Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Database Submission—Market Dynamics and User-Generated Content About Tablet Computers

Xin (Shane) Wang, Feng Mai, Roger H. L. Chiang

To cite this article:
Xin (Shane) Wang, Feng Mai, Roger H. L. Chiang (2013) Database Submission—Market Dynamics and User-Generated Content About Tablet Computers. Marketing Science

Published online in Articles in Advance 07 Nov 2013

. http://dx.doi.org/10.1287/mksc.2013.0821

MARKETING SCIENCE
Articles in Advance, pp. 1–10
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2013.0821

Database Submission

Market Dynamics and User-Generated Content About Tablet Computers

Xin (Shane) Wang
Department of Marketing, Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, Ohio 45221, wang2x5@mail.uc.edu

Feng Mai, Roger H. L. Chiang
Department of Operations, Business Analytics, and Information Systems, Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, Ohio 45221 {maifg@mail.uc.edu, roger.chiang@uc.edu}

Our Tablet Computer data set, collected from various websites, contains market dynamics related to 2,163 products, characteristics of 794 products, more than 40,000 consumer-generated product reviews, and information about 39,278 reviewers. The market dynamic information was collected weekly for 24 weeks starting February 1, 2012. Our Tablet Computer data set comprises four tables: the Market Dynamics of Products, Product Characteristic Information, Consumer-Generated Product Reviews, and Reviewer Information tables. In turn, it offers three unique properties. First, it contains both structured product information and unstructured product reviews. Second, it comprises product characteristic information and market dynamic information. Third, this data set integrates user-generated content with manufacturer-provided content. This integrated data set (available at http://pubsonline.informs.org/page/mksc/online-databases) is valuable for both academics and practitioners who conduct research related to marketing, information systems, computer science, and other fields using digital data readily available through the Internet.

Key words: market dynamics; online product reviews; tablet computers; user-generated content
History: Received: February 25, 2013; accepted: September 10, 2013; Preyas Desai served as the editor-in-chief and Puneet Manchanda served as associate editor for this article. Published online in *Articles in Advance*.

1. Introduction

Any marketing modeler who wants to use real data to test a model might have difficulties finding reliable, representative data sources. Information systems and computer science researchers who develop and evaluate text mining and sentiment analysis techniques to discover marketing intelligence from product reviews require representative sets of consumer-generated reviews. An emergent product (e.g., tablet computer) manufacturer that needs timely reports about competitors' consumers to improve and enhance the next version of its flagship product similarly must be able to identify ineffective strategies, without conducting time-consuming surveys.

In response to these demands and to encourage the use of user-generated content (UGC), we offer a tablet computer data set that contains product and review data collected from Amazon.com and manufacturers. It provides market dynamics describing 2,163 products (collected over a 24-week period in early 2012), characteristics of 794 products, more than 40,000 consumer-generated product reviews, and information about 39,278 reviewers. This data set can help academic researchers and practitioners understand the impact and value of digital data that are readily available online, as well as support the application of advanced data and text mining and sentiment analysis techniques to analyze UGC. Researchers can use this data set to conduct relevant research on product reviews, pricing, competition, new product development, and text analytics, for example. The method used to build the data set demonstrates not only that UGC is a viable external data source but also that this new data collection and integration approach can support research when reliable data are missing or cannot be easily obtained.

The use of structured data is common in mature marketing research, but because of the sheer volume and unstructured nature of market dynamics and UGC available on various websites, their use remains underdeveloped. For example, online product reviews contributed by consumers tend to be qualitative, which makes it challenging to quantify and convert the textual data into usable information. In the recent *Marketing Science* special issue on the emergence and impact of UGC (Fader and Winer 2012), two

separate contributions suggest new methodologies for analyzing UGC and thus discovering insightful marketing information. The ranking system implemented by Ghose et al. (2012) incorporates UGC directly into existing forms of product/service rankings, whereas Netzer et al. (2012) demonstrate that UGC can provide information about competitive market structures.

Lee and Bradlow (2011) have pioneered the integration of marketing and information systems disciplines through the use of UGC. They propose that text mining of online reviews might help automate the process of identifying the language that consumers use to describe products. Considering the opportunities to explore this external data source, we argue that potential collaborations among marketing scientists, computer scientists, and information systems researchers could facilitate its dissemination to marketing research (e.g., Das and Chen 2007, Lee and Bradlow 2011). Researchers thus could examine questions that have remained unanswered because of the lack of reliable and representative data. For example, an upcoming *Marketing Science* special issue on big data[1] seeks to promote research on big data analytics.

We present the structure and analysis of our Tablet Computer data set in §2. To illustrate the data set's significant, unique content, we include a list of figures and tables with summary statistics. In §3, we elaborate on three research opportunities available with this data set, and then we describe how to access and use it in §4. Section 5 concludes.

2. Tablet Computer Data Set

Existing research focuses on reviews of products such as digital cameras, movies, and smartphones; no data set comprises consumer-generated reviews of tablet computers. However, advances in information technology make tablet computers a prominent product category. A tablet is a one-piece mobile computer. Devices typically have a touch screen, with finger or stylus gestures replacing the conventional computer mouse. It is often supplemented by physical buttons or input from sensors such as accelerometers. According to a tablet buying guide published by *Consumer Reports*, tablets' main product attributes include their screen size and shape, wireless connectivity, display, operating systems, ports, and printing capability.[2]

Apple released its first iPad on April 3, 2010, which quickly became the first mobile tablet to achieve worldwide commercial success. In 2012, 31% of U.S. Internet users owned a tablet computer, up from 12% in 2011 (Moscaritolo 2012). The International Data Corporation (IDC) Worldwide Quarterly Smart Connected Device Tracker expects tablet shipments to surpass total PC shipments (desktop plus portable PCs) in the fourth quarter of 2013 and annually by 2015.[3] Among tablets available in 2012, the top-selling line of devices was Apple's iPad, with close 100 million sold by mid-October 2012; Amazon's Kindle Fire followed with 7 million, then Barnes & Noble's Nook with 5 million and the Google Nexus 7 with 3 million (Chen 2012). To reach wider audiences, more than 70% of mobile developers were targeting tablets in May 2013 (Developer Economics 2013). According to Bloomberg News (Kharif 2011), "On a single day in July, almost 18,000 fakes and clones resembling the iPad and Android devices were available for sale on 23 e-commerce sites." During the period we built this data set, more than 400 brands were listed under the tablet and tablet PCs category on Amazon—though that count includes a few miscategorized devices. The integration of consumer-generated product reviews with market dynamics related to this emerging product category can thus provide marketing researchers and practitioners with a valuable data source for answering interesting marketing questions.

In support of that objective, we collected data from the largest online retailer, Amazon, and tablet computer manufacturers. Based on the data we obtained, we were able to create the following four tables that make up our data set: the Market Dynamics of Products, Product Characteristic Information, Consumer-Generated Product Reviews, and Reviewer Information tables. Combined, the four tables offer three unique properties. First, they contain both structured product information and unstructured product reviews. Second, the data set comprises product characteristic information and market dynamic information. Third, the data set integrates UGC with manufacturer-provided content available on the Internet.

Compared with data sets of consumer-generated product reviews in prior research (e.g., Lee and Bradlow 2011), our offering provides at least four valuable contributions. First, it contains information about a prominent, emerging product category that has not been investigated previously. Second, the data set represents multiple, publicly accessible sources; it was built through data collection, cleaning, filtering, and integration processes, rather than just the simple, straightforward crawling. Third, it provides a large volume of consumer-generated product reviews (more than 40,000). Fourth, these product reviews are time-oriented, with the submission date specified. Therefore, this valuable data set can inform a range of marketing research issues, such as the price dispersion

[1] See the call for papers at http://dx.doi.org/10.1287/mksc.2013.0794.

[2] See http://www.consumerreports.org/cro/tablets/buying-guide.htm (accessed August 30, 2013).

[3] See IDC (2013).

in electronic markets, the multiple factors that affect overall sales (e.g., valences, volumes), and automated market structure analyses with text mining and sentiment analysis of product reviews. We elaborate on such research opportunities in §3.

Amazon features the tablets and tablet PCs category under its electronics and computers departments. A Java Web crawler obtained market dynamic information and product reviews from this category once weekly during February 1 through July 11, 2012. In addition, we manually searched the Internet and compiled manufacturer-provided product information about 794 tablet computers that prompted at least one review. The Tablet Computer data set thus offers two tables that organize product characteristic information and market dynamic information, respectively. We present detailed descriptions of the variables in our tables in §2.1 and then include the summary statistics of the data set in §2.2 to illustrate its significant and unique content.

2.1. Variable Descriptions

The variables related to the basic product and market dynamic information for 2,163 tablet computers, described in panel A of Table 1, are organized according to the classical marketing mix concept of the four P's (McCarthy 1960). For this study, "place" refers to the Internet platform, Amazon.com. The Web crawler gathered 11 variables related to "price" and "promotion" once weekly for the 24 weeks through Amazon's Web service, which provides programmatic access to product information.

Unlike traditional marketplace settings that provide only retail suggested and promotion prices, our data set reveals information about six price variables. *List_price* is the retail price suggested by the manufacturer; *Amazon_price* stands for the

Table 1 Product Information

Marketing mix (4 P's)	Variable name	Description
	Panel A: Market dynamics of products	
Product	*Item_ID*	Amazon Standard Identification Number (ASIN): Amazon assigns a unique identification number to each product
	Title	Title of the product
	Brand	Brand name of the product
	Model	Model number provided by the manufacturer
	UPC	Universal product code of the product
Price (24 weeks)	*List_price*	Retail price suggested by the manufacturer
	Amazon_price	Current selling price of the product
	Lowest_new_price	Lowest new product price quoted by merchants
	Lowest_used_price	Lowest used product price quoted by merchants
	Lowest_refurbished_price	Lowest refurbished product price quoted by merchants
	Trade-in_price	Trade-in price for used products quoted by Amazon
Promotion (24 weeks)	*Total_new*	Total number of merchants selling new products
	Total_used	Total number of merchants selling used products
	Total_refurbished	Total number of merchants selling refurbished products
	Num_reviews	Number of reviews submitted up to the collection day
	Sales_rank	Sales rank in the tablets and tablet PCs category
Place		http://www.amazon.com

Attribute name	Description	Attribute-level examples
	Panel B: Product characteristic information	
Item_ID	Amazon Standard Identification Number (ASIN)	
Date_first_available	Release date of the product	
Screen_size	Diagonal length of the screen	7, 8.9, 9.7, 10.1, or 10.4 inches
Wireless_type	Types of wireless protocol supported	802.11 a/b/g/n, 802.11 b/g, Bluetooth
3G	Whether 3G or 4G is supported	3G, 4G
Screen_resolution	Resolution of screen by pixels	$1{,}024 \times 768$, 800×480
Operating_system	Operating system of the tablet computer	Android, Windows XP, iOS
RAM	Size of memory	256 MB, 512 MB
Processor	Type and frequency of processor	1.66 GHz Intel Atom N450
Processor_brand	Brand of processor	Intel, VIA, TI
Storage_size	Size of hard drive or flash drive	16 GB, 32 GB
Average_battery_life	Manufacturer-rated battery life	4.5 hours, 9 hours
Item_weight	Weight of the product	4 pounds
Rear_webcam_resolution	Resolution in megapixels of webcam or camera if available	1.3 MP, 5 MP

current selling price of the focal product provided by Amazon. Compared with other electronic retailers, Amazon is not only the largest but also one of the few that offers an online marketplace as a channel for thousands of merchants to sell their own products. The variables *Lowest_new_price*, *Lowest_used_price*, and *Lowest_refurbished_price* reflect the lowest prices listed across all merchants for new, used, and refurbished products, respectively. *Trade-in_price* is the trade-in price for used products at Amazon.

Promotion refers to the communication method marketers use to provide information to different parties about their products, including advertising, public relations, personal selling, and sales promotion. For each product and across merchants, we categorize the total number of merchants selling new, used, and refurbished versions, as well as the sales rank. For example, each price variable has 24 values (i.e., each value contains the price in one particular week), so the maximal number of longitudinal observations of tablet computers is 571,032 ($11 \times 24 \times 2{,}163$). However, the 11 longitudinal variables may not feature price data for all 24 weeks; data might be missing for the period before the product entered the market or if it was removed from the category before the end of the data collection period. A missing *Amazon_price* value indicates that Amazon does not sell the focal product.

We organize *Item_ID*, *Date_first_available*, and 12 product characteristic variables (i.e., product attributes) provided by manufacturers in panel B of Table 1. In addition to *Screen_size*, *Wireless_type*, *Screen_resolution*, and *Operating_system*, which we gathered according to the buying guide published by *Consumer Reports*, we collected *3G*, *RAM*, *Processor*, *Processor_brand*, *Storage_size*, *Average_battery_life*, *Item_weight*, and *Rear_webcam_resolution*.

These characteristics of 794 products with at least one review submitted were collected manually and compiled in the Product Characteristic Information table. The third column of panel B shows examples of the various attribute levels. Each product characteristic variable features a list of levels (product attribute values). For example, regarding the *Screen_size*, Apple iPad offers a unique size (one attribute level) of 9.7 inches, but the Samsung Galaxy has three different sizes (three attribute levels: 7 inches, 8.9 inches, and 10.1 inches).

Table 2 consists of 12 variables to describe 40,741 consumer-generated reviews collected from Amazon. Each row of the Consumer-Generated Product Reviews table corresponds to a consumer-generated review of a particular tablet computer. Each review has eight basic variables: (1) *Review_ID*; (2) *Item_ID*; (3) *Base_item_ID*; (4) *Reviewer_ID*, which identifies the customer without disclosing personal information; (5) *Review_date*; (6) *Title*; (7) *Review*; and (8) *Rating*, which is out of a total of 5. Two binary variables, *Real_name* and *Verified_purchase*, indicate whether the reviewer used his or her real name and made a purchase on Amazon, respectively. Two additional variables, *Total_votes* and *Helpful_votes*, contain information about the helpfulness and social impact of the reviews. We collected these two variables weekly for 24 weeks; they were provided by reviewers who can vote as to whether each review has been helpful.

Table 2 Consumer-Generated Product Reviews of Tablet Computers

Variable name	Description
Review_ID	Unique identification number for each review
Item_ID	Amazon Standard Identification Number (ASIN) of the product
Base_item_ID	Item ID of the product's base or default model if the product's reviews are consolidated with other models
Reviewer_ID	Unique reviewer identification number assigned by Amazon
Review_date	Date the review was submitted
Title	Title of the review
Review	Textual content of the review
Rating	Numerical rating of the review (1–5 stars)
Real_name	Binary variable that indicates whether the reviewer used his or her real name when submitting the review
Verified_purchase	Binary variable that indicates whether the product purchase was made on Amazon
Total_votes (24 weeks)	Total number of reviewers who voted
Helpful_votes (24 weeks)	Number of reviewers who voted "Yes" on "Was this review helpful to you?"

To conduct research with both product information and consumer-generated reviews, researchers can link the textual reviews in the Consumer-Generated Product Reviews table to the product information in both the Market Dynamics of Products and Product Characteristic Information tables using the *Item_ID*. Furthermore, Amazon clusters certain product models from the same product line as a group and consolidates all reviews for product groups on the same Web page. For example, the reviews for two Acer Iconia A510 models, which differ only in their color (silver and black), are consolidated on the same product page. In Table 1, these two models are listed as different products, but a review written for one model should affect potential buyers for another model as well. Therefore, in Table 2, in addition to *Item_ID*, we provide *Base_item_ID* to indicate the default model for the product line (e.g., for the Acer Iconia A510, the *Base_item_ID* is the *Item_ID* of the silver model).

Finally, characteristic information about 39,278 reviewers who submitted tablet computer reviews is presented in the Reviewer Information table of this data set. Table 3 provides the descriptions of six variables of reviewer information. Variables such as *Reviewer_ranking* and *Total_helpful_votes* reflect

Table 3 Reviewer Characteristics

Variable name	Description
Reviewer_ID	Unique reviewer identification number assigned by Amazon
Total_reviews	Total number of reviews posted by this reviewer
Reviewer_ranking	Reviewer's ranking in "Amazon's Top Customer Reviewers" program;[a] helpfulness, number, and recency of reviews are taken into consideration
Total_helpful_votes	Total number of useful votes of reviews submitted by this reviewer
Location	The geographical location of the reviewer, if provided
Recent_ratings	A sequence of the 10 most recent numerical ratings associated with reviews written by the reviewer, on a scale from 1 to 5

[a]See http://www.amazon.com/gp/customer-reviews/guidelines/top-reviewers.html (accessed August 30, 2013).

the overall quality of the reviews submitted by the particular reviewer. Also, 40% of the reviewers provided their location information. The last variable, *Recent_ratings*, contains reviewers' 10 most recent ratings. To conduct research that combines consumer-generated reviews with the reviewers' own characteristic information, researchers can link the Consumer-Generated Product Reviews table with the Reviewer Information table using *Reviewer_ID*.

2.2. Analyses and Summary Statistics

To demonstrate the significant and unique content of the Tablet Computer data set, we conducted data analyses with descriptive statistics. We chose the top eight tablet computers in terms of the number of reviews, as shown in Table 4. However, we excluded the Kindle Fire, because Amazon is its dominant sales channel, so this product accounts for an overwhelming number of reviews on the site (46% of total reviews), which would distort the scale of the graphs. Two major competitors, the Apple iPad series and Samsung Galaxy, prompt fewer reviews and are not included in the top eight products list, so consumers apparently use other channels to purchase and provide comments (e.g., Apple's stores and website should be the dominant sales channels for the iPad series).

For the product information as described in Table 1, we provide descriptive statistics for the price and promotion variables in Tables 5 and 6, respectively. In Table 5, except for *List_price* (i.e., the retail price suggested by manufacturers), which is almost stable, considerable price variations occur across price variables. *Amazon_price* is usually lower than *List_price*, with the exception of the TouchPad. For this product, the manufacturer put it on "fire sale" and heavily discounted the manufacturer's suggested retail price (MSRP) (Deneckere et al. 1997), to the point that retailers ran out of stock quickly, and many consumers were willing to pay more than MSRP. This example reveals the imbalance between the supply and demand sides.

Prices provided by individual merchants who sell new, used, and refurbished products are usually lower than the *Amazon_price*. Using *Amazon_price*, the current selling price of the product by Amazon, and *Lowest_used_price*, the lowest selling price for used products provided by third-party merchants, we plot the top eight products' prices across 24 weeks in Figure 1. A discontinuity occurs for products that are out of stock for a specific week or that are no longer being manufactured. However, for a few

Table 4 Top Eight Tablet Computers by Number of Reviews

Label name	Product
Transformer	ASUS Eee Pad Transformer TF101-A1 10.1-inch tablet
XOOM	Motorola XOOM Android tablet (10.1-inch, 32 GB, Wi-Fi)
TouchPad	HP TouchPad Wi-Fi 32 GB 9.7-inch tablet computer
gTablet	ViewSonic gTablet with 10" multi-touch LCD screen, Android OS 2.2
IconiaTab	Acer Iconia Tab A500-10S16u 10.1-inch tablet computer
Thrive	Toshiba Thrive 10.1-inch 16 GB Android tablet AT105-T1016
PlayBook	Blackberry PlayBook 7-inch tablet (16 GB)
TC 970	Le Pan TC 970 9.7-inch multi-touch LCD Google Android tablet PC

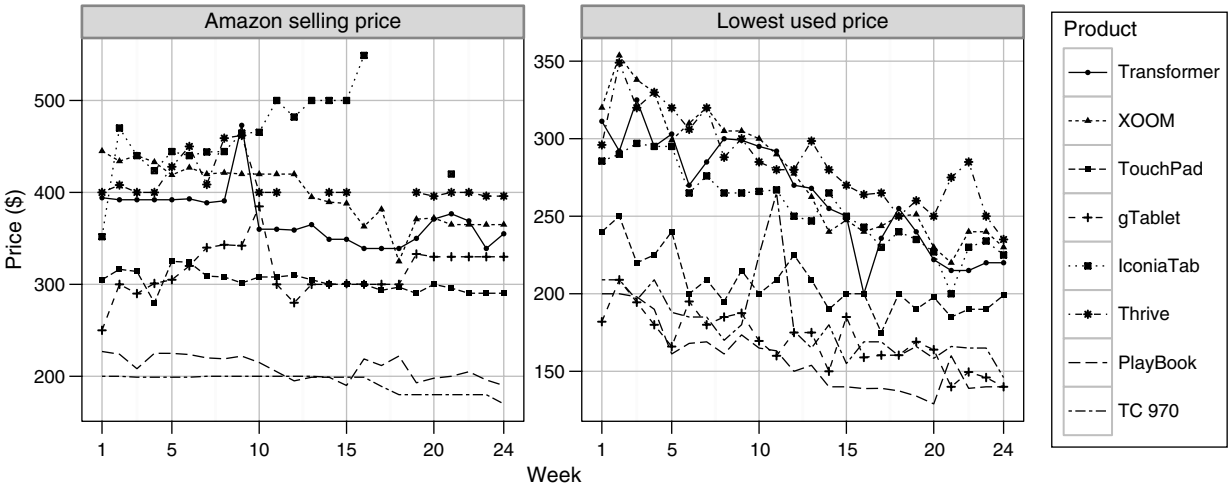
Table 5 Descriptive Statistics of Price Variables

	List_price				*Amazon_price*			
Product	Mean	SD	Min	Max	Mean	SD	Min	Max
Transformer	399.00	0.00	399.00	399.00	371.97	29.55	338.99	472.93
XOOM	499.00	0.00	499.00	499.00	398.50	31.96	325.00	444.99
TouchPad	149.99	0.00	149.99	149.99	302.55	10.99	279.97	324.99
gTablet	499.99	0.00	499.99	499.99	314.13	26.89	249.99	384.79
IconiaTab	468.81	18.96	400.00	478.14	461.12	44.08	351.86	549.00
Thrive	399.99	0.00	399.99	399.99	410.69	21.88	395.85	461.95
PlayBook	499.00	0.00	499.00	499.00	209.62	12.77	189.99	226.98
TC 970	249.99	0.00	249.99	249.99	192.99	9.89	169.99	199.99
	Lowest_new_price				*Lowest_used_price*			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Transformer	360.34	22.75	310.00	394.19	263.87	36.28	199.99	325.00
XOOM	372.40	39.94	320.00	415.00	276.79	39.93	220.00	353.71
TouchPad	280.76	16.51	230.00	299.00	206.40	18.78	175.00	249.99
gTablet	261.48	14.19	248.93	292.97	170.07	18.12	140.00	209.00
IconiaTab	403.54	36.81	325.00	469.99	255.93	25.96	200.00	296.97
Thrive	393.00	15.24	349.99	408.00	285.67	28.57	235.00	349.00
PlayBook	198.63	11.47	175.00	215.69	157.92	21.89	128.99	199.99
TC 970	189.30	11.96	169.28	199.99	181.72	26.50	145.95	265.00
	Lowest_refurbished_price				*Trade-in_price*			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Transformer	287.82	29.15	237.99	334.39	212.83	9.37	193.25	223.00
XOOM	311.71	36.34	235.00	348.99	204.11	54.00	104.50	278.25
TouchPad	228.30	22.19	198.00	270.00	98.50	15.36	84.75	125.00
gTablet	209.57	32.01	151.97	256.99	85.07	10.90	66.50	110.00
IconiaTab	265.08	22.86	230.04	310.00	123.08	51.76	73.00	253.00
Thrive	293.94	26.14	249.99	359.90	174.06	23.91	133.25	190.25
PlayBook	163.63	22.06	129.00	200.48	67.14	7.85	61.50	87.25
TC 970	165.00	7.07	160.00	170.00	71.02	4.74	60.25	75.50

Table 6 Descriptive Statistics of Promotion Variables

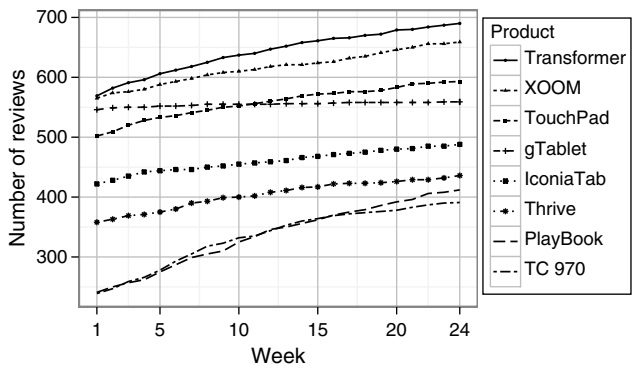
	Total_new				*Total_used*				*Total_refurbished*			
Product	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Transformer	43.13	13.58	25	68	27.29	13.22	7	51	7.33	3.69	0	15
XOOM	44.33	10.18	29	63	49.25	8.99	26	65	6.17	1.83	3	11
TouchPad	62.25	6.80	51	73	76.96	15.51	40	92	12.63	2.98	7	17
gTablet	18.00	2.70	14	23	26.71	4.69	15	33	2.54	1.32	0	5
IconiaTab	6.17	2.85	2	13	30.79	7.34	19	43	10.63	2.53	6	16
Thrive	8.58	3.67	4	18	20.13	7.01	7	31	14.50	5.19	4	23
PlayBook	142.58	26.00	102	180	43.58	9.60	28	61	6.42	2.59	2	11
TC 970	6.08	1.82	3	9	3.50	1.41	1	6	0.08	0.28	0	1

Figure 1 Weekly Trends of Selling Prices and Lowest Used Prices on Amazon



price variables such as *List_price*, no such discontinuity arises, because all prices are available throughout the data collection period. Table 6 shows the descriptive statistics for the total number of merchants selling new, used, and refurbished products. Across these three categories, used products are sold by the most merchants, whereas refurbished products have the fewest merchants. In Figure 2, we plot the total accumulated reviews for these eight products throughout the 24-week observation period. The first week captures the number of all prior reviews.

Figure 2 Total Accumulated Reviews



For the structured review information as described in Table 2, we first provide the descriptive statistics, such as *Rating*, *Real_name*, and *Verified_purchase*. In Figure 3 we present the distribution of review ratings, based on their numerical star ratings. A five-star rating describes a high proportion of all eight products. Table 7 uses two binary variables, *Real_name* and *Verified_purchase*, to indicate the percentages with which reviewers used their real names when submitting reviews and whether a purchase was made on Amazon. Most reviewers do not use their real names, but a high proportion of them make their purchases through Amazon.

To demonstrate the rich content of the qualitative product reviews, we provide the descriptive statistics for 40,741 consumer-generated reviews in Table 8, which offers researchers a unique opportunity to investigate the use of online product reviews. These reviews refer to 794 tablet computers. Most reviewers contribute only a single review (i.e., 40,741 reviews submitted by 39,278 reviewers). Each of the 794 tablet computers attracted an average of 51.3 reviews, each

Figure 3 Distribution of Review Ratings

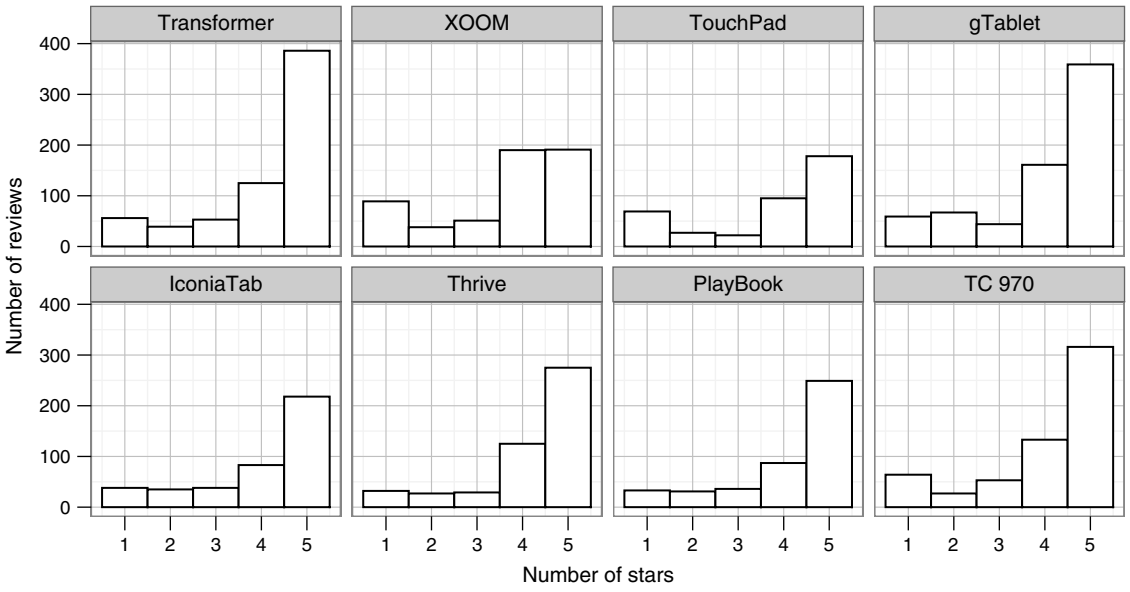


Table 7 Anonymity and Purchase Place

Product	*Real_name* (%)	*Verified_purchase* (%)
Transformer	31.30	65.22
XOOM	33.08	59.64
TouchPad	29.68	55.31
gTablet	36.67	62.08
IconiaTab	30.94	51.43
Thrive	28.90	62.84
PlayBook	27.67	45.39
TC 970	20.46	75.19

of which contained 11.5 sentences with 207.8 words and punctuation marks, on average. About two-thirds of the rating votes were positive (214,653 of 326,507). Because tablets are a prominent electronics product category, these statistics exemplify the rich content of online consumer-generated reviews, which can be used to answer interesting marketing questions. We conducted a simple text analysis of the 40,741 reviews, for which we extracted noun phases that contain up to two words, then ranked them according to frequencies. Table 9 lists the top 60 noun phrases. Figure 4 depicts these phrases as a tag cloud, highlighting those with higher frequencies by using larger fonts. For example, "tablet," "ipad," "kindle fire," and "device" emerged as the top four noun phrases in tablet reviews. Tablet computers' product features, such as "apps," "battery," "screen," "book," "video," "movie," "wifi," and "game," were also frequently mentioned. These noun phrases reflect consumers' opinions and preference measures in relation to tablet computers.

Table 8 Descriptive Statistics of Product Reviews

Attribute	Value
Number of products with reviews	794
Number of reviews	40,741
Number of distinct reviewers	39,278
Average number of sentences per review	11.5
Average number of words and punctuation marks per review	207.8
Average number of reviews for products with reviews	51.3
Total number of votes	326,507
Total number of helpful votes	214,653

Table 9 Top 60 Noun Phrases and Frequencies

Noun phrase	Frequency	Noun phrase	Frequency	Noun phrase	Frequency
Tablet	27,570	Game	5,208	Battery	3,221
iPad	18,687	Review	5,132	Day	3,203
Kindle fire	17,538	Something	5,030	Anything	3,005
Device	16,827	Issue	5,029	Work	2,923
One	16,030	Wifi	4,502	Apple	2,844
Fire	13,693	Way	4,364	Phone	2,830
Amazon	12,105	People	4,355	App	2,793
Apps	11,901	Laptop	4,019	Christmas	2,726
Screen	11,132	Music	3,964	Box	2,672
Book	9,709	Money	3,825	Wife	2,608
Thing	9,169	Little	3,680	Case	2,562
Kindle	8,998	Android	3,518	Keyboard	2,560
Problem	8,793	Web	3,513	Use	2,557
Time	8,054	Email	3,502	Feature	2,520
Lot	7,569	Computer	3,436	Touch screen	2,518
Product	7,079	Unit	3,433	Fact	2,499
Price	6,499	Internet	3,366	Hand	2,486
Everything	5,426	Battery life	3,353	Anyone	2,484
Video	5,325	Camera	3,248	Item	2,481
Movie	5,229	Bit	3,237	Purchase	2,405

Figure 4 **Tag Cloud Visualization of Top 60 Noun Phrases in Tablet Reviews**

3. Research Opportunities

3.1. Price Dispersion in Electronic Markets

The Tablet Computer data set contains six price variables, which provide a unique opportunity to address a stream of research questions regarding electronic price dispersion. Bakos (1997) first predicted that electronic markets would be more efficient and friction-free than traditional markets; since then, because of the reduced search costs associated with matching buyers and sellers, a growing body of research has examined the factors that cause price dispersions in electronic markets. Some notable factors include product and service bundling (Baye et al. 2004); differences in brand, reputation, and trust across merchants (Baye et al. 2006, Brynjolfsson and Smith 2000); product heterogeneity (Baye et al. 2006); price discrimination (Clemons et al. 2002); and randomized pricing strategies by firms (Ghose et al. 2007). Ghose and Yao (2011) also note that most existing work uses posted prices to estimate the price dispersion, which may lead to an overestimation of price dispersion. Instead, they suggest using actual transaction prices (i.e., market clearing prices) to evaluate the extent of price dispersion.

We believe prices may interrelate and influence price dispersions in a single electronic market. In addition, prices may interact with the preceding factors, which may serve as covariates that further affect the cross effects. For example, the interaction of the number of merchants and *Amazon_price* might cause a variation in other prices, such as the *Lowest_used_price*. Amazon, as a retailer and online marketplace, sets the price strategy, which implies a broad stream of interesting research questions. This data set also reveals the number of merchants that sell new, used, and refurbished products, each of which is a possible covariate that could drive the effects of price dispersion. Therefore, all six price variables could be dependent variables for marketing research.

3.2. Translating Sales Ranks to Sales Quantities

The integration of structured product information and unstructured consumer-generated product reviews is another unique feature of our Tablet Computer data set that can support a research trend in textual analysis (e.g., Onishi and Manchanda 2012, Tirunillai and Tellis 2012). That is, research has shown that consumer reviews relate closely to sales (Chevalier and Mayzlin 2006, Liu 2006). However, Amazon does not provide product-level sales information, and data regarding sales volumes of tablet computers are not available in public sources.

Amazon provides sales ranks of products within a particular product category. We could use the log of sales rank as a dependent variable (Archak et al. 2011), but growing empirical literature instead offers sophisticated methods to map Amazon's sales ranking data onto product-level sales quantities. For example, Chevalier and Goolsbee (2003) translate Amazon sales ranking data into Amazon sales using an experimental calibration. Continued research uses this technique in various contexts (e.g., Ghose et al. 2006, Smith and Telang 2009). By applying this translation, "sales" becomes available as a dependent variable. Existing literature also indicates that several factors may or may not matter for sales, such as valence (Chevalier and Mayzlin 2006), volume (Liu 2006), and the reviewer's decision to reveal his or her identity (Forman et al. 2008). Moe and Trusov (2011) also show that ratings have not only a direct effect on product sales but also a significant indirect effect through future ratings and through the rating-generation process.

In this sense, this data set provides sales research opportunities in at least two realms. First, positive and negative valences might be judged not solely by ratings but could also be obtained by mining the textual reviews. Second, other than dynamic rating data, variables such as valence and volume, text reviews, and identity information can be collected in a longitudinal setting.

3.3. Text Analytics and Market Structure Analysis

Unlike traditional surveys or transaction records collected from legacy systems, consumer-generated product reviews contain rich consumer insights and behavioral information. Researchers can examine qualitative consumer-generated reviews and automate market structure analyses by applying text mining and sentiment analysis techniques. Text mining the product reviews contributed freely by consumers can lead to the discovery of consumer usage situations, preferences, and sentiments toward product

attributes. Advanced text mining techniques support analyses of linguistic and semantic structures in the review sentiments (Turney and Pantel 2010). Thus, they uncover embedded consumer opinions (i.e., voices of consumers) that in turn reveal more sophisticated consumer preferences and insights to advance market structure analyses (e.g., Netzer et al. 2012).

4. Obtaining and Using the Data Set

The *Marketing Science* website (http://pubsonline.informs.org/page/mksc/online-databases) provides the link to the website maintained by the authors with additional instructions about accessing and using the Tablet Computer data set. To explore the textual content of the reviews, researchers can use natural language processing tools such as Stanford CoreNLP (Socher et al. 2013, Toutanova et al. 2003), Apache OpenNLP,[4] or the Natural Language Toolkit (Bird et al. 2009) to extract information. These free text analytic tools support common linguistic processing methods, including sentence boundary detection, tokenization, part-of-speech tagging, and named entity recognition. Noting the effort required to set up these tools, we provide the parsed content of 40,741 tablet computer reviews, obtained using Stanford CoreNLP (version 3.2.0). The parsed reviews are available in XML format, which most statistical packages can read and import.

Although this data set pertains to tablet computers, researchers can use the proposed data collection method to obtain even more recent reviews or data about different product categories. We also share our crawling program to facilitate research that seeks to collect data available on the Internet. This will benefit researchers who wish to obtain consumer-generated product reviews of a different product category.

Two research issues should be considered when using the consumer-generated product reviews. First, bogus (fake) reviews may have been submitted to Amazon. Bogus reviews seek to mislead both consumers and sentiment analysis systems, either to promote or to damage a particular product's reputation. Users should refer to research related to detecting and filtering out fake reviews (e.g., Lim et al. 2010, Mukherjee et al. 2012). Second, the nouns and noun phrases derived from the product reviews may contain synonyms or homonyms, because reviewers use various noun phrases to describe the same product attributes, and the same noun phrases extracted might refer to different product attributes. However, we expect this concern to be minimal, because this data set was collected for a particular product category.

[4] Available at http://opennlp.apache.org/ (accessed August 30, 2013).

5. Summary

The first decade of 21st century has experienced phenomenal, exponential growth in digital data. In the big data era, both UGC and manufacturer-provided content are readily available and can expand the horizons of marketing research. We offer a tablet computer data set for the marketing science community, which contains collected, filtered, and integrated product characteristic information, market dynamic information, consumer-generated product reviews, and reviewer information spanning a period of 24 weeks. This data set should help accelerate the drive toward big data analytics, as well as benefit the practice and study of marketing. Our approach for generating this data set provides academic researchers and practitioners with insights into an innovative data collection method (Mela 2011), supporting relevant research even when reliable and representative data are missing or difficult to obtain. In particular, our Tablet Computer data set can be used to answer many interesting research questions, even beyond those we proposed in §3.

Acknowledgments

All three authors contributed equally and are listed in reverse alphabetical order. The authors thank the editor-in-chief, Preyas Desai, the associate editor, and two anonymous reviewers with their constructive and insightful comments and suggestions in preparing the Tablet Computer data set and their article.

References

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Bakos JY (1997) Reducing buyer search costs: Implications for electronic marketplaces. *Management Sci.* 43(12):1676–1692.

Baye MR, Morgan J, Scholten P (2004) Price dispersion in the small and in the large: Evidence from an Internet price comparison site. *J. Indust. Econom.* 52(4):463–496.

Baye MR, Morgan J, Scholten P (2006) Persistent price dispersion in online markets. Jansen D, Elgar E, eds. *The New Economy* (University of Chicago Press, Chicago), 122–143.

Bird S, Klein E, Loper E (2009) *Natural Language Processing with Python* (O'Reilly, Sebastopol, CA).

Brynjolfsson E, Smith MD (2000) Frictionless commerce? A comparison of Internet and conventional retailers. *Management Sci.* 46(4):563–585.

Chen BX (2012) How are 7-inch tablets doing? *Bits* (blog), October 19, http://bits.blogs.nytimes.com/2012/10/19/7-inch-tablets?_r=0.

Chevalier JA, Goolsbee A (2003) Measuring prices and price competition online: Amazon and Barnes and Noble. *Quant. Marketing Econom.* 1(2):203–222.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(4):345–354.

Clemons EK, Hann I-H, Hitt LM (2002) Price dispersion and differentiation in online travel: An empirical investigation. *Management Sci.* 48(4):534–549.

Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the Web. *Management Sci.* 53(9):1375–1388.

Deneckere R, Marvel HP, Peck J (1997) Demand uncertainty and price maintenance: Markdowns as destructive competition. *Amer. Econom. Rev.* 87(4):619–641.

Developer Economics (2013) Developer attention in tablets catching up with smartphones. Report, VisionMobile, London. http://www.developereconomics.com/report/q3-2013-developer-attention-in-tablets-catching-up-with-smartphones.

Fader PS, Winer RS (2012) Introduction to the special issue on the emergence and impact of user-generated content. *Marketing Sci.* 31(3):369–371.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3):291–313.

Ghose A, Yao Y (2011) Using transaction prices to re-examine price dispersion in electronic markets. *Inform. Systems Res.* 22(2):269–288.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Ghose A, Mukhopadhyay T, Rajan U (2007) The impact of Internet referral services on a supply chain. *Inform. Systems Res.* 18(3):300–319.

Ghose A, Smith MD, Telang R (2006) Internet exchanges for used books: An empirical analysis of product cannibalization and welfare impact. *Inform. Systems Res.* 17(1):3–19.

IDC (2013) Tablet shipments forecast to top total PC shipments in the fourth quarter of 2013 and annually by 2015, according to IDC. Press release (September 11), IDC, Framingham, MA. http://www.idc.com/getdoc.jsp?containerId=prUS24314413.

Kharif O (2011) Fake iPads flood market as scammers target top technology gift. Bloomberg News (November 1) http://www.bloomberg.com/news/2011-11-01/fake-ipads-flood-market-as-scammers-target-top-technology-gift.html.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Lim E, Nguyen V, Jindal N, Liu B, Lauw H (2010) Detecting product review spammers using rating behaviors. *Proc. 19th ACM Internat. Conf. Inform. Knowledge Management* (ACM, New York), 939–948.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(4):74–89.

McCarthy EJ (1960) *Basic Marketing: A Managerial Approach* (Richard D. Irwin, Homewood, IL).

Mela CF (2011) Data selection and procurement. *Marketing Sci.* 30(6):965–976.

Moe W, Trusov M (2011) Measuring the value of social dynamics in online product ratings forums. *J. Marketing Res.* 49(3):444–456.

Moscaritolo A (2012) Survey: 31 percent of U.S. Internet users own tablets. *PC Magazine* (June 18) http://www.pcmag.com/article2/0,2817,2405972,00.asp.

Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. *Proc. 21st Internat. Conf. World Wide Web (WWW2012)* (ACM, New York), 191–200.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

Onishi H, Manchanda P (2012) Marketing activity, blogging and sales. *Internat. J. Res. Marketing* 29(3):221–234.

Smith M, Telang R (2009) Competing with free: The impact of movie broadcasts on DVD sales and Internet piracy. *Management Inform. Systems Quart.* 32(2):312–338.

Socher R, Manning CD, Ng A (2013) Parsing with compositional vector grammars. *Proc. 51st Assoc. Comput. Linguistics Conf., Sofia, Bulgaria*.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Toutanova K, Klein D, Manning C, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. *Proc. HLT-NAACL 2003* (ACL, Stroudsburg, PA), 173–180.

Turney PD, Pantel P (2010) From frequency to meaning: Vector space models of semantics. *J. Artificial Intelligence Res.* 37(1):141–88.

The data set described in this paper is maintained by the authors and available through http://pubsonline.informs.org/page/mksc/online-databases.