Digital Disintermediation and the Market for Ideas*

Christian Peukert[1] and Imke Reimers[2]

[1]UCP – Católica Lisbon School of Business and Economics, christian.peukert@ucp.pt

[2]Northeastern University, i.reimers@northeastern.edu

December 13, 2017

Preliminary Version

Abstract

Digitization has allowed inventors to circumvent traditional intermediaries and directly reach consumers, which may affect licensing outcomes and efficiency in the market for ideas. We study these impacts theoretically, and empirically in the book publishing industry, where the number of new books available to consumers has almost doubled due to the advent of digital self publishing platforms. Using data on 90,000 license deals between authors and publishers from 2002 to 2015, we identify disintermediation-related changes in this market from quasi-experimental variation across product types over time. Consistent with digital self-publishing improving the author's bargaining position, we find that authors get substantially more favorable license deals. We further show that ex-ante license fees reflect ex-post demand more precisely. This is consistent with the idea that additional entry due to self-publishing creates more information about the realized appeal of specific product types. In markets in which product appeal is difficult to predict, such improvements in the information environment can have a large impact on efficiency.

*We thank James Dana, Markus Reisinger, Michael Ribers, and Joel Waldfogel for valuable comments. The paper has benefitted 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, and seminars at Northeastern, Dartmouth, LMU Munich, NBER (Productivity Lunch), and the University of Zurich.

1 Introduction

Digitization has substantially decreased production and distribution costs in many industries. This has led to an upsurge of new products and an increase in the variety available to consumers (Waldfogel, 2012; Aguiar and Waldfogel, 2015; Waldfogel and Reimers, 2015; Aguiar and Waldfogel, 2017), which can have substantial welfare enhancing effects (Brynjolfsson et al., 2003). The impact of digitization on the market for ideas – the relationship between the creators of the ideas and intermediaries between creators and consumers – is not nearly as well understood.

Thanks to the emergence of online platforms and distribution services, inventors can circumvent traditional intermediaries and directly market their products to consumers. As a first-order effect, this leads to more products on the market, but it can also indirectly improve the inventor's bargaining position when a contract with a traditional intermediary is set up. At the same time, advances in data collection, storage, and analytics allow firms to observe the ex-post appeal of not only their own products, but also products of competitors, including those which reach consumers only because of lower entry costs (in digital markets). Intermediaries can use such data to predict the commercial success of a product 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 paper examines and quantifies the impact of digital disintermediation on the market for licensing contracts between inventors and firms, using the empirical setting of the book publishing industry. Because large investments are needed to produce 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 fixed costs of entry. If the expected demand for an idea is too low, it is rejected. If the expected demand is "large enough," publishers and authors agree on the terms of an exclusive license to market the book, typically including an upfront license payment that is proportional to the expected ex-post appeal, and per-unit royalty fees. Ex-ante expectations may not always correctly reflect ex-post appeal. While many rejected book ideas are unlikely to turn into bestsellers, 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 important inefficiencies, causing under-investment in products which would have created utility in the long run.

The arrival of digital self-publishing and the increasing diffusion of dedicated e-reading devices has brought about a new channel with which authors can reach consumers. Instead of relying on traditional publishing houses to recognize (bet on) a book's appeal, authors can now publish their works more directly. Digital self-publishing platforms often require a small fee for making the book available, and they take a share of the revenue from each book sold, although authors receive a much larger share of the revenues than they would in the traditional system.[1] At the same time, traditional publishers are better able to market the book, and since self-publishing is often limited to electronic books (which require an e-reader), traditional publishers, still selling mostly physical books which do not require any additional devices, can reach more consumers.[2]

Our paper introduces a simple theoretical model of competition and information in the market for ideas. A key insight from the model is that the option to self-publish directly improves the author's outside option and increases the level of competition among publishers, while it indirectly allows firms and authors to learn about an idea's ex-ante appeal through related products. The increased level of competition causes an increase in the publisher's payments to some authors, even if the quality of the idea remains unchanged. Yet other titles enter the market through digital self-publishing, even if traditional publishers would not have picked these ideas. Of course, there is information in the realized demand for these titles, which can be utilized by publishers and authors. The indirect effect of improved information could make both publishers and consumers better off: publishers are less likely to lose money on book deals if they accurately predict the book's success, and more "good" books enter the market, with traditional publishers allocating the most resources to those ideas which are worth the most.

[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, but no advance 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/`.

We test these predictions empirically. Empirical research 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 these ideas, which are usually hard to come by. Second, causal inference further demands plausibly exogenous variation in entry costs. Our empirical setting allows us to deal with both issues. E-books and self-publishing arrived fairly suddenly (driven by the large-scale diffusion of Amazon's Kindle), and this arrival affected different genres differently. While books of most genres are still predominantly published through traditional publishers, authors and readers of romance novels have largely embraced self-publishing platforms.[3]

We examine 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 book deals in a difference-in-differences model. We find that authors of romance novels experience a mean increase of 20% in license fees after the arrival of self-publishing than books of other genres, and we provide evidence that these increases are due to added competition on the publisher level, rather than supply- or demand-driven. We then analyze the publishers' ability to predict an idea's success by examining the correlation between advance payments and commercial success before and after the emergence of digital self-publishing. We find that publishers have become better at predicting the commercial success of authors of romance novels, compared to authors in other genres and compared to the pre-digitized era. We conclude that digital self-publishing can indeed improve the efficiency of the market by allowing firms to invest more money in the best ideas, while avoiding investment in ideas which will not be successful.

Our research is closely related to papers on the impact of digitization-induced entry on welfare, although most previous papers limit their attention to the consumer side. For example, Aguiar and Waldfogel (2017) show that lower fixed costs of production have not only increased the number of available music titles on the market, but also the average quality of the available titles. The rationale for such finding is clear: with lower fixed costs, firms introduce more songs, some of which turn out to be more successful than what the firm had predicted ex-ante. In addition to what Aguiar and Waldfogel (2017) do, our approach allows firms to make better predictions because of

[3]See Section 2 for evidence of as well as reasons for this variation.

more available data, and we explicitly take both false negatives – missed bestsellers – and false positives – flops – into consideration. We can therefore evaluate how digitization affects the static efficiency of the market for ideas. Moreover, because our analysis is at the pre-market licensing stage, we can gain insights into how digitization affects the dynamic efficiency of a market: the incentives to innovate in the first place.

In examining how inventors and firms interact in the market for ideas, our paper follows a long strand of literature analyzing the optimal commercialization strategy of new products (see Gans and Stern, 2003). 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. Hegde and Luo (2014) show that publication of patents through a credible, centralized institution mitigates information costs for buyers and sellers. Similarly, we argue that publication through traditional publishing institutions facilitates transaction in the products market.

Our results provide insights for many other settings in which powerful gatekeepers select those ideas which eventually reach the market as new products. Obvious other examples are music and movies (Luo, 2014), but our findings have implications for new ideas and products in any markets in which innovators (or inventors) license their products to downstream firms (Arora and Fosfuri, 2003). Most recently, traditional investors, who can be compared to traditional publishers, 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 the book, often with the help of an agent.[4] But since publishers have little incentive to publish books which will not sell, many authors cannot publish their books. If a match is found, a contract is set up in which the author licenses the book's rights to the publisher. The contract

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

includes a lump-sum payment to the author – to be paid out in advance of any copies being sold – as well as royalties for each copy that is sold beyond the advance payment. While royalty payments tend to be 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 predicted success of the book in the product market (see Greco, 2013).

This process of discovery and distribution can be highly inefficient if publishers do not accurately predict the success of an idea. While many rejected book ideas were likely inconsequential, some high-quality ideas may have been "falsely" rejected and never reached consumers despite a large market potential. At the same time, some published books did not sell as well as expected, losing the publishers substantial amounts of money. In addition to the obvious direct effects of some ideas never reaching consumers, such inefficiencies can lead to under-investment in types of ideas which could create large consumer surplus in the long run.

2.2 Digital Technology on the Production and Consumption Side

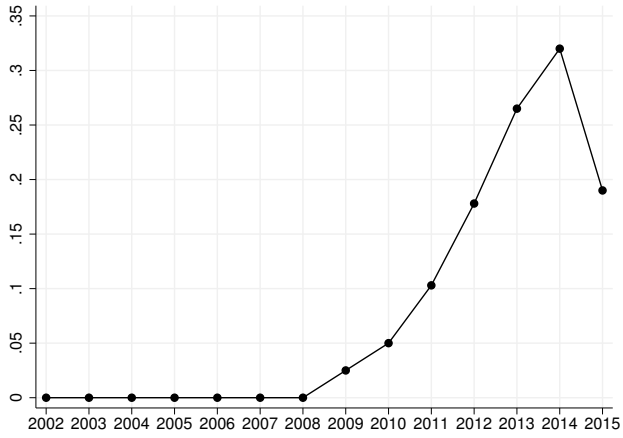
The increased use of digital technologies has decreased the cost of producing and distributing products significantly, leading to a large influx of new products in the market. The most recent such change was triggered by the introduction of e-reading devices, most notably the *Amazon Kindle* in November 2007, which made reading books in electronic format (e-books) a viable option.[5] The left-hand panel of Figure 1 shows the adoption of e-reading devices among US adults, according to a survey by the *Pew Research Center*. There is a steep increase in ownership after 2008, although ownership decreased in 2015 (presumably because e-readers are replaced by tablets).

At 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.[6] 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 any screening process, but also without an advance payment or major advertising efforts by the platform. In fact, authors oftentimes pay a nominal fee to the platform to

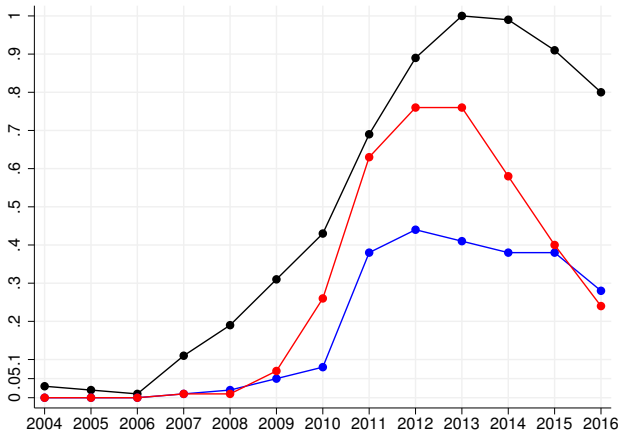
[5] While reading books electronically has been possible for years before (for example, as scanned PDFs), the Kindle and the accompanying e-ink technology improved the reading experience enough to trigger a large shift in reading behavior.

[6] 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).

Figure 1: Adoption of e-reading devices and digital self-publishing platforms in the US



• Share of US adults reporting to own an e-book reader.
Source: Pew Research, `http://tinyurl.com/q2lt5ou`



• *CreateSpace,* • *KindleDirect,* • *Smashwords*
Source: Google Trends US, relative frequency

have their books published. As such, self-publishing platforms are comparable to so-called Vanity Presses and Print-on-Demand services, which also allow any author to publish their physical books for a nominal fee.[7] 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 more popular among both authors and readers. While traditional publishers needed to "bet on the right horse" and would only take on those book ideas which they expected to be successful, self-publishing platforms provided a distribution channel for *all* books.

The right-hand panel of Figure 1 depicts the relative Google search volume for three major digital self-publishing platforms, *CreateSpace*, *KindleDirect*, and *Smashwords*. A steep rise in interest in these platforms coincides with the uptake of e-reading device ownership. For example, Amazon's *Kindle Direct Publishing* was introduced concurrently with the *Kindle*, in 2007, while *Smashwords* began operations in 2008. Searches for *CreateSpace*, which had existed as a distributor of physical books, CDs, and DVDs before being acquired by Amazon in 2005, also gained substantial attention after 2008.

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*

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

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.[8] But E.L. James was not the only aspiring author to find success in the self-publishing sphere before finding a traditional publisher to publish current or future work with. 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).[9]

Self-publishing has also served as an alternative for established authors to reach consumers (McCartney, 2016). Romance writer Jamie McGuire, for example, began her professional writing career by self-publishing *Beautiful Disaster* in May 2011. This book struck a "major" deal with Atria Publishing (an imprint of Simon and Schuster) in July 2012, including a sequel to be published a year later. In 2015 McGuire returned to self-publishing for another sequel, *Beautiful Redemption*, which even found shelf space at Walmart. She "still plan[s] to traditionally publish, but with books that [she] feel[s] are best suited for that route" (McCartney, 2016). Other authors pursue similar strategies of publishing some books traditionally while self-publishing others.

Yet other authors reject traditional deals outright. For instance, Vi Keeland and Penelope Ward rejected all offered deals from publishers because they felt they could make more money by self-publishing. Accordingly, their self-published novel, *Cocky Bastard*, was listed on the New York Times bestseller list in 2015. 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. We begin by assuming that authors and publishers have full information about an idea's appeal, and

[8] 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 which is supported in our data, and which we exploit in our analysis.

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

we then introduce uncertainty to show how self-publishing may affect the predictability of success.

3.1 Perfect Information

Consider a world with many authors who have ideas for books. Each idea has observable characteristics which determine a book's type m, drawn from some random distribution.[10] An idea's type may reflect vertical differentiation, as well as how the book aligns 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 consumer demand for a book of type m is linear, 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.[11] After obtaining the license, publisher j sells the book in the product market with demand as in Equation 1, paying an exogenously given royalty rate $s \in (0, 1)$ per dollar of revenue to the author.[12] 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*} is the quantity that maximizes the publisher's profit π_m^j – given the inverse demand function in Equation 1, it is $\frac{a_m}{2} - \frac{c^j}{2(1-s)}$. The author's profit from selling a license to publisher j – without the lump sum – is

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

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

[11] 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).

[12] 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).

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, which previously was prohibitively expensive. When using a digital self-publishing platform, the author's profit can be written as

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

where $q_m^{D*} = \frac{\lambda a_m}{2} - \frac{c^D}{2s}$ is the author's profit-maximizing quantity when publishing digitally, and the type-specific $\lambda_m \in [0, 1]$ reflects the possibility that the digital market is smaller than the physical market. The royalty rate s is the same for the digital platform as for traditional publishers.[13]

License Fees

For the traditional publisher to obtain the right to publish a book idea, her lump-sum payment must satisfy two conditions: 1) the author must find traditional publishing more profitable than her best outside option, which is either not to publish at all or to self-publish; and 2) the lump sum must be smaller than the publisher's profits before 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 regardless of the presence of self-publishing. In what follows, we consider a world with one traditional publisher, noting that all 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 will be made if the *joint* profits $\pi_m^j + \pi_m^A > 0$, and a lump-sum of $L_m = -\pi_m^A$ is likely.[14] 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 the (monopolistic) publisher j must increase her lump-sum payment L_m to the author, compared to the monopoly case without self-publishing.

Proposition 1 *Lump-sum payments increase for some books due to the advent of digital self-*

[13] 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 keep the number of parameters small, noting that a larger s for digital publishing would make our results stronger, as we explain below.

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

publishing platforms.

Equation 4 shows that $\pi^D > 0$ when $\lambda_m > \frac{c^D + 2\sqrt{sF_m^D}}{as}$. When the author's profit is positive, the traditional publisher has to increase her lump-sum payment in order to obtain the license to the author's idea.[15] □

Note that the second condition for traditional publishers to publish becomes more binding as the author's profit under digital self-publishing increases. The largest possible profit for the author when publishing traditionally 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. Hence, some books which would have entered the market with traditional publishing are now self-published.

Figure 2 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 $\pi_m^{D,\max}$ when publishing digitally (dotted line), and her maximum profit when publishing traditionally (solid line), as functions of the relative size λ of the digital market.[16] 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 remains published traditionally. When $\lambda_m > 0.81$, the author chooses to publish digitally.

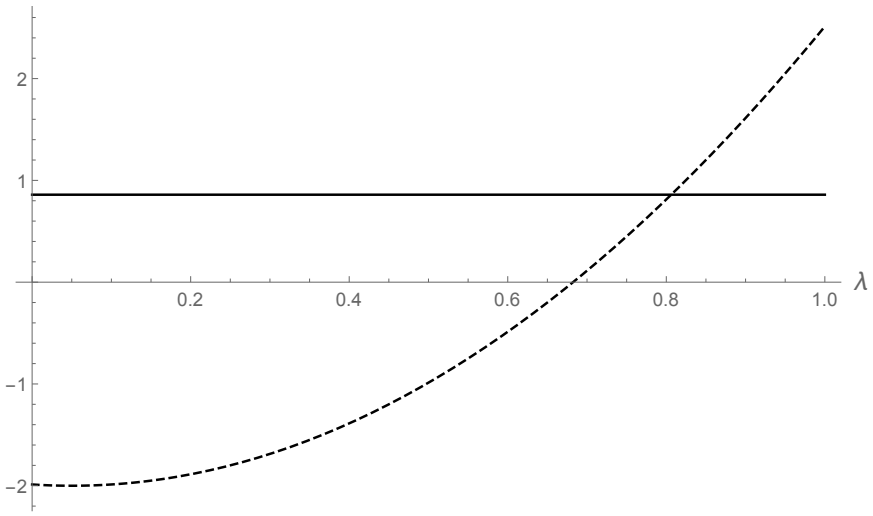
Finally, there are parameter values for which digital publishing is profitable for the author even though traditional publishers cannot publish the product profitably at all. Hence, our model further predicts the following relationship:

Corollary 1.1 *More books will appear on the market due to the advent of digital self-publishing*

[15] Of course, if the author receives a larger share s when publishing digitally, the lump-sum payment would increase for smaller values of λ_m as well.

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

Figure 2: Example of author profits under traditional publishing and digital self-publishing



Maximum author profit under traditional publishing ($\pi^{A,max}$), *Author profit under digital self-publishing (π^D)*
Note: *The x-axis shows λ – the relative size of the digital market. The vertical axis shows the author's respective profits. We use the following values: $a_m = 10, c^j = 5, c^D = 0.1, F^P = 5, F^D = 2$, and $s = 0.2$. When $\pi^D > 0$, the traditional publisher increases its lump-sum payment. When $\pi^D > \pi^{A,max}$, the author chooses to self-publish.*

platforms.

If $\pi_m^{A,\max} < 0 < \pi_m^D(q_m^{D*})$, the author self-publishes her book even if traditional publishers passed on it. In the above example, new books come to market if $F_m^j > 5.86 \approx 3 \times F_m^D$ and $\lambda > 0.68$. □ Intuitively, some books do not have enough appeal (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.

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: $a_m^i \sim N(\mu_m^i, \sigma_m^i)$. We assume for simplicity that publishers and authors are risk neutral.[17] Under uncertainty, all above results hold under the same conditions, with each agent's beliefs a_m^i substituted for the true a_m.

An agent can learn about book type m's typical appeal (the true average a_m across all ideas within that type) from data that she generates herself – her previous market entry decisions – or from

[17] If agents are risk averse, the "cutoff" expected profits will be lower under each publishing strategy.

data generated by other agents. We think of these data as being incorporated in future predictions of the market potential of type-m ideas in a Bayesian updating process.

Define the precision of the prior belief as $\tau_m^i = 1/\sigma_m^i$, and denote the precision of the true distribution of a_m as r_m. DeGroot (1970, p. 167) show that given an agent's prior distribution on the mean appeal $a_m^i \sim N(\mu_m^i, \sigma_m^i)$, and with n_m additional data points $X_{k,m} = x_{k,m}(k = 1, ..., n_m)$, the posterior distribution of a_m^i is also a normal distribution, with mean $\mu_m^{i'}$ and precision $\tau_m^i + n_m r_m$, where

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

In this equation, $\overline{x_m}$ 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 postulate the following:

Proposition 2 *Publishers make smaller errors when predicting the market potential of book ideas due to the advent of digital self-publishing platforms.*

Corollary1.1 shows that the number of new products increases due to digital self-publishing, which implies that more information is available to update priors – the number of observations n_m increases, and the predicted market appeal converges 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 not only leads to more accurate predictions on average, but also more accurate predictions for each individual idea within a type. The precision of the posterior distribution, $\tau_m^i + n_m r_m$, increases as $n_m \to \infty$, making any estimate of a type m product's appeal less noisy. □

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

effect.

4 Data and Identification

The theoretical model provides several predictions which we aim to examine empirically in the following sections. First, digital self-publishing allows more books to enter the market. Second, the additional competition on the platform level leads to contracts which favor the author more. And third, publishers become better at predicting the success of an idea.

4.1 Data Sources

To take the theoretical insights to the data, we use information from a variety of sources. 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 National* (France). We describe the two main datasets in detail here, and we refer to 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 authors, agents and publishers in which agents post information about book-related deals and the involved entities.[18] We observe all posted deals between January 1st 2002 and December 31st 2015 – a total of 100,772 deals. Table 1 shows two representative posts on that platform. 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). Most importantly, the database allows us to quantify the size of the lump-sum advances 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%).

[18] See `http://www.publishersmarketplace.com`. Agents have an incentive to post deals because these deals make them attractive to potential clients (authors).

Table 1: Examples from Publishers Marketplace

Date	Type	Text
12/04/2006	Fiction: General/Other	Lynn York's second novel, a follow up to her debut The Piano Teacher, to Trena Keating at Plume, in a very nice deal, by Suzanne Gluck at the William Morris Agency.
09/15/2014	International rights	German rights to Karin Slaughter's COP TOWN, the latest novel from the #1 international bestselling author, to Nicola Bartels at Blanvalet, in a major deal, for publication in 2015, by Victoria Sanders & Chris Kepner at Victoria Sanders & Associates.

After data cleaning, we observe 52,260 book deals and 39,584 rights deals.[19] From the posted information, we define six control variables, which describe deal-specific characteristics of authors and license deals:

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.

4.1.2 Success of Book Ideas

In an ideal research setting, we would track a book's licensing deal and its subsequent sales on the product market, both measured in absolute values, along with cost information. Then, we would compare license fee payments to net earnings in the spirit of the theoretical model above. We deviate from this setup slightly because we only observe categorical deal size variables, and because titles may have changed from the time the deal was made to the book's publication.

To increase the degrees of freedom, we connect our *categorical* licensing information with *categorical*

[19] For example, some (8,928) deals include two or more books, making it difficult to compare payments across deals. We therefore exclude these observations.

sales information, not on the title-level, but on the author-level. Specifically, we the date of an author's first appearance in the USA Today Top 150 bestseller list (using weekly data from 1993 to 2016) to create a measure of the extensive margin – whether an author is "sufficiently" successful or not. Exploiting the time dimension of these data, we construct a measure of whether an author who appears in the licensing dataset will become sufficiently successful at some point in the future, and we measure the time lag until success. The bottom panel of Table A.1 provides descriptive statistics regarding the ex-post success of the deals.

Of course, it is unclear what it means for an author to be "sufficiently" successful. We use the Nielsen Bookscan dataset, which offers titles' historical sales data, to inform inference on the commercial success of books which reach certain milestones in the USA Today weekly Top 150 bestseller lists. In particular, we estimate a title's lifetime revenues based on its ranking and the average as described in Section 5.3.1.[20]

4.2 Identification

Traditionally, 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 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, while the Amazon Kindle was introduced in the United States in November 2007, this device was only sold on the American platform (amazon.com), and the first Kindle with the ability to download content wirelessly outside of the United States (the Kindle 2 International) was not introduced

[20] We were only able to obtain historical sales information for a random sample of 100,000 books because Nielsen does not offer bulk access to all books. Out of those, we could match less than 10% with our licensing data. We therefore use a more complete dataset, noting that even if we had sales data on all books in the *Publishers Marketplace* database, we would still need to make assumptions about the book-specific cost structure of publishers (including advertising expenditure, variable license fees, retail mark-ups on wholesale prices, etc.).

until October 2009.[21] In addition, self-publishing is a phenomenon that is much more prevalent in the English-speaking world compared to non-English speaking markets.[22]

In 2011, the market share of e-books (fiction) was 13.6% in the US, and Amazon had announced that is has sold more e-books than print books.[23] In contrast, the market share of e-books was 0.5% in Germany, and 1.8% in France (Wischenbart, 2012). While about 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.[24] Amazon's self-publishing platform *Kindle Direct Publishing* started in Germany in April 2011, and was not rolled out in more countries until August 2013.[25] The commercial break-through of self-publishing did not happen until 2014 in France, and the Frankfurt book fair only recognized self-publishing in a distinct exhibition area in 2015.[26] This is suggestive evidence that the digital disruption happened in the United States earlier and more intensely than in other countries.

4.2.2 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 make use of quasi-experimental variation to help distinguish between groups of observations that are more strongly affected by the introduction and adoption of digital self-publishing platforms than others.[27]

To identify a quasi-experimental "treatment group", we conducted a number of interviews in the field, asking publishers and agents in the US 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.[28] Several reasons were given. First, romance books are reportedly relatively easy

[21] See `https://tinyurl.com/ybpdfmls`.
[22] See `http://tinyurl.com/ya3yn3w7`.
[23] See `http://www.nytimes.com/2011/05/20/technology/20amazon.html`.
[24] See `http://tinyurl.com/y9nvyjbh` and `http://tinyurl.com/ycuux8jc`.
[25] See `http://tinyurl.com/ycpspp6m` and `http://tinyurl.com/yd82wvst`.
[26] See `http://tinyurl.com/y98l5sy7` and `http://tinyurl.com/y82dox4s`.
[27] Note that the theoretical model equivalently distinguishes between book types with different values of λ, the relative size of the digital market.
[28] We use romance as a shorthand for romance/erotica in the remainder of the paper.

to write, among other reasons because they are less research-intensive than, say, historical non-fiction books. Second, advertising and marketing costs are lower than in other genres. Romance novel readers seem to be a close-knit group which often communicates via 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 for which the book's cover is not seen by others.

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.[30] In contrast, the share of the romance category among license deals with traditional publishers on *Publishers Marketplace* is only 16%.

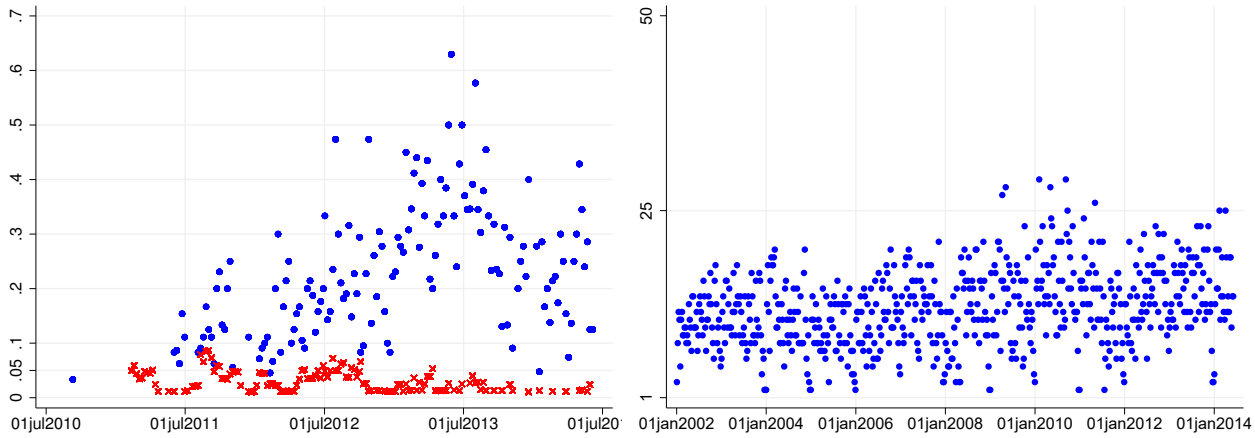
Of course, the number of titles by itself does not necessarily imply that the self-published novels are popular among consumers. The left-hand panel of Figure 3 depicts the share of originally self-published books among bestsellers in the *USA Today* Top 100 rankings, each week from 2009 to 2014, distinguishing between self-published romance books (blue) and all other self-published books (red). In 2011, up to 20% of bestsellers in the romance category had a self-publishing background, and this number increased to over 50% in 2012. During the entire observed period, much less than 10% of the bestsellers in other categories were originally self-published. Hence, the market potential for self-published books (λ in our theoretical model) seems larger for romance novels than for other book types.

The difference in popularity of self-published romance books could of course 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 like *Fifty Shades of Grey* did not change publishers' expectations regarding the profitability of romance novels in general. Again,

[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).

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

Figure 3: Romance books in USA Today Top 100 bestseller list



Share of originally self-published romance books in Top 100 *Number of traditionally published romance books in Top 100*

• *Romance,* × *Non-Romance* • *Romance*

we confirm this notion with quantitative data. The right-hand panel of Figure 3 shows the number of traditionally published romance novels among the USA Today weekly Top 100 rankings. The number of traditionally published romance novels remains stable throughout the studied time period.

The qualitative and quantitative evidence provides support for the idea that the realized appeal of romance novels did not change but for the introduction of self-publishing, and it is unlikely that changes in the market for book ideas are solely driven by downstream consumer demand. The fact that traditionally published romance novels do not become more popular overall leads us to believe that the mean quality of ideas which are presented to publishers does not increase either. Any changes in the market are likely the direct results of the introduction of digital self-publishing platforms.

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 the license deals for those 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.

5.1 Changes in the Number of Books

Figure 4 shows the total number of new titles (including re-editions) per country and year for the US, Germany and France from 2002 to 2010. During this time period, digital self-publishing de facto did not exist in Germany and France. While the number of new titles stagnates in France and Germany, we see an exponential increase in the US after 2008 (note the logarithmic scale). In a more formal analysis, we estimate a difference-in-differences model defined as

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

where ν_t and μ_i are year- and country-fixed effects, respectively, and $After_t$ is 1 in the year 2009 and beyond. The OLS estimate of δ is 1.044 with a standard error of 0.278 (p-value 0.000). Although the small number of observations in this annual dataset doesn't allow too strong an inference, the point estimate implies that digital self-publishing leads to almost a doubling of the number of books on the market ($exp(1.044) - 1 = 1.84$), compared to how the market might have evolved without it. And yet, this estimate is conservative because it does not take the almost exponential growth of the dependent variable into account.

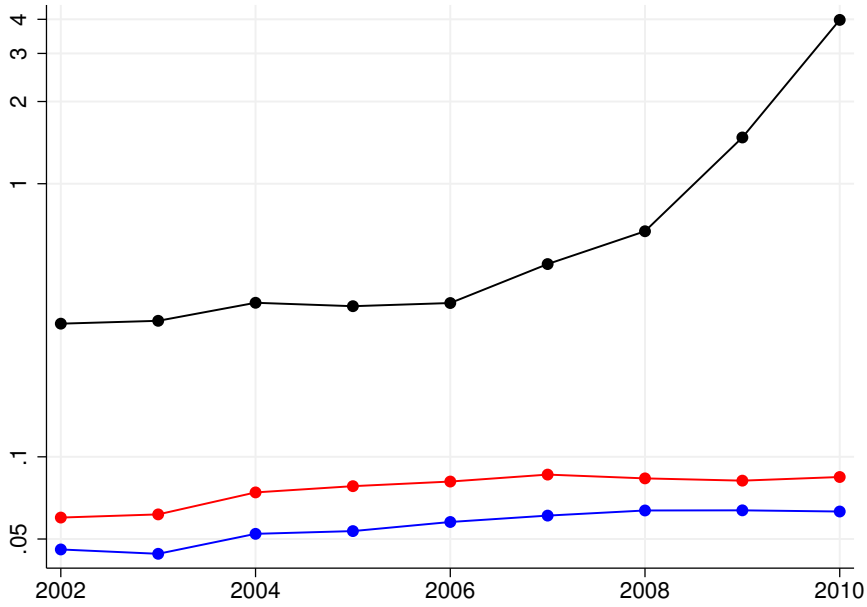
Since publishers can observe the sales performance of published books (including those which were published by competitors or self-published), this substantial increase in the number of books mechanically translates into more information that can be used to try to predict demand for new books. Anecdotal evidence suggests that publishers indeed use historical sales data to make business decisions, such as licensing new books.[31]

5.2 License size

Section 4.2 suggests that romance and erotica novels are inherently better suited for digital self-publishing than other genres, making them a prime candidate for the treatment group in a difference-in-differences analysis. We now estimate the impact of self-publishing on advances based on this identification strategy, and we later provide additional evidence that the identifying assumptions hold.

[31] See `http://tinyurl.com/y7dxzynh`.

Figure 4: Number of new books per year in USA, Germany and France



• *USA,* • *Germany,* • *France*

Note: Total number of new titles 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 national de France* (France).

5.2.1 Baseline Estimation

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

$$LogSize_{i,j,k,t} = \alpha + \beta_1 R_j + \delta(After_t \times R_j) + \kappa C_j + \mu_t + \gamma_t + \varepsilon_{i,j,k,t} \tag{8}$$

The unit of observation in this model is a license deal i closed on day t between author j and editor k (at publisher p). R_j indicates whether the author is ever listed on a deal in the romance category, and $After_t$ indicates whether the deal was closed after the year 2008. We are most interested in δ, the coefficient on the interaction of $After_t$ and R_j (our treated group).

We account for time-specific variation by including month-indicators μ_t in addition to year-fixed effects γ_t. To account for time-varying heterogeneity across authors, we include a vector of control variables $C_{j,t}$. We do not include author-fixed effects because we are more interested in identifying

the effect of digital self-publishing within genres and across authors, than in its effect within authors. However, we do want to absorb any unobserved heterogeneity across editors (and therefore publishers) and hence include editor fixed effects ν_k. For the same reason, and to avoid incorrect inference in the difference-in-difference model (Bertrand et al., 2004), we cluster standard errors at the editor level.

The main estimation results are reported in Table 2. The dependent variable in column (1) is the logarithm of the license payment, where we take the midpoints of each deal category in the data.[32] The coefficient on the interaction term (After × Romance) is positive and statistically significant, suggesting that digital disintermediation increased advance payments by 16.8%.

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 increase in advance payments of 24.9% and 18.9%, respectively.[34] Throughout specifications, the coefficients of all control variables have the expected positive signs as well.

5.2.2 The Mechanisms Behind Advance Increases

We show above that the emergence of self-publishing has increased the advances paid to authors. The theoretical model suggests that these increases are due to an increase in λ – the size of the digital self-publishing market relative to the traditional market – and a subsequent increase in the level of competition. Thus, one might also expect the increases in advances to be largest when publishers expect a large market appeal. Here, we investigate whether the impacts of digitization are strongest for certain types of book deals.

More precisely, we proxy for an idea's market potential a with our control variables ($a \in C$, the vector of control variables), and we interact each of these control variables with the difference-in-

[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).

[33] An ordered probit estimation provides nearly identical results but we report OLS results for ease of interpretation.

[34] Point estimates in columns (1) and (3) are transformed to percentage values according to $PercentageChange = (exp(Coefficient) - 1) * 100$. In column (2), we compare the coefficient to the sample mean, i.e. $(33.48/134.61) * 100 = 24.87$.

Table 2: Results: Changes in license deals

	(1) DV: Log(Size)		(2) DV: Size		(3) DV: Deal category	
Romance	-0.143***	(0.055)	-23.594**	(11.442)	-0.161**	(0.064)
After × Romance	0.144**	(0.062)	33.477**	(13.075)	0.173**	(0.072)
Acclaimed	0.150***	(0.034)	27.016***	(7.339)	0.174***	(0.039)
Bestseller	0.985***	(0.060)	201.890***	(14.185)	1.159***	(0.071)
Contested	0.667***	(0.061)	116.204***	(15.141)	0.760***	(0.073)
Debut	0.036	(0.055)	14.557	(11.763)	0.052	(0.063)
Self-published	0.385***	(0.140)	89.558***	(32.931)	0.459***	(0.166)
Sequel	0.165***	(0.054)	27.859**	(11.966)	0.186***	(0.062)
Observations	14771		14771		14771	
$\overline{R^2}$	0.541		0.409		0.526	

Notes: Editor, month and year fixed effects, and constant not reported.
Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

differences coefficient. That is:

$$LogSize_{i,j,k,t} = \alpha + \beta_1 R_j + \delta_1(After_t \times R_j) + \kappa C_j + \delta_2(After_t \times C_j) + \delta_3(After_t \times R_j \times C_j) + \mu_t + \gamma_t + \varepsilon_{i,j,k,t}. \quad (9)$$

The results in Table 3 show that the coefficient in the simple interaction ($After \times Romance$) remains very similar to the baseline estimates in Table 2. The interaction term of $After \times Romance \times Bestseller$ is positive and significant, suggesting that license fees increase when a book's potential market size is larger. Interestingly, the coefficient for $After \times Romance \times Debut$ is large and statistically significant as well, suggesting that publishers find it easier to predict sales for new authors among more inherently digital book types. However, there is no significant difference across genres among previously self-published authors. No other interaction terms are significantly different from zero.

Finally, the size of the digital self-publishing market relative to the traditional market (λ) is largely determined by the share of consumers who own an e-reading device. We measure the impact of e-reader ownership by replacing the $After_t$ indicator variable in equation (9) with a continuous measure of e-reading device adoption (see Figure 1). Column (1) in Table 4 shows the estimated

Table 3: Results: Changes in license deals, interacted

	(1)	
After × Romance	0.046	(0.059)
After × Romance × Acclaimed	0.095	(0.233)
After × Romance × Bestseller	0.352**	(0.175)
After × Romance × Contested	-0.500	(0.456)
After × Romance × Debut	0.472*	(0.278)
After × Romance × Self-published	-0.443	(0.446)
After × Romance × Sequel	0.114	(0.245)
Observations	14771	
$\overline{R^2}$	0.542	

Dependent variable: (Log+1) Dealsize.
Notes: *After* indicates years after 2008. Editor, month, and year fixed effects.
Lower-level interactions and constant included but not reported.
Standard errors in parentheses, clustered on the editor-level.
* $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

coefficient of interest of this regression. The results imply that advances for romance novels increase with device diffusion, and the estimated magnitudes are substantial. Moving from 0 to 100% adoption of e-reading devices increases license deals by 71%.[35]

Of course, this estimate needs to be interpreted with caution because consumers' adoption decisions are likely correlated with the number (and quality) of e-books available on the market. Thus, we also allow for a flexible time structure in the spirit of Autor (2003). We test, for each year, whether the size of license deals for romance authors is different from the size of license deals for non-romance authors:

$$LogSize_{i,j,k,t} = \alpha + \beta_1 R_j + \sum^{T} \delta^{\tau} \left(\gamma_{\tau} \times R_j\right) + \kappa C_j + \nu_k + \mu_t + \gamma_t + \varepsilon_{i,j,k,t} \tag{10}$$

The excluded year is 2008, to facilitate a comparison of pre- and post-years. Column (2) in Table 4 shows the results from this specification. 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 advance payments of close to 20%.

[35] Note that, by construction, the comparison here is against all years in which device diffusion was 0.

Table 4: Results: Changes in license deals, dynamics

	(1)		(2)	
Romance × E-Reader Adoption	0.536*	(0.279)		
Year=2002 × Romance			0.390	(0.275)
Year=2003 × Romance			0.078	(0.204)
Year=2004 × Romance			0.187	(0.152)
Year=2005 × Romance			0.064	(0.111)
Year=2006 × Romance			0.053	(0.108)
Year=2007 × Romance			0.001	(0.108)
Year=2009 × Romance			0.236**	(0.100)
Year=2010 × Romance			0.174*	(0.099)
Year=2011 × Romance			0.212**	(0.100)
Year=2012 × Romance			0.143	(0.112)
Year=2013 × Romance			0.252**	(0.116)
Year=2014 × Romance			0.265**	(0.128)
Year=2015 × Romance			0.236	(0.146)
Observations	14771		14771	
$\overline{R^2}$	0.541		0.541	

Dependent variable: Log(Dealsize+1).

Notes: Editor, month and year fixed effects, and constant not reported.

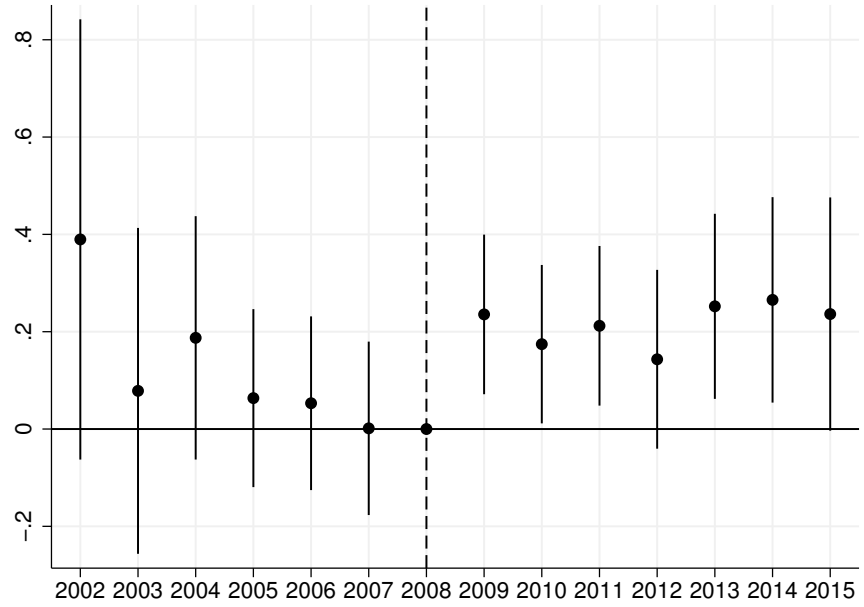
Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

The specification further provides evidence that a necessary condition for the identification of the effect holds: there are no differences in the trends of advance payments for romance novels and other genres before 2008. Figure 5 illustrates these dynamics. It plots the estimated year-specific difference coefficients (δ^{τ}), lending additional support to the identifying assumption.

5.2.3 Placebo Tests and Robustness Checks

Placebo Exercises Although Figure 5 suggests no differences in trends across genres before 2008, there might be factors beyond digitization which affected the size of advances, and which we have not accounted for. Here, we further test whether our results are a reflection of a pre-existing trend unrelated to the advent of digital self-publishing. We pretend that the uptake of self-publishing platforms took place in 2005, and estimate the model defined in equation (9) on the subsample spanning 2002–2008. A statistically significant and positive coefficient for $After \times Romance$ would be consistent with the presence of an increasing trend unrelated to the entry of digital self-publishing.

Figure 5: License deals, group differences over time



Note: OLS estimates of the δ^{τ} coefficients obtained from a regression of equation (10), i.e. yearly differences in *LogSize* between the treatment group (Romance authors) and the control group (non-Romance authors). The omitted year is 2008, standard errors clustered on the editor-level, and bars indicate 90% confidence bands.

On the other hand, if the estimated coefficient is statistically indistinguishable from zero, we can reject that hypothesis. We find the latter in columns (1) and (2) of Table 5. In column (1) we estimate $After \times Romance$ using the subsample that includes the years 2002 to 2008. The coefficient is negative and significant at the 5% level. Column (2) of Table 5 excludes the year 2002 – an outlier in our data. The point estimate remains negative but is no longer distinguishable from zero, providing further support for the identification strategy.

Moreover, while Figure 3 provides some indication that the romance genre did not become more popular overall, there may still be relative differences in popularity which our measure did not pick up. If romance books become more popular, we should also see an increase in payments for other commercialization channels, such as the market for international distribution (which typically involves translation and adaption for foreign markets), or the market for TV and film licenses. However, because it is substantially more difficult for authors to circumvent traditional publishers in these channels, traditional publishers do not (or to a much lesser degree) face competition from

Table 5: Results: Changes in license deals, placebo exercises

	(1) Placebo-Timing DV: Log(Size), books 2002–2008	(2) Placebo-Timing DV: Log(Size), books 2003–2015	(3) Placebo-Group DV: Log(Size), rights 2002–2008
After × Romance	-0.136*	-0.119	-0.108
	(0.074)	(0.076)	(0.109)
Observations	8166	7278	8194
$\overline{R^2}$	0.462	0.492	0.519

Dependent variables: Log(Dealsize+1), columns (1) & (2) are restricted to years < 2008, column (3) only includes deals concerning (audio, film, television, international) rights.
Notes: Editor, month and year fixed effects, and constant not reported.
Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

digital platforms here. Estimating the model defined in equation (9) on rights deals can therefore serve as a test of the hypothesis that unobserved factors specific to the romance genre, but unrelated to self-publishing, drive our results in Table 2. The estimated coefficient of $After \times Romance$ in column (3) of Table 5 is not different from zero and small in magnitude. Accordingly, it seems unlikely that advances for romance novels rose because more people wanted to read them.

Robustness Checks Our results suggest that the option to self-publish makes authors better off on average. But we observe deal sizes for only about 25% of all deals, and even then, we only observe these sizes as (arbitrary) categories. Thus, we examine the possibility of non-response bias and the dependence of our transformation of the categories in further robustness checks.

To investigate potential non-reporting issues, we run a linear probability model where our dependent variable indicates whether the deal size is reported at all.[36] While non-reporting overall might be correlated with the deal size or change over time, 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 of $After \times Romance$ in column (1) of Table A.2 confirms that there is no evidence of such non-random missing data.

Columns (2)–(6) of Table A.2 show results of linear probability models using indicator variables

[36] 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 do not find statistically significant differences between deals of romance and non-romance authors before 2008, providing support for our identification strategy.

for the respective size categories. Note that the comparison in each column is against all other categories, e.g. *<50* vs. *≥50* and *Size not reported*. The average effect in Table 2 is 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 of *nice* deals (<50k).

In Table A.3, we report results from an ordered logit model, where we include group-specific trends instead of editor fixed effects, which allows flexible estimation of thresholds between categories.[37] The results from this exercise are very similar to our main results as well. We see significant increases in the baseline probability of *significant* (250k–499k) and *major* (>500k) deals.

5.3 Changes in Predicting Ex-Post Appeal

The above analysis shows a digitization-related increase in the advances paid to authors and an increase in the number of books on the market. These additional books may provide useful information for authors and publishers who rely on correctly predicting an idea's commercial success ex ante. We examine this effect in more detail by comparing each idea's commercial success to its advance payments, for romance books and other genres, before and after the arrival of digital self-publishing.

5.3.1 Defining Success

Section 4 explains that we use data from two sources to determine an idea's success after a deal: weekly Top 150 ranking data from USA Today, and actual sales data from the Nielsen Bookscan database. We use the USA Today rankings because these data are clean and complete, but in order to determine whether a title is "sufficiently" successful, we need to map these rankings into monetary values.

To achieve this, we use weekly sales data from *Nielsen Bookscan* covering the top 100 bestselling books in any given week from 2004 to 2012, along with information on quantity sold and the suggested retail price.We regress cumulative sales on the number of weeks we observe a book in a given ranking position and a linear year trend. That is, we estimate $CumSales_i = \alpha +$

[37] The existing econometric theory regarding ordered logit/probit models with fixed effects allows dealing with individual-specific fixed effects, i.e. on the deal level, but has not considered multi-level fixed effects, like the editor-fixed effects in our case. See for example (Baetschmann et al., 2015).

$\sum_{r=1}^{100} WeeksRank_r + \theta y_t + \varepsilon_{i,t}$.[38] Using these estimates, we predict life-time sales of books we observe in the USA Today sales ranking. Table 6 reports means and corresponding standard errors for books which were ranked 1–10, 11–50, and 51–100 for at least one week, by the year of first appearance in the rankings. A few things become evident. First, standard deviations are substantial, reflecting the fact that some weeks contain extraordinarily successful books. Second, there is a general downward trend over time, perhaps indicating a shift away from physical books to e-books.[39] Third, a typical top 10 book generates around $7.0m in revenue (385k units), a typical top 50 books generates about $1.3m (58k units) and a typical top 100 book generates about $500k (14k units).

Table 6: Predicted life-time revenues

	Top 1–10		Top 11–50		Top 51–100	
	Mean	SE	Mean	SE	Mean	SE
Released in 2002	7018.04	968.63	2029.89	278.10	472.42	37.38
Released in 2003	15017.56	3154.61	2277.90	302.89	449.30	35.09
Released in 2004	8616.55	1105.32	1779.07	161.98	418.94	35.87
Released in 2005	8421.15	1075.66	1546.57	137.14	506.02	38.06
Released in 2006	11463.57	2497.07	1646.90	143.81	412.16	32.57
Released in 2007	8031.21	1297.41	1616.73	211.71	474.54	31.06
Released in 2008	9301.21	2031.29	1528.12	123.24	506.79	35.41
Released in 2009	8696.44	1637.61	1417.13	98.84	469.14	36.84
Released in 2010	7236.18	1239.29	1512.19	108.86	424.66	26.36
Released in 2011	5742.05	654.24	1483.69	110.39	457.05	26.20
Released in 2012	10122.47	2090.59	1285.78	81.72	421.71	21.07
Released in 2013	5666.45	624.90	1181.92	86.44	492.52	20.04
Released in 2014	4608.09	518.32	1108.75	65.28	482.08	18.67
Released in 2015	5029.08	943.21	944.85	45.72	489.10	17.55
Overall	7851.15	393.87	1454.35	36.15	464.90	7.25

Notes: Out-of-sample predictions of life-time sales predicted from a model specified as $CumSales_i = \alpha + \sum_{r=1}^{100} WeeksRank_r + \theta y_t + \varepsilon_{i,t}$, using weekly top 100 sales figures from 2004 to 2012. Means and standard errors of the mean given in thousands.

The profit margins of the five largest publishing houses – Hachette, HarperCollins, Penguin, Random House and Simon & Schuster – are remarkably stable between 2006 and 2013, with an average value of 9.7%.[40] We therefore assume that publishers can extract a fixed (percentage) rent in our en-

[38] To save space, the corresponding results table is not reported here but is available upon request.
[39] It is important to recognize that Nielsen's Bookscan database only covers sales of physical books.
[40] See `http://tinyurl.com/meao8hj`.

tire observation period, such that we define our empirical measure of $\pi^{A,\max}$, $\hat{\pi}^{A,\max} = 0.9(q^{j*}p^{j*})$. With a 90% confidence band, the life-time $\hat{\pi}^{A,\max}$ of a typical top 10 book is [$4852k, $8508k]. The estimate is [$1082k, $1407k] regarding a top 50 book and [$378k, $460k] regarding a top 100 book. Using this information, we define an empirical measure of prediction error. We set the variable *Error* to -3 if an author reaches the top 100 bestseller list at some point after we observe a deal in the smallest category (<50k). Similarly, we set *Error* to -2 in case of a deal between 50 and 99k, to -1 in case of a deal between 100 and 249k, to 0 in case of a deal between 250 and 499k and to 1 in case of a deal bigger than 500k. We follow the same logic regarding future top 50 and top 10 authors, and authors who do not enter the Top 100 at all.

5.3.2 Estimation and Results

The simplest way to test Proposition 2, i.e. whether digital self-publishing should result in less prediction error, is to compare the squared *Error* across treatment and control groups before and after the advent of digital self-publishing. To that effect, we repeat our analysis from equation (9), using our squared *error* measure as the dependent variable.

Column (1) of Table 7 shows the results of this specification. The coefficient on $After \times Romance$ is negative and statistically significant, providing strong support for Proposition 2. The advances paid to authors more accurately reflect the commercial success of an idea. The estimated effect size is a substantial 62.3% decrease in the squared prediction error.[41]

We run further regressions to distinguish between type I and type II errors. In spirit of Proposition 2, we should see fewer authors with *major* (> 500k) deals who never end up on the top 100 bestseller list, and we should see fewer authors with *nice* (< 50k) deals who later sell enough copies to show up on the top 10 bestseller list. In particuar, we test for these different error types using a triple-difference model, interacting $After \times Romance$ with the indicators $LargestDeal$ and $SmallestDeal$.

Results are reported in columns (2) and (3) of Table 7. We find negative and significant coefficients for each type of error. Independently, each result could be explained by increases in bargaining power, either with more bargaining power for authors (larger licensing deals for very successful

[41] Coefficient divided by sample average, (2.30/3.69)*100=62.3%.

Table 7: Results: Changes in predicting ex-post appeal

	(1) Error^2		(2) Not in Top 100		(3) In Top 10	
Romance	2.138***	(0.799)	-0.044***	(0.008)	0.000	(0.005)
Largest Deal			-0.355***	(0.016)		
Smallest Deal					0.000	(0.001)
Romance × Largest Deal			-0.179***	(0.061)		
Romance × Smallest Deal					0.008	(0.007)
After × Romance	-2.301***	(0.825)	0.034***	(0.008)	0.011	(0.007)
After × Romance × Largest Deal			-0.208***	(0.062)		
After × Romance × Smallest Deal					-0.016**	(0.007)
Observations	1569		58614		58614	
$\overline{R^2}$	0.502		0.250		0.064	

Dependent variable: Column (1): Linear measure of squared prediction error. Columns (2)–(3): Author appears (not) in the top 10/100 bestseller list. **Notes:** *After* indicates years after 2008. Editor, month, and year fixed effects. Lower-level interactions and constant included but not reported. Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

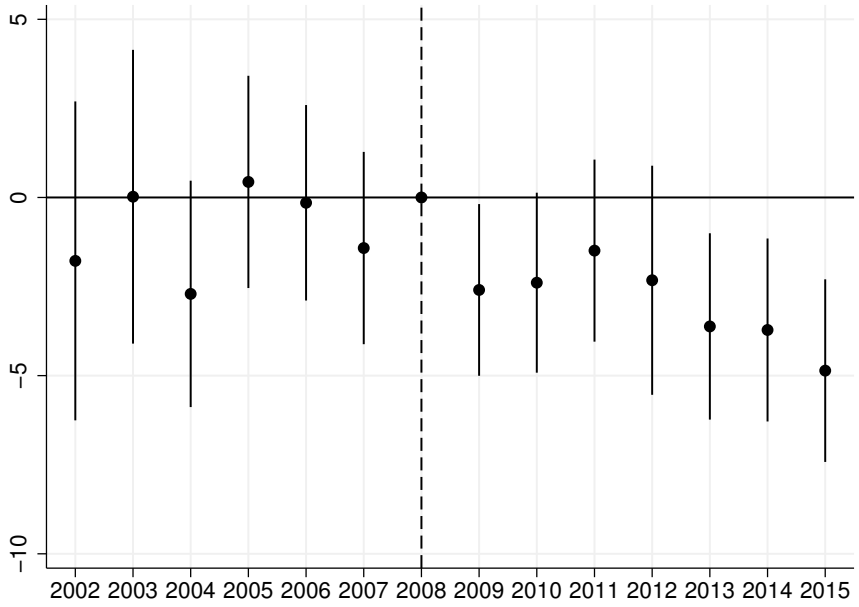
authors) or more bargaining power for publishers (smaller licensing deals for relatively unsuccessful authors). But bargaining power cannot increase simultaneously for both sides. Hence, the data are strongly indicative of improvements in the information environment: more entry produces more data to predict market outcomes in the future.

5.3.3 Robustness Checks

We test the necessary condition of the difference-in-differences model by estimating a model of annual differences similar to equation (10), using our squared *Error* measure as the dependent variable. Figure 6 plots the estimated year-specific difference coefficients (δ^{τ}). The 90% confidence band for all coefficients before 2008 includes zero. We cannot reject the hypothesis of a common trend of treatment and control group, lending support to the identifying assumption. In addition, the error for romance books decreases significantly and persistently after 2008, with the largest differences in the most recent years, reflecting a cumulative improvement based on an ever increasing number of published books.

Our results are robust to a number of different specifications, which we show in Tables A.4 and A.5. First, we show that estimates are similar in an ordered logit model that allows for non-

Figure 6: Prediction error, group differences over time



Note: OLS estimates of the δ^{τ} coefficients obtained from a regression of equation (10), i.e. yearly differences in *Error* between the treatment group (Romance authors) and the control group (non-Romance authors). The omitted year is 2008, standard errors clustered on the editor-level, and bars indicate 90% confidence bands.

linearity. Column (1) of Table A.4 reports individual $After \times Romance$ coefficients for each value of the *Error* variable, relative to the baseline case where license deal and (estimated) future sales are in the same monetary category. The results suggest that the advent of digital self-publishing has reduced smaller errors more than larger errors. However, the coefficients are not significantly different from each other. As per definition of the *Error* variable, it only includes deals of those authors that do appear on the top 100 bestseller list at some point. In column (2) of Table A.4, we extend our definition of *Error* to also include authors that never appear on the top 100 list after we observe the deal. In these cases, we set *Error* to 0 in case of *nice* ($<$ 50k) and *very nice* (50–99k) deals, 1 in case of a *good* (100–249k), 4 in case of a *significant* (250–499k), and 9 in case of a *major* ($\geq$ 500k) deal. The results remain very similar: we get negative coefficients, most of which are significant.

Second, although our definition of *Error* is based on quantitative sales data, it remains an estimate of actual sales. In an attempt to relax all assumptions we made to arrive at this sales estimate, we

run a linear probability model where the dependent variable indicates whether an author appears in the top 100 bestseller list at some point after the deal. We first test whether our results are driven by selective reporting. As described above, we only observe information about the size of license deals for about 25% of the deals in our data. In column (1) of Table A.5, we interact $After \times Romance$ with an indicator of whether a deal is reported. The triple interaction is not significantly different from zero, suggesting that our results are not driven by selective reporting. We then interact $After \times Romance$ with our linearized measure of deal size, $Log(DealSize)$. This is helpful to relax the somewhat arbitrary assumptions we made in our mapping of deal categories and (estimated) sales categories. The results are reported in columns (2) and (3) of Table A.5. When using a linear specification in column (2), we get a small, insignificant, negative coefficient. However, when we allow for non-linearity in column (3), we find evidence of a hump-shaped relationship. Authors with larger deals are more likely to appear in the top 100 bestseller list in the future, but at a decreasing rate. This provides further support to the idea derived in Proposition 2.

6 Discussion and Conclusion

The role of intermediaries has changed significantly as digital technology has gained importance in many markets. With both the production and distribution of products becoming cheaper, it has become 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 new contracts between the inventors and firms because the inventor's outside option changes. How these contracts change, and more importantly how the incentives to innovate and the distribution of products are affected has been difficult to assess in the past, although welfare implications are potentially large. A rigorous quantification of the welfare effects that our analysis implies is beyond the scope of this paper. However, we find that digital self-publishing has increased overall welfare, with some fine but important nuances.

First, some books which would not have been made available by traditional institutions become available now. Our most conservative estimates suggest that the advent of digital self-publishing has brought about 1.84 times more books, most of them at considerably lower prices. 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.[42] 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. Substantially greater product variety at lower prices suggests that the introduction of digital self-publishing and e-reading devices has the clear potential to be welfare enhancing for consumers.

Second, we find that authors benefit from larger license fees. Our most conservative estimates suggest that the average disintermediation-related increase in these payments is about 17%. This of course creates incentives for authors to produce more books. It is not necessarily clear how this affects overall welfare, though. Larger license fees most likely only reflect a redistribution of income from publishers to authors, making it a zero-sum game from a broader welfare perspective.

Finally, we find evidence that publishers become better at predicting the ex post appeal of book ideas. This of course makes the market for ideas more efficient. While authors who will become successful in the future can earn higher license fees, license fees for authors who will not become successful in the future decline. As a result, publishers and high-"quality" authors are better off due to digital self-publishing, although low-"quality" authors incur a welfare loss because publishers find it easier to reveal their ex post appeal and therefore pay lower licensing fees. The reallocation of resources towards "better" ideas also benefits consumers, as more high-quality products become accessible to more consumers.

Just as importantly, the reallocation of resources has large long-term implications in the market for ideas. Traditional institutions which have a comparative advantage in marketing products and reaching a lot of consumers can continue to exist alongside new platforms which allow inventors to reach consumers directly. Our research shows that both institutions have merit, and they can even complement each other by removing inefficiencies in the market.

[42] These data come from Author Earnings, see `http://www.authorearnings.com`.

References

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. (2015). "Revenue, new products, and the evolution of music quality since napster." Tech. rep., Institute of Prospective Technological Studies, Joint Research Centre.

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.

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

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*.

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.

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. (2012). "Copyright protection, technological change, and the quality of new products: Evidence from recorded music since napster." *Journal of Law and Economics*, *55*(4), 715–740.

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 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
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	6.509	6.016	3.712	5.163	3.489	5.061	2.649	4.146	3.693	5.103
Error^2 (incl.)	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	0.088	0.283	0.075	0.264	0.028	0.166	0.038	0.190	0.039	0.194
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, potential data issues

	(1) Size reported	(2) < 50	(3) 50–99	(4) 100–249	(5) 250–500	(6) > 500
After × Romance	0.023	-0.039**	0.003	0.019**	0.005	0.035***
	(0.019)	(0.017)	(0.007)	(0.008)	(0.004)	(0.008)
Observations	52260	52260	52260	52260	52260	52260
$\overline{R^2}$	0.207	0.287	0.026	0.028	0.010	0.075

Dependent variable: Column (1): dealsize reported (0/1), columns (2)–(6): deal size category (0/1).
Notes: *After* indicates years after 2008. Editor, month, and year fixed effects.
Lower-level interactions and constant included but not reported.
Standard errors in parentheses, clustered on the editor-level.
* $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table A.3: Results: Changes in license deals, ordered logit

	(1) DV: Deal category			
	50_99	100_249	250_500	_500
After × Romance	-0.050	0.390	0.909*	0.768*
	(0.375)	(0.335)	(0.511)	(0.397)
Observations	14771			

Notes: Group-specific time trends, month and year fixed effects, and constant not reported.
Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table A.4: Results: Changes in predicting ex-post appeal, ordered logit

	(1) Error^2		(2) Error^2 All deals	
1				
Romance	0.571**	(0.280)	-0.731***	(0.162)
After × Romance	-0.766**	(0.346)	-0.206	(0.219)
4				
Romance	1.193***	(0.355)	0.287	(0.222)
After × Romance	-0.898**	(0.439)	-0.489*	(0.259)
9				
Romance	1.528***	(0.332)	0.327**	(0.153)
After × Romance	-0.465	(0.453)	-0.611***	(0.222)
16				
Romance	1.986***	(0.355)	1.226***	(0.218)
After × Romance	-0.488	(0.563)	-0.823***	(0.275)
Observations	1569		14771	
$\overline{R^2}$				

Dependent variable: Linear measure of squared 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 5.3.1). Column (1) includes only observations where the author appears at least on the top 100 bestseller list at some point after the deal. Column (2) also includes observations where the author does not appear on the beststeller list. *After* indicates years after 2008. Editor, month, and year fixed effects. Lower-level interactions and constant included but not reported. Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table A.5: Results: Changes in predicting ex-post appeal, robustness

	(1) In Top 100		(2) In Top 100		(3) In Top 100	
After × Romance	0.009***	(0.003)	0.047	(0.207)	-3.711**	(1.784)
After × Romance × Deal Size Reported	-0.020	(0.017)				
After × Romance × Log(Deal Size)			-0.005	(0.019)	0.409**	(0.196)
After × Romance × Log(Deal Size)^2					-0.043**	(0.020)
Observations	52260		14771		14771	
$\overline{R^2}$	0.068		0.058		0.069	

Notes: Lower-level interactions and constant included but not reported. Editor, month and year fixed effects. Standard errors in parentheses, clustered on the editor-level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$