Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 10 / ISSUE 2 / MAY 2015 / 1-18

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

Busting Myths of Electronic Word of Mouth: The Relationship between Customer Ratings and the Sales of Mobile Applications

Sami Hyrynsalmi[1], Marko Seppänen[2], Leena Aarikka-Stenroos[1], Arho Suominen[3], Jonna Järveläinen[1] and Ville Harkke[1]

[1] University of Turku, Turku School of Economics, Turku, Finland,
{sthyry, leaari, jonna.jarvelainen, ville.harkke}@utu.fi
[2] Tampere University of Technology, Center for Innovation and Technology Research, Tampere, Finland,
marko.seppanen@tut.fi
[3]VTT Technical Research Centre of Finland, Innovation and Knowledge Economy, Turku, Finland,
arho.suominen@vtt.fi

Received 17 September 2013; received in revised form 13 August 2014; accepted 6 October 2014

Abstract

Business and academic research frequently highlights the power of electronic word of mouth, relying on the knowledge that online customer ratings and reviews influence consumer decision making. Numerous studies in different disciplines have been conducted to examine the effectiveness of electronic word of mouth communication. Previously, typically small sample studies suggest that positive electronic word of mouth increases sales and that the effects depend on the volume and valence of reviews and ratings. This study's contribution lies in testing the relationship between electronic word of mouth and the sales of applications in a mobile application ecosystem (Google Play) with an extensive dataset (over 260 million customer ratings; 18 months). The results show that higher values of valence of customer ratings correlate statistically significantly with higher sales. The volume of ratings correlates positively with sales in the long term but negatively in the short term. Furthermore, the relationship between electronic word of mouth and sales seems to be more important when the price of the application increases. The findings also underline the importance of the choice of a measurement period in studies.

Keywords: Products ratings, Consumer reviews, Consumer ratings, Sales, Electronic word of mouth, App stores, Micro-pricing, Big data

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 10 / ISSUE 2 / MAY 2015 / 1-18

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

1 Introduction

Hundreds, if not thousands, of new mobile applications are released every day in mobile application marketplaces, such as the Apple App Store, Google Play, and Windows Phone Marketplace. Together, these three dominant mobile application ecosystems provide over a million applications for customers. For a single developer of an application, differentiating one's offerings from the masses and increasing sales is an arduous task. However, as current research has argued [18], [26]-[28], earlier customers' comments delivered through electronic word-of-mouth, i.e., customer ratings and reviews, can improve sales. Hence, ratings might have a significant effect on the success of a product (see e.g. Site 1). Based on these earlier notions and the increasing relevance of the mobile app business, this paper addresses the relationship between customer ratings and sales in a mobile application marketplace. We use three datasets gathered from Google Play during a time span of 18 months. Together, these sets contain the information of over 800,000 unique applications and over 260 million ratings left by users.

Online recommendations and user reviews have become one of the most important sources of information for modern consumers [42]. Particularly in e-commerce, it is difficult for the customer to evaluate the product or the service and the benefits and value that it generates. Thus, new customers tend to rely on trustworthy independent information sources, such as customers who already have experienced the product. In other words, the customer relies on the *judgment* of other clients, experts, or actors in the field who share valuable information about the product through divergent ratings and review systems. Their experiences deliver information related to the user or customer perspective and thus reduce the risk and uncertainty perceived by consumers. There are divergent means to deliver consumer reviews and ratings, which are often termed as *electronic word-of-mouth* (E-WOM). E-WOM is communicated through discussion forums, blogs, online opinion sites, online communities, online product reviews, and comments written by consumers on web pages [11], [15]-[16] and includes divergent verbal and numeric practices of sharing customer experiences and judgments [60]. In this paper, we focus particularly on ratings, i.e., the numerical or star evaluations given by customers.

The rapidly increasing studies on customer reviews and ratings clearly show that the perception of the trustworthiness of a source can lead to the increased persuasiveness of the information, and that user-generated content, such as consumer reviews, are more influential than marketer-generated information on corporate websites (see, e.g., [7], [14], [16], [26], [48]).The extant studies have shown that electronic WOM has an effect on sales [26]-[28], customer value and loyalty [31], and online brand [2], as well as on the success of new product introductions [20]. Due to its rapidly expanding relevance and multiple effects in the context of contemporary business, e-WOM has given rise to much research in multiple disciplines. De Maeyer [26], for example, presents publications from marketing and management to psychology, information system sciences, and computer science. Thus, it can be argued that e-WOM is one of the cornerstones of e-commerce.

It is noteworthy that even though the previous studies have suggested that customer ratings improve sales [13], [18], [28], [60], there are also opposite results indicating that they do not have an impact on sales [38], [41], [44]. The key conceptualizations to measure the ratings' impact on sales have been *volume* (i.e., the number of ratings), *valence* (e.g., the average of the ratings), and *variance* (i.e., the dispersion of the ratings). There are, however, conflicting findings on how these measures indicate further sales: some studies have found a correlation between volume and sales but not between valence and sales [3], [13], whereas other studies have found support for the opposite [19], [29]. These incoherent results thereby strongly indicate that the measures estimating the ratings' impact on sales necessitate deeper investigations.

Most of the extant studies on this subject have been conducted with relatively small sample sizes (a few dozen items), either with books, DVDs, or movies; replication studies with different product categories have been requested by, e.g., De Maeyer [26]. The existing studies have typically focused on tangible products with considerable prices. Sometimes, all of the relevant qualities of the sample that can affect the result of the analysis might not have been described, such as the popularity of the product. Therefore, the purpose of this study is to examine the relationship between e-WOM-particularly the volume, valence, and variance of customer ratings- and sales of mobile applications. To achieve this purpose, we empirically study different dimensions of e-WOM with a large set of data in an established e-commerce domain. The extensive data of this study includes most of Google Play applications, from top-selling superstars to the niche products that have been downloaded only a few times. Furthermore, we study the moderating role of price and time in the mobile application ecosystems. Therefore, in this paper we set the following three research questions, the first of which focuses on the e-WOM's dimension and the applications' sales:

1. Do customer ratings correlate with the sales improvement in the mobile application marketplace?

The effect of time has been infrequently discussed in the extant e-WOM literature, and to the best of our knowledge, there are no previous studies addressing how e-WOM's relationship with sales changes over time. Furthermore, in e-WOM studies, the choice of a timeframe is an important decision; we are unaware of how long takes for the reviews to have an impact on the application sales. It is clear that there are a number of variables that explain changes in product sales, such as marketing efforts, visibility in social media or upgrades to the product. With selecting two time

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

frames, we look if the relationship between e-WOM and sales remains stable. Therefore, the second research question of this paper is as follows:

2. Do the different time period lengths affect the relationship between e-WOM and sales in the mobile application marketplace?

Finally, we assume that the price of the rated product may be relevant when examining the relationship between customer ratings and sales. The mobile application marketplaces offer products with a wide range of prices. While most of the products are cheap, costing only a few eurocents, there are also expensive applications, such as road maps, offered to consumers. The classic Prospect theory shows that the consumer makes a financial decision based on evaluated losses and gains [39]. That is, the consumers might be willing to just buy an application instead of exploring the reviews and ratings if the price is low enough. On the other hand, it is well known that consumers' economic behavior is emotional, which may lead to irrational decision-making [55]. Thus, we set the third research question as follows:

3. Does the price impact the relationship between e-WOM and sales in the mobile application marketplace?

This paper aims to contribute to e-WOM and e-commerce research by investigating in detail the relationship between e-WOM, i.e., consumer ratings, and sales improvement with an extensive dataset. The results show that the valence of ratings correlates positively with the sales in both the long and short time periods. The effect of e-WOM seems to increase when the price of an application grows, thus suggesting that it is easier for a user to take a risk and test a cheap application than to read through reviews. This paper is among the first, to the authors' knowledge, that has investigated the moderating role of price. The variance of ratings was found to correlate with sales in the short time period, although the results suggest that more work is needed in the operationalization of the theory. The volume of the ratings was found to correlate positively with sales in the long time period, while the correlation coefficient was negative in the short time period. Our results clearly indicate that e-WOM and its effect on sales is a more complex issue than previously presented. Further work is needed, particularly to understand the effect time has on e-WOM studies.

The rest of the paper is structured as follows. Section 2 will review the theoretical background of the study and, in order to address the presented research questions, put forth a set of hypotheses. Section 3 presents the research methodology and data-gathering process, and the fourth section focuses on the results. Section 5 will discuss the results, while section 6 summarizes the study's major contributions and presents the limitations and implications of the results with ideas for further research.

2 Theoretical Background and Hypotheses

In order to explicate and analyze the effects of e-WOM on sales, we will first discuss the current research knowledge of e-WOM and its effects. Based on this discussion, we derive hypotheses studied in this paper.

2.1 Electronic Word-of-Mouth and its Effects on Sales

Since the research tradition on online reviews, ratings, and e-WOM is growing quickly, and there are diverse means to deliver such information, overlapping and vague definitions and terms have been presented to describe the phenomenon. In this paper, we rely on the most commonly used definition of e-WOM, which was provided by Henning-Thurau et al. [35], according to which e-WOM refers to "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet." E-WOM occurs through customer ratings and reviews: a *customer rating* refers to the numerical or star value given by a consumer to express her/his satisfaction or dissatisfaction with the product, whereas the term *customer review* refers to a verbal message written by a consumer. In this paper, our focus is on customer ratings due to the utilized research method. Although, for example, Google Play allows a customer to leave both a verbal comment and a star rating at the same time, a customer can review an application without leaving text but not vice-versa.

Previous studies have found that the less popular the product, the more it will gain or lose through online reviews [53], [60]. Some of the first works that examined e-WOM [5], [35] have focused on e-WOM designated consumer-opinion platforms. As e-WOM has recently become an axiomatic source of information for consumers, research is needed to validate empirical results and create deeper understanding of the relevant concepts and phenomena [26], [35].

The extant research has substantially studied the effects of e-WOM on sales. The most often-mentioned dimensions of e-WOM in the literature are "verbal, valence, variance, volume and helpfulness of reviews" [26]. We focus on the valence, volume, and variance of ratings. *Valence* and *volume* are *classical dimensions* that have been the most frequently studied [26]. The *variance* dimension is less frequently addressed [26]; however, it might be able to explain consumers' perception of contradicting ratings. Therefore, it is included in our study. We exclude the *verbal*

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

dimension, since it would require content analysis that is not within the scope of our study, and the *helpfulness of a review* dimension, since it was not available in the studied marketplace at the time of the study.

The discrepancy between extant findings on how volume, valence, and variance of ratings indicate e-WOM's effect on sales within different product categories is displayed in Table 1. As shown in the table, previous studies have frequently found that valence is more important than volume and vice versa. There are studies that found a correlation between volume and sales, but not between valence and sales [3], [13], whilst there are also studies that have found support for the opposite [19], [29]. Therefore, which of these two (valence or volume) has a greater impact in on sales has been raised as a key question in e-WOM studies [40].

Table 1: A selection of studies and summaries of their results on e-WOM's effect on the sales of goods

Product Categ.	Results	Method and e-WOM data	Time period(s)	Study
Books	Volume has a positive impact while valence is not related.	Quantitative analysis of cross-sectional web crawled data from Amazon (N = 610)	Not studied	[13]
TV shows	Dispersion of e-WOM is more important than valence or volume.	Quantitative analysis of data collected from Usenet newsgroups between two episodes (usually a week) of a TV show (N = 44)	Short term	[30]
Books	A higher valence leads to higher sales, although negative reviews have a greater impact than positive reviews. Verbal dimension has an impact beyond numbers.	Quantitative analysis of data collected from Amazon and Barnes & Noble (N = 2,387) during two-day periods in May 2003, August 2003, and May 2004	Short (three months) and long term (one year)	[18]
Beers	Variance and valence are positively associated with sales growth; volume is not.	Quantitative analysis of review data between April 2000 and July 2004 from Ratebeer (N = 1,159) and sales data 2001–2003 from Association of Breweries	Not studied	[20]
Films	Volume over valence in predicting future box office revenues.	Quantitative analysis of review data from Yahoo Movies (N = 40) May–September 2002	From one to eight weeks	[44]
Films	Early volume can be used as a proxy of early sales; valence and dispersion have positive relationships with future sales.	Develop a sales forecasting model with data collected from Yahoo Movies in 2002, including weekly box office sales (N = 80)	From three to 51 weeks (mean = 14)	[28]
Films	Valence has a significant impact on a film's box office revenues but volume and variance does not.	Quantitative analysis of review data from Yahoo Movies, Variety and BoxOfficeMojo (N = 71) from July 2003 to May 2004	Two weeks	[29]
Films	Valence seems to matter, while volume and variance does not.	Quantitative analysis of review data from Yahoo Movies (N = 148) from November 2003 to February 2005	A 16-month period	[19]
Games	Volume seems to correlate better than valence.	Quantitative analysis of review data from GameSpot (N = 220) from March 2003 to October 2005	Monthly and over game lifecycle	[60]
eBooks	Valence does not correlate while volume correlates well with the sales.	Quantitative analysis of reviews from Amazon (N = 133) from February 2007	56 weeks	[3]
Various	Valence has an impact on search goods (consumer electronics) while volume affects experience products (games). Negative reviews have greater impact than positive reviews.	Analysis of panel data gathered from Amazon (N = 332) from August 2007 to April 2008	Max. 36 weeks (unbalanced set)	[24]
eBooks	Volume and valence have a positive effect on the sales; however, the effect of volume was consistent while valence's effect was not.	Quantitative analysis of web crawled data from Amazon (N = 851) from November 2012 to February 2013	12 weeks; test once a week	[22]
Cameras	Both valence and volume affect the sales; a negative review has higher impact than a positive review.	Quantitative analysis of web crawled data from Amazon (N_1 = 1,292 and N_2 = 428) from three one-day cross-sectional sets in 2009–2010	Static model (N_1) and 28 days (N_2)	[58]

N indicates the number of goods and services addressed in a study; the number of ratings and review comments is often considerably higher
N/A = Not available

Some of the contradicting results might be explained by the domain of the studies (e.g., expert opinion might weigh more in certain domains) or the reliability of the e-WOM platform under study (c.f., Amazon's feedback collection system that highlights good comments and systems that only show an average of ratings). In the following sections, we will discuss in detail each of the studied dimensions, i.e., valence, volume, and variance, as well as the moderating role of the price, and derive hypotheses regarding these dimensions.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

2.2 Hypotheses

In the following, we will present our four hypotheses and motivation for them. First, we will discuss on the valence, volume and variance of ratings. Finally, the fourth hypothesis assesses the moderating role of price.

2.2.1 Valence

Valence refers to the numerical value of a customer rating [26]. It can be, e.g., the average of numerical ratings given by customers. Valence is a rather classical dimension that is straightforwardly linked to the future sales of a product or a service; therefore, it is not surprising that it has been studied exhaustively. For example, Cui et al. [24] found by analyzing panel data of 332 new products from Amazon that the valence of reviews has a stronger effect on search products, whereas the volume of reviews is more important for experience products. Valence has been found to be positively correlated with sales in different product type domains (see, e.g., [18], [20], [28]-[29], [60]).

In addition to simple average, the positiveness or negativeness of ratings may have a different effect on sales. For example, Chen and Liu [12] found that negative e-WOM is more influential than positive e-WOM for mobile application sales in Apple's App Store. Although the conventional logic would argue that positive e-WOM or positive feedback for a product would result in higher product sales, this might not be the case. On the other hand [6], negative word-of-mouth can increase product sales through increasing awareness. The mobile application market is highly dynamic [38], and this might suggest that overall product awareness is even more important than positive reviews [60].

As discussed above, many previous studies support the idea that valence, i.e., the high average ratings, correlate with higher sales even though the opposite has also been concluded (e.g., [38], [44]). Nevertheless, valence is a classical dimension that should be studied when addressing e-WOM's relationship with the sales. Therefore, we formulate our first hypothesis by following the classic view that a high value of valence is positively correlated with sales:

H1: A high average value of ratings correlates positively with sales.

2.2.2 Volume

The *volume* of the ratings is based on the argument that the amount of feedback given is a signal of product popularity [26]. The results from the literature also show that the volume of reviews has a significant effect on sales (Table 1). Several researchers have argued that the volume of reviews matters and is more important than the ratings in predicting sales [3], [29], [44]. However, De Maeyer [26] notes that there is a lot of support for the notion that the volume correlates positively with sales, though there are conflicting results (see Table 1) about whether valence or volume is more important. Thus, we formulate our second hypothesis:

H2a: A high volume of ratings correlates positively with sales.

However, the volume of ratings clearly depends on the number of installations because the marketplace verifies that a reviewing customer has installed the application before a review or rating can be made. Therefore, it is likely that the volume of ratings only reflects the previous popularity (i.e., sales) of an application. Thus, we decided to study if the previous popularity, i.e., the overall number of installations, correlates with future sales. We formulate our third hypothesis as follows:

H2b: A high number of installations correlates positively with sales.

2.2.3 Variance

Variance refers to the dispersion of the customer reviews and ratings. For example, star-based customer ratings tend to follow a "J-shape" distribution [36]. That is, most of the ratings are highly positive; there are only a few in the middle and there are also some highly negative reviews. Variance of e-WOM ratings and reviews has been studied sparsely in the past and the results of such studies are inconclusive [26]. For example, variance is found to be a good predictor for the products of microbreweries [20]; variance of film critics' reviews has no impact on box office sales in the early weeks [59]; and variance has been found to have a negative impact on hotel reservations [57].

Recently, Sun [52] argued that *niche products* can be identified by the variance of user reviews. A *niche market* is defined as "consisting of an individual customer or a small group of customers with similar characteristics or needs" [25]; consequently, a *niche product* is seen as a product for this kind of a market. As those kinds of products provoke ratings from both those who love (i.e., those belonging to the niche market segment) and those who hate it (i.e., those who are not a part of that segment), the variance of ratings is high [52]. For an interested customer, this might be a good indicator.

In Sun's [52] theory, which we later refer to as a *variance theory*, she indicates that a variance of ratings, together with an average of ratings, can be used to identify niche products. That is, a high average value of ratings indicates a

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

good overall quality of product, so most of the customers consider the application to be very good, and therefore the variance is low. The high variance of ratings indicates a niche product, where the average of ratings is in the middle range, since the small number of customers gives conflicting reviews. These two variables together should indicate a good-quality niche product. In other words, if the overall quality of a product is good and meets the needs and expectations of the majority of customers, its variance is low and its valence is positive. Similarly, a niche product of good quality has a middle-range valence but high variance. Due to the interlinked nature of the variables, there cannot be a product that has both high valence and high variance.

Sun [52] found support for her theory in the book sales of Amazon and Barnes & Noble. However, it has been noted that more evidence is needed to verify this theory [26]. In this paper, we use the product of standard deviation (i.e., a square root of variance) and valence as an operationalization (i.e., standard deviation times valence) of the theory: for good-quality mainstream products and niche products, the product of these two variables is high. Furthermore, the product of these two variables for low-quality niche or mainstream products is low. This operationalization is the same that was used in the original study [52]. Thus, we formulate our hypothesis as follows:

H3: A high variance of ratings, together with a reasonably high average of ratings, correlates with sales improvements.

2.2.4 Moderating Role of Price

Several e-WOM studies have addressed different moderating variables that affect the effectiveness of e-WOM; for example, the effect of e-WOM may differ by product type [41], across products in the same category [60], or due to the credibility of the website [48]. However, to the best of the author's knowledge, the moderating role of price in the effectiveness of e-WOM has not been addressed in the previous studies.

In micro-pricing-with stakes of relatively low, nominal value- we may assume that risk-aversion bias [39] does not emerge and instead consumers may spend their money with no expectation of return. As the majority of the applications available in the marketplace are either free or cost less than one euro, price would impact the effectiveness of customers' ratings in the marketplace. In other words, we assume that when the price of an application increases, the customer spends more time to study the product and its reviews. When the product is reasonably cheap, the customer is more willing to save her/his time and take a risk with the application. Therefore, we formulate our hypotheses by assuming that for the cheap products, the impact of e-WOM is weaker:

H4a: A high average of ratings correlates more strongly with sales improvements when the price of the product increases.

H4b: A high volume of ratings correlates more strongly with sales improvements when the price of the product increases.

H4c: A high variance of ratings, together with a reasonably high average of ratings, correlates more strongly with sales improvements when the price of the product increases.

In the following chapter, we will discuss the previous e-WOM research in the mobile application ecosystems. In addition, we will present the operationalization of the measurement and hypotheses testing.

3 Research Context and Methodology

We will first discuss the domain of this study, the mobile application ecosystems. This is followed by a discussion of the research process and the variables included in the study.

3.1 Research Context: Mobile Application Marketplaces

Although mobile applications and their stores have been available for a while, the launch of Apple's App Store in 2008 quickened the development of-or arguably created- the new industry. Inspired by the success of the new "App Economy" [45], several large companies, such as Google, Microsoft, and Research in Motion followed Apple by publishing new mobile application marketplaces. The industry will, according to the ABI Research [1], reach US$ 25 billion total revenue in 2013.

While the publication process of an application from the developer's point of view varies between the different marketplaces [8], [23], the use of application marketplaces by the customers is rather standardized for all major ecosystems. A user can view and install the applications either on her/his smartphone or via a web browser; the application will be installed on-the-fly; and the user can, after installation, write a verbal review or give a star rating for the application. None of the major ecosystems actively seeks feedback from customers; on the contrary, leaving feedback is made rather difficult as the user has to, for instance, find the application in the marketplace before she/he can offer her/his judgment.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

In Google Play, for example, a user who has installed the application can leave a rating for the application. The marketplace uses a star-based review system where a user can give feedback for the product with positive stars, ranging from one to five. Google Play counts an average of all stars and shows the average star rating in the store. While there is no pre-selected rating for a user in this system, the previous studies have shown that the star-based rating leads to a J-shaped distribution [36].

In [56], Yan and Chen remark that rating in a mobile application marketplace requires laborious handwork, and thus reviews will be sparse and potentially lacking. In addition to star-rating, a user can write verbal feedback, which is then shown in the marketplace. At the time when the data for this study was collected, the marketplace did not filter the feedback in any way. However, the store was renewed in the autumn of 2013 and it can now show allocated feedback, based on, e.g., the device model of the reviewing user, to the consumer studying the application.

Many applications on the marketplace apply the freemium revenue model (see, e.g., [4], [50], [54]), in which users can experiment with the application for free and then get extra features or use the application after the tryout period expires for a fee. The freemium model can be a successful monetization strategy [38], [43]. This, together with micro-prices (applications that cost less than one euro), provides an intriguing setting for a e-WOM study. It is quite common with micro-priced applications for a customer to experiment with an application by herself/himself instead of spending her/his time reading others' reviews. Previously, Liu et al. [43] have shown that user reviews were not significant when a free version was offered in Google Play.

Previous work on user ratings in the online application marketplaces has contended that user ratings constitute a fundamental element in these marketplaces. It has been argued [34] that-*the app market where ratings play a central role in determining the consumer's ex ante perceived net utility as well as their willingness to pay*. Furthermore, Apple has stated in their iOS Developer (Site 1) guidelines that *[c]ustomer ratings and reviews on the App Store can have a big effect on the success of your app*-.

Although user ratings as such have been studied to a great extent, in the context of application or software marketplaces the existing literature is significantly thinner. A theoretical framework to assess the importance of ratings in the mobile application marketplaces was presented in [33]-[34]. Carare [9] showed that a bestselling rank has an effect as a determinant of demand. The results seem to indicate that the user ratings are not as crucial as visibility in the top list. The content of review comments in Google Play, focusing specifically on privacy and security issues, were analyzed in [32]. It was found that most of the comments are about the general quality of the application, and only a few concerned security issues.

The correlation between the paid application download categories and the average ratings in Google Play was studied in [38]. They assumed that, in free-to-install applications, user ratings are less relevant as it is easier to try the application than read the review comments. They found a small but statistically significant negative correlation between the number of downloads and the application's average rating. However, their dataset presented a static situation, and thus it cannot be estimated if the rating had an effect on the growth in popularity of an application.

3.2 Research Process and Method

We address the effects of e-WOM by gathering a large set of application data from Google Play (Site 2). This marketplace was chosen for the following three reasons: 1) availability of data, 2) the vast number of applications offered, and 3) its variety of different kinds of applications. Unlike several other digital marketplaces, Google Play shows a rough download category for every application. These characteristics allowed us to assess different aspects of e-WOM.

We used a web crawler (see, e.g., [10], [46]), implemented with the Python programming language and the Scrapy web scraping framework (Site 3), to gather data. The crawler started from the marketplace's front page and stored all links available from the page to a queue. From the queue, the crawler gets an address for a new page; if the page contains application information, the crawler stores this information in a NoSQL database. In every case, the crawler will go through all links available on the page and store them in the queue. If the links are already available in the data structure, they will be omitted to prevent the storage of duplicates. The crawler program will continue until there are no new pages to visit in the queue.

The data collection was repeated twice: once in February 2013 and once in May 2013. These sets were then exported from the database and stored in CSV files. In addition to these two datasets, we use a dataset collected in December 2011; however, this dataset was collected with a different data acquisition platform and it does not contain vote-specific information. The data-gathering process, utilized for the old set, differs only in the collection of application identifiers. The new system gathers them directly from the marketplace whereas the old system collected them from a third-party listing. See [37], [38] for more details on how the December 2011 dataset was gathered.

These three datasets allow us to study the impact of time on the effectiveness of e-WOM. In order to maximize the possibility of variance in the results, we decided to select two time windows: a long time span and a short time span. Based on an approximately 18-month period from December 2011 to May 2013, we are able to study the long-term effect of the e-WOM. The data calculated from these two datasets are referred to as the long time period data. As we

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

are not aware of a typical lifecycle of new generation mobile applications, we tested a few different options for the short time span. In the end, we decided to use a three-month period in order to show the relationship between e-WOM and sales while still maintaining a relatively short time span. Respectively, the data calculated from the datasets in February 2013 and May 2013 are referred to as the short time period data.

All datasets were imported to Microsoft Excel 2010 for visual examination and to calculate variables that were not offered directly at the marketplace. A Python script was implemented to combine the information of individual applications from several files via unique package names. Finally, the data file was imported in IBM SPSS Statistics 21 statistical software for analysis.

In this study, we use the following parsed and derived variables:

- Volume of ratings implies the number of people who have rated the application. A user can review the application with star ratings from one to five stars. In addition to the number of stars given, the user can also write a verbal comment about the application. The volume of the ratings is parsed from the marketplace.

- Valence of ratings is a value calculated and published by Google. As described by Google, this represents an arithmetic mean calculated from the star-based ratings. It should be noted that using an arithmetic mean as an average measure of a Likert-type value is challenging. However, we use the value as a proxy for a consumer rating, as this is the value that a consumer sees in the marketplace, simultaneously being aware of this limitation due to biased voting behavior (see, e.g., [36]) and its lack of rigorous statistical justification. This variable is parsed from the marketplace and it varies in the range of 1 to 5.

- Variance of ratings presents a distribution of ratings for a single application in the marketplace. The variable is calculated from the parsed data. We assume the star rating to be Likert-type categorical scale data. That is, we calculated the proportion of votes for each star class and, for these proportions, we then calculated variances and standard deviations.

- Product of an average rating and a standard deviation is used as a proxy when analyzing the variance theory. A relatively high average rating, together with a high variance, implies a niche product [52]. In this study, we used the product of a standard deviation and an average of ratings as a variable -similarly to the original work by Sun [52].

- Installation category illustrates the number of times an application has been downloaded and installed to a single device, divided in rough categories, published by the marketplace. The number of installations is defined as total installs per unique devices or users (Site 4). The published categories are on the half-logarithmic scale, i.e., 1-5, 5-10, 10-50, 50-100, etc. The values are coded, with numbers starting at 1. The highest category value in our dataset is 500,000,000-1,000,000,000 (coded with 19) by Google Play Service. A missing value is interpreted as zero downloads.

- Change in an installation category denotes differences between two installation category values. The variable is ordinal, as we calculate the number of steps that an application's installation category has advanced during the study period. For example, when an application installation category changes from 1-5 (coded with 1) to 10-50 (coded with 3), the value of this variable is 2.

- Median value of an installation category is used as a rough approximation of the actual number of installations for each application. The value is calculated from the installation category variable by taking the median from its extreme values, e.g., 75 for the download category 50-100 and 300 for the category 100-500.

- Change in median values of installation categories represents the sales increase between two measurement points. The value is calculated by subtracting a newer median value of an installation category from the older value. For example, the value 225 is used for an application in which the installation category changes from 50-100 to 100-500.

- Price of an application is parsed from the marketplace. The currency used is the euro.

Table 2 clarifies the used variables with a few example applications. It should be noted that the price may vary with time, as it is defined by the developer. We noted 11,657 cases in our dataset where the price was changed. We used the highest value for the price, as the lower ones possibly reflect an unusual situation, for instance, a seasonal sale.

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 10 / ISSUE 2 / MAY 2015 / 1-18

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

Table 2: Examples of variables used in this study

Package name	Price	Volume	Variance	Valence	Installation category in dataset Dec2011	Installation category in dataset May2013	Change in installation category	Product of rating and deviation
com. rovio. angry birds	0	618,597	0.767	4.6	10M-50M	50M-100M	+1	4.1
com. autoniq. vin scanner	0	346	2.740	3.8	10k-50k	50k-100k	+1	6.3
com. tinyredcloud. ma2003. v1	4.54	26	0.03	1	100-500	500-1.000	+1	0.2

In this study, we use Spearman's rank correlation coefficient in the hypotheses analysis. Spearman's rank correlation enables us to describe the relationship of the variables without making any assumptions regarding the frequency distribution or linear relationship of the variables.

4 Results

In this section, we will first describe the used datasets and their characteristics. The second subsection shows the results from the classical e-WOM dimensions, i.e., valence and volume of ratings. Subsection 4.3 tests the recent variance of ratings theory in the mobile application ecosystem. The effect of pricing on the relationship between e-WOM and sales is discussed in Subsection 4.4. The effect of time is analyzed as a cross-sectional aspect in the different dimensions.

4.1 Descriptive Statistics

Initially, we had the following three datasets: December 2011, February 2013, and May 2013. When duplicates were removed, there were 803,164 unique applications. The final dataset includes 316,965 applications that were present in at least two datasets. The used datasets, referred to here as *Long* and *Short*, are clarified in Figure 1.

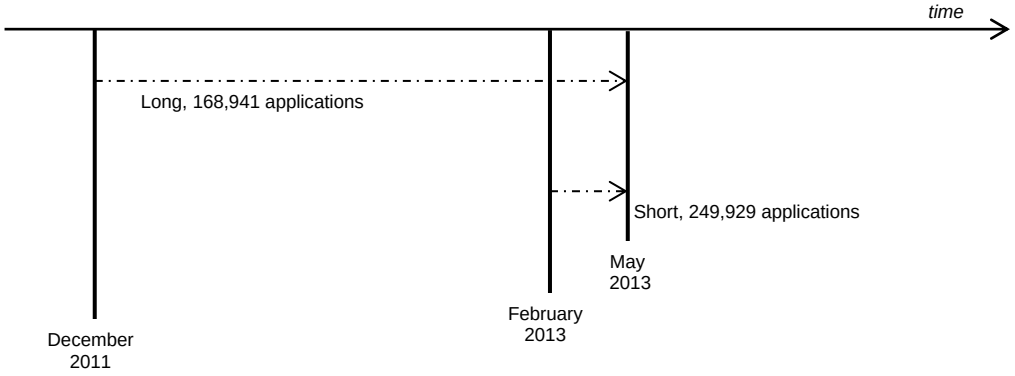


Figure 1: The used datasets on a timeline and the number of applications in both sets

Figure 2 illustrates the number of applications for which installation categories (i.e., popularity) have changed, measured in both short and long study periods. For instance, those applications that were initially in installation category number 7 (1,000-5,000 downloads) have increased the most often in their popularity (more than 21,000 applications). In other words, their installation category has increased in that period. For the most part, the shapes of these two histograms are rather similar; however, less popular applications (lowest installation categories) changed their categories more often in the long study period. Furthermore, it should be noted that for 77.7% of the applications in the short study period and 37.0% of the applications in the long study period, the installation category did not change during the measurement period.

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 10 / ISSUE 2 / MAY 2015 / 1-18

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

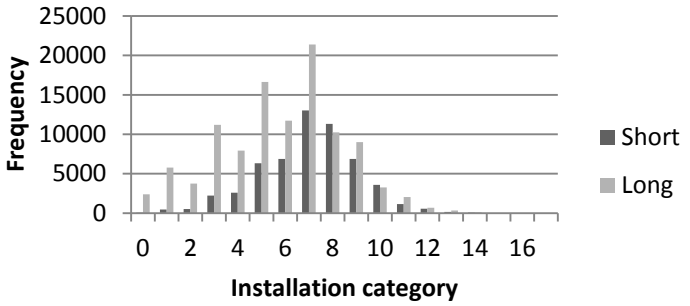


Figure 2: Histogram of applications for which installation category changed

We noted some anomalies in the final dataset. Firstly, for 6,126 applications, the volume of ratings decreased during the short time period. The reasons for the lost votes are unknown; however, it might be the result of the removal of inappropriate reviews and ratings, the deletion of reviewer accounts, or the withdrawal and re-launch of the application in the marketplace. Furthermore, it is possible that there are more applications that have lost votes but have received the same amount or more votes during the measurement period, and therefore we were not able to notice them. Therefore, these applications with anomalies are also still included in our analysis. Nevertheless, the number of applications with anomalies is low. Secondly, the datasets contain many (e.g., 112,826 in December 2011 and 141,524 in May 2013) applications without any ratings which are also included in the analysis. Ratings may be missing because applications are new and either no one has rated them yet or no one was interested enough to rate them. Thirdly, the ratings left in the marketplace are highly skewed toward the highest values. For example, in the original May 2013 dataset, of the more than 261.26 million ratings, about 67.36% are the highest possible rating (five stars). From all ratings in this dataset, 17.47% are four stars, 6.78% three stars, 2.19% two stars, and only 6.20% one star. The distribution follows the J-shaped curve often seen in other product review systems (see, e.g., [36]). This might indicate that the users only rate the applications when they are extremely satisfied.

4.2 The Effect of the Valence and Volume

For the classical dimensions, we examine three hypotheses (1, 2a and 2b): Do valence, volume, or previous popularity correlate with sales? In this analysis, we use the valence value of the first measurement point as a representative value for the observed time span. We use the change in median values in installation categories as a proxy of the sales. We studied the correlations for both the long and short time period.

We noted that several applications' average ratings varied a lot between the measurement points. Usually, these applications have a high valence value at the beginning when they have been rated by only a few users; however, the value drops quickly when more users have installed and reviewed the application. Therefore, we decided to include in this calculation only applications that have more than five ratings at the beginning of the measurement period. When examining the data, we noted that the applications' change in valence was a bit more stable when they have more than five votes.

The results of the analysis are presented in Table 3. The table presents Spearman's ρ correlation coefficients for download category, number of ratings, average ratings, and the change in median values in installation categories for both the short and long study periods for the applications that have at least five votes in the starting point of measurement. On the one hand, the results imply that the valence of ratings correlates positively with sales with both a long (ϱ [85,228] = 0.152, $p < 0.001$) and a short time period (ϱ [202,360] = 0.118, $p < 0.001$). In other words, a high (and low) average of star ratings correlates with high (and low, respectively) sales. Therefore, Hypothesis 1 seems to hold.

Table 3: Correlation coefficients for valence, volume, and the sales for both the short and long study periodst

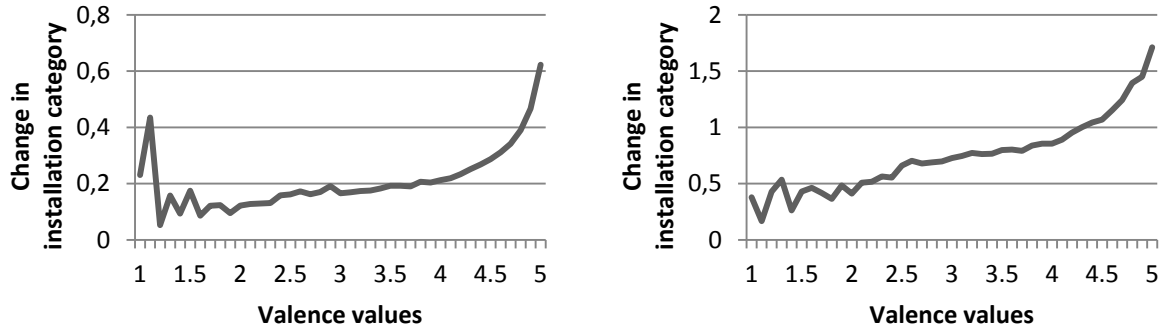
		Long	Short
Valence	coeff.	0.152***	0.118***
	N	85,228	202,360
Volume	coeff.	0.147***	-0.073***
	N	85,228	202,360
Installation Category	coeff.	0.194***	-0.142***
	N	85,228	202,360

***. Correlation is significant at the 0.001 level (2-tailed).

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 10 / ISSUE 2 / MAY 2015 / 1-18

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

This effect can be examined visually in Figure 3, which illustrates the averages of the installation category changes for each valence value for both the long (Figure 3a) and short (Figure 3b) time periods. Only those applications that had more than five votes at the first measurement point have been included. Figures show that the average of the installation category changes is considerably higher for applications with high valence values. However, the pattern for applications with extremely low valence values is unstable. This might be the result of highly skewed rating behavior in which low star ratings are rarely given and, as a result, there are only a few applications with the lowest valence values.



a) Long b) Short

Figure 3: Averages of installation category changes for each valence value

The volume of the ratings correlates positively with sales for the long time period, in which we found a statistically significant positive correlation (ϱ [85,228] = 0.147, p < 0.001). For the short time period, the correlation is rather small and negative (ϱ [202,360] = -0.073, p < 0.001). Similarly, the previous popularity correlates positively with the sales in the long time period (ϱ [85,228] = 0.194, p < 0.001) but not for the short period (ϱ [85,228] = -0.142, p < 0.001). Therefore, hypotheses 2a and 2b receive support only for the long time period but not for the short one.

We assume that for the long period, the positive correlation of previous popularity might be explained with the growing number of Android devices – and therefore the new installations of previously popular applications. In the short period, previous popularity, measured either by volume of ratings or previous number of installations, is negatively correlated with the sales. This might be a result of the users installing new highly rated applications. Unfortunately, our dataset is insufficient to test this assumption, but it should be analyzed in future studies when the number of devices begins to stabilize.

It also seems that the volume of ratings represents the popularity of an application only. These variables are highly correlated: ϱ [88,459] = 0.790 (p < 0.001) for a long period and ϱ [206,438] = 0.818, (p < 0.001) for a short period. This is expected as the marketplace requires that a user must install an application to rate it.

4.3 The Effect of Variance

As the earliest measurement point does not contain vote-specific information, we are able to calculate the impact of variance only for the short time period. The calculations are divided into two parts. First, we included all applications that have more than two votes, i.e., those applications for which a variance value can be calculated. In the second part, we study applications that have more than five votes to get a comparable result with the other dimensions. The results are shown in Table 4. The table presents Spearman's ρ correlation coefficients for variance of ratings and the product of standard deviation for applications that have more than two ratings and for applications that have more than five ratings

Table 4: Spearman's ρ correlation coefficients for variance of ratings and the product of standard deviation

		Volume > 2	**Volume > 5**
Variance	coeff.	-0.068***	-0.069***
	N	224,176	202,356
Product of standard deviation and valence	coeff.	0.014***	0.008***
	N	224,178	202,360

***. Correlation is significant at the 0.001 level (2-tailed).

The results show that Hypotheses 3 is supported, as the product of the average of ratings and the standard deviation correlates weakly with the sales (ϱ [224,176] = 0.014, p < 0.001) when all applications for which a variance can be

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

calculated are included. However, the variance of ratings itself correlates negatively with sales (ϱ [224,176] = -0.068, $p < 0.001$) in this case and therefore the operationalization can be justified. Nevertheless, it should be noted that the correlation coefficients of the variance of ratings and the operationalization used are extremely small and close to zero.

4.4 The Effect of Pricing

To test the impact of micro-prices on the relationship between e-WOM and sales, we divided applications into four categories based on price. We calculated the correlation coefficients for each category. The first one contains all free applications, and the other three categories each contain approximately one-third of the applications that are subject to charge before installation. Spearman's correlation coefficients for the different dimensions of e-WOM and sales for both the long and short time spans are presented in Table 5, grouped by the price categories. For the variance dimension, we used the product of standard deviation and valence as the operationalization. Furthermore, only the result from the short time period is presented for this dimension due to the lack of long time span data. In addition, only applications that have more than two ratings in the marketplace are included into the study of the variance dimension.

In the analysis, we only focus on the applications that are subject to a charge before installation. The free applications use a multitude of different monetization strategies, from advertisements to an in-application payment; as a result, their actual price for a consumer cannot be evaluated directly and they are thus not included in the analysis. The results from the applications that are subject to a charge for the long time period seem to support Hypothesis 4a-b, as the correlation coefficients grow constantly with the price group for both valence and volume. That is, e-WOM seems to be more effective for costlier applications. However, the results from the short time period are controversial as the correlation coefficient varies almost randomly with the price groups.

Table 5: Spearman's ρ correlation coefficients for e-WOM's dimensions and sales

			Valence		Volume		Variance
Price (€)			Long	Short	Long	Short	Short
Free	0.00	coeff.	0.158***	0.118***	0.231***	-0.077***	0.008**
		N	116,784	215,593	116,784	215,593	198,376
Applications subject to a charge	0.01–0.81	coeff.	0.121***	0.116***	0.088***	-0.033**	-0.026*
		N	18,505	10,353	18,505	10,353	7,403
	0.82–1.50	coeff.	0.165***	0.146***	0.138***	-0.046***	-0.060***
		N	17,791	11,273	17,791	11,273	8,339
	1.51–	coeff.	0.197***	0.137***	0.199***	-0.031***	-0.047***
		N	15,861	12,710	15,861	12,710	10,060

***. Correlation is significant at the 0.001 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

We used Cohen and Cohen's [21] (pp. 53-55) test to study differences between all of the independent correlation coefficients pairs of the different price groups in the long time period. The values are calculated with [49] and the results for valence and volume are shown in Table 6. The Bonferroni correction is used to correct the p-values for multiple comparisons. The table shows that differences between the price groups' correlation coefficients in the long time period are statistically significant. The same test was performed for the correlation coefficients of the short time period; however, the results were not statistically significant.

Table 6: The differences between independent price groups' correlation coefficients of the long time period data

	Valence			Volume		
Price categ.	**0.01-0.81 €**	**0.82-1.5 €**	**1.51-€**	**0.01-0.81 €**	**0.82-1.5 €**	**1.51- €**
0.01-0.81 €	-	4.278***	7.209***	-	4.814***	10.485***
0.82-1.5 €		-	3.030**		-	5.751***
1.51-€			-			-

***. Correlation is significant at the 0.001 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

From Table 6, we can confirm that a high average of ratings correlates more strongly with sales improvements when the price of the product increases since all results are greater than |1.96| and correlation coefficients are statistically significant (see [21]). The test shows that in the long time period data, the differences of correlation coefficients are

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

statistically significant. The study of correlation coefficients shows that the relationships between the sales and valence value and between the sales and volume value strengthen when the price of an application increases.

5 Discussion

In this section we discuss our major findings on e-WOM's relationship with sales in the mobile application marketplace. We address the following three research questions: the relationships of different e-WOM dimensions, and the effect of time and price. The results of the four hypotheses are presented in Table 7.

Summarizing our results regarding the first research question, we found significant correlations between different dimensions of e-WOM and sales in the mobile application marketplace in the long time period. Therefore, e-WOM might act as a predictor of future sales. While all studied correlations are statistically significant because of the large data set, the correlation coefficients are rather small and thus none of them seem to work as a good prediction tool for sales. There might be latent variables that can explain sales better than the studied variables. However, the previous popularity, measured by the total number of installations, has the highest correlation coefficients among studied variables in the dataset of the long time period. In other words, the previous popularity seems to still work as the best indicator of future sales.

Table 7: A summary of our hypotheses and their support

Hypotheses	**Long**	**Short**
H1: A high valence correlates positively with sales	Supported	Supported
H2a: A high volume of ratings correlates positively with sales	Supported	Not supported
H2b: A high number of installations correlates positively with sales	Supported	Not supported
H3: A high variance of ratings, together with a reasonably high average of ratings, correlates with sales	Not available	Supported
H4a: A high average of ratings correlates better with sales when the price of the product increases	Supported	Not supported
H4b: A high volume of ratings correlates better with sales when the price of the product grows	Supported	Not supported
H4c: A high variance of ratings, together with a reasonably high average of ratings, correlates more strongly with sales improvements when the price of the product increases	Not available	Not supported

The operationalization of Sun's [52] variance theory was only weakly supported while the raw variance value of the ratings has a negative correlation coefficient. This suggests that Sun's variance theory might capture some aspects of e-WOM. However, the correlation coefficients, although they are statistically significant, are near to zero in all cases. We studied in detail a dozen randomly selected applications with high average and high variance. Even though our analysis was based on numeric ratings, a brief analysis of the reviews written by the users indicates that most of these applications are receiving contradicting reviews; some of the comments were strongly negative, either based on technical problems or the uselessness of the product, while others praised the product's features. Interestingly, we found one application that had only negative verbal comments, but still more than half of the users (n = 97) had awarded five stars. We would assume that these reviews could have been done by, e.g., developers themselves, their close associates, or even by an outsourced review service-or by users with a strange sense of humor.

The second research question assessed the impact of time in the relationship between e-WOM and sales. Surprisingly, the results of the analysis differ substantially when we compare the short and long time period datasets. Only valence as an indicator of future sales was supported in both periods. In the short time span, the previous popularity or volume of reviews were not found to positively correlate with the sales. This might be the result of an ever-growing number of Android devices-and, thus, installations- in the marketplace, or just due to selection of the particular examination period. However, our study only covered two time windows and further work is needed to verify our results in other domains and to study if there are other patterns of changes beyond, or between, these time spans.

The third research question examined if the price impacts the relationship between e-WOM and sales. The majority of the applications available in this marketplace are either free or cheap. From the descriptive analysis of correlation coefficients in different price groups of the long time period dataset, we observed a small but constant increase when the price increases. This result is in line with previous studies on micro-prices and suggests that consumers are more willing to take a risk and try an application without reading the peer reviews when the price is low enough [55]. In our short period dataset this result was not found, which might be due to the selection of time period (data) or because the effect emerges slower and was therefore not yet identifiable in the data. Furthermore, early adopters

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

[47], [51] may be active in the short period whereas the majority of potential users make their moves (installations) later and thus a three-month examination period was too short to detect this behavior.

6 Conclusions

We studied common myths linked to the relationship between user ratings and product sales in the mobile application ecosystem. In this test, we used two measurement periods, a long one with a timeframe of approximately 18 months and a short one with a time period of three months. The used dataset contains over 300,000 unique applications that were present in at least two measurement points. The effect of the valence of ratings is confirmed in both periods. The volume of ratings remains as a plausible explanation. It was confirmed for the long dataset but busted for the short one. We showed that the valence of ratings seems to improve the sales of costlier products more than cheap ones in a long time period. Furthermore, we showed that the choice of time period is important in e-WOM studies.

6.1 Theoretical and Managerial Implications and Suggestions for Future Research

Our contribution is both to e-WOM and e-commerce research as we empirically investigated if customer ratings have an effect on sales. By analyzing a large dataset in terms of the volume, variance, and valence of ratings, we were able to extend current knowledge on the effects of e-WOM on sales (e.g., [24], [41], [52], [60]). Our result shows that the simple models of e-WOM seem to somewhat explain the phenomena; however, more complex theories, such as the variance of ratings, explain some partition of sales in the marketplace better than the classical dimensions, such as the volume and valence of reviews. Although there is a statistically significant correlation between high valence and sales, there are many different factors, most of which are outside of this study and the dataset used, that affect the outcome. For instance, in the most installed applications, i.e., the superstars of the ecosystem, there seem to be many negative and positive reviews, thus lowering the overall average of the ratings, while applications with the highest averages have been installed only a few times. As a result, we suggest that the use of ratings as a measure of quality or an indicator of future sales is a more complex issue than it has been considered to be and thus requires more focused operationalization and detailed research designs. In addition, the users' rating conventions-almost two-thirds of ratings were five stars- are hampering the usefulness of the reviews. In further studies, this skew and its potential effect with other studies of consumer/user-based assessments, e.g., in the fields of psychology or marketing, should be studied.

We assumed that the specific features, i.e., micro-pricing and fast delivery, of the mobile application ecosystems would cause consumer ratings to be less important than in online marketplaces for tangible products. Our assumption is supported, as e-WOM seems to be more important when the price of an application increases. That is, when a consumer is buying a new computer, she/he invests a considerable amount of money in that purchase. In order to avoid a wrong decision, she/he probably browses through review information, asks for references, sees ads, etc. A similar approach fits with many tangible products, such as books and movies, and with expensive intangible services that have been well studied. However, the characteristics of the mobile application ecosystems favor a simpler decision-making process during purchasing. Most of the applications are fairly cheap and even offer free trial versions that lower the buying barrier even more. However, further research would be needed to clarify the psychological processes related to micro-prices and intangible products, even including potential cultural differences regarding how people consider these factors. Furthermore, this paper is among the first that has investigated if the price of the product impacts the relationship between e-WOM and sales; further work is needed to replicate the results in other domains.

In a managerial sense, our result implies that the ratings are more complex than they seem to be at first sight. Although we found a statistically significant correlation, the high valence does not guarantee success. We found 24,696 (14.6% of the long-term dataset's applications) applications whose installation category did not change during the study period of 18 months, although they had an average rating of over four stars. Furthermore, the rating mechanisms are not completely open and trustworthy. One clear factor affecting the results and overall usefulness of the user ratings is that the marketplaces rarely ask users to review an application. The marketplace ensures that the reviewing users have installed the application before they can give the review. Therefore, the users are likely to evaluate applications only when they are either extremely disappointed or satisfied. Even more importantly, previous studies have shown that the effect of negative e-WOM is stronger than positive e-WOM. However, our data show that there is a strong bias toward taking review action only when a user is extremely satisfied. Furthermore, some applications have a built-in feature that occasionally reminds users to review the application. Some of these applications even filter users so that those with positive feedback are directed to the marketplace's rating service, while the users with negative feedback are directed to the application vendor's feedback page. In other words, the application producers aim to allow only positive feedback on the official forum, collecting all negative feedback in their own service. These kinds of behaviors distort data and decrease the usefulness of reviews as an indicator of quality but make it clear that the reviews are seen as having practical value.

Furthermore, the swift nature of the mobile applications-i.e., most of them are designed for only a short time interaction, and an user might install several applications, test them all, and decide to use only one and remove the res- in contrast to the tedious work of reviewing the application in the marketplace, might cause user reviews to be

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

sparse and uninformative. We saw this phenomenon when reviewing the verbal comments of users. Most of the reviews were only a few words long and focused on either praising the goodness or bashing the problems of the application. We did not find any long reviews on the pros and cons of a product, although these kinds of comments are quite common in other online marketplaces, such as Amazon. When we studied the niche applications, we noted that some of the comments might be more sarcastic than really helpful reviews.

In summary, although consumer reviews are important to a software developer as a quick feedback channel, we would argue that perhaps using time to optimize the average of ratings does not pay off. Furthermore, valence as an indicator of future sales might be questionable. We found a statistically significant correlation between a high valence and high sales; however, the coefficient is small and, most likely, some latent variable, e.g., visibility in social media or brand value, would explain the sales better. It is also possible that some other characteristics outside of this study, e.g., the high quality of an application, cause both high valence and high sales. We also showed that previous popularity has the highest correlation coefficients in the long run. Despite this, ratings are still a direct communication channel with the consumers, and the feedback is therefore valuable information to help the application developers improve the product. Finally, for the ecosystem orchestrators, our result implies that there are problems with the current rating system; a better rating mechanism for the digital marketplace should be developed.

6.2 Limitations

The wealth of data available in the World Wide Web enables researchers to access a plethora of data, but this comes with significant limitations. With web crawling we can only control the process of data gathering, but have no control over the information published online. To a certain extent we are required to take the data as is, but we have identified several factors that might influence our results. First, it is possible that some applications are filtered out due to the location of the computer from which the data collection was done. That is, the dataset contains the applications published for Finland's market, although the reviews are global, and thus the result might reflect some culture-specific features. Second, despite our best effort, during data collection we might have missed some applications that were presented in a previous data gathering point. Regardless of these issues, we have gathered a large dataset that represents a remarkable partition of data available in the marketplace. Third, the representativeness of the sample is most likely biased toward the popular applications. The data collection method used likely overemphasizes those applications that have been installed often, and emphasizes those applications that have been installed only a few times (or not at all) less. However, similarly the customers of the marketplace are more biased toward the popular applications than they are the vast majority of applications. Fourth, it should also be noted that data contain lots of *noise*; e.g., applications that are launched as a hobby and applications meant only for a small group of users. These applications most likely will not fit the presented theorems that suppose each application is meant for commercial use. Future work should develop and suggest guidelines on ways that research should deal with this *noise*. Fifth, it is not known how reliable ratings information or ratings systems are in this domain (see, e.g., [17], [27] for a discussion on the importance of credibility in e-WOM). Finally, this paper deals only with one mobile application ecosystem. It remains to be seen to what extent other mobile ecosystems contain similar structures or mechanisms.

Acknowledgments

The authors wish to express their gratitude to B.Sc. (Tech.) Miika Oja-Nisula for his contributions in the implementation of the data-gathering platform; Mrs. Satu-Päivi Kantola for her guidance in statistical analysis; and Dr. (Tech) Tuomas Mäkilä and Ph.D. Hongxiu Li for their comments on the early version of the manuscript. In addition, Sami Hyrynsalmi wishes to express his gratitude for the research grants given by the Nokia Foundation for his dissertation work on mobile ecosystems.

Websites List

Site 1: App Store Submission Tutorial
https://developer.apple.com/library/ios/navigation/

Site 2: Google Play
http://play.google.com

Site 3: Scrapy – An open source web-scraping framework for Python
http://www.scrapy.org

Site 4: Android Developers
http://developer.android.com/

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

References

[1] ABI Research, Tablets Will Generate 35% of this Year's 25 Billion App Revenue; Expected to Surpass Smartphones by 2018. London, United Kingdom: Press release, 2013.
[2] S. S. Alam and N. M. Yasin, What factors influence online brand trust: evidence from online tickets buyers in Malaysia, Journal of Theoretical and Applied Electronic Commerce Research, vol. 5, no. 3, pp.78-89, 2010.
[3] N. Amblee and T. Bui, Harnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital microproducts, International Journal of Electronic Commerce, vol. 16, no. 2, pp. 91-114, 2011.
[4] C. Anderson, Free: The Future of a Radical Innovation. New York, NY, USA: Hyperion, 1st edition, 2009.
[5] S. Balasubramanian and V. Mahajan, The economic leverage of the virtual community, International Journal of Electronic Commerce, vol. 5, no. 3, pp.103-138, 2001.
[6] J. Berger, A. T. Sorensen and S. J. Rasmussen, Positive effects of negative publicity: When negative reviews increase sales, Marketing Science, vol. 29, no. 5, pp. 815-827, 2010.
[7] B. Bickart and R. M. Schindler, Internet forums as influential sources of consumer information, Journal of Interactive Marketing, vol. 15, no. 3, pp. 31-40, 2001.
[8] P. R. J. Campbell and F. Ahmed, An assessment of mobile os-centric ecosystems, Journal of Theoretical and Applied Electronic Commerce Research, vol. 6, no. 2, pp. 50-62, 2011.
[9] O. Carare, The impact of bestseller rank on demand: Evidence from the app market, International Economic Review, vol. 53, no. 3, pp. 717-742, 2012.
[10] C. Castillo, Effective web crawling, Ph.D. dissertation, Department of Computer Science, University of Chile, Santiago, Chile, 2004.
[11] Y. Y. Chan and E. Ngai, Conceptualising electronic word of mouth activity: An input-process-output perspective, Marketing Intelligence & Planning, vol. 29, no. 5, pp. 488-516, 2011.
[12] M. Chen and X. Liu. Predicting popularity of online distributed applications: iTunes app store case analysis, in Proceedings of the 2011 iConference, iConference '11, New York, 2011, pp. 661-663.
[13] P.-Y. Chen, S.-y. Wu and J. Yoon, The impact of online recommendations and consumer feedback on sales, in Proceedings of Twenty-Fifth International Conference on Information Systems (ICIS 2004), 2004, pp. 711-724.
[14] X. Cheng and M. Zhou, Study on effect of e-WOM: A literature review and suggestions for future research, in Proceedings International Conference on Management and Service Science (MASS), 2010, pp. 1-4.
[15] C. M. Cheung and M. K. Lee, What drives consumers to spread electronic word-of-mouth on online consumer-opinion platform, Decision Support Systems, vol. 53, no. 1, pp. 218-225, 2012.
[16] C. M. K. Cheung and D. R. Thadani, The impact of electronic word-of-mouth communication: A literature analysis and integrative model. Decision Support System, vol. 54, no. 1, pp. 461-470, 2012.
[17] M. Y. Cheung, C. Luo, C. L. Sia, and H. Chen, Credibility of electronic word-of-mouth informational and normative determinants of on-line consumer recommendations, International Journal of Electronic Commerce, vol. 13, no. 4, pp. 9-38, 2009.
[18] J. A. Chevalier and D. Mayzlin, The effect of word of mouth on sales: Online book reviews, Journal of Marketing Research, vol. XLIII, no. 3, pp. 345-354, 2006.
[19] P. K. Chintagunta, S. Gopinath and S. Venkataraman, The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets, Marketing Science, vol. 29, no. 5, pp. 944-957, 2010.
[20] E. K. Clemons, G. G. Gao and L. M. Hitt, When online reviews meet hyperdifferentiation: A study of the craft beer industry, Journal of Management Information Systems, vol. 23, no. 2, pp. 149-171, Fall 2006.
[21] J. Cohen and P. Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc., 1983.
[22] D. A. Colvin, Effect of social media as measured by the dispersion of electronic word-of-mouth on the sales success of experience goods: An empirical study of kindle book, Ph.D. dissertation, Falls School of Business Administration, Anderson University, Anderson, IN, USA, 2013.
[23] F. Cuadrado and J. Dueñas, Mobile application stores: Success factors, existing approaches, and future developments, IEEE ASSP Communications Magazine, vol. 50, no. 11, pp. 160-167, 2012.
[24] G. Cui, H.-K. Lui and X. Guo, The effect of online consumer reviews on new product sales, International Journal of Electronic Commerce, vol. 17, no. 1, pp. 39-58, 2012.
[25] T. Dalgic and M. Leeuw, Niche marketing revisited: Concept, applications and some european cases, European Journal of Marketing, vol. 28, no. 4, pp. 39-55, 1994.
[26] P. De Maeyer, Impact of online consumer reviews on sales and price strategies: A review and directions for future research, Journal of Product & Brand Management, vol. 21, no. 2, pp. 132-139, 2012.
[27] C. Dellarocas, The digitization of word of mouth: Promise and challenges of online feedback mechanisms, Management Science, vol. 49, no. 10, pp. 1407-1424, 2003.
[28] C. Dellarocas, X. M. Zhang and N. F. Awad, Exploring the value of online product reviews in forecasting sales: The case of motion pictures, Journal of Interactive Marketing, vol. 21, no. 4, pp. 23-45, 2007.
[29] W. Duan, B. Gu and A. B. Whinston, Do online reviews matter? - an empirical investigation of panel data, Decision Support Systems, vol. 45, no. 4, pp. 1007-1016, 2008.
[30] D. Godes and D. Mayzlin, Using online conversations to study word-of-mouth communication, Marketing Science, vol. 23, no. 4, pp. 545-560, 2004.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

[31] T. W. Gruen, T. Osmonbekov and A. J. Czaplewski, E-WOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty, Journal of Business Research, vol. 59, no. 4, pp. 449-456, 2006.
[32] E. Ha and D. Wagner, Do android users write about electric sheep? Examining consumer reviews in google play, in Proceedings of the 10th Annual IEEE Consumer Communication & Networking Conference, CCNC 2013, IEEE, Flamingo, Las Vegas, 2013, pp. 149-157.
[33] L. Hao, X. Li, Y. Tan, and J. Xu, The economic role of rating behavior in third-party application market, in Proceedings of Thirty Second International Conference on Information Systems, ICIS 2011, Shanghai, 2011, pp. 1-15.
[34] L. Hao, X. Li, Y. Tan, and J. Xu. (2011, July) The economic value of ratings in app market. Social Science Research Network. [Online]. Available: http://ssrn.com/abstract=1892584
[35] T. Hennig-Thurau, K. P. Gwinner, G. Walsh, and D. D. Gremler, Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? Journal of Interactive Marketing, vol. 18, no. 1, pp. 38-52, 2004.
[36] N. Hu, J. Zhang and P. A. Pavlou, Overcoming the j-shaped distribution of product reviews, Communications of ACM, vol. 52, no. 10, pp. 144-147, 2009.
[37] S. Hyrynsalmi, T. Mäkilä, A. Järvi, A. Suominen, M. Seppänen, and T. Knuutila, App store, marketplace, play! an analysis of multi-homing in mobile software ecosystems, in Proceedings of the Fourth International Workshop on Software Ecosystems, IWSECO'2012, MIT Sloan School of Management, CEUR-WS, Cambridge, MA, USA, , 2012, pp. 55-68.
[38] S. Hyrynsalmi, A. Suominen, T. Mäkilä, and T. Knuutila, The emerging mobile ecosystems: An introductory analysis of Android Market, in Proceedings the 21st International Conference on Management of Technology, Hsinchu, Taiwan, International Association for Management of Technology, March 2012, pp. x-x.
[39] D. Kahneman and A. Tversky, Prospect theory: An analysis of decision under risk, Econometrica, vol. 47, no. 2, pp. 263-291, 1979.
[40] B. Lang and K. F. Hyde, Word of mouth: What we know and what we have yet to learn, Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior, vol. 26, pp. 1-18, 2013.
[41] J. Lee, J.-N. Lee and H. Shin, The long tail or the short tail: The category-specific impact of e-WOM on sales distribution, Decision Support System, vol. 51, no. 3, pp. 466-479, 2011.
[42] A. Lindgreen, A. Dobele and J. Vanhamme, Word-of-mouth and viral marketing referrals: what do we know? and what should we know?, European Journal of Marketing, vol. 47, no. 7, pp. 1028-1033, 2013.
[43] Z. Liu, Y. A. Au and H. S. Choi. An empirical study of the freemium strategy for mobile apps: Evidence from the google play market, in, Proceedings of 33rd International Conference on Information Systems (ICIS2012), Orlando, 2012, pp. 1-17.
[44] Y. Liu, Word-of-mouth for movies: Its dynamics and impact on box office revenue, Journal of Marketing, vol. 70, no. 3, pp. 74-89, 2006.
[45] D. MacMillan, P. Burrows and S. E. Ante. (2009, October) Inside the app economy. Bloomberg Businessweek Magazine. [Online]. Available: http://www.businessweek.com/magazine/content/09_44/b4153044881892.htm
[46] C. Olston and M. Najork, Web crawling, Foundations and Trends in Information Retrieval, vol. 4, no. 3, pp. 175-246, 2010.
[47] J. R. Ortt and J. P. Schoormans, The pattern of development and diffusion of breakthrough communication technologies, European Journal of Innovation Management, vol. 7, no. 4, pp. 292-302, 2004.
[48] C. Park and T. M, Lee. Information direction, website reputation and eWOM effect: A moderating role of product type, Journal of Business Research, vol. 62, no. 1, pp. 61-67, 2009.
[49] K. J. Preacher. (2002, May) Calculation for the test of the difference between two independent correlation coefficients. Computer software. [Online]. Available: http://www.quantpsy.org
[50] Pulkkanen and M. Seppänen, Freemium business models in technology product markets, in Proceedings of the XXIII ISPIM Conference - Action for Innovation: Innovating from Experience, Barcelona, Spain, 2012, pp. 1-9.
[51] E. M. Rogers, Diffusion of Innovations. New York, NY, USA: Simon and Schuster, 2010.
[52] M. Sun, How does the variance of product ratings matter? Management Science, vol. 58, no. 4, pp. 696-707, 2012.
[53] E. Vermeulen and D. Seegers, Tried and tested: The impact of online hotel reviews on consumer consideration, Tourism Management, vol. 30, no. 1, pp. 123-127, 2009.
[54] T. M. Wagner, A. Benlian and T. Hess, The advertising effect of free - do free basic versions promote premium versions within the freemium business model of music services?, in Proceedings 46th Hawaii International Conference on System Sciences (HICSS), Hawaii, 2013, pp. 2928-2937.
[55] T. Yamabe, V. Lehdonvirta, H. Ito, H. Soma, H. Kimura, and T. Nakajima, Activity-based micro-pricing: realizing sustainable behavior changes through economic incentives, in Proceedings of the 5th International Conference on Persuasive Technology, PERSUASIVE'10, Berlin, Heidelberg, Springer-Verlag, 2010, pp. 193-04.
[56] B. Yan and G. Chen, Appjoy: Personalized mobile application discovery, in Proceedings of the 9th International Conference on Mobile systems, Applications and Services, MobiSys '11, New York, NY, USA, ACM, 2011, pp. 113-126, .
[57] Q. Ye, R. Law and B. Gu. The impact of online user reviews on hotel room sales, International Journal of Hospitality Management, vol. 28, no. 1, pp. 180-182, 2009.
[58] L. Zhang, B. Ma and D. K. Cartwright, The impact of online user reviews on cameras sales, European Journal of Marketing, vol. 47, no. 7, pp. 1115-1128, 2013.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762015000200002

[59] X. M. Zhang, Tapping into the pulse of the market: Essays on marketing implications of information flows, Ph.D. dissertation, Alfred P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.
[60] F. Zhu and X. M. Zhang, The impact of online consumer reviews on sales: The moderating role of product and consumer characteristics, Journal of Marketing, vol. 74, no. 2, pp. 133-148, 2010