i.e. k^*

$$k^* = \max_k \left[V(W_{q,k}) - C(W_{q,k}) \right]$$

$$k^* = \max_k \left[\sqrt{q} + \sqrt{k} - \frac{k}{4} - k/q \right]$$

$$k^* = \frac{4q^2}{(q+4)^2}$$

Digitization and Copyright

Now consider that a digitization project makes it easier to access information to a certain topic, but that these reduction in costs depend on the copyright status of the underlying material. For topics that can benefit from out-of-copyright material, this reduction in cost is greater than it is for in-copyright material. A general way to parameterize this change is to assume that costs of adding information are reduced differentially for different copyright status groups.

Accordingly, let

$$C_{in-copy}(W_{q,k}) = \frac{C(W_{q,k})}{2} = \frac{k}{2q}$$

$$C_{out-of-copy}(W_{q,k}) = \frac{C(W_{q,k})}{4} = \frac{k}{4q}$$

Solving a similar maximization problem as before, we now obtain:

$$k^*_{in-copy} = \frac{4q^2}{(q+2)^2}$$

$$k^*_{out-of-copy} = \frac{4q^2}{(q+1)^2}$$

Therefore, $k^*_{out-of-copy} > k^*_{in-copy} > k^*$. This setup delivers the first two results that we obtained in the main part of the paper, i.e. digitization increased amount of information for both in-copyright and out-of-copyright pages, but this increase is significantly greater for out-of-copyright pages.

Differential Effects for Images vs. Text

While the previous section modeled the idea that copyright restrictions create differential cost reductions for digital information, the differential impact of copyright by media type were not discussed. However, while it is possible to paraphrase textual material without violating copyright, reusing copyrighted images without violating copyright is harder.

Accordingly, let

$$C^{images}_{in-copy}(W_{q,k}) = C^{text}_{in-copy}(W_{q,k}) = \frac{C(W_{q,k})}{2} = \frac{k}{2q}$$

$$C^{images}_{out-of-copy}(W_{q,k}) = \frac{C(W_{q,k})}{4} = \frac{k}{4q}$$

$$C^{text}_{out-of-copy}(W_{q,k}) = \frac{C(W_{q,k})}{2} = \frac{k}{2q}$$

Solving the maximization problem, we obtain:

$$k^{*text}_{out-of-copy} = k^{*text}_{in-copy} = \frac{4q^2}{(q+2)^2}$$

$$k^{*images}_{out-of-copy} = \frac{4q^2}{(q+1)^2} \quad \rangle \quad k^{*images}_{in-copy} = \frac{4q^2}{(q+2)^2}$$

Therefore, as is clear from this simple example, the differential cost reductions for images and text provides

a direct prediction: the impact of copyright on reducing information reuse is driven primarily by a difference in the reuse of images rather than the reuse of textual information.

Differential Effects by Quality Levels

Now consider the impact of the copyright law on affecting increase in knowledge for topics of different quality types.

For in-copyright topics, percent increase in knowledge $\Delta k_{in-copy} = \frac{k^*_{in-copy} - k^*}{k^*}$ and similarly, for out-of-copyright topics, $\Delta k_{out-of-copy} = \frac{k^*_{out-of-copy} - k^*}{k^*}$. Solving we get:

$$\Delta k_{in-copy} = \frac{4q^2}{(q+4)^2}\left[\frac{4(q+3)}{(q+2)^2}\right]$$

$$\Delta k_{out-of-copy} = \frac{4q^2}{(q+4)^2}\left[\frac{3(2q+5)}{(q+1)^2}\right]$$

$$\therefore \Delta = \Delta k_{out-of-copy} - \Delta k_{in-of-copy} = \frac{4q^2(2q+3)}{(q+1)^2(q+2)^2}$$

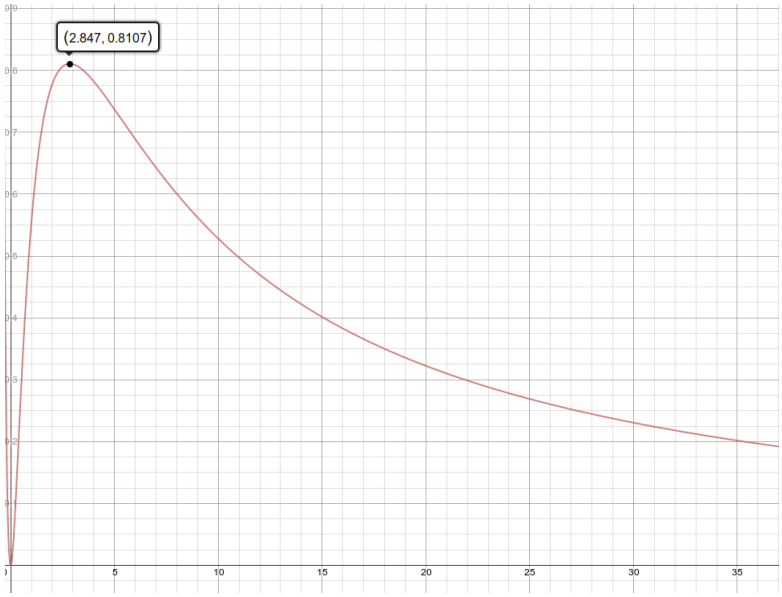
$$\therefore \frac{d\Delta}{dk} = -\left[\frac{8q(q^3-6q-6)}{(q+1)^3(q+2)^3}\right]$$

$$\implies \boxed{\frac{d\Delta}{dq} > 0 \quad \forall q \in (0, \approx 2.84)} \quad \textit{and} \quad \boxed{\frac{d\Delta}{dq} < 0 \quad \forall q \in (\approx 2.84, \infty)}$$

Therefore, under this simple model, while the increase in information reuse is greater for out-of-copyright topics than for in-copyright topics at the same quality level, the magnitude of this positive effect depends significantly on the quality level q of the topic. For low q (i.e. $0 < q < \approx 2.84$), out-of-copyright topics experience a greater increase in information reuse compared to in-copyright topics. The intuition for this effect is simple: returns to information are higher for higher quality topics, and therefore a greater reduction in cost of adding information due to a lack of copyright is most beneficial for these topics. However, after a certain threshold, this logic no longer applies, and an increase in topic quality reduces the benefit from out-of-copyright status. The intuition for this effect is the following: higher quality topics had higher levels of initial information, and returns to adding more information are decreasing. Therefore, it becomes more valuable to add information to medium-quality topics because these have less information to start with than high-quality topics. However, very low-quality topics are interesting to too few people to make the addition of information worthwhile and dont experience the same benefits of out-of-copyright status. The figure at the end of this section provides a plot of how Δ varies for different values of q.

In this way – the model builds intuition for the key results of the paper, (i) digitization improves the quality of Wikipedia information, (ii) Copyright law reduces the potential benefits from digitiziation (iii) copyright mainly operates through the reuse of images rather than text and (iv) Potential benefits from a lack of copyright on digital material are greatest for topics of "intermediate" quality.

A plot of how Δ varies with q



A.3 Appendix A3 : Sample B Construction

This appendix section provides details on data construction process for Sample B.

To build this sample, I first used the "Baseball Hall of Fame" voting dataset by Sean Lahman[20] to compile a list of 541 players who have been nominated for election to the Baseball Hall of Fame and who made their debut appearances between 1944 and 1984. The Hall of Fame nomination list allowed me to include players who had finished their careers and who had passed a screening committee judgment, but it also "removes from consideration players of clearly less qualification" (Abbott, 2011). Thus, the nomination list can be said to include only those players who merit encyclopedic inclusion. The dataset also provides biographical details of the players including date of debut and performance details like their experience, length of career and number of appearances in all-star games.

Having constructed a list of players who could have possibly benefited from magazine information, I then manually matched the names of players to their respective pages on Wikipedia. Manual matching helps to avoid problems where a player with a common name like "Jackie Robinson" is matched to the Wikipedia page for Jack Robinson the politician, or worse, Jackie Robinson the basketball player. After having completed this matching, similar to Sample A, for each player page I downloaded archival versions of each player's page as it appeared on December 1 for every year between 2004 and 2012. To measure the amount of information on each page, I then built an automated python parsing utility that allowed me to measure citations to *Baseball Digest* (as measured by references to *Baseball Digest* in the text), the number of images[21] on a page, and the number of words of text. These data do not count citations by year of publication, only the Baseball Digest magazine as a whole.

+Having constructed a list of players who could have possibly benefited from magazine information, I

[20] see `http://www.seanlahman.com/baseball-archive/statistics/`

[21] I detect images by looking for references to the following file extensions: `jpg,jpeg,gif,svg,tiff,png`

then manually matched the names of players to their respective pages on Wikipedia. Manual matching helps to avoid problems where a player with a common name like "Jackie Robinson" is matched to the Wikipedia page for Jack Robinson the politician, or worse, Jackie Robinson the basketball player. After having completed this matching, similar to Sample A, for each player-page I downloaded archival versions of each player's page as it appeared on December 1 for every year between 2001 and 2012. To measure the amount of information on each page, I then built an automated python parsing utility that allowed me to measure citations to *Baseball Digest* (as measured by references to *Baseball Digest* in the text), the number of images on a page, and the number of words of text (in thousands of words). I detect images by looking for references to the following file extensions: `jpg,jpeg,gif,svg,tiff,png`. Note that these data does not count citations by year of publication, only the Baseball Digest magazine as a whole.

For each page, I obtained web traffic data in the form of page-views from `stats.grok.se`. I also computed average monthly traffic data for every year from 2012 back to 2007, before which traffic data is not available. Additionally I constructed a *quality* metric for each player. *Quality* is calculated based on percentile rank in the list of all-star appearances within the sample under consideration. The All-Star game is an annual event that takes place between the "best" players of baseball's two leagues, and, therefore, provides a good indicator of a player's performance in a given year. *Quality* is a categorical variable with four values, indicating the player's ranking by percentile (top 25 percentile, 25-50 percentile, 50-75 percentile and bottom 25 percentile). Given that all the players in my sample have retired, the quality rankings do not change, and should be considered to be a time-invariant variable at the player-page level.