Amazon's Price and Sales-Rank Data

Jörn Boehnke & Brock Mendel

June 12, 2018

Abstract

This data set contains weekly price and sales-rank information for more than 400,000 products from a broad selection of product categories on Amazon.com. The data set ranges from 2010 to 2016 and includes, for each product, product title and product category. Additionally, most products have been successfully matched to their Universal Product Codes (UPC) and International Article Numbers (EAN) which allows researchers to link these data across different data platforms.

1 Introduction

Amazon.com is the dominant e-commerce company in the United States, accounting for roughly one quarter of the online retail market (Hadad, 2017). Given its importance, a representative analysis of online retail should include Amazon data. However, it is difficult to obtain reliable historic price and sales data as Amazon does not make this data available to outside researchers. We side-stepped this problem by collecting data from several price-tracking websites that display Amazon's historical prices in graph form.[1] We collected 13,092,414 of these graphs and successfully extracted the underlying time-series data from 12,158,783 of these. In addition to prices, for many products we observe the sales-rank, an ordinal describing the popularity of a product that can be used as a proxy for quantities sold. The combined data set contains weekly price and sales-rank information for 400,000 products offered on Amazon.com from 2010 to 2016.

In this article we introduce the data collected on Amazon's pricing patterns and sales-rank information. The structure of our Amazon data is very similar to that of Nielsen's Retail Scanner Data, and UPCs are provided whenever possible. This will enable researchers to link both data sets and answer questions of online and offline consumer behavior. More generally, these data are ideal for studying online pricing patterns and strategies. As a market platform Amazon hosts millions of 3rd-party sellers that independently set prices and sell goods.[2] The data contain both Amazon's own price and the price of 3rd-party sellers for the same items on Amazon.com.

The remainder of this article is structured as follows. Section 2 details the collection and structure of the data set and presents different approaches to translate sales-rank data to quantities. Section 3 discusses selected summary statistics and section 4 describes the potential uses of the data set for research projects and its limitations. We conclude with a short description of how to download and use the data.

[1] Specifically, most of these sites display Amazon's own price, the price of the cheapest 3rd-party seller offering the new product on Amazon's market platform, and the cheapest 3rd-party seller offering the used product on Amazon's market platform.

[2] `https://seekingalpha.com/article/3962561-amazon-com-third-party-sellers-drive-profitability`

2 Data Set

In order to collect Amazon.com's historic price and sales-rank data, we collected graphs from the 3rd-party price and sales-rank tracking sites `www.camelcamelcamel.com`, `www.ipricetracker.com`, `www.keepa.com`, `www.pricezombie.com`, and `www.thetracktor.com`. These sites obtain their data directly from Amazon and present it to users in graph form (cf. figure 1 for an example of a price history graph). We analyzed these graphs pixel-by-pixel to retrieve the underlying data points. In addition to these graphs, we spot-checked the real-time data by directly checking the prices and sales-ranks on Amazon.com in intervals of two to three weeks. We collected over 13 million graphs reflecting time series data on Amazon.com prices and sales-ranks ranging from 2010 to 2016. Together these panel data cover over 400,000 products and amount to over 2.7 billion data points in total. The raw data have heterogeneous and non-constant sampling intervals, which we standardize where possible. The version most convenient for research is weekly pricing and sales-rank information. We discuss this process in detail below.

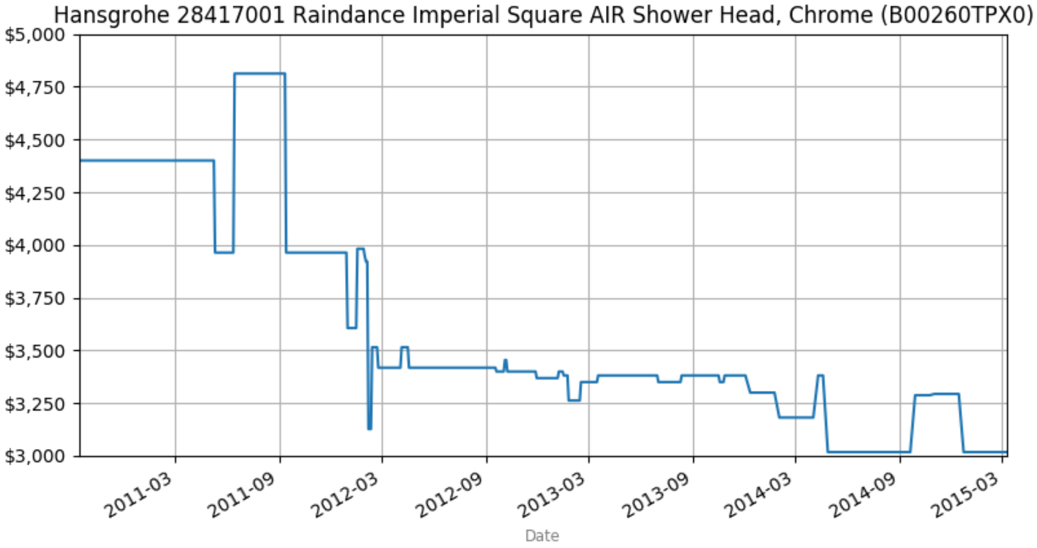

Figure 1: Price history sample graph.

Data Collection

We developed a web-crawler using Java, Python, and MySQL, and successfully downloaded 13 million graphs from Amazon price-tracking websites. Our crawlers utilized supervised learning techniques to overcome CAPTCHA[3] challenges posed by the websites. In addition, we developed a program using the OpenCV library[4] to back out price and sales-rank information from these graphs.

There is an element of measurement error in extracting numeric data from these graphs. We identify the vertical level of a pixel and compare it to the labels on the graph's y-axis. In many cases the pixel will lie on a grid line that is precisely marked. In others we have to interpolate the values between these markings. For products where prices move around frequently – or for which the time horizon is long so price movements appear very frequent on the graph – we may not be able to retrieve a single price, but only upper and lower bounds for the price. In order to mitigate these concerns, we form the weighted weekly average for all prices and sales-ranks.

We verified that the tracked price and sales-rank data matched those displayed on Amazon.com every two to three weeks. These checks correspond to a negligible fraction of the observations, but it is reassuring that they did not reveal any inconsistencies.

Data Layout

The data are stored in table form in two separate CSV files. Each product is uniquely identified by its `asin`, the Amazon Standard Identification Number. Table 1 describes the columns of the file "product_information.csv". The table contains the title and category for all items in our data set. Furthermore, it contains the UPC and EAN information for most items. All information is provided representing the time the information was collected. Amazon might have altered the category and title information since the period of data collection. The EAN and UPC information has been collected from `www.barcodelookup.com`, `www.barcodespider.com`, `www.ean-search.org`, `www.upcdatabase.com`, and `www.upcdatabase.org`. We only matched those EANs and UPCs that showed up in at least two different external product databases.

Table 2 describes the columns of the file "prices_and_ranks.csv". The table contains the time-period-weighted weekly average of prices and sales-rank.[5] Each row is identified by product `asin` and `week_end`, the date of Saturday completing the week. All prices are denoted in USD cents, making both prices and sales-rank integer values.

[3] Acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart"; test used online to determine whether or not the user is human, cf. Lillibridge et al. (2001).

[4] `http://opencv.org/`

[5] Prices and sales-rank are averaged over the period of one week. The average is weighted to adjust for different frequency at which the data were recorded. E.g., if we observed a week with one data point every 24h for the first 6 days and 3 data points for the remaining 24h, the average will be taken adjusting for the higher frequency of the last 3 observations.

Variable Name	Description	Data Type
`asin`	Amazon Standard Identification Number, Amazon's unique product ID	string of length 10
`title`	Amazon's product title	string
`category`	Amazon's hierarchical product category separated by "›"	string
`upc`	Universal Product Codes (UPC)	string of 12 digits
`ean`	International Article Numbers (EAN)	string of 13 digits

Table 1: Columns in "product_information.csv".

Variable Name	Description	Data Type
`asin`	Amazon Standard Identification Number, Amazon's unique product ID	string of length 10
`week_end`	last date of week, format YYYYMMDD	8 digit integer
`price_amazon`	weighted average of Amazon price in USD cents	integer
`price_new`	weighted average of 3rd party new price in USD cents	integer
`price_used`	weighted average of 3rd party used price in USD cents	integer
`sales_rank`	weighted average of sales-rank	integer

Table 2: Columns in "prices_and_ranks.csv".

A row contains the prices and sales rank of product `asin` during the week that ended on date `week_end` and is therefore uniquely defined by these two variables. Both of these variables are non-empty. Some of the four columns `price_amazon`, `price_new`, `price_used`, and `sales_rank` may be empty for a given `asin-week_end` pair. Unpopulated information is common, as there are many products for which only one type of seller – Amazon, 3rd-party new, or 3rd-party used – offers the product. For example, many items on Amazon.com are only available `new` such that `price_used` remains unpopulated. Moreover, a number of products are only offered by 3rd party sellers, in which case `price_amazon` is unpopulated. Furthermore, a price field will also be unpopulated if an item was out of stock. An item that has been introduced during our sample period will not have any data rows prior to its release date. Lastly, there are items for which we were unable to track the sales-rank over long periods of time.[6]

Translating Sales-ranks to Quantities

Amazon product pages display several sales-ranks for each good. These rankings are calculated across multiple subsets of Amazon products. For instance, at the time of writing, Robert Heinlein's "Stranger In

[6] We firmly believe that missing sales-rank observations occur non-systematically and can be treated as randomly missing observations.

A Strange Land" is ranked:[7]

- #120 in Books › Audible Audiobooks › Fiction & Literature › Classics
- #158 in Books › Audible Audiobooks › Science Fiction
- #176 in Books › Science Fiction & Fantasy › Science Fiction › Alien Invasion

The sales-rank is closely related to quantity sold: a lower sales-rank represents that this product sold more units than other products in the same category. For products where sales occur rarely (i.e. high-ranked products), the underlying quantity data can be determined by mere counting. There are stretches of time where the sales-rank undergoes an exponential-resembling decay; these are periods where the product is not selling. Subsequently the rank spikes up when a purchase is made, cf. figure 2.

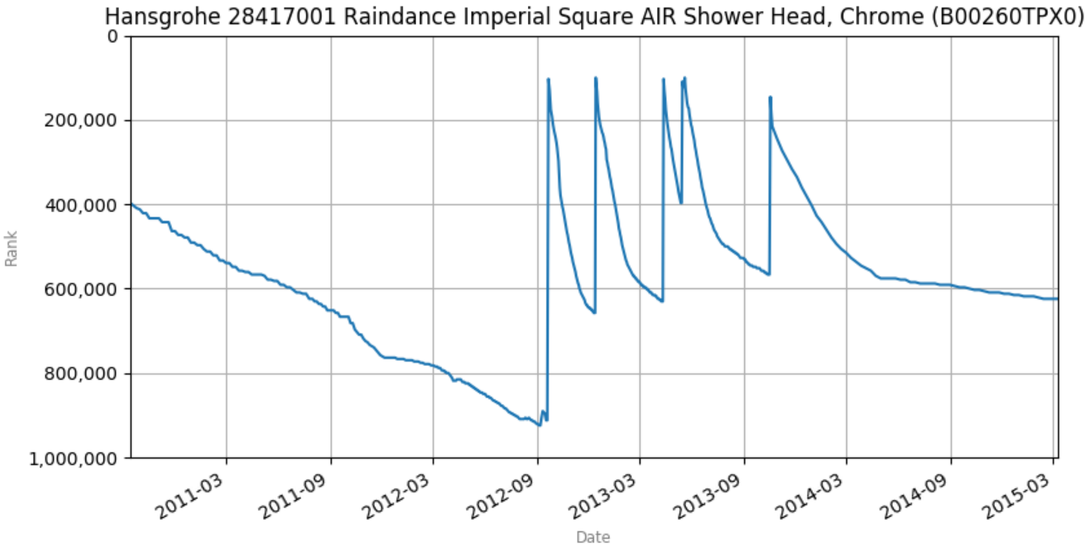

Figure 2: Sample sales rank history graph.

This "counting" does not work for products of lower sales-rank. Fortunately, there is a growing empirical literature that offers tools to map Amazon sales-ranks into product-level quantities. Perhaps the crudest way to transform sales-ranks into a quantity proxy is to use $-\log(\text{sales\_rank})$ (Archak et al., 2011). Chevalier and Goolsbee (2003) offer a more sophisticated approach to translate Amazon's sales-rank into quantity sold: they assume the probability distribution of sales to be Pareto and calibrate the resulting

[7] `https://www.amazon.com/Stranger-in-a-Strange-Land/dp/B00005QTH2` It remains a mystery though why Amazon classifies "Stranger In A Strange Land" as an alien rather than human invasion novel.

model, making quantity available as a dependent variable. Ghose et al. (2012) and Smith and Telang (2009) have used similar methods to derive quantities sold.

3 Summary Statistics

Table 3 displays the number of items for the largest ten Amazon categories in the data. Table 4 displays the number of items for the largest ten subcategories in "Books."

Category	Count
Books	121,299
Clothing, Shoes & Jewelry	41,747
Home & Kitchen	38,899
Electronics	36,573
Beauty & Personal Care	30,626
Sports & Outdoors	23,515
Movies & TV	21,361
Toys & Games	20,355
Tools & Home Improvement	17,880
Health & Household	14,179

Table 3: Top 10 Amazon categories.

Books' Subcategory	Count
Science & Math	51,111
Children's Books	29,438
Arts & Photography	19,496
Literature & Fiction	4,358
Cookbooks, Food & Wine	2,290
New, Used & Rental Textbooks	1,893
Comics & Graphic Novels	1,882
Computers & Technology	1,613
Business & Money	1,230
Biographies & Memoirs	1,160

Table 4: Top 10 subcategories in "Books."

Table 5 shows the average Amazon, 3rd-party new, and 3rd-party used prices in U.S. Dollars for the top ten Amazon categories, with standard deviations in parentheses. The averages are taken over the entire observed "lifetime" of a product, not adjusting for inflation.

Figures 3 and 4 display the average Amazon, 3rd-party new, and 3rd-party used prices in the Books and Electronics categories, respectively. The cross-sectional averages were taken quarterly for all items with existing Amazon, new, and used prices.

4 Research Opportunities

The weekly Amazon and 3rd-party prices and sales-rank information for 400,000 products from a broad selection of product categories can be used to investigate questions regarding online shopping behavior (e.g., Bhatnagar et al., 2000 and Manchanda et al., 2006), firm pricing decisions (e.g., Jaimovich et al.,

Category	Price Amazon	Price 3rd-part New	Price 3rd-party Used
Books	54.95 (75.09)	57.53 (3,122.58)	24.88 (106.63)
Clothing, Shoes & Jewelry	135.24 (265.63)	229.01 (1,969.72)	100.89 (297.23)
Home & Kitchen	77.27 (154.07)	107.84 (2,541.94)	89.21 (518.71)
Electronics	240.20 (544.26)	288.73 (1,194.21)	199.85 (537.50)
Beauty & Personal Care	43.37 (315.12)	51.49 (367.55)	71.92 (1,896.74)
Sports & Outdoors	92.35 (230.51)	171.29 (5,261.29)	106.75 (1,110.11)
Movies & TV	23.38 (27.92)	27.82 (142.83)	22.18 (170.24)
Toys & Games	37.21 (83.07)	153.45 (9,897.94)	35.83 (136.41)
Tools & Home Improvement	90.81 (254.93)	149.05 (3,476.23)	96.46 (835.91)
Health & Household	26.40 (30.15)	29.79 (115.70)	34.51 (209.45)

Table 5: Average Amazon, 3rd-party new, and 3rd-party used prices in USD for the top ten Amazon categories. Standard deviation in parenthesis.

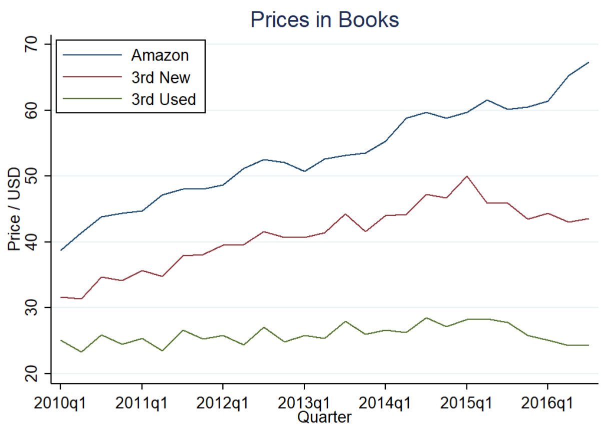

Figure 3: Average Amazon, 3rd-party New, and 3rd-party Used prices in the Books category

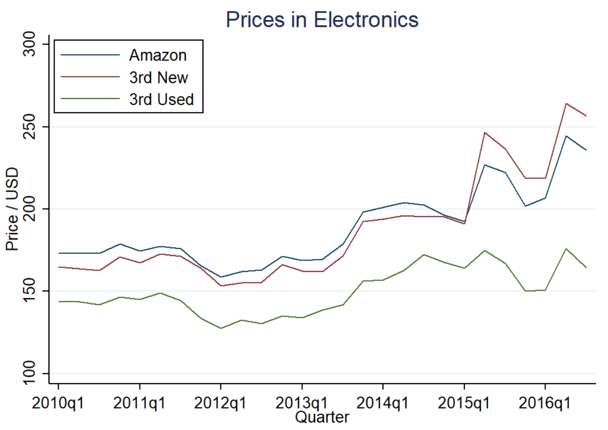

Figure 4: Average Amazon, 3rd-party New, and 3rd-party Used prices in the Electronics category

2014), intertemporal price discrimination (e.g., Nakamura and Steinsson, 2011), the economics of Amazon as a platform market and a retailer (e.g., Wang et al., 2013), the interaction between new and used prices (e.g., Houser and Wooders, 2006), product depreciation, etc.

The structure of our Amazon data is very similar to that of Nielsen's Retail Scanner Data.[8] The week_end variable is defined identically to that in Nielsen, and UPCs are provided whenever possible. The products in the Amazon categories "Home & Kitchen," "Movies & TV," and "Beauty & Personal Care" has a significant overlap with the Nielsen's product selection. This compatibility between Amazon's and Nielsen's data

[8] Nielsen's Retail Scanner Data covers 2.6 million grocery items, health and beauty aids, and selected general merchandise from more than 35,000 participating stores (https://research.chicagobooth.edu/nielsen/datasets#simple2).

enables researchers to link both data sets and analyze differences between online and offline consumer behavior.

More generally, these data are ideal for studying online pricing patterns and strategies. As a market platform Amazon hosts millions of 3rd-party sellers that independently set prices and sell goods. Recently, Amazon has started to subsidize prices of its platform's 3rd-party sellers for certain items in order to keep its overall pricing below that of competing sites.[9] Researchers may be interested in how Amazon and 3rd party seller pricing strategies interact, or in patterns of cross-product pricing behavior.

In addition to the marketing and industrial organization topics mentioned above, this data can be used by macroeconomists to measure price durations for the purpose of studying price stickiness.[10] Different firms use different pricing techniques and may vary their strategies over time. Which class of pricing strategy best describes actual firm pricing behavior is an empirical question, one which we hope these data can help address.

Limitations

It is worth explicitly noting some caveats about the data that may make it unsuitable for certain research purposes:

- All prices and sales-ranks are weighted weekly averages. The exact timings of changes in price and sales-rank have limited precision. Moreover, fluctuations in prices and sales-ranks may be lost due to the averaging (these issues are similar to e.g. Nielsen's Retail Scanner Data and BLS' Consumer Price Index data).
- This is not a representative sample of products. The sampling probability is weighted towards products that users explicitly search for.
- The data are not balanced over time. The number of items observed increases over most of the sample horizon.

[9] https://www.wsj.com/articles/amazon-snips-prices-on-other-sellers-items-ahead-of-holiday-onslaught-1509883201

[10] Most models can be categorized as either state-dependent or time-dependent (Klenow and Kryvtsov, 2008). State-dependent models include menu cost models (e.g., Mankiw, 1985). In these models the times at which prices are changed are exogenous and firms only choose the magnitude of price changes. Time-dependent models include Calvo (1983). In these models, firms endogenously choose which prices to change. State-dependent models are easier to motivate theoretically, but have – in general – the counterfactual implication that Monetary Policy is ineffective.

5 Accessing and Using the Data Set

The data set is available for academic use only. It can be downloaded from `https://www.dropbox.com/s/48kemxl07m4uvmh/data_sample.7z`.[11] Uncompressed, the data are stored in two CSV files that are approximately 5 GB in total size. The character set is UTF-8. Given the large data size, the researcher should either analyze these data on a machine with sufficient RAM, load the entire file into a database, or make use of streaming / partial file reading tools. While the file "prices_and_ranks.csv" is too large to be opened in Microsoft Excel, most statistical analysis packages such as R or Stata will be able to open and process it.

[11] This file contains a sample of the "Amazon price and sales-rank data" for evaluation purposes. We are in talks with the Kilts Center for Marketing at Chicago Booth to host the entire data set.

References

Archak, Nikolay, Anindya Ghose, and Panagiotis G Ipeirotis, “Deriving the pricing power of product features by mining consumer reviews,” *Management science*, 2011, *57* (8), 1485–1509.

Bhatnagar, Amit, Sanjog Misra, and H Raghav Rao, “On risk, convenience, and Internet shopping behavior,” *Communications of the ACM*, 2000, *43* (11), 98–105.

Calvo, Guillermo A, “Staggered Prices in a Utility-Maximizing Framework,” *Journal of Monetary Economics*, 1983, *12* (3), 383–398.

Chevalier, Judith and Austan Goolsbee, “Measuring prices and price competition online: Amazon.com and BarnesandNoble.com,” *Quantitative marketing and Economics*, 2003, *1* (2), 203–222.

Ghose, Anindya, Panagiotis G Ipeirotis, and Beibei Li, “Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content,” *Marketing Science*, 2012, *31* (3), 493–520.

Hadad, Jonathan, “IBISWorld Industry Report 45411a: E-Commerce & Online Auctions in the US,” 2017. [Online; accessed 2-Jan-2018].

Houser, Daniel and John Wooders, “Reputation in auctions: Theory, and evidence from eBay,” *Journal of Economics & Management Strategy*, 2006, *15* (2), 353–369.

Jaimovich, Nir, Marty Eichenbaum, Sergio Rebelo, and Josephine Smith, “How Frequent are Small Price Changes,” *American Economic Journal: Macroeconomics*, 2014.

Klenow, Peter J and Oleksiy Kryvtsov, “State-Dependent or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation?,” *The Quarterly Journal of Economics*, August 2008, *123* (3), 863–904.

Lillibridge, Mark D, Martin Abadi, Krishna Bharat, and Andrei Z Broder, “Method for selectively restricting access to computer systems,” 2001. US Patent 6,195,698.

Manchanda, Puneet, Jean-Pierre Dubé, Khim Yong Goh, and Pradeep K Chintagunta, “The effect of banner advertising on internet purchasing,” *Journal of Marketing Research*, 2006, *43* (1), 98–108.

Mankiw, N Gregory, "Small menu costs and large business cycles: a macroeconomic model of monopoly," *The Quarterly Journal of Economics*, 1985, *100* (2), 529–537.

Nakamura, Emi and Jón Steinsson, "Price setting in forward-looking customer markets," *Journal of Monetary Economics*, 2011.

Smith, Michael D and Rahul Telang, "Competing with free: the impact of movie broadcasts on DVD sales and internet piracy," *MIS Quarterly*, 2009, *33* (2), 321–338.

Wang, Xin, Feng Mai, and Roger HL Chiang, "Database submission – market dynamics and user-generated content about tablet computers," *Marketing Science*, 2013, *33* (3), 449–458.