Interdisciplinary Institute for Innovation

Superstars and Outsiders in Online Markets: An Empirical Analysis of Electronic Books

David Bounie
Bora Eang
Marvin Sirbu
Patrick Waelbroeck

Working Paper 12-TS-01

April 19, 2012

Department of Economics and Social Sciences
TELECOM ParisTech
46 rue Barrault, 75013 Paris, France
Email: wpecon@telecom-paristech.fr

Superstars and Outsiders in Online Markets: An Empirical Analysis of Electronic Books

David Bounie[1], Bora Eang[2], Marvin Sirbu[3] and Patrick Waelbroeck[4]

April 19, 2012

Abstract

Recent rapid growth in electronic book sales has raised a critical question for publishers and book stores: do e-books cannibalize or increase print sales? In this article we compare the best-selling titles sold on Amazon.com in print or electronic (Kindle) formats during the period from November 2007 to July 2010. Using econometric methods, we find that the cannibalization of print sales by e-books is more likely to occur for superstar titles written by successful authors. However, we find that a new segment of successful electronic titles that are not best-selling in print format emerge; these books would probably have been unpopular without the new Kindle store and therefore this new distribution channel has expanded the market. We refer to these titles as digital outsiders. The latter are characterized not only by lower prices but also by older release dates. They also include titles that are only released in electronic format. We then argue that electronic books increase the market viability of old print releases. Finally we identify a category that we call "print preferred" of books that are top sellers in print but not as e-books for reasons of color, graphics, or the need to navigate non-linearly, a style to which the current generation of e-book readers are not well adapted.

[1] Telecom ParisTech, Department of Economics and Social Sciences, 46 rue Barrault, 75634 Paris Cedex 13, France; Tel.: + 33 1 45 81 73 32; Email address: david.bounie@telecom-paristech.fr; Corresponding author.

[2] Telecom ParisTech, Department of Economics and Social Sciences, 46 rue Barrault, 75634 Paris Cedex 13, France; Email address: bora.eang@telecom-paristech.fr.

[3] Carnegie Mellon University, Department of Engineering and Public Policy and Tepper School of Business, 5000 Forbes Avenue Pittsburgh, PA 15213-3891; Email address: sirbu@cmu.edu.

[4] Telecom ParisTech, Department of Economics and Social Sciences, 46 rue Barrault, 75634 Paris Cedex 13, France; Email address: patrick.waelbroeck@telecom-paristech.fr.

1 Introduction

The book industry is facing a new revolution. According to most experts, electronic books (e-books) will shake the industry in the same way that the printing press did.[5] Recent numbers published by the *Association of American Publishers* show that 49.5 million e-books have been sold in the US in 2010 for a value of USD 441 million (an increase of 291% compared to 2009). Sales of e-books now represent 8.3% of total sales of books in the US in 2010 (against 3.2% in 2009).[6] Amazon currently sells more e-books for Kindle, their proprietary e-book reader, than print books.[7] Many actors in the industry view this major technological change with apprehension for various reasons. First, sales of e-books could threaten retail stores selling print books, given the income constraint of consumers. Secondly, this potential cannibalization of print sales by e-books is reinforced by the entry of new online retailers such as Amazon, Apple or Google. The bargaining power between traditional retail stores and new online merchants is changing the way companies conduct business. For instance, new technologies allow writers to publish e-books directly without signing with a publisher.[8] Finally, editors and publishers fear illegal copying on the Internet of copyrighted material that could reduce their income.

This article directly addresses the issue of the cannibalization: do e-books cannibalize or instead increase print sales? The issue of cannibalization or disintermediation has been a central question since the early days of the Internet and has taken a new turn in recent years. The Internet first created a new sales channel for *physical* cultural products and rapidly a growing empirical literature has emerged to assess how the Internet has impacted the sales of *physical* goods on off-line channels (Balasubramanian, 1998). This research led to the formulation of the "long tail" theory (Anderson, 2004). According to the latter, cultural industries which are characterized by superstar products and "winner-takes-all" phenomenon (Rosen, 1981) – *i.e.* a small portion of artists/writers/films account for most sales in the market – should be impacted by the Internet, which lower search costs and allows internet users to share information within online communities. As a result, the Internet should make cumulative sales of "niche" (obscure or unknown) products profitable and thus flatten the distribution of total sales of cultural products (Brynjolfsson *et al*., 2003). This prediction however remains controversial in the economics literature (Elberse and Oberholzer-Gee, 2007). In the specific context of the book market, Chevalier and Goolsbee (2003) and Ghose, Smith, and Telang (2006) have found that online channels result in a relatively small

[5] See for instance The NewYork Times website, "Using E-Books to Sell More Print Versions", June 26, 2011; or, The Wall Street Journal, "How the E-Book Will Change the Way We Read and Write", April 20, 2009; etc.

[6] The press release is available at the following address: http://www.publishers.org/press/24/ (last visit: 20/05/2011).

[7] A press release available at the following address outlines: "Since April 1 [2010], for every 100 print books Amazon.com has sold, it has sold 105 Kindle books" (last visit: 20/05/2011).

[8] See two illustrative papers retrieved from the Guardian.co.uk, 12 January 2012. "Amanda Hocking, The writer who made millions by self-publishing online" and guardian.co.uk, 8 February 2012, "Self-published ebook author becomes Amazon's top seller; "Amazon's success with self-published authors in the UK follows the US arm of the retailer's announcement last year that two self-published authors, John Locke and Amanda Hocking, had sold more than 1m books on the Kindle. Hocking went on to sign a reported $2m deal with St Martin's Press, while Locke has signed up with Simon & Schuster."

cannibalization of print sales.[9] Likewise, Bounie et *al.* (2011) have compared the top 100 weekly best-selling books in France via three different distribution channels: Amazon, Amazon Marketplace and physical stores during the period from March to August 2006. They show that on average, only 11% of all books ever entered all three lists of weekly Top 100 best selling items. Old releases are more successful in the electronic markets (Amazon and Amazon Marketplace) than in traditional physical stores.

The issue of cannibalization has now taken a new turn with the increasing availability of *digital* cultural goods. Indeed, new formats have appeared on the Internet such as digital music (iTunes, Spotify, etc.) or online films and movie downloads (iTunes, Amazon Instant Video, Google Play, Netflix, Youtube, etc.). Likewise, the development of e-books has lead to an increasing popularity of the Kindle store. The emergence of these new formats raises a new question: do *electronic* formats cannibalize or instead increase the sales of *physical* cultural products (that can be sold through physical or online channels)? Indeed, consumers not interested in print formats could now be interested in purchasing new electronic formats. This market expansion could not only be the result of decreasing search costs and online communities but could also be driven by a new demand for electronic formats (increased portability, library of books always available, bookmarks and annotations, etc.). The long tail theory should therefore be extended to account not only for the effect of online distribution channels on print sales but also the impact of new electronic formats on the variety of cultural products purchased by Internet users. Several empirical papers have addressed this question. Focusing respectively on the introduction of online newspapers or music, Deleersnyder *et al*. (2002) and Biyalogorsky and Naik (2003) find a relatively small cannibalization effect on physical newspaper circulation or record sales. Even though the conclusions of the latter earlier studies could be today questioned, these findings seem to be supported by more recent research on videos and TV: for instance, Waldfogel (2009) shows that the Youtube platform that allows users to stream videos online has only a small negative impact on television viewing and Danaher *et al*. (2010) find that the iTunes distribution channel has no significant statistical impact on DVD sales.

In this study, we focus on the characteristics of books that are: 1) best sellers in both print and electronic formats, 2) books that are best sellers in print, but not as e-books, and 3) a group of books that are successful as e-books but either currently less successful in print or have no print equivalent. We refer to the first group as "superstars", the second as "print preferred" and the third as "digital outsiders". Paper may be preferred for reasons of color, graphics, or need to navigate non-linearly, a style to which the current generations of e-book readers are not well adapted. To do so, we have run automated scripts to collect all titles that have appeared in the monthly top 100 list of best-selling print and electronic books from the US Amazon websites. The period of the sample ranges from November 2007 to July 2010 (121 months). There are 1861 unique titles in the set of print and electronic books that we compare. The focus on the US market is justified by the faster development of the US e-book market compared to other countries. Using econometric methods, we find that the cannibalization of print sales by sales of electronic books is more likely to occur for superstar titles written by successful authors. However, we find that a new segment of successful electronic titles that are not best-selling in print format emerge; these books would probably have been unpopular without the new Kindle store and therefore this new distribution channel creates a market expansion effect. These digital outsiders are characterized not only by lower prices but also by

[9] Ghose, Smith and Telang (2006) have especially assessed the impact of used print books on the sales of new print books on Amazon MarketPlace during the period 2002-2003. These authors show that used print books are imperfect substitutes to new ones in the sense that only 16% of used books cannibalize sales of new print books.

older release dates. They also include titles that are only released in electronic format. We then argue that electronic books increase the duration of popularity of older print releases.

To the best of our knowledge, few studies have analyzed the market expansion/cannibalization issues in the book industry due to the emergence of an electronic format.[10] Jiang and Katsamakas (2010) use a game theoretical model to study the effects of the entry of an e-book seller on prices and total book readership. They find that prices in the book market may increase after the entry of an e-book seller and that the total readership may decrease, if the e-book seller is owned by one of the print book sellers. In an empirical context, Oestreicher-Singer and Sundararajan (2010) examine the e-book industry but only focus on the pricing strategy of e-book publishers and their choices of technological protection (with or without piracy). The closest paper to our research is probably Hu and Smith (2011). They study the impact of the decision of a publisher to stop releasing Kindle e-books on the Amazon website. This leads to a natural experiment where a title is initially only available in print format and later on available in both print and electronic formats. These titles can be compared to a control group of titles that were available in both formats during the full period of observation. They find that delaying the release of e-books causes an insignificant change in overall hardcover sales but a significant decrease in e-book sales, total sales, and likely total revenue and profit to the publisher. Unfortunately, they only have one to eight weeks of observations of titles that were delayed in electronic format compared to two months of observations for titles in the control group, which makes a comparison based on sales difficult to interpret. Moreover, they find a negative cross-price elasticity between books in print and electronic formats, which suggests that the formats are complements. On the contrary we estimate a positive cross-price elasticity, which implies that both formats are rather substitutes. Finally, they restrict their sample to avoid outliers with unusually large sales (average weekly sales higher than 1600 copies), while we specifically focus on these superstar titles. In addition, they do not analyze digital outsiders that are only available in electronic formats, whereas, they are a main focus of this article.

The remainder of the article is organized in three sections. First, we describe the dataset and compare the best selling books in print and electronic formats. Secondly, we explain the probability that a best selling print book is also best selling in its electronic version by a set of book characteristics. Finally, we characterize digital outsiders and show that they either correspond to electronic books that do not exist in print formats or to old releases that Internet users want to rediscover through the electronic store.

2 Data and methodology

Electronic books refer both to a title and to a reader of specific formats such as the Kindle reader.[11] In this article, we use the word e-book for the content. Several formats coexist, proprietary or not: ".azm", ".epub", ".pdf", ".txt", etc. Each format has its own specific features (reading images, printing content, bookmarking, searching indexes, etc.). We use data from the Kindle store owned by Amazon.com that are readily available from the Internet.

[10] Huang and Hsieh (2012) show that consumers' perceived innovative attributes not only directly affect their acceptance behaviour, but also influence behaviour via their perception of switching costs.

[11] Two types of readers are available: the first one, called e-Reader, is exclusively dedicated to the reading of electronic texts such as the Amazon Kindle, the Sony Reader, etc. The second one is a multifunctional "touch pad" which allows reading electronic text files as well as playing videos, listening music, etc. The well known iPad by Apple is an example. Many e-book sellers also provide software based readers for PCs and smartphones.

Indeed, Amazon keeps archives of the monthly top 100 best-selling print books (since 2000) and electronic books (since 2007). We ran automated PHP scripts to collect characteristics of print and electronic books (including rank, author, ISBN and so on) that entered at least once in the respective monthly top 100 list. Top print books were followed from July 2000 to July 2010, and top e-books from November 2007 (first available date of the archives) to July 2010. Amazon is the leading seller of e-books in the US.

Overall, the dataset covers a period of 121 months for print books and 33 months for e-books. In the rest of the article, we mainly focus on the period where both print books and e-books are available, *i.e.* November 2007 to July 2010.

2.1 Data collection

The characteristics of the books are the following: title, author, ISBN (for the print format) or ASIN (for the electronic format) which stands for Amazon Standard Identification Number, monthly sales rank (from 1 to 100), genre, publication date, observation date, the average rating of customers, and price. It is important to stress several data issues related to online data collection of book characteristics. First, the same book can have a different title between its print and electronic versions; for instance, "Harry Potter 7" and "Harry Potter seven". These two titles are the same and we treat them as such in our database by recoding them to a single title. Secondly, we face a similar problem for the name of the author of a title that might be spelled differently in each format: "J.K. Rowling" and "Joanne Rowling" for instance. We also recoded these fields to a single field. Thirdly, there are different versions of the same book like "soft paperback", "hard paperback" versions, limited edition, audiobook edition, etc. These versions have different unique identifiers (ISBN or ASIN) but correspond to the same content. In what follows, we treat these different versions as a single informational content, which is what we are eventually interested in. At this point, we have a unique identifier for a book that is the combination of the author's name and title. Fourthly, the publication date is sometimes misreported on Amazon websites. Indeed, Amazon.com reports a publication date of a specific version of a print book or an e-book that depends on the edition. While this is a minor issue for recent titles, we need to correct publication dates of books that were reprinted or that exist in different versions. We define the publication date of a title as the earliest publication date of existing versions sold on the Amazon website. For books in the public domain, we use the publication date of the first edition collected from Wikipedia. We have excluded books that were published before the 19th century to avoid biases due to a small number of observations. For reference, 7 out of 2340 unique titles can be considered "historical". Finally, books are assigned to genre categories on Amazon. We have grouped these genres in 6 categories for e-books: Practical (including self-help and hobbies) essays, science and biography (non-fiction); books for young readers; guidebooks and how-to; fiction; reference books and textbooks. Print books have an additional category: "Comics and graphic novels". A title that belongs to several categories is assigned to the category under which it had the highest sales rank. When Amazon does not assign a category to a title, we used keywords to assign a category. To sum up, the categories that we have assigned to books does not necessarily match the categories listed on the website.

2.2 Descriptive statistics

1,244 e-books appeared at least once in the top 100 list of best-selling titles versus 1,097 print books. There are only a few cases of multiple versions of the same book as the total number of unique titles is fairly similar (1,238 for e-books and 1,041 for print books). We observe

more e-books than print books due to faster turnover of e-books in the top 100 list (2.5 months average residence time against 3 months for print books). The average publication date of e-books is significantly older than for print books, 2002 vs. 2005. The oldest e-book title in our dataset dates to 1811 (while excluding titles published before 1800). The oldest print book was published in 1900. It thus seems that older releases can still be successful in new electronic formats.

Market	Print book	E-Book	Unique content
Number of items (ISBN or ASIN)	1,097	1,244	-
Number of unique books	1,041	1,238	1,861
Number of unique authors	761	833	1,290
Number of books by author	1.44	1.49	-
Average date of release[12]	2005	2002	-
Minimum date of release	1900	1811	-
Average rank of entry	56.4	50.2	-
Average rank of exit	62.6	61.8	-
Average best rank	46.2	44.3	-
Average lifespan (in month)	3.0	2.5	-
Average number of comments by book	4.1	3.8	-
Average Amazon price	16.0	10.9	-
Average Amazon price: hardcover (n=780)	17.6	-	-
Average Amazon price: paperback (n=299)	11.8	-	-

Table 1 : Descriptive statistics

Print and electronic books belong to specific genres, print books having more varied genres (Figure 1). The "fiction" category (27%) is the main category of print books, followed by "practical" (24%), and "non-fiction" (23%). Electronic books are predominantly "fiction" (70%), followed by "non-fiction" (12%) and "practical" (8%). The category "reference and textbooks" is more popular among print books than electronic book readers. A really small fraction of best-selling books correspond to "Comics", currently unavailable in electronic format. Finally, electronic books are on average cheaper than print books: USD 10.9 vs. USD 15.9.

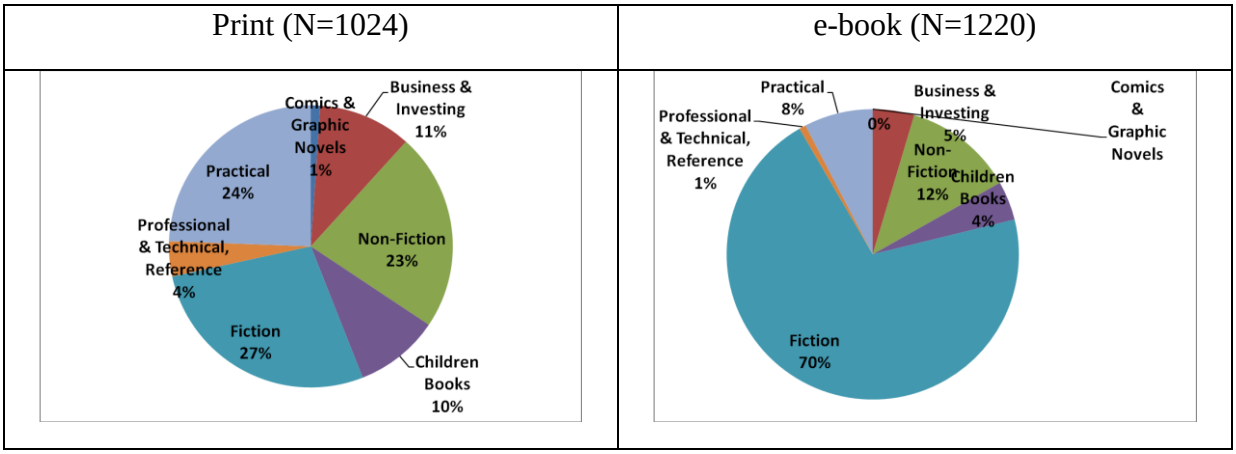

Figure 1 : Genres of print and electronic books

[12] We excluded all dates prior to 1800 for this calculation, a total of five Kindles and 2 paper books.

We now compare the list of best-selling print books to the corresponding list of e-book titles. Among the 1,861 titles in our dataset, 1,515 or 81.4% are available in electronic and print formats, 79 or 4.2% are only available in kindle edition and 267 or 14% are only available in print format. Among the 79 electronic books that do not have a print equivalent, we find two categories of titles: classics and pure e-books. The first category of titles (27 e-books) are old titles released in print with a specific edition and which have not the exact counterpart on Amazon kindle store. The second category (54 titles), are specific to the kindle edition and often correspond to "Harlequin-like" e-books (see below).

We now analyze the list of titles available in print and electronic format. Among the 1,515 unique titles available in both formats, 418 (or 27.6%) have appeared at least once in the monthly top 100 best-selling list of both print and electronic books. We refer to these books as belonging to the common list. 741 titles (or 48.9%) have appeared only in the monthly top 100 best-selling list of electronic books and 356 titles (or 23.5%) only in the monthly top 100 best-selling list of print books. In other words, a large fraction of successful titles are specific to a format (72.4% of all the 1,515 unique titles), in addition to the 346 (= 267+79) books that are only available in one format. In addition, for the 774 (= 418 + 356) print titles also available in kindle format, 46% have never entered the monthly top 100 list of best-selling electronic books. Conversely, 64% of e-book titles (out of 1,159) have never entered the monthly best-selling list of print books. Figure 2 illustrates this phenomenon.

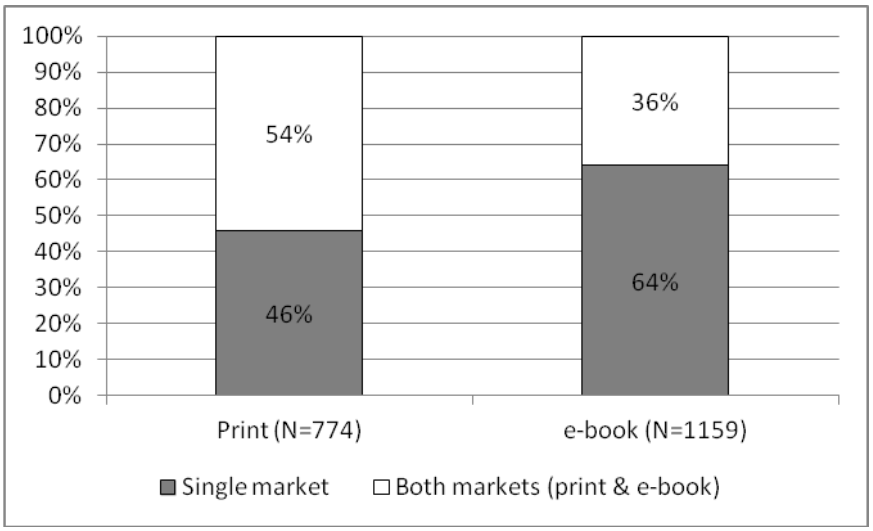

Figure 2 : Percentage of titles in both monthly print and electronic top 100 lists

The electronic market is characterized by both superstar and long tail properties. In Tables 2 and 3, we have divided the set of all books into 3 categories. These are defined by the best monthly sales rank achieved by a book during the period of observation. The first category includes books whose best ranking over all observations was in the top 33; the second, books whose best ranking was between 34 and 66; the third category is books whose best ranking was between 67 and 100. In Table 2, the first column corresponds to print books that never entered the monthly top 100 list of e-books, namely "print preferred". For instance, in Table 2, 623 print books (out of 1,041 or 60%) never entered the list of monthly best-selling e-books. In Table 3, the first column corresponds to e-books that never entered the monthly top 100 list of print books, *i.e.* "digital outsiders".

	Print preferred	% total	Superstars	% Total	Total
Top 1 to 33	193	18.6%	209	20.1%	402
Top 34 to 66	212	20.4%	130	12.5%	342
Top 67 to 100	218	21%	79	7.5%	297
Total	623		418		1,041

Table 2 : Comparison of sales ranks of print books that entered the top 100 list of best selling e-books with those that did not

We use the new ranking to characterize the 418 titles that are both popular in print and electronic formats. We first find that these titles were written by well-known authors who have published many best-sellers. Superstars in the print format remain superstars in the new electronic format. Secondly, we observe that within the top 33 best-selling e-books, 303 titles (24.5% of the available e-books in our dataset) never entered in the list of monthly best-selling print books (against 17.8% of the books that entered both top 100 lists). These titles that we name "digital outsiders" are lesser known and correspond to a digital "long tail". This suggests that there is not a complete cannibalization of print books by e-books.

	Digital outsiders	% total	Superstars	% Total	Total
Top 1 to 33	303	24.5%	220	17.8%	523
Top 34 to 66	244	19.7%	129	10.4%	373
Top 67 to 100	272	22%	69	5.6%	341
Total	819		418		1,238

Table 3 : Comparison of sales ranks of e-books that entered the top 100 list of best selling print books with those that did not

To conclude this section, we would like to stress that a proportion of e-books (4.2% or 79) do not exist in print format. These titles were on average published in 1998 (1990 if we include titles published before 1800) and correspond to two types of books. On the one hand, we have old releases and titles that are no longer available in print format. On the other hand, we have new titles that are first published in electronic format to test the market and then possibly published in print format. In the latter case, electronic books represent a new way to promote and sell new titles. For instance, in Table 4 we give examples of electronic books that are not available in print on the Amazon or marketplace websites.

Title	Writer	Publication year	Category	Best rank	Average customer review	Average lifespan (in month)
72 Hours	Shannon Stacey	2006	fiction	3	3	1
My Soul To Lose	Rachel Vincent	2009	children's book	5	3.5	9
The Babysitter'S Code	Laura Lippman	2008	fiction	6	2	1
Icy Heat: A Heat Series Story	Leigh Wyndfield	2008	fiction	7	4	1
When Night Falls	Margaret Daley	2009	fiction	10	2	6
Haley'S Cabin	Anne Rainey	2007	fiction	12	2.5	1
Talking With The Dead	Shiloh Walker	2006	fiction	14	3.5	2
Thin Blood	Vicki Tyley	2010	fiction	15	4	2
Look What Santa Brought	Annmarie McKenna	2007	fiction	16	3	1
Believe	Daniel Oran	2007	fiction	24	4.5	2

Table 4 : Example of e-books not available in print on Amazon

To briefly summarize this section, we found in Tables 2 and 3 that the demand for print and electronic titles are rather different and that there are more superstar titles that belong to both top 100 list, while there is a significant number of top selling electronic titles that do not have the same popularity in print format. In the following section, we formally test the existence of superstars and digital outsiders using econometric methods.

3 Superstar titles

In this section, we analyze the factors that determine whether a successful print book is also successful in electronic format. We are especially interested in best-selling print books that enter the monthly top 100 list of best-selling e-books. We run a regression on the sample of print books that have an e-book equivalent (1,041 unique print books – 267 print books that do not have an e-book equivalent) and that entered the monthly top 100 list at some date during our period of observation.

The methodology is the following. We explain the probability that a print book in our dataset belongs to the top 33 percentile of best-selling e-books. The dependent variable is a binary variable that is equal to 1 if the title has reached an *average* rank below the 33rd percentile during the full period of observation and is equal to 0 otherwise. We use a linear probability model that assumes that the probability of observing the dependent variable equal to 1 is a linear function of a set of explanatory variables.[13]

We use most characteristics of a book as explanatory variables. In addition, we create the variable *DIFFPRICE* that computes the difference between the price of the print book and the price of the electronic version. Recall that since we treated the different versions of a print book, namely hardcover and softcover (paperback/mass paperback), as a single informational content, the price of a print book can be related to its softcover or hardcover format (see the previous Data Collection section). As a result, the price difference that we compute can either be attributed to a difference between a paperback and an e-book or between a hardcover and a

[13] The linear probability model has the advantage of giving directly the partial effects and their standard deviations. We have checked the robustness of the results by estimating probit models as well. Results are similar and available upon request.

e-book version.[14] Therefore, this variable is not homogenous among observations. Nevertheless, we expect that a larger price difference increases the probability that a print title is very successful in the Kindle store. Next, we compute the number of different books published by the same author over the period of observation. We expect that the more published the author the higher the probability that a print book is also successful in electronic format. We also use the rating of a title by customers (between 0 and 5) and hypothesize that higher ratings should increase the probability that the dependent variable is equal to 1. We finally introduce a variable that computes the best rank reached by a title on the Amazon website. This variable is defined between 1 and 100, 1 being the best possible rank. All three of the preceding variables capture a superstar effect in the print book market. Next, we control for the number of days that a title has been available on the shelves, *TOM* for time on market, by calculating the difference between the day the title exited the monthly top 100 list and the first publication date. Finally, we use several binary variables to account for the different genres that we defined in the previous section; the base category is "fiction"; the other categories are "non-fiction", "reference and textbooks", "practical"; "guidebooks and how-to", "young readers".

We test two specifications (Model A and Model B). In model A, we estimate the probability that our dependent variable equals 1 over the sample of all print books that enter the monthly top 100 list of best-selling e-books. We specify the probability of observing a superstar as:

$$p_i = P(y_i = 1) = \mathrm{x}_i'\beta + \varepsilon_i$$

where y_i equals 1 with probability p_i and 0 with probability $1 - p_i$, x_i is the set of explanatory variables and ε_i is an unobserved variable with mean 0 and variance σ^2. Because we are missing observations for the variables average rating of comments and time on market, we end up with 806 observations.

In Model B, we estimate the coefficients of the regressions on all print titles that exist in electronic version and for which both prices exist. This constraint on prices reduces the number of observations further. Before we comment on the estimation results, we note that the distribution of the number of books by author is asymmetric with 75% of authors publishing less than 2 books and 5% publishing more than 6 (Table 5).

Model A	mean	min	max	p 75	p 95	nb observations
Nb. of electronic books by author	1.5	1	17	1	4	567
Minimum ranking	45.4	1	100	70	93	774
Tom (in months)	33.9	-4	2669	17	126	773
Average rating	4.05	2	5	4.5	4.5	771
Model B						
Number of electronic books by author	1.99	1	17	2	6	227
Diffprice	3.03	-5.89	18.92	5.83	9.76	332
Minimum ranking	36.73	1	100	56	85	332
Tom (in month)	24.64	-1	1324	22	81	332
Average rating	3.91	2.5	5	4.5	4.5	332

Table 5 : Descriptive statistics of the models A and B

[14] For a small number of cases, the softcover and the hardcover formats of the same title entered together in the monthly top 100 list of best-selling books. In this case, we computed the average price of both formats to obtain the price of the paper version.

In addition, we observe that the distribution of the price difference (*diffprix* in Figure 3) is bimodal with two peaks at USD -1 and USD 5. This is precisely due to what we explained before on the existence of two formats for print books, namely softcover and hardcover.

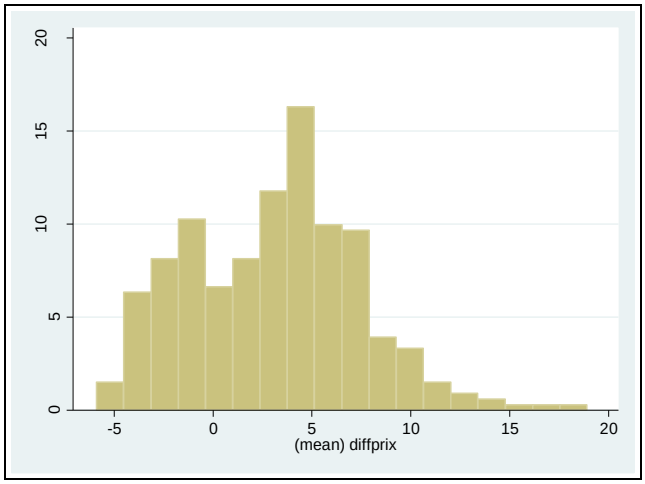

Figure 3 : Distribution of price difference print/electronic

Table 6 gives the estimation results that confirm the descriptive statistics of the previous section. The lowest (best) ranking has a negative effect on the probability that a successful print book is also successful in electronic format. For instance a variation of (-)45 ranks (from top 1 to 46), the probability increases by 27% (-45×-0.006) in Model B.

Dependant variable	Probability that an e-book is also a superstar		Probability that an e-book is also a superstar	
Model	A		B	
	Coeff.	Std. Err.	Coeff.	Std. Err.
Minimum Ranking	-0.00342***	(0.000424)	-0.00586***	(0.000908)
Fiction (reference)				
Non-fiction	-0.236***	(0.0331)	-0.256***	(0.0586)
Reference and textbooks	-0.241***	(0.0818)	-0.298	(0.240)
Practical	-0.260***	(0.0349)	-0.267***	(0.0767)
Guidebooks and how-to	-0.230***	(0.0421)	0.0226	(0.121)
Young readers	-0.232***	(0.0531)	-0.204**	(0.0966)
Tom	-0.0000534	(0.0000736)	-0.000195	(0.000249)
Average rating of	-0.0391	(0.0250)	-0.0355	(0.0453)
Nb. of books by author	0.0173***	(0.00395)	0.0188***	(0.00596)
Diffprice			0.0153***	(0.00557)
Constant	0.602***	(0.101)	0.669***	(0.186)
R^2	0.253		0.249	
N	770		332	

Table 6 : Estimation results: – Probability that a print book belongs to the top 33 percentile of best-selling e-books

This superstar effect is confirmed by looking at the effect of the number of books by an author on the probability of e-book success: for an author with 12 books, this probability increases

by 22.5% (12×0.0188). Fiction is the most common genre for electronic books and this translates to negative coefficients for all other genres (-25%). In Model B, we only keep observations for which *DIFFPRICE* is not missing. We find a negative and significant effect of the price difference on the probability of e-book success; for instance, for the peak of the distribution at USD 5, the probability increases by 7.5% (5×.015): an e-book that cost USD 5 less than its print version increase the probability of belonging to the first 33rd percentiles by 7.5%. This effect would be reinforced if we only included hardcovers that have a higher price than paperbacks. Finally, the time on market of books does not significantly influence the probability of e-book success.

4 Digital outsiders

In this section, we analyze successful electronic books that do not appear in the monthly top 100 list of best-selling print books. We analyze the factors that influence the probability of that event. As explained in section 2, we refer to these titles as digital outsiders. The sample of books is the set of all electronic books that exist in print format (in the previous section, we analyzed the set of print books that exist in electronic format). The dependent variable is a binary variable that is equal to 1 if the electronic book has never appeared in the monthly top 100 list of best-selling print books and 0 otherwise. We estimate a linear probability model.

In addition to the explanatory variables used in the previous section, we include in the regression the following variables. First, we create two variables that capture the price difference between the electronic version of a book and the paperback version (*DIFFPRICE1*)[15]. Next, we compute the number of electronic books published by authors over the sample period. We also capture this superstar effect by computing the best monthly rank achieved by a title during the whole period. This rank varies from 1 to 100. Finally, we use 4 categories to describe the first publication date: recent (published after 2005), contemporary (from 2000 to 2005), old (before 2000) and public domain (classified as such by Amazon).

Table 6 describes electronic books in the dataset. There are 550 authors who publish 1.6 books on average with noticeable differences between authors, since the 75th and the 95th percentile correspond to 3 and 8 books respectively.

Variable	Mean	Min	Max	p 75	p 95	Nb. observations
Nb. of electronic books by author	1.63	1	24	3	8	550
Best ranking	43.19	1	100	68	93	774
Time on market (days)	87.02	-3	3411	17	224	774
Contemporary books	0.084	0	1			774
Old books	0.0633	0	1			774
Public domain	0.0349	0	1			774

Table 7 : Descriptive statistics of the variables

Moreover, we have four variables that are related to the first date of publication of a title. The first variable is computed as the difference between the month the book entered the top 100

[15] We also run the same regression by using the price difference between the electronic book and its *hardcover* version. The results remain the same and are not reproduced in this version. Diffprice1 is limited to 774 observations since only 774 e-books had a price for its kindle version and its paperback version at the time of the data collection.

list and its first publication date (Model 3a in Table 8). Next, we create 3 binary variables that account for contemporary books (8.4%), old books (6.3%) and public domain (3.5%). The first two binary variables are included in Model 3b and the public domain binary variable is used in model 3c.

Table 7 gives descriptive statistics about the prices of electronic books. First, electronic books and paperbacks are similarly priced (0.7 USD between the two formats). Secondly, the price of a title increases with its monthly ranking. Hence, the price of titles at the bottom of the monthly top 100 list is 1.3 higher than the price of the best-selling titles. Thirdly, electronic books that are not sold in print format through the Amazon websites (including Amazon *Marketplace*[16]) are significantly cheaper than others (5.47 USD vs. 10.9 USD respectively). These titles are mainly ranked in the first group of best-selling titles (TOP 33). This suggests that consumers are very sensitive to prices of electronic books and that new marketing techniques (such as price reductions with promotional links) to exploit titles in the long tail make a lot of sense.

Variable	mean	min	max	p 75	p 95	Nb. observations
Diffprice1 (Kindle-paperback)	0.69	-18.46	14.19	3.04	7.31	774
Electronic price top 33	10.16	0	19.98	17	224	345
Electronic price top 66	11.15	0	19.98			308
Electronic price top 100	11.47	0	20.99			326
Price of electronic books that are not sold in print format	5.47	0	9.53			8

Table 8 : Descriptive statistics of the price variables

We specify the probability of observing an outsider as:

$p_i = P(y_i = 1) = x_i'\beta + \varepsilon_i,$

where y_i equals 1 with probability p_i and 0 with probability $1 - p_i$, x*i* is the set of explanatory variables and ε_i is an unobserved variable with mean 0 and variance σ^2.

Estimation results are reported in Table 8. Decreasing the best rank by one (increasing its popularity) has a negative effect on being a digital outsider. In other words, digital outsiders are more likely to be in the middle or the bottom of the charts. Secondly, digital outsiders are more likely to belong to the fiction category since all the coefficients with the other categories are negative. Next, the number of electronic books by authors increases the probability that a best-selling electronic book is also a best-selling print book. In addition, the price difference between the electronic and the paperback version decreases the probability of being an outsider. In other words, when the price of a paperback is lower than the price of the Kindle version, the probability of reaching a high rank in the top 100 list of best-selling print books increases. Finally, old titles and books that have been on the market for a long time have a lower probability of becoming high sellers. In the three specifications (models 3a-3c), the variables related to the first publication date decrease the probability to become a print best-seller. New electronic distribution channels favor the renewal of old titles.

[16] Amazon *MarketPlace* is a platform that offers new and used cultural goods by individuals and professionals.

Electronic Outsiders	3a		3b		3c	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Minimum ranking	0.0046029***	0.00055	0.0045559***	0.0005527	0.0046068***	0.0005504
Fiction (reference)						
Non-fiction	-0.3619841***	0.0462848	-0.3555218***	0.0464852	-0.3602719***	0.0463558
Reference and textbooks	0.307701	0.2200137	0.3115275*	0.2201278	0.3062269*	0.2201842
Practical	-0.1564586***	0.0620218	-0.156972***	0.0620991	-0.1524583***	0.0621171
Guidebooks and how-to	-0.5157635***	0.0831962	-0.5223381***	0.0836213	-0.5165165***	0.0832586
Young readers	-0.1798194***	0.0797264	-0.1812723***	0.0798364	-0.1856402***	0.0798905
Nb. e-books by author	-0.0270077***	0.0046739	-0.0270054***	0.0046796	-0.0269479***	0.0046792
Diffprice1	-0.008346***	0.0037466	-0.0093724***	0.0037248	-0.0082351***	0.00376
Tom	0.0001566***	0.0000436				
Contemporary (2000-2005)			0.1061116**	0.0572066		
Old (before 2000)			0.2148113***	0.0658561		
Public domain					0.3001125***	0.0877897
Constant	0.5510354***	0.0347051	0.544235***	0.0351264	0.5535055***	0.0346658
R^2	0.2228		0.2233		0.2216	
N	774		774		774	

Table 9 : Estimation results – Probability of being an outsider

5 Conclusion

In this article, we have compared the best-selling list of print and electronic books sold on the US Amazon website from November 2007 to July 2010. Many people from the industry have feared the cannibalization of print sales by electronic books. If there is cannibalization we should see high cross-price elasticity between electronic and print. Our estimation results show high cross-price elasticity for a portion of the top selling print books. However, the Internet channel also creates a market expansion effect for titles that we referred to as digital outsiders. The latter fall into one of two categories: e-books that are not popular in print or are older and e-books that do not have a print equivalent (in our dataset, 4.2% of electronic books correspond to titles that do not exist in print format). The first category corresponds to best-selling electronic books that never entered the monthly top 100 list of best-selling print books during the observation period of our study. It includes print titles from the public domain or released before 2000; for these titles, the electronic format increases the duration of popularity of print books and contributes to the renewal of old titles. There are two possible reasons why we are seeing older books appearing as best selling e-books. First, people who already own paper copies of these books are now buying e-book versions because they value unique properties of the e-book version (searchability, portability, ease of bookmarking/annotating). If this is the explanation, it is a short-lived (e.g. 5-10 year) phenomenon that will disappear once paper has been replaced by e-books. This phenomenon would be similar to converting one's CD library to iTunes. Secondly, there is a continuing interest in the classics. Readers who are new to these books are more likely to acquire their first edition as an e-book (lower cost) rather than as print, thus raising the ranking of the e-book sales so they appear in our top 100. This phenomenon should be ongoing, and is another instance of cannibalization. Indeed, Amazon offers public domain titles as e-books at a price of zero. The second category of digital outsiders corresponds to electronic books written in majority by contemporary authors specializing in the romance genre (Harlequin-like). Online distribution of electronic books

allows them to cheaply self-release new titles and to test the potential readership of each title, leading to a print version for successful electronic titles.

The existence of digital outsiders has both research and strategic implications.

First, our results direct research on the long tail in a new direction. Current research has mainly focused on the Internet as a new channel for distributing physical products in niche markets. Our results show that titles published before 2000 and new titles published exclusively in electronic format challenge the standard view of the long tail theory. Secondly, the increasing commercial duration of a title has implications for marketing research, requiring new ways to measure cumulative audiences and readership of titles that are sold in different formats over a long period of time.

Secondly, our results have managerial implications as well for writers, publishers and online intermediaries. On the one hand, the new electronic format creates an opportunity for emerging niche or independent authors to reach their audience as well as to test new markets and pricing strategies. Successful writers of best-sellers in print format can see the electronic format as a mean to increase revenues from new readers who have acquired an e-book reader. New readers are also exposed to older titles that benefit from the increasing life "on the shelves". On the other hand, online intermediaries such as Amazon see the new electronic format as an opportunity to publish authors who may not have signed a contract with a traditional publisher, as well as a new way to differentiate their services by offering both print and electronic versions of a title. It remains to be seen who will provide the traditional publisher/editor services if authors directly go to digital. One possibility is that e-book distributors, such as Amazon, will integrate backwards into editing/publishing, capturing more of the value added. However, editing services are far from an e-book distributor's core competencies, and it is more likely that either traditional publishers, or new online intermediaries will eventually come to dominate over e-book distributors. In any event it seems probable that traditional publishers are likely to lose bargaining power as authors find that they can self-publish their electronic books using online intermediaries.

6 References

Anderson, C. 2004. The Long Tail. Wired, October.

Balasubramanian, S. 1998. Mail Versus Mall: A Strategic Analysis of Competition Between Direct Marketers and Conventional Retailers. Marketing Science. 17(3), 181-195.

Biyalogorsky, E., Naik P., 2003. Clicks and Mortar: The Effect of On-line Activities on Off-line Sales. Marketing Letters. 14(1), 21-32.

Bounie, D., Eang, B., Waelbroeck P., 2011. Les plateformes de ventes sur Internet: une opportunité pour les industries culturelles. Revue Economique. 62(1), 101-112.

Brynjolfsson, E., Hu, Y., Smith M.D, 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. Management Science, 49(11).

Chevalier, J., Goolsbee A., 2003. Measuring Prices and Price Competition Online: Amazon and Barnes and Noble. Quantitative Marketing and Economics. 1(2), 203-222.

Danaher, B., Dhanasobhon, S., Smith, M.D., Telang, R. 2010. Converting Pirates without Cannibalizing Purchasers: The Impact of Digital Distribution on Physical Sales and Internet Piracy. Marketing Science. 29(6), 1138-1151.

Deleersnyder, B., Geyskens, I., Gielens, K., Dekimpe, M.G, 2002. How Cannibalistic is the Internet Channel? A study of the newspaper industry in the United Kingdom and The Netherlands. International Journal of Research in Marketing. 19(4), 337-348.

Elberse, A., Oberholzer-Gee, F., 2007. Superstars and underdogs: An examination of the long tailphenomenon in video sales. Marketing Science Institute, 4, 49-72.

Ghose, A., Smith, M.D., Telang, R. 2006. Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact. Information Systems Research. 17(1), 3-19.

Huang, L.Y., Hsieh, Y.J., 2012. Consumer electronics acceptance based on innovation attributes and switching costs: The case of e-book readers. Electronic Commerce Research and Applications. forthcoming.

Jiang, Y., Katsamakas, E., 2010. Impact of e-book technology: Ownership and market asymmetries in digital transformation. Electronic Commerce Research and Applications. 9, 386-399.

Oestreicher-Singer G., Sundararajan A., 2010. Are Digital Rights Valuable? Theory and Evidence from Ebook Pricing. CeDER Working Paper No. 06-01 Working Paper Series.

Rosen, S. 1981. The Economics of Superstars. American Economic Review. 71 (5), 845-858.

Waldfogel, J., 2009. Lost on the web: Does Web Distribution Stimulate or Depress Television Viewing? Information Economics and Policy. 21(2), 158-168.