Digital Disintermediation and Efficiency in the Market for Ideas*

Christian Peukert
Católica Lisbon School of Business and Economics
christian.peukert@ucp.pt

Imke Reimers
Northeastern University
i.reimers@northeastern.edu

June 29, 2018

Abstract

Digital technology has allowed inventors to circumvent intermediaries, which affects licensing outcomes and efficiency in the market for ideas. We study these impacts theoretically and empirically, using data on over 90,000 license deals for books. Identification comes from quasi-experimental variation across product types over time. Consistent with digital self-publishing improving an author's outside option, authors get more favorable license deals. In addition, ex-ante license fees reflect ex-post demand more accurately. This is consistent with additional entry generating more information. Such improvements can have large impacts on welfare in any markets in which product appeal is difficult to predict.

JEL: D22, D83, L82

*We thank James Dana, Markus Reisinger, Michael Ribers and Joel Waldfogel for valuable comments. 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), and the University of Zurich. We acknowledge support from FCT – Portuguese Foundation of Science and Technology for the project UID/GES/00407/2013.

1 Introduction

Digitization has substantially decreased production and distribution costs in many industries. This 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, 2017; Brynjolfsson et al., 2003). The impact of digitization on the market for ideas – the relationship between the creators of the ideas and their distributors – is not nearly as well-understood.

Online platforms and distribution services have allowed inventors to circumvent traditional intermediaries and directly market their products to consumers. As a first-order effect, more products arrive in the market, and the platforms can improve the inventor's bargaining position when a contract with a traditional intermediary is set up. At the same time, firms can observe the ex-post appeal of not only their own products, but also of their competitors' products, including those which only reach consumers because of lower entry costs. Firms can use such data to predict a product's commercial success before entering into a licensing agreement. In markets in which an idea's success is notoriously difficult to predict, such an improvement can have a large impact on efficiency.

This article examines and quantifies the impacts of such digital disintermediation on the market for licensing contracts between inventors and firms in an environment where these developments are particularly salient: book publishing. Because large investments are needed to produce and distribute physical books, books could traditionally only reach consumers via (large) publishing companies. Publishers would select book ideas based on a comparison of expected downstream consumer demand and the costs of entry. If the expected demand is large enough, publishers and authors agree on the terms of an exclusive license to market and sell the book, including per-unit royalty fees and an upfront license payment that is proportional to the expected ex-post appeal.

With imperfect prediction, some high-quality ideas were likely falsely rejected in this market and never reached consumers despite considerable ex-post appeal. Of course, poor prediction can also lead to false positives. A published book which does not sell as well as expected may represent a loss to the publisher, and a misallocation of publishing resources more generally. Errors of any type can lead to significant inefficiencies, causing under-investment in products which would have created utility in the long run, and over-investment in ideas which do not.

The arrival of digital self-publishing and the increasing diffusion of dedicated e-reading devices have introduced a new channel for authors to reach consumers directly, without relying on traditional publishing houses to recognize (bet on) a book's appeal. Digital self-publishing platforms often require a small fee for making the book available, and they take a (small) share of the revenue from each book sold.[1] However, traditional publishers may be better able to market the book, and because self-publishing is often limited to electronic books (e-books, which require an e-reader), traditional publishers, who continue selling physical books in addition to e-books, can reach more consumers.[2]

To examine the impacts of such disintermediation, this article introduces a simple theoretical model of competition and information in the market for ideas. The key insights from the model are that digital self-publishing directly improves the author's outside option, and it indirectly allows firms and authors to learn about an idea's likely ex-post appeal through related products. The author's improved outside option causes an increase in license payments to authors whose works are well-suited for self-publishing. Other titles enter the market via digital self-publishing even if traditional

[1] The most popular 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 percent of revenue (depending on the chosen price), but no upfront license payments. See `https://kdp.amazon.com/help?topicId=A30F3VI2TH1FR8`.

[2] In 2015, 45% of American adults owned a tablet computer and 19% of American adults owned a designated e-book reader. See `http://www.pewinternet.org/2015/10/29/technology-device-ownership-2015/pi_2015-10-29_device-ownership_0-01/`.

publishers would not have picked them up, and publishers and authors utilize information about their realized demand when making investment decisions.[3] This indirect effect – improved information – could make both publishers and consumers better off: publishers are less likely to incur losses on book deals if they predict the book's success more accurately, and more "good" books enter the market, with traditional publishers allocating more resources to more valuable ideas.

Empirical evidence on the relationship between creators and distributors in the market for ideas is scarce for two reasons. First, the researcher has to observe data on ideas, which are hard to come by. Second, causal inference demands exogenous variation in entry costs. Our setting allows us to deal with both issues. We examine a unique dataset covering contracts of over 90,000 book and rights deals from 2002 to 2015, and we utilize variation in the genres' propensity for self-publishing to estimate the impact of this disintermediation on traditional book deals in a difference-in-differences model. E-books and self-publishing arrived fairly suddenly (driven by the large-scale diffusion of Amazon's Kindle), and they affected different genres differently. Whereas books of most genres are still predominantly published by and consumed through traditional publishers, authors and readers of romance novels have largely embraced self-publishing platforms.[4]

We find that license fees for authors of romance novels increase on average by 15% after the arrival of self-publishing, compared to changes in license fees for authors in other genres. We further provide evidence that these increases are due to digital self-publishing, rather than other supply- and demand-side factors. The increases in advances may go to books which were previously self-published, to authors who have proven their quality with previous, self-published titles, or to authors who write books that are similar to successful self-published works.

[3] Anecdotal evidence suggests that publishers use historical sales data to make business decisions, such as licensing new books. See `http://tinyurl.com/y7dxzynh`.

[4] See sections 2.2 and 4.2 for evidence of as well as reasons for this variation.

We then analyze the publishers' ability to predict an idea's success by examining the relationship between upfront license payments and commercial success. We find that publishers have become better at predicting the commercial success of romance novel authors, compared to authors in other genres and to the pre-digital era, decreasing error rates by about 25%. Importantly, we show a decrease in both false positives (flops) *and* false negatives (missed bestsellers). These results further support our identification strategy, confirming that books by romance authors perform neither unexpectedly well nor unexpectedly poorly. We conclude that digital self-publishing can improve the market's efficiency by making license deals more accurately reflect the value of an idea, with firms investing more money in better ideas, while avoiding investment in ideas which turn out to be less successful.

Our research is closely related to papers on the impact of digitization-induced entry on welfare. Out of those, surprisingly few papers focus on the supply-side. For example, Aguiar and Waldfogel (2017) argue that with lower fixed costs, firms introduce more products, some of which turn out to be more successful than what the firm had predicted ex-ante. In addition to that, our paper explicitly acknowledges that firms may make better predictions due to additional available data. Moreover, because our analysis is at the pre-market licensing stage, we gain insights into how digitization affects the market's static efficiency as well as the incentives to innovate in the first place. In examining how inventors and firms interact in the market for ideas, this article follows a long strand of literature analyzing the optimal commercialization strategy of new products (see Gans and Stern, 2003). Closely related to our paper, Ellison (2011) argues that the role of scientific journals in disseminating research has declined with the internet, especially for high-profile authors. In the market for literary works, even if digital self-publishing facilitates the discovery of high-quality ideas, most of these ideas are eventually published through traditional channels regardless

of their original path. This is in line with Hegde and Luo (2014), who show that publication of patents through a credible, centralized institution mitigates information costs for buyers and sellers. Our results provide insights for many other settings in which powerful gatekeepers select ideas which eventually reach the market. Obvious other examples are music and movies (Luo, 2014), but our findings have implications for new products in any markets in which inventors license their products to downstream firms (Arora and Fosfuri, 2003; Katz and Shapiro, 1985). Most recently, traditional investors increasingly face competition from crowdfunding platforms concerning the financing of ideas and inventions (Agrawal et al., 2013, 2015). Determining the driving forces behind changes in the contracts between authors and publishers provides insights into innovators' incentives to create new products and their optimal commercialization strategies beyond book publishing.

2 Industry Background

2.1 Traditional Institutions in the Book Publishing Industry

Traditionally, authors could reach consumers only if they found a publisher who was willing to publish their book.[5] But because publishers have little incentive to publish books which they believe will not sell, many authors never reached consumers. If a match is found, a contract is set up in which the author licenses the book's rights to the publisher. The contract includes a lump-sum payment to the author – to be paid out before any copies are sold – as well as royalties for each copy that is sold beyond the advance payment. While royalty payments have 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).

[5] See Moldovanu and Tietzel (1998) for a historical perspective on book publishing in the late 18^{th} century.

2.2 Digital Technology in Production and Consumption

The increased use of digital technologies has significantly decreased the cost of producing and distributing products (Waldfogel and Reimers, 2015). The most recent such change was triggered by the introduction of e-reading devices, most notably the Amazon Kindle in November 2007, which made reading books in electronic format (e-books) a viable option.[6] The left-hand panel of Figure I shows the adoption of e-reading devices (e-readers and tablets) among US adults, according to a survey by the *Pew Research Center*. Ownership of designated e-readers increased steeply after 2008, with a decrease in 2015 presumably because e-readers are replaced by tablets.

Around the same time, the number of new book editions (ISBNs) per year has increased significantly, from 284,370 in 2007 to 703,378 in 2012.[7] Many of these new books and titles were published through a new distribution channel altogether: online self-publishing platforms allow authors to publish their books without a screening process, for a small fee but without major advertising efforts by the platform. Such self-publishing platforms are comparable to so-called Vanity Presses and Print-on-Demand services that allow any author to publish their physical books for a fee as well.[8] However, the full automation and digital distribution of the self-publishing platforms drive costs down far below what these more traditional outlets could offer, making self-publishing popular among both authors and readers.

The right-hand panel of Figure I depicts the number of books published on the self-publishing platform *Smashwords*.[9] The supply of new books has grown significantly since 2008, although the

[6] While reading books electronically has been possible for years before (for example, as scanned PDFs), the Kindle and the accompanying e-ink technology improved the reading experience enough to trigger a large shift in reading behavior.

[7] Data provided by Bowker. See `http://manuscritdepot.com/documentspdf/autoedition_usa_bowker.pdf` and `http://www.bowker.com/assets/downloads/products/isbn_output_2002_2013.pdf` (accessed March 31, 2016).

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

[9] We thank Dainis Zegners for sharing the data used in Zegners (2016).

increase is mostly driven by one particular genre. Before 2010, the number of Romance books was similar to the number of books in other genres (fantasy, children, religion, mystery, self-improvement, biography). In 2011, there were roughly twice as many romance books, and after 2011 the supply of romance novels on the self-publishing platform is roughly five times as large as that of the second-most represented genre.

Self-publishing allowed many new authors to reach consumers, and some of these self-published authors became largely successful. A "poster-child" of self-publishing, E.L. James' *Fifty Shades of Grey* was originally released as an e-book 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.[10] 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 and to inspire a major motion picture starring Matt Damon. Weir sold the rights to his next book on September 8, 2014, again to Crown Publishing, in a "major" deal (at least $500,000).[11]

Self-publishing has also served as an alternative for established authors to reach consumers (McCartney, 2016). Romance writer Jamie McGuire, for example, struck 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. In 2015 McGuire returned to self-publishing for another (successful) sequel. She "still plan[s] to traditionally publish, but with books that [she] feel[s] are best suited for that route" (McCartney, 2016). Yet other authors reject traditional deals outright. For instance, Vi Keeland and Penelope Ward rejected all offered deals from publishers, instead self-publishing their

[10] See `https://www.nytimes.com/2014/02/27/business/media/for-fifty-shades-of-grey-more-than-100-million-sold.html`. Interestingly, this article suggests that this title did not lead to an increase in sales for other romance novels – a fact that is supported in our data, and which we exploit in our analysis.

[11] Information on individual deals is taken from our dataset, which is described in detail in section 4.

novel, *Cocky Bastard*, which was listed on the New York Times bestseller list in 2015 (McCartney, 2016). Self-publishing can serve both as a stepping stone for aspiring authors and as a lucrative alternative for established authors. We explore these functions both theoretically and empirically in the following sections.

3 Theory

In what follows, we introduce a simple model of competition in the market for book ideas to provide intuition for the empirical analysis. We begin by assuming full information about an idea's appeal, and we then introduce uncertainty to show how self-publishing impacts the predictability of success.

3.1 Perfect Information

Consider a world with many authors who have ideas for books. Each idea has observable characteristics that determine a book's type m, drawn from some random distribution.[12] An idea's type reflects vertical differentiation, as well as the book's alignment with consumer taste. It is identified through variables such as the book's genre, its characters, its setting and tone, the length of the text, and author-specific characteristics (perhaps reflecting their identity). For simplicity, suppose the consumers' inverse demand for a book of type m is given by

$$p_m = a_m - q_m. \tag{1}$$

Traditional Publishing

Authors can sell licenses for the exclusive right to publish their book of type m for a lump-sum payment L_m.[13] After obtaining the license, publisher j sells the book in the product market with

[12] Relaxing the assumption of exogenous arrival of ideas complicates the analysis but strengthens our results. See the discussion below.

[13] Publishers typically offer an *advance* against royalties, rather than a lump-sum payment. Our theoretical model extends to such a world. The lump-sum payments are designed to increase the author's utility. This utility is monetary, but can also be derived from certainty, or from receiving money earlier (Beck, 2012).

demand as in equation (1), paying an exogenously given royalty rate $s \in (0,1)$ per dollar of revenue to the author.[14] Publisher j's profit without the lump-sum payment is

$$\pi_m^j = q_m^{j*}(a_m - q_m^{j*})(1-s) - c^j q_m^{j*} - F_m^j, \tag{2}$$

where c^j is the marginal cost of production and distribution, F_m^j is the publisher's fixed cost of publishing a book of type m, and $q_m^{j*} = \frac{a_m(1-s)-c^j}{2(1-s)}$ is the quantity that maximizes the publisher's profit π_m^j. The author's profit from selling a license to publisher j – without the lump sum – is

$$\pi_m^A = q_m^{j*}(a_m - q_m^{j*})s. \tag{3}$$

Self-publishing and the Digital Age

Digital technologies decrease fixed and marginal costs significantly, to F_m^D and c^D, allowing authors to use existing digital technologies to distribute their books directly. When using a digital self-publishing platform, the author's profit is

$$\pi_m^D = q_m^{D*}(\lambda_m a_m - q_m^{D*})s - c^D q_m^{D*} - F_m^D, \tag{4}$$

where $q_m^{D*} = \frac{\lambda_m a_m s - c^D}{2s}$ is the author's profit-maximizing quantity when publishing digitally, and $\lambda_m \in [0,1]$ reflects the possibility that the digital market is smaller than the physical market. Note that λ_m can vary across types and over time, for example as more e-reading devices are adopted or digital self-publishing platforms become more popular. For simplicity, we assume the royalty rate

[14] Following book industry standards, we let s be exogenous: traditional publishers offer relatively fixed royalties, while offers vary on marketing and the lump-sum payment (the advance) (Levine, 2016; Greco, 2013).

s is the same for the digital platform as for traditional publishers.[15]

License Fees

For a traditional publisher to obtain the right to publish a book idea, the lump-sum payment must satisfy two conditions: 1) the author must find traditional publishing more profitable than her best outside option; and 2) the lump sum must be smaller than the publisher's profit without the transfer. With a single traditional publisher, the first condition determines the size of the lump-sum payment. With multiple identical publishers and full information, competition erodes all publisher profits. In what follows, we consider a world with one traditional publisher, noting that our results hold with multiple, non-identical publishers, and/or with imperfect information.

Before digital self-publishing, the author's outside option is to not publish at all, for a profit of zero. A deal is made if the *joint* profits $\pi_m^j + \pi_m^A > 0$, and a lump-sum of $L_m = -\pi_m^A$ is likely.[16] With digital self-publishing, the author's best outside option may have changed. If the author's profit from publishing digitally is larger than zero ($\pi_m^D > 0$), then publisher j must increase the lump-sum payment L_m, compared to the monopoly case without self-publishing.

Proposition 1 *Lump-sum payments increase for some books due to digital self-publishing.*

Equation 4 implies that $\pi^D > 0$ when $\lambda_m > \frac{c^D + 2\sqrt{sF_m^D}}{a_m s}$. When π^D is positive, the traditional publisher has to increase her lump-sum payment in order to obtain the license for the author's idea.[17] Further, because $\frac{\partial \pi_m^D}{\partial a_m} > 0$, the pressure for the publisher to increase the lump-sum payment is larger for books with a higher appeal a_m.

[15] In reality, authors can retain a larger part of the revenue when self-publishing, with an s up to 1 (Levine, 2016). We assume an unchanged s to minimize the number of parameters, noting that a larger s for digital publishing would strengthen our results.

[16] This type of deal is common enough that industry jargon talks about author-subsidized books or vanity presses.

[17] Recall that λ_m (and therefore the author's profit from digital self-publishing) can change over time as a function of e-reader adoption and the popularity of digital self-publishing platforms.

Finally, the largest possible profit for the author when publishing traditionally – including the lump-sum fee – is

$$\begin{aligned} \pi_m^{A,\max} &= \pi_m^A(q_m^{j*}) + \pi_m^j(q_m^{j*}) \\ &= q_m^{j*}(a_m - q_m^{j*}) - c^j q_m^{j*} - F_m^j. \end{aligned} \tag{5}$$

If $\pi_m^{A,\max} < \pi_m^D$, the author will choose to self-publish even if she receives an offer to publish traditionally.[18]

Figure II illustrates conditions under which 1) lump-sum transfers remain unchanged, 2) lump-sum transfers increase due to self-publishing, and 3) authors choose self-publishing despite receiving an offer to publish traditionally. The figure shows the author's profit π_m^D when self-publishing digitally (dashed line), and her maximum profit $\pi_m^{A,\max}$ when publishing traditionally (solid line), as functions of the relative size of the digital market λ_m.[19] Given our parameters, the lump-sum payment remains unchanged for values of λ_m below 0.68 because $\pi_m^D < 0$. For values of λ_m between 0.68 and 0.81, the lump-sum payment increases but the book is still published traditionally. When $\lambda_m > 0.81$, $\pi_m^D > \pi_m^{A,\max}$ and the author chooses to publish digitally.

Note that digital publishing may be profitable for some authors even when traditional publishing is not, i.e. $\pi_m^D > 0 > \pi_m^{A,\max}$. Intuitively, some books do not have enough market potential to cover the relatively high fixed costs of traditional publishing, but are appealing enough to cover the relatively low fixed costs of digital self-publishing. Hence, our model further predicts the following relationship:

[18] This happens when $\lambda_m > \frac{a_m c^D + \sqrt{a^2 s\left(\frac{c^2(1-2s)}{(1-s)^2} + a^2 - 2ac + 4(F_m^D - F_m^j)\right)}}{a_m^2 s}$.

[19] We set parameters to $a_m = 10, s = 0.2, c^j = 5, c^D = 0.1, F_m^j = 5$, and $F_m^D = 2$.

Corollary 1.1 *More books appear on the market due to digital self-publishing.*

In the above numerical example, new books come to market if $F_m^j > 5.86 \approx 3 \times F_m^D$ and $\lambda_m > 0.68$. Of course, higher royalty rates when self-publishing and asymmetric information across authors and publishers about the book's ex-post appeal can further strengthen this result.

3.2 Imperfect Information

Now suppose neither the authors nor the publishing firms have perfect information about the market potential of a type-m idea, and instead each agent i (author, publisher) forms a belief about a_m, drawn from a normal distribution: $a_m^i \sim N(\mu_m^i, \sigma_m^i)$. Under uncertainty, the central results above hold with each agent's beliefs a_m^i substituted for the true a_m.[20]

A publisher learns about book type m's typical appeal – the true average a_m across all ideas within that type – from her previous entry decisions and from data generated by others. These data are incorporated in future market potential predictions of type-m ideas in a Bayesian updating process. Given an agent's prior distribution, $a_m^i \sim N(\mu_m^i, \sigma_m^i)$, we define the precision of the prior belief as $\tau_m^i = \frac{1}{\sigma_m^i}$, and denote the precision of the true distribution of a_m as r_m. Next, let there be n_m additional data points $X_{k,m} = x_{k,m}(k = 1, ..., n_m)$. Then the posterior distribution of a_m^i is also a normal distribution, with mean $\mu_m^{i'}$ and precision $\tau_m^i + n_m r_m$ (see DeGroot, 1970). In particular,

$$\mu_m^{i'} = \frac{n_m r_m}{\tau_m^i + n_m r_m} \overline{x_m} + \frac{\tau_m^i}{\tau_m^i + n_m r_m} \mu_m^i, \tag{6}$$

where $\overline{x_m}$ is the sample mean of the additional data. Thus, $\mu_m^{i'}$ is a weighted average of an estimate of a_m^i formed from data ($\overline{x_m}$), and an estimate of a_m^i formed from the prior distribution (μ_m^i).

From this relationship, we infer the following implication about the prediction error:

[20] We assume for simplicity that publishers and authors are risk neutral. If agents are risk averse, the "cutoff" expected profits will be lower under each publishing strategy.

Proposition 2 *Publishers predict the market potential of book ideas more precisely due to digital self-publishing.*

Corollary 1.1 shows that the number of new products increases due to digital self-publishing. As the number of observations n_m increases, the predicted market appeal will converge to the true market potential because $\lim_{n_m \to \infty} \frac{n_m r_m}{\tau_m^i + n_m r_m} = 1$ and $\lim_{n_m \to \infty} \frac{\tau_m^i}{\tau_m^i + n_m r_m} = 0$. Moreover, if the true market appeal a_m is normally distributed with a known variance, an accurate prediction of the mean appeal leads not only to more accurate predictions on average, but also to more accurate predictions for each individual idea within a type because the precision of the posterior distribution $\tau_m^i + n_m r_m$ increases as $n_m \to \infty$.

Finally, note that the speed of convergence to the true mean a_m can increase for two reasons. First, an increase in n_m over time leads to quicker convergence as described above. Second, if authors learn about a type's appeal, they may respond by writing more books of popular types, which in turn results in more titles of this type on the market (an increase in n_m). In the empirical analysis, we cannot distinguish between these two mechanisms, but we determine the overall effect.

4 Data and Identification

4.1 Data Sources

We collect data from a variety of sources to test the predictions from the model. First, we obtain information about book licensing deals from the industry database *Publishers Marketplace*. Second, we have weekly sales rankings published by the newspaper *USA Today*, which we combine with proprietary sales data from *Nielsen Bookscan* to determine an author's ex-post success. In addition, we obtain counts of published books from *Bowker* (US), *Börsenverein des deutschen Buchhandels* (Germany), and *Bibliothèque Nationale de France* (France). We describe the two main datasets

here, and we explain the remaining data in more detail when used.

4.1.1 License Deals

Data on license deals come from *Publishers Marketplace*, a professional online community for the book industry, in which literary agents post information about book-related deals and the involved entities. We observe all posted deals between January 1st 2002 and December 31st 2015 – a total of 100,772 deals. We extract names of authors and editors along with genres and types of deals (digital or print book deals, and rights for audio books, film, TV, and international distribution). Importantly, the database allows us to quantify the size of the lump-sum payments for a subset of about 25% of these deals across five categories: (1) less than \$50k (*"nice"*, 62% of all deals), (2) \$50k to \$99k (*"very nice"*, 9%), (3) \$100k to \$249k (*"good"*, 14%), (4) \$250 to \$499k (*"significant"*, 5%), and (5) more than \$500k (*"major"*, 10%).

After data cleaning, we observe 52,260 book deals and 39,584 rights deals.[21] From the posted information, we define six control variables that describe deal-specific characteristics. In the empirical estimation, these characteristics help control for author and book quality:

Acclaimed is a dummy set to 1 if the text includes the words *award*, *edgar*, *nominee*, *winner*, *finalist*, *pulitzer*, *NYT notable*, *acclaimed*, *syndicated* or *star*.

Bestseller indicates if the text includes the words *bestselling* or *bestseller*.

Contested is a dummy set to 1 if the text includes the words *at auction* or *preempt*.

Debut indicates if the text includes the words *debut*, *first-time* or *first novel*.

Self-published is a dummy set to 1 if the text includes the word *self-published*.

Sequel indicates if the text includes the words *sequel*, *prequel*, *next book* or *follow-up*.

Descriptive statistics of deal sizes and characteristics can be found in the top panel of Table A.1.

[21] For example, 8,928 deals include multiple books, making it difficult to compare payments across deals. We exclude these observations.

4.1.2 Success of Book Ideas

We link our categorical licensing information with categorical sales information from the USA Today Top 100 bestseller list (using weekly data from 2002 to 2016). Specifically, we use the dates of an author's appearances in the bestseller list to create a measure on the extensive margin: whether an author is sufficiently successful after we observe a deal with that author in the Publishers Marketplace dataset.[22] To clarify what "sufficiently successful" means in monetary terms, i.e. beyond the ordinal information of (not) reaching a Top 100/50/10 position in the bestseller list, we use sales information from the proprietary Nielsen Bookscan database. In particular, we estimate a title's lifetime revenues based on its ranking as described in appendix section A.2. The bottom panel of Table A.1 provides descriptive statistics regarding a deal's probability of reaching the Top 100 and Top 10 bestseller lists.

4.2 Identification

Empirical work on the impact of disintermediation on the market for ideas has been hampered not only by a lack of data in the market for ideas, but also by a lack of exogenous variation across products. We circumvent the latter issue with two strategies. First, some countries adopted e-readers and self-publishing earlier than others. Second, some genres are better suited for digital self-publishing than others.

4.2.1 Identification Across Countries

Digital self-publishing platforms are largely country-specific, and the country-specific roll-out of e-books and self-publishing platforms happened at different points in time. For example, although the Amazon Kindle was introduced in the United States in November 2007, the device was only sold on the American platform (amazon.com), and the first Kindle with the ability to download

[22] We create sales categories to allow a direct comparison with the dead size data.

content wirelessly outside the United States (the Kindle 2 International) was not introduced until October 2009.[23]

In 2011, e-books accounted for 13.6% of all fiction novel sales in the US, and Amazon had announced that is has sold more e-books than print books.[24] In contrast, e-books had a market share of 0.5% in Germany, and 1.8% in France in the same year (Wischenbart, 2012). While around 18% of the US population owned e-reading devices in 2012, hardware diffusion was much lower in Europe with only about 5% of the German and French population, respectively.[25] Finally, Amazon's self-publishing platform *Kindle Direct Publishing* started in Germany in April 2011, and was rolled out in additional countries in August 2013.[26] Accordingly, the Paris and Frankfurt Books Fairs did not recognize self-publishing in distinct exhibition areas until 2014 and 2015, respectively.[27] This is suggestive evidence that the digital disruption happened in the United States earlier and more intensely than in other countries.

4.2.2 Identification Across Genres

While a comparison across countries allows us to study the impact of disintermediation on the extensive margin (whether books were published at all), an analysis of the intensive margin (which book deals are made, and the size of these deals) requires more detailed data. We use quasi-experimental variation to identify observations that are more strongly affected by the introduction and adoption of digital self-publishing platforms than others. Note that the theoretical model equivalently distinguishes between book types with different values of λ_m, the relative size of the digital market.

[23] See `https://tinyurl.com/ybpdfmls`.
[24] See `http://www.nytimes.com/2011/05/20/technology/20amazon.html`.
[25] See `http://tinyurl.com/ycuux8jc` and `http://tinyurl.com/y9nvyjbh`.
[26] See `http://tinyurl.com/ycpspp6m` and `http://tinyurl.com/yd82wvst`.
[27] See `http://tinyurl.com/y98l5sy7` and `http://tinyurl.com/y82dox4s`.

To define a quasi-experimental treatment group, we conducted several interviews in the field, asking industry insiders for their opinions about the typical characteristics of digitally self-published books. The experts argued that romance and erotica novels – such as E. L. James' *Fifty Shades of Grey* and Nicholas Sparks' *The Notebook* – are especially well-suited for self-publishing, for several reasons.[28] First, romance books are reportedly relatively easy to write, because they do not tend to be research intensive. Second, romance novel readers are a close-knit group that communicates extensively in online communities, allowing readers to learn about new books via word-of-mouth, rather than through costly advertising campaigns often employed by traditional publishers.[29] Finally, the experts argued that the nature of many romance novels might make readers reluctant to read them in public. Self-publishing platforms circumvent this problem by predominantly publishing e-books, a format which does not show the book's cover when read.

Quantitative data also indicate that romance enjoys a special status. Meta data from the popular self-publishing platform *Smashwords* suggest that romance/erotica novels are by far the most frequently published type, representing 28% of the 431,307 books published between 2008 and 2016. In contrast, the share of the romance category among license deals with traditional publishers on *Publishers Marketplace* is 16%.

We further find that self-published romance books are more popular among consumers than self-published books of other genres. The left-hand panel of Figure III displays the share of originally self-published books among bestsellers in the *USA Today* weekly bestseller lists, each week from 2010 to 2014, distinguishing between self-published romance books and all other self-published books. Beginning in 2011, between 20% and 50% of bestsellers in the romance category had a self-

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

[29] For example, the most used tag on the review platform `goodreads.com` is "romance" (4,553 times). The second most used tag is "fiction" (3,984 times; numbers as of October 2, 2017).

publishing background. During the entire observed period, less than 5% of the bestsellers in other categories were originally self-published. Hence, the market potential for digitally self-published books (λ_m in our theoretical model) seems larger for romance novels than for other book types.

The rise in popularity of self-published romance books could coincide with romance books becoming more popular *per se*. Again, interviews with experts suggest that romance has always been a popular genre, and that immensely successful books (such as *Fifty Shades of Grey*) did not change the publishers' expectations regarding the profitability of romance novels in general. Again, we confirm this notion quantitatively. The right-hand panel of Figure III shows the total number of (print) books sold by genre according to a presentation by Nielsen at the 2014 Digital Book World.[30] It shows that print book sales decreased after 2008, and no less for romance novels than for other genres. Further information by the Romance Writers of America confirms that the share of romance book sales in the US remains constant, around 13% over the observed time period.

Qualitative and quantitative evidence supports the idea that the realized appeal of romance novels did not change despite the introduction of self-publishing, and it is unlikely that changes in the market for book ideas are driven by downstream consumer demand. The fact that traditionally published romance novels do not become more popular overall suggests that the mean market potential of ideas which are presented to publishers does not increase either.

5 Estimation and Results

Our empirical estimation aims to provide evidence on three levels: the number of books that may reach consumers (Corollary 1.1), the size of license deals for ideas that receive contracts (Proposition 1), and the predictability of success (Proposition 2). To examine the first effect, we take advantage of variation across countries, and for the latter two, we utilize variation across genres.

[30] See `https://tinyurl.com/y8yxu7st`.

5.1 The Number of Books

Although self-publishing platforms and e-reading devices emerged in the United States starting in 2008, digital self-publishing was de facto non-existent in non-English speaking countries until 2010. Accordingly, Figure IV shows the total number of new books (including re-editions) per country and year for the United States, Germany and France from 2002 to 2010. The number of new books remains relatively unchanged in France and Germany over that time period, but we see an exponential increase in the US after 2008 (note the logarithmic scale). In a formal analysis, we estimate a difference-in-differences model defined by the equation

$$Log(Books_{it}) = \alpha + \delta(After_t \times US_i) + \nu_t + \mu_i + \varepsilon_{it}, \tag{7}$$

where ν_t and μ_i are year- and country-fixed effects, respectively, and $After_t$ is 1 beginning in 2009 – just over one year after the introduction of the Kindle. The OLS estimate of δ is 1.044 with a standard error of 0.278 (p-value 0.000). The point estimate implies that digital self-publishing leads to almost a doubling of the number of books in the market ($exp(1.044) - 1 = 1.84$), compared to how the market might have evolved without it. This substantial increase in the number of books mechanically translates into more information which can be used to predict demand for new books.

5.2 The Size of License Payments

We continue with an analysis of the intensive margin: what happens to those license deals which are made? Based on the evidence reported in section 4.2, we identify romance writers as the treatment group in a difference-in-differences analysis. We estimate the impact of self-publishing on upfront license payments, and we later provide additional evidence that the identifying assumptions hold.

5.2.1 Baseline Estimation

Our baseline model for testing Proposition 1 estimates the following difference-in-differences model:

$$LogSize_{i,j,k,t} = \alpha + \beta R_j + \delta(After_t \times R_j) + \kappa C_j + \mu_t + \varepsilon_{i,j,k,t} \tag{8}$$

The unit of observation in this model is a license deal i between author j and editor k (at publisher p) on day t. R_j indicates whether the author ever published a book in the romance category ("romance author"). The definition on the author level helps us identify the overall impact of digital self-publishing, including within books, across books by the same author, and across books in the same genre.[31] For the same reason, we do not include author fixed effects. $After_t$ indicates whether the deal was closed after the year 2008, coinciding with the beginning of the wide-spread adoption of e-reading devices as shown in Figure I.

We account for time-specific variation by including month-year fixed effects μ_t, and we include a vector of the control variables $C_{j,t}$ introduced in section 4 to account for time-varying heterogeneity across authors. 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).

The main estimation results are reported in Table I. The dependent variable in column (1) is the logarithm of the license payment, measured as the midpoints of each deal category in the data.[32] The coefficient on the interaction term (After × Romance) is positive and statistically significant,

[31] The main results remain almost identical when alternatively categorizing the treatment group at the deal level.

[32] Recall the deal categories are "nice" (up to $50,000); "very nice" (between $50,000 and $99,000); " "good" ($100,000 to $249,000); "significant" ($250,000 to $499,000); and "major" (above $500,000). For "major" deals, we set the midpoint at $750,000, noting that results are robust.

suggesting that digital disintermediation increased license payments by 14.8%.

Columns (2) and (3) confirm that the results are robust to different specifications, including using the untransformed midpoints of the deal categories (in thousands, column 2) and categorical size variables – ordered from 1 to 5 (column 3).[33] The coefficient of interest suggests an average increase in license payments of 23.8% and 8.5%, respectively.[34] Throughout specifications, the coefficients of all control variables have the expected positive signs.

5.2.2 Timing

On average, license fees increase more for romance authors than for non-romance authors after 2008, but it is unclear whether these increases are immediate and lasting. We allow for a flexible time structure in the spirit of Autor (2003) to estimate the changes in deal sizes for romance authors compared to those for non-romance authors in each individual year:

$$LogSize_{i,j,k,t} = \alpha + \beta R_j + \sum_{\tau \in T} \delta_\tau \left(\gamma_\tau \times R_j\right) + \kappa C_j + \nu_k + \mu_t + \varepsilon_{i,j,k,t}, \tag{9}$$

where γ^τ denotes annual dummy variables. The omitted year is 2008, to facilitate a comparison of pre- and post-years.

Figure V plots the estimated year-specific difference coefficients (δ_τ). The coefficients are not statistically different from zero in any year before 2008, and they become large and significantly positive immediately after 2008, with a digitization-related increase in license payments of close to 20% in each year, compared to the years before 2008. Note that the specification provides further evidence that the identifying assumption holds: we find no differences in trends across treated and

[33] An ordered logit estimation provides nearly identical results (see section 5.2.3) but we report OLS results for ease of interpretation.

[34] Point estimates in column (1) are transformed to percentage values according to $PercentageChange = (\exp(\delta) - 1) \times 100\%$. In columns (2) and (3), we compare the coefficient to the sample mean, i.e. $(31.99/134.61) \times 100\% = 23.76\%$ in column (2) and $(0.167/1.975) \times 100\% = 8.46\%$ in column (3).

control groups before 2008.

5.2.3 Placebo Exercises

Although Figure III suggests that romance novels did not become more popular leading up to the introduction of self-publishing, there may still be relative changes in popularity that coincided with the introduction of e-readers and self-publishing platforms.

We test for this possibility by looking at deals for rights to existing books. If romance books became more popular after 2008, we would also see an increase in payments for international distribution rights (which typically involve translation and adaption for foreign markets) and for TV, film and audiobook licenses. However, because circumventing traditional publishers is substantially more difficult for these products (for example, there are no self-publishing platforms for international rights), traditional publishers face much less competition from digital platforms. Thus, to test whether unobserved factors specific to the romance genre but unrelated to self-publishing drive the results in Table I, we estimate variations of the model defined in equation (10) on rights deals, rather than book deals.

Table II shows the results from these regressions. The estimated coefficient of $After \times Romance$ in all three columns is negative, small in magnitude, and statistically insignificant. Accordingly, it is unlikely that license fees for romance novels rose because demand for them increased disproportionately. We examine (and confirm) the robustness of our results in further robustness checks – including an analysis of heterogeneous effects and an examination of the role of contracts between the publisher and retailer – in appendix section A.1.1.

5.3 Predicting Ex-Post Appeal

Above, we have provided evidence that license fees paid to authors in fact increase (Proposition 1). Here, we examine the ability to predict success by comparing an author's ex-post commercial

success to her ex-ante license deal, for romance authors and non-romance authors, before and after the arrival of digital self-publishing (Proposition 2).

In particular, we estimate the impact of digital self-publishing on both false positives (license fees reflective of more than the idea's ex-post value) and false negatives (license fees reflecting less than the idea's ex-post value) to account for the author's improved bargaining position. With a better outside option, the author can negotiate a deal that reflects the book idea's ex-post value more accurately, decreasing the rate of "false negatives" even if there is no additional information. On the other hand, a reduction in the rate of false positives is consistent with improved information, but not with changes in competition across publishers or in the author's bargaining power.

5.3.1 Measuring Prediction Precision

To determine how well a book's ex-post success matches the publisher's ex-ante prediction, we use categorical data from the USA Today bestseller lists. Further, to understand what reaching a top 100/50/10 position in the bestseller list typically means in terms of a book's overall success, we use information from Nielsen's Bookscan database to predict life-time sales for each book we observe in the USA Today bestseller list. The exact procedure is described in detail in appendix section A.2. We find that an average book reaching a top 10 spot in the bestseller list earns a revenue ($\widehat{Rev}$) of \$7.8 million throughout its life-time, an average title peaking between 11 and 50 in the rankings earns \$1.5 million, and an average title peaking between 51 and 100 earns \$464 thousand.[35]

Our theoretical model indicates that the lump-sum license payments may range between the author's outside option and the publisher's profit π^j, depending on relative bargaining powers. Publicly available information suggests that a publisher's per-title profit π^j (including e-book sales) is close to 50% of its revenue from physical books (see appendix section A.2). Assuming equal bar-

[35] Note that Nielsen's Bookscan database only covers sales of physical books.

gaining powers between publisher and author, we would thus expect the "correct" license payment L^* to be 0.25 times the ex-post revenue: $L^* = \theta\widehat{Rev}$, where $\theta = 0.25$. However, because we can neither determine the true profit margin nor the true relative bargaining powers, we repeat our analyses with values of θ ranging from 0 to 1.

Given θ, we define the prediction error as illustrated in Figure VI, comparing a deal's ex-post "correct" license fee L^* to its observed license fee ("deal size"). The *Error* is zero if the two are equal. If the observed license payment is larger, the error is positive, and if the payment is smaller, it is a false negative. The size of the error increases with the distance between the two measures. With five categories of deal sizes/ex-post profit, the *Error* ranges between -4 and 4.

5.3.2 Estimation and Results

We first repeat the analysis from equation (10) with different functions of the *Error* as dependent variables. We let the post-disintermediation period begin after 2010 because this is the first year in which the supply of new self-published romance novels at Smashwords was significantly larger than that of other genres (see the right panel of Figure III), allowing publishers to obtain disproportionate amounts of information about the demand for romance books.[36]

Table III shows the results from several specifications. Column (1) estimates the change in the absolute value of the *Error*. The coefficient on $After \times Romance$ is negative and statistically significant, providing strong empirical support for Proposition 2. The license fees paid to authors reflect an idea's commercial success 28.8% more accurately due to self-publishing.[37] Column (2) examines the extensive margin – whether publishers become less likely to make a mistake at all – in a linear probability model. The negative and significant coefficient indicates that the likelihood

[36] Note that this is consistent with our theoretical model. For the author's outside option to improve, it is sufficient that publishers *expect* that self-publishing will become more profitable in the future, whereas to learn from market data generated by self-published books, these already have to be on the market.

[37] Coefficient divided by sample average, 0.236/0.819=28.8%.

of making an error decreases by 9.1 percentage points, or 24.4% fewer errors at the mean.

Of course, the finding that the prediction error decreases can be explained by the increase in license fees estimated in the previous section. Without improvements in the information environment, this change would mechanically lead to *fewer* false negatives, and *more* false positives. We investigate these mechanisms in columns (3) and (4) of Table III. While the negative and significant impact on false negatives is expected, we also find a significant negative impact on false positives. At their mean values, the coefficients correspond to an 81.9% decrease of false negatives, and a 13.4% decrease of false positives. While the relatively large decrease in false negatives is likely due in large part to the increase in license payments, the fact that publishers less often pay too large a license fee suggests that publishers can indeed better predict an idea's commercial success.

5.3.3 Timing

To determine how quickly the additional products improve prediction, we estimate annual differences in the absolute value of the error term similar to equation (9). Figure VII plots the estimated year-specific difference coefficients (δ_τ). Most estimates before 2010 are statistically insignificant, supporting the identifying assumption of the difference-in-differences model. Estimates decrease significantly and persistently after 2011, with the largest drops in the most recent years. Prediction improvements increase as more romance novels enter the market.

5.3.4 Robustness to Assumptions

In the above analysis, we set the optimal license fee $L^* = 0.25\widehat{Rev}$, i.e. $\theta = 0.25$, based on publicly available data on costs and on assumptions about bargaining powers. The true L^* may lie anywhere between the publisher's profit (net of the license fee), and a value that makes the author indifferent between the deal and the outside option.

We examine our results' sensitivity to our assumptions by repeating the analyses from columns

(3) and (4) in Table III for values of θ ranging from 0.1 to 1.0, noting that the true θ is unlikely to rise above 0.5 due to the industry's cost structure. Figure VIII plots the coefficients on the interaction term ($After \times Romance$) against θ. The point estimates for both error types remain negative for all values, and they are statistically significant for all values that seem realistic given public information about costs and license fees.

These results suggest that the estimated improvements in prediction precision reported in Table III are not driven by specific assumptions. In addition, the fact that our results are robust to (almost) any assumed levels of competition and bargaining power provides further evidence that the decreased errors are at least partly due to improvements in the information environment. Finally, we present additional robustness checks in appendix section A.1.2.

6 Conclusion and Welfare Implications

Decreasing costs of production and distribution have made it easier for the creators of products – the inventors – to become entrepreneurs and bring their products to consumers without the help of intermediaries. Intuitively, this leads to changes in the terms of contracts between inventors and firms. How these contracts change, and how the incentives to innovate are affected, has been difficult to assess in the past, although potential welfare implications are large. We find evidence that digital self-publishing has increased overall welfare, with some important nuances.

First, some products that would not have been published by traditional institutions now become available. Our estimates suggest that the advent of digital self-publishing has almost doubled the number of books in the market, most of them at considerably lower prices.[38] Substantially greater product variety at lower prices suggests that the introduction of digital self-publishing and e-reading

[38] The average price of all 431,307 books published on *Smashwords* between 2008 and 2016 is \$3.29. In contrast, the average price of 457,404 e-books scraped from *Amazon* at four points in time in 2014 and 2015 is \$10.99, according to Author Earnings (see `http://www.authorearnings.com`). Even more, the average price of the 6,413 physical books we observe in our *Bookscan* sample, covering the weekly top 100 bestselling titles 2004–2012, is \$17.32.

devices can be welfare-enhancing for consumers.

Second, we find that authors receive significantly larger license fees. While this likely creates incentives for authors to produce more books, the overall welfare impact is not obvious. Larger license fees reflect a redistribution of income from publishers to authors without directly creating additional value.

Finally, we find evidence that publishers become better at predicting an idea's ex-post appeal, improving efficiency in the market for ideas. While authors who will be successful in the future earn higher license fees, license fees for other authors decline. As a result, publishers and "high-quality" authors are better off due to digital self-publishing, whereas "low-quality" authors incur a welfare loss. The reallocation of resources toward "better" ideas also benefits consumers, as more high-quality products become accessible to more consumers.

Just as importantly, the reallocation of resources has long-term implications for the market for ideas. Traditional institutions, which are better able to market products and reach more consumers, can continue to exist alongside new platforms that allow inventors to reach consumers directly. Our research shows that these institutions may complement each other.

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.

Agrawal, A., Catalini, C., and Goldfarb, A. (2015). "Crowdfunding: Geography, social networks, and the timing of investment decisions." *Journal of Economics & Management Strategy*, *24*(2), 253–274.

Agrawal, A. K., Catalini, C., and Goldfarb, A. (2013). "Some simple economics of crowdfunding." Tech. rep., National Bureau of Economic Research.

Aguiar, L., and Waldfogel, J. (2017). "Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music." *Journal of Political Economy*, *forthcoming*.

Arora, A., and Fosfuri, A. (2003). "Licensing the market for technology." *Journal of Economic Behavior & Organization*, *52*(2), 277–295.

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.

Baetschmann, G., Staub, K. E., and Winkelmann, R. (2015). "Consistent estimation of the fixed effects ordered logit model." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(3), 685–703.

Beck, J. (2012). "Advance contracting, word-of-mouth, and new-product success in creative industries: a quantification for books." *Journal of Media Economics*, *25*(2), 75–97.

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.

De los Santos, B., and Wildenbeest, M. R. (2017). "E-book pricing and vertical restraints." *Quantitative Marketing and Economics*, *15*(2), 85–122.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. Wiley.

Ellison, G. (2011). "Is peer review in decline?" *Economic Inquiry*, *49*(3), 635–657.

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

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

Hegde, D., and Luo, H. (2014). "Patent publication and the market for ideas." *Management Science*.

Katz, M. L., and Shapiro, C. (1985). "On the licensing of innovations." *The RAND Journal of Economics*, 504–520.

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.

Levine, R. (2011). *Free ride: how the Internet is destroying the culture business and how the culture business can fight back*. Random House.

Luo, H. (2014). "When to sell your idea: Theory and evidence from the movie industry." *Management Science*, *60*(12), 3067–3086.

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).

Moldovanu, B., and Tietzel, M. (1998). "Goethe's second-price auction." *Journal of Political Economy*, *106*(4), 854–859.

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.

Wischenbart, R. (2012). "The global ebook market: Current conditions and future projections 2011." *https://www.publishersweekly.com/binary-data/ARTICLE_ATTACHMENT/file/000/000/522-1.pdf*.

Zegners, D. (2016). "Voluntary disclosure of product information: The case of e-book samples." *Working Paper*.

Table I: Results: Changes in license deals

	(1) DV: Log(Size)		(2) DV: Size		(3) DV: Deal category	
Romance	-0.141**	(0.051)	-22.895**	(7.669)	-0.156**	(0.055)
After2008 × Romance	0.163***	(0.036)	31.990***	(8.202)	0.167***	(0.042)
Acclaimed	0.160***	(0.027)	27.112***	(4.486)	0.175***	(0.032)
Bestseller	0.992***	(0.084)	201.510***	(12.368)	1.158***	(0.089)
Contested	0.675***	(0.069)	117.900***	(13.810)	0.766***	(0.081)
Debut	0.043	(0.054)	15.997	(10.790)	0.063	(0.062)
Self-published	0.390*	(0.190)	92.313**	(33.069)	0.481*	(0.217)
Sequel	0.166***	(0.050)	26.972**	(11.614)	0.181**	(0.058)
Observations	14369		14771		14771	
$\overline{R^2}$	0.540		0.410		0.526	

Notes: Editor, month-year fixed effects, and constant not reported. The After period begins in 2009, the first year of reported e-reader ownership (see the left panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table II: Results: Changes in rights deals (placebo exercises)

	(1) DV: Log(Size)	(2) DV: Size	(3) DV: Category
After2008 × Romance	-0.061	-2.487	-0.062
	(0.109)	(19.223)	(0.123)
Observations	8194	8194	8194
$\overline{R^2}$	0.527	0.423	0.515

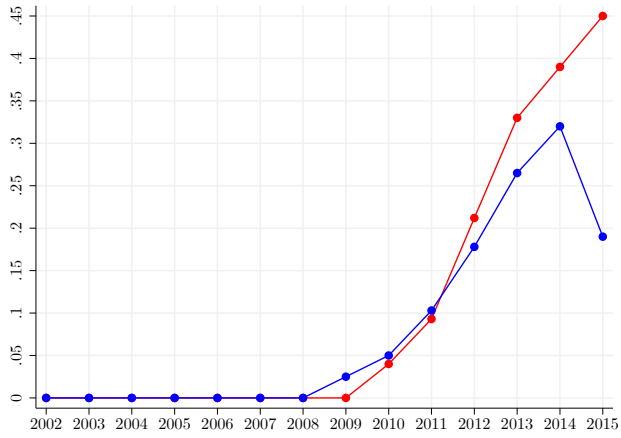
Notes: Editor and month-year fixed effects and coefficients of control variables not reported. The After period begins in 2009, the first year of reported e-reader ownership (see the left panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

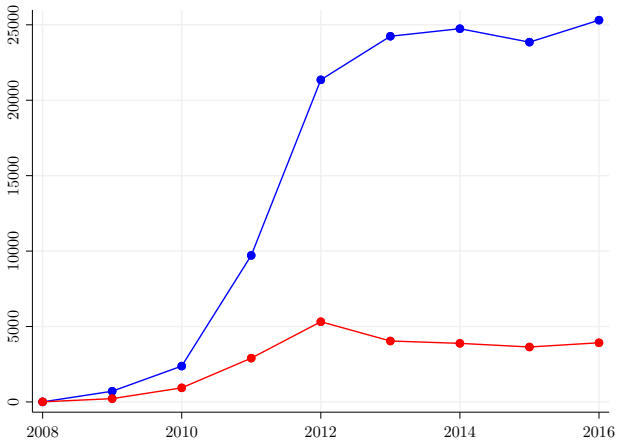
Table III: Results: Changes in predicting ex-post appeal

	(1) Abs(Error)	(2) I(Error)	(3) False Neg	(4) False Pos
After2010 × Romance	-0.236***	-0.091***	-0.046**	-0.045***
	(0.070)	(0.011)	(0.017)	(0.011)
Observations	14771	14771	14771	14771
$\overline{R^2}$	0.336	0.380	0.076	0.396

Notes: Editor and month-year fixed effects. Controls and constant included but not reported. The after period begins in 2011 based on availability of books on Smashwords (right panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Figure I: Adoption of e-reading devices and supply on digital self-publishing platforms in the US





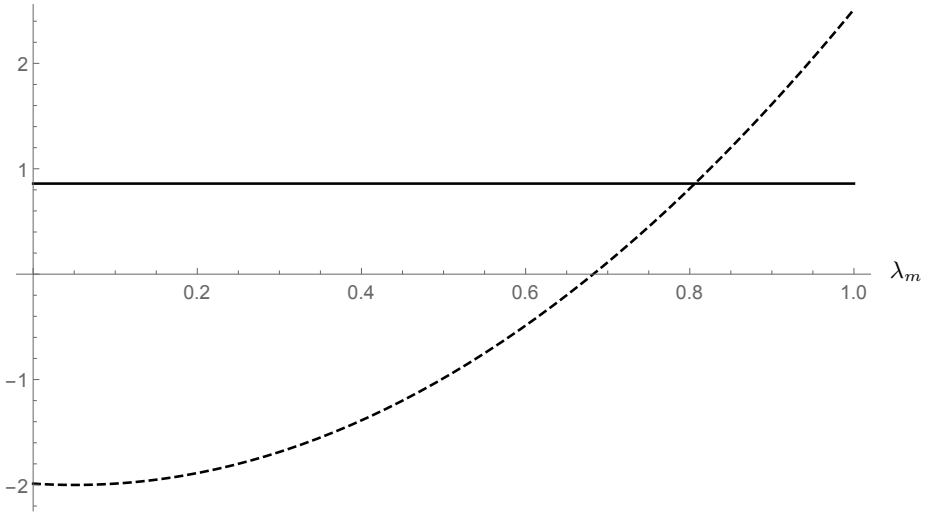
- Share of US adults owning an e-book reader
- Share of US adults owning a tablet computer

Source: Pew Research, `http://tinyurl.com/q2lt5ou`

- New romance titles on Smashwords per year
- New titles in other genres on Smashwords per year

Source: Smashwords

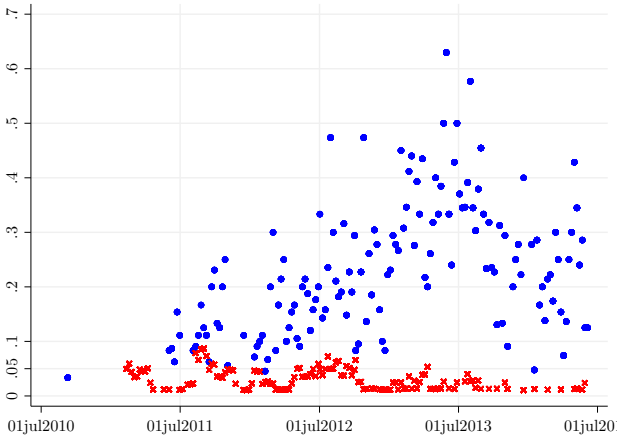
Figure II: Author profits under traditional publishing and digital self-publishing



— *Maximum author profit under traditional publishing* ($\pi_m^{A,max}$), - - - *Author profit under digital self-publishing* (π_m^D)

Notes: The x-axis shows λ_m – the relative size of the digital market. The vertical axis shows the author's respective profits. We use the following values: $a_m = 10, c^j = 5, c^D = 0.1, F^j = 5, F^D = 2$, and $s = 0.2$. When $\pi_m^D > 0$, the traditional publisher increases its lump-sum payment. When $\pi_m^D > \pi_m^{A,\max}$, the author chooses to self-publish.

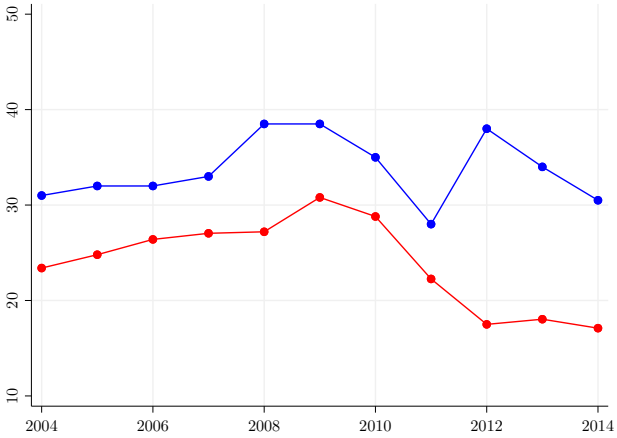
Figure III: Demand for romance books – self-published and traditionally published



Weekly share of originally self-published books in Top 100

• *Romance*, × *Non-Romance*

Source: Waldfogel and Reimers (2015)

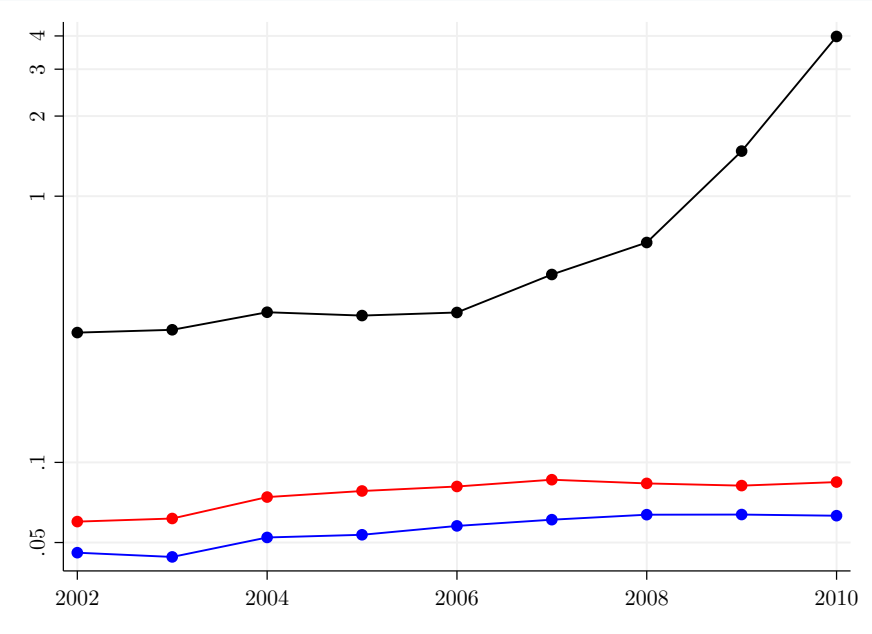


Total unit sales (in millions)

• *Romance*, • *Non-Romance*

Source: Nielsen

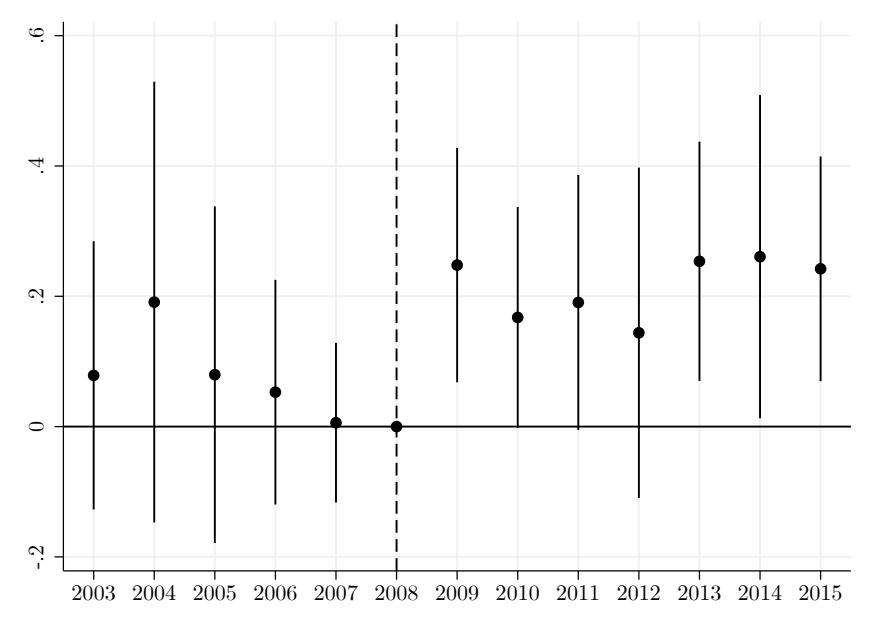
Figure IV: Number of new books per year in USA, Germany and France



• *USA,* • *Germany,* • *France*

Notes: Total number of new books per year (includes re-editions) in millions, vertical axis is in log-scale. Data source: Bowker ISBN counts (US), *Börsenverein des deutschen Buchhandels* (Germany), and *Bibliothèque Nationale de France* (France).

Figure V: License deals, group differences over time



Notes: OLS estimates of the δ_τ coefficients obtained from a regression of equation (9), i.e. yearly differences in *LogSize* between the treatment group (Romance authors) and the control group (non-Romance authors). The omitted year is 2008 based on e-reader ownership as illustrated in the left panel of Figure I. Standard errors are clustered on the genre-level, and bars indicate 90% confidence bands.

Figure VI: Deal size, ex-post profit and prediction error

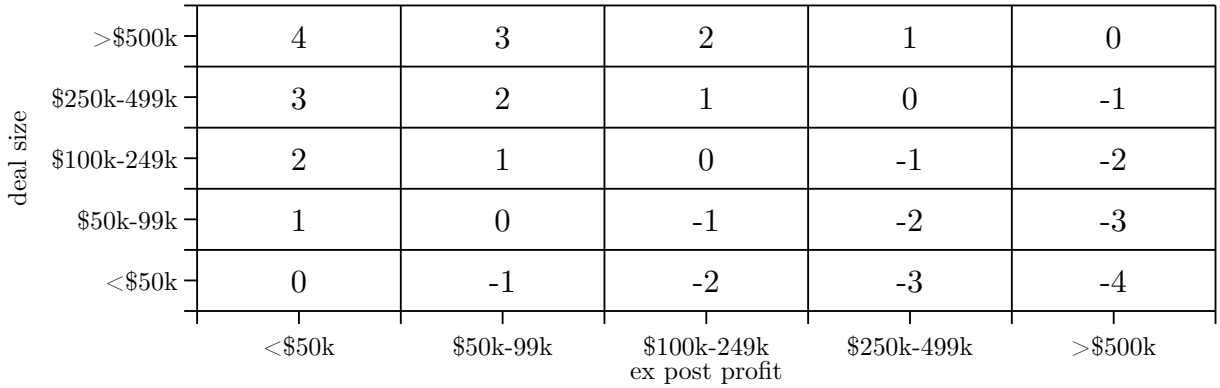
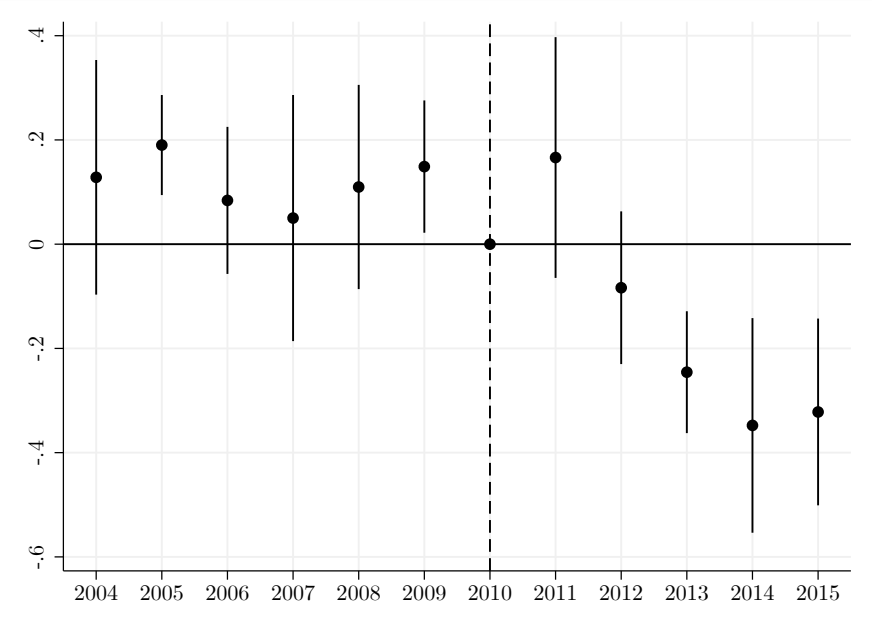
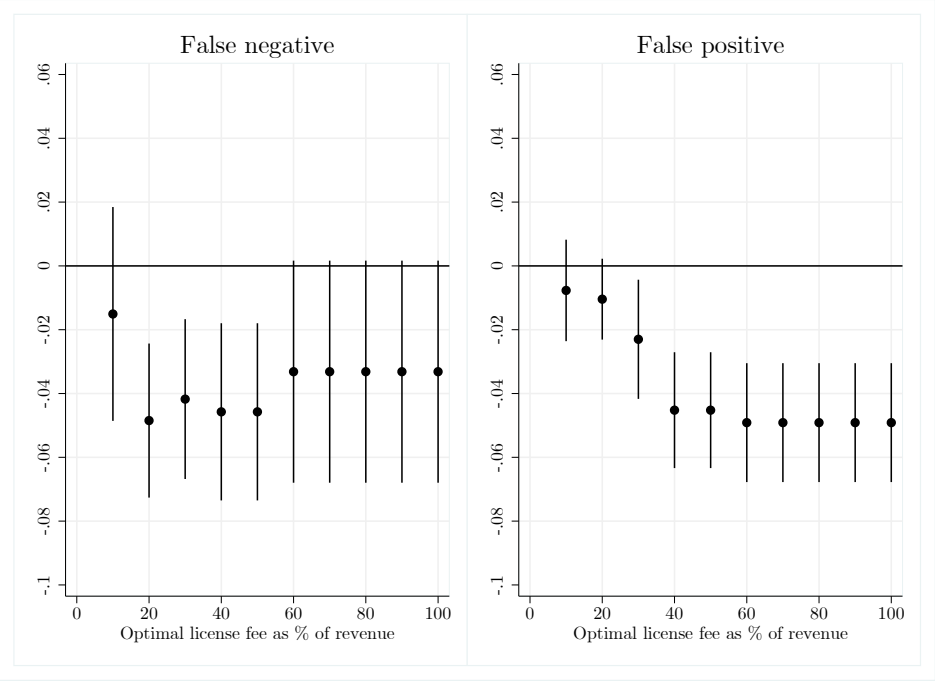


Figure VII: Prediction error, group differences over time



Notes: OLS estimates of the δ_τ coefficients obtained from a regression of equation (9) with the absolute value of the error as the dependent variable, i.e. yearly differences in $|Error|$ between the treatment group (Romance authors) and the control group (non-Romance authors). The omitted year is 2010 based on availability of books on Smashwords (right panel of Figure I). Standard errors are clustered on the genre-level, and bars indicate 90% confidence bands.

Figure VIII: Prediction error by profit margin



Notes: LPM estimates of the coefficients on $After \times Romance$ obtained from a regression of equation (10), using $1\{Error < 0\}$ and $1\{Error > 0\}$ as the dependent variables, respectively. The after period begins in 2011 based on availability of books on Smashwords (right panel of Figure I). Standard errors are clustered on the genre-level, and bars indicate 90% confidence bands.

A Technical Appendix

A.1 Additional Robustness Checks

A.1.1 License Payments

Heterogeneous Effects The theoretical model suggests that the increases in license payments are due to an increase in λ_m – the size of the digital self-publishing market relative to the traditional market – which determines the outside option for authors. The model further predicts that the outside option improves more for better ideas (larger a_m), and hence one might expect the increases to be larger when publishers expect a larger market appeal. Accordingly, we investigate whether the impacts of digitization are stronger for certain types of book deals. We interact each of our control variables ($a \in C$, the vector of control variables), with the difference-in-differences coefficient in a triple-differences analysis. That is, we estimate

$$\begin{aligned} LogSize_{i,j,k,t} = \alpha + \beta R_j + \delta_1(After_t \times R_j) + \kappa C_j + \delta_2(After_t \times C_j) \\ + \delta_3(After_t \times R_j \times C_j) + \mu_t + \varepsilon_{i,j,k,t}, \end{aligned} \tag{10}$$

where all variables are as described in the main text.

The results in Table A.2 show the coefficients of the triple interactions from this regression. While most coefficients are not statistically significant, their point estimates suggest that the impact might be largest for well-known authors (bestselling and acclaimed authors, sequels), as well as for those who are not yet well-established (debut). Interestingly, licenses for contested books do not seem to increase. This is consistent with our model, which assumes that publishers have market power, whereas they do not have market power for contested ideas by definition.

Agency and Wholesale Pricing Models The main text focuses on the relationship between publishers and authors, assuming implicitly that contracts on the retail level remain unchanged. However, the time period of our study includes changes in the contracts between publishers and retailers. In 2010, five of the six largest American publishers moved from a wholesale model (in which the retailers set book prices) to an agency model (in which the publisher directly sets the

book price), a change which was reversed in 2012.[39] De los Santos and Wildenbeest (2017) show that book prices are significantly larger under the agency model (from 2010 to 2012) than under the wholesale model. This may have impacted license deals as well, and if romance novels are disproportionately represented by major publishers, such changes may have been mistaken for impacts of digital self-publishing.

To test for this possibility, we create an indicator variable that is 1 if the deal is made with a major publisher, and we interact it with the time fixed effects in equation (10), testing both Propositions 1 and 2 again. Tables A.3 and A.4 show the results from these specifications. All results are almost unchanged, if not slightly stronger than in the main specifications. This suggests that the changes in license deals and predictability of success are not driven by changes in the retail environment.

Additional Robustness Checks Finally, our results that self-publishing increases the license payments may be driven by the peculiarities of our data. For example, we observe deal sizes for only about 25% of all deals, and even then, we only observe them in (arbitrary) categories. Here, we examine the possibility of non-response bias and the dependence of our results on the chosen transformations of the size categories.

To investigate potential non-reporting issues, we estimate the probability that the deal size is reported at all in a linear probability model. Note that our identification strategy would only fail if there was a systematic difference in reporting *trends* between romance deals and those in other genres. The small and insignificant coefficient on $After \times Romance$ in column (1) of Table A.5 shows no evidence of such issues.[40]

In addition, columns (2)–(6) of Table A.5 show results of linear probability models using indicator variables for the respective size categories. The comparison in each column is against all other categories, e.g. $<50k$ vs. $\geq 50k$ and *Size not reported*. The increases in license fees are mainly driven by increases in the probability of *major* deals, i.e. deals with volumes $> 500k$, but also by an increase in *good* deals (100–249k) and a decrease in *nice* deals (<50k).

[39] These publishers are Harper Collins, Hachette, Simon & Schuster, Penguin, and MacMillan. Random House adopted the agency model a little later, in early 2011.

[40] We also test whether the necessary condition for the identifying assumption (common pre-trends) is satisfied in this context. In results not reported but available upon request, we find no statistically significant differences between deals of romance and non-romance authors before 2008, providing additional support for our identification strategy.

Finally, Table A.6, reports results from an ordered logit model, which allows for flexible estimation of thresholds between categories.[41] The results from this exercise are similar to our main results as well, showing significant increases in the baseline probability of *significant* (250k–499k) and *major* (>500k) deals.

A.1.2 Prediction Precision

Similarly, our finding that additional entry improves the publishers' ability to predict an idea's ex-post appeal is robust to several different specifications. We explain these in more detail here.

We first estimate an ordered logit model to allow for the possibility that errors of different sizes are impacted differently. Column (1) of Table A.7 reports individual $After \times Romance$ coefficients for each value of the *Error* variable as described in Figure VI. The results suggest that the advent of digital self-publishing has reduced smaller errors more than larger errors, although the coefficients are not significantly different from one another.

Finally, we test whether our results are driven by selective reporting by interacting $After \times Romance$ with an indicator variable which is 1 if the deal size is reported in a regression which estimates the whether the author later appeared in the USA Today Top 100 bestseller lists. Table A.8 shows that this triple interaction is not significantly different from zero, suggesting that our results are not driven by selective reporting.

A.2 Defining Success

Section 5.3 utilizes sales information from two sources to determine how well ex-ante license payments match ex-post market appeal: weekly sales data from Nielsen's Bookscan database (we observe 462 weeks from 2004 to 2012), and the weekly USA Today bestseller lists (we observe these from 2002 to 2016).[42] Here, we describe how we determine unit sales, revenues and profits.

We first use the weekly Top 100 bestseller lists from Nielsen's Bookscan database to estimate each bestseller's cumulative revenue (unit sales times suggested retail price) as a function of its life-time observed ranking positions (between 1 and 100), adding a linear time trend to allow for

[41] The existing econometric theory regarding ordered logit/probit models with fixed effects allows for individual-specific fixed effects, i.e. on the deal level, but has not considered multi-level fixed effects. See for example (Baetschmann et al., 2015). We therefore include group-specific trends instead of year-month and editor fixed effects.

[42] The two datasets are very highly correlated as they are largely based on sales information from the same sources, although the USA Today bestseller lists include e-book sales. We account for these e-book sales in our analysis.

out-of-sample prediction before 2004 and beyond 2012. Formally, we estimate

$$Rev_i = \alpha + \sum_{r=1}^{100} \beta_r WeeksAtRank\_r_i + \theta y_t + \varepsilon_{i,t}.$$

The regression provides reasonable estimates for the bestsellers' cumulative sales and revenues.[43] An average Top 10 bestseller earns a revenue of $7.8 million throughout its life-time, an average title peaking between 11 and 50 in the rankings earns $1.5 million, and an average title peaking between 51 and 100 earns $464 thousand. With average prices of $17 per book, these revenues correspond to life-time unit sales of 460k, 88k, and 27k, respectively.

We use the estimated parameters of this model to estimate the ex-post appeal for each observed license deal, based on the ranking information in the USA Today data. We thus create a dataset which allows us to compare ex-ante license deals to ex-post revenue of the same author after the license deal, for all license deals from 2002 to 2015.[44]

Next, we map the estimated revenues ($\widehat{Rev}$) into per-title profits (π^j).[45] Determining the true size of π^j requires some detective work. First, at least among physical books, publishers and retailers typically use the wholesale model, in which the publisher sells books to the retailer for about 50% of the suggested retail price.[46] The marginal cost of production is roughly estimated at another $3 for hardcover and paperback books, and the cost of possible returns is reported to be around $1 (Levine, 2011, p. 167). Given the average prices, a physical book's profit margin is therefore close to 30%, so that $\hat{\pi}^j \approx 0.3 \times \widehat{Rev}$. Adding sales through channels not captured by Nielsen Bookscan (e.g. e-book sales) raises the true $\hat{\pi}^j$ to $0.5 \times \widehat{Rev}$ or even higher. Assuming equal bargaining powers between publisher and author, we would therefore expect the "correct" license payment (L^*) to be close to 0.25 times the ex-post revenue: $L^* = 0.25\widehat{Rev}$.

[43] The regression provides 102 coefficients. The corresponding results are not reported but available upon request.

[44] We cannot match ex-ante and ex-post appeal on the individual book level. We instead use the author's first appearance in the top 100 list after the book-specific deal. A manual check of a random sample, using information on the book's plot and publisher, suggests that the corresponding books very likely match across the datasets.

[45] The model predicts that the lump-sum license payments may lie between the author's outside option and the publisher's profit π^j, depending on relative bargaining powers. We therefore repeat our analysis for assumed optimal license payments ranging from zero to $\widehat{Rev}$ in the paper.

[46] In recent years, publishers have switched between agency and wholesale models (see, for example, De los Santos and Wildenbeest, 2017). We abstract away from these changes because they are not central to our arguments. In appendix section A.1.1, we show that our results are robust to a specification that explicitly takes those changes in the retail market into account.

Table A.1: Descriptive statistics

	Romance (N=2,164)				Non-Romance (N=12,607)				Total (N=14,771)	
	Before		After		Before		After		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Log(Deal Size)	10.761	1.075	10.832	1.185	10.977	1.184	11.003	1.179	10.961	1.178
Deal Size	106.598	187.508	130.174	223.602	137.036	214.069	137.072	209.120	134.610	211.715
Deal Size Categories	1.683	1.205	1.776	1.351	1.919	1.336	1.943	1.326	1.901	1.328
Acclaimed	0.081	0.273	0.072	0.258	0.087	0.282	0.109	0.312	0.093	0.291
Bestseller	0.098	0.297	0.186	0.389	0.033	0.179	0.065	0.247	0.062	0.241
Contested	0.020	0.140	0.012	0.108	0.045	0.207	0.068	0.252	0.049	0.216
Debut	0.055	0.227	0.018	0.134	0.036	0.185	0.026	0.159	0.032	0.176
Orig. Self-published	0.001	0.033	0.015	0.122	0.005	0.068	0.004	0.067	0.005	0.072
Sequel	0.032	0.177	0.044	0.206	0.023	0.149	0.027	0.163	0.027	0.162
Error^2 (All deals)	2.673	5.081	2.393	5.078	1.370	3.299	1.437	3.490	1.561	3.700
Not in Top 100	0.746	0.435	0.761	0.426	0.928	0.258	0.903	0.296	0.894	0.308
In Top 100	0.254	0.435	0.239	0.426	0.072	0.258	0.097	0.296	0.106	0.308
In Top 10	0.050	0.218	0.057	0.232	0.016	0.127	0.022	0.146	0.024	0.153

Table A.2: Results: Changes in license deals, interacted

	(1)	
After2008 × Romance	0.046	(0.054)
After2008 × Romance × Acclaimed	0.059	(0.142)
After2008 × Romance × Bestseller	0.349	(0.332)
After2008 × Romance × Contested	-0.465*	(0.235)
After2008 × Romance × Debut	0.419	(0.277)
After2008 × Romance × Self-published	-0.294	(0.357)
After2008 × Romance × Sequel	0.154	(0.192)
Observations	14771	
$\overline{R^2}$	0.542	

Dependent variable: Log(Dealsize+1).

Notes: Editor and month-year fixed effects. Lower-level interactions and constant included but not reported. The After period begins in 2009, the first year of reported e-reader ownership (left panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.3: Results: Changes in license deals, controlling for agency model changes

	(1) DV: Log(Size)		(2) DV: Size		(3) DV: Deal category	
Romance	-0.145***	(0.043)	-22.563**	(7.763)	-0.149**	(0.052)
After2008 × Romance	0.167***	(0.030)	32.991***	(7.434)	0.170***	(0.037)
Major publisher	1.357	(0.755)	162.787**	(66.160)	0.957**	(0.359)
Acclaimed	0.141***	(0.029)	26.569***	(4.221)	0.170***	(0.033)
Bestseller	0.976***	(0.095)	200.914***	(13.083)	1.155***	(0.090)
Contested	0.646***	(0.067)	117.116***	(13.483)	0.759***	(0.081)
Debut	0.032	(0.049)	16.038	(10.545)	0.061	(0.061)
Orig. Self-published	0.445*	(0.235)	89.843**	(31.938)	0.481**	(0.208)
Sequel	0.177**	(0.054)	26.663*	(12.568)	0.177**	(0.064)
Observations	12188		14771		14771	
$\overline{R^2}$	0.552		0.413		0.530	

Notes: Editor, month-year and major-month-year fixed effects are included. The After period begins in 2009, the first year of reported e-reader ownership (left panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.4: Results: Changes in predicting ex-post appeal, controlling for agency model changes

	(1) Abs(Error)	(2) I(Error)	(3) False Neg	(4) False Pos
After2010 × Romance	-0.228**	-0.094***	-0.043**	-0.051***
	(0.076)	(0.012)	(0.018)	(0.011)
Observations	14771	14771	14771	14771
$\overline{R^2}$	0.338	0.384	0.076	0.398

Notes: Editor, month-year and major-month-year fixed effects are included. Controls and constant included but not reported. The after period begins in 2011 based on availability of books on Smashwords (right panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.5: Results: Changes in license deals, individual deal categories

	(1) Size reported	(2) < 50	(3) 50–99	(4) 100–249	(5) 250–500	(6) > 500
After2008 × Romance	0.023	-0.039	0.003	0.019**	0.005	0.034***
	(0.021)	(0.023)	(0.006)	(0.006)	(0.004)	(0.007)
Observations	52259	52259	52259	52259	52259	52259
$\overline{R^2}$	0.208	0.288	0.027	0.030	0.011	0.075

Dependent variable: Column (1): dealsize reported (0/1), columns (2)–(6): deal size category (0/1).
Notes: Editor and month-year fixed effects not reported. Control variables and constant included but not reported. The After period begins in 2009, the first year of reported e-reader ownership (left panel of Figure I). Standard errors in parentheses, clustered on the genre-level.

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

Table A.6: Results: Changes in license deals, ordered logit

	(1) DV: Deal category	
Deal_50_99		
After2008 × Romance	-0.050	(0.375)
Deal_100_249		
After2008 × Romance	0.390	(0.335)
Deal_250_500		
After2008 × Romance	0.909*	(0.511)
Deal_____500		
After2008 × Romance	0.768*	(0.397)
Observations	14771	

Notes: Group-specific time trends, month-year fixed effects, control variables, and constant not reported. The After period begins in 2009, the first year of reported e-reader ownership (left panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.7: Results: Changes in predicting ex-post appeal, ordered logit

	(1) Error	
-4		
Romance	1.864***	(0.010)
After2010 × Romance	-0.463***	(0.113)
-3		
Romance	0.176**	(0.082)
After2010 × Romance	0.337***	(0.030)
-2		
Romance	1.649***	(0.067)
After2010 × Romance	-1.453***	(0.005)
-1		
Romance	0.762***	(0.056)
After2010 × Romance	-0.492**	(0.239)
1		
Romance	-0.575***	(0.021)
After2010 × Romance	-0.314***	(0.013)
2		
Romance	-0.978***	(0.023)
After2010 × Romance	-0.586***	(0.002)
3		
Romance	-0.999***	(0.007)
After2010 × Romance	-0.848***	(0.032)
4		
Romance	-1.418***	(0.004)
After2010 × Romance	-1.073***	(0.030)
Observations	14771	
$\overline{R^2}$		

Dependent variable: Absolute value of the error.

Notes: Comparison group are observations where *Error* is zero, i.e. where the deal size and future sales match (according to our estimate of future sales, see section 5.3). The after period begins in 2011 based on availability of books on Smashwords (right panel of Figure I). Editor and month-year fixed effects. Control variables and constant included but not reported. Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table A.8: Results: Changes in predicting ex-post appeal, robustness

	(1) In Top 100	
After2010 × Romance	0.025***	(0.007)
After2010 × Romance × Deal Size Reported	-0.002	(0.029)
Observations	52260	
$\overline{R^2}$	0.201	

Notes: Lower-level interactions and constant included but not reported. Editor and month-year fixed effects. The after period begins in 2011 based on availability of books on Smashwords (right panel of Figure I). Standard errors in parentheses, clustered on the genre-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$