Digitization, Prediction and Market Efficiency: Evidence from Book Publishing Deals[*]

Christian Peukert
University of Lausanne, HEC
christian.peukert@unil.ch

Imke Reimers
Northeastern University
i.reimers@northeastern.edu

Forthcoming in Management Science

Abstract

Digitization has given creators direct access to consumers as well as a plethora of new data for suppliers of new products to draw on. We study how this affects market efficiency in the context of book publishing. Using data on about 50,000 license deals over more than ten years, we identify the effects of digitization from quasi-experimental variation across book types. Consistent with digitization generating additional information for predicting product appeal, we show that the size of license payments more accurately reflects a product's ex-post success, and more so for publishers that invest more in data analytics. These effects cannot be fully explained by changes in bargaining power or in demand. We estimate that efficiency gains are worth between 10% and 18% of publishers' total investments in book deals. Thus, digitization can have large impacts on the allocation of resources across products of varying qualities in markets in which product appeal has traditionally been difficult to predict ex-ante.

JEL: D22, D83, L82

[*] We thank James Dana and Joel Waldfogel for valuable comments. We are grateful to Ilia Azizi for excellent research assistance. The paper has benefited from the feedback of participants at Florence Media Workshop, IIOC, Southern Economic Association Annual Meeting, ICT Conference Paris, SEARLE Internet Conference, ZEW ICT Conference, CESifo Economics of Digitization Conference, Toulouse Digital Economics Conference, NBER Digitization Winter Meeting, and seminars at Northeastern, the Dartmouth Winter IO Workshop, LMU Munich, NBER (Productivity Lunch), University of Melbourne, Digital Economy Workshop Lisbon, University of Zurich, University of Glasgow, University of Minnesota, Wake Forest University and Harvard Business School. Peukert acknowledges support from FCT-Portuguese Foundation of Science and Technology (FCT-PTDC/EGE-OGE/27968/2017) and from the Swiss National Science Foundation (100013_197807).

1 Introduction

Digitization has substantially changed markets for creative content. First, it has decreased production and distribution costs, which has led to an emergence of new products and an increase in the variety available to consumers (Waldfogel, 2017), with substantial welfare-enhancing effects (Aguiar and Waldfogel, 2018; Brynjolfsson et al., 2003). Second, digitization has provided additional avenues for content creators to reach consumers, allowing them to circumvent traditional intermediaries and directly market their products to potential customers (Gu and Zhu, 2018). Third, it has significantly improved the information environment in which creators and distributors of content as well as consumers make decisions: In addition to the information made available through recommendation algorithms and ratings systems (Claussen et al., 2019; Reimers and Waldfogel, 2021), the digitization-induced influx of content can further inform creators and distributors about the appeal of new products before they reach a mass market. For example, musicians such as Justin Bieber or Ed Sheeran achieved YouTube fame before securing major record deals, and books such as Andy Weir's *The Martian* and E.L. James' *Fifty Shades of Grey* became poster-children of self-publishing platforms before being picked up by major publishing houses.

While the first-order effects of digitization on consumers have been studied at length, the impacts of digitization on licensing markets and their efficiency are not nearly as well-understood. We examine and quantify the effects of digitization in the context of the book publishing industry, looking at deals between authors and traditional publishers. We ask whether digitization has allowed publishers to identify the most promising books as reflected in a stronger correlation between the size of a deal and its ex-post success. We then examine the role of improvements in the information environment as opposed to changes in bargaining positions and demand effects.

Book publishing is a particularly salient industry for studying the effects of digitization on the efficiency of licensing markets. Because the production and distribution of physical books traditionally required large investments, books were only able to reach consumers via (large) publishing companies. With imperfect prediction, however, some high-quality ideas were likely falsely rejected in this market and never reached consumers despite considerable ex-post appeal. Likewise, some books would not live up to their expectations, implying that the publishers could not recoup their initial investments. Digitization has changed the relationship between authors and traditional publishers. First, it has allowed authors to reach consumers via self-publishing platforms, so they no longer rely on traditional publishing houses.[1] Second, digitization has facilitated access to large-scale crowd-sourced reviews, with large effects on sales (Chevalier and Mayzlin, 2006) and consumer surplus (Reimers and Waldfogel, 2021); and publishers may have benefited as well. For example, a book's (digitized) text can be analyzed and its prospects can be evaluated based on the success of similar texts, as has been highlighted in academic and popular presses (Archer and Jockers, 2016; Phillips, 2016). These developments in the book industry may have significantly changed the nature of publishing deals.

Large-scale empirical evidence on the effect of digitization on the relationship between creators and distributors in any intermediary market is scarce for two reasons. First, data on contracts for inventions and creations are hard to come by; and second, causal inference demands exogenous variation in the prevalence of digitization. Our setting allows us to deal with both issues. We examine a unique dataset covering contracts of over 50,000 book deals from 2002 to 2015, including the advance that the publisher paid to the author. We utilize variation in book genres' attractiveness of the ebook format and self-publishing to estimate the effects of digitization on traditional book deals in a difference-in-differences framework. Ebooks and self-publishing arrived fairly suddenly, driven by the large-scale diffusion of Amazon's Kindle e-reader in 2008. While books of most genres were still predominantly published by and bought through traditional publishers by 2016, authors and readers of romance novels have largely embraced self-publishing platforms. We thus argue (and provide qualitative and quantitative evidence) that romance novels are a good treatment group for our analyses.

Our empirical results are consistent with a conceptual framework that is informed by the prior

[1] The most popular self-publishing platforms include Lulu, Smashwords, and Amazon's Kindle Direct Publishing, and all offer similar deals for authors. Amazon, for example, offers authors a platform to publish their work with a royalty rate of up to 70% of revenue (depending on the chosen price), but no upfront license payments. See `https://kdp.amazon.com/help?topicId=A30F3VI2TH1FR8`.

literature.

To study the relationship between a book's advance payment and its ex-post success, we estimate each title's ex-post success as a function of the deal size, and we compare the sizes of the residuals across genres, before and after the onset of digitization. Measuring ex-post success as a book entering the weekly Top 150 bestseller lists, we find that advances become more accurate predictors of such realized success for authors of romance books compared to other genres. Further, advances and ex-post success become more highly correlated: A smaller share of deals with advances in our lowest category (below $50,000) become bestsellers ('false negatives'), and a smaller proportion of major deals (above $500,000) does not reach best-selling status ('false positives'). Our estimates imply that 10% of publishers' investments in small deals are no longer misallocated and that savings from the reduction in false positives amount to more than 18% of publishers' investments in major deals.

We explore several mechanisms that could explain such increases in the correlation between advances and ex-post success, and we conclude that improvements in the information environment are a driving force of these changes. Specifically, alternative mechanisms such as digitization-induced changes in demand or bargaining power are – by themselves – inconsistent with the patterns we observe. First, while we find that digitization led to a substantial increase in the advances paid to authors, these increases cannot explain the decrease in false positives. Second, increases in demand cannot explain the decrease in false negatives.

Our analyses on the predictability of success suggest that traditional and self-publishing can have complementary roles. Self-publishing can function as a talent pool and source of information for traditional publishers, who can then ensure that those creators and ideas with the largest appeal reach the largest audience (Ordanini, 2006; Benner and Waldfogel, 2016). Additional analyses across authors and publishers show that reaping these benefits requires some effort by the publishers though. We find that the improvements in predictability are strongest among major publishers – those who dedicate the most resources into work related to data analytics (as measured by job postings). These changes in the market for book ideas can have further implications on the downstream product market. If larger publishers are better able to learn from the increasingly available market information, then digitization could lead to even more market concentration among publishers. Thus, digitization could amplify any anti-competitive effects of the recent high-profile mergers in the publishing industry.

Our results relate to the emerging literature on the economics of data and information, by providing empirical evidence on the theoretical notion that publicly available information can have broad effects across firms (Jones and Tonetti, 2020). Specifically, we relate to evidence showing that more information can increase efficiency in health care (Ribers and Ullrich, 2019), and that forcing firms to share data likely improves welfare by increasing competition (Reimers and Shiller, 2019).

As such, the findings in this paper reach beyond creative industries to any market in which a product's quality is not known to all market participants, especially those where intermediaries play an important role (Spulber, 1996). For example, traditionally, used car dealers served as curators who vouched for a car's quality, and real estate brokers both helped buyers and sellers find a good match and advised buyers on the quality of a house or apartment. Digitization has democratized these processes by allowing consumers to find good matches more easily (Reimers and Waldfogel, 2021) *and* by removing some of the uncertainty inherent in these markets.

Showing that digitization has increased efficiency in the licensing market adds to the literature on the broader welfare effects of digitization. So far, surprisingly few papers focus on impacts on the supply-side. Most closely related to our paper, Aguiar and Waldfogel (2018) argue that with lower fixed costs, firms introduce more products, some of which turn out to be successful. Going beyond this mechanism, we explicitly address the information channel: Firms may make better predictions due to the availability of additional data, allowing them to invest more in the best ideas. Such improved ex-ante selection of high-quality ideas would increase consumer surplus even without cost decreases; and in addition to improving the market's static efficiency, our results also suggest that digitization affects the incentives to create in the first place.[2]

[2] Similarly, Yang (2018) estimates the extent to which learning from competitors can attenuate the negative impacts of competition on profits in retail markets. This article thus relates to previous literature on information externalities (Caplin and Leahy, 1998; Toivanen and Waterson, 2005; Shen and Xiao, 2014), although these papers mostly examine impacts on the extensive entry margin, rather than intensive investment decisions.

2 Industry Background and Research Framework

Publishers have traditionally assumed the role of an important intermediary between authors and consumers in the book industry. Authors would try to license the rights to their books to publishers, who then market the final product to consumers. For authors who are lucky enough to find a match, the contract typically includes a lump-sum (advance) payment to the author – to be paid out before any copies are sold – as well as a royalty for each copy that is sold beyond the advance payment. While royalty rates have mostly remained constant across books and over time, the lump-sum payments vary significantly across books, authors, and publishers, from a few thousand dollars to over a million, depending on the book's predicted success in the product market (see Greco, 2013; Levine, 2016).

2.1 Digitization in the Book Publishing Industry

Digitization has affected the book publishing industry and its structure in a variety of ways. Notably, output and prices have changed substantially since the late 2000's. The number of new book editions (ISBNs) issued per year rose from 561k in 2008 to 2.35 million in 2012 – an increase of 319%.[3] Over the same period, US total sales (physical and electronic books) increased by 27%, from 2.16 billion units in 2008 to 2.70 billion units in 2012, and average prices (adjusted for inflation) decreased by 10%, from $14.13 to $12.68 (Waldfogel and Reimers, 2015). Here we discuss the main drivers of these changes: substantial reductions in cost, a changing role of intermediaries in the licensing market, and changes in discovery and consumption patterns in the final product market.

Lower Costs: Traditionally, marginal costs in book publishing included creating the copies as well as shipping the printed books to stores and returning unsold inventory. Consolidation of physical book stores and centralized distribution of large online retailers lowered the costs of distributing physical books significantly, and the arrival of the electronic book format has all but removed these

[3] Data from Bowker. See `https://web.archive.org/web/20150406062314/http://www.bowker.com/assets/downloads/products/isbn_output_2002_2013.pdf`.

costs altogether. Waldfogel and Reimers (2015) estimate that average costs fell from about $7 for a physical book to about $2 for ebooks, where the remaining costs mainly comprise the royalties paid to the author. Provided that competition across publishers did not decrease significantly, these cost decreases are consistent with a shift of the aggregate supply curve to the right.

New Avenues of Consumption: Digitization in book publishing was largely spurred by the introduction of high-quality e-reading devices. Most notably, the Amazon Kindle, which was launched in November 2007 and used the e-ink technology, improved the reading experience enough to trigger a large shift in reading behavior. According to a survey by the *Pew Research Center*, ownership of designated e-readers increased steeply and steadily from essentially zero percent in 2008 to over 30% in 2014. In addition, almost 40% of US adults owned a tablet by 2014 – up from zero in 2009, the year before Apple introduced the iPad.[4] The ebook format also likely attracted new readers. By 2010, ebook sales had reached 69 million units, and by 2014 that number had risen to 234 million.[5]

Disintermediation: Digital printing technologies lowered the average costs of physical books, which gave rise to print-on-demand services that allowed any author to publish their books.[6] More recently, digital self-publishing platforms emerged as a new channel for authors to publish electronic books for a small fee, with no screening process and no major advertising efforts by the platform.[7] While such platforms are akin to Vanity Presses and Print-on-Demand services, the full automation and digital distribution offered by self-publishing platforms drove costs down far below what traditional publishers and other services could offer, making digital self-publishing a popular alternative for authors.

[4] Evidence shows that attributes such as ease of use and portability are positively related to consumers' stated intent to use e-reading devices (Jung et al., 2012; Torres et al., 2014), and that consumers' value from e-reading devices depends on the availability of ebooks (Torres et al., 2014; Huang et al., 2017).

[5] See `https://www.statista.com/statistics/426799/e-book-unit-sales-usa/`, which lists NPD Bookscan and NPD PubTrack Digital as sources.

[6] See Laquintano (2013) for a detailed discussion of the differences between traditional vanity presses and print-on-demand services that entered the market using digital technology.

[7] Authors can promote their works on self-publishing platforms as well. See Zegners (2017) for an example.

Anecdotal evidence suggests that authors and traditional publishers have taken advantage of these platforms. A "poster-child" of self-publishing, E.L. James' *Fifty Shades of Grey* was originally released as an ebook and a print-on-demand paperback through the Australian independent virtual publisher The Writer's Coffee Shop in May 2011. It was then picked up by Vintage Books, an imprint of Random House (the largest publisher in the United States), in March 2012.[8] Similarly, Andy Weir's *The Martian*, which was originally self-published in 2011, attracted enough demand to be published by Crown Publishing (a subsidiary of Random House) in February 2014. Weir sold the rights to his next book in September 2014 to Crown Publishing.

Information and Discovery: Entertainment industries have widely been described as markets where "nobody knows anything" in the sense that product success is notoriously difficult to predict (Caves, 2000). Book publishing is no exception. In the licensing market, publishers had limited opportunities to test a books's appeal before its launch. In the final product market, consumers have traditionally learned about new books through word-of-mouth, bestseller lists and reviews by professional critics, but many books remained "hidden." Digitization has significantly improved this information environment. First, sales data have become readily available through subscription-based point-of-sale data aggregators such as *Bookscan* and *Bookstat.* Second, sales are increasingly made in digital channels that – through 1:1 communication – allow distribution platforms and publishers to identify purchases by individual consumers. Third, online retailers such as Amazon publish sales rankings, prices, and consumer reviews in a more transparent way than was traditionally available. Fourth and perhaps most importantly, the massive increase in the number of available books has allowed such data to become available for a much larger set of products.

2.2 The Special Case of Romance Books

The effects of digitization and self-publishing are not equally distributed across genres. On the supply side, the left-hand panel of Figure 1 shows that the number of books published on the

[8] See `https://www.nytimes.com/2014/02/27/business/media/for-fifty-shades-of-grey-more-than-100-million-sold.html`.

popular self-publishing platform *Smashwords* has grown significantly since 2008, mostly driven by the romance genre. Before 2010, the number of romance books on the platform was similar to the number of books in the other top genres (fantasy, children, religion, mystery, self-improvement, biography). In 2011, there were roughly twice as many romance books, and by 2011 the supply of romance novels became roughly five times as large as that of the second-most represented genre. In sum, romance/erotica novels represent 28% of the 431,307 books published on the site between 2008 and 2016.[9]

We see similar patterns on the demand side. The right-hand panel of Figure 1 displays – separately for romance and for non-romance books – the share of all titles in the *USA Today* weekly bestseller lists within the genre that were originally self-published, for each week from 2010 to 2014. Beginning in 2011, between 20% and 50% of all bestsellers in the romance category had a self-publishing background. In contrast, during the entire observed period, the share of originally self-published bestsellers in other categories never exceeded 5%.

To understand the sources of these differences, we conducted several interviews in the field and asked industry insiders about the typical characteristics of digitally self-published books. The experts argued that romance and erotica novels – ranging from Nicholas Sparks' *The Notebook* to E. L. James' *Fifty Shades of Grey* – are especially well-suited for self-publishing for several reasons. First, romance books are relatively easy to write because they do not tend to be research-intensive. Second, the nature of many romance novels might make readers reluctant to read them in public. As the experts argued, e-reading devices, which prevent others from seeing the book's cover, could remove social frictions (Goldfarb et al., 2015) and make self-publishing particularly popular among romance books. Finally, romance readers form a close-knit group that communicates extensively in online communities, making word-of-mouth particularly effective and traditional advertising campaigns less crucial. This is consistent with data from book reviews. For example, the most used tag on the review platform `goodreads.com` is "romance" (4,553 times) and the second most used tag is "fiction" (3,984 times).[10] In addition, data from Reimers and Waldfogel (2021) indicate

[9] We use romance as shorthand for romance/erotica in the remainder of the paper.

[10] Numbers as of October 2, 2017.

that, even when conditioning on a book's age and ranking history, romance books receive more than 50% more consumer ratings on Amazon than books of other genres (t-value=5.8).

The rise in the popularity of self-published romance books – relative to traditionally published books – could coincide with romance books becoming more popular per se. We examine this possibility quantitatively. First, the left-hand panel of Figure 2 shows the total number of print books sold by genre, based on data by Nielsen.[11] It shows that print book sales of all genres decreased after 2008, although romance saw a renaissance in 2012 – the year in which *Fifty Shades of Grey* was licensed to the traditional publishing house Vintage Books.[12] Second, additional information by the Romance Writers of America suggests that the share of romance book sales in the US remained constant, around 13% over the observed time period.[13] Third, the right-hand panel of Figure 2 exhibits the same demand patterns in the USA Today Top 150 bestseller lists. It plots the share of pseudo-sales in the lists – calculated as $\frac{1}{rank}$ – that is attributed to romance books. This share remains constant throughout the digitization period, with the exception of mid 2012 to mid 2013 – again largely driven by the success of the *Fifty Shades* trilogy.

2.3 The Effect of Digitization on Licensing Efficiency

The mechanisms described above guide our research framework on how digitization has affected the relationships between authors and traditional publishers.

2.3.1 The Information Enviroment and Ex-ante Prediction

Digitization has both facilitated access to information about *existing* products and spurred growth in *new* products from which firms can learn (see Bergemann and Välimäki, 2000 and Amador and

[11] See their presentation at the 2014 Digital Book World, `https://tinyurl.com/y8yxu7st`.

[12] While interviews with industry insiders suggest that *Fifty Shades of Grey* did not change the publishers' expectations regarding the profitability of romance novels, we examine the role of these immensely popular books in Section 4.4.2.

[13] Information obtained through the wayback machine, for `http://www.rwanational.org/cs/the_romance_genre/romance_literature_statistics`, `http://www.rwa.org/p/cm/ld/fid=580`, and `https://www.rwa.org/page/romance-industry-statistics`. According to these sources, the revenue share (print) of romance books over all fiction books was 13.5% in 2008, 13.2% in 2009, 13.4% in 2010, 14.3% in 2011, 16.7% in 2012, 13.0% in 2013 and 9.8% in 2014. The average from 2008 to 2014 is 13.4%, and with the exception of 2012, the share remained quite constant.

Weill, 2012 for detailed mechanisms); and previous papers point to organizations and individuals learning both from their own and others' successes and failures (Srinivasan et al., 2007; Madsen and Desai, 2010; Benner and Tripsas, 2012; KC et al., 2013). Thompson (2013) confirms that similar patterns likely hold in book publishing, highlighting that "considerable time and energy is invested [..] in trying to come up with other books to which the new book they want to sell as an agent or buy as an editor can be compared" (p. 202).

The large influx of new books, along with more readily available data about success and consumer preferences, should therefore generate more information from which publishers can draw when making decisions about book advances. Specifically, the new sources of information made available by digitization can help publishers avoid "false negatives" (investing too little in high-appeal books) as well as "false positives" (investing too much in low-appeal books). This leads us to conjecture that digitization could affect advance payments such that they better reflect a book's ex-post success.

Further, the improvements in prediction accuracy depend on how authors and publishers use the information. We study two mechanisms. First, if publishers mostly learn from a specific book's or author's past success, the improvements are likely driven by deals with established authors. However, if the improvements in information are driven by the successes and failures of *others*, they may be particularly large among new authors. Whether digitization leads to larger improvements in prediction accuracy among new or established authors is therefore an open question.

Second, major publishers are likely better able to take advantage of the new information than smaller, independent publishers. Interpreting the inflow of new information may require substantial investments in technology and human capital, which are easier to take on for less cash-constrained firms. For example, large manufacturing firms have been quick to adopt data-driven decision making in the mid- to late 2000's (Brynjolfsson and McElheran, 2016). Evidence further shows that data assets and employees with data analytics skills are complements (Tambe, 2014), and firms with experience in process innovation receive the most benefits from data analytics (Wu et al., 2020).

2.3.2 Changes in the Licensing and Final Product Markets

The changes brought about by digitization can affect publishers' relationships both with authors on the licensing market and with consumers on the final product market; and both can strengthen the correlation between license deals and ex post success beyond the information mechanism we describe above.

First, a growing literature suggests that digital platforms have changed the outside options available to creators and complementors. For example, gig economy platforms such as Uber have introduced new opportunities for self-employment, which crowded out low-quality entrepreneurial activity (Burtch et al., 2018). Other papers have highlighted the role of competition in platform markets. Platform owners can glean information from transactions on the platform to enter more successful product spaces (Zhu and Liu, 2018), and such entry has been shown to increase the complementors' incentives to innovate in some cases (Foerderer et al., 2018), and to circumvent the platform altogether in others (He et al., 2020).

Anecdotal reports show that these patterns also apply to book publishing. Romance writer Jamie McGuire, for example, signed a major deal with Atria Publishing (an imprint of Simon and Schuster) for her previously self-published *Beautiful Disaster* and a sequel in July 2012, but returned to self-publishing for another (successful) sequel in 2015. She "still plan[s] to traditionally publish, but with books that [she] feel[s] are best suited for that route" (McCartney, 2016). These patterns suggest that digitization can lead to higher advance payments for authors.

Second, digitization and the option to read books in electronic formats may have led to increased demand for certain types of books. Consequently, publishers may pick and choose certain types of romance books that they know to be successful, for example those that are akin to Stephenie Meyer's *Twilight* or E.L. James' *Fifty Shades* series.

3 Data

To test the ideas derived in our research framework, we use a novel dataset collected from a variety of sources. We describe the two main sources in detail here and introduce additional data when used.

3.1 License Deals

Data on license deals come from *Publishers Marketplace*, a professional online community for the book industry, from which we observe information about book-related deals and the involved entities. We observe all posted deals between January 1$^{\text{st}}$, 2002, and December 31$^{\text{st}}$, 2015 – a total of 100,772 book and rights deals. We focus on the 58,782 book deals in the dataset, and we extract names of authors and editors along with information on genres and types of deals. Importantly, the database allows us to quantify the size of the advances for a subset of about 30% of these deals (17,136 deals), based on the reporting entity's decision. Each deal can be reported by any (verifiable) participant, including the author, the agent, and the publishing editor.[14] The advance is reported in categories pre-specified by the platform: (1) less than $50k (*"nice"*, 58% of all book deals), (2) $50k to $99k (*"very nice"*, 11%), (3) $100k to $249k (*"good"*, 16%), (4) $250 to $499k (*"significant"*, 6%), and (5) more than $500k (*"major"*, 10%).

From the posted information, we define additional control variables that describe deal-specific characteristics such as whether it included multiple books, and if it was subject to competition among publishers; author-specific characteristics such as whether the author has won awards (is acclaimed), has previously written a bestseller, has self-publishing experience, or is a debut author; and book-specific characteristics such as whether it is part of a series.[15] In addition, we use

[14] For our analysis, it is important that the share of unreported deals does not vary systematically over time and across genres. We test for this by regressing an indicator that is 1 if the deal size is reported on year dummies as well as their interactions with a romance dummy. The interactions are statistically insignificant at the 5% level for all years in our data. If we aggregate the years to before-2008 and after-2008 periods, the coefficient on the interaction of "after" and "romance" is 0.0099 (standard error = 0.011).

[15] To create these variables, we search for keywords in the text. They are: *at auction, preempt; award, edgar, nominee, winner, finalist, pulitzer, NYT notable, acclaimed, syndicated, star; bestselling, bestseller; self-published; debut, first-time, first novel; sequel, prequel, next book, follow-up.*

the publisher information provided in the deals database to determine whether the deal involves (imprints of) one of the five biggest publishers (major publishers), and we collect information on each author's previously published titles from the *Bowker Books-in-Print* directory.[16]

Descriptive statistics of deal sizes and characteristics for book deals for which the deal size is posted – separately for romance novels and other genres and before and after 2008 – can be found in Table 1. This dataset contains 2,987 romance deals, including 1,191 (170 per year) between 2002 and 2008 and 1,796 (257 per year) between 2009 and 2015, and 14,149 deals in non-romance categories (1,126 per year before 2008, and 895 per year after 2008).[17] The romance category exhibits a rise in the share of sequels, as well as in deals with previously bestselling authors and authors with self-publishing experience, although both remain far below 5%. The share of deals with debut authors decreased after 2008 and the average number of previous books by the author increased. Non-romance deals display similar trends in these variables, although the changes are smaller in magnitude. Moreover, we consider the midpoints of each deal size category and choose a deal size of $750k for deals above $500k to calculate average advances. While the average deal size remains relatively stable at around $144,000 to $152,000 for non-romance deals, the average deal size increases from about $106,000 to $134,000 for romance deals. Specifically, the share of "good" deals in non-romance genres increases the most at the expense of "nice" deals. For romance deals, the largest relative increase occurred in "major" deals, mostly at the expense of "good" deals.

3.2 Ex-post Success

We link the licensing information with ex-post success information from the USA Today weekly bestseller lists for 2002 to 2016. We use these bestseller lists because they are – to our knowledge – the most reliable source on success at the title level, and the only one combining sales of both physical and electronic formats.[18] The information at the title level is crucial for our analyses,

[16] The major publishers are: Hachette, HarperCollins, Macmillan, Penguin/Random House, and Simon and Schuster.

[17] Including deals without information on the advance fee, the number of romance deals rises from 353 per year before 2008 to 863 after 2008, and the number of non-romance deals increases from 2,813 to 4,345 per year over the same time periods.

[18] The NPD (formerly Nielsen) Bookscan database, for example, only reports sales of physical editions.

because sales of ebooks are a direct consequence of digitization and are therefore likely to vary significantly across genres.

While the USA Today weekly bestseller lists provide a large upside by incorporating sales from all editions, their use creates two challenges. First, matching between deals and bestsellers is complicated by the fact that the deal data from Publishers Marketplace include working titles that may be different from the final book titles. We treat a bestselling book as a match for a deal if the author matches and the book entered the bestseller lists after the deal. Second, the USA Today data only allow us to determine whether a book became a bestseller at all – a measure on the extensive rather than the intensive margin. We test whether our coarse success measure creates issues of selection toward the most successful works in a separate analysis of the weekly top 100 bestsellers from the NPD (formerly Nielsen) Bookscan database. Here, we can observe (physical) sales numbers in addition to ranks, and we find that 95% of these bestsellers warrant at least a "very nice" deal.[19] Therefore, we pay particular attention to ex-post success among the two groups that suggest the strongest priors: Which "nice" deals become bestsellers (implying a 'false negative'), and which "major" deals do not make the list (implying a 'false positive')?

The bottom of Table 1 provides summary statistics of the deals' ex-post appeal. Overall, about 12% of all book deals with size information become bestsellers after the deal, so that our measure provides ample variation despite its coarse nature. About 30% of the *romance* deals in our data appear in the USA Today bestseller lists until 2008, and about 26% after. The share of *non-romance* deals entering the top 150 list is about 8% before and after 2008. In addition, the advances for romance deals seem to become more accurate over time. The share of small deals that become bestsellers decreases from 20% to 10%, and the share of major deals that "flop" decreases from 30% to 9%. These shares also decrease among non-romance deals, albeit to a much smaller extent. The patterns laid out in Table 1 are suggestive of large effects of digitization on the ability of advances to predict ex-post success. In what follows, we examine these relationships more formally.

[19] Their cumulative sales numbers while in the Top 100 list most often exceed $100,000, so that they at least pay back a 'very nice' advance.

4 Estimation and Results

4.1 Identification Strategy

Based on the evidence laid out in Section 2.2, we use romance novels as a treatment group and the years following 2008 as the treatment period in a number of difference-in-differences analyses to examine how digitization has affected book contracts. Specifically, we estimate variants of the following equation:

$$Y_{i,j,k,t} = \alpha + \beta R_j + \delta(After_t \times R_j) + \kappa C_{jt} + \mu_t + \nu_k + \varepsilon_{i,j,k,t} \tag{1}$$

The unit of observation is a license deal i between author j and editor k (at publisher p) in month t. The dependent variable, $Y_{i,j,k,t}$, describes a function of how well the advance explains the book's realized ex-post success. We explain these functions in more detail below. $After_t$ indicates whether the deal was made after the year 2008, and R_j indicates whether the author ever published a book in the romance category ("romance author").[20] To account for time-varying heterogeneity across authors, we include our control variables $C_{j,t}$.[21] We also account for time-specific variation by including month-year fixed effects μ_t. In addition, we absorb any unobserved heterogeneity across editors (and hence publishers) by including editor fixed effects ν_k. Finally, we cluster standard errors at the genre level to avoid incorrect inference in the difference-in-differences model (Bertrand et al., 2004; Abadie et al., 2017). We first use this setup to measure how closely advances and sales are aligned, and we later adopt variants of it to explore underlying mechanisms.

4.2 Predicting Ex-Post Appeal

We first ask whether digitization facilitates prediction of ex-post success in general. We then specifically examine whether publishers use the improved information environment to make better

[20] The main results remain almost identical when we instead categorize the treatment group at the deal level.

[21] Some unobserved heterogeneity across authors may remain. However, we cannot include author fixed effects because only a small number of authors appear in our data more than once.

decisions.

4.2.1 General Improvements in the Information Environment

To ask whether the information environment becomes better in general, we create a dependent variable in two steps. We first estimate the author's ex-post success – an indicator that equals one if the author had a new book enter the USA Today Top 150 weekly bestseller list after the deal was made – as a function of market-level information available to everyone. In this step, the explanatory variables include indicators for the number of the author's previously published titles, genre indicators, our set of deal-specific control variables, month-year indicators, and interactions of these terms. We record the residuals from this regression. In the second step, we then estimate equation (1), where the dependent variable is the absolute value of these residuals.

The results reported in column (1) of Table 2 suggest that the residuals decrease economically and statistically significantly, by 0.05 units, or 17% at the pre-digitization mean for romance deals.[22]

4.2.2 Do Advances Better Explain Success?

We next examine the possibility that publishers' ex-ante predictions (as measured by the advance payment they offer) become more precise. Our main strategy here is similar to our strategy above. We first estimate the author's ex-post success as a function of the deal size indicators, interacted with genre indicators and year dummies:

$$success_{i,j,k,t} = \alpha + \sum_{\tau \in T} \sum_{g \in G} \sum_{s \in S} \delta_{\tau,g,s} (\gamma_\tau \times \zeta_g \times \phi_s) + \nu_k + \epsilon_{i,j,k,t}, \tag{2}$$

where γ_τ, ζ_g, and ϕ_s denote dummy variables for the year, the deal's genre, and the advance size category, respectively. As above, we are interested in the residuals from this regression. We then run our baseline difference-in-differences regression from Equation (1) with the absolute values of these residuals as the dependent variable. The coefficients of interest are shown in column (2) of

[22] With the exception of the "previous bestseller" indicator, which is correlated with larger residuals, the coefficients on the control variables are statistically insignificant in most specifications. We include but do not report the coefficients on the control variables in all following regressions.

Table 2. The coefficient on After 2008 × Romance is 60% larger in magnitude than that from the "public information" regression, at -0.08. Given the mean of the absolute residuals before 2008 (0.290), this coefficient corresponds to an improvement of almost 28%.

We go on to study the nature of the reduction in prediction errors more closely. We treat "nice" deals – those with advances under $50,000 – as those for which publishers do not expect a Top 150 bestseller; and we treat "major" deals – with advances over $500,000 – as those that are expected to succeed. Consequently, we describe a "nice" deal that enters the Top 150 as a "false negative," and we treat a major deal that does not enter the Top 150 as a "false positive." To determine the impact of digitization on false negatives, we estimate a linear probability model on the set of all "nice" deals, where the dependent variable equals 1 if the book becomes a bestseller; and to determine the impact on false positives, we estimate the likelihood of *not* becoming a bestseller on the set of all "major" deals.

Columns (3) and (4) of Table 2 show significant decreases in both false negatives *and* false positives. First, those deals that *shouldn't* become bestsellers become much less likely to succeed. At the pre-digitization mean of 0.159 false negatives, the coefficient of -0.094 implies a decrease in the probability of a false negative by 59%. At the same time, deals that *should* become bestsellers become less likely to remain obscure. At the pre-digitization mean of 0.540, the coefficient of -0.184 implies a decrease in the probability of a false positive by 34%. Overall, our results suggest that digitization allows the ex-ante advances to better reflect ex-post success.

4.2.3 Timing

While we choose 2008 as the treatment year in our main regressions, the improvements in prediction may only materialize with a lag if information only gradually becomes available. To study the effects on prediction in each individual year, we allow for a flexible time structure in the spirit of Autor

(2003):

$$Y_{i,j,k,t} = \alpha + \beta R_j + \sum_{\tau \in T} \delta_\tau \left(\gamma_\tau \times R_j\right) + \kappa C_j + \mu_t + \nu_k + \varepsilon_{i,j,k,t}, \quad (3)$$

where γ^τ denotes annual dummy variables, and $Y_{i,j,k,t}$ describes the same functions of precision as above. The omitted year is 2008, to facilitate a comparison of pre- and post-years.

We report the coefficients of interest in Figure 3. The two top panels are parallel to columns 1 and 2 of Table 2 and describe changes in the residuals from regressions of ex-post success. Both panels show that the yearly coefficients of the difference between romance and non-romance deals decrease sharply after 2008, reaching their lowest points by 2010. For the specification based on public information, they decrease to about -0.10, for a decrease of about 34%; and the coefficients based on deal advances sink even lower, to -0.15. The results suggest that publishers learn significantly more about an idea's potential than outsiders (with publicly available data) can.

The bottom panels of Figure 3 reflect columns 3 and 4 of Table 2. They show that while the reduction in false negatives (successful "nice" deals, in the left panel) is immediate, the share of major deals that do not make it to the Top 150 (false positives) decreases only gradually, with the annual coefficients only becoming statistically significant several years into the digitization period. These dynamics suggest that the changes in the information environment may not set in immediately, perhaps due to changes in the size of advances.

To quantify differences over time, we separately estimate the effects of digitization in the first two years after the digitization treatment (2009 and 2010) and in the following years (2011 and later). Table 3 shows that the improvements in prediction become larger after a few years in all specifications. For false positives, the effect only becomes statistically significant after 2010. We examine the possible sources of this pattern in Section 4.4.

4.3 Heterogeneity and Spillovers

We have established that publishers can utilize the additional information made available through digitization, but it remains unclear whether all authors and publishers benefit equally or if there is significant heterogeneity in improvements across deals. Here, we examine information spillovers to an author's other books, as well as the role of improvements in information on the intensive and extensive level (established and new authors, respectively), and the role of resources available for data analytics (major publishers versus independents).

Spillovers Within Authors Learning can be specific to a certain book or story, or it can pertain to the appeal of a certain author more generally. To examine the relative importance of these effects, we estimate the separate changes in the explanatory power of advances on romance books and on non-romance books by romance authors, in Table 4. Interestingly, the residuals from the deal size regressions (column 2) decrease significantly for both groups. For romance deals, they decrease by about 30.7% at the baseline average (=0.092/0.300); and for non-romance books by authors with experience in the romance genre, they decrease by about 22.3% (=0.059/0.262). The same holds true for the other specifications as well. That is, the improvements in information apply both to specific content and to the creators of this content more generally.

Varying Effects Across Authors Publishers may be able to glean insights from *other*, related authors. Any information spillovers across authors are likely to be valuable for first-time authors, but they may also significantly raise precision among established writers if publishers haven't previously taken advantage of available information about their histories. We examine these potential differences by estimating the effects across new ("rookie") and established ("veteran") authors. We first estimate a variant of equation (2) that also includes interactions with a dummy variable that is 1 if the author has published a book before the deal. We then estimate the absolute values of the residuals in a variant of equation (1) that includes interactions of the After × Romance variable with the two author experience indicators.

The results are reported in columns (1) through (3) of Table 5. We see significant decreases in the sizes of the residuals (column 1) among all authors, and no significant difference between types of authors. At baseline residual sizes of 0.279 for rookies and 0.289 for veteran authors, the coefficients imply improvements of 31% and 28%, respectively. Columns (2) and (3) tell a similar story, although the decreases in false negatives and false positives seem slightly more pronounced among rookie authors.

We further examine whether the role of the author's experience depends on *where* the author has previously published. We classify authors as previously self-published if they had published at least one book with one of the five most well-represented self-publishing platforms before the deal.[23] If we drop all authors with a self-publishing background based on this definition, the estimated effects on precision remain statistically significant but decrease slightly, which suggests that publishers gain even more information through the self-publishing channel.

Varying Effects Across Publishers While all publishers have access to more information, larger publishers may be able to allot more resources to using this information for business decisions. To examine whether larger publishers benefit more from the information, we add interactions of all variables with major-publisher dummies to both steps of the regressions. The results, in columns (4) through (6) of Table 5, confirm our expectations. While both independent and major publishers benefit, major publishers seem to benefit even more. From the baseline sizes of the residuals (0.25 for independents and 0.34 for majors), we estimate a decrease of 18% (to 0.21) among independents and 33% (to 0.23) among majors. Note that independent publishers had much smaller residuals before digitization, and that the residuals for the two groups are almost identical after 2008. These initial differences may be due to the major publishers' bargaining positions coupled with higher success rates. While major deals were less likely to flop before digitization if they involved a major publisher (49% vs. 63%), the share of false negatives – errors in favor of publishers – was almost

[23] We find this information through the Bowker Books-in-Print directory, which lists all titles that receive an ISBN in a given month. Our measure for self-publishing histories is imprecise because not all self-published titles are reported in this database.

twice as large among majors (24% vs. 13%) before 2008, and decreased much more in absolute terms (15.2 vs. 6.7 percentage points) after digitization.

The result that the improvements in precision are larger for major publishers could be due to differences in romance-specific editor experience or to differences in resources allocated to data analytics. In separate regressions, we treat editors as experienced in romance if they are listed on at least one previous romance deal on the Publishers Marketplace database. We find some evidence that editors with romance experience see slightly larger improvements in prediction precision than other editors. However, this does not explain differences in improvements between independent and major publishers: while the share of experienced editors increases over time (by construction), there is virtually no difference in the annual shares between the two publisher types.

By contrast, we find large differences in the allocation of resources toward data analytics. Using data from 46,271 job advertisements posted from 2004 to 2015 by publishers on the Publishers Marketplace website, we count the number of job advertisements that are for data-related positions, separately for independent and major publishers.[24] We plot the cumulative number of data-related positions per posting publisher in Figure 4. On average, major publishers had advertised more than one data-related position on Publishers Marketplace by 2015 – an order of magnitude more than independent publishers.[25] Accordingly, an analysis mirroring column (4) of Table 5, which uses the publisher-type specific cumulative number of data jobs instead of independent and major publisher dummies, underlines this correlation and suggests an important role of data-driven decision making. The coefficient on the triple interaction, After × Romance × Data Jobs, is -0.228 (std. error = 0.142).

[24] Our definition of data-related positions include all ads where the text includes variants of "analytic," "data," "intelligence," "finance," and "tax."

[25] The *share* of postings that are for data-related positions is also larger among major publishers than others. For example, from 2011 to 2015, about 3 precent of all postings by majors advertised a data-related position, compared to just 1% among independents.

4.4 Mechanisms

Beyond improvements in the information environment, our result that publishers' advance payments better reflect a book's ex-post success could be the net effect of two underlying mechanisms: changes in the relationships between publishers and authors that lead to changes in the license payments; and/or changes in demand in the final product market. In this section, we provide evidence that both types of mechanisms may be at play, but that neither can fully explain our main results.

4.4.1 Changes in the Licensing Market

We first focus on the interactions between publishers and authors in the licensing market. As discussed above, digitization may have changed the authors' outside options through disintermediation by self-publishing, which could in turn induce publishers to offer more attractive deals in the licensing market. Here, we study this possibility empirically, using variants of equation (1).

The estimation results are reported in Table 6. Our preferred specification, in column (1), uses the natural logarithm of the license payment, measured at the midpoints of each deal category in the data. For "major" deals, we set the midpoint at $750,000. The coefficient on the interaction term (After × Romance) is positive and statistically significant, suggesting that digitization increased license payments by 12.1% ($=e^{0.114}-1$). Columns (2) through (4) confirm the results with different functions of the dependent variable, including the untransformed midpoints of the deal categories (in thousands, column 2), categorical size variables (ordered from 1 to 5, column 3), and the base ten log transformation (column 4). The latter takes additional information on the deal sizes from unstructured text data (e.g. "six-figure deal") into account and thus alleviates concerns arrising from the fact that the reported deal sizes are capped at "$500,000 and above." The estimated increases in advance payments vary between 7.2% (column 3) and 20.3% (column 2).[26]

We plot the estimated year-specific difference coefficients from a variant of equation (3) in Figure 5. The coefficients are not statistically different from zero in any year before 2008, providing evidence

[26] In columns (2) and (3), we compare the coefficient to the sample mean, i.e. $(28.32/139.27) \times 100\% = 20.3\%$ in column (2) and $(0.139/1.94) \times 100\% = 7.15\%$ in column (3). In column (4), the estimated percentage impact is $10^{0.039} - 1 = 9.3\%$.

that the identifying assumption of the difference-in-differences model holds. They become large and significantly positive immediately after 2008, with a digitization-related increase in license payments between 10% and 20%, compared to the years before 2008.

An additional strategy for attracting an author is to include several books in one deal – thus providing the authors more certainty. Column (5) of Table 6 shows that publishers utilize this strategy more after digitization, as the probability that a deal includes multiple books increases by 5.6 percentage points. In addition to providing authors more certainty, this result implies an increase in efficiency by limiting negotiations per book. The changes in the types of deals are a direct consequence of the mechanisms we describe in our research framework. First, publishers may shoulder a larger portion of the risk to compete with the author's improved outside option. Second, they may be particularly willing to do so because the improved information environment has significantly lowered their own risk. Third, deals including multiple books may also become more popular because of a shift in both supply and demand in the consumer market, if books and stories tend to become shorter and more serialized overall.

Importantly, while the increases in advances to romance authors could be driving the decrease in false negatives shown in column 3 of Table 3, they cannot explain the decrease in false positives (shown in column 4 of that table). Specifically, the immediate increase in advances could explain the gradual decrease in false positives shown in the bottom right panel of Figure 3. It is possible that publishers reacted too strongly to changes in the market at first, but they learned over time, offering better deals only to those authors who would go on to draw larger audiences.

4.4.2 Changes in the Final Product Market

We now turn to the dynamics of demand in the final product market: the interactions between publishers and consumers. It is possible that the introduction of e-readers and the ability to read books more discretely disproportionately affected the demand for romance books (Goldfarb et al., 2015), which could in turn affect the accuracy of advance deals.

Inspired by the descriptive analyses summarized in Figure 2, we examine the role of two major successes in the romance category after 2008: Twilight (released as a movie in 2008, based on the 2005 novel) and Fifty Shades of Grey (originally published in 2011). Publishers may have seen an opportunity to "piggyback" off the success of these titles and therefore been willing to invest more money in similar books. To examine this possibility, we identify authors in our dataset who write stories similar to Twilight and Fifty Shades of Grey ("T&F").[27] We use this information to separately estimate the effects on contracts for romance books that are similar to these book series and for other romance books.

The first two columns of Table 7 show the estimated coefficients from the deal size regressions. While deal sizes for books like T&F increased substantially, and highly significantly, by over 150% (column 1) or over $200,000 (column 2), the effect is not limited to these books. Advances for other romance books also increased significantly, albeit just by about 8% or close to $20,000. That is, publishers increased advance payments for all books – not just those that would most obviously see the largest demand increases.

Columns (3) through (5) of Table 7 examine the changes in precision separately for the two types of romance books/content. Unlike the effects on the size of the advance, advances predict ex-post success *more* precisely among romance books that are unrelated to T&F. Among these books, the sizes of the residuals decrease by about 23% (=0.063/0.27, column 3), and both false negatives and false positives become less likely to occur (columns 4 and 5). In contrast, predictions seem to become *less* accurate among romance books that are similar to T&F. The coefficient of +0.049 implies a statistically significant increase in the size of the residuals by 41%. The next two columns suggest that this increase is driven by changes in medium-sized deals: The smallest deals become less likely to be successful (fewer false negatives, column 4), and the largest deals remain similarly successful (column 5).

Together, these results suggest that the improvements in prediction cannot be driven entirely by

[27] We find these authors by searching through lists on the major book review site Goodreads. The lists we consult are: 437 books from "If you liked Twilight" (`https://www.goodreads.com/list/show/5322.If_You_Liked_Twilight`; 3118 books from "Best M/F Erotic Romance like Fifty Shades of Grey (not paranormal, high school, gay or sci-fi)" (`https://www.goodreads.com/list/show/18698.Best_M_F_Erotic_Romance_like_Fifty_Shades_of_Grey_not_paranormal_high_school_gay_or_sci_fi`); and 105 books from "Books Like Fifty Shades of Grey" (`https://www.goodreads.com/list/show/31201.Books_Like_Fifty_Shades_of_Grey`).

changes in demand. In fact, deals for books that were ex-ante most likely to experience demand increases may have become less precise. This could be because the expected demand increases didn't consistently materialize, so that publishers paid licensing fees that were too high for some of the books that are similar to T&F. Regardless, even if demand for romance novels increased disproportionately, note that such effects by themselves are inconsistent with the immediate decrease in false negatives that we estimate above. Rather, publishers offer more money for those books that become successful, and are less likely to offer large advances for books that do not – a clear indication that publishers are better able to make use of the available information.

4.4.3 Alternative Explanations: Effects of Entry and Competition

Although the presented evidence is strong, one might be worried about alternative explanations for our results. Specifically, massive-scale entry due to digitization and corresponding changes in competition may play a role in explaining the improved accuracy of the contracts between authors and publishers. One possibility is that an increase in the number of romance authors could lead to increased competition among authors for a limited number of publishers and less market power on *both* sides of the deals market. In that case, what we interpret as a decrease in the prediction error could in fact be an improvement in efficiency due to increased competition. We think this explanation is unlikely, for two reasons. First, increased competition among authors is not consistent with an overall increase in license payments. Second, the number of (posted) romance and non-romance deals increased substantially after 2008, and more so for romance deals. This suggests an overall market expansion and implies that publishers most likely had no binding resource constraints. That is, we find no evidence that publishers become a bottleneck that would lead to fiercer competition for license deals among authors. Consistent with this, our finding that license deals increased rather than decreased further suggests that competition among authors has not intensified with digitization.

We identify one more possible reason for the improvements in precision. The influx of authors (through self-publishing) likely led to increased competition for consumers of romance novels and perhaps to less demand for each individual book. If the added competition lowers expectations of demand for individual romance novels, one would expect license payments for each romance book to decrease, perhaps leading to fewer false positives. However, again, this explanation is inconsistent with the overall increase in deal sizes and with the decrease in the likelihood of false negatives.

5 Discussion and Conclusion

Digitization has largely affected demand and supply in the creative industries. In addition to effecting a large influx of new products, it has allowed creators of content to circumvent traditional gatekeepers and directly reach consumers, and it has facilitated access to information for both consumers and producers of content. We find that these developments have changed the contracts between the creators (authors) and distributors (publishers) of content in the book publishing industry. Advances better explain a book's ex-post success, likely because market participants have better access to sales of similar books and information about consumer preferences. These changes yield a number of interesting implications for efficiency and market structure, and therefore ultimately for managers and overall welfare as well.

5.1 Managerial and Welfare Implications

We can use our deal-level estimates to speculate about the overall amount of money that is now spent more efficiently. The decrease in false negatives of about 60% implies that over the time period we study, about $2.5 million of all the money spent on "nice" deals (that is, almost 10% of $25.9 million) will no longer be misallocated. Conversely, the decrease in false positives implies that publishers no longer falsely spend (or waste) $15.3 million on deals that do not become successful.[28] This corresponds to savings of more than 18% of the total amount spent on major deals.

The improvements in prediction precision can have large impacts on the composition of publishers in the market. On one hand, independent publishers traditionally carry more risk because they have

[28] We observe 165 "nice" romance deals before 2008 that are false negatives, for an assumed average advance of $25k. We also observe 60 "major" romance deals ($750k) before 2008 that are false positives. If we assume that "major" deals imply average advances of $500k instead of $750k, the savings would be $10.2 million.

smaller portfolios of books, so that improving prediction precision can particularly benefit small publishers. On the other hand, our results suggest that major publishers see larger prediction improvements than independents do. Thus, it is unclear which publishers benefit the most from improved information. There has been a large influx of independent publishers since the onset of digitization (Waldfogel and Reimers, 2015), but we do not know if these new publishers are consequential, nor if they pursue publishing as their only source of income. The share of books by major publishers in the weekly Top 150 bestseller lists has remained quite constant between 2003 and 2014, around 60 to 70%. Any observed changes in the publisher landscape could be due to decreased entry costs or to reduced risk from improved information. We leave the distinction between these effects for future research.

The improvements in the information environment also affect the tasks carried out by traditional publishers. If the information improvements facilitate scouting and discovery of "good" content, publishers can allocate the resources that were traditionally used for identifying "winners" toward editing and marketing books. Yet, we see larger improvements in predictions among the types of publishers who invest more in data analytics (as opposed to editing). The best allocation of resources within firms likely incorporates a combination of both traditional and data-driven tasks. Finally, the reduction in false negatives implies that more (good) books will enter the market through traditional publishers, who continue to allocate more resources to marketing the books than self-publishing allows. By examining the effects of digitization on the contracts between the creators and distributors of creative content, we also expand on the existing literature on the effects of digitization on the long tail. For example, the welfare benefits in Aguiar and Waldfogel (2018) are due to reductions in fixed costs and the fact that appeal is not predictable. Our finding that appeal becomes more predictable implies that those products that *will* be produced are "better," so that improvements in consumer surplus would even be possible without decreases in traditional publishers' fixed cost. That is, the improvements in the information environment make the publishing market more efficient.

5.2 Implications Beyond Traditional Book Publishing

Our study is limited to observing the interactions between authors and traditional publishers, but we argue that the efficiency improvements we find have long-term implications for new entrants and digital self-publishing as well. Traditional institutions, which are better able to market products, can exist alongside new platforms that allow creators to reach consumers directly, albeit in a new role. Before digitization, traditional publishers served as gatekeepers who ensured that consumers could read the books they deemed good while those books that didn't pass the bar would not be published. Now that virtually any book can be published, traditional publishers become curators who signal a book's quality; and if publishers allocate more resources to marketing and editing, they can provide additional incentives for authors to publish traditionally.

Our finding that improvements in information can lead to more efficient market outcomes applies to settings beyond the creative industries. For example, in health care, information about a patient's history can help physicians choose the most effective treatment (Ribers and Ullrich, 2019). Policy makers should consider the tradeoff between these benefits and privacy protection when deciding about access to personal or firm data. Our paper adds empirical evidence to a theoretical discussion of data externalities (Jones and Tonetti, 2020). Forcing firms to share data likely improves welfare by increasing competition (Reimers and Shiller, 2019), but it can also improve welfare dynamically by enabling better prediction about future markets and increasing incentives to innovate in socially valuable products.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). "When should you adjust standard errors for clustering?" Tech. rep., National Bureau of Economic Research.

Aguiar, L., and Waldfogel, J. (2018). "Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music." *Journal of Political Economy*, *126*(2), 492–524.

Amador, M., and Weill, P.-O. (2012). "Learning from private and public observations of others' actions." *Journal of Economic Theory*, *147*(3), 910–940.

Archer, J., and Jockers, M. L. (2016). *The bestseller code: Anatomy of the blockbuster novel*. St. Martin's Press.

Autor, D. H. (2003). "Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing." *Journal of Labor Economics*, *21*(1), 1–42.

Benner, M. J., and Tripsas, M. (2012). "The influence of prior industry affiliation on framing in nascent industries: The evolution of digital cameras." *Strategic Management Journal*, *33*(3), 277–302.

Benner, M. J., and Waldfogel, J. (2016). "The song remains the same? technological change and positioning in the recorded music industry." *Strategy Science*, *1*(3), 129–147.

Bergemann, D., and Välimäki, J. (2000). "Experimentation in markets." *The Review of Economic Studies*, *67*(2), 213–234.

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

Brynjolfsson, E., Hu, Y., and Smith, M. D. (2003). "Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers." *Management Science*, *49*(11), 1580–1596.

Brynjolfsson, E., and McElheran, K. (2016). "The rapid adoption of data-driven decision-making." *American Economic Review*, *106*(5), 133–39.

Burtch, G., Carnahan, S., and Greenwood, B. N. (2018). "Can you gig it? an empirical examination of the gig economy and entrepreneurial activity." *Management Science*, *64*(12), 5497–5520.

Caplin, A., and Leahy, J. (1998). "Miracle on sixth avenue: information externalities and search." *The Economic Journal*, *108*(446), 60–74.

Caves, R. E. (2000). *Creative industries: Contracts between art and commerce*. Harvard University Press.

Chevalier, J. A., and Mayzlin, D. (2006). "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research*, *43*(3), 345–354.

Claussen, J., Peukert, C., and Sen, A. (2019). "The editor vs. the algorithm: Returns to data and externalities in online news."

Foerderer, J., Kude, T., Mithas, S., and Heinzl, A. (2018). “Does platform owner’s entry crowd out innovation? evidence from google photos.” *Information Systems Research*, *29*(2), 444–460.

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

Greco, A. N. (2013). *The book publishing industry.* Routledge.

Gu, G., and Zhu, F. (2018). “Trust and disintermediation: Evidence from an online freelance marketplace.” *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (18-103).

He, S., Peng, J., Li, J., and Xu, L. (2020). “Impact of platform owner’s entry on third-party stores.” *Information Systems Research*, *31*(4), 1467–1484.

Huang, L.-C., Shiau, W.-L., and Lin, Y.-H. (2017). “What factors satisfy e-book store customers? development of a model to evaluate e-book user behavior and satisfaction.” *Internet Research.*

Jones, C. I., and Tonetti, C. (2020). “Nonrivalry and the economics of data.” *American Economic Review*, *110*(9), 2819–58.

Jung, J., Chan-Olmsted, S., Park, B., and Kim, Y. (2012). “Factors affecting e-book reader awareness, interest, and intention to use.” *New media & society*, *14*(2), 204–224.

KC, D., Staats, B. R., and Gino, F. (2013). “Learning from my success and from others’ failure: Evidence from minimally invasive cardiac surgery.” *Management Science*, *59*(11), 2435–2449.

Laquintano, T. (2013). “The legacy of the vanity press and digital transitions.” *Journal of Electronic Publishing*, *16*(1).

Levine, M. (2016). *The Fine Print of Self-Publishing: A Primer on Contracts, Printing Costs, Royalties, Distribution, Ebooks, and Marketing.* North Loop Books.

Madsen, P. M., and Desai, V. (2010). “Failing to learn? the effects of failure and success on organizational learning in the global orbital launch vehicle industry.” *Academy of management journal*, *53*(3), 451–476.

McCartney, J. (2016). “Self-publishing preview: 2016.” *https://www.publishersweekly.com/pw/by-topic/authors/pw-select/article/69156-self-publishing-preview-2016.html*, Publishers Weekly (January 15, 2016).

Ordanini, A. (2006). “Selection models in the music industry: How a prior independent experience may affect chart success.” *Journal of Cultural Economics*, *30*(3), 183–200.

Phillips, S. (2016). “Can big data find the next ‘harry potter’?” *The Atlantic.*

Reimers, I., and Shiller, B. R. (2019). “The impacts of telematics on competition and consumer behavior in insurance.” *The Journal of Law and Economics*, *62*(4), 613–632.

Reimers, I. C., and Waldfogel, J. (2021). “Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings.” *forthcoming in American Economic Review.*

Ribers, M. A., and Ullrich, H. (2019). "Battling antibiotic resistance: can machine learning improve prescribing?"

Shen, Q., and Xiao, P. (2014). "Mcdonald's and kfc in china: Competitors or companions?" *Marketing Science*, *33*(2), 287–307.

Spulber, D. F. (1996). "Market microstructure and intermediation." *Journal of Economic perspectives*, *10*(3), 135–152.

Srinivasan, R., Haunschild, P., and Grewal, R. (2007). "Vicarious learning in new product introductions in the early years of a converging market." *Management Science*, *53*(1), 16–28.

Tambe, P. (2014). "Big data investment, skills, and firm value." *Management Science*, *60*(6), 1452–1469.

Thompson, J. B. (2013). *Merchants of culture: the publishing business in the twenty-first century.* John Wiley & Sons.

Toivanen, O., and Waterson, M. (2005). "Market structure and entry: Where's the beef?" *RAND Journal of Economics*, 680–699.

Torres, R., Johnson, V., and Imhonde, B. (2014). "The impact of content type and availability on ebook reader adoption." *Journal of Computer Information Systems*, *54*(4), 42–51.

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

Waldfogel, J., and Reimers, I. (2015). "Storming the gatekeepers: Digital disintermediation in the market for books." *Information Economics and Policy*, *31*, 47–58.

Wu, L., Hitt, L., and Lou, B. (2020). "Data analytics, innovation, and firm productivity." *Management Science*, *66*(5), 2017–2039.

Yang, N. (2018). "Learning in retail entry." *Available at SSRN 1957992.*

Zegners, D. (2017). "Building an online reputation with free content: Evidence from the e-book market." *Available at SSRN 2753635.*

Zhu, F., and Liu, Q. (2018). "Competing with complementors: An empirical look at amazon. com." *Strategic management journal*, *39*(10), 2618–2642.

Table 1: Descriptive Statistics: Book Deals

	Romance (N=2,987)				Non-Romance (N=14,149)				Total	
	Before		After		Before		After			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Acclaimed	0.052	0.222	0.036	0.187	0.056	0.229	0.062	0.241	0.056	0.229
Prev. Bestseller	0.100	0.300	0.199	0.399	0.035	0.183	0.073	0.260	0.070	0.256
Competition	0.027	0.162	0.016	0.126	0.049	0.216	0.078	0.269	0.055	0.228
Orig. Self-published	0.001	0.029	0.017	0.130	0.005	0.067	0.005	0.071	0.006	0.076
Sequel	0.026	0.159	0.037	0.190	0.022	0.146	0.027	0.163	0.026	0.158
Debut	0.055	0.229	0.032	0.177	0.042	0.202	0.045	0.208	0.043	0.204
Prev. Titles	3.352	5.245	6.621	7.265	3.074	5.427	4.441	7.139	3.965	6.390
Major publisher	0.408	0.492	0.374	0.484	0.507	0.500	0.441	0.497	0.462	0.499
Log(Deal Size)	10.782	1.069	10.888	1.193	11.027	1.202	11.120	1.205	11.030	1.197
Deal Size	106.045	182.623	134.117	221.677	144.293	219.912	151.656	217.820	143.260	217.229
Deal Size Categories	1.701	1.193	1.834	1.357	1.974	1.360	2.069	1.360	1.975	1.352
Future Top 150	0.295	0.456	0.258	0.438	0.081	0.273	0.085	0.279	0.116	0.320
Pr(Success | small deal)	0.203	0.402	0.097	0.296	0.027	0.162	0.014	0.116	0.045	0.208
Pr(No success | major deal)	0.301	0.462	0.089	0.286	0.662	0.473	0.599	0.490	0.561	0.496
Observations	1,191		1,796		7,885		6,264		17,136	

Table 2: Baseline Results: Predicting Ex-Post Appeal

	Abs(residuals)		Cond. on size	
	(1) Public info	(2) Deal size	(3) False Neg	(4) False Pos
Romance	0.078***	0.103***	0.149***	-0.104
	(0.007)	(0.007)	(0.014)	(0.069)
After 2008 × Romance	-0.049***	-0.080***	-0.094***	-0.184***
	(0.010)	(0.010)	(0.014)	(0.037)
Acclaimed	0.001	0.003	-0.007	0.097
	(0.009)	(0.009)	(0.009)	(0.084)
Prev. Bestseller	0.109**	0.063**	0.109**	-0.250***
	(0.036)	(0.020)	(0.045)	(0.053)
Competition	0.012	0.030**	-0.011	0.038
	(0.009)	(0.010)	(0.027)	(0.051)
Debut	-0.021	-0.005	-0.033	0.089
	(0.012)	(0.006)	(0.024)	(0.107)
Orig. Self-published	0.068*	0.031	-0.045	-0.182
	(0.031)	(0.018)	(0.028)	(0.129)
Sequel	0.023*	0.019	0.014	-0.004
	(0.012)	(0.011)	(0.020)	(0.070)
Observations	17136	17136	10000	1648
$\overline{R}^2$	0.424	0.393	0.087	0.282

Notes: Editor and month-year fixed effects not reported. The dependent variables in columns (1) and (2) describe the absolute value of the residuals from regressions of ex-post success on author and book characteristics (1) and deal sizes (2). Column (3) looks at "nice" deals (below $50k) and the dependent variable is 1 if the author becomes successful. Column (4) looks at "major" deals (above $500k) and the dependent variable is 1 if the author does *not* become successful. Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Predicting Ex-Post Appeal Over Time

	Abs(residuals)		Cond. on size	
	(1) Public info	(2) Deal size	(3) False Neg	(4) False Pos
Years 2009 & 2010 × Romance	-0.029**	-0.048***	-0.070***	-0.202
	(0.011)	(0.009)	(0.013)	(0.115)
After 2010 × Romance	-0.060***	-0.098***	-0.110***	-0.178**
	(0.010)	(0.014)	(0.019)	(0.057)
Observations	17136	17136	10000	1648
$\overline{R}^2$	0.424	0.394	0.088	0.281

Notes: Editor, month-year fixed effects and control variables not reported. The dependent variables in columns (1) and (2) describe the absolute value of the residuals from regressions of ex-post success on author and book characteristics (1) and deal sizes (2). Column (3) looks at "nice" deals (below $50k) and the dependent variable is 1 if the author becomes successful. Column (4) looks at "major" deals (above $500k) and the dependent variable is 1 if the author does *not* become successful. Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Author vs. Book Effects

	Abs(residuals)		Cond. on size	
	(1) Public info	(2) Deal size	(3) False neg	(4) False pos
After 2008 × romance author				
. . . × Romance deal	-0.054***	-0.092***	-0.110***	-0.232**
	(0.011)	(0.007)	(0.021)	(0.091)
. . . × Non-romance deal	-0.050**	-0.059***	-0.063**	-0.119
	(0.020)	(0.017)	(0.026)	(0.098)
Observations	17136	17136	10000	1648
$\overline{R}^2$	0.416	0.392	0.087	0.280

Notes: Editor, month-year fixed effects and control variables not reported. "Romance deal" and "Non-romance deal" are mutually exclusive groups. The omitted category is non-romance deals with non-romance authors. The dependent variables in columns (1) and (2) describe the absolute value of the residuals from regressions of ex-post success on author and book characteristics (1) and deal sizes (2). Column (3) looks at "nice" deals (below $50k) and the dependent variable is 1 if the author becomes successful. Column (4) looks at "major" deals (above $500k) and the dependent variable is 1 if the author does *not* become successful. Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Predicting Ex-Post Appeal: Author and Publisher Types

	Author experience			Publisher size		
	(1)	(2)	(3)	(4)	(5)	(6)
	Residuals (Deal size)	False neg	False pos	Residuals (Deal size)	False neg	False pos
After 2008 × Romance						
. . . × Rookie author	-0.087***	-0.120***	-0.355*			
	(0.017)	(0.039)	(0.195)			
. . . × Veteran author	-0.079***	-0.080***	-0.134			
	(0.012)	(0.028)	(0.132)			
. . . × Indie publisher				-0.046***	-0.067**	-0.308
				(0.013)	(0.026)	(0.218)
. . . × Major publisher				-0.110***	-0.152***	-0.128
				(0.015)	(0.048)	(0.125)
Observations	17136	10000	1648	17136	10000	1648
$\overline{R^2}$	0.379	0.088	0.285	0.391	0.089	0.282

Notes: Editor, month-year fixed effects and control variables not reported. The dependent variables in columns (1) and (4) describe the absolute value of the residuals from regressions of ex-post success on deal sizes. Columns (2) and (5) look at "nice" deals (below $50k) and the dependent variable is 1 if the author becomes successful. Columns (3) and (6) look at "major" deals (above $500k) and the dependent variable is 1 if the author does *not* become successful. "Rookie" and "Veteran" authors are mutually exclusive groups, as are "Indie" and " Major" publishers. Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Changes in License Deals

	(1) Ln-Size	(2) Size	(3) Deal cat	(4) Log-Size	(5) Multibook
Romance	-0.134**	-22.971**	-0.151**	-0.055***	0.015
	(0.048)	(8.086)	(0.053)	(0.016)	(0.018)
After 2008 × Romance	0.114***	28.321***	0.139***	0.039***	0.056***
	(0.029)	(5.182)	(0.033)	(0.012)	(0.011)
Observations	17136	17136	17136	17139	17136
$\overline{R^2}$	0.541	0.408	0.526	0.453	0.214

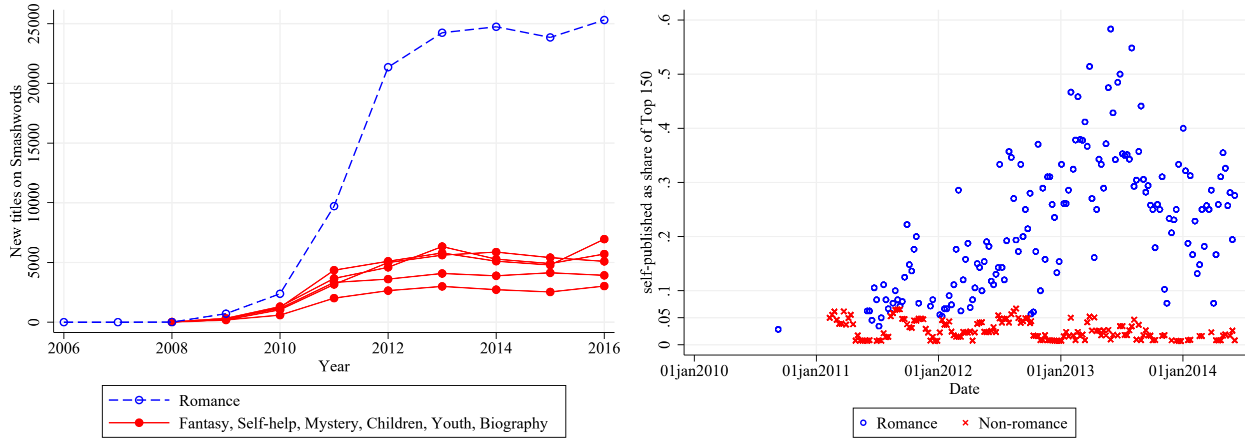
Notes: Editor, month-year fixed effects and control variables not reported. Column headers indicate the dependent variables, where sizes are reported at the midpoints of the deal size categories; and multibook is a dummy that is 1 if the deal included more than one book. Standard errors in parentheses, clustered at the genre-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: The Effects of Twilight and Fifty Shades of Grey

	Deal size		Precision		
	(1)	(2)	(3)	(4)	(5)
			Residuals		
	Ln-Size	Size	(Deal size)	False neg	False pos
After 2008 × Romance					
. . . × Like T&F	0.942***	211.005***	0.049**	-0.371***	-0.024
	(0.194)	(45.390)	(0.019)	(0.014)	(0.384)
. . . × Not like T&F	0.077**	19.380***	-0.063***	-0.082***	-0.188***
	(0.027)	(4.859)	(0.013)	(0.017)	(0.033)
Observations	14547	14547	17136	10000	1648
$\overline{R^2}$	0.542	0.412	0.381	0.145	0.280

Notes: Editor, month-year fixed effects and control variables not reported. "Like T&F" and "Not Like T&F" are mutually exclusive groups of romance novels. The omitted category is non-romance authors. Columns (1) and (2) describe results from the deal size estimations; columns (3)–(5) report results from precision estimations. Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

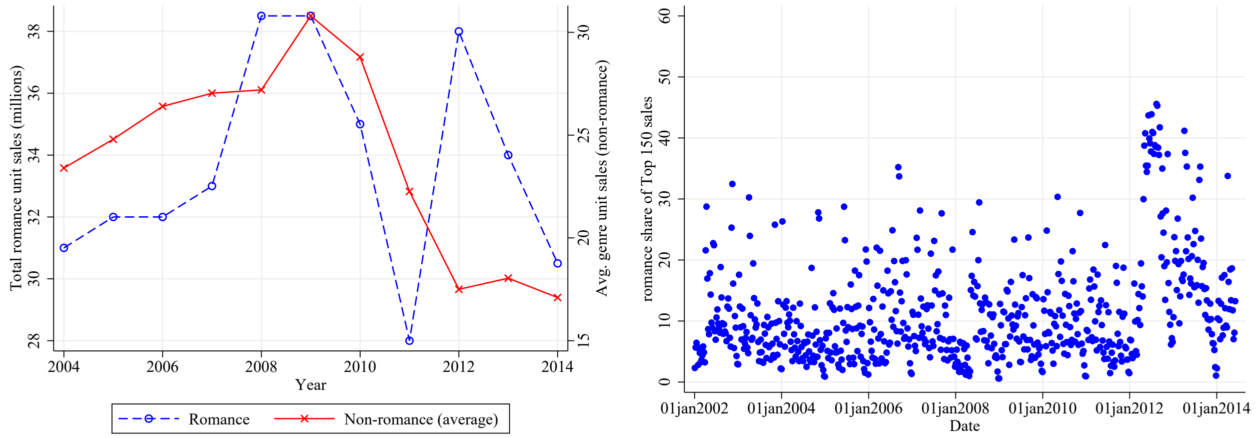
Figure 1: Supply of and Demand for Self-published Titles by Genre



New titles on self-publishing platform Smashwords per year
Source: Smashwords, January 2017

Share of originally self-published books in Top 150
Source: Waldfogel and Reimers (2015)

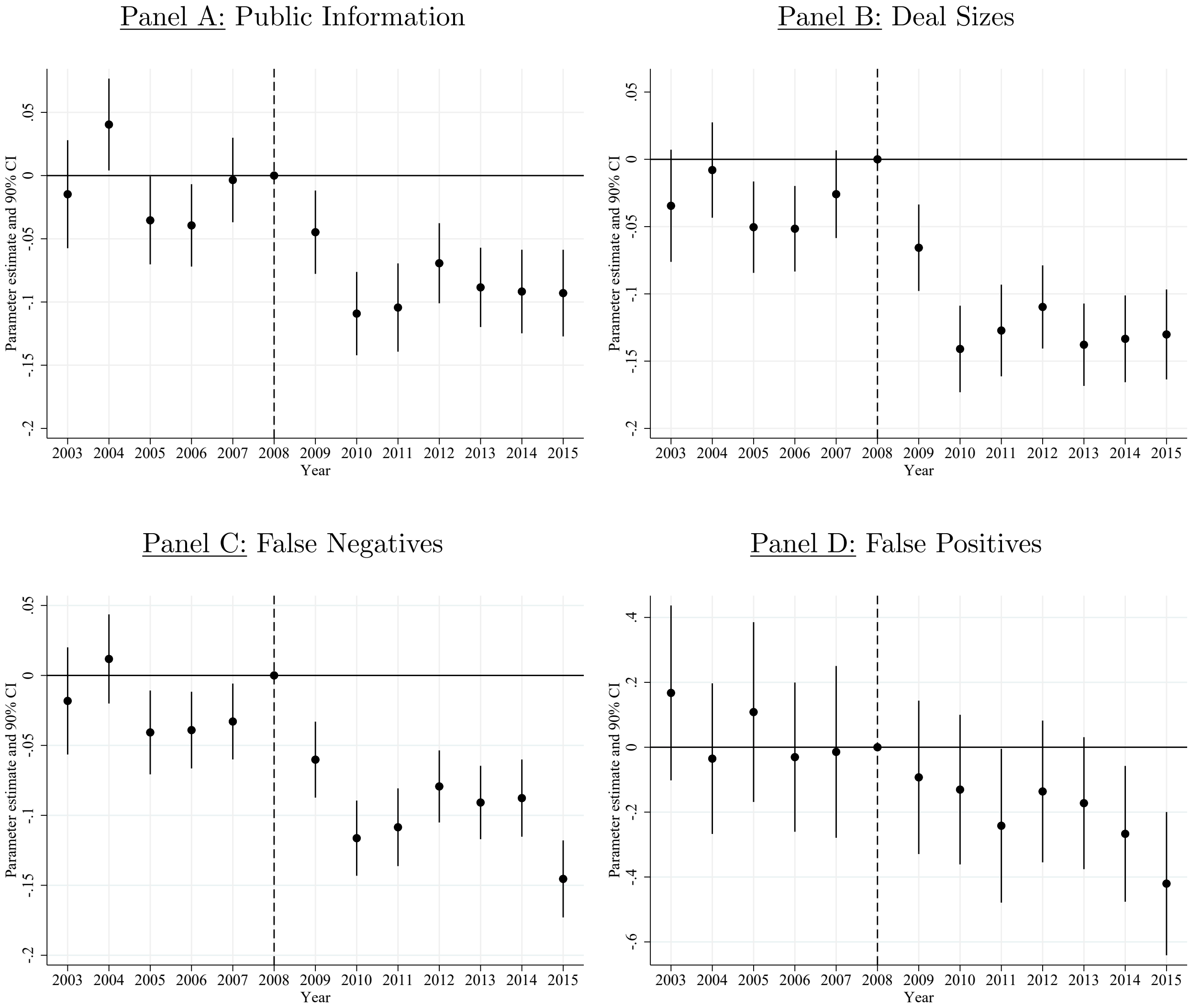
Figure 2: Demand for Romance Books – Traditional and Self-published



Total annual physical unit sales (in millions)
Source: Nielsen

Romance share of pseudo-sales (=1/rank) in Top 150
Source: USA Today

Figure 3: Predicting Ex-Post Appeal – Romance vs. Not



Panel A: Public Information

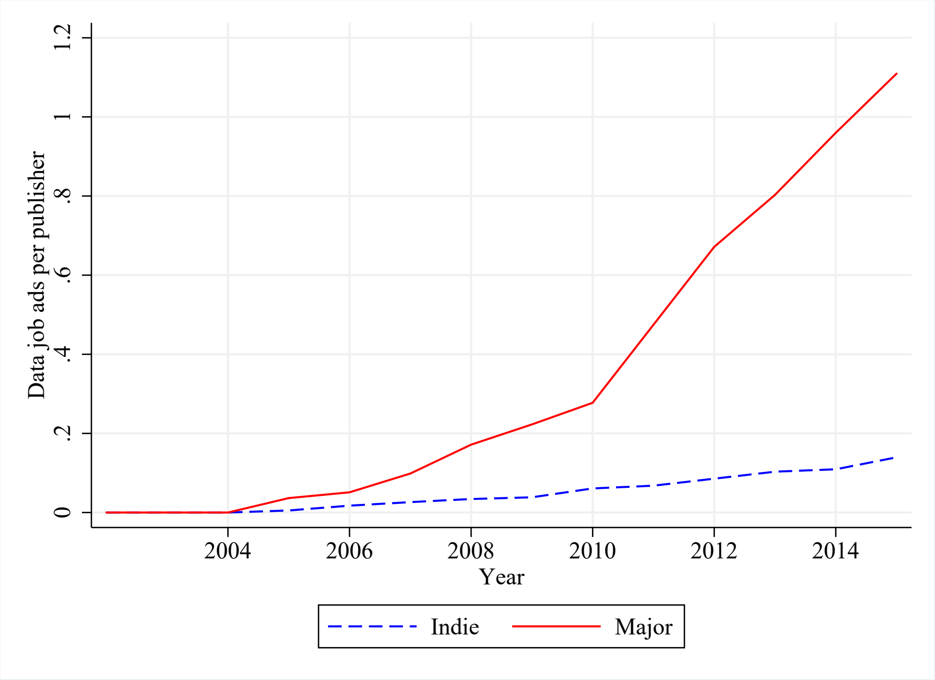
Panel B: Deal Sizes

Panel C: False Negatives

Panel D: False Positives

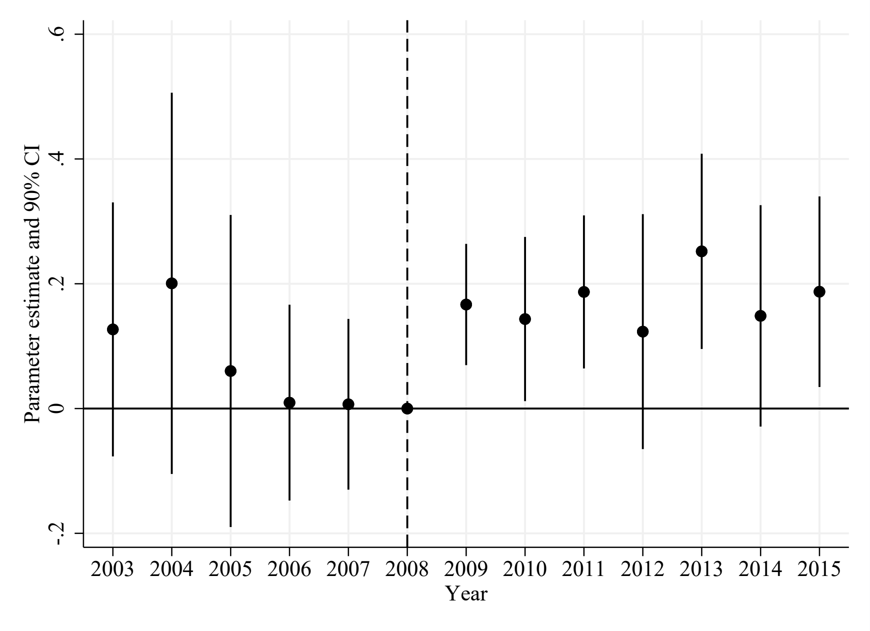
Note: OLS estimates of the δ_τ coefficients obtained from variants of equation (3), using the absolute value of the residuals from regressions of ex-post success on author and book characteristics (panel A) and deal sizes (panel B) as the dependent variable. In panel C we use a dummy that is 1 if "nice" deals (less than \$50k) became successful, and in panel D we use a dummy that is 1 if "major" deals (more than \$500k) did not become successful. The omitted year is 2008. The dots reflect year-specific point estimates between the treatment group (Romance authors) and the control group (non-Romance authors). Standard errors are clustered on the genre-level, and bars indicate 90% confidence bands.

Figure 4: Analytics Job Postings per Publisher



Notes: Cumulative number of data-related job postings per publisher, separately for independent publishers and major publishers. Job posting data is collected from Publishers Marketplace. The numerator includes all job postings up to year t that include data-related keywords. The denominator is the number of distinct publishers in the group that have posted at least one job ad by year t.

Figure 5: License Deals, Group Differences over Time



Notes: OLS estimates of the δ_τ coefficients obtained from a regression of equation (3) with ln(size) as the dependent variable. The omitted year is 2008. The dots reflect estimated yearly differences in ln(size) between the treatment group (Romance authors) and the control group (non-Romance authors). Standard errors are clustered on the genre-level, and bars indicate 90% confidence bands.