The Impact of the Agency Model on E-book Prices: Evidence from the UK

Maximilian Maurice Gail* Phil-Adrian Klotz*†

May 24, 2021

Abstract

This paper empirically analyzes the effect of the widely used agency model on retail prices of e-books sold in United Kingdom. Using an unique cross-sectional data set of e-book prices for a large number of book titles across all major publishing houses, we exploit cross-genre and cross-publisher variation to identify the mean effect of the agency model on e-book prices. Since the genre information is ambiguous and even missing for some titles in our original data set, we also apply a Latent Dirichlet Allocation (LDA) approach to determine detailed book genres based on the book's descriptions. We find that e-book prices for titles sold under the agency model are on average 36% cheaper compared to titles sold under the wholesale model. Our results are robust to different specifications, the Lewbel instrumental variable approach, as well as machine learning techniques.

Keywords: e-books, agency, resale price maintenance, Amazon, double machine learning, Latent Dirichlet allocation

JEL Classification: D12, D22, L42, L81, L82, Z11

*Chair for Industrial Organization, Regulation and Antitrust, Department of Economics, Justus Liebig University Giessen. Licher Strasse 62, 35394 Giessen, Germany. E-mail: maximilian.m.gail@wirtschaft.uni-giessen.de. We would like to thank Georg Götz, Daniel Herold, Jan Thomas Schäfer, Jona Stinner and Xiang Hui for the fruitful discussions.

†Corresponding author. E-mail: phil.a.klotz@wirtschaft.uni-giessen.de.

1 Introduction

Over the last years, the book industry has remained a massive, greatly influential global market. In the US alone, the book publishing industry generated an annual revenue of $25.93 billion in 2019, having sold more than 2.7 billion books.[1] For comparison, total revenue in the United Kingdom (UK) amounted to £2.95 billion in 2018.[2] Simultaneously, the digitization stimulates the competition in the book market through the established retail channel e-Commerce and particularly the book format e-book. In the US book market, e-Books accounted for 7.5% of the overall revenues in 2019 [3], while in the UK the share of digital audio books and e-books even belonged to 20% in 2018.[4]

Books are experience goods because readers can ascertain the quality only after reading a given book (Nelson 1970; Reimers and Waldfogel 2020). Besides, in some countries such as Germany, France or Japan, book prices are fixed whereas countries without fixed book price systems include the UK and the USA.[5] Fixed book prices are a form of resale price maintenance (RPM), where publishers set retail prices and price competition between retailers is restricted or eliminated. Mostly the motivation behind the introduction of fixed book price systems is the assurance for a broad and diverse supply of books, available through a geographically wide network of bookstores.

[1] see https://bit.ly/34hIY9l (last accessed March 16, 2021).
[2] see https://bit.ly/2YhByPw (last accessed March 16, 2021).
[3] see footnote 1.
[4] see footnote 2.
[5] Presently, 15 OECD countries have regulation for fixing the prices of printed books. The fixed price for printed books typically lasts 18-24 months after a book has been published.

With the advent of e-books, countries with a fixed price system for print books had to decide whether to extend existing legislation to e-books. It is questionable whether the same cultural policy arguments and legal considerations apply, in particular because a geographically wide network of bookstores is irrelevant for e-books. Nevertheless, eight OECD countries with a fixed price for printed books also have a fixed price for e-books, while no country is known to have a RPM regulation for e-books but not for print books (Poort and van Eijk 2017).

However, in many countries without fixed prices for e-books (such as the UK) this digital product is partly sold under a so-called agency model, which is also used for many other digital products (e.g., smartphone apps) (Gilbert 2015). The agency model gives publishers the ability to directly set retail prices. Thereby, the retailer merely acts as an agent for the publisher and receives a commission for every e-book sold. Hence, the agency model has similar effects as RPM between a manufacturer and a retailer.[6] In contrast, under the wholesale model publishers sell e-books to retailers at a wholesale price and retailers set the retail price at which they sell e-books to the consumers.

With regard to e-books, Apple in co-operation with the six largest publishing multinationals have been the first who adopted the agency model in 2010 in response to Amazon's aggressive pricing strategy to gain market share. In April 2012, the Department of Justice (DOJ) sued Apple and five of the six

[6]They are economically similar in the sense that the upstream firms control the retail price. However, a major difference is that for the agency pricing the downstream firms individually delegate retail pricing to the upstream firms, whereas under the classical case RPM is imposed at the level of the market for any given good.

publishing houses for conspiring to raise e-book prices by using the agency model in conjunction with most-favored nation (MFN) clauses (which prohibited publishers from selling their e-books at higher retail prices on Apple's iBookstore than they sold for elsewhere).[7] Three of the publishers settled shortly after the antitrust case was filed, while the other two followed later the same year, which meant that the five publishers could not restrict a retailer's ability to set e-book prices for a period of two years.

Empirical evidence on the price effects of RPM and fixed book prices as well as of the agency model is scarce. While systematic empirical evidence on RPM is limited to case studies (see MacKay and Smith 2017; Ippolito 1991), the only study investigating the empirical effect of the agency model is the one from De los Santos and Wildenbeest (2017). They used data on e-book prices of bestselling book titles for the years 2012 and 2013 and the Apple case as an exogenous shock to show that the agency model in combination with MFN clauses led to an average increase in prices between 8-18% (depending on the retailer).

The goal of this study is to analyze the price effect of the agency model using a larger and more detailed data set (especially not only incorporating bestselling book titles) to check whether similar effects also occur absent a court decision as in the Apple case. Our cross-sectional data set contains prices for 10,048 e-books published on Amazon UK between March, 2010 and March, 2020.[8] Using data from Amazon ensures a high market coverage, since Amazon accounted for 50 percent of UK sales of physical and digital books in

[7] See United States v. Apple Inc., 12 Civ. 2826 (DLC).

[8] There are some older books in our data. Nevertheless, most of the books are from 2018-2020.

2018.[9] We further use publisher and book genre variation to estimate the effect of digital books sold under the agency sales model on the price of an e-book.

However, the relationship of an e-book price and one important explanatory variable, the sales rank, may be bi-directional, since the sales rank of a book indicates its sold quantity. To account for resulting endogeneity, we employ an instrumental variable approach using the number of consumer reviews on an e-book as an instrumental variable. Moreover, we apply the heteroscedasticity based IV approach as suggested by Lewbel (2012) and various Double Machine Learning (DML) approaches as robustness checks for potential endogeneity.

Our findings indicate that e-books sold under the agency model on Amazon.co.uk are on average 36.4% cheaper than digital books sold under the wholesale model. The Lewbel approach and various DML concepts support our main finding. This result contradicts the empirical outcome from De los Santos and Wildenbeest (2017), but fits into explanations put forward by the theoretical literature on agency versus wholesale models. In particular, it supports the study from Johnson (2020) who finds that even though prices may initially be higher under the agency model, consumers are likely to be worse off in the long run under the wholesale model than under the agency model. This is based on the pricing behavior of retailers who use the wholesale model to initially set low prices and lock in consumers, but find it optimal to raise prices in the long run after having locked in a sufficient number of

[9] See Nielsen (2018), "Books & Consumers - UK Industry Standard Report Q4 2018", p. 13.

readers.

The rest of the paper is structured as follows. In Section 2, we describe the related literature. We present our unique data set in Section 3.1. Descriptive statistics are given in Section 3.2 and our text mining approach to determine book genres is explained in Section 3.3. Section 4 presents our main estimation strategy and results. In Section 5, our robustness checks are outlined. In Section 6, we conclude and outline the contributions of our paper.

2 Related Literature

Our article contributes to several strands of literature. First and foremost, it is related to studies which investigate the competitive effects of the agency model. While the empirical literature on the economic effects of the agency model is rather scarce (an exception is the study of De los Santos and Wildenbeest 2017), several recent theoretical papers have analyzed differences in retail prices between the agency and the wholesale model. One strand of this literature is focused on a lock-in effect of consumers in this context. Johnson (2020) finds that when publishers set retail prices instead of retailers, prices may be higher in early periods but lower in later periods suggesting that retailers will initially set low prices to lock in consumers, but find it optimal to raise prices once a sufficient number of consumers are locked in.

Another strand of the theoretical literature on agency models assumes that complementary devices are necessary for the enjoyment of the main products (e.g., an e-book reader in the case of e-books). Gaudin and White (2014) point out that the incentive of a retailer to set high prices is higher when

she has monopolistic control over a complementary device, as it was the case in the e-book market when e-books from Amazon could only be read on a Kindle device. In another model-theoretical setup, Abhishek et al. (2016) show that agency selling is more efficient than the wholesale model and leads to lower retail prices, although retail prices may be higher under the agency model if there are positive externalities from sales of associated products (such as e-readers in the case of e-books).

Foros et al. (2017) show that the agency model is always anti-competitive (leads to higher retail prices) when it is adopted by the platforms on a market-by-market basis. To be more specific, they find that upstream firms (publishers) will set higher retail prices than downstream firms (retailers) would set if they were in control as long as competition is greater among retailers than among publishers. Moreover, they point out that a retailer who sets retail prices independently (wholesale model) benefits when a horizontal rival is restricted by the agency model, since the latter creates a price umbrella which makes it profitable for the independent price-setting retailer to increase prices. Condorelli et al. (2018) present a theory that makes the decision whether to use agency or wholesale models endogenously in an environment where the retailer has privileged information about the valuations of consumers and show that retailers prefer the agency model.

Our article also contributes to the literature on MFN clauses. Gans (2012) examines the pricing of mobile applications on platforms and finds that a hold up problem may arise if consumers have to purchase a device to access the platform. Nevertheless, he shows that restrictive conditions on application providers, such as MFN clauses, may help overcome this problem. In

another study from Boik and Corts (2016), it is shown that the adoption of MFN clauses by retailers can lead to higher retail prices when there is competition both upstream and downstream. Shaffer (2012) finds that it does not matter for this conclusion whether the MFN clauses are adopted by the downstream or upstream firms. Beyond, Boik and Corts (2016) also point out that MFN clauses can eliminate the retailer's incentives to compete in revenue shares. Retail prices rise when the upstream firms are in charge of setting prices because with strictly positive marginal costs the smaller revenue shares received by these firms act like a marginal tax increase.

Third, our analysis is also related to studies examining the substitutability between e-books and physical books. Using sales data from Amazon, Chen et al. (2019) find that delaying e-book availability results in a 43.8% decrease in e-book sales but no increase in print book sales. Crosby (2019) uses a market segmentation approach to show that there are three different class of readers and only one of them, the 'technological adopters', substitute the printed books by e-books. The other two groups of readers (accounting for over half of the market) are steadfast in their preferences for traditional formats. Overall, these results suggest a fundamental lack of substitutability between e-books and printed books.[10]

More generally, our article contributes to the broader literature on RPM. RPM can lead to lower retail prices because of the internalization of vertical externalities such as double marginalization. Since a manufacturer chooses the wholesale price given its costs and the retailer as a result chooses the

[10] A closely related topic is the substitutability between the online and offline sales channels, which has also been examined for the book market by previous studies (see, e.g., Brynjolfsson et al. (2009) or Goetz et al. (2020)).

retail price given the wholesale price, both firms will add their mark-up and in the end the consumers will pay too high prices. An obvious possibility to solve this problem is RPM because the manufacturer simply imposes the resale price on the retailer (see Spengler 1950 and Tirole 1988). Moreover, Telser (1960) and Yamey (1954) were the first who noted that strong intra-brand competition can be detrimental to retailer's incentives to invest in free-rideable services. Suppose a situation in which a retailer benefits from the service provision of a competitor. In these circumstances, a firm will think twice before investing in services because the competitor would have an incentive to avoid the cost of this effort, free ride on the provision of services and offer a better price. Many authors have shown that RPM can be used to correct for service externalities (see Mathewson and Winter 1984, Perry and Porter 1986 and Winter 1993). Besides, Dearnley and Feather (2002) and Davies et al. (2004) find that there is a larger number of brick-and-mortar (B&M) stores in regimes with RPM compared to regimes with free prices. However, Rey and Stiglitz (1988) and Rey and Stiglitz (1994) point out that vertical restraints that eliminate intra-brand competition can also be used to mitigate inter-brand competition and then would be anti-competitive.

Fifth, our study is related to existing work examining the impact of professional reviews and of word of mouth reviews on product sales. There are a couple of studies investigating the impact of professional reviews on movie and book sales (e.g., Reinstein and Snyder 2005, Sorensen 2007, Berger et al. 2010, Garthwaite 2014 and Reimers and Waldfogel 2020). For instance, Berger et al. (2010) show that a negative review in the *New York Times* hurts sales of books by well-known authors, but increases sales of books that had

lower prior awareness. Prominent examples studying the impact of word of mouth reviews (=recommendation networks) include Chevalier and Mayzlin (2006), Helmers et al. (2019), Oestreicher-Singer and Sundararajan (2012a), Oestreicher-Singer and Sundararajan (2012b) and Reimers and Waldfogel (2020). Chevalier and Mayzlin (2006) make use of cross-platform comparison of book sales ranks and star ratings to show that books with a higher average star rating also have a higher market share. Oestreicher-Singer and Sundararajan (2012a) find that customer reviews can shift demand towards niche titles and the same authors show that recommendations can increase demand for new titles (2012b). In another study, Reimers and Waldfogel (2020) indicate that the effect of star ratings on consumer surplus is roughly 15 times the effect of professional reviews.

Finally, our article also contributes to the newer literature on machine learning and text mining approaches. Varian (2014) and Athey and Imbens (2019) provide an overview of important ML methods. Wang et al. (2019) uses the *Learning to Place* ML approach to predict book sales and find that a strong driving factor of book sales across all genres is the publishing house. For a broad overview on text mining approaches see Gentzkow et al. (2019). In our article, we use a latent Dirichlet allocation (LDA) model to find book topics by analyzing the description of a book (see, e.g., Larsen and Thorsrud 2019).

3 Data

Our data set contains prices of e-books for a large number of titles. In this section, we will present this data set for our empirical analysis. We first

describe the construction of our data set in Section 3.1 for which we derive descriptive statistics including information on prices, ratings, reviews and the digital size (in KB) of e-books in Section 3.2. In Section 3.3, we present our LDA text mining approach to derive book genres based on the descriptions of books.

3.1 Data Set Construction

The data generating process is structured as follows. We have scraped the Amazon.co.uk webpage for a list of publisher and imprint names for the year 2019 starting mid February 2020 taking two weeks to get e-book prices as well as further book characteristics available on the website. Thus, we use the methods of web-scraping to generate a cross-sectional data set. For creating this data set, we use *a priori* a list of publishing houses, publishers and imprints which is taken from a historical *Sunday Times* bestseller list. This procedure ensures that our sample only contains books from publishers with a relatively high market size.[11]

With this list of publishers we have first searched on Amazon.co.uk for e-books published between January, 2019 and March, 2020. After that, we have updated the original list, since the respective publisher names on Amazon are often written in different ways but the search itself must match every character exactly. Following this, the data set with publisher names has risen. With this newly obtained list, the actual data set is constructed on the above mentioned time frame by searching for all publisher names of the

[11]The used bestseller list contains entries from January, 2006 until the end of March, 2019.

updated list.[12] This proceeding also incorporates books into our data set which have been published before 2019 because we have done the publisher search on Amazon.co.uk independently of the format. Thus, it may have happened that for a book title, which we have found within our observation period, another format of the same title has been published years before. However, we have ensured that no book is included in our working data set which has been published before March, 2010.

All book titles on Amazon.co.uk have different ASINs (Amazon Standard Identification Number), which leads to an identification problem for different format types of the same book. To solve this problem, we first extract as much data from the webpage of a book as possible. Then, we open the list of all formats from the observed book available on Amazon and collect the information about them, which only contain their prices. Unfortunately, in some cases different editions of a book title are sold by different publishers, which enforces us to be even more precise for the identification of those books. Therefore, we take the newest, oldest and cheapest version of the format list and scrap detailed information for them. To prevent data failures, the ASINs of those three versions are combined to create a unique identifier. Finally, within the data conversion we only use the newest book of each format when there are various editions available.[13]

Our raw data set consists of roughly one million row entries, whereby each entry contains several information on different prices, formats, descriptions,

[12] In the end, we will have around 3000 entries of publisher/imprint names, which will collapse to six big publishing houses and one category of small/independent publishers we could not attribute to one of the large publishing houses.

[13] Books with versions have been sold after March 3, 2020 are not considered.

ratings, reviews etc. being available on the Amazon website. Besides, for every book title there are three entries if all formats (hardcover, paperback, e-book) are available for the respective book title. However, due to the usage of web-scraping methods the data set contains of some entries that are duplicates or not of interest for our analysis. Hence, after the data cleansing process our working data set consists of $77,629$ entries, respectively $47,161$ unique books. In the former number, all three format types (hardcover, paperback, and e-book) are included. For our empirical analysis, we drop the hardcover and paperback book titles (see Section 4). Besides, we only use e-book observations in our estimation approach for which all explanatory variables (book characteristics) are available. Thus, for our empirical analysis 10,048 e-book titles remain in the final working data set.

Our variables of interest are the retail price, which is the price a consumer has to pay for a respective book, and the dummy variable *Agency*, which takes the value one if there is a text-field on the Amazon web-page of a book title saying *'This price was set by the publisher'* and zero otherwise.[14] Moreover, we have data on several control variables for our empirical analysis. These variables comprise book characteristics as the book format, the book genre, the number of pages/ size of the e-book in KB, variables on book reviews as the star rating and the number of consumer and expert reviews of a book, variables containing information on the author or publisher of a book title and other variables as the publication date or the recommended retail price (RRP). Table 1 summarizes the descriptions for all variables included in our data set.

[14] See Figure 8 in Appendix A for an example.

Variables	Information
Price	Retail price from the upper right *Buy-Box*
Format	Hardcover, paperback, Kindle
Star rating	Average rating normalized to be between 0 and 1
No. customer reviews	Number of consumer reviews
No. expert reviews	Number of expert reviews on Amazon
Series	Dummy variable whether book is part of a series
Description and reviews	Detailed text-information on the book and by different reviewers
Genre	Constructed by LDA from the descriptions and reviews (see Section 3.3)
RRP	Recommended retail price which is the print RRP. For Kindle it is either related to the hardcover or paperback RRP
Agency	Dummy variable to be one if the price was set by the publisher and zero otherwise. Only possible for e-books
Seller	Sample is restricted to be sold by Amazon
Author	Information on the author of a book
Title	Information on the title of a book
Kindle_Size	Kindle file size (in KB)
Pages	Number of pages in the print format
Publisher	Name of the publisher. We have different levels of aggregation (Imprint,Publisher,Publishing House)
Amazon rank	Uncategorized Amazon bestseller rank for either print books or e-books
Bestsellers	Number of bestsellers in the Sunday Times Bestseller List conditional on the Author's name
WeekInChart	Average number of weeks in the bestseller charts conditional on the Author's name
Identifier	Aggregation of ASINs to verify the books
Publication Date	Publication Date of a book title
Other	Other small control variables

Table 1: Relevant Variables per book and the information content they provide.

We have also matched the data set obtained from Amazon with a historical *Sunday Times* bestseller list to identify authors who have already written a bestselling book in the past. This variable is important for our empirical analysis, in which we estimate the price on an e-book, since the name of a bestseller author is an important quality signal for the book readers.

Each book is a unique product written by an author and mostly published by one publisher. Thus, books are heterogeneous goods which make it impossible to actually compare the value of one specific book with one another. In order to provide an acceptable analysis, it is therefore also necessary to control for the genres of the several books. Hence, we use a Latent Dirichlet Allocation (LDA) to determine book genres from the descriptions and reviews of the individual books available on the Amazon web-page. This control variable should be able to capture specific effects within the individual genres. In Section 3.3, this text mining approach will be explained in more detail.

3.2 Descriptive Statistics

As already mentioned above, our working data set contains 47,161 book titles. Our sample consists of titles that have been published on Amazon.co.uk by the publishers Bloomsbury, Hachette, Harper Collins, Pan Macmillan, Penguin Random House, Simon & Schuster, and smaller/independent publishers between March, 2010 and March, 2020. However, in overall there are 77,629 observations in our data set, since there are several book formats available for some titles. Even though the focus of our empirical analysis is on the price of e-books, in this section we also present some descriptive statistics for the book formats hardcover and paperback to show the relation between those three book formats.

	Publisher	Bloomsbury	Hachette	Harper Collins	indie/small	Pan Macmillan	Penguin Random House	Simon & Schuster
Retail Price	mean	10.32	5.37	6.03	5.94	7.89	6.49	7.78
	std	6.77	2.88	3.11	6.14	3.94	2.60	3.11
Sales Rank	mean	557530.94	248169.45	454336.31	750943.38	491780.02	323492.44	556914.72
	std	604677.41	378211.81	537049.55	717915.76	545185.33	478824.90	667546.00
Star Rating	mean	0.90	0.89	0.89	0.86	0.89	0.88	0.90
	std	0.10	0.08	0.08	0.12	0.09	0.08	0.08
No. Customer Reviews	mean	49.94	121.56	126.52	79.24	118.07	134.22	123.09
	std	111.99	179.90	180.46	158.39	201.39	198.28	200.45
Pages	mean	293.20	370.43	335.89	301.95	336.41	324.64	333.61
	std	127.89	1714.53	135.05	783.93	118.82	139.96	127.21
Kindle Size	mean	13267.91	14728.26	9275.93	9879.06	12432.30	16643.93	16657.32
	std	27466.70	50569.89	24888.57	34405.40	39533.77	36918.61	28966.55
RRP	mean	15.49	13.06	13.38	12.56	13.81	14.45	14.63
	std	9.76	5.22	5.51	11.65	5.73	5.74	5.12
Date Retail	mean	1.53	1.68	2.01	1.89	1.56	2.09	1.92
	std	1.59	1.88	2.26	1.86	1.65	2.13	1.95
No. Expert Reviews	mean	1.74	2.31	1.43	1.38	2.29	1.70	1.31
	std	0.97	1.02	1.08	1.51	1.59	1.38	0.90

Table 2: Summary Statistics

Table 2 gives descriptive statistics for the variables we use for our empirical analysis, summarized by publishers. In addition to e-book retail prices and RRP, we observe several characteristics for each title, such as the rank of a title on Amazon, the customer ratings, the number of customer and expert reviews, and the number of pages. As shown in the table, e-books from Hachette exhibit the lowest average price, while the e-books from Bloomsbury

have the highest average prices. Beyond, the titles from Hachette also have the lowest average book rank and the book titles published by Penguin Random House exhibit the highest average number of customer reviews. Most of the other book characteristics are very similar across publishers.

The several publishers use different sales models for their e-books. While the major publishing houses all have adopted the agency model, Bloomsbury and the smaller/independent publishers still use the wholesale model. Amazon mentions on its product pages whether it or a publisher has set the price of a particular e-book. The Figures 8 and 9 in the Appendix A present examples of this by showing screenshots for the books *Elon Musk: How the Billionaire CEO of SpaceX and Tesla is Shaping our Future* as well as *Pulse*. In Figure 8, it can be seen in the first box on the right hand side of the Amazon webpage that the 'price was set by the publisher', so that this is an example for the application of the agency model. On the contrary, in Figure 9 this information is missing, which means that Amazon sets the retail price for this e-book and it is an example for the wholesale model.

Figure 1 shows the frequency distribution of the retail prices for e-books (top), paperbacks (centre) and hardcover books (bottom) below £100. It is obvious that e-book prices are in a range between £0 and £10, paperback prices concentrate mostly in the £10-£20 interval and hardcover prices are generally higher. Furthermore, while the distributions of e-books and paperbacks are more compressed, the hardcover prices exhibit a higher volatility. And third, all three distributions have significant mass points at candidate focal points (e.g., £0.49 (e-books), £9.99 (paperback) and £15.99 (hardcover)).

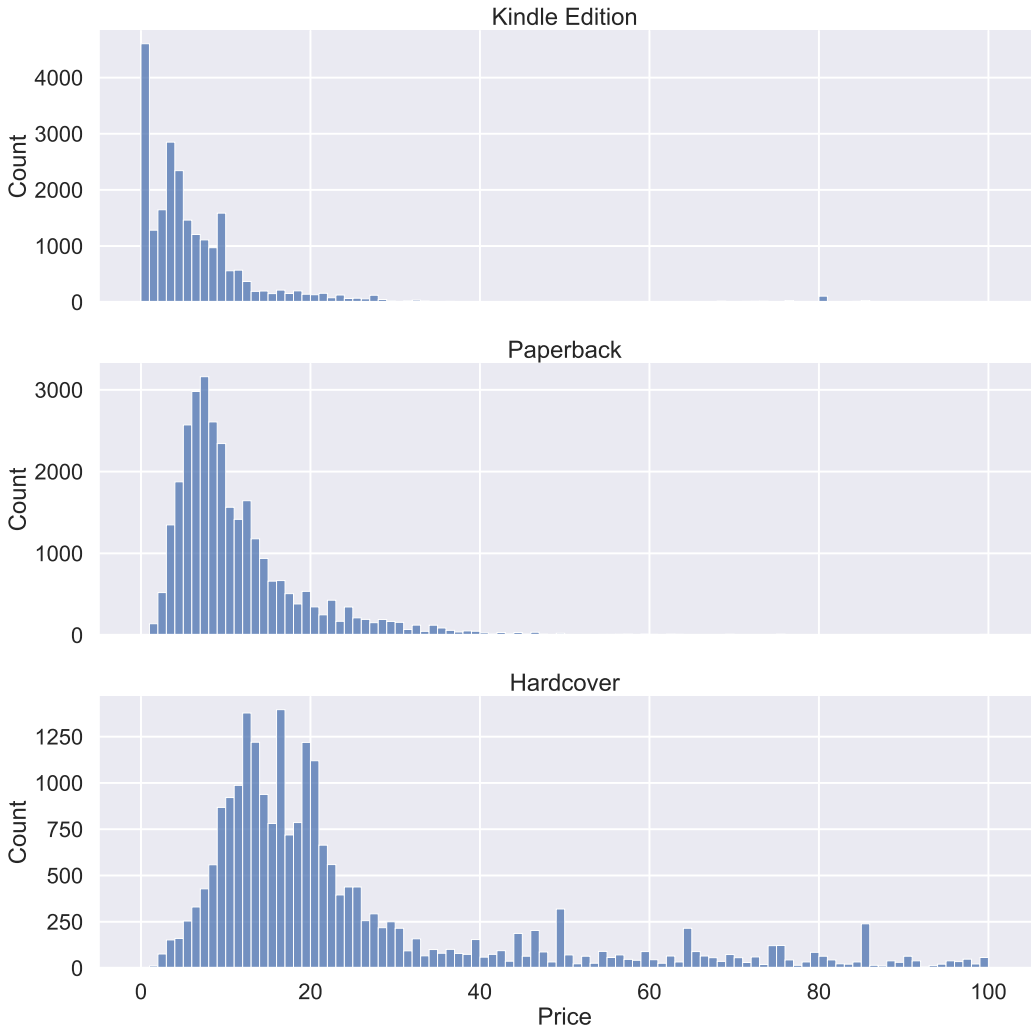

Figure 1: Distribution of 77,629 retail prices by book format.

Table 3 presents the descriptive statistics for Figure 1. As expected, with an average price of £8.4, e-books are the cheapest of the three book formats, followed by paperbacks (£12.34) and hardcover books (£28.42). The high standard deviation for hardcover books confirms its high volatility, which we have already detected in Figure 1. In overall, the descriptive statistics on book formats prove that hardcover books exhibit the highest quality of the three formats and confirm the results from Li (2019), who has found that

e-books and paperbacks are close substitutes.

Book Format	Observations	Mean	Std. Dev.	Min	25%	50%	75%	Max
Hardcover	23,307	28.42	41.57	1.05	12.99	18.95	28.78	1575.0
e-book	23,495	8.4	14.02	0.25	1.99	4.99	9.3	467.8
Paperback	30,827	12.34	18.86	0.31	6.55	9.09	13.95	1899.0

Table 3: Prices grouped by book format.

Table 4 uses one further level to highlight the importance of different publishers. Here the distribution of our data can be seen. The data set contains books from major British publishers (Bloomsbury, Hachette, HarperCollins, Pan Macmillan, Penguin, Simon & Schuster) and one last category called Indie/Small, which contains all smaller or independent publishers in the data set. The third column in Table 4 shows that a large share of the book titles in our data set belongs to the category Indie/Small. Furthermore, it can be seen that the book titles published by Bloomsbury and smaller/independent publishers are on average significantly more expensive than the titles of the other publishers for the formats hardcover and paperback, but concerning the format e-book only the titles from Bloomsbury are significantly more expensive.

Figure 2 visualizes the statistics from Table 4 for the format e-books even further. Prices are obviously more dispersed for book titles published by Bloomsbury, whereas the other major publishers mostly have books in the range up to £20. The Indie/Small category has a significantly larger fraction of e-Books in the cheapest price range (about £0.49). In Section 3.3, we will present this figure combined with assigned book genres.

Book Format	Publisher	Observations	Mean	Std. Dev.	Min	25%	50%	75%	Max
Hardcover	Bloomsbury	1,074	61.08	37.97	3.63	20.0	69.2	85.0	395.0
	Hachette	1,161	14.27	5.04	2.03	11.12	13.59	16.99	85.56
	Harper Collins	1,376	14.83	5.92	1.63	10.99	14.17	16.99	117.01
	Indie/Small	16,810	30.85	46.4	1.05	14.32	20.0	33.27	1575.0
	Pan Macmillan	711	14.02	5.62	1.73	10.14	13.21	18.16	44.6
	Penguin Random House	1,592	14.51	7.94	1.79	10.65	13.26	16.99	191.16
	Simon & Schuster	583	13.80	6.73	2.03	10.27	12.99	15.25	99.35
e-book	Bloomsbury	1,173	31.88	29.13	1.42	9.98	19.94	56.62	123.5
	Hachette	2,514	5.39	2.96	0.99	3.99	3.99	5.99	27.99
	Harper Collins	2,091	5.90	3.70	0.75	2.99	5.49	6.99	62.58
	Indie/Small	13,670	7.62	14.38	0.25	0.99	3.79	8.50	467.80
	Pan Macmillan	855	7.57	3.76	0.99	4.91	6.99	9.99	32.30
	Penguin Random House	2,611	7.21	3.48	0.49	4.99	6.52	9.49	40.0
	Simon & Schuster	581	8.01	3.12	0.99	4.99	8.49	9.99	20.99
Paperback	Bloomsbury	769	17.37	8.89	1.0	8.99	16.99	25.2	57.25
	Hachette	1,648	8.13	2.84	1.0	6.55	7.37	8.99	31.0
	Harper Collins	1,945	9.43	12.3	1.0	5.39	7.35	9.99	192.0
	Indie/Small	23,171	13.23	21.18	0.31	6.79	9.99	14.99	1899.0
	Pan Macmillan	628	9.09	3.52	2.58	6.55	8.19	11.90	30.99
	Penguin Random House	2,100	9.27	4.83	1.0	6.55	8.19	11.15	76.47
	Simon & Schuster	566	9.03	7.69	1.55	6.18	7.99	10.97	167.39

Table 4: Prices grouped by format and publisher.

As already mentioned above, the e-books on Amazon.co.uk are sold under different sales models, mostly depending on the publisher. Table 5 illustrates those sales models for all publisher in our sample. If the number in the column 'Agency' takes the value one, the respective book titles are sold under the agency model, otherwise the title is sold under the wholesale model. The table shows that e-books published by Harper Collins and Simon&Schuster are sold under the agency model on Amazon. For the other three of the five big publishers (Hachette, Pan Macmillan, and Penguin Random House), most of the titles are sold under the agency model. However, some of the e-books published by them are classified to be sold under the wholesale model, even though these publishers have an agreement with Amazon to sell e-books under the agency model. This phenomena either is due to a database-error, falsely classified imprints or some imprints of the big

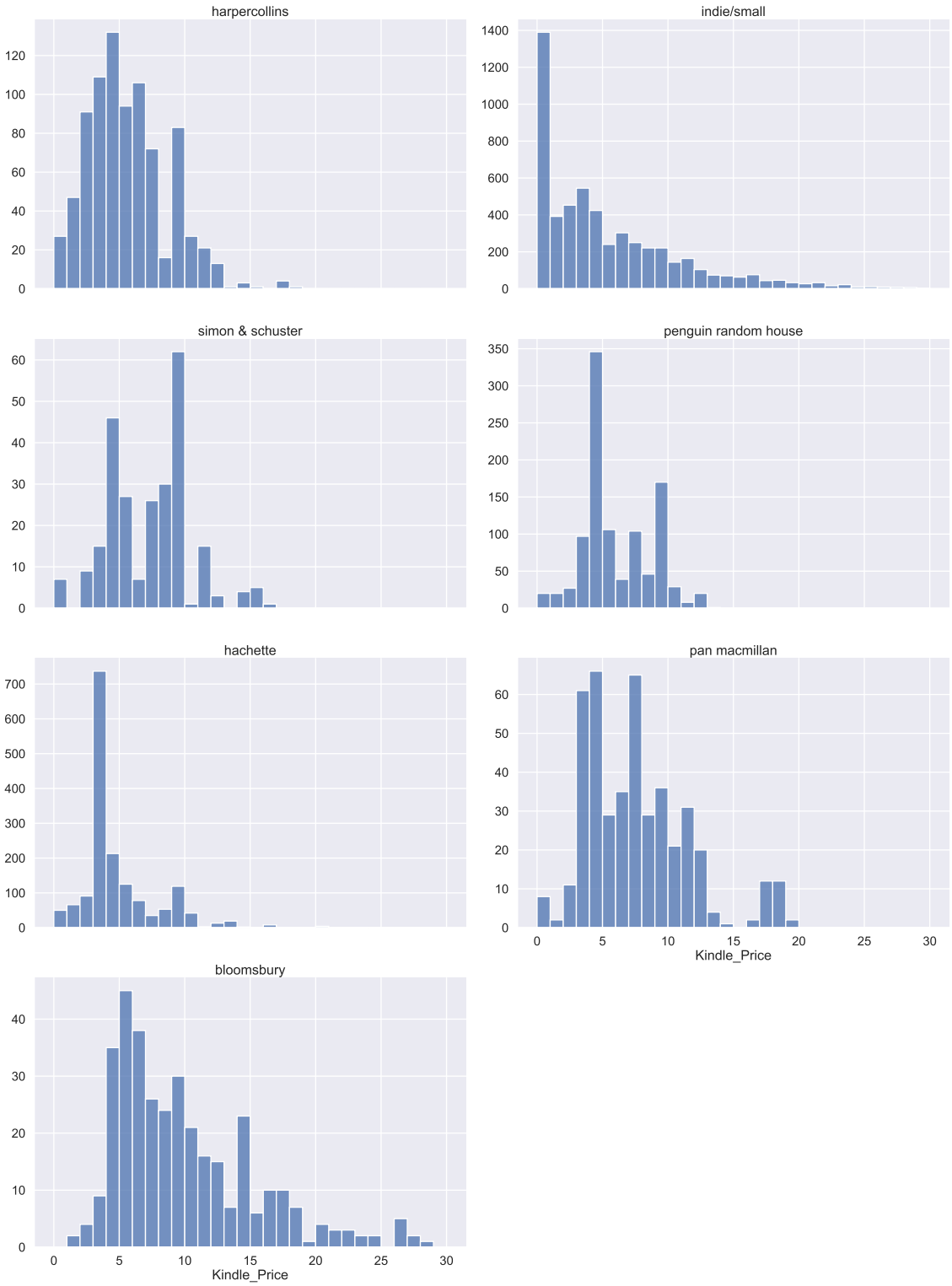

Figure 2: Prices for e-books grouped by publishers. The interval size for each bar is 1 Pound. For illustration purposes the figures are censored at 30 Pound.

publishers actually do not use the agency model. For example, Pan Macmillan has some American imprints, which are not sold under the agency regime. The smaller/independent publishers use the wholesale and the agency model to sell e-books on Amazon.co.uk.

Publisher	Agency	Distribution	Split
Bloomsbury	0	353	1.0000
Hachette	1	1659	0.9994
	0	1	0.0006
Harpercollins	1	848	1.0000
Indie/small	0	3869	0.7102
	1	1579	0.2898
Pan Macmillan	0	233	0.5201
	1	215	0.4799
Penguin Random House	1	1020	0.9874
	0	13	0.0126
Simon & Schuster	1	258	1.0000

Table 5: Distribution of the agency variable by publishers.

Finally, we want to illustrate the relationship between retail prices for e-books and their book sales rank on Amazon, which is illustrated in Figure 3 by using a scatter-plot with a simple regression line. Obviously, there is a positive relation between the retail price and the rank of an e-book in our sample, since the regression line has a positive slope. This finding in our sample is in line with the study of Fishwick (2008), who states that 'substantial discounts' (p. 370) have become prominent for bestselling books in the British book market after the abandonment of the Net Book Agreement in 1997. Regarding the book rank, we have to stress how the rank on Amazon is determined. According to sources of Amazon, the ranks are internally updated hourly, but it does not appear immediately. The rank includes current

and *all* past sales with higher weights on current sales.[15] This information on the definition of book ranks on Amazon will be important for our estimation approach because the price today might be affected by current sales or past sales, but not necessarily vice versa.

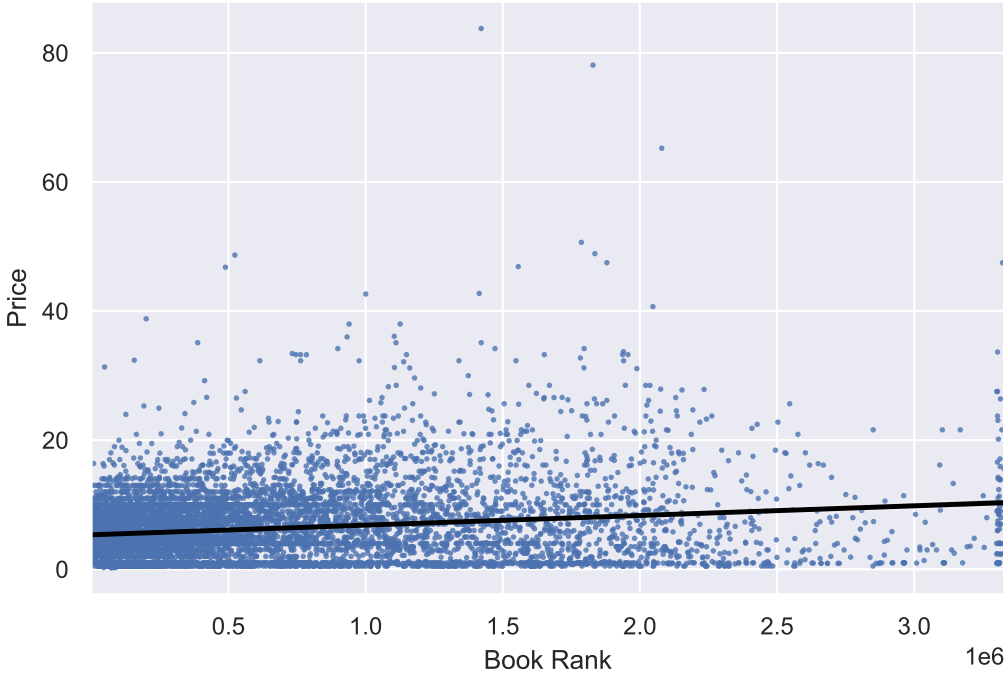

Figure 3: Relation of Amazon book ranks for e-books and retail prices.

3.3 Latent Dirichlet Allocation (LDA)

In the recent past, new technologies have made it possible to use text as data and, therefore, as an input to economic research. Text data, which is inherently high-dimensional, can capture relevant economic concepts not covered by "hard" economic data. In the last years, there has been an explosion of empirical economics research using text as data (e.g., see Larsen and

[15] See https://kdp.amazon.com/en_US/help/topic/G201648140.

Thorsrud (2019) for an Latent Dirichlet Allocation (LDA) approach or Lenz and Winker (2020) for paragraph vector topic modelling). We have decided to use an LDA approach to generate book genres and to assign every single title from our data set into one of these genres. Such a text mining approach is necessary because on the Amazon webpage the genre information is ambiguous and even not available for some book titles. For this purpose, we use the descriptions from the individual books in our data set as text data input. We further rely on natural language processing (NLP) to extract the relevant information.

We apply several Python-Modules to clean and prepare the raw data set.[16] Thereby, we remove common words and surnames, eliminate stop words, remove punctuation and pronouns as well as reduce all words to their respective word stems. We note here that around 45,819 unique tokens are kept after this filtering process.

This cleaned descriptions corpus is decomposed into book genres using the already mentioned LDA model. The LDA provides a statistical framework for the generation of documents based on topics. It is an unsupervised topic model that clusters words into topics/ genres, which are distributions over words, while at the same time classifying descriptions as mixtures of topics/ genres. The term "latent" is used because the words are intended to communicate a latent structure, namely, the subject matter (topic) of the description. The term "Dirichlet" is used because the topic mixture is drawn from a conjugate Dirichlet prior (see Thorsrud 2020).

[16]Base module is gensim (Řehůřek & Sojka, 2010) with a wrapper called Mallet, which is a Java-based open-source NLP text analytics tool (see McCallum (2002) or http://mallet.cs.umass.edu/).

The structure of the LDA model is as follows: the whole corpus is represented by M distinct documents (descriptions) and $N = \sum_{m=1}^{M} N_m$ is the total number of words in all documents. Assume K latent topics/ genres, each topic is given by a probability vector $\boldsymbol{\phi_k} = (\phi_{k,1}, ..., \phi_{k,N})$ with $\sum_{n=1}^{N} \phi_{k,n} = 1$ indicating the probability that each word shows up in this topic. Further, each document $m \in \{1, ..., M\}$ contains all topics with different probabilities (weights) $\boldsymbol{\theta_m} = (\theta_{m,1}, ..., \theta_{m,K})$ with $\sum_{k=1}^{K} \theta_{m,k} = 1$. Both $\boldsymbol{\phi_k}$ and $\boldsymbol{\theta_m}$ are assumed to have conjugate Dirichlet distributions with hyper parameters (vectors) α and β, respectively.

Given $\boldsymbol{\phi_k}$ and $\boldsymbol{\theta_m}$, a document is generated by drawing for each word a topic $k \in \{1, ..., K\}$ according to the probabilities $\boldsymbol{\theta_m}$ and one word from the selected topic according to its distribution $\boldsymbol{\phi_k}$. This procedure is repeated until the length of the document is reached. To solve the LDA model, we a priori determine $\alpha = 50$ and $\beta = \infty$. The hyper parameter optimization is executed by using Gibbs simulations. Gibbs sampling (also known as alternating conditional sampling) is a specific form of Markov chain Monte Carlo and simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others (see, e.g., Steyvers and Griffiths 2007)

The sampling is done sequentially and proceeds until the sampled values approximate the target distribution. We set the number of sampling iterations equal to 1,000. Then, based on the coherence value across the estimated LDA models using smaller numbers of genres, we find that 12 topics/genres provide the best statistical decomposition of our book description corpus.[17]

[17]For 12 different topics, the coherence value exhibits a local peak. We have also

A detailed list of all 12 genres is presented in Table 10 of Appendix B.

One caveat of the LDA estimation procedure is that it does not give the topics/genres any names or labels. Thus, labels are subjectively given to each genre based on the most important words associated with each topic. In the most cases, it is conceptually simple to classify the genres. Besides, the exact labeling plays no material role in our empirical approach, it is just used as a convenient way of referring to the different topics instead of only using topic numbers.

It is more important that the LDA decomposition gives a meaningful and easily interpretable genre classification of the book descriptions, which it does because our LDA approach identifies all important book genres and clearly delineates the topics. This is shown by the Figures 4 and 5, which are two examples of our 12 word clouds to visualize within the topic distribution of words by assigned probabilities through the LDA. We have labelled the topic in Figure 4 *Crime Novel/Thriller* and the genre in Figure 5 *Politics*. The larger the size of a word in these clouds is, the higher is its weight within the respective topic.

considered 9 and 17 different topics, but in the end there was no real change in the effects on the other variables. In the final prediction stages, we only want to use the variable book genre as a control.

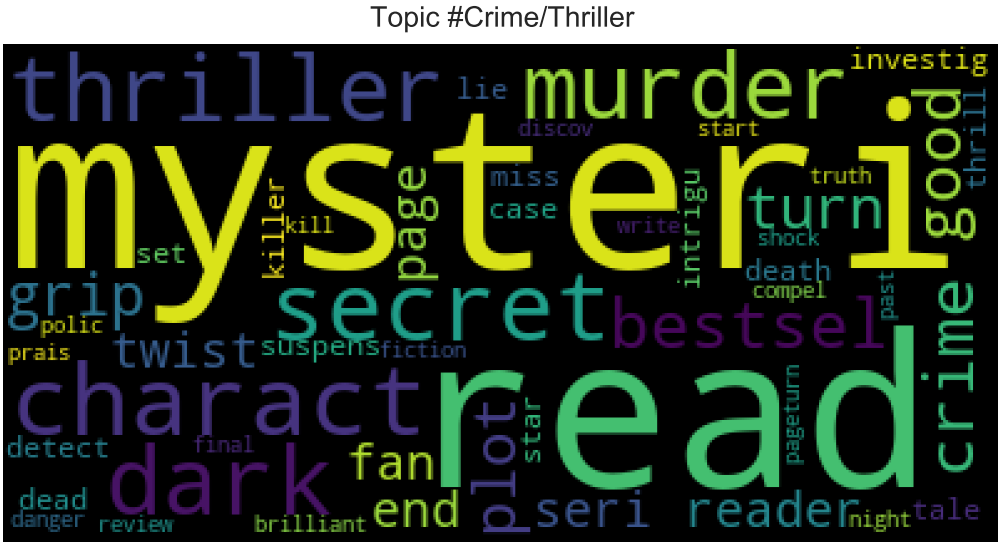

Figure 4: Topic Crime Novel/Thriller from the LDA. Size is according to weight of word within the topic.

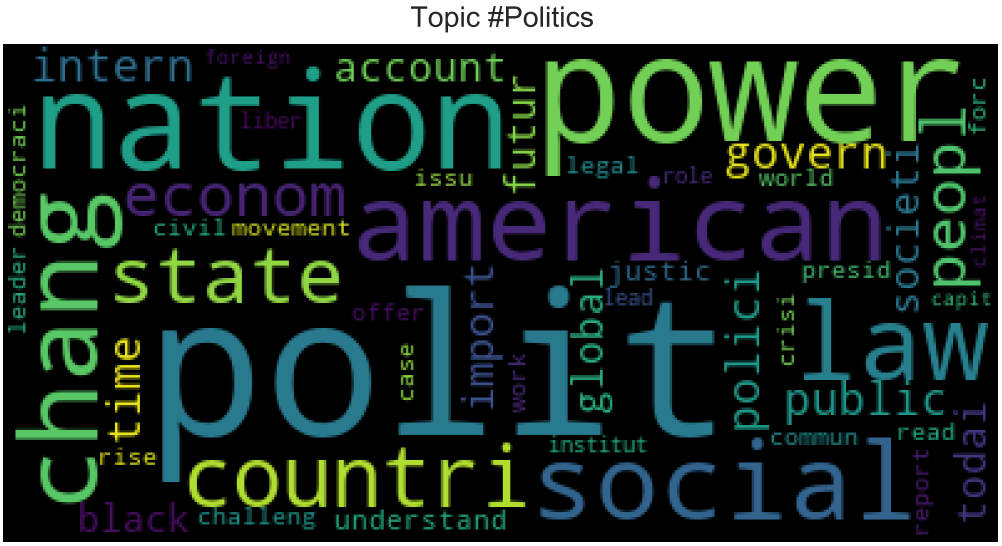

Figure 5: Topic Politics from the LDA. Size is according to weight of word within the topic.

The LDA provides the ability to give each document (book description) from our corpus (data) a probability of being a respective topic. For the visualization purpose in Figure 6, which is based on Figure 2, the largest

probability value is chosen to highlight the distribution of topics over e-book prices and publishers. This distribution of prices by genres exhibits the high comparability between the several publishers in our data set because they are not specialised in certain topics, but all publishers sell book titles from different genres. Nevertheless, it is obvious that the individual publishers have distinct main topics. For instance, Pan Macmillan primarily publishes fiction titles like crime novels, thrillers or society novels whereas HarperCollins has a focus on the topics drama as well as children & youth. However, it is important not to take these topics at face value because the LDA assigns a probability to each individual topic. That is why we directly include these probabilities as difference to one reference topic in the estimation procedure (see footnote 18 in Section 4).

As already mentioned above, one major takeaway from Figure 6 is that all publishers in our data set publish e-books with different genres so that no publisher is specialized in a certain book genre. This is an advantage for our empirical approach because otherwise we would get multicollinearity issues and from an economic point of view we would fail in the sense that we could not compare publishers at all, if there were only specific topics from specific publishers. Fortunately, using the probabilities can also deal with this issue, since we use a continuous value for determining to which genre an e-book should be attributed to.

Another interesting insight is to analyze the Amazon ranks for e-books grouped by publishers and genres, which is illustrated in Figure 7. Due to the scale of the abscissa the plot may be misleading. There are only 30 bins for ranks going into millions which skews the perspective.

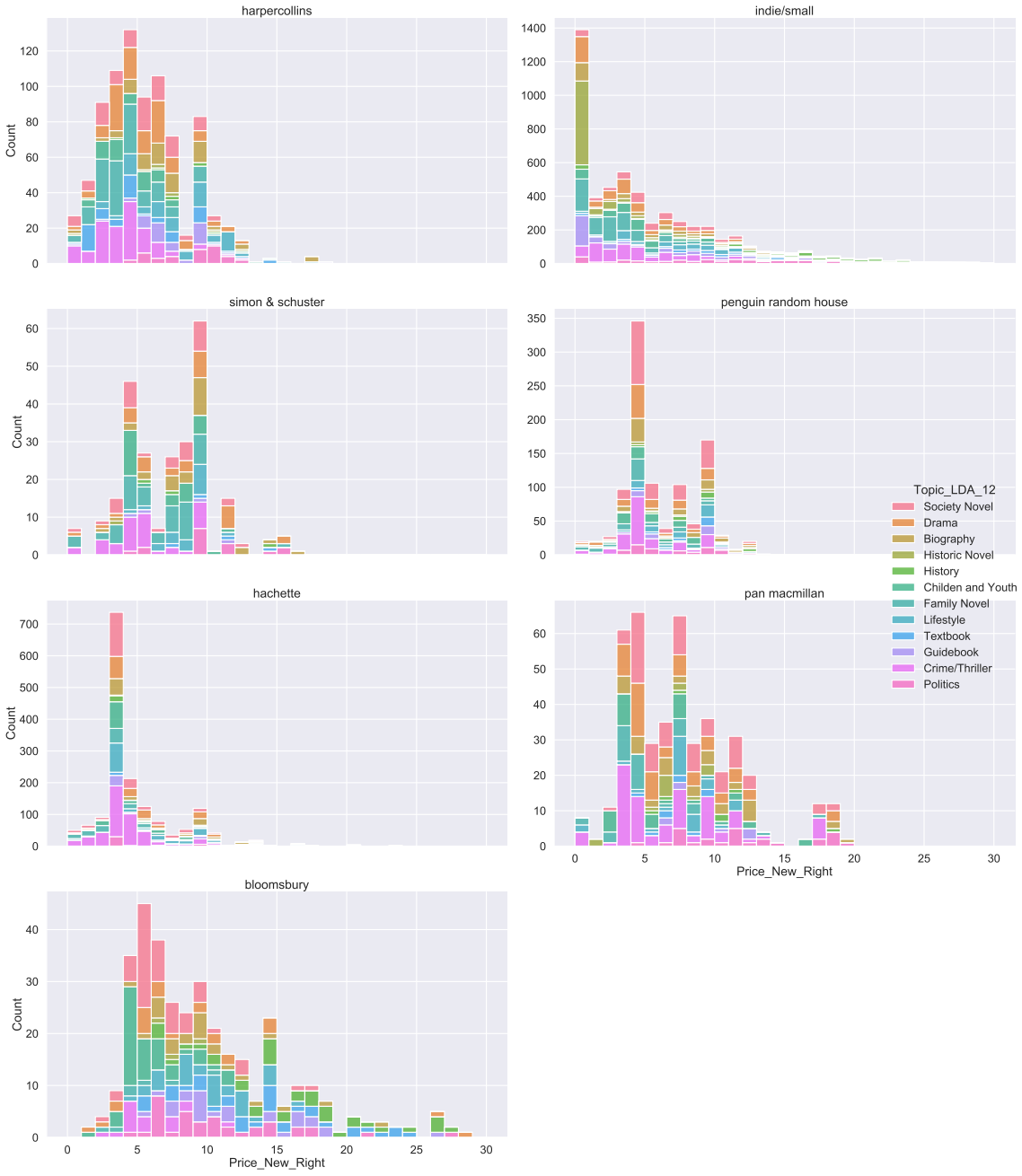

Figure 6: Prices for e-books grouped by publisher and genre. The ordinate is scaled differently for each subplot.

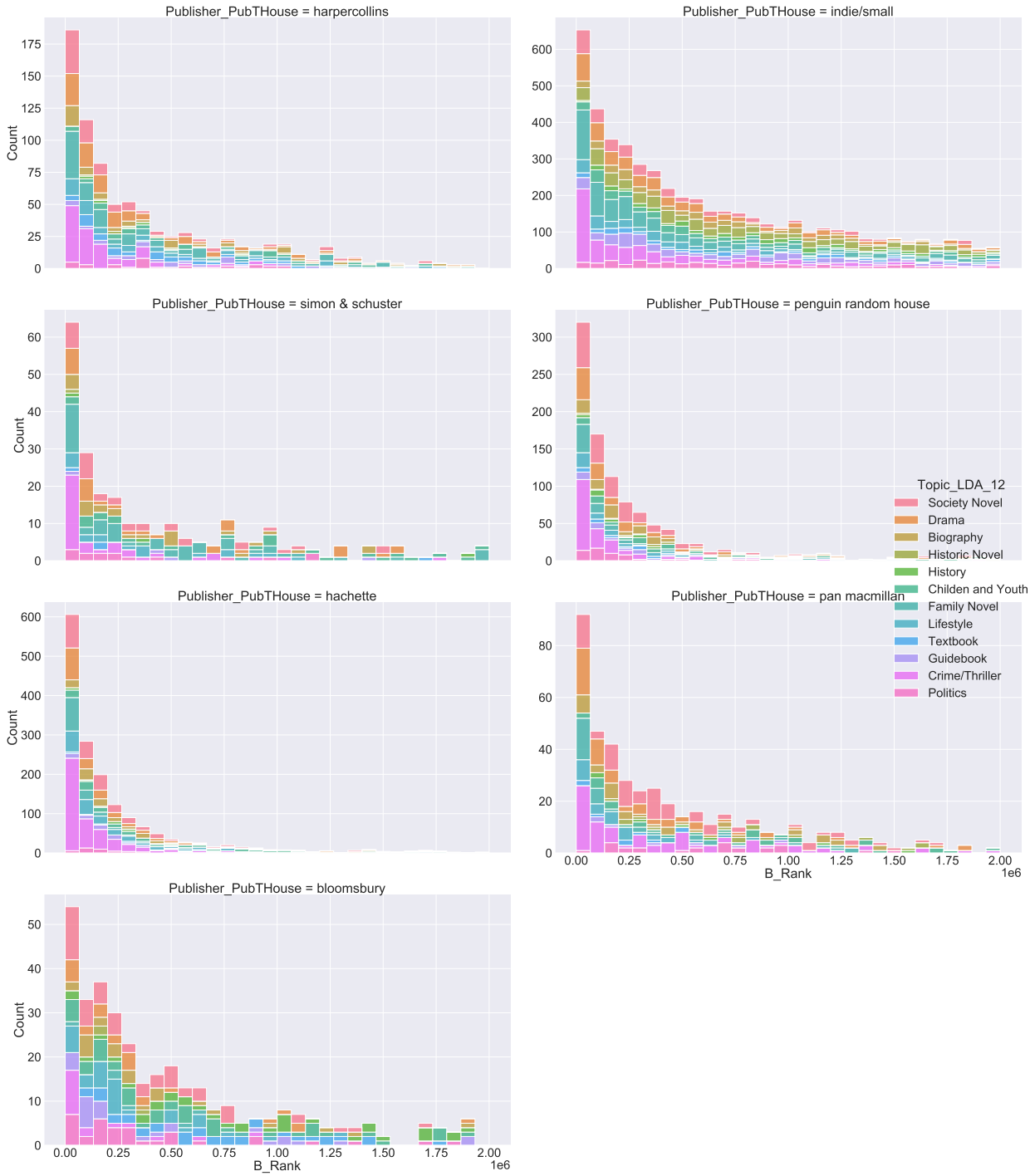

Figure 7: Book rank distribution of e-books by publisher. The bin size is determined by $2,000,000/30 \approx 66.667$.

Nevertheless, the distribution of Bloomsbury and Indie/Small are different in shape compared to the other publishers. They especially have more worse ranked books and the distribution is more dispersed. In general, it can be seen that a majority of the bestsellers on Amazon belong to the book genres crime/thriller and society novel.

4 Empirical Analysis

In this chapter, we present our main empirical analysis. We first describe our estimation strategy in Section 4.1. In Section 4.2, we present the results of our regressions, in which we estimate the impact of the sales model (agency or wholesale) on the retail price of an e-book.

4.1 Estimation Strategy

As already described in Section 1, the goal of our study is to analyze the impact of the sales model on the retail prices of e-books sold on Amazon.co.uk. Therefore, we use publisher and book genre variation of our cross-sectional data to estimate the price effect of e-books sold under the agency model. Before turning to the presentation of our estimations, we formalize the hypothesis that is to be tested. If there was no difference between the two online sales models 'agency' and 'wholesale' regarding the price of an e-book, the opportunity for a publisher to set the retail price of an e-book should (*ceteris paribus*) not have any impact on the prices. Hence, the hypothesis to be tested is:

Hypothesis 0 (H_0): *The retail price of an e-book is independent of the used sales model.*

If there will be a positive correlation between the agency model and the price of e-books, H_0 can be falsified and e-books sold under the agency model are more expensive on average. Observing a negative correlation would also lead to a falsification of H_0, but e-books sold under the agency model would be cheaper on average.

In our baseline estimation approach, we use the standard hedonic modeling approach in the spirit of Rosen (1974), which relies on observing differences in market prices to infer the value or implicit price of underlying characteristics. Thus, we estimate the following log-log OLS model with heteroscedasticity-consistent standard errors:

$$p_i = \alpha_0 + \alpha_1 A_i + \alpha_2 P_i + \alpha_3 G_i + \alpha_4 P_i \times G_i + \alpha_5 R_i + \alpha_6 D_i + \alpha_7 RRP_i + W\theta + \eta_i. \quad (1)$$

In equation (1), the dependent variable p_i is the logarithm of the retail price for an e-book i sold on Amazon.co.uk and the primary variable of interest, A_i, is a dummy variable which takes the value one if an e-book is sold under the agency model and zero otherwise. Beyond, P_i contains the publisher fixed-effects, G_i is a continuous variable containing differences of probabilities to a certain reference genre (genre fixed-effects) and R_i is a continuous variable for the e-book sales ranks on Amazon. D_i reflects the time since a title has been published the first time (in years) and RRP_i gives the recommended retail price of book title i. All other book-specific

covariates are collected in the matrix W (see Table 1 in Section 3.1).

However, there may be issues of reverse causality in regression equation (1). Not only does the rank of an e-book R_i affect the retail price of an e-book p_i but also does the retail price reflect the demand side and, therefore, affect the rank of an e-book which mirrors its sold quantity. As already explained, e-book ranks on Amazon.co.uk are internally determined by overall weighted sales ranks. Thus, ranks might be driven by the quantities sold today but the relation is ambivalent, since the rank is also affected by quantities sold in the past. The e-book prices can be affected by current and past sales but the impact of prices on total (current and past) sales is not so clear. Nevertheless, we cannot clearly reject the endogeneity issue due to a potential reverse causality. To resolve this potential source of endogeneity between the e-book price and its sales rank, we will also present an instrumental variable approach in the following section.

4.2 Main Results

The results of the main log-log OLS model are outlined in Table 6. We estimate four different specifications for our baseline estimation approach: in the first column of Table 6, we present a naive OLS regression model without publisher and genre fixed effects, in column (II) we include publisher fixed effects, in the third column we additionally integrate genre fixed effects[18] and,

[18] We have described the process to generate book genres by using an LDA approach in Section 3.3. Thereby, we have identified a topic for every e-book title based on the largest probability assigned by the LDA. However, this procedure may be miss-leading because sometimes the genre probabilities for a title could be very similar. Therefore, we have solved this issue within our estimation approach by using one reference category and build the difference of the probabilities to the other categories. So by differentiating them, a positive value means that the the respective category has a higher probability assigned

finally, column (IV) also contains the interaction term *publisher* × *genre*.

It is obvious that there is a negative and significant effect of the sales model 'Agency' on the retail price of e-books across all four different specifications. According to the amount, the effect is between 18.4% and 40% depending on the exact specification. For the regression in column (I), an e-book which is sold under the agency model on Amazon.co.uk is approximately 18% cheaper than an e-book which is sold under the wholesale model on average.[19] This is the lowest effect across the four different specifications but, however, this coefficient is probably biased due to omitted variables, since neither publisher nor genre fixed effects are included there, although these variables are crucial to explain the retail price of an e-book. Including publisher and genre fixed effects (as well as their interaction term, see column (IV) in Table 6), increases the agency effect to 36.25% (on amount).

The estimated coefficients for the other controls shown in the table are very similar across the specifications. The sign of the variable *log sales rank* indicates that e-books with higher sales ranks are sold at higher prices which we have already demonstrated descriptively in Figure 3 of Section 3.2 and which confirms the results of Fishwick (2008) whereupon bestsellers are sold cheaper in the UK due to 'substantial discounts'. However, we will discuss potential endogeneity issues concerning this control variable below.

The publisher dummy variables in the columns (II)-(IV) show the price differences between the several publishing houses whereby the publisher *Bloomsbury* is the reference category for those specifications. It is worth emphasizing

by the LDA than the reference category.

[19] To calculate the exact effect of the dummy variable 'Agency' on the dependent price variable, the formula $100 \times (e^{\beta} - 1)\%$ must be used.

that e-books sold by the small and independent publishers are significantly cheaper than the titles of the other publishers in all model specifications. Beyond, there is a significant and positive relation between the RRP of an e-book and its retail price (see *log RRP*) which is not a surprise. We can further observe that the memory space of an e-book (given in KB) has a positive and significant price effect (see *log Kindle Size*).

Most of the remaining control variables in Table 6 are related to consumer ratings and reviews as well as expert reviews of book titles on Amazon.co.uk. Concerning the consumer recommendation networks, the effect of the star rating (*log star rating*) on the price of an e-book seems to be higher than the impact of more consumer reviews (*log no. customer reviews*) because the latter is insignificant in three of the four specifications. However, there is an additional common effect on prices of e-books with a high number of consumer reviews and a high star rating (see interaction variable *log reviews x log stars* in Table 6). Besides, we have found a positive and significant impact of the number of expert reviews on e-book prices (see variable *No. expert reviews*).

There are three control variables remaining in Table 6. First, the covariate *Date Retail* reflects the time period since the publication date of a book title (in years). This variable is only significant in specification (IV) and implies that e-books on average become cheaper (*c.p.*) if they are already available for a longer period of time. Second, the explanatory variable *WeekInChart* reflects the average number of weeks former bestsellers of an e-book's author have last in the bestseller charts of the *Sunday Times*. As expected, this variable has a positive and significant effect (except in specification (II)) on

	(I)	(II)	(III)	(IV)
Constant	-1.62489***	-1.26831***	-1.30165***	-1.16120***
	(0.10395)	(0.09873)	(0.09947)	(0.10130)
Agency	-0.20329***	-0.51041***	-0.44818***	-0.45027***
	(0.01212)	(0.02016)	(0.02015)	(0.02053)
log sales rank	0.03928***	0.06148***	0.06763***	0.06817***
	(0.00491)	(0.00472)	(0.00476)	(0.00478)
Hachette		0.08175***	0.05209*	-0.02843
		(0.02750)	(0.02796)	(0.03452)
HarperCollins		0.12280***	0.11430***	0.08781**
		(0.02999)	(0.03022)	(0.03595)
Indpt/Small		-0.41372***	-0.34868***	-0.45013***
		(0.01762)	(0.01798)	(0.02105)
Pan Macmillan		0.09616***	0.10712***	-0.04101
		(0.02514)	(0.02512)	(0.04312)
Penguin		0.17617***	0.16128***	0.11932***
		(0.02862)	(0.02909)	(0.03291)
Simon & Schuster		0.27029***	0.27555***	0.25721***
		(0.03598)	(0.03660)	(0.05907)
log RRP	1.11609***	1.00566***	0.91672***	0.88570***
	(0.01294)	(0.01237)	(0.01355)	(0.01395)
log Pages	-0.03176**	-0.02663**	-0.01306	-0.00845
	(0.01331)	(0.01210)	(0.01230)	(0.01228)
log Kindle Size	0.04477***	0.03527***	0.03336***	0.03509***
	(0.00428)	(0.00418)	(0.00456)	(0.00450)
log star rating	0.21363***	0.15435***	0.12431**	0.10499*
	(0.06237)	(0.05791)	(0.05789)	(0.05785)
log no. customer reviews	0.00972*	0.00321	0.00197	0.00295
	(0.00583)	(0.00562)	(0.00557)	(0.00552)
log reviews x log stars	0.24541***	0.21841***	0.20619***	0.19178***
	(0.02811)	(0.02579)	(0.02588)	(0.02561)
No. expert reviews	0.03121***	0.02066***	0.02259***	0.01850***
	(0.00482)	(0.00450)	(0.00449)	(0.00454)
Date Retail	0.00496	-0.00032	0.00025	-0.00521*
	(0.00320)	(0.00302)	(0.00301)	(0.00303)
WeekInChart	0.00558**	0.00344	0.00474*	0.00434*
	(0.00258)	(0.00242)	(0.00244)	(0.00232)
Bestsellers	0.00155***	0.00103***	0.00111***	0.00082**
	(0.00041)	(0.00038)	(0.00037)	(0.00037)
R-squared	0.63427	0.67600	0.69169	0.70453
Adj. R-squared	0.63369	0.67529	0.69068	0.70159
Number of observations	10,048	10,048	10,048	10,048
Publisher	No	Yes	Yes	Yes
Genre	No	No	Yes	Yes
Publisher x Genre	No	No	No	Yes

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Baseline regressions (log-log OLS). Dependent variable is the logarithm of the retail price for e-books sold on Amazon.

the retail price of an e-book exhibiting a quality signal . Third, the continuous variable *Bestsellers* exhibits the number of bestselling book titles the author of an e-book has written in the past. This variable also has a positive and significant effect on the retail price of an e-book for all four specifications in Table 6 which can also be interpreted as a quality signal increasing the price of a book title.[20]

However, the results of our OLS estimation in Table 6 might be biased and inconsistent due to endogeneity issues regarding the explanatory variable for the book sales ranks (see Section 4.1 for a discussion). Hence, we will follow an instrumental variable approach in the following estimations to resolve this potential source of endogeneity. We use the logarithmized number of customer reviews (*log no. customer reviews*) as an instrument for the book sales rank to avoid inconsistent estimates due to reverse causality.

Our instrumental variable *log no. customer reviews* is highly correlated with our endogenous regressor book sales rank but should have no partial effect on the price of an e-book (orthogonality assumption). Customer reviews can enhance the awareness and information quality for a consumer and, thus, change the tendency for a consumer to purchase a book. However, the absolute number of customer reviews does not affect the purchasing decision of a consumer for a book title, but only surprisingly positive (negative) reviews can increase (decrease) the consumption of a given good (see Reimers and Waldfogel (2020)). Hence, the absolute number of customer reviews should also have no direct impact on e-book pricing, even though our instrument

[20]The variables *WeekInChart* and *Bestsellers* are based on a historical *Sunday Times Bestseller* list. The matching process was conducted via Python's Fuzzy Matching.

is highly correlated with the book sales rank (as it is an indicator for past sales).

Following the approach explained above, the linear projection in the first stage regression of our 2SLS estimation can be formalized as follows:

$$R_i = \beta_0 + \beta_1 A_i + \beta_2 P_i + \beta_3 G_i + \beta_4 P_i \times G_i + \beta_5 RRP_i + \beta_6 CR_i + W\theta + \xi_i. \quad (2)$$

In equation (2), the dependent variable R_i refers to the book rank on Amazon.co.uk of title i. The covariates A_i, P_i, G_i, RRP_i, and W have already been described in the context of our baseline estimation in equation (1). Our instrumental variable *log no. customer reviews* is displayed by CR_i.

The structural equation of our basic model then takes the following form:

$$p_i = \gamma_0 + \gamma_1 A_i + \gamma_2 P_i + \gamma_3 G_i + \gamma_4 P_i \times G_i + \gamma_5 RRP_i + \gamma_6 \hat{R}_i + W\theta + \varepsilon_i, \quad (3)$$

where the dependent variable p_i is the retail price of e-book i and the fitted values from the first-stage are captured by $\hat{R}_i$.

The regression results based on our structural equation are presented in Table 7. For reasons of comparison, column (1) in the table outlines the results of a comparable OLS estimation using publisher and genre fixed effects (as well as their interaction term). The regression results of our 2SLS estimation (equation (3)) are given in column (2) of Table 7. Also the results of our IV-approach confirm that e-books sold under the agency model at Amazon.co.uk are on average significantly cheaper than titles sold under the wholesale model. The estimated coefficients for the variable *Agency* only

differ in their magnitude between the OLS and the IV estimations.

While the OLS regression states that e-books sold under the agency model are on average 36.36% (given by $100 \times (e^{-0.452} - 1)\%$) cheaper, this effect is only slightly larger for the IV-approach (36.43%, given by $100 \times (e^{-0.453} - 1)\%$). Also the effects of the other explanatory variables barely differ between the OLS and the IV estimation approaches, only the impact of the book ranks and the number of expert reviews on e-book prices have become greater for the IV estimation. Hence, we can reject the H_0-hypothesis that the retail price of an e-book is independent of the sales model at the 0.1%-level based on our estimation results.

	(1) Price	(2) Price
Constant	-1.291***	-1.745***
	(-13.65)	(-14.89)
Agency	-0.452***	-0.453***
	(-27.90)	(-21.84)
log RRP	0.905***	0.896***
	(73.09)	(66.97)
log sales rank	0.0762***	0.109***
	(20.39)	(13.16)
log star rating	0.313***	0.331***
	(8.74)	(7.20)
No. expert reviews	0.00775	0.0119**
	(1.81)	(2.69)
log Kindle Size	0.0356***	0.0341***
	(7.52)	(7.63)
Number of observations	10,048	10,048
Publisher	Yes	Yes
Genre	Yes	Yes
Publisher x Genre	Yes	Yes

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Model Comparison of OLS- (1) and IV-approach (2). Woolridge-Test for endogeneity suggests with $\chi^2 = 18.5$ that there is endogeneity. F-Stats of first stage is 2423.

5 Robustness Checks

To check the robustness of our results presented in Section 4, we apply two further estimation approaches. First, we apply the heteroscedasticity based IV approach as suggested by Lewbel (2012) in Section 5.1. Following, we

present a double machine learning (DML) approach in Section 5.2.

5.1 Lewbel Approach

The Lewbel approach from Lewbel (2012) is a relatively new method, which tackles the issue of endogeneity in linear systems. This approach serves to identify structural parameters in regression models with endogenous or mismeasured regressors in the absence of traditional identifying information (e.g., external instruments). Thereby, identification is reached by having regressors that are uncorrelated with the product of heteroskedastic errors. We use this method to instrument the potentially endogenous variable *log sales rank*.

Column (1) in Table 8 shows the estimation results of our standard IV approach from equation (3) in Section 4.2 (c.f., column (2) in Table 7) to give a comparison. The second column contains the result of the corresponding Lewbel's IV regression without forcing any of the given exogenous variables to be an instrument. Column (3) builds on the previous instruments of the first column to run another Lewbel approach. Based on the Hansen-J-Statistic both Lewbel regressions perform worse than the standard IV approach, since this criterion suggests rejection of the overidentifying restrictions.

Nevertheless, the Lewbel approach confirms the robustness of our estimation results. The effect of the sales model 'Agency' on the retail prices of e-books is still negative and significant. Only the magnitude of the effect has slightly changed. Also the regression results for the other covariates barely differ between the standard IV approach and the Lewbel regressions.

	(1) Price	(2) Price	(3) Price
Constant	-1.745***	-1.641***	-1.665***
	(-14.89)	(-11.40)	(-15.04)
Agency	-0.453***	-0.453***	-0.453***
	(-21.84)	(-21.91)	(-21.91)
log sales rank	0.109***	0.102***	0.104***
	(13.16)	(11.80)	(16.71)
log RRP	0.896***	0.896***	0.898***
	(66.97)	(63.68)	(64.68)
log star rating	0.331***	0.332***	0.328***
	(7.20)	(7.21)	(7.16)
Date Retail	-0.00690*	-0.00656*	-0.00677*
	(-2.27)	(-2.14)	(-2.23)
No. expert reviews	0.0119**	0.0123**	0.0113**
	(2.69)	(2.70)	(2.58)
log Kindle Size	0.0341***	0.0348***	0.0345***
	(7.63)	(7.66)	(7.69)
Bestsellers	0.000950*	0.000929*	0.000923*
	(2.56)	(2.53)	(2.51)
WeekInChart	0.00632*	0.00595*	0.00595*
	(2.51)	(2.41)	(2.40)
log no. customer reviews		-0.00330	
		(-0.55)	
Number of observations	10,048	10,048	10,048
Publisher	Yes	Yes	Yes
Genre	Yes	Yes	Yes
Publisher x Genre	Yes	Yes	Yes

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Lewbel approach. The first column is the IV-approach from column (2) in Table 7, the second column shows Lewbel with no specific instrument and the third column is the Lewbel approach with the previously defined instrument (log no. customer reviews). The Hansen-J-Statistic suggests to reject H_0 of overidentification for both Lewbels, but not for the standard instrumental variable approach.

5.2 Double Machine Learning

There are further methods beyond the established approaches in the standard econometric analysis. We have already applied standard econometric methods as the OLS or the instrumental variable approaches. Recent advances in

machine learning approaches also offer a larger toolbox for empirical analysis in economics (see, e.g., Athey (2018) and Athey and Imbens (2019) for a broad overview).

Due to recent developments in the machine learning literature there are many approaches, e.g. the DML technique, that gives the possibility to deal with common econometric issues as confounding variables or variable selection via cross-validation and non-parametric models. This technique also allows to make use of non-standard modelling of relations between variables, like the independent variables have a specific, often assumed linear or quadratic effect on the dependent variable. It permits to use any arbitrary machine learning technique relying on algorithms to find a fitting model for some chosen score functions like the mean squared error.

The reason for relying on further non-parametric/semi-parametric regressions is to circumvent the imposition of a model structure based on the underlying data generating process but to let the algorithm choose the best fitting model under some restrictions or parametrization. We will then use estimation approaches via, e.g., regularized linear regression techniques like the least absolute shrinkage and selection operator (LASSO) or regression Trees/Forests. These methods help to compare our standard econometric approaches with models that are able to ignore irrelevant variables or include non-linear effects (Athey & Imbens, 2019).

Therefore, we apply DML to compare previous estimations with these approaches to provide further robustness checks. From a prediction's perspective, the estimations will be split into three different regression models, which is proposed by Chernozhukov et al. (2017) and Chernozhukov et al.

(2018). The models use the DML-framework to deal with high dimensional or possibly confounding variables based on the *DML-Conditional-Average-Treatment-Effect-Estimator*.[21]

The equation system we estimate has the following form:[22]

$$\begin{aligned} Y &= \theta T + q(W) + \varepsilon \\ T &= f(W) + \eta \\ s.t.\ E[\varepsilon|W] &= E[\eta|W] = E[\varepsilon \cdot \eta|W] = 0. \end{aligned} \tag{4}$$

The estimation of the equation-system (4) is based on partial linear regression models described in Robinson (1988) which will be estimated by arbitrary functional forms for $q(W)$ and $f(W)$ and is shown in the following equation-system:

$$\begin{aligned} \tilde{Y} &= Y - q(W) \\ \tilde{T} &= T - f(W) = \eta \\ \tilde{Y} &= \tilde{T} + \epsilon. \end{aligned} \tag{5}$$

The variable of interest still is the agency dummy variable (given by the parameter T in (4) and (5)). Hence, we regress the retail price of e-books on this agency dummy variable ($\hat{p} = \hat{Y}$ on $\hat{A} = \hat{T}$) using arbitrary machine learning functions in the two first stages of the equations above. Thereby,

[21] For the estimation implementation we follow the Python Module *econml* provided by Microsoft Research (2019).

[22] The formal approach of the model is described at https://econml.azurewebsites.net/spec/estimation/dml.html#overview-of-formal-methodology.

W contains the previously introduced covariates to predict the new prices $\hat{p}$ and the new agency dummies $\hat{A}$ in the first stage.

In Table 9, the column *Model* represents the applied functional form. The first entry in this column refers to the functional form of predicting $\tilde{Y}$ and the second for classifying $\tilde{T}$. Therefore, *Lin Logit* relies on OLS and a logistic regression (including L_2 penalty), *Lasso Logit* uses a Lasso and a logistic regression with L_2, *Lasso RFC* is a Lasso and a random forest classifier, *RFR Logit* combines a random forest and a logistic regression, *RFR RFC* uses a random forest for both stages and XGBoost means Extreme Gradient Boost for both stages. The columns *Score* represents the mean squared error of the final stage. In the final stage, a simple linear regression is used to get the conditional average treatment effect. There are many more possible estimation techniques but these are sufficient for highlighting the stability of our results. The hyperparameters for each model are chosen from a reasonable set and then we use $3-5$ cross-fold validation within Python's Sklearn GridSearch. Besides, we also do another 5-fold splitting in each estimation. The selected results outline the best estimation (lowest score) for each model class.

Model	Agency	Std. Dev.	p-value	Score	Perc Change
Lin Logit	-0.4184	0.021	0.00	0.2722	-34.1901
Lasso Logit	-0.4282	0.021	0.00	0.2852	-34.8319
Lasso RFC	-0.4721	0.024	0.00	0.2853	-37.6309
RFR Logit	-0.3188	0.019	0.00	0.2325	-27.2979
RFR RFC	-0.4122	0.022	0.00	0.2288	-33.7808
XGBoost	-0.3396	0.025	0.00	0.1999	-28.7945

Table 9: Double-Machine Learning. Dependent Variable E-Book-Prices. Variable of interest is *Agency_Set*. The effect for the respective model is given in the column *point_estimate*. Column *Score* refers to mean squared error. Logistic Regression with penalization exhibits perfect classification.

The point estimates (see column Agency in Table 9) are in the vicinity of the estimations presented above and the relative percentage changes in relation to the intercept are given in the column *Perc Change*. Again, one can calculate the exact percentage change by the formula $100 \times (e^{Agency} - 1)\%$. For instance, e-books sold under the agency model on Amazon.co.uk are cheaper with a mean value of 34.8% when using the estimation model Lasso Logit. In overall, the DML approaches confirm the results of our main estimations presented in Section 4.2 and prove the robustness of our regressions.

6 Conclusion

In this paper, we have provided evidence that e-books sold under the agency model on Amazon.co.uk are on average significantly cheaper than e-books sold under the wholesale model. Our results are based on an appropriate data set containing many characteristics of an e-book, which has been scraped from the Amazon.co.uk webpage. To measure the relationship between the retail price of an e-book and the used sales model, we have relied on classical econometric techniques like the IV-approach as well as newer methods as the DML approach. We have found an robust and statistically significant effect that e-books sold under the agency model are (on average) 36% cheaper than digital books sold under the wholesale model.

The results of our empirical analysis are in line with many theoretical papers studying the price impact of the agency model. Those theoretical analyses argue that retail prices for e-books sold under the agency model are lower due to a lock-in effect exploited by retailers (Johnson 2020), the

monopolistic power of retailers over a complementary device as it is the case for Amazon when e-books could only be read on a Kindle device (Gaudin and White 2014) or because agency selling is just more efficient than the wholesale model and leads to lower retail prices (Abhishek et al. 2016). Our results also match to the model-theoretical analysis from Foros et al. (2017) if one assumes that competition should be greater among publishers than among retailers, which most likely is the case due to the quasi-monopolistic power of Amazon.

To the best of our knowledge, our paper is the first empirical analysis estimating the price effect of the agency model concerning e-books not only incorporating bestselling titles but a broader cross-sectional data set of digital books. Besides, in contrast to previous empirical analyses regarding the retail price of e-books, we apply an LDA approach to determine detailed book genres. Nevertheless, a limitation of our approach is that we use cross-sectional data instead of panel data. Cross-sectional data only provides a small snapshot at a certain point of time and we cannot analyse any dynamic effects on the retail price of e-books. Moreover, we only include one online platform, namely Amazon, instead of comparing various online retailers and most of the contractual details between the publishers and the retailers, which could also play a role in explaining the retail prices of e-books, are unobservable to us.

The dynamic impact of the agency model on e-book retail prices including bestselling as well as "long tail" book titles remains an open question. Future research should concentrate on panel data to address such long-run effects of book sales models. Thereby, also multiple online platforms selling e-books

should be included in such an analysis to identify effects across the online retailers, even though Amazon provides a large market coverage for e-book sales. Finally, the long-run impact of the agency model on consumer welfare is an interesting research area. Consumer welfare not only depends on the price effect but also on other factors such as the number, variety, and quality of book titles written and published.

References

Abhishek, V., Jerath, K., & Zhang, Z. J. (2016). Agency selling or reselling? channel structures in electronic retailing. *Management Science*, *62*(8), 2259–2280.

Athey, S. (2018). The impact of machine learning on economics. *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, *29*(5), 815–827.

Boik, A., & Corts, K. S. (2016). The effects of platform most-favored-nation clauses on competition and entry. *The Journal of Law and Economics*, *59*(1), 105–134.

Brynjolfsson, E., Hu, Y., & Rahman, M. S. (2009). Battle of the retail channels: How product selection and geography drive cross-channel competition. *Management Science*, *55*(11), 1755–1765.

Chen, H., Hu, Y. J., & Smith, M. D. (2019). The impact of e-book distribution on print sales: Analysis of a natural experiment. *Management Science*, *65*(1), 19–31.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., et al. (2017). *Double/debiased machine learning for treatment and causal parameters* (tech. rep.).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, *21*(1).

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

Condorelli, D., Galeotti, A., & Skreta, V. (2018). Selling through referrals. *Journal of economics & management strategy*, *27*(4), 669–685.

Crosby, P. (2019). Don't judge a book by its cover: Examining digital disruption in the book industry using a stated preference approach. *Journal of Cultural Economics*, *43*(4), 607–637.

Davies, S., Coles, H., Olczak, M., Pike, C., & Wilson, C. (2004). Benefits from competition: Some illustrative uk cases.

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

Dearnley, J., & Feather, J. (2002). The uk bookselling trade without resale price maintenance an overview of change 1995–2001. *Publishing research quarterly*, *17*(4), 16–31.

Fishwick, F. (2008). Book prices in the uk since the end of resale price maintenance. *International journal of the economics of business*, *15*(3), 359–377.

Foros, Ø., Kind, H. J., & Shaffer, G. (2017). Apple's agency model and the role of most-favored-nation clauses. *The RAND Journal of Economics*, *48*(3), 673–703.

Gans, J. S. (2012). Mobile application pricing. *Information Economics and Policy*, *24*(1), 52–59.

Garthwaite, C. L. (2014). Demand spillovers, combative advertising, and celebrity endorsements. *American Economic Journal: Applied Economics*, *6*(2), 76–104.

Gaudin, G., & White, A. (2014). On the antitrust economics of the electronic books industry. *Available at SSRN 2352495*.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74.

Gilbert, R. J. (2015). E-books: A tale of digital disruption. *Journal of Economic Perspectives*, *29*(3), 165–184.

Goetz, G., Herold, D., Klotz, P.-A., & Schäfer, J. T. (2020). *The substitutability between brick-and-mortar stores and e-commerce-the case of books* (tech. rep.). Joint Discussion Paper Series in Economics.

Helmers, C., Krishnan, P., & Patnam, M. (2019). Attention and saliency on the internet: Evidence from an online recommendation system. *Journal of Economic Behavior & Organization*, *161*, 216–242.

Ippolito, P. M. (1991). Resale price maintenance: Empirical evidence from litigation. *The journal of law and economics*, *34*(2, Part 1), 263–294.

Johnson, J. P. (2020). The agency and wholesale models in electronic content markets. *International Journal of Industrial Organization*, *69*, 102581.

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, *210*(1), 203–218. https://doi.org/https://doi.org/10.1016/j.jeconom.2018.11.013

Lenz, D., & Winker, P. (2020). Measuring the diffusion of innovations with paragraph vector topic models. *PLOS ONE*, *15*(1), 1–18. https://doi.org/10.1371/journal.pone.0226685

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, *30*(1), 67–80.

Li, H. (2019). Intertemporal price discrimination with complementary products: E-books and e-readers. *Management Science*, *65*(6), 2665–2694.

MacKay, A., & Smith, D. A. (2017). Challenges for empirical research on rpm. *Review of Industrial Organization*, *50*(2), 209–220.

Mathewson, G. F., & Winter, R. A. (1984). An economic theory of vertical restraints. *The RAND Journal of Economics*, 27–38.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. *http://mallet. cs. umass. edu*.

Microsoft Research. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation [Version 0.8.0b1].

Nelson, P. (1970). Information and consumer behavior. *Journal of political economy*, *78*(2), 311–329.

Oestreicher-Singer, G., & Sundararajan, A. (2012a). Recommendation networks and the long tail of electronic commerce. *Mis quarterly*, 65–83.

Oestreicher-Singer, G., & Sundararajan, A. (2012b). The visible hand? demand effects of recommendation networks in electronic markets. *Management science*, *58*(11), 1963–1981.

Perry, M. K., & Porter, R. H. (1986). *Resale price maintenance and exclusive territories in the presence [of] retail service externalities*. Department of Economics, State University of New York at Stony Brook.

Poort, J., & van Eijk, N. (2017). Digital fixation: The law and economics of a fixed e-book price. *International Journal of Cultural Policy*, *23*(4), 464–481.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Reimers, I. C., & Waldfogel, J. (2020). *Digitization and pre-purchase information: The causal and welfare impacts of reviews and crowd ratings* (tech. rep.). National Bureau of Economic Research.

Reinstein, D. A., & Snyder, C. M. (2005). The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The journal of industrial economics*, *53*(1), 27–51.

Rey, P., & Stiglitz, J. (1994). *The role of exclusive territories in producers' competition* (tech. rep.). National Bureau of Economic Research.

Rey, P., & Stiglitz, J. E. (1988). *Vertical restraints and producers' competition* (tech. rep.). National Bureau of Economic Research.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of political economy*, *82*(1), 34–55.

Shaffer, G. (2012). The economics of parities and differentials. *Final report prepared for the Office of Fair Trading*.

Sorensen, A. T. (2007). Bestseller lists and product variety. *The journal of industrial economics*, *55*(4), 715–738.

Spengler, J. J. (1950). Vertical integration and antitrust policy. *Journal of political economy*, *58*(4), 347–352.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Telser, L. G. (1960). Why should manufacturers want fair trade? *The journal of law and economics*, *3*, 86–105.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *38*(2), 393–409.

Tirole, J. (1988). *The theory of industrial organization*. MIT press.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., & Barabási, A.-L. (2019). Success in books: Predicting book sales before publication. *EPJ Data Science*, *8*(1), 31.

Winter, R. A. (1993). Vertical control and price versus nonprice competition. *The Quarterly Journal of Economics*, *108*(1), 61–76.

Yamey, B. S. (1954). *The economics of resale price maintenance*. Pitman.

A Appendix

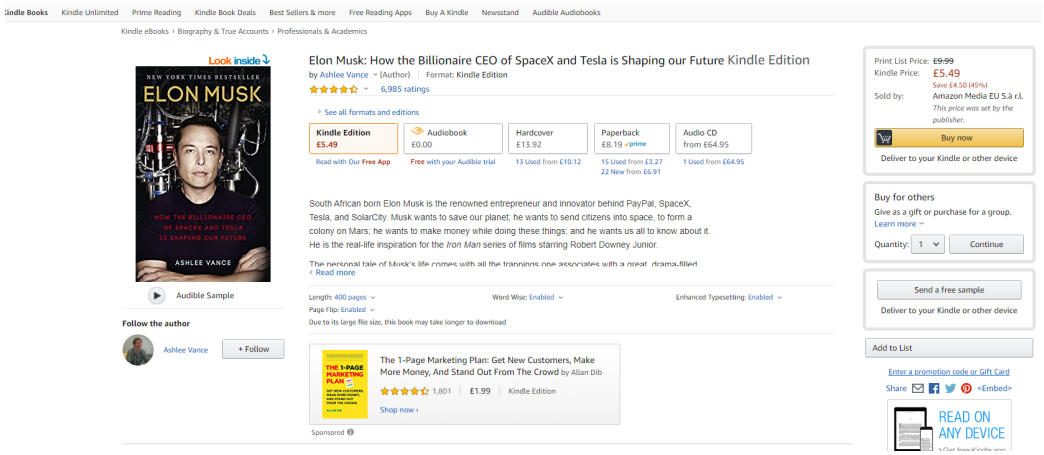

Figure 8: Screenshot of *Elon Musk: How the Billionaire CEO of SpaceX and Tesla is Shaping our Future* (Amazon.co.uk).

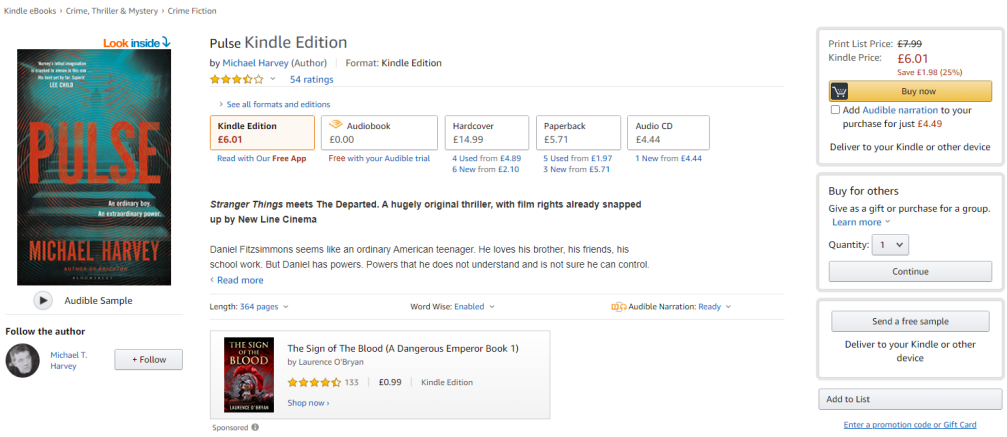

Figure 9: Screenshot of *Pulse* (Amazon.co.uk).

B Appendix

Topic	Genre
0	History
1	Guidebook
2	Childen and Youth
3	Society Novel
4	Lifestyle
5	Crime Novels/Thriller
6	Politics
7	Historic Novel
8	Drama
9	Family Novel
10	Biography
11	Textbook

Table 10: 12 different genres identified by our LDA approach.