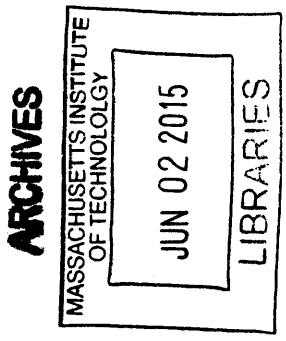
Does Copyright Affect Reuse?
Evidence from the Google Books Digitization Project

by

Abhishek Nagaraj

B.Tech., College of Engineering Pune (2008)
PGDM, Indian Institute of Management Calcutta (2010)



Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

S.M. in Management Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author .....
Sloan School of Management
June 6, 2014

Signature redacted

Certified by .....
Scott Stern
David Sarnoff Professor of Management of Technology
Thesis Supervisor

Signature redacted

Accepted by .....
Ezra Zuckerman
Chair, Sloan PhD Committee

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

by

Abhishek Nagaraj

Submitted to the Sloan School of Management
on June 6, 2014, in partial fulfillment of the
requirements for the degree of
S.M. in Management Research

Abstract

While digitization projects like Google Books have dramatically increased access to digital content, in this study I show how the ability to reuse such information and deliver value to end-users depends crucially on features of copyright law. I use the digitization of both copyrighted and non-copyrighted issues of one publication digitized under Google Books, Baseball Digest, to measure the impact of copyright on a prominent venue for reuse: Wikipedia. I find that digitization causes a significant increase in content on Wikipedia pages, but copyright hurts both the extent of reuse and thereby the level of internet traffic to affected Wikipedia pages. Specific features of copyright law like "fair use" produce nuanced effects: the impact of copyright is more pronounced for images compared to text and becomes economically significant only post-digitization.

Thesis Supervisor: Scott Stern
Title: David Sarnoff Professor of Management of Technology

1 Introduction

Digitization has dramatically transformed both the amount of information available for reuse, and the ability to rapidly leverage such information for new applications (Brynjolfsson et al., 2011). Businesses are increasingly using digitized information for decision-making and a number of startups, mobile applications and web-services are creating immense value by reusing content produced by third-parties and professional agencies in innovative new formats. Through reuse in various services, digitized information is being used in a wide number of industries ranging from health (Miller and Tucker, 2011; Freedman et al., 2013; Dranove et al., 2012), business strategy (Shiller, 2013; Bharadwaj et al., 2013), food safety (Luca, 2013), human-resource management (Aral et al., 2007; Tambe et al., 2008) to crime (Lee, 2011), academic research (Kim, 2012) and ecommerce (Dellarocas, 2003; Ghose and Ipeirotis, 2007; Smith et al., 2012; Chen et al., 2008) to improve decision-making and enhance productivity.

While there has been enthusiasm for the potential of such "big data" to transform the nature of business and technology both in the media and research (Scott and Varian, 2013; Freeland, 2012), there remain important legal questions that regulate the ability to reuse digital information for useful applications (Goldfarb and Tucker, 2011). Specifically, *copyright law* has the potential to shape the value that digitized information could deliver and yet, whether copyright enables smooth and legal exchange of data or hinders reuse through transaction costs remains an open question (Greenstein et al., 2010).

Economic models (Varian, 2006; Watt and Towse, 2006) give ambiguous predictions on whether copyright impedes or promotes the reuse of information once it has been created. Consider a simple example where two data-sets, otherwise similar in quality, differ only in their copyright status. Under one logic (Kitch, 1977; Hardin, 1968; Landes and Posner, 2002; Liebowitz and Margolis, 2004) the copyrighted data-set is *more* likely to be exploited in an innovative application because copyright provides incentives for investments in maintenance and marketing of the data-set, while the out-of-copyright dataset is likely to languish in the public domain, forgotten, un-

used or even abused. Under a competing logic however, copyright imposes transaction costs that prevent the exchange of information, and these transaction costs are likely to be significant given the inherent uncertainty in the value of information (Arrow, 1962) and the increasingly digital nature of knowledge production including declining costs of access and reuse (Lessig, 2005; Benkler, 2006; Zittrain, 2009; Lemley, 2004). When transaction costs are significant, out-of-copyright material is more likely to get reused and built upon. These two theories, while diametrically opposed, are clear in their respective predictions: copyright lubricates the market for information in one view, while it impedes the free reuse of data in the other.

Given contradictory theoretical predictions, systematic empirical analysis of the impact of copyright on reuse might help us make progress on this question. Unfortunately, such analysis has been largely absent in the economics, IT and management literature on the topic, perhaps due to empirical challenges. Issues of data and measurement are central. Unlike scientific output which can be measured through patents, the diffusion of information is more informal and harder to track. When data can be located, copyright applies to all information by default and persists often for over a hundred years, making it difficult to observe diffusion in the absence of copyright. Finally, even if these problems can be resolved, comparing reuse of copyrighted and non-copyrighted information is challenging because unobserved variables like the underlying quality of the data are often correlated with reuse, i.e. we do not know if information was reused because it was off-copyright or because it was of higher inherent quality.

In this paper, I exploit a natural experiment that occurred during a marquee digitization project in the history of the internet – the digitization of about 30 million works by Google Books. Specifically, during the digitization project in December 2008, Google Books digitized all existing issues of Baseball Digest, a prominent baseball magazine, and made them available online to readers for free.[1] Apart from the fact that Baseball Digest is perhaps one of the most important reference sources on the game of baseball, it is also interesting for my purposes because Baseball Digest

[1]Google has since taken the digitized version of Baseball Digest offline at least as of January 2013

issues printed before 1964 are out of copyright while those printed after are still under copyright (see Section 6.1). In other words, issues published after 1964 cannot be freely reused or modified. This study exploits this idiosyncratic variation in the copyright status of the digitization project to measure the impact of copyright on the reuse of underlying content on Wikipedia. Wikipedia is a natural venue to investigate this question being not only the fifth most visited website on the internet receiving about 10 billion page-views every month[2] but also because it is favorable for measurement, offering access to both past and present versions of information thereby allowing careful measurement of change in information in response to digitization.

To fix ideas, consider the example laid out in Figure 1. The example shows scanned pages from two issues of Baseball Digest, one about a feature on Felipe Alou published in 1963 (in an out-of-copyright issue) and another on Johnny Callison published in 1964 (in a copyrighted issue). For these two players Figure 1 depicts pages on Wikipedia as they appeared on December 2012. Neither of these two pages had the players' images before Baseball Digest was digitized, but in 2012, while Felipe Alou's page has an image from Baseball Digest, Johnny Callison's page has no images at all. Despite being printed in two issues that are close to each other in time, one image finds use in a broader context while the other seems lost among the pages of Baseball Digest. When the length of the text is compared however, both pages seem to have similarly detailed textual descriptions (containing over 50,000 characters of text in both cases). In terms of internet traffic, investigation reveals that the number of page-views to the Out-of-Copyright page increased by about 72 visitors per month (or 121%) between December 2008 and 2012, while it increased by only about 5 visitors per month (or 23%) for the In-Copyright page. While many alternative explanations could account for the differences in this case (differences in player popularity for example), the statistical analysis isolates the role of copyright in establishing these patterns.

Specifically, the estimation proceeds as follows. First, I identify a set of five hundred prominent baseball players, each of whom was active between 1944 and 1984.

[2]http://stats.wikimedia.org/EN/SummaryEN.htm

Further, because there was no comparable basketball magazine that was digitized at this time, I also include comparable set of basketball players in my analyses to provide an additional layer of controls. For these players I collect data on their Wikipedia pages before and after the digitization event including the number of images, text and internet traffic. The research design then proceeds in a differences-in-differences-in-difference (DDD) framework by constructing three levels of differences that provides me with an estimate of the impact of copyright on reuse in Wikipedia that is robust to a number of different alternative explanations. Specifically, I first construct differences-in-differences estimates that compare the change in content between In-Copyright pages (player debuts after 1964) and Out-of-Copyright pages (player debut before 1964) for both baseball players and basketball players separately, and then compare these two estimates to arrive at a DDD estimate of the impact of copyright on reuse.

The results show that the amount of content on baseball Wikipedia pages indeed increased post-digitization. However, copyright moderated this effect in important ways. First, before Baseball Digest was digitized I find no significant difference in the amount of content on In-Copyright and Out-of-Copyright players suggesting that, in this setting, copyright has little bite in a world without digitization and high costs of access. Second, once the content has been digitized, Out-of-Copyright pages benefit disproportionately as compared to In-Copyright pages as far as the reuse of images is concerned but not for text. Interviews with participants suggest that "fair use" features of Copyright law allow for the use of copyrighted text by permitting the paraphrasing of facts, while under the law it was difficult to reuse images without infringement. Finally, I find that changes in the amount of information have a meaningful impact on traffic to Wikipedia pages. Instrumental variables estimation suggests that an increase in images for Out-of-Copyright pages is associated with about a 25% increase in traffic. My calculations estimate that the lower bound of the loss to society from this diminished value of Wikipedia is on the order of $300,000 annually. More broadly, this research shows that digital platforms are creative constructions that rely on a number of different data sources (much like an academic

paper relies on the studies that it cites) and that copyright policy ultimately shapes both the nature of a novel digital platform and the value that it is able to generate for consumers.

The present study joins the nascent empirical literature on copyright. Of note is work in the legal literature on copyright that estimates the impact of copyright on availability and compares works produced before and after the US copyright cutoff date of 1923 (Heald, 2007, 2009a; Buccafusco and Heald, 2012). Other work that has also studied the impact of copyright on prices and access in the market for historical books (Li et al., 2012; Reimers, 2013) is also relevant. This paper adds to this literature by focusing on the causal impact of copyright on *reuse*, by using micro-data on knowledge flows and by separately identifying the interaction between digitization and copyright.

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 impact of the Baseball Digest copyright experiment. Section 4 concludes.

2 Empirical Context and Data

2.1 Empirical Context

Google Books, Baseball Digest and Copyright

Google Books is a Google initiative that has as its objective the digitization of all books ever published and currently offers a catalog of about 30 million books. 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.

Apart from its salience as a source of digital information, the case of Google Books is interesting because on 9th December 2008, Google Books announced that it would digitize magazines in addition to books,[3] including all issues of "Baseball Digest"[4]

[3]Official Google announcement can be found at this address: `http://bit.ly/googlebaseball`

[4]Other magazines digitized included New York magazine and Popular Mechanics

published between 1942 and 2008. Issues of Baseball Digest published before 1964 are out of copyright, while those published after will retain copyright till 2019 (see section 6.1 for more details). Therefore, even though all digital issues of Baseball Digest were freely available to be read, only those published before 1964 could be legally reused, providing a potential natural experiment for the impact of copyright on reuse.

Further, the experiment is likely to be economically meaningful given the widespread interest in both the game of baseball and in Baseball Digest. Over 45% of all Americans identify as baseball fans and revenues from the sport of baseball in 2010 were estimated to be approximately 7 billion USD. Baseball Digest has provided this vast fan-base with information and news about the game over seven decades since its founding in 1942.[5] Important for this study, in addition to news the magazine often contains profiles of baseball players and teams, particularly in the form of detailed articles, interviews and player images. Apart from being a source of entertainment for fans, magazines like Baseball Digest also prove to be a valuable resource for historians and writers who produce books, movies, TV and radio shows and in the case of Wikipedia, encyclopedic content around baseball topics. It is also useful to note that no similar publication about basketball was digitized around this time, a fact that will be exploited in the empirical analysis.

The Google Books digitization of Baseball Digest therefore represents a unique case where depending on the publication date of the periodical (before or after 1964) and the date of access (before or after December 2008), a widely-read information source differs across both the nature of digitization and copyright. Digitized issues are easily accessed after December 2008, but not before while issues published before 1964 are out of copyright while those published after remain under copyright. The role of digitization and the impact of copyright on the reuse of content is the empirical focus of the paper.

[5]See Jones (2007) and Brown (2011)

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 an analysis. First, Wikipedia is the pre-eminent source of information on the internet. 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 and the use of magazines like Baseball Digest is typical. Finally, each revision of a Wikipedia page is archived and publicly accessible allowing me to collect repeated panel data on Wikipedia pages both before and after a digital version of Baseball Digest was made available. This work builds on recent papers in the literature that have used similar data successfully for other questions (Zhang and Zhu, 2010; Greenstein and Zhu, 2012; Nagaraj et al., 2009; Algan et al., 2013; Gorbatai, 2012).

The dataset that I build is based on four different sources. First, I use the "Baseball Hall of Fame" voting dataset by Sean Lahman[6] to build a list of 514 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 allows me to include players who had finished their careers, and who had passed a screening committee judgment but also "removes from consideration players of clearly less qualification"(Abbott, 2011) and thereby includes those who merit encyclopedic inclusion. The dataset also provides biographical details of the players including date of debut and performance details like number of appearances and length of playing career. Baseball players in the sample have played an average of 1290 games in their career (sd=760.99), made their debut around 1966 and played for an average of 14.74 years (sd=3.8) in their career.

In order to measure the impact of digitization on baseball players against a control group of basketball players, I obtain data from `databasebasketball.com` which provides names, year of debut, point scored, assists, appearances and total career

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

minutes played. To identify a set of players comparable to baseball players, I use the dataset for the Top 1000 players by career minutes played and choose those who made their debut between 1944 and 1984.[7] Basketball players in the sample played about 658.74 games (sd=259), scored 8375.04 points (sd=5446.50) over their career and made their debut around 1970.

Having constructed this data, I then manually match the names of players to their respective pages on Wikipedia. Manual matching helps me avoid problems where a player with a common name like "Jackie Robinson" is matched to the Wikipedia page for Jack Robinson the politician, or worse, Jackie Robinson the basketball player. After having completed this matching, for each player page I download archival versions of the same page as it appeared on December 1, 2008 and December 1, 2012. I then build an automated parsing utility that allows me to measure the number of images and number of characters on a given Wikipedia page. For each page, I count images above 75 pixels in height (in order to avoid counting small images like icons and logos) and count characters that pertain to its core text (excluding Wikipedia template text but including formatting markup code that is likely to constant across all pages). This procedure extracts the number of images and the number of characters of text for a given Wikipedia player page at a point in time. Finally, I obtain web traffic data in the form of page-views from `stats.grok.se`. For each player page, I compute average monthly traffic data for the quarter September-December 2008 and September-December 2012 giving me a measure of average monthly page-views, a commonly adopted metric to measure traffic to web pages.[8]

Table 1 lists summary statistics for both baseball (Panel A) and basketball (Panel B) players. The data show that baseball Wikipedia pages contained on average 0.72 images (sd=1.37), 61503.36 characters (sd=58932.32) and receive on average 90.07 page-views per month (sd=162.95). For basketball players on the other hand, Wikipedia pages have on average 0.24 images (sd=0.81), 46287.76 characters (sd=54835.53)

[7]Using Baseball Hall of Fame nominations data is not feasible because the voting process does not begin until 1959

[8]Results are similar if I use only the average traffic level from the month of December, but measuring traffic in this way provides greater statistical power.

and receive on average 118.40 page-views per month (sd=673.84).

3 Empirical Results

Means of outcome variables by treatment and control groups are presented in Table 2. Analyzing the differences in the number of images for Out-of-Copyright and In-Copyright pages *before digitization* reveals that copyright has a small but significant impact on the use of images. While the difference in the number of images is close to zero for basketball players, the "before digitization" difference in the number of images for baseball players is about 0.093 images.

Table 2 also indicates that Out-of-Copyright baseball players have on average 0.303 images before digitization which increases to 1.672 images after. Correspondingly, traffic also increases from 75.568 to 141.226 page-views per month. For In-Copyright players however the gains are more modest. Images increase from 0.21 images per page to 0.906 images per page, while traffic increases from 57.188 to 100.554 page-views per month. The differences are not so stark for basketball players. Out-of-Copyright basketball players experience a gain of 0.262 images as compared to a similar 0.186 images for In-Copyright basketball players. Differences in mean traffic to these groups are somewhat different however, though standard errors are large. Out-of-Copyright basketball players gain about 67.317 page-views per month while In-Copyright players gain about 86.037 page-views per month. This suggests that more recent players are experiencing a greater increase in traffic as compared to older players over time.

The presentation of these data in the style of Gruber (1994) also allows us to estimate preliminary difference-in-difference estimates of the impact of copyright on baseball and basketball players in addition to the triple differences estimate of copyright on Wikipedia pages. Calculations indicate that Out-of-Copyright baseball players have a greater increase in images and traffic as compared to In-Copyright baseball players, while there seems to be no similar effect for basketball players. The triple difference estimate of the impact of copyright on reuse is about 0.598 for images and 41.012 for internet traffic. This section probes these results econometrically and ex-

plores these results further, notably by allowing player fixed-effects that helps control for systematic differences in player quality.

3.1 Impact of Digitization on Reuse

Before analyzing the impact of copyright, I first study the role the digitization of Baseball Digest on Wikipedia. I exploit the fact that while baseball pages were at risk of being affected by a digital baseball magazine, there was no comparable digital "*Basketball Digest*" that basketball player pages could have benefited from.

I estimate OLS models with the following specification:

$$Y_{itg} = \alpha + \beta_1 \cdot Post2008_t + \beta_2 \cdot Post2008_t \times Baseball_g + \gamma_i + \epsilon_{itg}$$

where the dependent variable is either $IMAGES_{itg}$, $TEXT_{itg}$ or $TRAFFIC_{itg}$ for player i for sport g in year t. The variable $Baseball_g$ is an indicator variable that equals one for baseball players and zero for basketball players. $Post2008_t$ is an indicator variable that equals one if the observation is from December 2012, and zero if it is from December 2008. Under the assumption that changes in dependent variable after December 2008 would have been comparable for both basketball and baseball players, the coefficient on $Post2008_t \times Baseball_g$ estimates the causal effect of digitization on the reuse of Baseball Digest content. Standard errors are clustered at player level; this allows for the potential serial correlation in the error terms at the level of players and uses repeated observations on pages of the same player to estimate standard errors robust to this problem. This is similar to the commonly used method of clustering standard errors at the state level in difference-in-differences analyses with individual–state–time level observations (Bertrand et al., 2004). Standard errors are clustered in a similar way for all the specifications reported in this study.

The results of the OLS regression models (Table 3), suggest that digitization has a significant impact on the amount of images and text for baseball pages on Wikipedia over time. Pages in the Post period have on average 0.202 more images (sd=0.038) and 30622.7 more characters (sd=1704.1). However, baseball pages have a much

greater increase in the number of images and text as compared to basketball pages. The estimates suggest that baseball players have 0.749 more images (sd=0.070) and 6917.5 more characters (sd=2345.5) in the Post period as compared to basketball players. This represents an increase of 157% in images (as compared to an average of 0.479 images per page) and an increase of 12.82% for text (as compared to an average of 53925.28 characters per page) due to digitization. However, these changes do not seem to be associated with an increase in traffic on baseball pages. In fact, estimates suggest that baseball players see a reduction of about 30.43 page-views a month, even though estimates are noisy and not significant. Table B.1 in the Appendix establishes the robustness of these measures to log specifications.

3.2 Impact of Copyright on Reuse

Having established that Wikipedia pages benefit from digitization in terms of content, I now turn to analyzing the central question of this study: does copyright affect the reuse of digitized content on Wikipedia?

Graphical Analysis

Figure 2 (Panel A) plots the data for change in images after digitization for baseball players in the Out-of-Copyright (debut before 1964) and In-Copyright (debut after 1964) groups. First, for each player the change in the number of images between 2008 and 2012 is calculated. Then for each player debut cohort, I calculate the mean change in the number of images after digitization. The mean change is plotted in dark-gray bars for players in the Out-of-Copyright group, while the bars in light-gray represent players in the In-Copyright group.

The resulting plot seems to indicate that baseball players in the Out-of-Copyright group have a consistently larger increase in the number of images after digitization as compared to players in the In-Copyright group. The mean change for players in the Out-of-Copyright group is 1.369 images as compared to a mean increase of 0.696 images for players in the In-Copyright group. As an illustration, all baseball players

in the data who made their debut in 1957, for example, gained about 1.6 images on average while those making their debut in 1972 had gained only about 0.47 images. The corresponding increases in traffic are equally striking. 1957 players gained about 43.67 page-views on average while 1972 players gained only 4.1. Basketball player data does not seem to show a similar pattern (Figure 2 Panel B). Out-of-Copyright players have a mean change of 0.262 images while In-Copyright players have a similar change of 0.187 images.

If the effects of copyright are truly discontinuous at 1964, then one concern might be the lack of a similar discontinuity in Figure 2. Specifically, the reuse of images seems to be relatively high for players classified between the years 1944-1960, and then reduces uniformly between 1960 and 1964, posing, perhaps a threat to the identification strategy. However, further investigation reveals that the reduction in reuse for players between the years 1960-1964 is likely caused by my conservative classification rule: i.e. because these players made their debut in the year 1960, images for "Out-of-Copyright" players are more likely to appear in copyrighted issues of the magazine published after 1964 rather than in out-of-copyright issues published before. Given that I do not have detailed data about the images featured in Baseball Digest by issue, the coarse (but conservative) classification measure that I use, is likely to count players affected by copyright in the "Out-of-Copyright" category, especially if they made their debut close to 1964.

In order to make sure that this misclassification does not upward bias my estimates in anyway, in the regressions I present, I will include individual player and year fixed effects and compare baseball players to basketball players which alleviates concerns that the estimates are being driven by a systematic difference in quality between baseball players who played before 1964 and those who played after. Further, I present a version of Figure 2 that adjusts more directly for issues in the classification of players who made their debut close to the cutoff year of 1964 in Figure 3. The adjustments I make are as follows. First, in order to understand the extent of misclassification, for each player, I calculate a measure of misclassification ϕ, i.e. the percent of a given player's career that has been played before 1964. By this measure a player who makes

his debut in 1960, and retires in 1980 would've played 25% of his career in the out-of-copyright period. Having calculated ϕ for each player, I calculate the mean change in the number of images by debut cohort (as in Table 2), but scale these differences by dividing by the *average* value of ϕ for every debut cohort.

Figure 3 plots calculated values of "Percent In-Copyright" $(1-\phi)$ as the black line, and the rescaled differences for each debut cohort in the bar chart. As expected, for early debut cohorts, the value of $(1-\phi)$ is close to zero, but approaches one close to 1964. Further, once these differences are accounted for, we see a more discontinuous difference between players who made their debut before and after 1964. This provides confidence that differences in levels of reuse are likely to be driven by a difference in copyright status before and after the year 1964, rather than other confounding factors.

Within-Baseball Estimates

Overall, Figure 2 (A) seems to indicate that Out-of-Copyright baseball players seem to have benefited disproportionately as compared to players in the In-Copyright group after digitization.

The following regression tests this idea formally using only the set of baseball players:

$$Y_{it} = \alpha + \beta_1 \cdot Post2008_t + \beta_2 \cdot OutOfCopyright_i \times Post2008_t + \gamma_i + \epsilon_{it}$$

where the dependent variable is either $IMAGES_{it}$ or $TRAFFIC_{it}$ on a Wikipedia page for player i in year t. The indicator variable $OutOfCopyright_i$ equals one for players who made their debut appearances between 1944-1964 and zero for players who made their debut appearances between 1964 and 1984. Indicator variable $Post2008_t$ equals one if the observation is from December 2012 and zero if it is from December 2008. γ_i indicates player fixed effects. Under the assumption that changes in dependent variable after December 2008 would have been comparable for both In-Copyright and Out-of-Copyright players, the coefficient on $OutOfCopyright_i \times Post2008_t$ estimates the causal effect of a lack of copyright on the reuse of content on

Wikipedia.

Baseline regressions indicate that content changes differentially for In-Copyright and Out-of-Copyright players after the digitization of Baseball Digest. Pages after digitization have on average 0.696 more images (sd=0.063, Table 4, Column 1) than player pages before digitization. However, post-digitization the number of images for Out-of-Copyright players increased by an additional 0.673 compared to players in the In-Copyright group (sd=0.128, Table 4, Column 1). Relative to an average level of 0.721 for the number of images on a given page, this represents a 93% increase in response to the absence of copyright on issues of Baseball Digest. This increase in the number of images is also accompanied by a positive effect on traffic. Players in the Out-of-Copyright group have about 22.29 more page-views per month (sd=9.59, Table 4, Column 3) as compared to players in the In-Copyright group, indicating an increase of about 24.74% in page-views due to a lack of copyright. While the impact on text is also reported, this estimate is shown to not be robust to across-sport comparisons as explained in Section 3.2. Proportional models where the dependent variables are transformed to logs reported in the appendix seem to be in accordance with these results (Table B.2).

Triple Difference Estimates

Having established that digitization affects baseball pages positively and that copyright seems to hurt the extent of such reuse for baseball players, I now turn to the central set of results from this study by comparing baseball players against a control set of basketball players. The validity of the results in Table 4 is in question if Out-of-Copyright players were accruing content at an increasing rate on their pages as compared to In-Copyright players even before the digitization of Baseball Digest. Different time trends could be as a result of a revival in interest over time in an older generation of players in sport due to changing fashions (Hahl, 2012) or due to newer players gaining popularity faster than older players for example.

Graphical Analysis

First, in order to further investigate this issue I collect additional data that measures the amount of images for player pages in the month of December annually between 2006 and 2012. This allows me to estimate time-varying coefficients of the impact of copyright.

Time varying coefficients for Images (Figure 4 Panel A) reveal no discernible evidence in the increase in images for players in the Out-of-Copyright group before the digitization event. Time varying coefficients are close to zero before the digitization event and two years after. In the final two years of analysis, the number of images for Out-of-Copyright players increases substantially indicating that, in this context, reuse occurs after a lag of about two years. In interviews I conducted with Wikipedia editors, the process seems to have been one where editors discovered the existence of non-copyrighted source of images in about 2011 and were then motivated to transfer a large number of images en masse in a short space of time. While all Out-of-Copyright players in the sample are unlikely to have been photographed in the Baseball Digest, given the popularity of the players in the sample, a large number of them are likely to have been found. A similar analysis for basketball players (Panel B) reveals no similar pattern. For basketball players there is no discernible difference in the mean between the two groups over the period of analysis, providing further confirmation that the estimates for the impact of copyright on images in the within-baseball analysis is unlikely to have been driven by changing fashions.

Regression Estimates

In order to test formally the role of a growing preference towards older players over time as a source of the differences in time varying estimates, I consider the difference-in-difference estimates for baseball players and compare them with similar estimates for basketball players in a triple differences framework. Such a framework allows us to control for separate intercepts on time-group combinations, allowing the model to freely account for period level changes in a group's popularity over time (say).

Specifically, I estimate OLS regressions using the following specification:

$Y_{itg} = \alpha + \beta_1 \cdot Post2008_t + \beta_2 \cdot OutOfCopyright_i \times Post2008_t + \beta_3 \cdot Post2008_t \times Baseball_g + \beta_4 \cdot Post2008_t \times Baseball_g \times OutOfCopyright_i + \gamma_i + \epsilon_{itg}$

where Y_{itg} represents the number of Images, Text or Traffic for player page i in year t and in sport group g (either basketball or baseball). $OutOfCopyright_i$ is an indicator variable equal to one if a player made his debut in the period between 1944 and 1964, $Post2008_t$ is an indicator variable that equals one if observations are from December 2012 and $Baseball_g$ is an indicator variable that equals one for baseball players. β_4 is the coefficient of interest, indicating the difference-in-differences-in-differences estimate of the impact of copyright on Out-of-Copyright baseball players in the Post period. Specifically, β_4 compares baseball and basketball players with respect to the change in content over time between In-Copyright and Out-of-Copyright player groups.

The results (Table 5) indicate that the reuse of images on Wikipedia pages does seem to be linked causally to the digitization of Baseball Digest. The DDD estimate of β_4 is 0.598 (sd = 0.149, Table 5) in the case of images and is significant at the 1% level. This represents a nearly 124% increase in the reuse of images and is comparable to the coefficient of 0.673 from baseline estimates (Table 4). The results also indicate that Wikipedia pages in the Out-of-Copyright group experience a greater increase in traffic as compared to the In-Copyright group. The DDD estimate of β_4 is 41.01 (sd = 32.10), but this coefficient loses significance when player fixed effects are added. Internet traffic tends to be highly volatile, leading perhaps to nosier estimates in this case. Despite this caveat, taken together with the difference-in-difference estimates, it is reasonable to interpret that copyright seems to have a negative effect on internet traffic to pages that did not benefit from Baseball Digest content. Finally, I also consider the impact of copyright on text. The results indicate that the estimates for β_4 in regressions that consider the reuse of text are slightly negative and not significant at the 10% level (Table 5) stressing that copyright does not seem to impact the reuse of text in the same way that it does images. Interviews with Wikipedia editors suggests that while it is easy to summarize text without violating copyright, doing so

for images is hard. The intricacies of "fair use" law, and the differing nature of the underlying media therefore seems to be one reason for this variation on the impact of copyright. Proportional models that use log-transformed dependent variables (Table B.3) seem to be in accordance with these results.

Combined, the within-baseball estimates, the time variation in impact and the triple differences estimates support the theory that copyright hurts the ability of creative inputs to be reused. Copyrighted issues of Baseball Digest are less likely to be a source of images on Wikipedia and copyright negatively impacts the level of traffic that a page receives over time.

3.3 Is an Increase in Images Related to an Increase in Traffic?

Results in Table 5 seem to suggest that In-Copyright players seem to suffer a penalty in terms of internet traffic after the digitization of Baseball Digest. This section tests the hypothesis that In-Copyright pages have lower increases in traffic after digitization because a lower level of images makes such pages less valuable, making it less likely than a user will access the page. Anecdotal evidence seems to suggest that the role of image search engines like Google Images that drive substantial traffic to Wikipedia pages through users looking for images of a particular player, and of search engines that list "thumbnail" images next to search results (thereby increasing the probability of being clicked on) could both be important channels for such an effect.

I perform three tests to investigate the possible link between the number of images and level of traffic. First, Figure 5 plots the change in traffic on the vertical axis against the the change in the number of images for Hall of Fame players (Panel A) and all players (Panel B) on the horizontal axis. If traffic increases with the number of images then one expects these variables to be positively correlated. Figure 5 shows that the correlation coefficient between the change in traffic and change in the number of images is 0.432 for Hall of Famers and 0.485 for the full sample. This indicates that players whose pages are associated with a larger change in images are also associated with an increase in Internet traffic. Mickey Mantle's page (debut 1951) for example gains 6 images over this period and is associated with an increase in over 600 hits per

month.

While this evidence is suggestive, in order to control for player-level differences, I run a simple OLS regression at the page level, where I regress the level of traffic on the number of images for baseball player pages and a suite of player level controls. Table 6 indicates that an increase of one image is associated with an increase in about 64.55 page-views. However, when player level fixed effects are added this estimate reduces to 36.14 page-views per image, suggesting that player level differences not captured by the controls is driving a large part of this effect. While player fixed effects are useful, a problem persists in that images are possibly added to player pages when there is greater interest in the player. Wikipedia authors are often final consumers of the finished product and might be motivated to upload an image precisely when a player is in the news (and his page has higher levels of traffic), for example. Therefore, naive OLS estimates of the impact of images on traffic, even with player fixed effects, could be biased upwards if the timing of the addition of new images is endogenously affected by the level of traffic.

In the final step of the analysis, I use instrumental variables estimation to tackle this problem. Specifically, I instrument for the number of images in a given year with a dummy for $OutOfCopyright_i \times Post_t$. We expect this instrument to be correlated with the number of images (because images were more likely to be available for Out-of-Copyright pages in the *Post* period) but the timing of digitization is independent of the level of traffic to baseball pages on Wikipedia. Before presenting the results, caution must be exercised in interpreting them to be the causal impact of images because the exclusion restriction is tentative here. Specifically, while the amount of text did not change substantially following digitization, other unmeasured aspects of the page content (like the *quality* of the text) could have be affected by the instrument leading to higher levels of traffic.

Table 6 shows first stage estimates indicating that the chosen instrument is strongly correlated with the number of images; the instrument seems to explain a large portion of the variation in the number of images (and the corresponding F-Statistic is also reasonably large). The IV estimates confirm the expected upward bias in OLS

estimates indicating that for every image added, traffic seems to increase by about 33.11 page-views per month (sd=12.62). Note however that this estimate is quite similar to the OLS estimate of 36.14 when fixed effects are added, indicating that the timing of the addition of images is driving only a small portion of the bias in the OLS estimates.

Ultimately, the evidence from these three tests taken together seems to suggest strongly that an important channel through which copyright affects internet traffic to Wikipedia pages is the reduced number of images on a given page caused by copyright.

4 Discussion

While the increased use of digitized information in media and business in increasingly common, this study invesigates how copyright law might shape how economic value from digitized information is derived. Without the answer to this question, copyright policy remains stalled. A report released by the Republican Study Committee (Khanna, 2012) on copyright states the basic problem quite succinctly:[9]

> We frankly may have no idea how [copyright] actually hurts innovation, because we don't know what isn't able to be produced as a result of our current system.

4.1 Findings

This paper suggests one way to estimate the impact of copyright on the ability to reuse digitized information in the light of a plausible counterfactual. The study reveals that copyright does seem to have a negative impact on digital reuse. A number of results of the study point to this conclusion. First, reuse of content on Wikipedia from Baseball Digest increases substantially post-digitization. When compared to basketball players, baseball players have a 157% increase in images and 12.82% increase in the amount of text on their pages. Second, copyright reduces the level of

[9] See `http://bit.ly/TewlRN` for details.

reuse once digitized content is made available – players who make their debut after 1964 are less likely to benefit from digitization than are players who played before 1964. Third, the impact of copyright on reuse is particularly salient for the reuse of images, while text seems less affected. Out-of-Copyright players had on average 93%-124% greater number of images after digitization as compared to In-Copyright players. Finally, copyright on Baseball Digest post-1964 is shown to have an impact on internet traffic. Out-of-Copyright player pages receive approximately 25% more visits per month as compared to In-Copyright players. I also find a positive correlation between the increase in the number of images and internet traffic to pages. Instrumental variables estimates suggest that the increase in images leads directly to higher levels of internet traffic on baseball player's pages. Finally, a back-of-the envelope calculation suggests that a lower bound on the loss to social welfare from copyright is about $267,335 annually for Wikipedia. See Appendix 6.2 for details.

4.2 External Validity

One concern with the results could be 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 could be unusually stringent in enforcing copyright causing my estimates to be biased upwards. While anecdotal evidence does suggest that Wikipedia is vigilant about enforcing copyright, it is hardly an isolated case. A number of platforms where one might expect reuse of digitized information have extensive programs for enforcing copyright including YouTube (Seidenberg, 2009), Amazon, all major mobile application stores and even Google's search engine (Dillon Scott, 2011). For example, Apple's AppStore rejected about a thousand application in August 2009 because they were using copyrighted images and books in their applications[10] and hosts an online tool where firms might report copyright violation. Google removed about 26 million links in October 2013[11] from its search index including links that provided access to copyrighted books, music and data. An

[10]See bit.ly/1aXpksj

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

extensive literature on piracy and DRM (Bechtold, 2004) in the entertainment industry has also shown that copyright-related interventions that limit the availability of digitized content are quite common, and often quite effective (Danaher et al., 2010; Danaher and Smith, 2013).

To further ease concerns about external validity, my study builds on the emerging empirical literature on the effects of copyright that suggests that copyright has a negative effect on access, a first step to reuse. Paul Heald and coauthors (Heald, 2007, 2009b; Buccafusco and Heald, 2012) have shown that that works produced before 1923, a period in which works are generally in the public domain, are much more accessible today as compared to works produced after 1923. Books produced before 1923 for example, are more easily accessible on Amazon and Audible.com and are more likely to be digitized (Brooks, 2005). A more recent study in the economics literature (Reimers, 2013), analyzes the market for books in a similar time period and finds that a copyright extension decreases welfare from fiction bestsellers by decreasing variety, causing a decrease in consumer surplus that outweighs the increase in profits. Similarly, a study of the market for fiction in the 1820s, also shows that an important impact of copyright is likely to be an increase in the price of books, perhaps reducing 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.

4.3 Contributions

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. While existing studies address an important margin of copyright law, i.e. the ability of readers to access copyrighted content, an important channel through which existing knowledge might impact productivity is through its digitization and reuse in novel applications. This study measures, to my knowledge, the first causal measurement of copyright on reuse, where the underlying work is a source for

creating derivative work. The result that copyright harms the reuse of content is also novel, and stands in contrast to a previous study (Heald, 2009b) that measures the impact of copyright on reuse, and suggests that copyright has little effect in preventing reuse.[12] The results are in line with a similar stream of work on the empirical effects of intellectual property on the diffusion of scientific knowledge (Murray and Stern, 2007; Murray et al., 2009; Williams, 2013; Furman and Stern, 2011) that find a generally negative effect of intellectual property on follow-on use. From a policy point of view, this paper is able to address directly questions that are likely to be important going forward: (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?

In addition to the precise estimation exercise I perform, methodologically, the 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 (see Garg et al. (2011) for another example). 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 and as the internet itself is shaped by external institutions including copyright laws. Second, in the 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 – i.e. while patents protect the idea, copyright protects only the "expression", i.e. the impacts of copyright are likely to vary not simply by the quality of the data, but also but the medium of expression. This distinction is likely to be important going forward. Finally, the specific case of periodicals is useful because it allows for the measurement of the impact of copyright *within* a given publication and

[12]Heald (2009b) analyzes the differences in the rates of reuse of popular public domain and copyrighted songs between 1909-1932 in movies and finds little difference, but controlling for the underlying quality of the music works is challenging in this setting. The paper itself outlines this challenge quite well, stating that "a substantial majority of the compositions (44 out of 70) were published in the six-year period from 1926-31, indicating the significance of the golden age of Tin Pan Alley, an extraordinary time period which marked the publication of many enduringly familiar works ...". The present study is able to address such identification challenges directly.

because a large number of important periodicals saw their copyright expire in 1964 (including journals like the the American Economic Review) providing avenues for similar studies in other domains going forward.

Finally, while this paper does not evaluate the overall welfare consequences of copyright (specifically the impact of copyright on the incentives to create new content is not examined), it is still useful because it helps us estimate what potential incentives copyright needs to provide to creators in order to justify the losses estimated to society. Even within the case that I study, we might be skeptical of the overall welfare gains from removal for copyright protection of archival Baseball Digest issues to be large if producer surplus in this case is significant, i.e. if licensing content from Baseball Digest is a major source of revenue for its producers. However, I made a number of reasonable attempts to contact the publishers of Baseball Digest in order to investigate the possibility of licensing content for reuse, but my requests were met with no response, suggesting that in this case producer surplus from licensing archival material is fairly low. More broadly, copyright issues often arise in policy discussions for works already created such as with the "Mickey Mouse" law of 1998[13] or for orphan works whose creators cannot be located (Smith et al., 2012). This paper helps estimate the impact of the losses to society from retroactive extensions or difficulties in locating missing authors.

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

5 Tables and Figures

Table 1: Summary Statistics

Panel A : Baseball Players

	Mean	SD	Median	Min	Max
Images	0.72	1.37	0.00	0.00	9.00
Traffic	90.07	162.95	33.33	0.00	1849.67
Text	61503.36	58932.32	40932.00	2382.00	379995.00
Debut Year	1966.17	10.26	1966.00	1944.00	1984.00
Debut Before 1964 (pct)	0.38	0.49	0.00	0.00	1.00

Panel B : Basketball Players

	Mean	SD	Median	Min	Max
Images	0.24	0.81	0.00	0.00	11.00
Traffic	118.40	673.84	17.83	0.00	16295.67
Text	46287.76	54835.53	28802.00	2428.00	520000.00
Debut Year	1970.51	9.24	1971.00	1947.00	1984.00
Debut Before 1964 (pct)	0.20	0.40	0.00	0.00	1.00

Note: This table present summary statistics of the sample of players and corresponding Wikipedia pages used in this study. Panel A presents data from the sample which includes all 514 baseball players who made their debut between 1944 and 1984 and were nominated for the Baseball Hall of Fame. Panel B presents data from the sample which includes all 510 basketball players who made their debut between 1944 and 1984 and are included in the list of Top 1000 players by career minutes played. For each player the data includes outcome measures for page revisions as they appeared in December 2008 and December 2012. For each page revision, *Images* measures the number of images above a width of 75 pixels, *Text* measures the number of characters (including formatting code) and *Traffic* measures the average monthly page-views for a given Wikipedia page.

Table 2: Difference-in-Difference-in-Difference Estimates of the Impact of Copyright on Wikipedia

PANEL A. Treatment Individuals : Baseball players who played between 1944 and 1964

Player Group	Before Digitization		After Digitization		Time Difference	
	Images	Traffic	Images	Traffic	Images	Traffic
Debut (1944 – 1964)	0.303	75.568	1.672	141.226	1.369	65.658
N=195	(0.051)	11.234	(0.142)	(17.886)	(0.151)	(21.122)
Debut (1964 – 1984)	0.21	57.188	0.906	100.554	0.696	43.366
N=319	(0.035)	(4.908)	(0.080)	(8.318)	(0.087)	(9.658)
Group Difference	0.093	18.379	0.766	40.672		
	(0.060)	(10.793)	(0.151)	(17.567)		
		Images	*Traffic*			
Difference-in-Difference		0.673	22.292			
		(0.128)	(9.586)			

PANEL B. Control Individuals : Basketball players who played between 1944 and 1964

Player Group	Before Digitization		After Digitization		Time Difference	
	Images	Traffic	Images	Traffic	Images	Traffic
Debut (1944 – 1964)	0.136	56.032	0.398	123.35	0.262	67.317
N=103	(0.073)	(19.057)	(0.107)	(39.284)	(0.130)	(43.662)
Debut (1964 – 1984)	0.135	82.645	0.321	168.682	0.186	86.037
N=407	(0.033)	(24.227)	(0.043)	(45.608)	(0.054)	(51.518)
Group Difference	0.001	-26.613	0.076	-45.332		
	(0.075)	(49.133)	(0.101)	(92.843)		
		Images	*Traffic*			
Difference-in-Difference		0.0754	-18.72			
		(0.076)	(30.668)			
		Images	*Traffic*			
DDD estimate		**0.598**	**41.012**			
		(0.149)	(32.125)			

Note: This table presents mean outcomes (images and traffic) for Wikipedia page revisions, by date (before or after digitization) by sport (baseball or basketball) and by player group (Out-of-Copyright or In-Copyright). Unadjusted standard errors are reported for all group means while standard errors for differences of means have been adjusted for clustering at the player level.

Table 3: Difference-in-Difference Regressions Estimating Impact of Digitization on Wikipedia

	(1) Images	(2) Traf.	(3) Text
Baseball X Post	0.749***	-30.43	6917.5***
	(0.0703)	(18.99)	(2345.5)
Post 2008	0.202***	82.26***	30622.7***
	(0.0377)	(18.47)	(1704.1)
Constant	0.190***	70.69***	36877.8***
	(0.0176)	(4.732)	(586.3)
Player FE	Yes	Yes	Yes
adj. R^2	0.272	0.0481	0.455
N	2048	2048	2048

Robust standard errors clustered at player level are reported
* p < 0.10, ** p < 0.05, *** p < 0.01
Mean(Img) = .479, Mean(Text) = 53925.279
Mean(Traf) = 104.18

Note: This table investigates the impact of digitization of Baseball Digest on Wikipedia page-level outcomes. Level of observation is a player-page and data includes pages on all 1024 baseball and basketball players in the sample from December 2008 and 2012. Estimates are reported from an OLS regression with the specification: $Y_{itg} = \alpha + \beta_1 \cdot Post2008_t + \beta_2 \cdot Post2008_t \times Baseball_g + \gamma_i + \epsilon_{itg}$ where γ_i represents fixed effects for each individual player. Co-efficient on $Baseball_i$ is not estimated because it is collinear with player fixed effects.

Table 4: Difference-in-Difference Regressions Estimating Impact of Copyright on Wikipedia

	(1) Images	(2) Traffic	(3) Text
Out-of-Copy X Post	0.673***	22.29**	7287.8**
	(0.128)	(9.586)	(3484.8)
Post 2008	0.696***	43.37***	34775.3***
	(0.0628)	(5.039)	(1855.1)
Constant	0.245***	64.16***	42733.3***
	(0.0288)	(2.200)	(802.8)
Player FE	Yes	Yes	Yes
adj. R^2	0.372	0.219	0.518
N	1028	1028	1028

Robust standard errors clustered at player level are reported

* p < 0.10, ** p < 0.05, *** p < 0.01

Mean(Images) = .721, Mean(Traffic) = 90.072, Mean(Text)=61503.361

Note: This table investigates the impact of copyright applied to Baseball Digest issues published after 1964 on Wikipedia page-level outcomes. Level of observation is a player-page and data includes pages on all 514 baseball players in the sample from December 2008 and 2012. Estimates are reported from a regression with the specification: $Y_{it} = \alpha + \beta_1 \cdot Post2008_t + \beta_2 \cdot OutOfCopyright_i \times Post2008_t + \gamma_i + \epsilon_{it}$ where γ_i represents fixed effects for each individual player. Co-efficient on $OutOfCopyright_i$ is not estimated because it is collinear with player fixed effects.

Table 5: DDD Regressions Comparing Impact of Copyright between Baseball and Basketball Wikipedia Pages

	(1) Images	(2) Traffic	(3) Text
Post X Baseball X Out-of-Copy	0.598***	41.01	-7299.0
	(0.149)	(32.10)	(5709.6)
Out-of-Copy X Post	0.0754	-18.72	14586.9***
	(0.0759)	(30.64)	(4524.1)
Post X Baseball	0.509***	-42.67*	7098.6***
	(0.0770)	(23.11)	(2607.9)
Post 2008	0.187***	86.04***	27676.7***
	(0.0446)	(22.55)	(1833.9)
Constant	0.190***	70.69***	36877.8***
	(0.0172)	(4.732)	(581.6)
Player FE	Yes	Yes	Yes
adj. R^2	0.302	0.0481	0.463
N	2048	2048	2048

Robust standard errors clustered at player level are reported

* p < 0.10, ** p < 0.05, *** p < 0.01

Mean(Images) = .479, Mean(text) = 53925.279 and Mean(Traffic) = 104.18

Note: This table uses the triple differences framework to estimate the impact of copyright on Baseball Digest issues published after 1964 on Wikipedia page-level outcomes. Level of observation is a player-page and data includes pages on all 514 baseball players and 510 basketball players in the sample and Wikipedia page revisions from December 2008 and 2012. Estimates are reported from an OLS regression with the specification:

$Y_{itg} = \alpha + \beta_1 \cdot Post2008_t + \beta_2 \cdot OutOfCopyright_i \times Post2008_t + \beta_3 \cdot Post2008_t \times Baseball_g + \beta_4 \cdot Post2008_t \times Baseball_g \times OutOfCopyright_i + \gamma_i + \epsilon_{itg}$

for player i in year t and in sport group g where γ_i represents player fixed effects. Main effects for $Baseball_i$ and $OutOfCopyright_i$ and interaction term $Baseball_i \times OutOfCopyright_i$ are not estimated because they are collinear with player fixed effects.

Table 6: Is an Increase in Images Associated With an Increase in Traffic?

	OLS		First Stage		IV Estimates	
	Traffic	Traffic	Images	Images	Traffic	Traffic
Images	64.55***	36.14***			53.11**	33.11**
	(6.693)	(6.579)			(19.73)	(12.62)
Out-Of-Copy. X Post			0.766***	0.673***		
			(0.163)	(0.128)		
Controls	Yes	Player FE	Yes	Player FE	Yes	Player FE
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	1028	1028	1028	1028	1028	1028
adj. R^2	0.519	0.397	0.157	0.373	0.405	0.395
F-Stat			22.19	27.6		

Standard errors clustered at player level are reported

* p < 0.10, ** p < 0.05, *** p < 0.01

Note: This table investigates the impact of lost images due to copyright on traffic. Level of observation is player-page and the sample is restricted to baseball players. *Controls* include player characteristics: Number of appearances, Hall of Fame status, Age at debut and Career length.

$Out - Of - Copy \times Post$ instruments for the number of images on a given page. IV models use a two-step least squares (2SLS) estimator for estimation.

Figure 1: An Illustration of How Copyright Might Affect Content Reuse

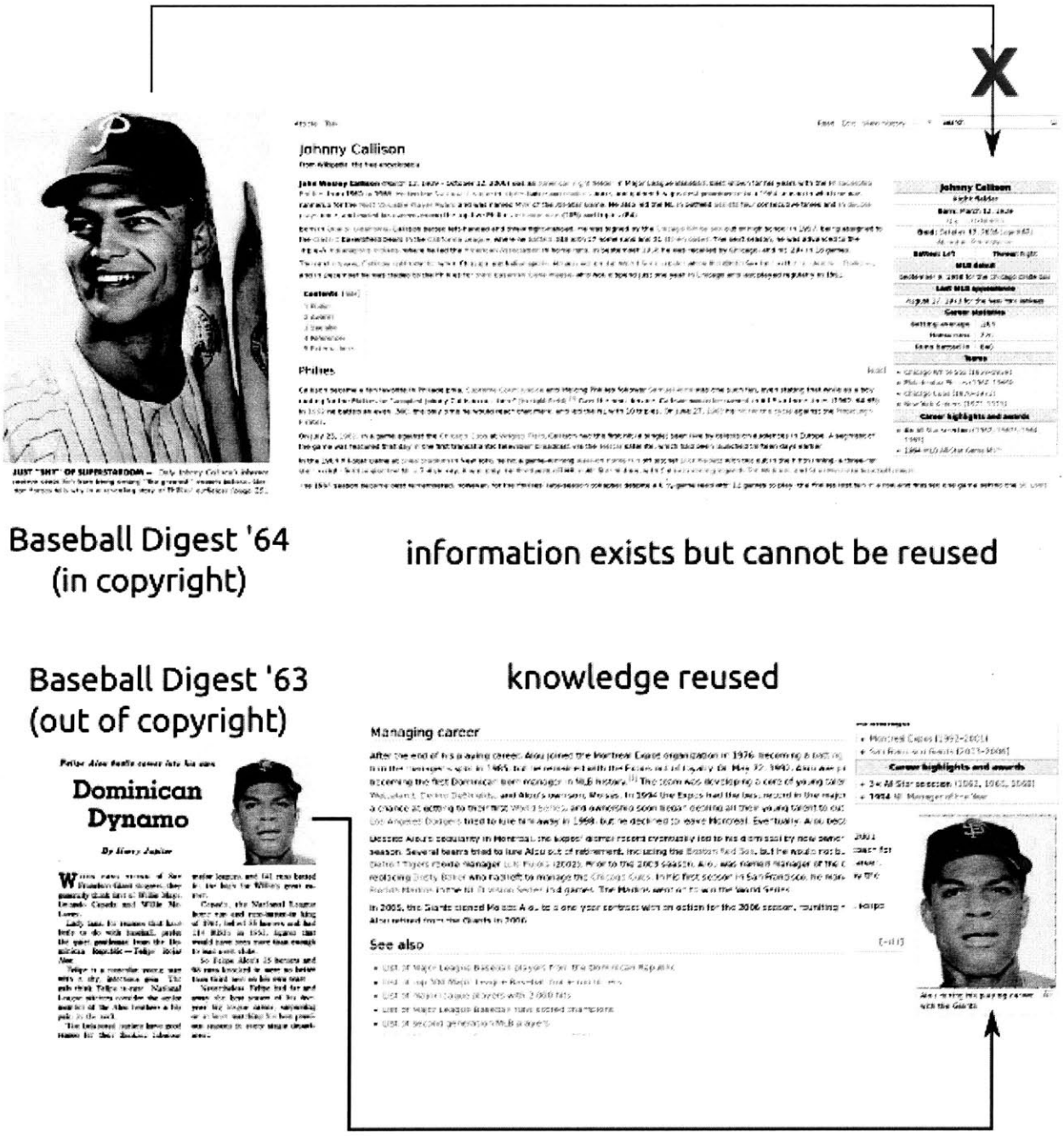
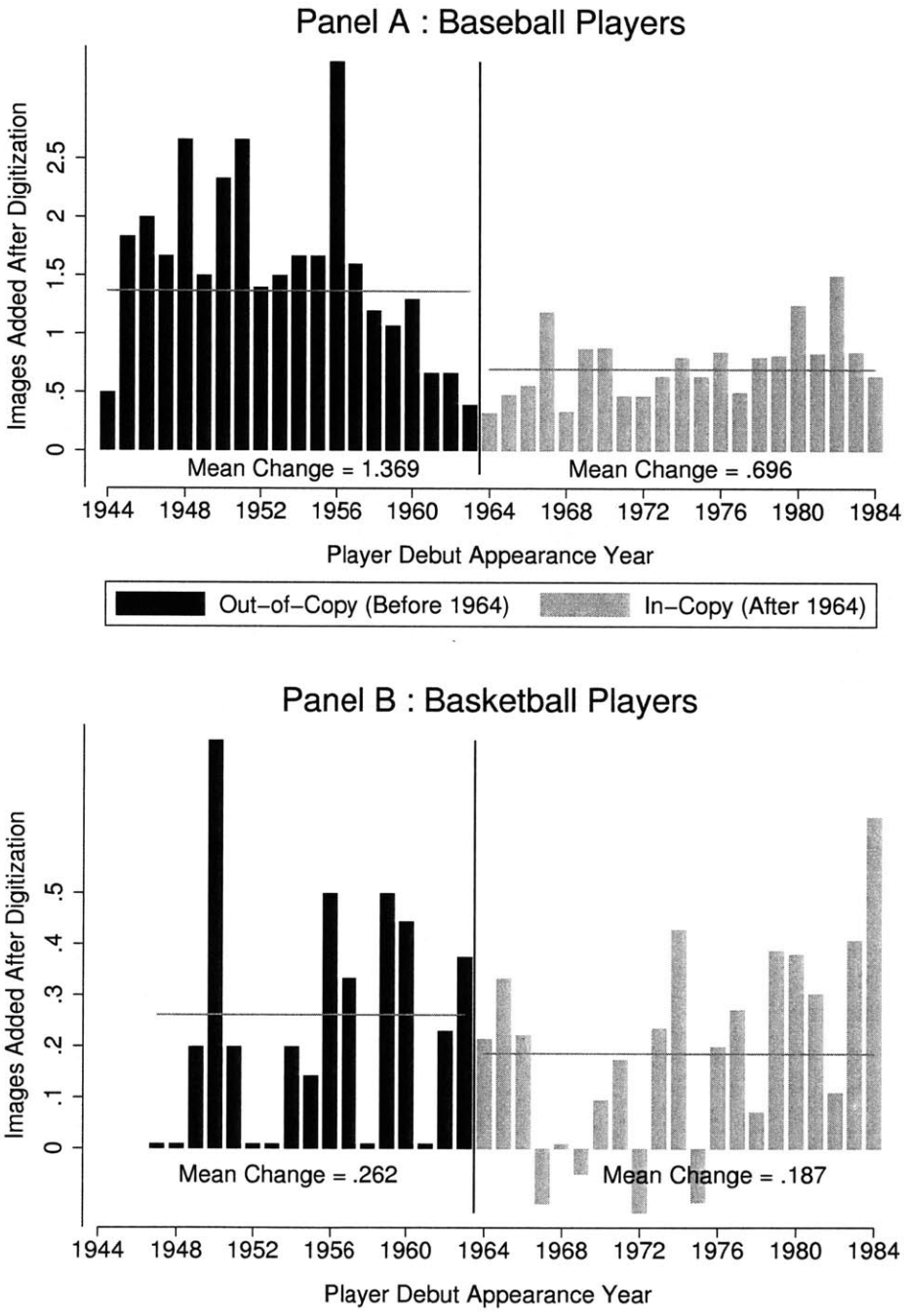
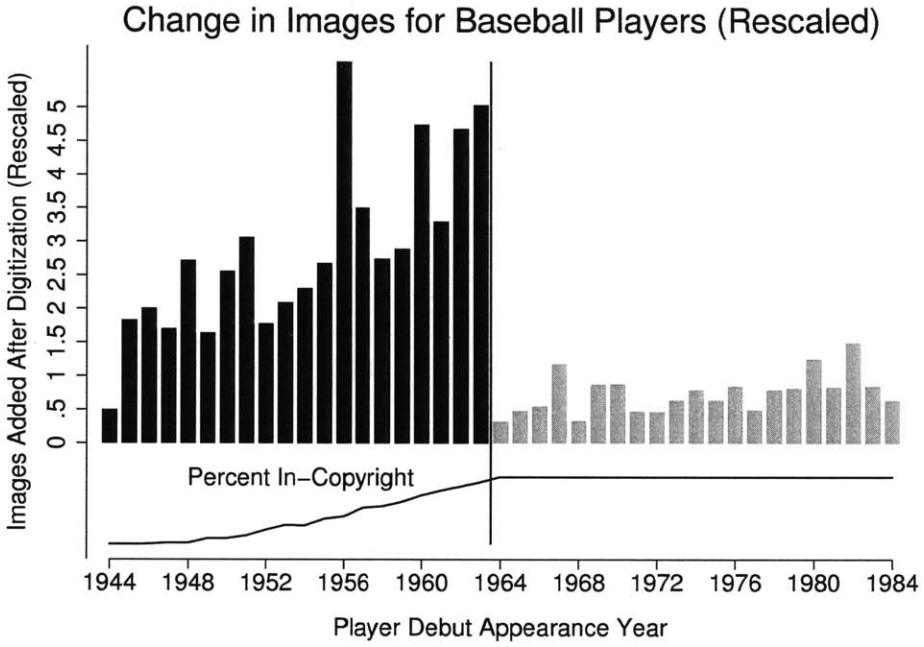


Figure 2: Mean Images Added (By Debut Cohort) After Digitization of Baseball Digest



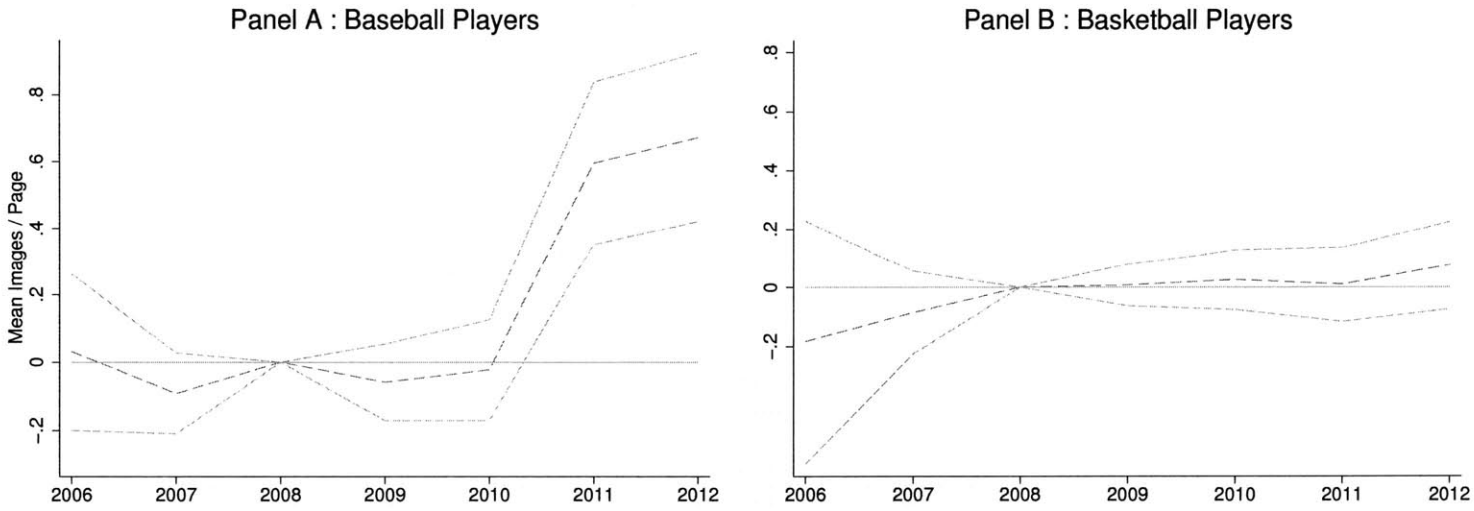
Note: For every debut cohort of baseball and basketball players, the mean number of images added on their Wikipedia page between December 2008 and December 2011 is calculated and plotted, i.e. because 28 baseball and 12 basketball players made their debut in 1959; this figure plots the mean change in the number of images on a given Wikipedia page for this cohort for the year 1959. See Section 3.2 for a discussion of why the impact of reuse of images for baseball players seems to be reducing between the years 1960-1964.

Figure 3: Mean Images Added (By Debut Cohort) Adjusting for Issues in Player Classification



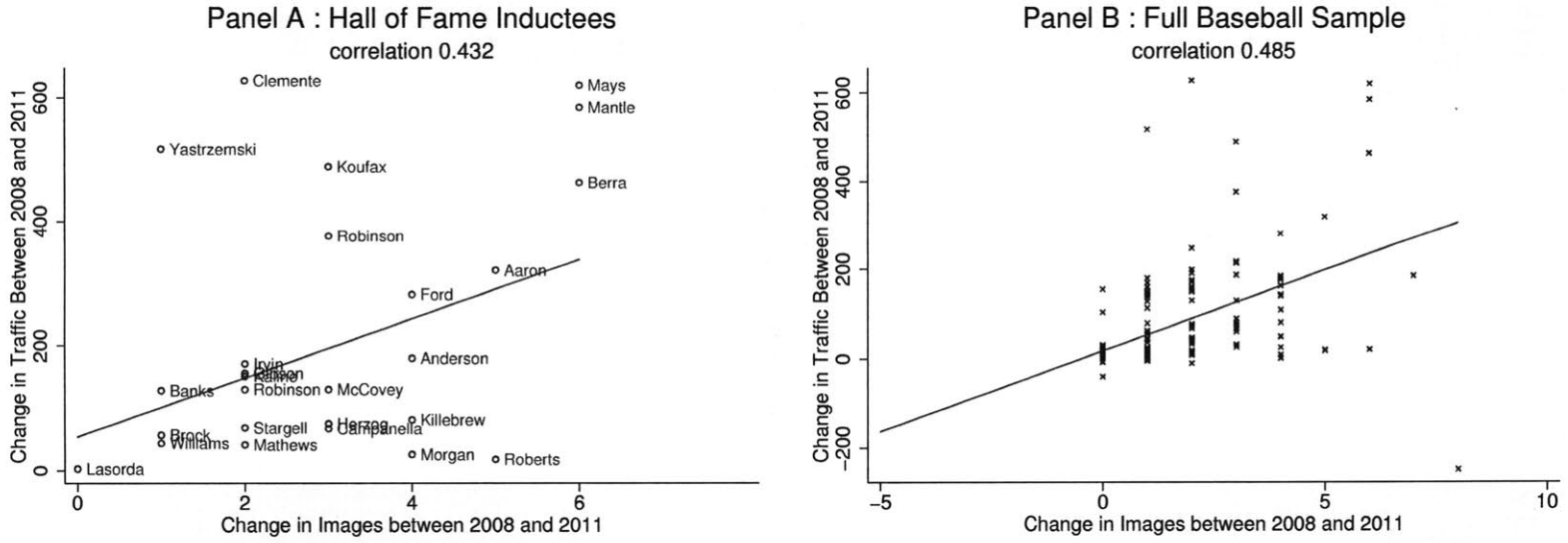
Note: This figure plots the mean change in the number of images by debut cohort after scaling these differences to adjust for the conservative classification scheme adopted. Specifically, the mean change in the number of images for any given cohort is scaled by dividing by the *average* value of ϕ where, ϕ is my measure of player misclassification representing the percent of player's career played before 1964. The line chart plot the value of $(1 - \phi)$ for each debut cohort.

Figure 4: Time Varying Estimates of the Impact of Copyright on Reuse of Images for Baseball and Basketball Players



Note: Estimates are derived from a regression with specification $IMAGES_{it} = \alpha + \beta \cdot OutOfCopyright_i \times YEAR_t + \gamma_i$ where $YEAR_t$ represents a full set of dummies for years 2006-2011, with 2008 as the omitted year and γ_i represents a complete set of player fixed effects. Robust standard errors clustered at player level are reported. Estimates are calculated separately for baseball and basketball players for this analysis. For this analyses data is obtained for number of images on Wikipedia page revisions for December 2006 to 2012.

Figure 5: Relation between Change in Traffic and Images for Players Inducted into the Hall of Fame



Note: For all baseball players in the "No Copyright", this chart plots the Change in Traffic (2008-2012) on the Y-axis vs. the Change in Images (2008-2012) on the X-axis. Panel A indicates last name of all Hall-of-Fame players included in the analysis, while Panel B includes all baseball players in the sample.

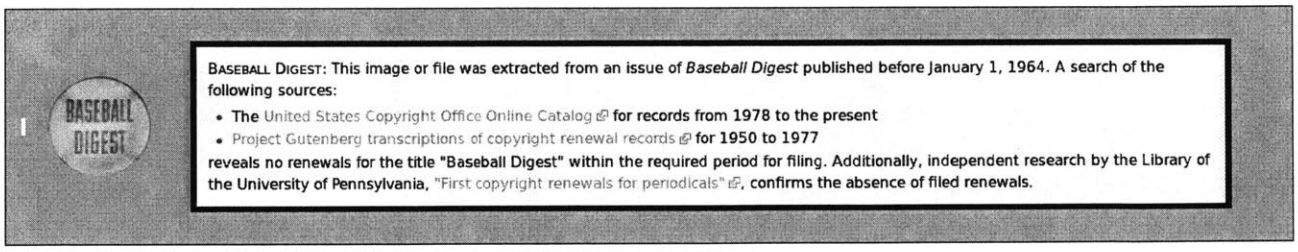
6 Appendices (For Online Publication)

6.1 Appendix A1 : The 1964 Copyright Exception

Legal opinion about copyright law concerning periodicals from University of Pennsylvania Libraries. See `http://onlinebooks.library.upenn.edu/cce/firstperiod.html` for more details.

> For works that received their copyright before 1978, a renewal had to be filed in the work's 28th year with the Library of Congress Copyright Office for its term of protection to be extended. The need for renewal was eliminated by the Copyright Renewal Act of 1992, but works that had already entered the public domain by non-renewal did not regain copyright protection. Therefore, works published before 1964 that were not renewed are in the public domain. With rare exception (such as very old works first published after 2002) no additional copyrights will expire (thus entering the public domain) until at least 2019 due to changes in the applicable laws.

The following screenshot is from a Wikipedia banner explaining the legal use of an image sourced from Baseball Digest.



6.2 Appendix A2 : Back-of-the-Envelope Welfare Calculation

This section outlines the methodology through which I arrive at my estimate of about $267,335 annually for a lower bound on the loss to social welfare from copyright being about $267,335 annually for Wikipedia.

In order to arrive at this estimate two pieces of data are needed: (a) the approximate value of a page-view to society and (b) estimated page-views lost due to copyright. For piece (a) I use `webindetail.com` which provides the estimated daily earnings of Wikipedia from potential advertising which would equal to $2.2 million dollars daily for about 400 million daily page-views.[14] This translates into a value to Wikipedia of about $0.0055 per page-view from advertising. For piece (b) results from this study suggest that for every missing image, a Wikipedia page receives about 33.1 fewer page-views per month (Table 6) and that pages have 0.598 fewer images on average (Table 5) due to copyright. Therefore, a page affected by copyright is expected to lose about $0.1088 per month. For the set of 319 pages affected by copyright in this study therefore, this translates into an annual loss of about $416 or a net present value of about $20,800.[15] Assuming that about 5% of all 4.1 million articles on Wikipedia are affected in a similar way, this translates into an annual loss of about $267,335 or a net present value of about $13.36 million. These estimates are economically significant for Wikipedia in the light of estimates of the economic value of Wikipedia itself which is pegged to be at about $43.5 million per year (Greenstein, 2013). Further, these estimates represent only a lower bound on lost surplus because advertising rates capture only the valuation advertisers place on reader eyeballs and do not calculate value to readers including value of derivative works of Wikipedia pages.

[14] While Wikipedia does not accept advertising, `webindetail.com` arrives at this estimate based on a comparables analysis based on other similar websites with a comparable user base.

[15] discounted at a rate of 2% per year over a perpetual life term

6.3 Appendix A3 : Robustness – Classification of "Out-of-Copyright" players

This section presents robustness checks that take seriously issues around the classification of players in the "Out-of-Copyright" group, but who played in both Out-of-Copyright and In-Copyright periods.

First, in order to understand the extent of misclassification, for each player, I calculate a measure of misclassification ϕ, i.e. the percent of a given player's career that has been played before 1964. By this measure a player who makes his debut in 1960, and retires in 1980 would've played 25% of his career in the out-of-copyright period. Having calculated ϕ for each player, I calculate the mean change in the number of images by debut cohort (as in Table 2), but scale these differences by dividing by the *average* value of ϕ for every debut cohort. Figure 3 plots calculated values of "Percent In-Copyright" $(1-\phi)$ as the black line, and the rescaled differences for each debut cohort in the bar chart. As expected, for early debut cohorts, the value of ϕ is close to one, but drops off fairly quickly close to 1964. Further, once these differences are accounted for, we see a more discontinuous difference between players who made their debut before and after 1964. This provides confidence that differences in levels of reuse are likely to be driven by a difference in copyright status before and after the year 1964, rather than other confounding factors.

6.4 Appendix B : Robustness Check using Log Models

Table B.1: Log Models : DD Regressions Estimating Impact of Digitization on Wikipedia

	(1) Ln(Images)	(2) Ln(Traf.)	(3) Ln(Text)
Baseball X Post	0.306***	-0.0547	0.0445
	(0.0275)	(0.0377)	(0.0276)
Post 2008	0.122***	0.702***	0.641***
	(0.0173)	(0.0274)	(0.0200)
Constant	0.103***	3.086***	10.16***
	(0.00688)	(0.00942)	(0.00690)
Player FE	Yes	Yes	Yes
adj. R^2	0.339	0.556	0.693
N	2048	2048	2048

Robust standard errors clustered at player level are reported

* p < 0.10, ** p < 0.05, *** p < 0.01

Note: This table provides a robustness check for the impact of digitization on Wikipedia using a log transformed dependent variable. The same specification and data as in Table 3 is used. Dependent variable is $Log(Y+1)$ where Y denotes either $Images_{itg}$, $Traffic_{itg}$ or $Text_{itg}$ depending on the model.

Table B.2: Log Models : DD Regressions Estimating Impact of Copyright on Wikipedia

	(1) Ln(Images)	(2) Ln(Traf)	(3) Ln(Text)
Out-of-Copy X Post	0.246***	0.153***	0.0169
	(0.0437)	(0.0575)	(0.0379)
Post 2008	0.335***	0.589***	0.679***
	(0.0252)	(0.0277)	(0.0254)
Constant	0.141***	3.330***	10.29***
	(0.0103)	(0.0129)	(0.00951)
Player FE	Yes	Yes	Yes
adj. R^2	0.473	0.555	0.717
N	1028	1028	1028

Robust standard errors clustered at player level are reported

* p < 0.10, ** p < 0.05, *** p < 0.01

Note: This table provides a robustness check for the impact of copyright on Wikipedia using a log transformed dependent variable. The same specification and data as in Table 4 is used. Dependent variable is $Log(Y+1)$ where Y denotes either $Images_{it}$, $Traffic_{it}$ or $Text_{it}$ depending on the model.

Table B.3: Log Models : DDD Regressions Comparing Impact of Copyright between Baseball and Basketball Wikipedia Pages

	(1) Ln(Images)	(2) Ln(Traffic)	(3) Ln(Text)
Post X Baseball X Out-of-Copy	0.218***	0.130	-0.0968
	(0.0582)	(0.0881)	(0.0605)
Out-of-Copy X Post	0.0287	0.0236	0.114**
	(0.0386)	(0.0667)	(0.0472)
Post X Baseball	0.218***	-0.108***	0.0610*
	(0.0322)	(0.0414)	(0.0341)
Post 2008	0.116***	0.697***	0.618***
	(0.0200)	(0.0309)	(0.0227)
Constant	0.103***	3.086***	10.16***
	(0.00675)	(0.00938)	(0.00689)
Player FE	Yes	Yes	Yes
adj. R^2	0.363	0.559	0.694
N	2048	2048	2048

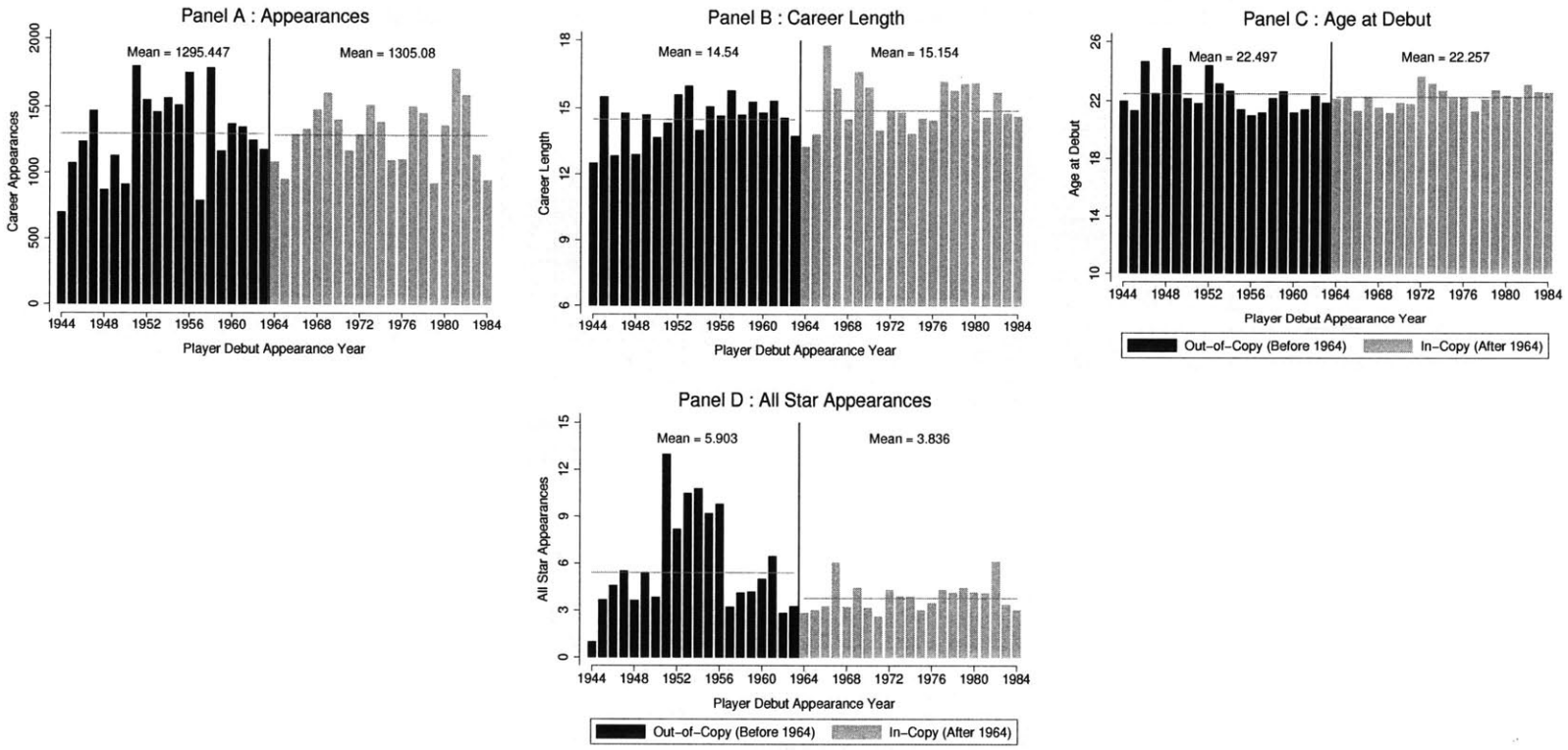
Robust standard errors clustered at player level are reported

* p < 0.10, ** p < 0.05, *** p < 0.01

Note: This table provides a robustness check for the impact of copyright on Wikipedia in a DDD framework using a log transformed dependent variable. The same specification and data as in Table 5 is used. Dependent variable is $Log(Y + 1)$ where Y denotes either $Images_{itg}$, $Traffic_{itg}$ or $Text_{itg}$ depending on the model.

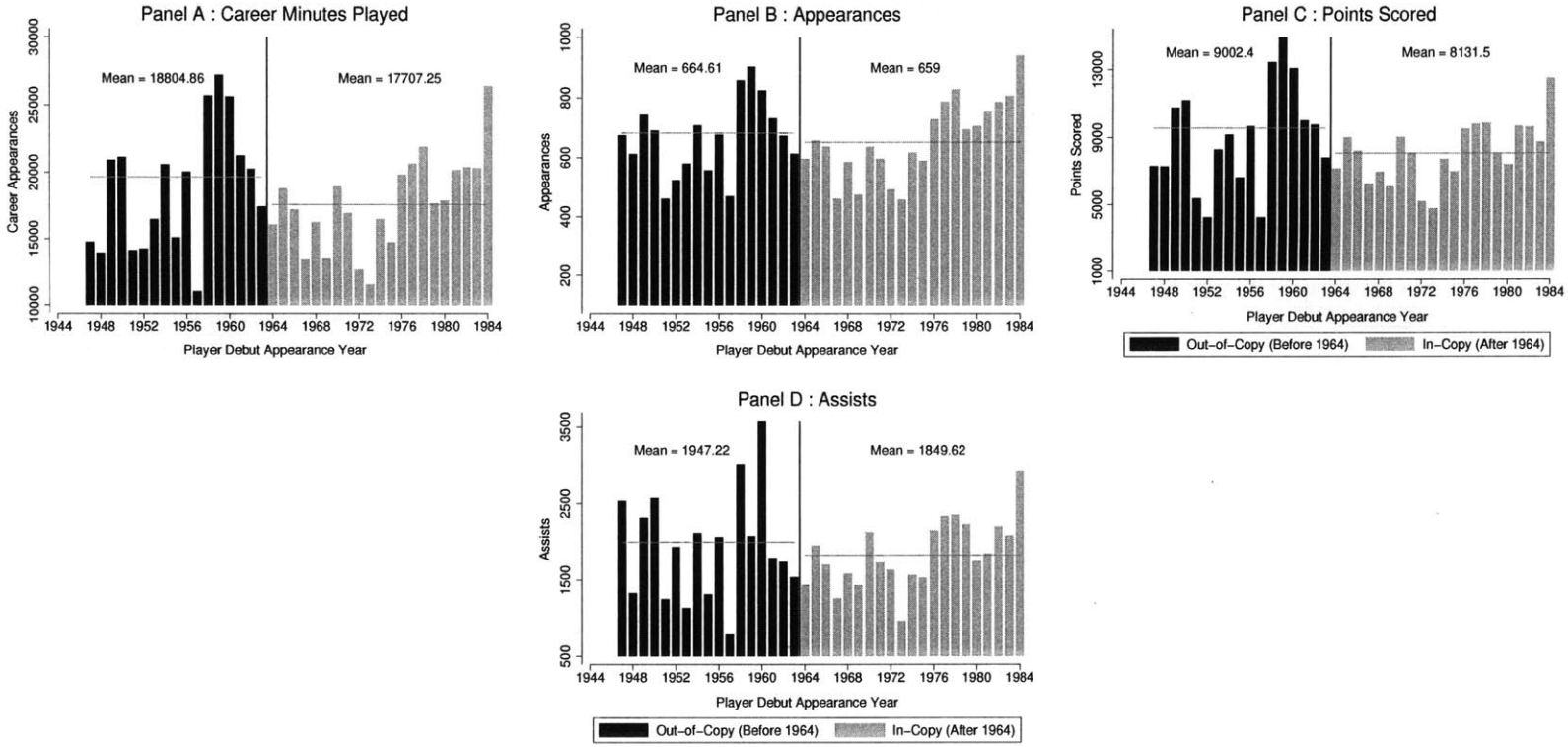
6.5 Appendix C : "Placebo" Tests for 1964 Cutoff using Player Characteristics

Figure C.1: Baseball Player Covariates by Debut Cohort



Note: These set of charts analyze the "placebo" test, where copyright cutoff of 1964 is assumed to be related to baseball player characteristics. Panel A plots the total number of appearances made by a baseball player over his career. Panel B plots the total number of years that a player was active. Panel C plots the age of a baseball player at debut and Panel D plots the number of All Star games played for a given player. A discrete jump at year 1964 would indicate that copyright seems to be related to these player level covariates.

Figure C.2: Basketball Player Covariates by Debut Cohort



Note: These set of charts analyze, for basketball players, the "placebo" test, where copyright cutoff of 1964 is assumed to be related to player characteristics. Panel A plots the total number of minutes spent of court for a basketball player over his career. Panel B plots the total number of games that a given baseball player was involved. Panel C plots the total points scored over a player's career and Panel D plots the number of assists for a given player. A discrete jump at year 1964 would indicate that copyright seems to be related to these player level covariates.

Bibliography

Abbott, L. (2011, September). Future baseball hall of fame players who did not appear in a world series.

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. 1âĂŞ31.

Aral, S., E. Brynjolfsson, and M. Van Alstyne (2007). *Information, technology and information worker productivity: Task level evidence*. National Bureau of Economic Research Cambridge, Mass., USA.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors*, pp. 609–626. Nber.

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.

Bharadwaj, A., O. A. El Sawy, P. A. Pavlou, and N. Venkatraman (2013). Digital business strategy: Toward a next generation of insights. *MIS Quarterly 37*(2), 471–482.

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., L. Hitt, and H. Kim (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

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*.

Chen, P.-Y., S. Dhanasobhon, and M. D. Smith (2008, May). All reviews are not created equal: The disaggregate impact of reviews and reviewers at Amazon.Com. SSRN Scholarly Paper ID 918083, Social Science Research Network, Rochester, NY.

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*.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science 49*(10), 1407–1424.

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

Dranove, D., C. Forman, A. Goldfarb, and S. Greenstein (2012). The trillion dollar conundrum: Complementarities and health information technology. Technical report, National Bureau of Economic Research.

Freedman, S., H. Lin, and J. Prince (2013). Are there heterogeneous effects of electronic medical record adoption on patient health outcomes? *NBER Summer Institute*.

Freeland, C. F. |. (2012, January). In big data, potential for big division. *The New York Times*.

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.

Garg, R., M. D. Smith, and R. Telang (2011). Measuring information diffusion in an online community. *Journal of Management Information Systems 28*(2), 11âĂŞ38.

Ghose, A. and P. G. Ipeirotis (2007). Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, pp. 303–310. ACM.

Goldfarb, A. and C. Tucker (2011). Privacy and innovation. Technical report, National Bureau of Economic Research.

Gorbatai, A. (2012, September). Social structure and mechanisms of collective production: Evidence from wikipedia.

Greenstein, S. M. (2013, March). Technology: Measuring consumer surplus online | the economist.

Greenstein, S. M., J. Lerner, and S. Stern (2010). The economics of digitization: An agenda for NSF. *NBER Working Papers*.

Greenstein, S. M. and F. Zhu (2012). Is wikipedia biased? *The American Economic Review 102*(3), 343âĂŞ348.

Gruber, J. (1994, June). The incidence of mandated maternity benefits. *The American Economic Review 84*(3), 622–641.

Hahl, O. (2012, November). Turning back the clock: Authenticity crises and retro fashion cycles. *MIT Sloan Working Paper*.

Hardin, G. (1968). The tragedy of the commons. *New York*.

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.

Jones, J. M. (2007, October). Less than half of americans are baseball fans. *Gallup*.

Khanna, D. (2012, November). RSC policy brief: Three myths about copyright law and where to start to fix it.

Kim, H. H. (2012). The effect of free access on the diffusion of scholarly ideas. In *ICIS*.

Kitch, E. W. (1977). Nature and function of the patent system, the. *JL & Econ. 20*, 265.

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

Lee, D. N. (2011). The digital scarlet letter: The effect of online criminal records on crime. Working Paper 1118, Department of Economics, University of Missouri.

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

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

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*.

Liebowitz, S. and S. Margolis (2004). Seventeen famous economists weigh in on copyright: The role of theory, empirics, and network effects. *bepress Legal Series*, 397.

Luca, M. (2013). Digitizing disclosure: The case of restaurant hygiene grades. *Harvard Business School Working Paper*.

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

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., 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)*.

Reimers, I. (2013). The effects of intellectual property on the market for existing creative works. *Working Paper*.

Scott, S. L. and H. R. Varian (2013). Bayesian variable selection for nowcasting economic time series. Technical report, National Bureau of Economic Research.

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

Shiller, B. R. (2013). First degree price discrimination using big data. Technical report.

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*.

Tambe, P., L. M. Hitt, and E. Brynjolfsson (2008). The extroverted firm: How external information practices affect productivity. In *ICIS*, pp. 12.

Varian, H. R. (2006, December). Copyright term extension and orphan works. *Industrial and Corporate Change 15* (6), 965–980.

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*.

Zhang, X. and F. Zhu (2010). Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, 07âĂŞ22.

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