Does Copyright Affect Reuse? Evidence from Wikipedia and the Google Books Digitization Project

Abhishek Nagaraj *
nagaraj@berkeley.edu

September 13, 2016

Abstract

While digitization projects like Google Books have dramatically increased access to information, this study examines how the ability to reuse digital information and diffuse knowledge to a wider audience critically depends on features of copyright law. I use the digitization of both copyrighted and non-copyrighted issues of *Baseball Digest* by the Google Books Digitization Project to measure the impact of copyright on a prominent venue for reuse: Wikipedia. A specific feature of the 1909 Copyright Act, *copyright renewal requirements* ensure that material published before 1964 is out of copyright, which allows for causal estimation of the impact of copyright on Wikipedia across this sharp cutoff. I find that digitization by Google Books increased the reuse of Baseball Digest on Wikipedia, but copyright hurt reuse by reducing citations to copyrighted issues by over 135%. Copyright also had real impacts on readership – affected pages had, on average, 20% lower gains in internet traffic than pages unaffected by copyright. Finally, the impact of copyright on digitized material was highly uneven. Copyright mostly affected the reuse of images rather than text, and prevented reuse of information about less-popular players as compared to more popular ones.

*Haas School of Business, University of California, Berkeley. F565, 2220 Piedmont Ave, Berkeley, CA 94720-1900.
I am thankful to Erik Brynjolfsson, Daniel Fehder, Jeff Furman, Avi Goldfarb, Shane Greenstein, Joshua Krieger, Matt Marx, Aruna Ranganathan, Fabian Waldinger, Pai-Ling Yin, participants of the Economic Sociology Working Group and MIT Economics Third Year Lunch and especially Pierre Azoulay, Scott Stern, Catherine Tucker and Heidi Williams for comments. All errors remain mine.

1 Introduction

The increasing 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 enabled massive changes in the ways that creative producers build upon and reuse existing information (Goldfarb et al., 2014). For example, the digitization of music has enabled new forms of production, remixing and reuse which has resulted in a new wave of innovation in the music industry (Waldfogel, 2014). Similarly, the digitization of "prior art" about traditional medicine has significantly affected new innovations in this area (Choudhury and Khanna, 2015). Similarly, the availability of new digital maps of the earth's surface have led to a new wave of discoveries in the gold exploration industry (Nagaraj, 2015).

Despite its enormous potential, the digitization process is governed by intellectual property 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., 2012). Some firms have argued for strengthening of 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 importance of these debates (for example, see Supreme Court case *Authors Guild v. Google, Inc.*), 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., 2013).

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 intellectual property 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, digitization impedes the free diffusion of digital information and needs to be reigned-in. Further, given contradictory theoretical predictions, empirical research on the impact of copyright on reuse also faces significant challenges. First, it is difficult to cleanly measure the reuse and diffusion of digital information. Second, there exists little variation in copyright – it often applies by default for over a hundred years. And finally, even when variation exists, comparing the reuse of out-of-copyright and in-copyright material is problematic because such variation is often correlated with the quality or recency of the underlying material.

In this paper, I make empirical progress on the question 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. About 4 years after the project was underway, in December 2008, Google Books digitized all existing issues (i.e. every issue published between 1940 to 2008) of *Baseball Digest*, a prominent baseball magazine, and made them available online to readers for free. The digitization of *Baseball Digest* is particularly 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. At the same time, however, both pre-1964 and post-1964 issues of the magazine were digitized by Google Books and could be accessed and read online. By exploiting this idiosyncratic variation in copyright, this study sheds light on the broader question of how intellectual property could affect the potential benefits from digitization. In particular, I focus on the impact of copyright on affecting the reuse of magazine material on Wikipedia, a natural venue to investigate this question. Not only is Wikipedia the fifth most visited website on the internet (receiving about 10 billion page-views every month) as well as a common source of information about the history of baseball, it also stores all past versions of a given page, allowing the analyst to track how information changes in response to the Google Books digitization event and copyright.

To make things clearer, consider the example in Figure 1. The example shows scanned pages from Google Books for two issues of *Baseball Digest*. Panel (1) is about a feature on Felipe Alou published in 1963, rendering it "out-of-copyright", while Panel (2) is on Johnny Callison published

in 1964, making it "in-copyright." For these two players, Figure 1 depicts pages on Wikipedia as they appeared in December 2012. Neither of these two pages displayed the players' images before *Baseball Digest* was digitized, but in 2012, while Alou's page had an image from Baseball Digest, Callison's page had no images at all. Despite being printed in two issues that were published around the same time, one image finds use in a broader context while the other seems lost among the pages of Baseball Digest. Further analysis shows that, over 51 new citations were made between 2008 and 2012 to issues of Baseball Digest published in 1963, while only 17 new citations were made to magazine issues published in 1964 in the same time period. Furthermore, at the page level, Felipe Alou's Wikipedia page in 2012 contained a citation to the 1963 issue of *Baseball Digest*, and experienced about a 121% increase in traffic since 2008 as compared to only a 23% increase for Callison's page. While many alternative explanations (such as differences in player popularity) could account for the differences, the statistical analysis isolates the role of copyright in establishing these patterns.

Specifically, I track Wikipedia citations to issues of Baseball Digest published between 1944 and 1984 before and after the Google Books project on Wikipedia (i.e. between 2004 and 2012). The unit of analysis is a "publication-year," i.e. all issues of the magazine published in any 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 as compared to in-copyright issues (publication-years 1964-1984) after the Google Books digitization event in a differences-in-differences framework. This specification helps to isolate the causal impact of copyright on the reuse of Baseball Digest information after the launch of the digitization project.

Figure 2 provides some simple descriptive statistics that foreshadow the core econometric results. Panel (1) plots total citations to all issues of Baseball Digest on Wikipedia between 2004-2012, while Panel (2) plots total citations separately for in-copyright and out-of-copyright publication-years. As Panel (1) indicates, after the digitization of Baseball Digest in late 2008, the average number of citations for all publication-years of the magazine between 1944-1984 increased dramatically, suggesting the large and positive impact of digitization on information reuse. However, when this increase is examined separately in Panel (2) for in-copyright and out-of-copyright publication-years, the gains from digitization are heavily concentrated for out-of-copyright issues. The econometric estimates indicate that citations to out-of-copyright publication-years increase by about 135% as compared to citations to in-copyright publication-years after digitization, even after controlling for

publication-year and calendar-year fixed effects. Combined, Figure 2 and the econometric estimates convey the headline findings of the paper – 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.

Having understood the impact of the copyright law at the publication-year level, I construct an additional sample at the Wikipedia player-page level (Sample B) which tracks Wikipedia pages for a set of about five hundred prominent baseball players who were active between 1944 and 1984. Using this sample I provide robustness for the baseline analysis and extend it in two different ways. First, I evaluate the impact of copyright on the value that Wikipedia pages deliver to end users, by incorporating data on user traffic to Wikipedia pages. If, despite the limited diffusion of information from in-copyright Baseball Digest issues, Wikipedia editors are able to create high-quality pages using information from alternate sources and attract similar levels of readership, then the welfare consequences of copyright on Wikipedia are likely to be less severe. However, my econometric estimates indicate that pages affected by copyright had on average 20% lower gains in internet traffic, suggesting a permanent and significant loss to Wikipedia from the inability to reuse copyrighted information.

Second, I hypothesize and test the idea that while copyright lowers the overall level of reuse, its impact is more concentrated for certain types of information as compared to others. Building on the digitization and intellectual property literature, I argue that the impact of copyright on preventing reuse is most salient when the underlying information is harder to duplicate using alternate sources. Specifically, I hypothesize that the impact of copyright is most salient for the reuse of images rather than text because it is harder to duplicate visual information without violating copyright, while textual information is easily paraphrased. Similarly, I also hypothesize that copyright restrictions on digital information are most harmful for Wikipedia pages about players who are less well-known because information about them is harder to obtain from alternate sources. I test the heterogeneous impact of the 1964-copyright experiment on both these margins in a regression framework and find support for my hypotheses. These results highlight the important role of copyright law in affecting digitization by shaping the distribution of reuse for different types of information.

This research contributes to the literature on digitization (Goldfarb et al., 2012; Miller and Tucker, 2011; Waldfogel, 2014) by evaluating the role of intellectual property law in influencing the economic impacts 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, 2014; Nagaraj, 2015; Brynjolfsson et al., 2006). Finally, I also contribute to the nascent empirical literature on copyright including some work in the legal domain estimating the impact of copyright in the publishing context (Heald, 2007, 2009a; Buccafusco and Heald, 2012), work that studies the impact of copyright on prices (Li et al., 2012; Reimers, 2013; Mortimer, 2007) and literature on copyright enforcement and piracy (Aguiar and Waldfogel, 2014; Luo and Mortimer, 2015; 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

Google Books is 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). It is perhaps the most prominent of a number of ongoing digitization projects which include US government efforts to make available digitized records from government transactions and the Library of Congress "National Digital Library" project. 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).

In order to understand the role of copyright in influencing the broader impact of the Google Books project, I focus on an event that occurred on 9 December 2008, when Google Books announced that it would make available all of the past issues of *Baseball Digest*, along with a number of other popular magazines. All published issues since the magazine's founding were made available at one time with consent from the publishers (Foulser, 2008). Furthermore, while it is quite common for Google Books to provide only a few pages of a given book on their website under "limited

preview," Baseball Digest was not subject to this restriction. Instead every page of every issue of the magazine printed before December 2008 was made available online. 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. Specifically, while existing work in copyright has mostly considered differences in copyright status between different publications (Reimers, 2013; Heald, 2013), the digitization of Baseball Digest is more attractive because some issues of the very same publication are available in the public domain, while others are copyrighted. This subsection describes, in more detail, the specific legal changes that led to this copyright variation.

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, by considering magazines and periodicals, additional variation within a single publication can be obtained. This is because, for periodicals, the copyright term is not defined in terms of the author's life term, but rather the publication date. Specifically, prior to 1964, periodicals were subject to the *copyright renewal* requirement. As per this rule, under the 1909 Copyright Act, two copyright terms were provided: a 28-year initial term and a 28-year renewal term (Landes and Posner, 2002). 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 had 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 1963 plus 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, issues of a single publication could have widely varying copyright protection, despite being printed only a few months apart from each other.

Although copyrights for some movies and books were renewed successfully, because this requirement was not well-known, it *"tripped up many smaller publishers, and a failure to renew caused many works to lapse into the public domain"* (Andrade, 2014). According to research by the University

of Pennsylvania library system (Ockerbloom, 2006), a failure to renew copyrights has meant that issues of many periodicals published before 1964, have fallen into the public domain including issues of journals like the *American Economic Review*, *Bell System Technical Journal*, *Biometrika* and periodicals like *The Baltimore Sun*, *The Los Angeles Evening Herald*, *The Kansas City Star* etc.[1]

I leverage the The University of Pennsylvania library's "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,[2] no issues of Baseball Digest published before 1964 were ever renewed. In other words, all issues of the magazine published before 1964 could safely be assumed to be in the public domain, while for issues published in or after 1964, copyrights will start to expire starting in 2020 and remain copyrighted until that date.

While a large number of magazines were under the copyright exemption that I study, only a very small number of publications were digitized by Google Books. Of these, Baseball Digest happens to be the only one that I'm aware of that meets both criteria of being digitized and also falling under the copyright exemption. In addition, given my focus on Wikipedia, baseball as a topic is a good choice because it has a thriving editor community on Wikipedia. Further, the experiment is also 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. *Baseball Digest* has provided baseball's vast fan-base with information and news about the game for over seven decades since its founding in 1942 (Jones (2007), Brown (2011)).

The Google Books digitization of *Baseball Digest*, therefore, represents a unique case. 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. The role of differing copyright status on the reuse of digitized content is the empirical focus of the paper.

[1] While a number of different publications did not renew copyright, I focus on Baseball Digest, because it was one of the few that was also digitized in its entirety by the Google Books project.

[2] http://onlinebooks.library.upenn.edu/cce/firstperiod.html

2.2 Data

In order to understand the impact of *Baseball Digest*'s copyright status on reuse, I turn to 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 contributions, making citations to magazines like *Baseball Digest* typical on the site. Third, *Baseball Digest* often contains profiles of baseball players and teams, particularly in the form of detailed articles, interviews and player images. Such biographical information forms the foundation of any encyclopedia (Greenstein and Zhu, 2014) 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. This, in turn, allowed me to trace the diffusion of information using micro-data, which would be difficult to do in another setting. While one aim of this paper is to show how Wikipedia can be used to study the diffusion of information, others have used similar data to study collaboration and digital knowledge production (Zhang and Zhu, 2010; Greenstein and Zhu, 2012; Nagaraj et al., 2009; Algan et al., 2013; Gorbatai, 2012; Aaltonen and Seiler, 2013). Specifically, I use two different strategies to collect Wikipedia data and build two samples, Sample A and Sample B. This section describes these samples, which form the core of my analysis, in greater detail.

2.2.1 Sample A: Publication-Year Level

In order to construct the workhorse sample for the main specification (referred to throughout as Sample A), I followed a five-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.[3] Third, having collected multiple snapshots of thousands of pages, I wrote python scripts to detect citations to Baseball Digest magazine on each of the individual snapshots. I searched for these citations using pattern

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

matching techniques for a variety of different citation formats because Wikipedia does not follow a standard way to cite source material. However, I found that Wikipedia editors most commonly use the format *Baseball Digest (1963)*, rather than referring to a specific issue, or article within the magazine.[4] Accordingly I harmonized all citations by the publication-year of the magazine that was included in a particular citation for all publication-years between 1944 and 1984.[5] At this step, I also noted whether the citation was made to an image was being reused, or whether the citation was in the main text of the article. As a fourth and final step, I manipulated these raw data to build Sample A which tracks total citations on Wikipedia to every publication-year of Baseball Digest with one observation for every 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. In this way, Sample A allowed me to detect the exact year when new citations to a publication-year were added to Wikipedia and to analyze the impact of the 1964 copyright experiment on citations to Baseball Digest before and after the Google Books digitization event.

Table 1, Panel (1) lists summary statistics for this sample. The main independent variable of interest is *1(Out-of-Copy)*, which I include in the analysis as $out-of-copy$ and which indicates whether a given publication-year is in the period where no copyrights apply (i.e. all publication-years before 1964). The second independent variable of interest is, *1(Year>2008)* which I include in the analysis as *post* and which indicates that the observation is from a time period after the Google Books digitization event. The main dependent variable is *Total Citations* which counts total citations to any given publication-year in a calendar year. Two related dependent variables are *Image Citations* and *Text Citations* which splits these citations by whether a citation is being made for reusing an image, or in the main text of the article respectively. By construction, almost half of the publication-years are in the out-of-copyright period and the median Wikipedia-year in the study is 2008. As far as outcome variables are concerned, the data show that, on average, a publication-year of Baseball Digest receives about 4.19 citations of which about 3 are for text, while the rest are for images.

[4] Although, citations to specific articles or issues can also be found on Wikipedia, I rely on citations to the publication-year to get the largest possible sample of reuse for Baseball Digest.

[5] I choose these years because they form a window of 20 years before and after the copyright cutoff year of 1964.

2.2.2 Sample B: Player-Page Level

While Sample A helps to measure citations to Baseball Digest at the publication-year level and helps to provide an accurate assessment of the reuse of Baseball Digest information, it is still inadequate to understand the aggregate impact of the Baseball Digest digitization and copyright restrictions on Wikipedia. For example, if copyright on Baseball Digest issues reduces the reuse of magazine information, but Wikipedians are able to create high-quality pages for a given topic using alternate sources, then the overall impact of copyright on Wikipedia would be less severe. We are unable to assess this overall impact without directly collecting data on the quality of Wikipedia pages that could have been affected by the Baseball Digest digitization and copyright restrictions. To understand the overall impact of the copyright restriction for Wikipedia, I build another sample (referred to throughout as 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 or quantity of images) as well as proxies for quality, such as player-page level traffic. A player-page level analysis is also helpful because it 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 as compared to less well-known ones, the heterogeneous effects of copyright on different types of topics can be better understood. Such an analysis would not be possible at the publication-year level. For these reasons, I rely on a player-page sample (referred to throughout as Sample B) of the Wikipedia data to estimate the impact of copyright on the reuse of digitized information.

To build this sample, I first used the "Baseball Hall of Fame" voting dataset by Sean Lahman[6] 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.

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

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.[7] 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[8] to *Baseball Digest* (as measured by references to *Baseball Digest* in the text), the number of images[9] on a page, and the number of words of text (in thousands of words).

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.[10] *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.

Table 1 Panel (2) lists summary statistics for Sample B. The main independent variables of interest are *1(Out-of-Copy)* which I include in the analysis as $out-of-copy$, which assigns a player to the out-of-copyright group if they made their debut before 1964. The second is, *1(Year>2008)* which I include in the analysis as *post*, defined in a similar way to Sample A, and is set to one if an observation is from a time period after the Google Books digitization event. The main dependent variables are *Citations, Images, Text* and *Traffic*. The summary statistics show that baseball Wikipedia pages contain an average of 0.12 citations, 0.50 images and 890 words (or .89 times a thousand words) and that they receive, on average, 101.49 page views per month.

[7] 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.

[8] Note that this data does not count citations by year of publication, only the Baseball Digest magazine as a whole.

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

[10] 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.

3 Empirical Results

3.1 Descriptive Analysis

In order to estimate the impact of the 1964 copyright experiment on the reuse of the digitized Baseball Digest magazine, 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

Similar in spirit to Figure 1, Table 2 provides simple descriptives comparing the likelihood of reuse of information in out-of-copyright and in-copyright issues of Baseball Digest. Specifically, Table 2 Panel (1) compares citation and reuse outcomes for in-copyright and out-of-copyright publication-years of Baseball Digest using Sample A, while Table 2 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, as Panel (1) makes clear, 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 21.05 citations. 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) finds that out-of-copyright players make 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 some 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, setting aside the 1964 copyright experiment. Figure 2 Panel (1), plots the average citations to each Baseball Digest publication-year for all issues in Sample A, between 2004

and 2012. Average citation counts after the Google Books Digitization event are shown using dark grey bars, while those from before are indicated using light grey bars. As these data indicate, the average Baseball Digest publication-year received a very small number of citations (about 0.125 cites according to my data in 2008) before the Google Books project made this material more accessible. However, after the Google Books digitization project, citations to Baseball Digest issues increased dramatically and discontinuously. In 2009, the average citations to Baseball Digest publication-years increased to 1.7 from 0.125 in 2008 and this number ultimately increased to about an average of 15.4 citations per publication-year by 2012. This dramatic and discontinuous increase in citations to Baseball Digest suggests large potential 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 Appendix, I include some analysis that helps to provide a causal estimate of the impact of digitization on increasing 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 as 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 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.

Having confirmed that there were potential benefits to Wikipedia from Baseball Digest, Table 2 Panel (2) plots the average number of citations to Baseball Digest separately for out-of-copyright (red dots) and in-copyright (blue crosses) publication-years between 2004 and 2012. As this figure makes clear, while digitization dramatically increased citations to the average publication-year, these gains were heavily concentrated among issues published before 1964, i.e. those in the out-of-copyright regime. 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.

When taken together, Table 2 and Figure 2 suggest important cross-sectional and temporal patterns in the data. First, there seem to be important cross-sectional differences in reuse outcomes for out-of-copyright and in-copyright Baseball Digest publication-years (Table 2 Panel (1)), and in the amount of material on in-copyright and out-of-copyright player-pages (Table 2 Panel (2)).

Similarly, Figure 2 provides extremely striking descriptive data suggesting that while the Google Books digitization event greatly increased the reuse of material from Baseball Digest on Wikipedia, gains from digitization were disproportionately concentrated for issues of Baseball Digest published before 1964 as compared to those published after.

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, and, therefore, it is difficult to conclude from the descriptive analysis that the difference in copyright status of pre- and post-1964 issues of Baseball Digest are the primary drivers 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 this 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, in this section I use a regression framework to directly test the central hypothesis of this paper that differences in copyright status of pre- and post-1964 issues were the primary drivers of differences in the levels of reuse of magazine material on Wikipedia.

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

$$Cites_{it} = \alpha + \beta_1 \times post_t \times out-of-copy_i + \gamma_i + \delta_t + \epsilon_{it}$$

where γ_i and δ_t represents 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 $post_t \times out-of-copy_i$ 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.

In sample A, the unit-of-observation fixed effect controls for publication-year fixed effects, which flexibly controls for time-invariant differences between publication-years 1944-1984. Further, $out-of-copy_i$ 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, $out-of-copy_i$ is an indicator variable that equals one if a player makes his debut before 1964. Table 3 presents estimates from 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.[11] The first OLS model is estimated without publication-year (Sample A) or player-page (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 as compared to in-copyright issues across the different specifications. The estimate in Column(2) 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 5.7 more citations as compared to issues published in or after 1964. Compared to the average citation level of about 4.2, this is an increase of almost 135%. These estimates are somewhat muted although they remain large and significant in the Log specification in Column (3), which suggests that citations to out-of-copyright issues increase by about 34.7% after the digitization event. The difference between the Log and OLS estimates is important to note,[12] and the Log estimates should be preferred for a more conservative estimate of the impact of copyright. I will include both OLS and Log specifications in the remaining analysis.

Sample B provides additional evidence for the negative effect of copyright law on preventing the diffusion of material to Wikipedia.[13] Specifically, the estimates from Sample B suggest that after the Google Books digitization event, players who made their debut before 1964 (i.e. out-of-copyright

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

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

[13] 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.

players) are more likely to cite material from Baseball Digest as compared to in-copyright players. The OLS estimate from Sample B, Column (2) suggests about 0.2 additional citations as 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 7.5%. This analysis reveals two findings: first, 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, i.e. players who were most likely to be affected by the 1964 copyright experiment.

Taken together, estimates from both Sample A and Sample 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 reuse on Wikipedia.

3.3 Checking for Pre-Trends and Other Robustness Checks

When they are combined, the descriptives and the regression analysis help to build confidence that out-of-copyright issues of Baseball Digest were significantly more likely to be reused as compared to in-copyright issues. The regression analysis includes both fixed effects for wikipedia-year, which controls for general time trends in Wikipedia citation patterns to Baseball Digest, and importantly also publication-year or player-page level fixed effects, depending on the sample employed. These fixed effects control for time-invariant differences between publication-years (for example, if certain publication-years had more or higher-quality magazines than others) and between player-pages (for example, if certain players were more well-known as compared to others).

3.3.1 Time-Varying Estimates

However, while this specification controls for time-invariant differences, it is unable to control for time-varying differences between in-copyright and out-of-copyright groups. 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 differences-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 upwards. 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 3 presents graphical versions of the following event study specification separately for Sample A and Sample B:

$$Cites_{it} = \alpha + \gamma_i + \delta_t + \Sigma_t \cdot \beta_t \cdot out - of - copy_i \times 1(t) + \epsilon_{it}$$

for unit i in wikipedia-year t.

Time-varying coefficients in Figure 3 reveal no discernible evidence in the increase in citations for out-of-copyright groups as compared to in-copyright groups before 2009. Panel (1) that estimates this specification with Sample A, finds virtually zero difference in the level of citations between pre- and post-1964 issues, and 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 less precisely estimated. 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 and does not seem to be changing before 2009. Further, after 2009, we see a large positive effect, even though none of the coefficients are individually significant. In addition to the graphical analysis, Appendix Table A.2 presents in tabular format, each of the different point estimates from the β_t used to construct Figure 3. Inspecting the different "pre-trend" coefficients helps to further justify the similarity of pre-trends across groups.

Overall, the evidence from this analysis, especially from the precisely estimated coefficients from 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 the pre-trends directly, it is possible to check for robustness of the main result using another feature of the setting that has so far been under-exploited. Specifically, the setting allows for examination of the impact of the intellectual property 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 simply by the higher likelihood of older issues to 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 would provide a robust check of the main specification because it provides a non-parametric method to examine the impact of copyright on reuse that does not rely on the "pre-trends" assumption inherent to the difference-in-difference specification.

Accordingly, Figure 4 plots the net increase in citations to individual publication-years between 1944 and 1984 between 2008 and 2012. Each bar represents the total number of new citations to a given publication-year on Wikipedia before and after the Google digitization project was launched. Observations from the out-of-copyright period are in the darker shade of grey, while those from the in-copyright group are in the lighter shade.

As Figure 4 indicates, the number of new citations added by publication-year does not display a steady time trend that decreases from 1944 to 1984. Instead, issues published right before 1964 have a significantly higher gain in the number of new citations as compared to issues published right after. For example, issues published in 1963 gained about 51 citations as compared to those published in 1964, which gained only about 17 citations. Similarly, issues published in 1962 gained about 38 citations as compared to only 16 citations for issues published in 1965. This sharp discontinuous difference[14] in the likelihood of new citations for out-of-copyright publication-years significantly helps to increase 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 other confounding factors such as different pre-trends in citation patterns.

3.3.3 Additional Falsification Checks

In addition to examining the pre-trends between in-copyright and out-of-copyright groups directly and exploiting the discontinuity in the cutoff around 1964, in the Appendix, I present a few other robustness checks that help build confidence in the main results. First, I conduct a falsification

[14]It must be noted that while the increase in citations after 1963 is relatively small and similar between 1964-1984, issues published closer to the copyright cutoff before 1964 are more likely to be cited as compared to issues published closer to 1944. This is likely to be because of Wikipedians focusing on the issues published right before the 1963 cutoff and "working backwards", thereby paying less attention to older issues. It is also possible that older issues were of lower quality and were less useful and informative.

analysis,[15] where I restrict the sample to the "pre" period only and assume that the treatment year is 2007, rather than 2009. If out-of-copyright groups are experiencing an increasing rate of citations as compared to in-copyright groups, then we 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) 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 pre-trends between in-copyright and out-of-copyright groups might be driving the main results. In addition to the pre-trends, it is also important to address the pattern of the time-trend in the estimates after the digitization event. Specifically, Figure 3 suggests that while the digitization happened in late 2008, the positive impact of out-of-copyright status seems to become apparent around 2011. Some qualitative evidence and some robustness checks I conducted 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 (through other 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, 2012). For example, a "power" Wikipedian I interviewed told me:

> 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 around certain calendar years rather than being evenly distributed. In addition, for robustness, I include two specifications in the Appendix where I shorten the time-frame of the analysis and re-estimate the main specifications excluding later years in my sample. Specifically, in Appendix Table A.4 Panel (1) I estimate the specification using a sample from 2005-2011, and in Panel (2) using a

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

sample from 2006-2010 rather than estimating the panel using years 2004-2012. 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

I have so far established that citations to out-of-copyright issues increased at a significantly higher rate as compared to citations to in-copyright issues. While this evidence is important to understanding the impact of copyright law on the diffusion on digitized material, I now turn to traffic information to directly inform the welfare impact of copyright on Wikipedia. Specifically, if Wikipedia contributors are able to supplement copyrighted information not available from the magazine with information from other sources, then we might find that a reduced likelihood of citations to Baseball Digest does not translate into lower quality for Wikipedia pages. However, if a lack of citations to in-copyrighted issues of Baseball Digest also translates into lower traffic for Wikipedia pages, then the overall implications for copyright to impact welfare are magnified.

Table 2 presents some simple cross-sectional comparisons for the traffic information, indicating that, on average, $out-of-copy$ player-pages have about one and half times more traffic as compared to $in-copy$ 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 as compared to their post-1964 counterparts. In this section, I utilize traffic data[16] from Sample B in order to shed light on the impact of the 1964 copyright experiment on traffic in a regression framework. The specification follows exactly the estimating equation laid out 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 4 player quality quartiles. Similarly, Models (2) and (3) also include additional player-page level fixed effects. The estimates in Column (2)

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

indicate that on average, out-of-copyright pages receive a boost of about 20 hits per month after controlling for player and year fixed effects. The coefficient reduces slightly when $quality \times year$ fixed effects are included. Against a mean of about 101 page-views per month, this represents an increase of about 20%. In the appendix, I examine the robustness of these estimates to log models. Columns (3) and (4) of Table A.6 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 copyright on traffic to affected pages would be to boost page-views in the order of 20%, a significant difference.

Overall, my estimates suggest a significant positive impact of the digitization program for 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 also 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 work has found that one important channel through which digitization affects markets is by reducing costs of access to information (Goldfarb et al., 2014; Bakos, 1997). The digital availability of information has been theorized to make it much easier to locate and build upon relevant knowledge (Shapiro and Varian, 1999; Chiou and Tucker, 2011). 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 observed in Figure 2 Panel (1). In parallel, a robust literature in intellectual property has argued that IP might introduce transaction costs that mitigate benefits from the relative reduction in costs of access (Waldfogel, 2012; Williams, 2013; Murray and Stern, 2007). Taken 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 Figure 2 Panel (2) 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 media (notably text vs. images) and different player-pages on Wikipedia. I then proceed to empirically testing these hypotheses.

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 transaction costs required to prevent copyright infringement in the case of reuse are significantly different for images vs. text. For work to be reused without copyright infringement, some evidence of "transformative reuse" is often necessary. While copyright on text only prohibits verbatim reuse of large sections of text, for images, this rule does not apply. 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, in order to prevent infringement, the reuse of textual information with sufficient modifications is typically possible at low costs and relatively simple. However, for images to be legally reused, large modifications need to be made to satisfy the "transformative reuse" criteria, 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) and, therefore, end-users often avoid the reuse of copyrighted images without explicit permission from the creators.

I argue that differences in practical and legal difficulties in reuse impose different transaction costs from intellectual property 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 as 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 rather 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 affecting the level of information on Wikipedia pages is most pronounced for pages of lower player quality, as compared to pages of higher player quality.

In order to understand how copyright affects player-pages, I argue that the optimal level of knowledge on Wikipedia pages depends both on the value of new information to a 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, 2011; Strömberg, 2007). In line with this research, we can expect value of information about players to be directly proportional to player quality, while the costs of sourcing information to be inversely related. Higher quality players attract higher interest from end users and therefore the value of information for these players is greater. On the contrary, there are a number of alternate sources covering higher quality players which makes it cheaper to source information about them. The assumption for lower quality players is exactly the opposite: it is more expensive to source information and the value of such information is also lower. In this framework, given the costs and benefits of adding new information, there are greater incentives to add information for players of higher quality and this information is easier to obtain even before the digitization effort.

Now, consider the reduction in the cost of access to information due to digitization (Goldfarb et al., 2014) and increased transaction costs due to intellectual property (Williams, 2013). For players of the highest quality, information already exists 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 a middle tier of player quality: players who are good enough to merit encyclopedic inclusion, but information about whom was relatively difficult to source before digitization. Therefore the prediction is 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 tier of player quality is unlikely to 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, reduced cost of music due to digital distribution most benefited smaller 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 the Bitnet "democratized" innovation by benefiting lower-ranked university collaboration (Agrawal and Goldfarb, 2008). Similarly, the innovation literature has also found that reduced cost of access to scientific material through the establishment of scientific institutions are most for countries in the developing world, where such access is harder to obtain (Furman and Stern, 2011).

4.2 Comparing the Reuse of Images vs. Text

The next part of the analysis is to test the predictions emerging from the theoretical framework for the differential effects of copyright on the reuse of images vs. text.

As a first step, I recreate the simple descriptive analysis in Figure 4 separately for image and text citations using data from Sample A. Image citations are made for the reuse of an image, while text citations are citations made to Baseball Digest in the main text of a page. As before, for each publication-year of Baseball Digest, I measure the total number of new (text or image) citations between 2008 and 2012 on Wikipedia. This chart is displayed as Figure 5.

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 from an issue published in or after 1964. In other words, the likelihood that an image will cite material from a post-1964 issue of Baseball Digest after digitization is very close to zero. However, this picture changes dramatically for text citations. In this case, there are a significant number of citations both before and after the 1964 cutoff, and the copyright status seems to have little impact on influencing the level of Wikipedia citations. 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, therefore, estimate difference-in-difference specifications similar to the baseline specification used

for Table 3. The main outcome variables are citations to images and text, rather than the total number of citations. The estimates from this analysis are presented in Table 5. The estimates indicate that the 1964 copyright cutoff has a more significant impact on the reuse of images rather than text. For example, the second set of estimates for images, which includes both publication-year fixed effects and wikipedia-year fixed effects, suggests that on average out-of-copyright publication-years experience about 5.4 more image citations as compared to in-copyright publication-years. However, the analogous estimate for text citations indicates only a small and statistically insignificant difference in citations between in-copyright and out-of-copyright publication-years. This conclusion is justified even when Log-OLS models are considered. Image citations experience about a 117% increase while the corresponding increase for text citations is statistically indistinguishable from zero.

Finally, similar to Figure 3, I also test the validity of these estimates by plotting time-varying coefficients separately for image and text citations. These are represented in Figure 6. As Panel (1) indicates, the difference in the reuse of images for in-copyright and out-of-copyright issues is again close to zero before the digitization event. However, after 2008, there was an immediate increase in this difference and by 2012, out-of-copyright publication-years had a significantly higher levels of image reuse. However, as Panel (2) indicates, this pattern does not hold for text citations. The difference in text citations pre-digitization is close to zero and constant, however, this pattern does not change significantly, even 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

Finally, having tested the prediction that copyright law will have a greater effect for images as compared to text, I now turn to testing the differential impact of copyright law across different player quality categories leveraging data from Sample B.

In order to examine the heterogeneous impact of the copyright experiment by player quality, I

estimate the following specification:

$$Y_{it} = \alpha + \beta_1 \times post_t \times out-of-copy_i + \sum_{m=2}^{4} \beta_m \times post_t \times 1(quality_i = m) + \sum_{n=2}^{4} \hat{\beta}_n \times post_t \times out-of-copy_i \times 1(quality_i = n) + \gamma_i + \delta_t + \epsilon_{it}$$

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 percentile "quality" rank of a player as a number between 1 and 4 (top 25 percentile, 25-50 percentile, 50-75 percentile and bottom 25 percentile). The key coefficients of interest $\hat{\beta}_n$ estimate the difference between the $out-of-copy_i \times post_t$ coefficient and $out-of-copy_i \times post_t \times 1(quality = n)$ 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.

Figure 7 plots these coefficients separately for each quality percentile.[17] Panel (1) validates the hypothesis that the impact of copyright on the reuse of images is larger for the 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 second 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, also shows a similar pattern, indicating that the impact on copyright on increasing traffic is also most relevant for players in the lower tier of player quality. 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.

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

[17] I plot the coefficient on $out-of-copy \times post_t$ for quality=1 and add this estimate to coefficients for other quality levels to compute marginal effects.

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*
> *Motion Picture Association of America (MPAA)*

This paper suggests an empirical framework to answer this question and suggests one mechanism through which copyright might influence the benefits from digitization: by prohibiting reuse of digitized material, particularly within open, community-based innovation projects like Wikipedia.

There are three major sets of 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, as compared to 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 terms of traffic. Finally, the impact of copyright on affecting reuse was uneven–it mostly affected the reuse of images, while textual material was not affected, and out-of-copyright material was most helpful for less well-known player's Wikipedia pages as 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 Wikipedia represents an idiosyncratic setting to analyze the impact of copyright on reuse because Wikipedia is a non-profit and could have lower benefits from reusing copyrighted work, which would cause my estimates to be biased upwards. This is a valid concern, and is certainly an important limitation of this study. However, the following dicussion helps reduce this concern to some extent.

First, Wikipedia does not seem to be alone in enforcing copyright, since a number of other digital platforms, where one might expect reuse of digitized information, also have extensive programs for copyright enforcement. These include YouTube (Seidenberg, 2009), Amazon, all major mobile

application stores and even Google's search engine (Dillon Scott, 2011). For instance, Apple's AppStore rejected about a thousand applications in August 2009 because they used copyrighted images and books in their applications.[18] Apple also hosts an online tool where firms can report copyright violation. Meanwhile, Google removed about 26 million links from its search index in October 2013[19] including links that provided access to copyrighted books, music and data. An extensive literature on piracy (Bechtold, 2004) in the entertainment industry has also shown that copyright-related interventions that limit the availability of digitized content are quite common, and are often very effective (Danaher et al., 2010; Danaher and Smith, 2013).

Furthermore, Wikipedia's non-profit status also does not seem to completely prohibit it from licensing content similar to other commercial for-profit entities. For example, online volunteers are known to negotiate for licenses by leveraging Wikipedia's General Counsel which acts similar to a company's legal counsel.[20] In this way, despite being a non-profit, Wikipedia does have some mechanisms for licensing information similar to a firm. In addition, Wikipedia is also a direct input of information to other for-profit organizations. For example, Google's "Knowledge Engine" (which provides information on certain individuals and events in response to Google searches) relies heavily on Wikipedia.[21]

Third, there also seem to be a number of other anecdotal examples where copyright on other types of content is having an impact on Wikipedia that is similar to the case of Baseball Digest. Some preliminary research that I have done seems to indicate that a large portion of the anatomical images on Wikipedia seem to be sourced from a 1918 edition of Gray's Anatomy,[22] 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 non-renewal, and, therefore, a large number of images from TIME magazine from before 1964 find reuse on Wikipedia.[23] Finally, in the context of Amazon, Heald (2013) documents how copyright restrictions shape the availability of book reprints on Amazon. He finds that a random sample of new books for sale on Amazon.com shows more books for sale from the 1880's than the 1980's

[18] See bit.ly/1aXpksj

[19] https://www.google.com/transparencyreport/removals/copyright/

[20] see: https://wikimediafoundation.org/wiki/User:GeoffBrigham_%28WMF%29

[21] https://www.technologyreview.com/s/520446/the-decline-of-wikipedia/

[22] see this page for a listing of these images: https://commons.wikimedia.org/wiki\/Category:Gray's_Anatomy_plates

[23] for example: https://commons.wikimedia.org/wiki/File:Shidehara_Kijuro_on_TIME_magazine_cover.jpg

suggesting that out-of-copyright works are more available on digital bookshelves as compared to more recent copyrighted works.

Finally, to further ease concerns about external validity, my study builds on 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, 2007, 2009b; Buccafusco and Heald, 2012) has shown that works produced before 1923, which are generally in the public domain, are much more accessible today than works produced after 1923. For example, books produced before 1923, are more easily accessible on Amazon and Audible.com and are more likely to be digitized (Brooks, 2005) than those produced afterward. A more recent study in the economics literature (Reimers, 2013) 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. Similarly, a study of the fiction market in the 1820s also shows that an important impact of copyright is likely to be an increase in the price of books and that such an increase may reduce access (Li et al., 2012).

In light of the anecdotes from the previous section and 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, given Wikipedia's prominence, the estimates presented represent a significant part of the gains due to innovation in the digital economy.

5.2 Contributions and Managerial Implications

Going beyond the question of external validity, this paper makes a number of contributions to the nascent empirical literature at the intersection of intellectual property and digitization. A significant literature has analyzed the implications of digitization for important economic activities like consumer search, pricing and targeting. This literature generally posits 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 role of intellectual property in influencing the economic effects of digital information. In particular, I argue 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. Digitization, therefore, is posited to have a "democratization" effect in economic markets. For example, it has been found that file-sharing increases live performance revenues for small 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 the brands that were less well-known at the time of infringement and newer brands (Qian, 2014).[24] My work adds to this literature by looking at the impact of copyright on inequality in the reuse of information for less and more well-known topics. My findings follow the intuition of the infringement and file-sharing literature–when access costs are reduced through digitization and public domain status, less well-known topics benefit disproportionately but the presence of copyright could prevent this "democratization."

The results are also related to another stream of work on the empirical effects of intellectual property on the diffusion of knowledge. Some studies (Murray and Stern, 2007; Murray et al., 2009; Williams, 2013; Furman and Stern, 2011) find a generally negative effect of intellectual property on follow-on use, while more recent evidence seems to be mixed (Sampat and Williams (2014), Galasso and Schankerman (2015)). This study provides direct evidence to this literature that the costs of access (i.e. digitization) seem to matter for the impacts of IP on reuse. From a policy point of view, this paper is able to directly 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, and (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 intellectual property 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 useful because there is often a concern that piracy is so rampant on the internet that tools other than traditional intellectual property (like DRM) are often necessary (Zhang, 2014) to supplement toothless copyright law. Second, for managers who are interested in using user communities like Wikipedia to generate innovation (Boudreau et al. (2011), Franzoni and Sauermann (2014)) or open innovation more broadly (Fleming and Waguespack, 2007), this study suggests that provision of external, un-

[24] I thank a reviewer for pointing us towards this research.

copyrighted but digitized material can be extremely beneficial. This study finds that knowledge within user communities is often sourced from external sources, and policy measures that affect the availability and legal status of sources can either boost or retard innovative activity within such communities.

Methodologically, this paper provides a number of suggestions for measuring copyright's effects going forward. First, I show how the internet provides a fertile ground for estimating carefully the impacts of copyright on reuse using micro-data. Not only is the internet an important venue where future copyright battles will be waged, but the digital and quasi-permanent nature of digital content allows for the detailed measurement of the creation of new products and services on the internet. In addition, in light of the finding that copyright impacts images more than it does text, this research points to the importance of the key difference between patents and copyrights, namely that patents typically protect the underlying idea while copyright protects only the "expression". In other words, the impacts of copyright are likely to vary not simply by the quality of the data, but also by the medium of expression. This distinction is likely to be important in the future.

5.3 Limitations and Welfare Calculation

Finally, this paper does not evaluate the overall welfare consequences of the impact of copyright on digital information, but helps to make progress in that direction. In a static setting, where new digital information does not build upon pre-existing work, stronger copyright law should incentivize the production of digital information (Watt and Towse, 2006). However, in the more dynamic setting, where the production of new knowledge depends upon pre-existing information (Scotchmer, 1991) (for example: the presence of Google Books helps the production of new knowledge on Wikipedia), whether stronger copyright will boost knowledge production is unclear. If transaction costs imposed by copyright prevent the reuse of existing work, then optimal copyright policy should provide for weaker intellectual property than it would without a significant burden from transaction costs. In this framework, it becomes critical that credible empirical measurements of the cost of copyright on preventing follow-on use of digital information be obtained. Without such measurements, we do not know, "whether copyright protection would need to be strengthened or weakened" in the digital age (Waldfogel, 2012). 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 pre-existing work, then a weakening of copyright will hurt overall knowledge creation as well. In such as 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.[25] However, I am not able to directly estimate the role of lost copyright on hurting incentives for the production of new knowledge by the publisher which is an important limitation of this work.

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 further the diffusion of information. Such an argument was a feature of the "Mickey Mouse" law of 1998.[26] Further, in the case of archival material, the question of compensating authors does not arise because of the so-called "orphan works" problem, when the authors of material cannot be identified or contacted (Smith et al., 2012). Even under the possibility of copyright creating incentives for creating new material, the estimates help measure welfare losses from retroactive extensions of copyright or from difficulties in locating missing authors.

References

Aaltonen, A. and S. Seiler (2013). Cumulative knowledge and open source content growth: The case of wikipedia.

Abbott, L. (2011, September). Future Baseball Hall of Fame Players Who Did Not Appear in a World Series.

Agrawal, A. and A. Goldfarb (2008). Restructuring Research: Communication Costs and the Democratization of University Innovation. *American Economic Review 98*(4), 1578–90.

Aguiar, L. and J. Waldfogel (2014). Digitization, Copyright, and the Welfare Effects of Music Trade. *Copyright, and the Welfare Effects of Music Trade (December 3, 2014)*.

Algan, Y., Y. Benkler, M. F. Morell, and J. Hergueux (2013). Cooperation in a Peer Production Economy Experimental Evidence from Wikipedia. In *Workshop on Information Systems and Economics, Milan, Italy*, pp. 1–31.

[25] I did make a number of reasonable attempts to contact the publishers of *Baseball Digest*, including emailing them, filling out a contact form and calling their office in order to investigate the possibility of licensing content for reuse, but my requests were met with no response. This suggests, that in this case, producer surplus from licensing archival material is fairly low.

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

Anderson, N. (2007, May). New Copyright Alliance hopes to strengthen copyright law.

Andrade (2014). Copyright renewal - when it had to happen, or else.

Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management science 43*(12), 1676–1692.

Bechtold, S. (2004). Digital rights management in the United States and Europe. *Am. J. Comp. L. 52*, 323.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bertrand, M., E. Duflo, and S. Mullainathan (2004, February). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Boudreau, K. J., N. Lacetera, and K. R. Lakhani (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science 57*(5), 843–863.

Brooks, T. (2005). How Copyright Law Affects Reissues of Historic Recordings: A New Study. *ARSC Journal*.

Brown, M. (2011, April). MLB Revenues Grown From $1.4 Billion in 1995 to $7 Billion in 2010. *Biz of Baseball*.

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

Brynjolfsson, E., Y. J. Hu, and M. D. Smith (2006). From niches to riches: The anatomy of the long tail.

Buccafusco, C. J. and P. Heald (2012). Do Bad Things Happen When Works Enter the Public Domain?: Empirical Tests of Copyright Term Extension. *Berkeley Technology Law Journal*.

Chiou, L. and C. Tucker (2011, December). Copyright, Digitization, and Aggregation. *SSRN eLibrary*.

Choudhury, P. R. and T. Khanna (2015). Ex-ante Information Provision and Innovation: Natural Experiment of Herbal Patent Prior Art Adoption at the USPTO and EPO.

Danaher, B., S. Dhanasobhon, M. D. Smith, and R. Telang (2010). Converting pirates without cannibalizing purchasers: the impact of digital distribution on physical sales and internet piracy. *Marketing Science 29*(6), 1138–1151.

Danaher, B. and M. Smith (2013). Gone in 60 Seconds: The Impact of the Megaupload Shutdown on Movie Sales. *Available at SSRN 2229349*.

Dillon Scott, P. (2011). Google Transparency Report: UK requests removal of nearly 100,000 items from index.

Fleming, L. and D. M. Waguespack (2007). Brokerage, boundary spanning, and leadership in open innovation communities. *Organization science 18*(2), 165–180.

Foulser, D. (2008, December). Search and find magazines on Google Book Search.

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

Furman, J. and S. Stern (2011). Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production. *American Economic Review 101*(5), 1933–63.

Galasso, A. and M. Schankerman (2015). Patents and Cumulative Innovation: Causal Evidence from the Courts*. *Quarterly Journal of Economics 130*(1).

Gallini, N. and S. Scotchmer (2002). Intellectual property: when is it the best incentive system? In *Innovation Policy and the Economy, Volume 2*, pp. 51–78. MIT Press.

Gans, J. S. (2015). Remix rights and negotiations over the use of copy-protected works. *International Journal of Industrial Organization 41*, 76–83.

Gans, J. S. and S. Stern (2003). The product market and the market for "ideas": commercialization strategies for technology entrepreneurs. *Research policy 32*(2), 333–350.

Goldfarb, A., S. Greenstein, and C. Tucker (2014). Introduction to" Economic Analysis of the Digital Economy". In *Economic Analysis of the Digital Economy*, pp. 1–17. University of Chicago Press.

Goldfarb, A., R. C. McDevitt, S. Samila, B. Silverman, and others (2012). The Effect of Social Interaction on Economic Transactions: An Embarrassment of Niches? In *AEA 2013 Annual Meeting*.

Gorbatai, A. (2012, September). Social Structure and Mechanisms of Collective Production: Evidence from Wikipedia.

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

Greenstein, S. and F. Zhu (2014). Do experts or collective intelligence write with more bias? Evidence from Encyclop\a edia Britannica and Wikipedia. *Geography*.

Greenstein, S. M. and F. Zhu (2012). Is Wikipedia Biased? *The American Economic Review 102*(3), 343–348.

Heald, P. (2007). Property rights and the efficient exploitation of copyrighted works: an empirical analysis of public domain and copyrighted fiction best sellers. *UGA Legal Studies Research Paper* (07-003).

Heald, P. (2009a). Testing the Over-and Under-Exploitation Hypotheses: Bestselling Musical Compositions (1913-32) and their Use in Cinema (1968-2007). *Review of Economic Research on Copyright*.

Heald, P. J. (2009b). Does the Song Remain the Same-An Empirical Study of Bestselling Musical Compositions (1913-1932) and Their Use in Cinema (1968-2007). *Case W. Res. L. Rev. 60*, 1.

Heald, P. J. (2013, July). How Copyright Keeps Works Disappeared. SSRN Scholarly Paper ID 2290181, Social Science Research Network, Rochester, NY.

Jones, J. M. (2007, October). Less Than Half of Americans Are Baseball Fans. *Gallup*.

Kitch, E. W. (1977). Nature and Function of the Patent System, The. *JL & Econ. 20*, 265.

Kupferman, T. R. (1944). Renewal of Copyright. Section 23 of the Copyright Act of 1909. *Columbia Law Review 44*(5), 712–735.

Landes, W. M. and R. A. Posner (2002). Indefinitely renewable copyright. *U Chicago Law & Economics, Olin Working Paper* (154).

Lemley, M. A. (2004). Ex ante versus ex post justifications for intellectual property. *The University of Chicago law review*, 129–149.

Lessig, L. (2004). *Free culture: How big media uses technology and the law to lock down culture and control creativity*. Penguin.

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

Leval, P. N. (1990). Toward a fair use standard. *Harvard Law Review 103*(5), 1105–1136.

Li, X., M. MacGarvie, and P. Moser (2012, November). Dead Poets' Property - The Copyright Act of 1814 and the Price of Books in the Romantic Period. *Working Paper*.

Luo, H. and J. H. Mortimer (2015). Copyright Enforcement: Evidence from Two Field Experiments.

Mazzoleni, R. and R. R. Nelson (1998, December). Economic Theories about the Benefits and Costs of Patents. *Journal of Economic Issues 32*(4), 1031–1052.

Merges, R. P., P. S. Menell, and M. A. Lemley (2012). Intellectual property in the new technological age.

Miller, A. R. and C. E. Tucker (2011). Can health care information technology save babies? *Journal of Political Economy 119*(2), 289–324.

Mortimer, J. H. (2007, August). Price Discrimination, Copyright Law, and Technological Innovation: Evidence from the Introduction of DVDs. *Quarterly Journal of Economics 122*(3), 1307–1350.

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

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

Murray, F. and S. Stern (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization 63*(4), 648–687.

Nagaraj, A. (2015). The Private Impact of Public Maps—Landsat Satellite Imagery and Gold Exploration.

Nagaraj, A., P. Seetharaman, R. Roy, and A. Dutta (2009, December). Do Wiki-pages Have Parents? An Article-Level Inquiry into Wikipedia's Inequalities. *Workshop on Information Technology Systems (WITS)*.

Ockerbloom, M. J. (2006). The Next Mother Lode for Large-scale Digitization? Historic Serials, Copyrights, and Shared Knowledge. *Scholarship at Penn Libraries*, 65.

Prat, A. and D. Strömberg (2011). The political economy of mass media.

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

Reimers, I. (2013). The Effects of Intellectual Property on the Market for Existing Creative Works. *Working Paper*.

Rob, R. and J. Waldfogel (2007). PIRACY ON THE SILVER SCREEN*. *The Journal of Industrial Economics 55*(3), 379–395.

Sampat, B. and H. Williams (2014). How do patents affect follow-on innovation. *Evidence from the human*.

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

Samuelson, P. (2009). Google Book Search and the future of Books in Cyberspace. *Minn. L. Rev. 94*, 1308.

Scotchmer, S. (1991). Standing on the shoulders of giants: cumulative research and the patent law. *The Journal of Economic Perspectives*, 29–41.

Seidenberg, S. (2009). Copyright in the Age of YouTube. *ABAJ 95*, 46.

Shapiro, C. and H. R. Varian (1999). Information Rules.

Silverwood-Cope, S. (2012, February). Wikipedia: Page one of Google UK for 99% of searches | IP Blog: SEO, SMO and web development insights.

Smith, M., R. Telang, and Y. Zhang (2012). Analysis of the Potential Market for Out-of-Print eBooks. *Available at SSRN 2141422*.

Strömberg, D. (2007). Natural disasters, economic development, and humanitarian aid. *The Journal of Economic Perspectives 21*(3), 199–222.

Waldfogel, J. (2012, May). Copyright Research in the Digital Age: Moving from Piracy to the Supply of New Products. *American Economic Review 102*(3), 337–342.

Waldfogel, J. (2014). Digitization and the Quality of New Media Products: The Case of Music. In *Economic Analysis of the Digital Economy*, pp. 407–442. University of Chicago Press.

Watt, R. and R. Towse (2006, December). Copyright Protection Standards and Authors' Time Allocation. *Industrial and Corporate Change 15*(6), 995–1011.

Williams, H. (2013). Intellectual Property Rights and Innovation: Evidence from the Human Genome. *Journal of Political Economy*.

Wu, T. (2015, September). What Ever Happened to Google Books? *The New Yorker*.

Zhang, L. (2014). Intellectual property strategy and the long tail: Evidence from the recorded music industry. *Available at SSRN*.

Zhang, X. and F. Zhu (2010). Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review*, 07–22.

Zittrain, J. (2009). *The future of the internet–and how to stop it.* Yale University Press.

6 Tables and Figures

Table 1. **Summary Statistics**

(1) Sample A: Unit of Observation – Publication-Year (N=360)

	Mean	SD	Median	Min	Max
Publication-Year	1964.50	11.56	1964.50	1945	1984
Wikipedia-Year	2008.00	2.59	2008.00	2004	2012
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	4.16	7.79	0.00	0	53
Image Citations	1.19	4.03	0.00	0	30
Text Citations	2.97	4.78	0.00	0	23

(2) Sample B: Unit of Observation – Player-Page (N=4869)

	Mean	SD	Median	Min	Max
Player Debut-Year	1966.12	10.19	1966.00	1944	1984
Wikipedia-Year	2008.00	2.58	2008.00	2004	2012
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	3395
Quality Percentile	2.62	1.29	3.00	1	4

Note: 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 presented above, 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 4869 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 is only available for years 2007 to 2013 and data is missing for other observations. See text for detailed data and variable descriptions.

Table 2. **Cross-Sectional Comparison of Reuse Outcomes**

(1) Sample A : Baseball Digest Publication-Years

	(1)out-of-copy $\bar{y}$	**(2)in-copy $\bar{y}$**	**(3)diff**	**(4)p-val**
Total Citations	20.89	10.33	10.56	0.00
Image Citations	10	0.0952	9.905	0.00
Text Citations	10.89	10.24	0.657	0.68

(2) Sample B : Wikipedia Player-Pages

	(1)out-of-copy $\bar{y}$	**(2)in-copy $\bar{y}$**	**(3)diff**	**(4)p-val**
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

Note: This table compares outcomes for out-of-copyright and in-copyright groups using cross-sectional data from Wikipedia data from 2012. $N = 40$ for Panel (1) and $N = 541$ for Panel (2). In Panel (1), column (1) includes publication-years 1944-1963, while 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 Column (1) and (2). See text for more detailed data and variable descriptions.

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 X post	5.667	5.609	0.316	0.216	0.200	0.0747
	(1.830)***	(1.809)***	(0.153)**	(0.0592)***	(0.108)*	(0.0364)**
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.637	0.703	0.908	0.0477	0.0757	0.0986
N	360	360	360	4869	4869	4869

*+: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 citations to Baseball Digest before and after digitization in a differences-in-differences framework using OLS. The estimates presented use both Sample A (columns 1-3) and Sample B (columns 4-6). *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). See text for specification, detailed data and variable descriptions.

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

	(1) Traffic	(2) Traffic	(3) Traffic
out-of-copy X 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 X Year FE
adj. R^2	0.0137	0.0810	0.0899
N	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 regression estimates the impact of the 1964 copyright exception on traffic to Wikipedia player-pages before and after digitization in a differences-in-differences 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 $debut - year < 1964$. See text for specification, detailed data and variable descriptions. In Column (3) $Quality\ X\ Year\ FE$ controls for separate time-trends by each of the four quartiles of player quality.

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

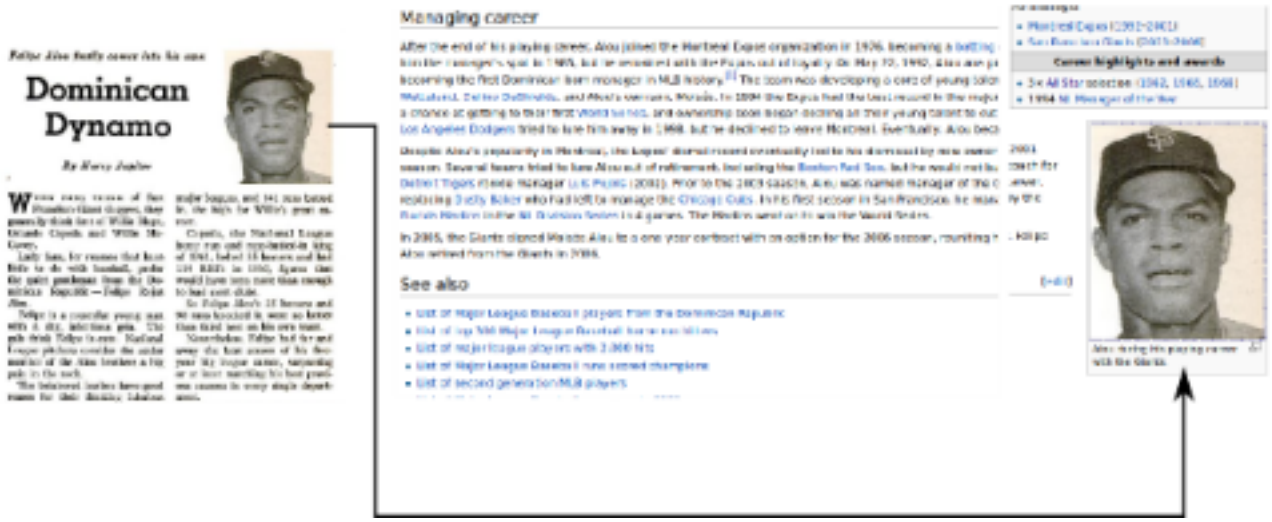
	Images			Text		
	OLS	OLS	Log-OLS	OLS	OLS	Log-OLS
out-of-copy X post	5.444	5.444	1.173	0.222	0.165	-0.0275
	(1.094)***	(1.094)***	(0.151)***	(1.041)	(1.026)	(0.142)
Publication-Year FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.413	0.416	0.574	0.723	0.801	0.913
N	360	360	360	360	360	360

*+: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 A. *post* refers to all Wikipedia-years after 2008, and *out-of-copy* refers to *publication* − *year* < 1964. See text for specification, detailed data and variable descriptions.

Figure 1. **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

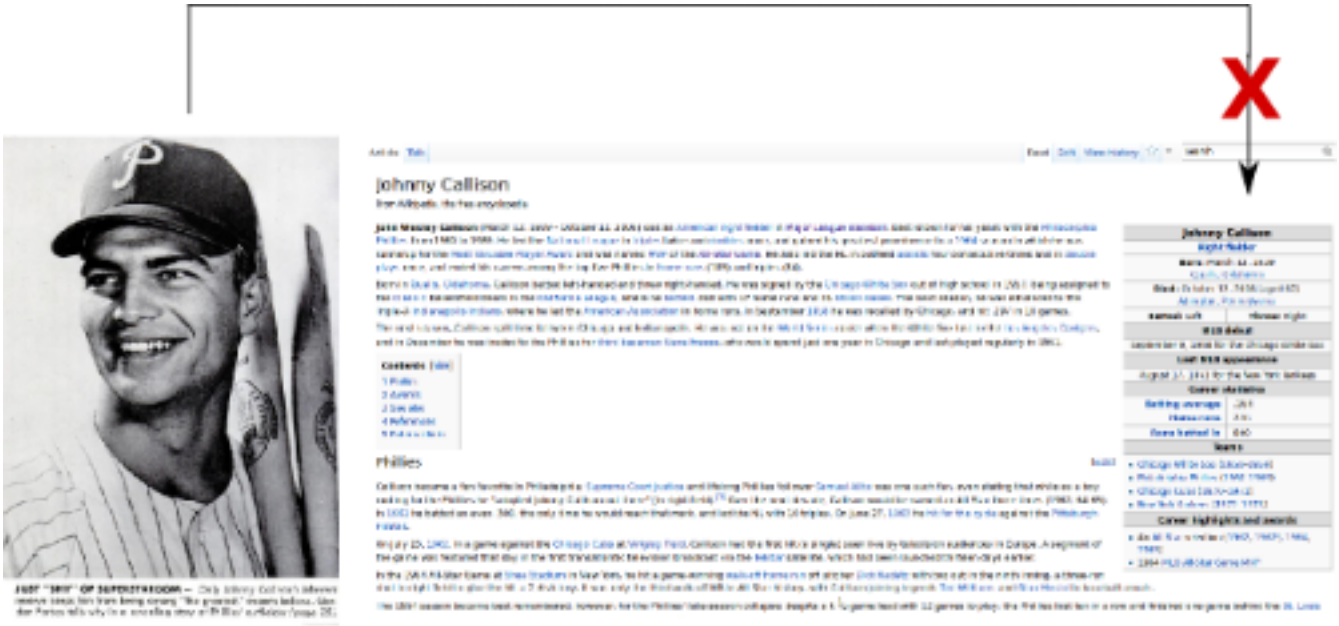
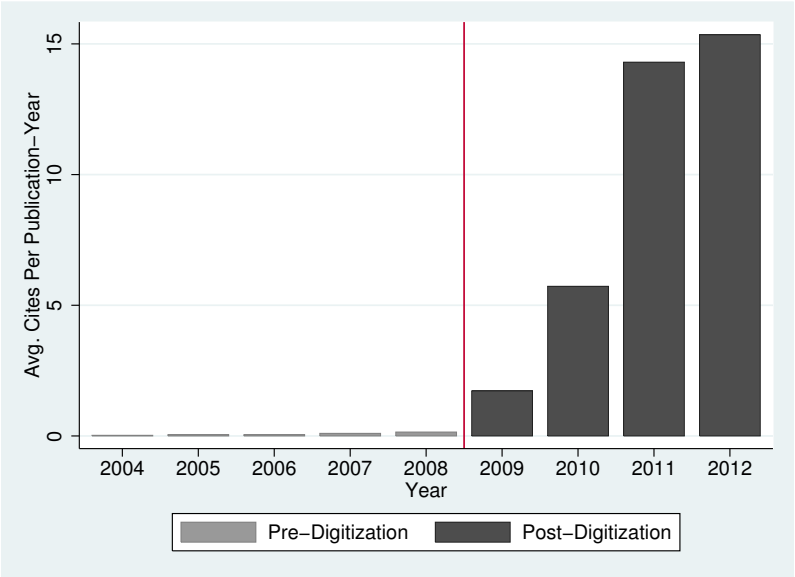
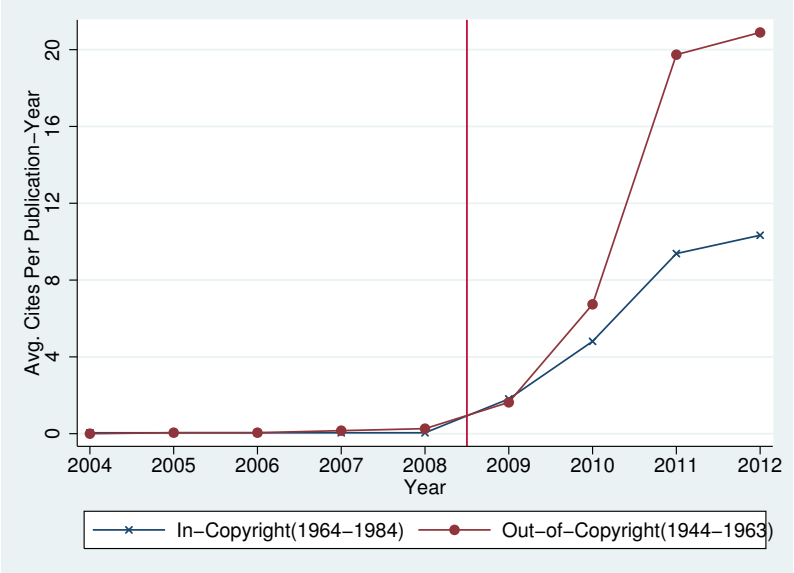


Figure 2. **Citations to Baseball Digest on Wikipedia (Sample A)**

(1) Citations to Baseball Digest before and after Digitization



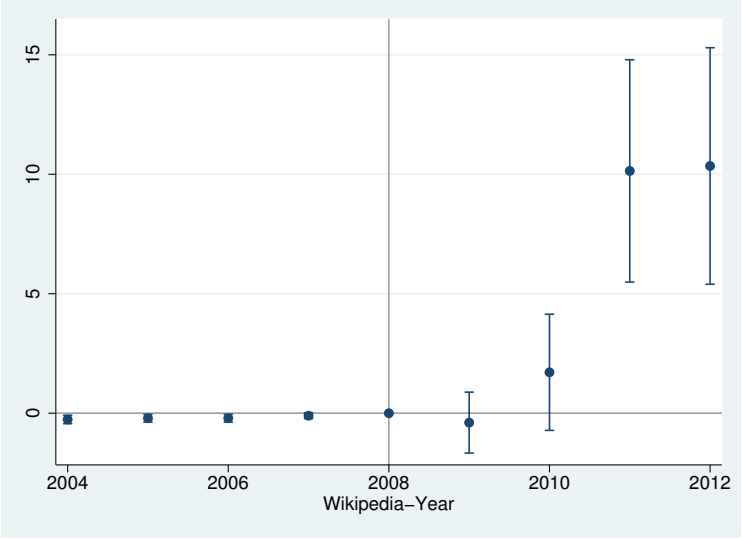
(2) Citations to Baseball Digest for issues published before and after 1964



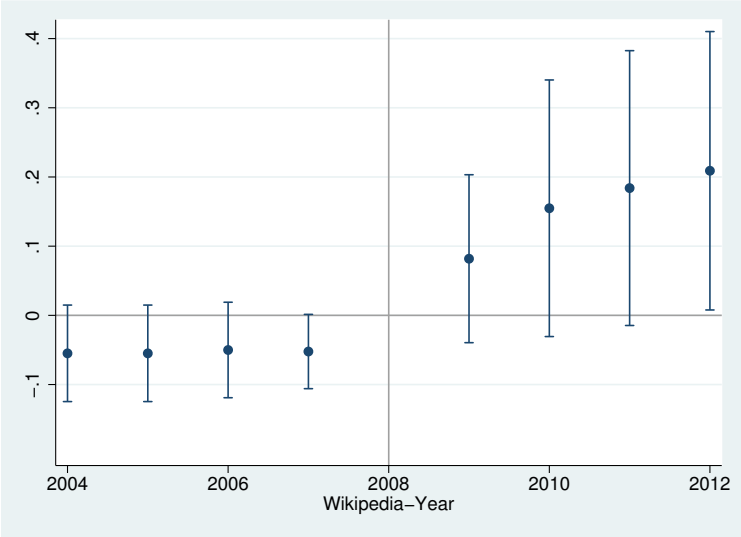
Note: This plot presents some simple descriptive data on citations to Baseball Digest issues. Panel (1) presents average citations per publication-year of Baseball Digest on Wikipedia before and after the Google Books digitization event in 2008. Panel (2) presents similar information, but data are presented separately for all publication-years before 1964 (out-of-copyright) and those in or after 1964 (in-copyright). See text for more detailed data and variable descriptions.

Figure 3. **Time-Varying Estimates of the Impact of Copyright on Citations to Baseball Digest**

(1) Sample A : Total Citations

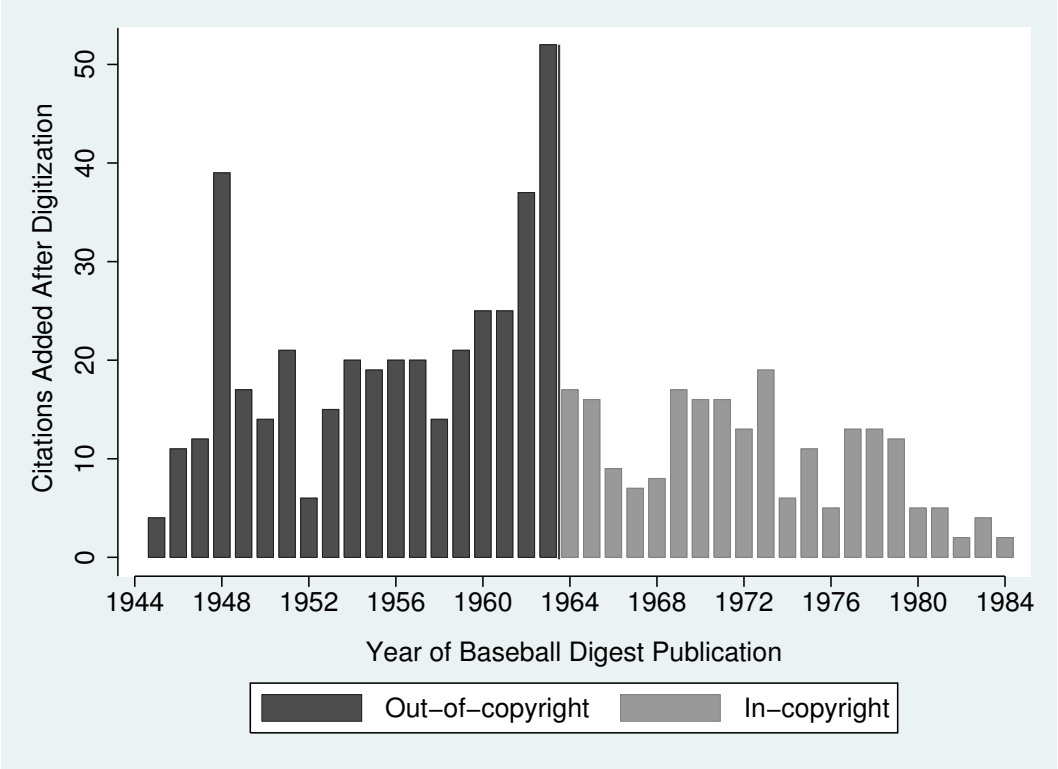


(2) Sample B : Total Citations



Note: This figure plots coefficients (and 95 percent confidence intervals) from the event study specifications described in Section 3.3.1. On the x axis is the Wikipedia-year and the reference year is 2008, the year of the digitization event. This specification is based on Sample A for Panel (1) and Sample B for Panel (2), the coefficients are estimates from ordinary-least-squares (OLS) models, and standard errors are clustered. The dependent variable in both panels is the total number of citations in a calendar year. See text for more detailed data and variable descriptions.

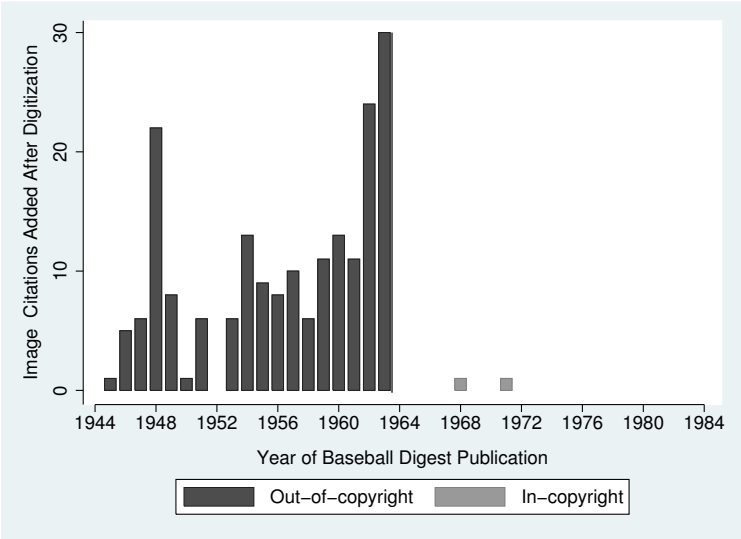
Figure 4. **Citations to Baseball Digest Published Before and After the 1964 Copyright Cutoff (Sample A)**



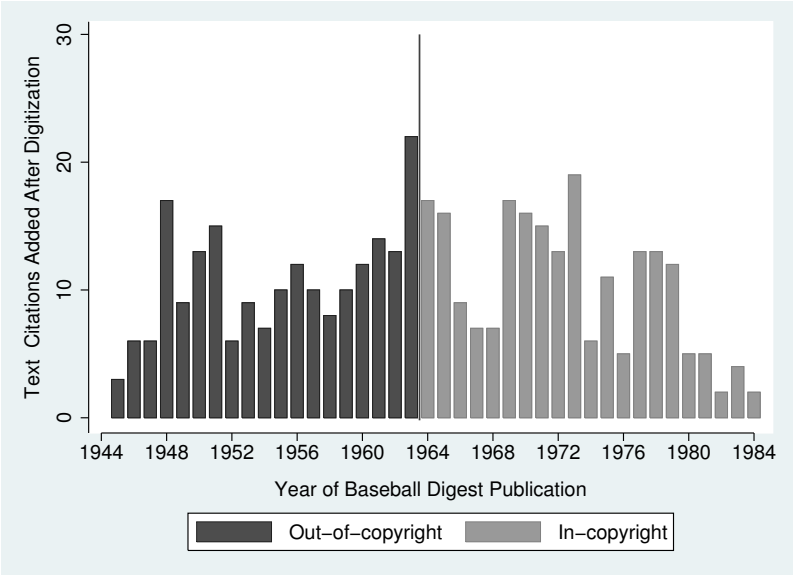
Note: This figure plots the growth in citations to Baseball Digest publication-years in 2012 as compared to 2008. Out-of-copyright publication-years (1944-1963) are shown in dark grey, while in-copyright publication-years (1964-1984) are shown in light grey. See text for more detailed data and variable descriptions.

Figure 5. **Impact of Copyright on Image and Text Citations (Sample A)**

(1) Image Citations



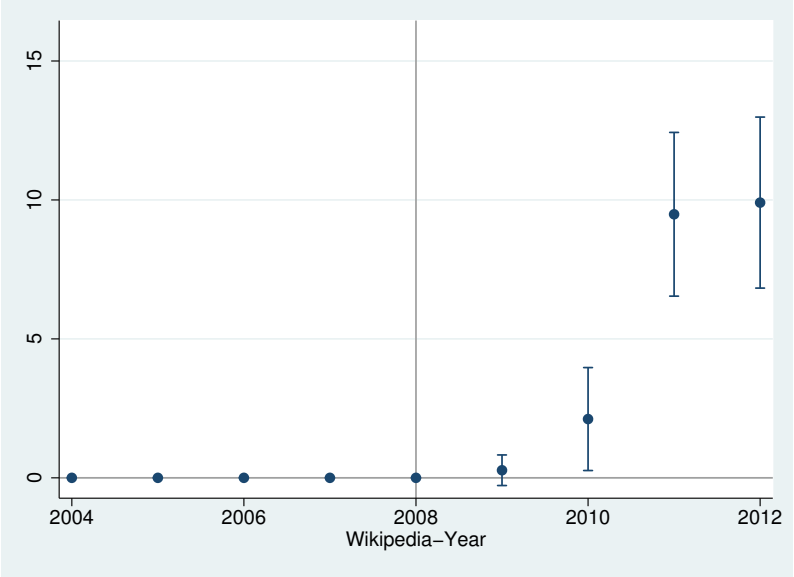
(2) Text Citations



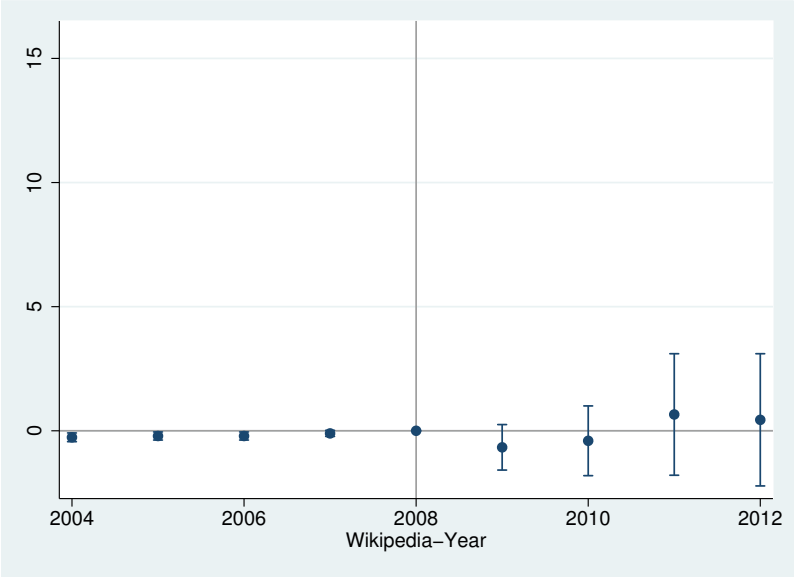
Note: This figure plots the growth in citations to Baseball Digest publication-years in 2012 as compared to 2008. Panel (1) plots the growth in Image citations, while Panel (2) plots the growth in Text citations. Out-of-copyright publication-years (1944-1963) are shown in dark grey, while in-copyright publication-years (1964-1984) are shown in light grey. See text for more detailed data and variable descriptions.

Figure 6. **Time-Varying Estimates for Image and Text Citations (Sample A)**

(1) Image Citations



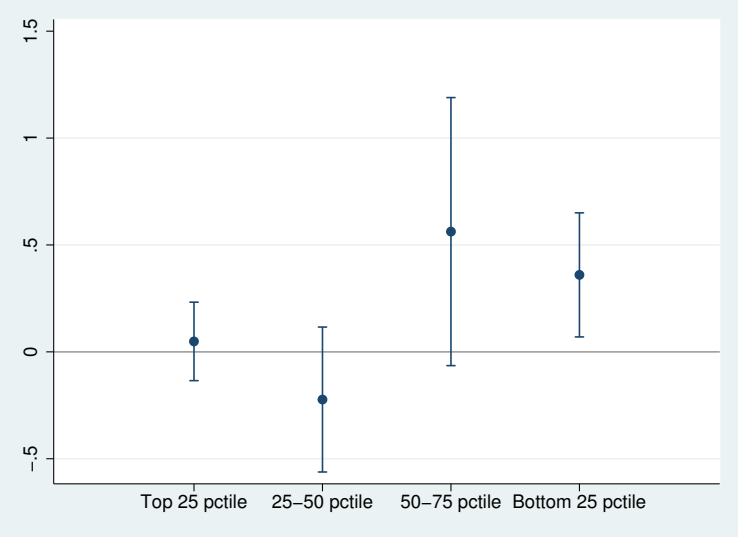
(2) Text Citations



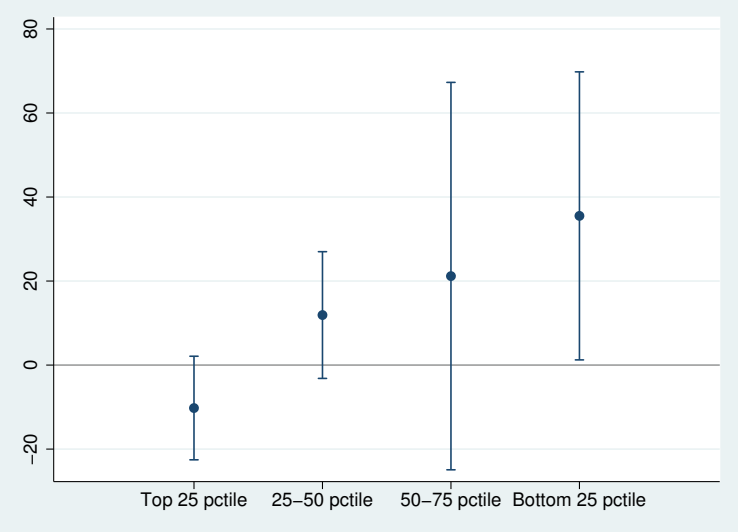
Note: This figure plots coefficients (and 95 percent confidence intervals) from the event study specifications described in Section 3.3.1 separately for image (Panel 1) and text citations (Panel 2). On the x axis is the Wikipedia-year and the reference year is 2008, the year of the digitization event. This specification is based on Sample A for Panel (1) and (2). The coefficients are estimated from ordinary-least-squares (OLS) models, and standard errors are clustered. The dependent variable in both panels is the total number of citations in a calendar year. See text for more detailed data and variable descriptions.

Figure 7. **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* as described in Section 4.3. 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. See text for detailed data and variable descriptions.

A Appendices

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. See text for detailed data and variable descriptions.

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
$Digitization_{-3}$	-0.000	-0.000	-0.000	-0.008	-0.194	-0.642
	(.)	(0.000)	(0.000)	(0.004)*	(0.024)***	(0.031)***
$Digitization_{-2}$	-0.000	-0.000	-0.000	-0.008	-0.075	-0.359
	(.)	(0.000)	(0.000)	(0.004)**	(0.027)***	(0.022)***
$Digitization_{-1}$	-0.000	-0.000	-0.000	-0.003	-0.033	-0.106
	(0.000)	(0.000)	(0.000)	(0.003)	(0.019)*	(0.009)***
$Digitization_{+1}$	1.762	0.095	1.667	0.046	0.060	0.088
	(0.424)***	(0.066)	(0.421)***	(0.015)***	(0.016)***	(0.007)***
$Digitization_{+2}$	4.762	0.095	4.667	0.112	0.140	0.220
	(0.680)***	(0.066)	(0.681)***	(0.027)***	(0.020)***	(0.012)***
$Digitization_{+3}$	9.333	0.095	9.238	0.128	0.220	0.397
	(1.064)***	(0.066)	(1.060)***	(0.028)***	(0.026)***	(0.017)***
$Digitization_{+4}$	10.286	0.095	10.190	0.132	0.259	0.472
	(1.195)***	(0.066)	(1.189)***	(0.028)***	(0.026)***	(0.020)***
$Digitization_{-3}$ x out-of-copy	-0.211	0.000	-0.211	-0.040	-0.093	-0.180
	(0.097)**	(0.000)	(0.097)**	(0.027)+	(0.064)+	(0.084)**
$Digitization_{-2}$ x out-of-copy	-0.211	0.000	-0.211	-0.037	-0.013	-0.127
	(0.097)**	(0.000)	(0.097)**	(0.027)	(0.062)	(0.067)*
$Digitization_{-1}$ x out-of-copy	-0.105	0.000	-0.105	-0.035	0.033	-0.114
	(0.073)	(0.000)	(0.073)	(0.021)*	(0.052)	(0.050)**
$Digitization_{+1}$ x out-of-copy	-0.393	0.273	-0.667	0.077	-0.038	0.050
	(0.757)	(0.327)	(0.544)	(0.045)*	(0.042)	(0.024)**
$Digitization_{+2}$ x out-of-copy	1.712	2.115	-0.404	0.160	-0.013	0.124
	(1.443)	(1.100)*	(0.834)	(0.068)**	(0.055)	(0.039)***
$Digitization_{+3}$ x out-of-copy	10.140	9.484	0.657	0.187	0.383	0.101
	(2.762)***	(1.748)***	(1.453)	(0.073)**	(0.083)***	(0.046)**
$Digitization_{+4}$ x out-of-copy	10.346	9.905	0.441	0.206	0.454	0.117
	(2.938)***	(1.826)***	(1.581)	(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	0.748	0.573	0.799	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	0.0493	0.0618	0.0311	0.0685	0.0631	0.0207
	(0.300)	(0.271)	(0.0855)	(0.0323)**	(0.0398)	(0.0145)
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.256	0.319	0.525	0.0265	0.0350	0.0441
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. Please see text for detailed data and variable descriptions.

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	4.035	3.951	0.198	0.199	0.180	0.0655
	(1.487)***	(1.459)***	(0.150)	(0.0657)***	(0.102)*	(0.0345)*
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.619	0.678	0.889	0.0444	0.0695	0.0898
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	0.875	0.764	-0.0325	0.177	0.152	0.0540
	(0.984)	(0.960)	(0.159)	(0.0745)**	(0.0914)*	(0.0316)*
Unit of Obs. FE	No	Yes	Yes	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.478	0.570	0.820	0.0381	0.0595	0.0757
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. Please see text for detailed data and variable descriptions.

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 whose careers span before and after the copyright-cutoff year of 1964 and estimates the model using players who 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. See text for more detailed data and variable descriptions. All estimates are from ordinary-least-squares (OLS) models.

Table A.6. **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 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.

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

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^*

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

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

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

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:

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

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

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

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

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

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, this magnitude of this positive effect depends significantly on the quality level q of the topic. For low q (i.e. $0 < q <\approx 2.84$), higher quality topics experience a greater increase in information reuse as compared to lower quality 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 lower quality topics because these have a lower starting starting point, as compared to adding information to topics that already have higher levels of information to begin with.

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.

Fig: A plot of how Δ varies with q